BİLGİSAYAR BİLİMLERİ VE TEKNOLOJİLERİ DERGİSİ
JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGIES
http://dergipark.org.tr/tr/pub/bibted

Research Article

Clustering Hotels and Analyzing the Importance of Their Features by Machine Learning Techniques

Mert Akyol[*1]

[1]*Setur, R&D Department, Istanbul, Turkey*

Keywords:
Machine Learning,
KMeans Clustering,
Principal Component Analysis,
Elbow Method,
Random Forest

ABSTRACT

The firms which are specialized in hotel bookings generally have huge amounts of hotels with hundreds of features in their database. To be able to get the most meaningful insights from that data, it is vital to use the right machine learning techniques for segmenting those hotels into meaningful groups and finding their most important features. In this study, hotels data from Setur firm have been used for clustering, dimensionality reduction and feature selection analysis. Firstly, hotels were clustered by KMeans Clustering algorithm according to the similarity of their features. To see the effect of dimensionality reduction technique on the clustering process of hotels data, PCA(Principal Component Analysis) method was applied on hotels data and KMeans Clustering algorithm was applied to this processed data in order to observe the differences between the clustering results when PCA is applied and not applied. After that, multivariate and univariate feature selection techniques were applied to the clustered hotels data for identifying the most important features of hotels which have effect on clustering process. As a multivariate feature selection technique, Random Forest algorithm was used. For the univariate technique, SelectKBest algorithm with chi2 score function was used as a filter-based feature selection method.

Otellerin Kümelenmesinin ve Özelliklerinin Önem Derecelerinin Makine Öğrenmesi Teknikleri Kullanılarak Analiz Edilmesi

Anahtar Kelimeler:
Makine Öğrenmesi,
KMeans Kümeleme,
Temel Bileşenler Analizi,
Elbow Metodu,
Rastgele Orman

ÖZ

Büyük veri kavramı, otel rezervasyon sektöründe çalışan firmalar için çok yüksek sayıda farklı otel ve bu otellerin yüzlerce farklı özelliği olarak yer almaktadır. Firmaların veritabanlarında tuttuğu bu büyük veri kullanılarak anlamlı içgörülerin çıkarılması, bu firmaların gelişimi açısından büyük öneme sahiptir. Bu çalışmada, Setur firmasından alınan ve içerisinde anlaşmalı oldukları otellerin ve özelliklerinin bulunduğu veri seti kullanılarak, makine öğrenmesi algoritmaları ile veriden anlamlı çıkarımlar yapılmıştır. Kullanılan bu makine öğrenmesi algoritmaları kümeleme, boyut indirgeme ve özellik seçimi algoritmalarıdır. Öncelikle oteller, özelliklerinin benzerliklerine göre KMeans Kümeleme algoritması kullanılarak kümelenmiştir. Oteller verisinin üzerinde, bir boyut indirgeme algoritması olan Temel Bileşenler Analizi methodu uygulanmıştır ve bu işlenmiş verinin üzerinde de KMeans Kümeleme algoritması uygulanarak, boyut indirgeme yönteminin otellerin kümelenme işlemi üzerindeki etkisi gözlemlenmiştir ve sonuçlar karşılaştırılmıştır. Daha sonra, kümelenmiş oteller verisinin üzerinde çok değişkenli ve tek değişkenli özellik seçimi teknikleri uygulanmıştır. Bu özellik seçimi tekniklerinin uygulanmasındaki amaç, otellerin kümelenmesi işleminde en çok etkisi olan otel özelliklerinin belirlenmesidir. Çok değişkenli özellik seçimi yöntemi olarak Rastgele Orman algoritması kullanılmıştır. Tek değişkenli özellik seçimi yöntemi olarak ise, filtre-tabanlı bir özellik seçimi yöntemi olan ve skor fonksiyonu olarak 'chi2' nin kullanıldığı SelectKBest algoritması kullanılmıştır.

