# ÇOK DEĞİŞKENLİ VERİ KÜMELERİNDE İLGİNÇ ÖRÜNTÜ TESPİTİ İÇİN BİLEŞEN ANALİZİ

***

# COMPONENT ANALYSIS FOR INTERESTING PATTERN DETECTION IN MULTI-VARIABLE DATA SETS

DOI: 10.33461/uybisbbd.802938

**Ahmet YÜCEL***

**Öz**

*Çağımızın yeni güç kaynağı haline gelen veri kavramı üzerine, son yıllarda büyük gelişmeler elde edilmiştir. Hem kodlama hem de mekanik düzeyde ulaşılan yeni yöntem ve teknikler sayesinde, verinin aktarımı, depolanması ve işlenmesi konusunda muazzam hızlara ulaşılmıştır. Veri aktarımı ve depolama hızlarındaki gelişmeler, dijital platformlardaki en küçük bilgiyi dahi veri olarak depolamayı günlük hayatın doğal bir parçası haline getirmiştir. Aile fotoğraflarından sağlık verilerine, ticari kayıtlardan akademik yayınlara, Twitter'da paylaşılan bir yorumdan Youtube'da paylaşılan bir videoya kadar, hemen her alanda değişik boyutlarda veri anlık olarak depolanmaktadır. Depolanmış verinin içinde bulunan ilginç örüntüler ve açığa çıkarılmayı bekleyen bilgi, veri madenciliğinin temel hedeflerindendir. Veri madenciliği çalışmalarında, veri boyutunun büyüklüğü, karşılaşılan en yüyük sorunlardan biridir. Bu tarz verilerin yapısal hale getirilme süreçlerinin uzunluğu ve sonrasında oluşturulacak bir modelin çalıştırılması sırasında yaşanabilecek sıkışmalar, büyük boyutlu verilerde karşılaşılan sorunlardan bazılarındır. Büyük veri boyutundan kaynaklanan problemlerin üstesinden gelebilmek için birçok boyut indirgeme algoritması geliştirilmiştir. Bu çalışmada, çok değişkenli bir veri üzerine, yeni bir boyut indirgeme yaklaşımı geliştirilmiştir. Bu yaklaşım genel olarak Temel Bileşen Analizine (TBA) dayalı örüntü tanıma adımlarından oluşur. Oluşturulan modeller, birbirlerinden ayrık ve dengeli alt veri kümelerine uygulanmış ve tümü 0.05 anlamlılık düzeyinde anlamlı sonuçlar göstermiştir. Modellerin açıklayıcı performansları; Çoklu R-Kare ölçeğinde [0.819, 0.888]aralığında, ve R-Kare ölçeğinde [0.804, 0.878] aralığında gerçekleşmiştir.*

***Anahtar Kelimeler**: Temel Bileşen Analizi, Örüntü Tanıma, Çok Değişkenli Veri Analizi.*

**Abstract**

*In recent years, great advances have been made on the concept of data, which has become the new power source of our age. Thanks to new methods and techniques at both coding and mechanical level, tremendous speeds have been achieved in the transfering, storing, and processing of data. Thanks to those digital developments, storing even the smallest information on digital platforms has become a natural part of daily life. From family photos to health history, from commercial records to academic publications, from a comment shared on Twitter to a video shared on Youtube, data in almost every field is stored instantly in different sizes. Interesting patterns and information in stored data waiting to be revealed are the main goals of data mining. In data mining studies, the size of data is one of the biggest problems encountered. Some of the problems encountered in large-scale data are the length of the processes of structuring such data and the jams that may occur during the execution of a model to be created afterward. Many dimension reduction algorithms have been developed to overcome the problems arising from large data sizes. In this study, a new dimension reduction approach has been developed on multivariate data. This approach generally consists of pattern recognition steps based on Principal Component Analysis (PCA). The created models were applied on disjoint and balanced sub-datasets and all produced significant results at the 0.05 confidence level. Explanatory performances of the models; They are in the range of [0.819, 0.888] on the multiple R-Square scale and in the range of [0.804, 0.878] on the R-Square scale.*

***Keywords:** Principal Component Analysis, Pattern Recognition, Multivariate Data Analysis.*

* Dr, Öğretim Üyesi, Şereflikoçhisar Uygulamalı Bilimler Fakültesi, Ankara Yıldırım Beyazıt Üniversitesi, Ankara, Türkiye, ayucel@outlook.com, ORCID: 0000-0002-2364-9449



Geliş Tarihi / Received: 30.09.2020
Kabul Tarihi / Accepted: 17.01.2021

## 1. INTRODUCTION

Computer and electronics-based technologies are rapidly becoming a part of our lives thanks to the speed and convenience they provide. This rapid integration of digital developments in our lives is an important factor in the changes in the sociology of society. In addition to the social and cultural effects of digital technology, some significant effects are also seen in economic, political, or organizational contexts. This situation is also efficient in determining the social purposes of the developed technologies. In other words, sociological processes also determine the direction and scope of digital Technologies (Musik et. al., 2019). New contents produced in almost every field from health to education are recorded on digital platforms. This is a normal practice of daily life in the digital sociology of the new age (De Reuver et. al., 2017).

