ISSN 2063-5346

AI AND DATA ANALYTICS IN HEALTHCARE: NEED FOR SKILLS DEVELOPMENT

M. Prema [1*], Dr. V. Raju [2] and Dr. S. Pitchumani Angayarkanni [3]

Article History: Received: 01.02.2023 **Revised:** 07.03.2023 **Accepted:** 10.04.2023

Abstract

In recent times, healthcare has become an extremely important social and economic issue to be dealt with in the global arena. As the demographics has been changing in many parts of the world, healthcare has become a major factor in the industrial output and the overall growth of the economies everywhere. Healthcare has also become a major factor in the long-term security and the overall prosperity of nations. From Japan to the countries of Europe and the North Americas, the aging population and its health conditions, the cost of healthcare in those regions, and the availability of a viable workforce to meet the service needs have pushed the significance of healthcare to the top of the political, economic and the national security agenda. With the emergence of new technologies, including Artificial Intelligence, Data Analytics, Internet of Things and others, there is the promise that the challenges in healthcare will be overcome considerably and the countries everywhere will have their citizens being healthier, safer and prospering in thriving economic conditions. This paper presents an overview of a broader set of AI and Data Analytics skills needed for the health care industry in promoting innovation in medicine, predicting diseases, supporting drug discovery and development, improving patient care, delivery of services, operational efficiency of healthcare systems, and the means to control the cost of healthcare. Further, it presents a survey on the preparations of the current undergraduates of engineering programs to meet the skill demands of the healthcare industry. An online survey was developed and administered to the students of third and fourth year Bachelor of Technology (B.Tech) and Bachelor of Engineering (B.E) degree programs that belong to 4 different streams namely Computer Science and Engineering (CSE), Information Technology (IT), Electrical and Electronics Engineering (EEE) and Electronics and Communication Engineering (ECE). The statistical approach involved in the analysis of the survey data are Principal Component Analysis (PCA). The proposed approach with PCA yielded five components that are essential for an entry level employee to perform well in the health care industry. The five components are named as Artificial Intelligence and Data Analytics, Health Care Functions, Soft Skills, Math and Statistics, and Programming Skill

Keywords: Data Analytics, Artificial Intelligence, Healthcare, Skill sets, Parallel analysis, and Principal Component Analysis (PCA)

[1*] Vice Principal, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India. m.prema@sriramachandra.edu.in , ORCID: 0000-0001-8970-6533

[2] Provost, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India. provost@sret.edu.in, ORCID: 0000-0001-8456-6615

[3] Associate Professor, Sri Ramachandra Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India. angayarkanni@sret.edu.in, ORCID: 0000-0001-9621-190X

DOI: 10.31838/ecb/2023.12.s1.046

1. Introduction

Healthcare can be broadly defined as the combination of all of the activities that keep individuals and the global community healthy and safe. At one level, they cover the pharmaceutical activities, ranging from recognizing and identifying diseases, discovering and developing drugs, testing and validating their use in curing diseases and manufacturing and distributing the drugs everywhere. At another level, healthcare involves diagnosing diseases, treating patients with medicines and medical procedures, and managing patient related information for disease prevention, control and effective treatment. And yet at another level, it involves managing healthcare facilities and services at a single unit to multiple global level operations. There are also management of healthcare at a national or at the global level. When the outbreak of a communicable disease in a region or a pandemic at a global level, the management becomes a public health challenge. In all these cases, enormous volumes of data are generated and accumulated at all levels. Utilizing such data for disease identification, prevention, and control, development of drugs and medical procedures, and prediction of patterns in diseases, procedures, and patient recovery are all ideal applications where Artificial Intelligence (AI) and Data Analytics have started making enormous impacts. It is believed that we are at the early stages of application artificial intelligence and data analytics and related technologies to healthcare. There remains a whole array of possibilities where such technologies will change the future of healthcare.