[*]**Correspond Author**

[*](mert.akyol@setur.com.tr) ORCID ID 0000-0002-3499-0001

e-ISSN: : 2717-8579
Arrıval Date: 22/02/2021; Acceptance Date: 20/03/2021

1. INTRODUCTION

Nowadays, almost all of the big companies, especially in tourism industry, store vast amount of data in their database. It is highly significant to analyze that big data with the right machine learning techniques in order to get the most meaningful insights from that data. For the companies, learning the insights from their stored data is vital for their both technological enhancement and economical growth in their business.

In tourism industry, the big data of companies like Setur, which are specialized in hotel bookings, are mostly composed of hotels and their features. For these companies, their priority is displaying the best fit hotels to best fit customers in order to increase their sales. To achieve this process, the most important thing is learning as much information as possible about both their hotels and customers, so that they will know which hotels would be the best fit for which customers. Clustering algorithms would be a good solution for identifying similar hotels in the data. On the other hand, hotels data contain hundreds of different hotel features. The noise and sparsity of this data is very high, that's why the accuracy of machine learning algorithms while using this data could be low. Feature selection algorithms would be a good solution for identifying the most and least important features in the data, so that redundant features can be removed. After getting the necessary knowledge and preprocessing the data accordingly, they can build different types of recommendation engines on top of it which can make a positive impact on their hotel sales.

In this study, hotels data from Setur have been used for clustering, dimensionality reduction and feature selection analysis. The processed result data of this analysis can be used as a data infrastructure for a hotel recommendation engine or it can be used to get more insights about similar hotels which are grouped together and their most important features in order to do further analysis.

Firstly, hotels data of Setur were processed by KMeans clustering algorithm in two different scenarios. In the first scenario, KMeans clustering algorithm was applied directly to the data. For the second scenario; firstly, Principal Component Analysis(PCA) dimensionality reduction algorithm was applied to the data and after that, KMeans clustering algorithm was applied to the processed data. Elbow method was used for identifying the number of clusters for both scenarios. Different clustering results of hotels were observed and comparison was made between two scenarios. After this step, multivariate and univariate feature selection techniques were applied to the clustered hotels data for identifying the most important features of hotels which have effect on clustering process. As a multivariate feature selection technique, Random Forest algorithm was used. For the univariate technique, SelectKBest algorithm with chi2 score function was used as a filter-based feature selection method.

This study consists of "Introduction", "Literature Review", "Method", "Findings" and "Conclusion" sections.

2. LITERATURE REVIEW

In this part of the study, 4 different hotel recommendation system in literature were analyzed and compared with our study in the context of data, techniques and results.

In the study by Sayar and Turdaliev, a machine learning based dynamic hotel recommendation system has been developed with the aim of increasing customer satisfaction about hotel prices. Support Vector Machines(SVM) machine learning algorithm was used for the classification process of hotels while developing a recommendation system in this study. SVM classification algorithm which was used in this study is a supervised machine learning technique. Whereas, in our study, KMeans clustering algorithm was used which is a unsupervised machine learning technique. Dataset used in this study is similar to our dataset in the context of containing binary hotel feature values. Some additional features were included in the dataset of this study such as hotel ratings and descriptions. (Sayar and Turdaliev, 2018)

Turker et al. have proposed a hotel recommendation system based on collaborative filtering and user profiles. Content-based and collaborative filtering approaches have been combined for developing a hybrid hotel recommendation system. A dataset which contains 7 years of hotel reservation records of customers and hotel features from a firm in tourism industry has been used in this study. For the data preprocessing step, Principal Component Analysis method was used for reducing the number of hotel features to 11 from 220 features. Precision value was chosen as the accuracy value and precision values in different scenarios for content-based, collaborative filtering and hybrid recommendation methods were calculated and comparisons were made between these methods and scenarios. KMeans clustering algorithm was used for clustering the hotels according to their features. Precision scores of different recommendation methods were calculated by using the clustered hotels data and by using the non-clustered hotels data. As a result, precision scores by using non-clustered hotels data were higher than the scores calculated by using clustered data. However, processing load was much higher in the non-clustered scenario than the clustered scenario. Precision score results for the hybrid recommendation system scenario came up to be much higher than the other scenarios. 2 of the machine learning algorithms are mutual in this study and our study, which are Principal Component Analysis and KMeans clustering. PCA was used for reducing the number of hotel features and KMeans

clustering was used for clustering the hotels by their features in both studies. Also datasets used in both studies are from different tourism firms and are similar in the context of containing hotel features. (Turker et al., 2019)