Data mining is the process of discovering non-lean and important information hidden among large data heaps using statistics-based methods such as machine learning (Sumiran, 2018). The enormous size in instant new data generation, reliability, and fast possibilities in data storage processes offer an unlimited resource for data mining. However, this situation brings some new challenges in terms of researchers' literature background.

Predictive decisions are made using mathematical models built on existing data. In these decision processes, large data sets from almost every field from production to finance, from social sciences to natural sciences can be analyzed. This diversity of fields may cause difficulties in terms of inexperience and lack of knowledge of researchers related to the literature, as well as some other problems. Content heterogeneity, measurability, database shortage, data security, aiming right data and applying right methodology are just some of the challanges that researches have to deal with during a data mining process (Kalra et. al., 2014). In addition to some difficulties encountered in the preliminary process from storage to analysis, there are also some difficulties in the analysis phase. In particular, the complexity level of the relationship between variables of large-scale data reveals problems such as overfitting, collinearity or biasness in the mathematical model to be created (Garg et. al., 2013). For this purpose, many data size reduction methods have been developed in the literature (Varghese et. al., 2012).

Principal component analysis (PCA) is one of the most popular and strong unsupervised dimension reduction methods in data mining literature (Sharifzadeh et al., 2017). Sehgal et. al. has utilized a PCA-based algorithm to shorten the analysis process. The main motivation of the study on choosing PCA as the dimension reduction tool is that PCA is a very strong method for reducing the data size while causing a tiny information loss from the original data (Sehgal et al., 2014). In another study, multivariate stock exchange data (Shanghai stock exchange 50 index (SSE50)) is analyzed for screening the linear transformation of random vectors by utilizing functional PCA which is an insight pattern exploring tool through the variables and a data dimension reduction method. PCA provides the ability to reduce the dimension of the data down to a smaller size and detect the statistically significant ones existing among extracted features. In the study, the performance of PCA is checked on different data sizes. The method produced convenient results for each level of data size (Wang et al., 2014).

Pattern recognition is an information extraction process that includes machine learning and has a wide range of applications from health to social media (Farahnaz et. al., 2020). Thanks to the statistical pattern recognition method, a data set from the health domain may be analyzed and some vital information about patients detected by doctors (Chen et. al., 2019). PCA is a widely used method in such pattern recognition processes (Washizawa, 2009). In a study designed by Vidhyavathi, PCA is utilized as pattern recognition tool to identify some significant patterns in multivariate medical data (Vidhyavathi, 2017).

In addition to the medical and economics domain, PCA is also a popular statistical tool in statistical image processing and face recognition fields (Sarkar et al., 2014). Vajčnerová et. al. compares the performances of PCA and cluster analysis on an e-commerce dataset. The primary motivation of the study is to determine the advantages and disadvantages of each method through comparison. The findings of the study show that the performances of the methods are similar to each other (Vajčnerová et al., 2016).

The main purpose of this study is to present a new data analysis approach that can be applied by any researcher despite having very limited experience in the field. Therefore, the data used in the study have both social and medical content. One of the main sources of inspiration for the study is the article published by Pitombo and Gomes in 2014. In the study, the behavioral patterns of society in terms of economic and cultural-class factors were detected. A PCA integrated model was introduced for the pattern detection process (Pitombo et al.,2014).

A subset of data is determined based on a particular severity scale in order to maximize the analysis performance level by solving size and density problems. A feature selection method should be applied to determine this subset (Marta et. al., 2019). In literature, from Gain ratio to Information Gain, there are many feature selection algorithms (Ahmed et. al., 2020). In the methodology of this study, a classification based feature selection is applied. Especially in machine learning and pattern recognition studies on high dimensional datasets, feature selection has a very major contribution to reduce the data processing time and improve the analysis performance (Zhang et. al., 2018). Determining the most suitable feature selection method is very critical. In Chandrashekar and Şahin's study, they applied many different feature selection algorithms on datasets from different fields, and they analyzed the performance differences of the algorithms by comparing them (Chandrashekar et al., 2014).

Feature selection is applied to find an ideal subset of extracted features. In the study of Dash et al., feature-selection-based subsets are compared in terms of inconsistency, and classification accuracy rates, before and after the feature selection (Dash et al., 2003).

In this study, a new multi-variable data analysis approach that is integrated with PCA and f-value-based Feature Selection algorithms is presented.

## 2. MATERIAL AND METHODS

### 2.1. Pricipal Component Analysis (PCA)

The main idea of the PCA is to reduce the size of a data set, and so to make it easier to understand and construct a model on it while maintaining the variability in the set. Measurements of variance or covariance between variables are the main components that allow PCA to generate data with a smaller size than the original one without losing any information (Tibco, 2020). Creating a smaller-sized subset from the original data has some basic goals such as selecting the most critical information, making the data easier to interpret, and identifying interesting patterns in the data (Abdi et al., 2010). The mathematical expression of the PCA is as follows:

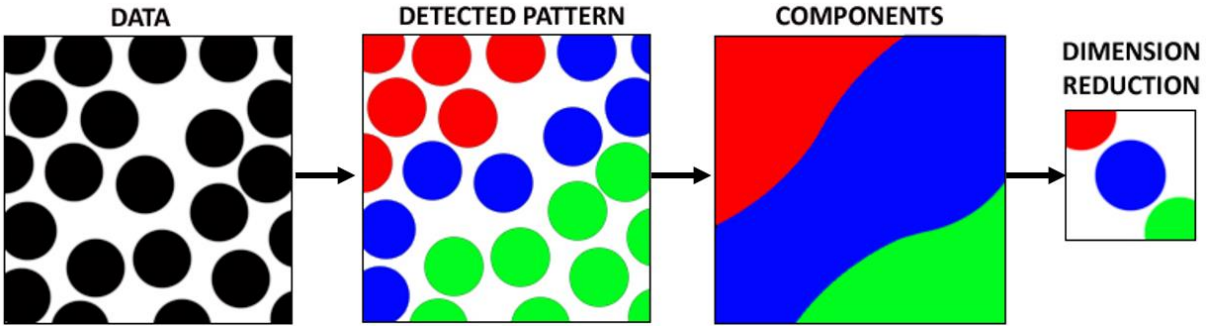


**Figure 1.** PCA Methodology

Let $X$ be an $n \times m$ matrix having the columns (vaiables) $x_1, x_2, \dots, x_m$. Let Y be the matrix having principal components of $X$ as variables; $y_1, y_2, \dots, y_k$ $(k < m)$. Let $\mu_1, \mu_2, \dots, \mu_m$ be the mean values of each variable of $X$. By using the mean values, the covariance matrix ( $\sum$ ) of $X$ can be calculated. Then, eigen-decomposition of $\sum$ will give the eigenvectors of $X$. Each vector is principal component of $X$. After choosing the best components which are not close to zero, we can form a new matrix called Singular Value Decomposition (SVD) of $X$. The projection of $X$ is $P = Y^T X$ where $X$ is the original data, $Y^T$ is the transpose of the chosen principal components (Brownlee, 2018).

### 2.2. General Linear Regression (GLM)

Generalized linear models (GLM) are an extended form of the simple linear regression concept. According to the linear model, the conditional expected value of Y (dependent or independent variable) is equal to a linear composition of $X^T\beta$, i.e.

$$E(Y|X) = X^T\beta. \quad (1)$$

Where Y is the dependent variable matrix, and X is the matrix of independent variables. The equation $Y = X^T\beta + \varepsilon$ is equivalent to Eq.1 (Müller, 2004).

### 2.3. Feature Selection

The ANOVA test based on the F-statistics is applied to determine the sensitivity level of a function. The ratio of sum of squares is used in the F-statistics calculation. Twice the sum of squares are calculated. The first is for the observations within the samples, and the second is between the samples. Sum of squares within the samples can be given by

$$SS_w = \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - X_i)^2 \quad (2)$$

Where, m is the number of independent variables, n is the number of cases in the each independent variable, Xi is the mean of the ith sample. Sum of squares between samples can be given by,

$$SS_b = \sum_{i=1}^{m} (X_{ij} - X_i)^2 \quad (3)$$

The following equation gives the F-statistic.

$$F = \frac{SS_w(m-1)}{SS_b(nm-m)}$$

A feature selection is made according to the F-statistic. If the F value of a feature is too low, then it may be removed from the model (Madhavi et. al., 2016).

## 3. EXPERIMENTAL STUDY

### 3.1. Data Set

The data used for the study were obtained from Kaggle.com, a worldwide data-sharing platform. Data have a "Public Domain Dedication" CC0 1.0 Universal (CC0 1.0) license. The original name of the data is ‘India - Annual Health Survey (AHS) 2012-13’. The data consist of the results of a health survey. The survey was conducted in nine states called the 'Enhanced Action Group' (EAG), with almost half of the total country population. After data cleaning, the data set consists of 215 variables and 206 cases, including ID variables.