Some of the main areas where AI and Data Analytics is used in Healthcare are Pharmaceutical Process and Drug Development, Disease Diagnosis and Treatment, Healthcare System Management and Public Health Management

2. Review of Related Literature

Valenta et. al. (2017) specified that the foundational domains required for health care industries are the Information Science and Technology, Health information science and technology with Human factors and socio-technical systems.

Sirajudeen and Mohamed (2017) have analyzed the impact of ergonomics which is the science of designing the workplace to fit the worker and has come out with the findings that majority of the subjects were unaware of ergonomics (32.8%), cumulative trauma disorders (18.6%), healthy postures related to elbow (34.4%), wrist & hand (39.5%), Level of Monitor (35%), Position of mouse (47.4%) and Mini breaks (42.9%). This research highlighted the necessity of Ergonomic analysis and training pertaining to healthy postures and the measures to reduce the risk of musculoskeletal disorders for the students.

Swathi et. al. (2017) has analyzed the nine most crucial employability skills which are essential in medical services of Healthcare Industry through keyword search from 105 papers. The analysis revealed that the top nine skills required are the communication skills, ICT skills, Work Psychology skills, Teamwork skills, Interpersonal skills, Critical Thinking and Problem-Solving skills, Self-management skills, Planning and Organizing skills and Conceptual and Analytical skills. These skills form the backbone for any successful organization. These skills are very much needed for Medical Services since the services deal directly with the patients both at the front end and back-end. The organization need to assure that these mandatory skills are updated frequently and the employees are trained from time to time.

3. Items/ Skill required for entry level role in health care industry

A preliminary review of the items/skills needed for an entry level employee for the use of AI and data analytics in health care industry, lead to the identification of 65 various items/skills. These 65 items are listed in Table 1.

Table 1: Items/Skill essential for entry level employee in health care industry

Python with Numpy	Central tendencies and measures	Big Data Analytics	Patient rights, privacy protection and legal matters
Python with Pandas	Probability and Statistical distributions	NOSQL Database	Decision making process
Python with MatPlotlib	Differential Equations	Apache Spark	Business operations
Python with Seaborn	Sampling Methods	Predictive analytics	Financial matters
Python with ScikitLearn	Design of Experiments	Tensor Flow	Human resource matters
Python with PyTorch	Regression and Correlation	Keras	Organizational Structure / Decision making process
R programming	Hypothesis Testing and Decision making	Cloud Environment for ML: Google Colab	Communication Skills
Data Structures and Algorithms	Time Series Analysis	Cloud Environment for ML: Kaggle	Interpersonal Skills
DBMS SQL	Data Modelling	Cloud Environment for ML: Github	Team Building
Cyber Security	Data mining	Drug Design and Development	Team work
Web Programming and Development	Data Visualization Tableau	Genome Sequence Analysis	Decision making
Cloud Computing	Data wrangling and Preproceesing	Medical Imaging	Lifelong learning
Computer Vision and Image Processing	Pattern Recognition	Health Information Management System	Design Thinking
Block Chain Technology	Machine Learning	Functions of a health research facility	Ethical practices
IoT	Deep Learning	Patient care operations	
Calculus	Reinforcement Learning	Hospital operations	
Linear Algebra	Neural Networks	Healthcare regulations	