Mavalankar et al. have proposed a hotel recommendation system which finds and recommends 5 different best fit hotel clusters to a user among 100 hotel clusters. A dataset which contains hotel reservation records of customers and hotel features from Expedia firm has been used in this study. Different techniques and models have been used in this study, which includes Naive Bayes, SGD Classifier, XG Boost, Principal Component Analysis and Random Forest algorithm. In the data preprocessing step, Principal Component Analysis method was used for reducing the number of hotel features to 20 from 149 features. As a result of this study, Random Forest algorithm gave better results than the other techniques. 2 of the machine learning algorithms are mutual in this study and our study, which are Random Forest algorithm as a feature selection method and Principal Component Analysis as a dimensionality reduction technique. Random Forest algorithm performed better than other methods in this study, whereas it didn't perform well in our study. Also datasets used in both studies are from different tourism firms and are similar in the context of containing hotel features. (Mavalankar et al., 2019)

Jalan and Gawande have proposed a hybrid hotel recommendation system in order to solve the cold start problem in recommendation process. Cold start problem may occur, for example when a new hotel is added to the database. No users had any interaction with that hotel before, and therefore it doesn't have any similary rank among other hotels. A hybrid recommendation approach has been proposed by combining the collaborative filtering technique with sentimental analysis. A dataset which contains the general information of hotels, their ratings and reviews from Tripadvisor firm has been used in this study. OpenNLP tools were used to parse the hotel review sentences in order to extract the words which are hotel features and user opinions. Semi-supervised clustering algorithm was used to cluster hotel features which have similar meaning. Opinion words were classified as negative and positive sentiments and an orientation score was assigned to each feature accordingly. Weights of the features were assigned according to the number of times that feature occurs. Features were combined with their weights and orientation scores for assigning a score to each review. Recommendation scores were calculated based on user's selected features and review scores of hotels. In this study, a semi-supervised clustering algorithm was used in order to cluster hotel features into meaningful groups, whereas in our study KMeans clustering algorithm was used which is a unsupervised machine learning technique in order to cluster hotels. Datasets are from different tourism firms in both studies but hotel features were extracted manually from the hotel reviews in this study, whereas in our study the hotel features already exists in the dataset. (Jalan and Gawande, 2017)

3. METHOD

In this work, hotels data of Setur has been used which contains 1621 hotels and 32 features.

3.1. Data Preprocessing

Firstly, the features of hotels which will not be used in this analysis were removed, which were 'HotelID', 'HotelName' and 'City'. Categorical values in the data were converted to numeric labels, which were in 'HotelCategory' and 'HotelType' columns. All the values of the data except 'Price' column were discrete numeric labels after this process. 'HotelCategory' has 9 different labels and 'HotelType' has 2. 'Price' column contains continuous numeric price values of hotels. All of the other features have only 2 different labels which are '0' or '1' according to if the hotel has that feature or not. Data was scaled in order to get all the features in the same scale. This process is essential before applying KMeans as it is a distance based algorithm, because scale of the variables affect distance based algorithms.

3.2. Elbow Method

Elbow Method is one of the techniques which can be used to determine the number of clusters to use in KMeans clustering algorithm. In this study, Elbow Method was used to determine the number of clusters. The basic idea behind KMeans clustering is to define clusters such that the total Within-Cluster Sum of Square(WCSS) is minimized. The total WCSS measures the compactness of the clustering. The number of clusters should be chosen such that adding another cluster doesn't improve much better the total WCSS. (Kassambara, 2017)

KMeans algorithm was executed in a loop for the number of clusters between 2 to 60 and WCSS scores for each cluster value were visualized in Figure 1. In the Elbow Method, the location of a bend in the plot is mostly considered as an indicator of the suitable number of clusters. (Kassambara, 2017) By using this knowledge and analyzing the plot in Figure 1, the number of clusters to use for KMeans in the first scenario of this study was determined as 20.