### 3.2. Methodology

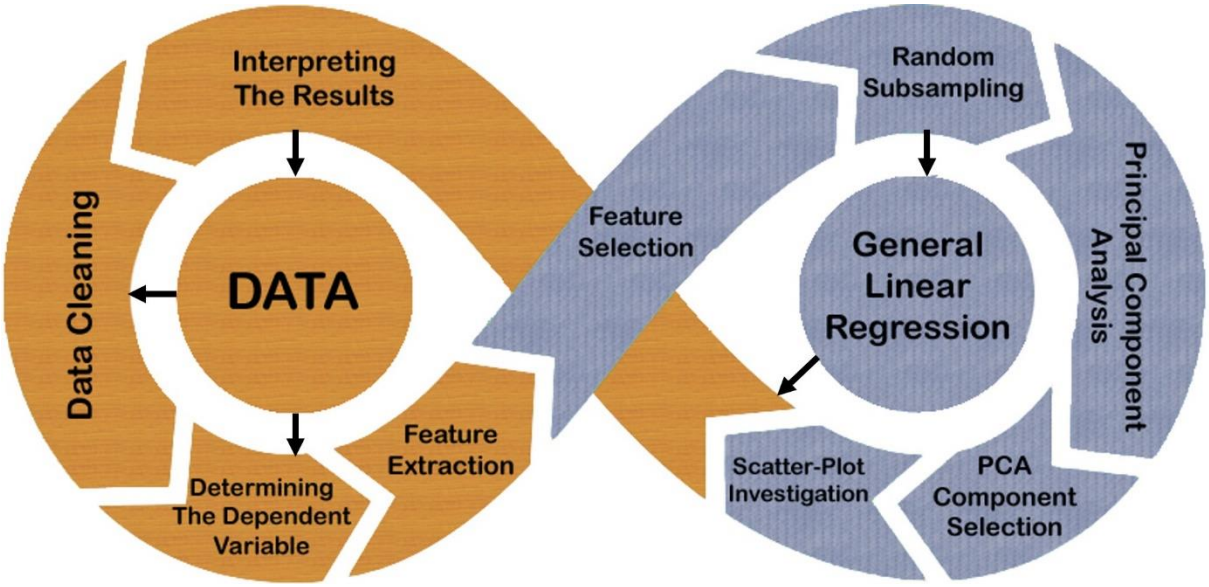


**Figure 2.** Methodology

The general purpose of this study is to produce a multivariate pattern recognition modeling approach, although researchers have little technical knowledge about the data content. First of all, irrelevant or unnecessary details (web links, image extensions, page numbers, etc.) existing in the data are removed. Besides, variables that contain too many missing (more than 90%) or invariant variables are purged from the data. In the next step, researchers determine one of the variables as the dependent variable, either by random selection or based on their knowledge of the subject. After that, a feature selection is applied to the data based on the specified dependent variable. Thus, independent variables having a significant relationship with the dependent variable are determined from the feature-selection-based subset. This process produces smaller data that consists of only important features. After this stage, the process splits in two directions. The first one is to visualize the pattern of the relationship between the dependent and independent variables, and the second one can be considered as the mathematical expression of the correlation level and aspect of the independent variables with the dependent variable. In the first one, we use principal component analysis (PCA) to reduce the data size and identify the pattern among the variables. In the PCA process, the reduced-size forms of the original data are created using relational regression and there is no loss of information. Thus, the components obtained at the end of the PCA process contain the same amount of information as the variables in the original data. However, the first component contains the most information and the last one contains the least. In other words, from the first

component to the last one, a sequence is determined so that each one contains more information than the next one. For this reason, the first and second components are used as the axes of the scatter plot that will help us determine the pattern in the data. In the second one, a linear regression model is created using the determined dependent and independent variables. By interpreting the visual and mathematical results together, it is tried to detect the possible pattern in the data.

### 3.3. Results

By a random selection, *'abortion_taking_place_in_institution*' is determined as the dependent variable. In the next step, F-values are calculated according to the determined dependent variable, and a feature selection based on the F-value is performed at the 0.01 significance level. Twelve features have been selected according to the determined parameters. The selected features and order of importance are given in Table 1.

**Table 1.** Variable Importance Order

| **Dependent Var.: Abortion_Taking_Place_In_Instit.** | **F-value** | **p-value** |
|---|---|---|
| Abortion_Performed_By_Skilled_Health_Personnel | 125,09 | 0,000 |
| Women_Who_Received_Any_Anc_Before_Abortion | 4,43 | 0,000 |
| Child_Who_Rec_Foods_Other_Than_BreastMilk_Dur_1st_6_Mon_Animal_Formula_Milk | 4,07 | 0,000 |
| Women_Who_Went_For_Ultrasound_Before_Abortion | 3,69 | 0,000 |
| Child_Aged_5_14_Years_Engaged_In_Work_F | 3,75 | 0,000 |
| Num_Of_Injured_Persons_By_Type_Of_Treat_Receivd_Per_10E5_Population_Severe_F | 3,58 | 0,002 |
| Current_Usage_F_Sterilization | 3,11 | 0,002 |
| Avrg_Household_Size_All | 2,78 | 0,004 |
| Persons_Suffer_From_Acute_Illness_Per_10E5_Population_Any_TypeOf_Acute_Illness_F | 2,91 | 0,006 |
| Current_Usage_Emergency_Contraceptive_Pills | 3,14 | 0,009 |
| Persons_Suffer_From_Acute_Illness_Per_10E5_Popul_Any_TypeOf_AcuteIllness_Person | 2,75 | 0,009 |
| Avrg_Mon_By_Which_Child_Receivd_Food_Other_Than_Breast_Milk_Vegetables_Fruits | 2,50 | 0,009 |

The PCA model is applied to the original data and consists of thirteen variables including the dependent variable. At the end of this process, components that explain 100% of the data are composed. Since the explanatory ratios of the components decrease from the first to the last, the first and second components with the highest explanatory ratios are used as the axis-parameters of the scatter-plot.

**Table 2.** Principal Components Analysis Summary

| **Component** | **$R^2X$** | **$R^2X$(Cumul.)** | **Eigenvals.** | **$Q^2$** | **Limit** | **$Q^2$(Cumul.)** | **Iterations** |
|---|---|---|---|---|---|---|---|
| **1** | 0,210973 | 0,21097 | 2,73653 | 0,02909 | 0,08187 | 0,02909 | 11 |
| **2** | 0,155678 | 0,36665 | 2,01627 | -0,08154 | 0,08833 | -0,05008 | 50 |

The scatter-plot created in this way provides the best possible image of interaction models between variables. PCA results for the first two components are given in Table 2. According to the results, 36.6% (cumulative **$R^2X$ = 0,36665**) of the original data is explained by the two components.