These items are essential for an entry level employee in the health care industry to perform efficiently. Hence, the graduates need to be well prepared in the essential items/skills to perform efficiently at the work place. This study focusses on the engineering graduates' preparedness to work in the health care industry. For assessing the graduates' preparedness, their skill proficiencies need to be determined. The skill proficiency for these 65 items/skills among under graduate students in different streams of Engineering is studied. An online survey questionnaire is developed as the research tool. The proficiency level of the student for each of these 65 items/skills are surveyed. Likert Scale is used for checking the skill proficiency with 1 indicating least proficiency and 5 indicating most proficiency for a given item. In Tamilnadu, India, 10 engineering colleges were randomly selected. For feasibility of the study, in each college, it was decided to survey only four engineering streams, Computer Science and Engineering (CSE), Information Technology (IT), Electrical and Electronics Engineering (EEE) and Electronics and Communication Engineering (ECE). Also, only the third year and fourth year students were included in the study. This ensures that the student had enough years of study and opportunity to equip themselves with the skills. A web survey link was sent to the head of the department of the four streams in each of the 10 engineering colleges. They were asked to circulate it to all the students in their stream. 487 usable responses were obtained from the survey. The demographics of the survey respondents included gender, year of study and stream of study. To ensure confidentiality, no personal information about the survey respondent or any institutional information was collected.

4. Methodology

In order to reduce the dimensionality of the data, principal component analysis is used. Principal Component Analysis (PCA) is an unsupervised machine learning technique that helps to find the most important features in the dataset and makes the interpretation easier. Principal Component Analysis reduces larger number of variables (65 variables in this study) in to fewer new variables that are linear combination of the variables. The sample size in the study is 487. There are several guidelines for sample size adequacy. The sample size adequacy suggested by Comfrey and Lee (1992) is used as a guideline in this study and is given as follows:

50 – very poor; 100 – poor; 200 – fair; 300 – good; 500 – very good; 1000 or more – excellent (p. 217). Based on this guideline, the sample size of 487 cases in the study is considered as very good. PCA is done with varimax rotation using IBM SPSS Statistics for Windows, version 28. This yielded six components. Loadings greater than or equal to 0.50 are considered significant. The six components with their loadings are given in Table 2 in Appendix.

4.1 Retention of factors

Retention of number of factors is an important step in PCA. Theoretically PCA yields as many factors as the number of variables. Retaining the right number of factors makes the interpretation easy and meaningful. Retaining few factors result in loss of important information (Zwick & Velicer, 1986). Retaining more factors leads to concentrating on less significant factors at the cost of significant factors and interpretation of factors becoming difficult (Zwick & Velicer, 1986). Retention of factors is normally done based on the eigen values and scree plot. . The most commonly used methods is the Kaiser or the eigen value greater than 1 criterion (K1). This criterion retains factors with eigenvalues greater than 1 (Kaiser, 1960). From Table 3 (listed only the first 8 components), there are 7 factors that have eigen values greater than 1.

Table 3: Eigen Values and Variance

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	31.294	48.144	48.144	31.294	48.144	48.144
2	5.921	9.109	57.253	5.921	9.109	57.253
3	3.421	5.263	62.516	3.421	5.263	62.516
4	2.470	3.799	66.315	2.470	3.799	66.315
5	1.678	2.582	68.897	1.678	2.582	68.897
6	1.543	2.374	71.271	1.543	2.374	71.271
7	1.186	1.825	73.096	1.186	1.825	73.096
8	0.929	1.430	74.526			

Cattell's Scree test (1966) is another widespread method used in determining the number of retained factors. This involves plotting the eigen values against the respective components, then retaining the factors that lie in the steep cliff region and ignoring the ones that lie in the shallow scree region. Figure 1 gives the scree plot, the short intersecting line indicates the point at which the scree begins and the long flat line indicates the entire scree region. There are five factors that lie in the cliff region and are above the scree point. Hence, it is decided to retain 5 factors.