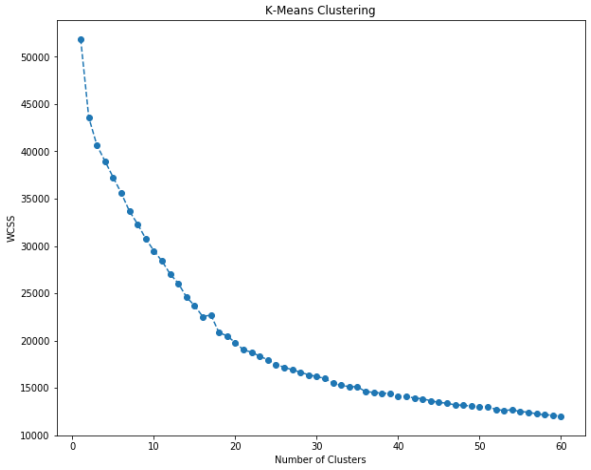

Figure 1. WCSS Scores vs the number of clusters in Elbow Method for the first scenario

3.3. KMeans Clustering

KMeans algorithm is an iterative algorithm that tries to partition the dataset into k pre-defined distinct non-overlapping clusters where each data point belongs to just one group. It tries to form the intra-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid(arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation within clusters, the more similar the data points are within the same cluster. (Jain, 2010)

For the first scenario, KMeans Clustering algorithm was applied to the preprocessed hotels data to cluster hotels into 20 different groups according to the similarity of their features.

3.4. Principal Component Analysis

PCA(Principal Component Analysis) is one of the dimensionality reduction algorithms, which are used for reducing the number of input variables in training data.

When dealing with high dimensional data, it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the "essence" of the data. This is called dimensionality reduction. (Murphy, 2012)

The difference between feature selection and dimensionality reduction is: Feature selection simply selects and excludes given features without changing them. However, dimensionality reduction transforms features into a lower dimension. As an example, PCA reduces dimensionality by making new synthetic features from linear combination of the initial ones, and then discarding the less important ones.

For the second scenario, before applying PCA on the preprocessed hotels data, number of features to keep on the data should be decided. To give this decision, the amount of variance captured in the data for the number of components between 2 and 32 (which is the number of our features) were calculated and visualized in Figure 2 below. The rule in this method is generally preserving approximately 80 percent of the variance. So, number of components was determined as 17 by analyzing the plot in Figure 2.

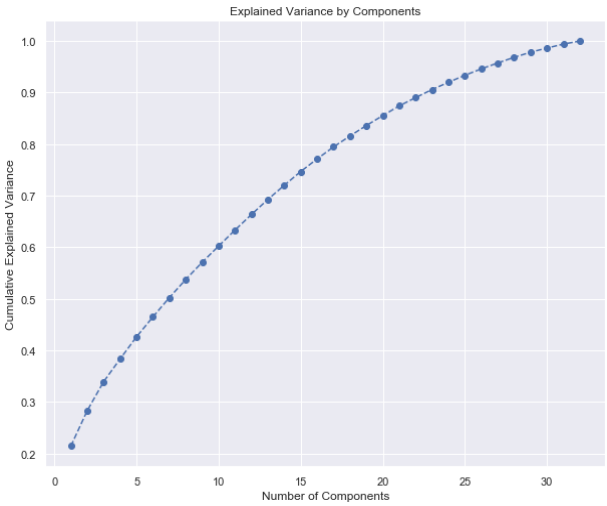

Figure 2. Explained variance by number of components

After this process, PCA was applied to the preprocessed hotels data by using 17 as the number of dimensions to keep. After applying PCA, Elbow Method was used on this dimensionally reduced data in order to determine the number of clusters to use in KMeans clustering algorithm. The WCSS scores for each cluster value were visualized in Figure 3 below.