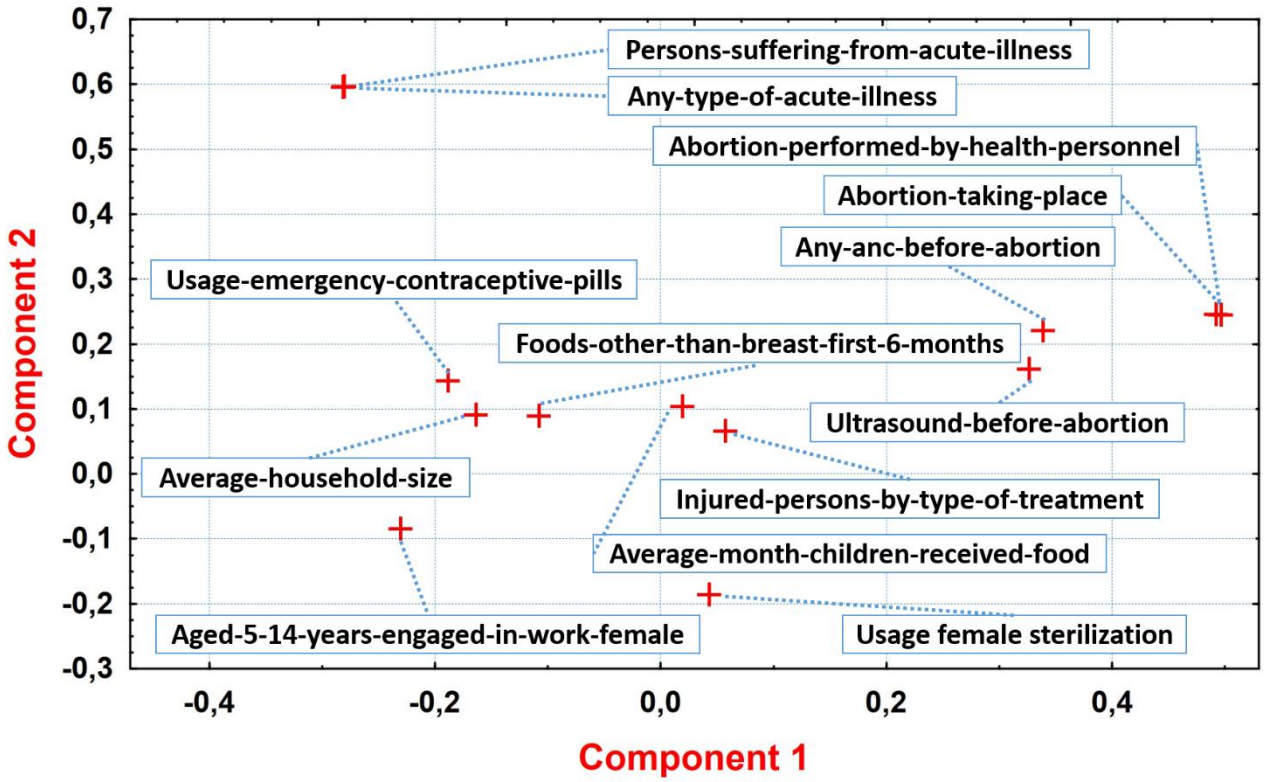


**Figure 3.** Scatter-plot based on the PCA components

A scatter-plot is designed using the selected components (see Fig.3). Each point on the graph corresponds to a variable. The distance between any two points is an important indicator that expresses the correlation level between the variables corresponding to those points. In other words, it can be said that closely located points have a stronger correlation. From this point of view, by looking at the locations of the variables, correlation patterns can be determined. Considering the variables numbered sequentially according to their distance from the dependent variable in the graph, it is seen that all independent variables are in a relationship with the dependent variable at certain levels. The *'abortion-performed-by-skilled-health-personnel*' variable has the strongest relationship with the dependent variable. Also, *'women-who-received-any-anc-before-abortion*' and *'women-who-went-for-ultrasound-before-abortion*' variables have a strong relationship with each other, and their level of correlation with the dependent variable is second.

Variables *'child-who-receivd-foods-other-than-breastmilk-during-1st-6-mon-animal-formula-milk', 'child-aged-5-14-years-engaged-ın-work-famel a', 'num-of-injured-persons-by-type-of-treat-receivd-per-10e5-population-severe-femal a', 'avrg-household-size-all', 'avrg-mon-by-which-children-receivd-foods-other-than-breast-milk-vegetables-fruits',* and *'current-usage-emergency-contraceptive-pills'* form a relationship pattern. Finally, *'current-usage-f-sterilization'* is not included in any relationship set with other variables, '*persons-suffering-from-acute-ıllness-per-10e5-population-any-type-of-acute-ıllness-female',* and *'persons-suffering-from-acute-ıllness-per-10e5-popul-any-type-of-acuteıllness-person*' form a set of relationships with the weakest level of correlation with the dependent variable relative to other variables. Also, interesting patterns can be captured by examining the distribution of independent variables around the determined dependent variable. The correlation pattern based on the dependent variable is given in Figure 4.