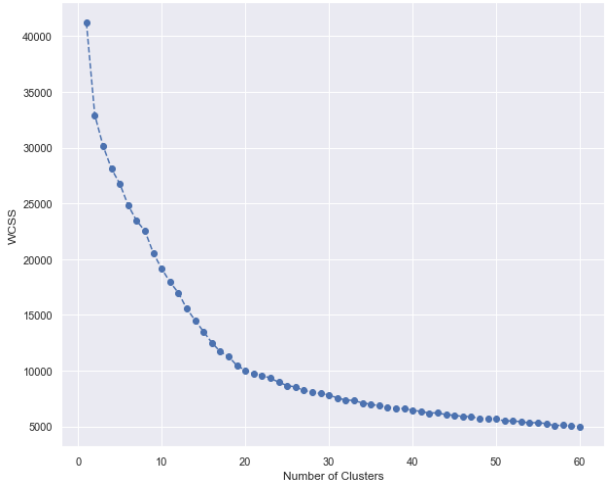

Figure 3. WCSS Scores vs the number of clusters in Elbow Method after applying PCA

By analyzing the plot in Figure 3, the number of clusters to use for KMeans after applying PCA was also determined as 20. Although there is a slight difference between the plots in Figure 1 and Figure 3, the location of the bends indicate the same number of clusters. That's why number of clusters didn't change after applying PCA.

KMeans clustering algorithm was applied to the dimensionally reduced data in order to observe the differences between the clustering results when PCA is applied and not applied.

3.5. Multivariate vs Univariate Feature Selection Algorithms

The aim of feature selection is to find a subset of most relevant variables for a prediction task. To this end, univariate filters, such as a t-test or chi-square test, are commonly used because they are fast to compute and their associated p-values are easy to interpret. (Paul et al., 2013) p-value refers to the hypothesis of the significance level, which is the amount of change a feature will affect towards the final output i.e. how important is this feature and how much it affects the ultimate output. ("URL-1") However, univariate feature selection algorithms don't take into account the possible interactions between variables, whereas multivariate feature selection algorithms take into account the interactions between variables while calculating the importance scores of features.

3.6. Random Forest as a Multivariate Feature Selection Algorithm

In contrast to univariate feature selection methods, a feature selection procedure embedded into the estimation of a multivariate predictive model typically captures interactions between variables. A representative example of such an embedded variable importance measure has been proposed by Breiman with its Random Forest algorithm. (Breiman, 2001)

Random Forest consists collection of decision trees. Each of the trees are built over random extractions of the observations from the dataset and a random extraction of the features. Not every tree sees all the features or all the observations, and this guarantees that the trees are de-correlated and thus less vulnerable to over-fitting. Each tree is additionaly a sequence of yes-no questions based on a single or combination of features. At each node (this is at each question), the tree divides the dataset into 2 buckets, each of them hosting observations that are more similar among themselves and different from those within the other bucket. Therefore, the importance of every feature is derived from how "pure" each of the buckets is. ("URL-2") Within the Random Forest algorithm which was utilized in this study, feature importance was calculated as the decrease in node impurity weighted by the probability of reaching that node. The more a feature decreases the impurity, the more important that feature is. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. ("URL-3")

In this study, Random Forest was used as a multivariate feature selection algorithm. Random Forest algorithm was applied on clustered hotels data. Hotel features were used as the input and cluster number was used as the target variable in order to calculate the importance score of each feature to the clustering process of hotels.

3.7. SelectKBest With Chi2 Score Function as Univariate Feature Selection Algorithm

Univariate feature selection algorithms don't look at all the features collectively. Which means, they don't take into account the interaction between features while calculating their importance score. They determine if there is a significant relationship between the feature and the target variable by looking at each feature separately.