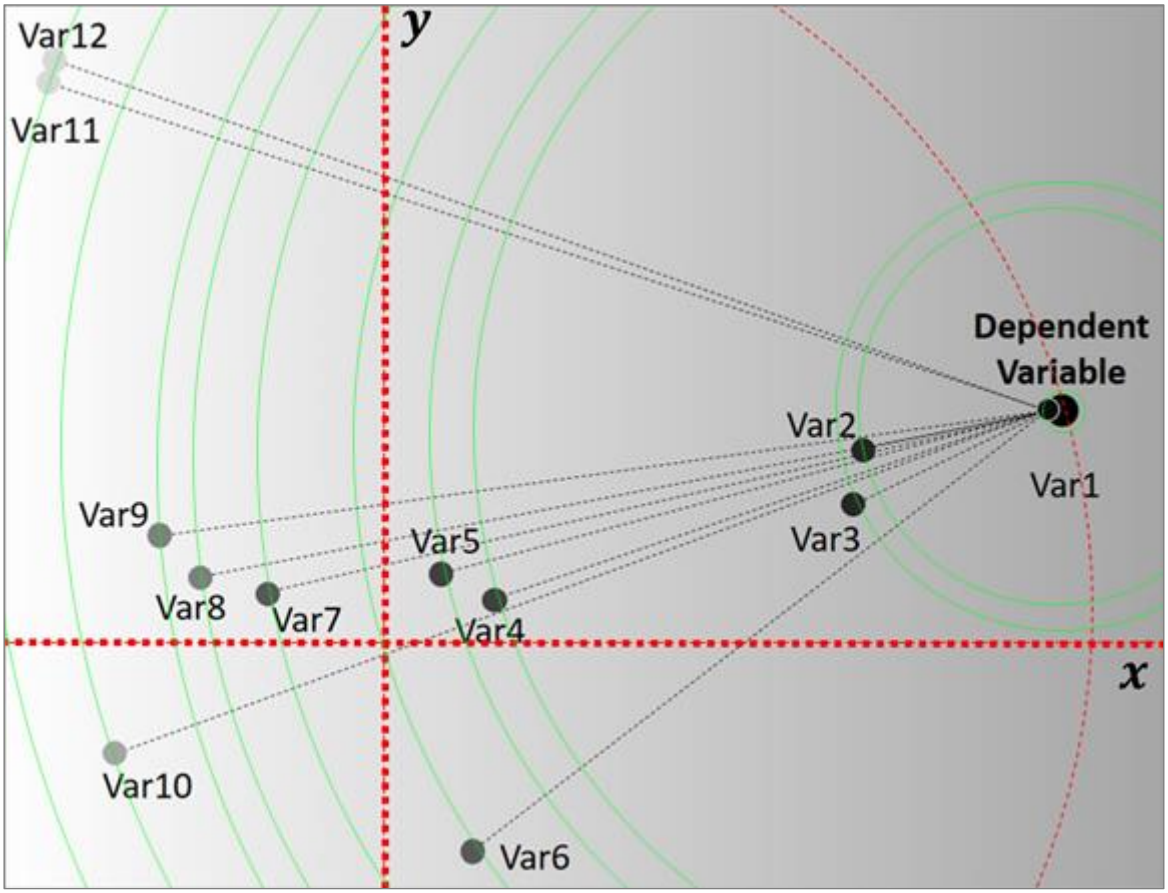


**Figure 4.** Distribution pattern of the independent variables around the dependent variable

In the next step, we create linear regression models with the selected independent and dependent variables. First of all, the data is split into five equal-sized, stratified, and mutually disjoint subsets. The purpose of this procedure is to control the overfitting problem and to observe the level and trend of the correlation between the independent and dependent variables. The applied process is called 5-fold cross-validation. Results are being given in Table 3.

**Table 3.** General linear regression (5-fold cross-validation)

| | Dependent Variable: Abortion_Taking_Place_In_Institution_Urban | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Multiple (R) | Multiple ($R^2$) | Adjusted ($R^2$) | SS (Model) | df (Model) | MS (Model) | SS (Res.) | df (Res.) | MS (Res.) | F-Value | P-Value |
| Fold 1 | 0,916 | 0,839 | 0,825 | 50335,730 | 12 | 4194,644 | 9685,972 | 138 | 70,188 | 59,763 | 0,000 |
| Fold 2 | 0,915 | 0,837 | 0,822 | 49934,541 | 12 | 4161,212 | 9759,122 | 139 | 70,210 | 59,268 | 0,000 |
| Fold 3 | 0,939 | 0,882 | 0,872 | 53628,358 | 12 | 4469,030 | 7143,018 | 141 | 50,660 | 88,217 | 0,000 |
| Fold 4 | 0,942 | 0,888 | 0,878 | 57935,073 | 12 | 4827,923 | 7325,062 | 136 | 53,861 | 89,637 | 0,000 |
| Fold 5 | 0,905 | 0,819 | 0,804 | 44936,863 | 12 | 3744,739 | 9927,708 | 141 | 70,409 | 53,185 | 0,000 |

According to the results, all five models are significant at alpha = 0.05 confidence level. In addition, multiple-R, multiple R-square, and adjusted-R-square ratios have been examined to determine the statistical significance of the models. All of them are in the range of [0, 1] and statistically significant. 0 means 0% compatibility (performance) and 1 means 100% compatibility. According to the results, Multiple R-Square takes values varying between [0.905, 0.942], Multiple

R-Square [0.819, 0.888] and Corrected R-Square [0.804, 0.878]. These results show that modeling is very accurate and successful. Results for each fold are given in Figure 4.