SelectKBest is a method to rank features of a dataset by their importance with respect to the target variable. This importance is calculated using a score function. ("URL-4")

Chi Square is a univariate feature selection algorithm, as it only evaluates a single variable and doesn't take into account the interaction among more than one variable upon the outcome. ("URL-5") It is a statistical test applied to the groups of categorical features in order to evaluate the likelihood of correlation or association between them using their frequency distribution. ("URL-6")

In this study, SelectKBest method was used with chi2 score function as a univariate feature selection algorithm. Firstly, as the data preprocessing step for this section, 'Price' column was removed from the clustered hotels data. The reason of this process is because chi2 score function gives accurate results only when used with categorical input and target variable. Discrete numerical variables can be used as categorical, but continuous variables can't be used. The only continuous variable in clustered hotels data is 'Price', that's why it needs to be removed before applying the SelectKBest algorithm. After this process, same procedures were applied as in the multivariate feature selection section, which are: applying the SelectKBest algorithm on clustered hotels data while selecting the hotel features as the input and cluster number as the target variable in order to calculate the importance score of each feature to the clustering process of hotels.

4. FINDINGS

4.1. Clustering Results for When PCA is Applied and Not Applied

Table 1 and Table 2 show the distributions of 1621 hotels in 20 different clusters when PCA is applied and not applied. 7 mutual clusters were observed between 2 tables by analyzing the contents of each cluster label in these clustering results. Labels of these mutual clusters are 7-6-5-12-15-14-10 For Table 1 and 6-9-7-8-11-19-4 for Table 2. The other 13 clusters were unique for Table 1 and Table 2.

Table 1. Clustering results without applying PCA (first scenario)

Cluster Labels	Number of Hotels
16	251
2	207
19	200
1	118
0	113
3	109
4	108
17	101
18	93
8	75
10	68
14	33
15	31
11	30
13	28
9	23
12	12
5	12
6	7
7	2

Table 2. Clustering results after applying PCA (second scenario)

Cluster Labels	Number of Hotels
5	294
1	261
13	177
16	128
14	110
0	108
3	105
12	92
4	68
15	48
2	41
18	37
19	33
11	31
17	29
10	26
7	12
8	12
9	7
6	2

By analyzing the contents of each cluster in Table 1 and Table 2; Table 1, which contains the clustering results without applying PCA, was chosen as the table which gives more meaningful cluster results. Clustering process after applying PCA couldn't cluster some of the very important hotel features like 'Ski Hotels' or 'Golf Hotels', whereas it managed to cluster these hotel types without applying PCA. In Table 1, all 23 hotels in cluster label 9 were ski hotels and all 12 hotels in cluster 5 were golf hotels. All the other clusters in Table 1 contain specific hotel types and most of the bigger clusters contain hotels which all have multiple specific hotel features. For example, cluster label 8 contains 75 hotels which all are 'Summer Holiday Hotels' and all of them has 'Gravelly Beach'.

By doing the clustering process without PCA, more information have been captured from the data and that resulted in more meaningful clusters. Having more features in the data didn't have any negative impact in this study.

4.2. Feature Importance Results for Multivariate and Univariate Feature Selection Methods

Table 3. Feature importance scores from SelectKBest algorithm with chi2 score function

Features	Importance Scores
Gravelly Sea	1619.000000
Golf Hotel	1614.000000
Sandy Sea	1609.000000
Casino Hotel	1609.000000
Non-Alcoholic All Inclusive	1590.000000
Adult Friendly	1563.161765
Full Pension	1485.396789
Gravelly Beach	1343.512778
Business Hotel	1234.681015
Conservative	1225.432640
Ski Hotel	1201.614398
Only Room	1030.961088

Table 4. Feature importance scores from Random Forest algorithm

Features	Importance Scores
HotelCategory	0.110779
Outdoor Pool	0.088018
Only Room	0.085372
HotelType	0.066826
Indoor Pool	0.058330
Sauna-Hammam	0.057315
Business Hotel	0.052067
Gravelly Beach	0.045766
Fitness	0.043373
Half Pension	0.040327
All Inclusive	0.037867
Bed & Breakfast	0.037489

Table 3 and Table 4 show the highest 12 importance score results from the SelectKBest algorithm with chi2 score function and Random Forest algorithm. Most noticeable difference between 2 feature importance results is that, all of the features in Table 3 are the most significant feature in one of the hotel clusters which are shown

in Table 1. For example; cluster label 12 contains only casino hotels in Cyprus so 'Casino Hotel' is the most significant feature of this cluster, or cluster label 9 contains only ski hotels, cluster 5 contains only golf hotels and so on. However, only 3 of the features in Table 4 are the most significant feature in one of the hotel clusters. Other features like indoor/outdoor pool or fitness are more general features. Which means most of the hotels in Setur database have a fitness center or pool, that's why these features have less effect to differentiate the hotels, so these features are not the most significant feature in any of the clusters.