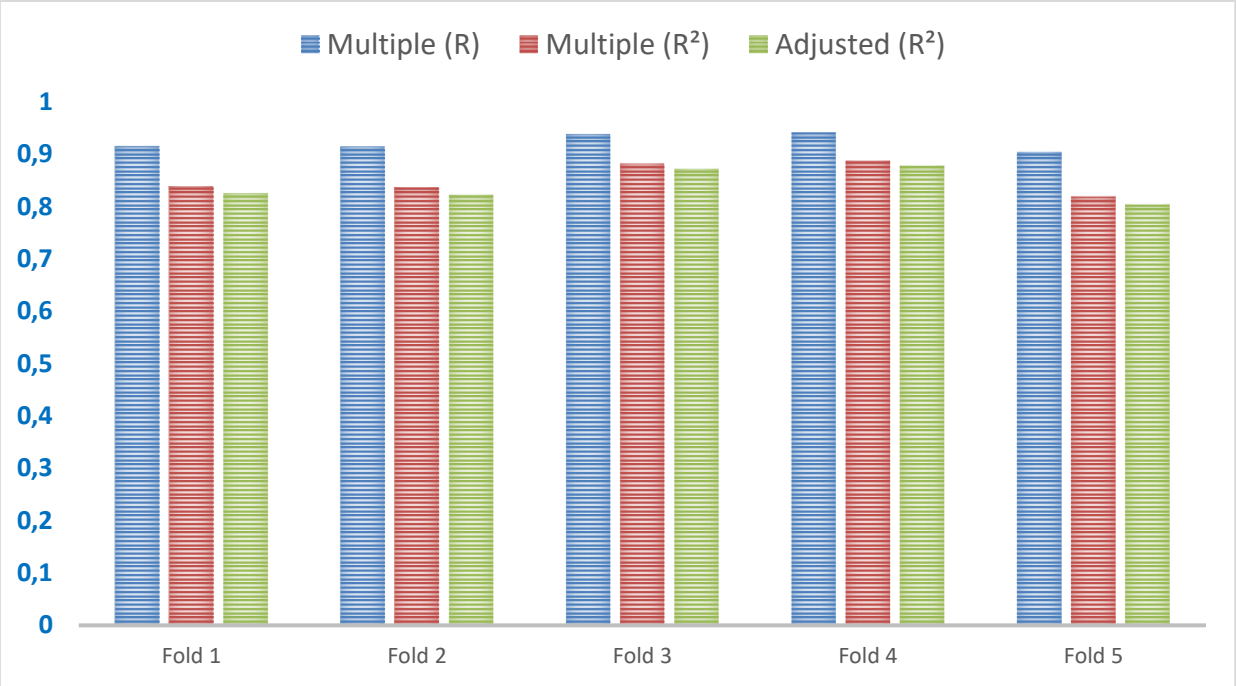


**Figure. 4.** Explanatory performances of the models

## 4. DISCUSSION

In this study, data analysis has been performed on health research data collected from several states of India. Since the data contains too many variables and there is very limited literature knowledge about its content, in other words, the researcher is not familiar with the content of the data, it is not possible to create a linear approach. Therefore, the process of data analysis and revealing the pattern within the data is quite challenging. Thanks to digital technologies, data is generated at tremendous speeds. Therefore, it is quite possible that data analysts have no technical knowledge about data content or do not know which steps to take to access information in the complexity labyrinth structure of the data without a specific entrance gate. The study aims to provide a data analysis approach that will help the researcher analyze such data. The success of the data analysis models produced based on this new approach is very high. The general purpose of this study is to ensure that researchers have a multivariate pattern recognition modeling approach, although they have little technical knowledge about data content. With a method called 5-Fold cross-validation, the overfitting problem that can be found in modeling has been overcome. Visual and mathematical results were evaluated together to determine the possible pattern in the data. Linear regression models were produced with selected independent and dependent variables. The results reveal that the new approach produced is successful at alpha = 0.05 confidence level. In the article published by Pitombo and Gomes in 2014 (Pitombo et al., 2014), which is one of the main inspiration sources of the study, a PCA integrated model was applied in the pattern recognition process to determine the behavior shape of society in terms of class factors. A prediction model based on the integration of PCA and decision tree methods was applied in their study. In this study, after using PCA as a data reduction tool in the first stage of the process, visual arguments are used in the second stage. This preserves the model from the difficulties encountered in applying decision trees to big data and makes it easier to implement.

## REFERENCES

Abdi H., Williams L. J., (2010)."Principal component analysis", Volume 2, John Wiley & Son s, In c. doi. 10.1002/wics.101

Ahmed, M. R., Tahid, S. T. I., Mitu, N. A., Kundu, P., & Yeasmin, S. (2020, July). A Comprehensive Analysis on Undergraduate Student Academic Performance using Feature Selection Techniques on Classification Algorithms. In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE. doi: 10.1109/ICCCNT49239.2020.9225341.

Brownlee, J., (2018). "How to Calculate Principal Component Analysis (PCA) from Scratch in Python".