5. CONCLUSION

In this study, hotels data from Setur firm have been used for clustering, dimensionality reduction and feature selection analysis by applying machine learning methods. Elbow Method was used as one of the most accurate methods for determining the number of clusters to use for KMeans clustering. As a result of this method, number of clusters were determined as 20 which indicates the bend points in Figure 1 and Figure 3 plots. 2 different scenarios were analyzed during the clustering process; which are: applying KMeans Clustering algorithm after applying PCA method on the hotels data and without applying PCA. PCA was used as a dimensionality reduction algorithm on hotels data to see the effect of dimensionality reduction on the clustering process. Number of features to keep on the hotels data was determined as 17 by analyzing the plot in Figure 2 in order to preserve 80 percent of the variance in the data. By applying KMeans Clustering algorithm in both scenarios and analyzing the results, clustering process without applying PCA was selected as it resulted in more meaningful hotel clusters. PCA didn't have any positive impact for the clustering results in this study. Random Forest was used as the multivariate and SelectKBest with chi2 score function was used as the univariate feature selection algorithm on the clustered hotels data. By analyzing the feature importance results, SelectKBest algorithm with chi2 score function gave more meaningful feature importance results as all the most important features in Table 3 were the features which were used to differentiate hotels the most during the clustering process.

In the future studies to be carried out after this study, the results from this study can be used as the infrastructure data for a hotel recommendation engine. Similar hotels were identified and clustered based on the similarity of their features. Also the features which have the highest effect on this clustering process were determined, so features can be chosen according to these results.

REFERENCES

Sayar, A., and Turdaliev, N. (2018). Makine Öğrenmesi ile Adaptif Otel Öneri Sistemi. *12th Turkish National Software Engineering Symposium*, Istanbul, Turkey.

Turker, B. B., Tugay, R., Kizil, I., & Oguducu, S. (2019). Hotel Recommendation System Based on User Profiles and Collaborative Filtering. *2019 4th International Conference on Computer Science and Engineering (UBMK)*, Samsun, Turkey, pp. 601-606.

Mavalankar, A. A., Gupta, A., Gandotra, C., & Misra, R. (2019). Hotel Recommendation System. *Internal Report.*

Jalan, K., and Gawande, K. (2017). Context-Aware Hotel Recommendation System based on Hybrid Approach to Mitigate Cold-Start-Problem. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, Chennai, India, pp. 2364-2370.

Kassambara, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. In *Multivariate Analysis, (1),* 101.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. In *Pattern recognition letters, 31(8),* 651–666.

Murphy, K. (2012). Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series), 11.

Breiman, L. (2001). Random Forests. In *Machine Learning, 45(1), 5–32.*

Paul, J., Verleysen, M., & Dupont, P. (2013). Identification of Statistically Significant Features from Random Forests, 1.

Breiman, L. (2004). Consistency for a Simple Model of Random Forests. Technical Report 670, Technical report, Department of Statistics, University of California, Berkeley, USA.

URL-1: https://medium.com/@abhinav.mahapatra10/ml-basics-feature-selection-part-2-3b9b3e71c14a
[Date of Access: 27.01.2021]

URL-2: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f
[Date of Access: 29.01.2021]

URL-3: https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3
[Date of Access: 29.01.2021]

URL-4: https://medium.com/swlh/feature

-importance-hows-and-why-s-3678ede1e58f
[Date of Access: 03.02.2021]

URL-5: https://science.jrank.org/pages/1401/Chi-Square-Test.html
[Date of Access: 03.02.2021]

URL-6: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/
[Date of Access: 05.02.2021]