Chandrashekar G., Sahin F., (2014)."A survey on feature selection methods", Computers and Electrical Engineering Vol. 40, Issue 1, pp. 16-28. doi.org/10.1016/j.compeleceng.2013.11.024

Chen, H. Yan, J. Zhang, G. Hong, H. Zhu, X. "Human target respiration pattern recognition based on vital-SAR-imaging". Asia-Pacific Microwave Conference Proceedings, APMC, Proceedings of the 2019 IEEE Asia-Pacific Microwave Conference, APMC 2019. (Asia-Pacific Microwave Conference Proceedings, APMC, December 2019, 2019-December:865-867)

Dash M., Liu H., (2003)."Consistency-based search in feature selection", Artificial Intelligence, Volume 151, Issues 1–2, Pages 155-176, doi.org/10.1016/S0004-3702(03)00079-1

De Reuver, M., Sørensen, C., Basole, R. (2017). "The Digital Platform: A Research Agenda". Journal of Information Technology. 33. 10.1057/s41265-016-0033-3.

Farahnaz G. M., Huthaifa A. Al_Issa. (2020). "Developing Machine Learning Model for Disambiguate Pattern Recognition on Social Media". 2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM) Computation, Automation and Knowledge Management (ICCAKM), 2020 International Conference on. :547-551 Jan, 2020

Garg, A., Tai, K.. (2013). "Comparison of statistical and machine learning methods in modelling of data with multicollinearity". Int. J. of Modelling. 18. 295_312. 10.1504/IJMIC.2013.053535.

Kalra, B., Yadav, S., Chauhan, D. (2014). "A Review of Issues and Challenges with Big Data". 2. 97-101. International Journal of Computer Science and Information Technology Research. ISSN 2348-120X (online) Vol. 2, Issue 4, pp: (97-101), Month: October - December 2014

Madhavi B. Desai, S. V. Patel & Bhumi PrajapatI. (2016). "ANOVA and Fisher Criterion based Feature Selection for Lower Dimensional Universal Image Steganalysis". International Journal of Image Processing (IJIP), Volume (10) : Issue (3) : 2016 145.

Marta L., Mauro F. (2019) "Statistical analysis of proteomics data: A review on feature selection", Journal of Proteomics, Volume 198, 2019, Pages 18-26, ISSN 1874-3919, https://doi.org/10.1016/j.jprot.2018.12.004.

Musik, C. And Bogner, A. (2019) "Book title: Digitalization & society: A sociology of technology perspective on current trends in data, digital security and the internet", Österreichische Zeitschrift für Soziologie: Vierteljahresschrift der Österreichischen Gesellschaft für Soziologie, 44(Suppl 1), p. 1. doi: 10.1007/s11614-019-00344-5.

Müller, M., (2004). "Generalized Linear Models". 10.1007/978-3-642-21551-3_24.

Pitombo C. S., Gomes M. M., (2014)."Study of Work-Travel Related Behavior Using Principal Component Analysis", Open Journal of Statistics, 4, 889-901. doi.org/10.4236/ojs.2014.411084

Sarkar J., Saha S., Agrawal S., (2014). "An Efficient Use of Principal Component Analysis in Workload Characterization-A Study, AASRI Conference on Sports Engineering and Computer Science (SECS 2014), AASRI Procedia 8 (2014) 68 – 74, doi: 10.1016/j.aasri.2014.08.012

Sehgal S., Singh H., Agarwal M., Bhasker V., Shantanu, (2014)."Data analysis using principal component analysis," International Conference on Medical Imaging, m-Health and Emerging Commun. Systems, Greater Noida, 2014, pp. 45-48. doi.10.1109/MedCom.2014.7005973

Sharifzadeh S., Ghodsi A., Clemmensen L. H., Ersbøll B. K., (2017)."Sparse supervised principal component analysis (SSPCA) for dimension reduction and variable selection", Engineering Applications of Artificial Intelligence, Volume 65, Pages 168-177, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2017.07.004.

Sumiran, K. (2018). "An Overview of Data Mining Techniques and Their Application in Industrial Engineering". Asian Journal of Applied Science and Technology (AJAST) (Open Access Quarterly International Journal) Volume 2, Issue 2, Pages 947-953, 2018

TIBCO Product Documentation. (2020) "Principal Component Analysis (PCA) and Partial Least Squares (PLS) Technical Notes"

Vajčnerová I., Šácha J., Ryglová K., Žiaran P.,(2016). "Using The Cluster Analysis And The Principal Component Analysis In Evaluating The Quality Of A Destination", Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis, Vol. 64, No. 2, doi.org/10.11118/actaun201664020677

Varghese N., Verghese V., Gayathri P., Jaisankar N., (2012)."A Survey Of Dimensionality Reduction And Classification Methods", International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3.

Vidhyavathi R., (2017)."Principal Component Analysis (Pca) In Medical Image Processing Using Digital Imaging And Communications In Medicine (Dicom) Medical Images", International Journal of Pharma and Bio Sciences; 8(2): (B) 598-606 ISSN 0975-6299, doi.org/10.22376/ijpbs.2017.8.2.b.598-606

Wang Z., Sun Y., Li P., (2014)."Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index", Hindawi Publishing Corporation Discrete Dynamics in Nature and Society, Article ID 365204, pp. 7 doi.org/10.1155/2014/365204

Washizawa, Y. (2009). "Subset kernel PCA for pattern recognition". 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. :162-169 Sep, 2009

Zhang, P., Gao, W. And Liu, G. (2018) "Feature selection considering weighted relevancy", Applied Intelligence, 48(12), p. 4615. Available at: http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=132904824&lang=tr&site=eds-live (Accessed: 14 January 2021).