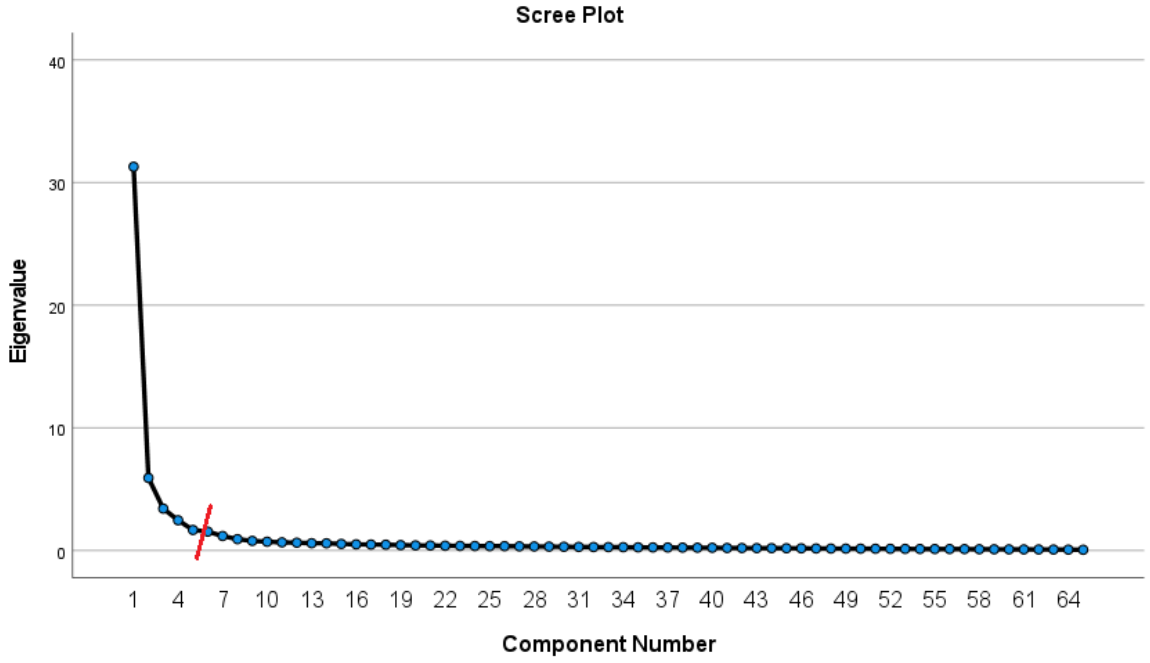

Figure 1: Plot of Eigen values with the respective components

The retained number of components differ as per the K1 criterion (7 components) and Cattell's scree test (5 components). Hence, another more accurate method namely Parallel Analysis PA (Horn, 1965) is done to find the number of components to be retained. In this parallel analysis (PA), a random dataset with the same sample size and number of variables is generated. The eigen values of the significant components from the real data should be greater than that of those from parallel components from the random dataset. This forms the basis of PA. (Ford et al., 1986). Table 4 gives the results of the parallel analysis conducted using IBM SPSS Statistics for Windows, version 28. The eigen values of the real data, eigen values of the random data and the 95th percentile of random data given in the Table 4 are compared. Those components whose eigen values of the real data are greater than that of the random data and 95th percentile are retained. Table 4 shows only the first six components and their eigen values and others are suppressed. It is found that the eigen values of the real data for the first five components is greater than that of the random data and the 95th percentile. Hence, parallel analysis yields five components.

Table 4: Parallel Analysis

Component	Real Data Eigen	Random Data Eigen	Random Data Percentile
1	31.294	1.801395	1.857202
2	5.921	1.735841	1.792646
3	3.421	1.682869	1.728206
4	2.470	1.638138	1.677296
5	1.678	1.602691	1.64066
6	1.543	1.566263	1.599092

Both the Cattell's scree test and parallel analysis yields the retention of five components. Hence, it is decided to retain five components. These retained five components explain a cumulative variance of about 69%. These components have to be named. Naming the components is subjective. The retained five components are named as follows:

Component 1: There are 22 items that loaded in components 1 and they are given in the Table 5.

All these items pertain to artificial intelligence and data analytics and hence this factor is named as Artificial Intelligence and Data Analytics. This component explains 31% of the total variance.

Table 5: Component 1: Artificial Intelligence and Data Analytics

Predictive analytics	Keras	Cloud Environment for ML: Github	Block Chain Technology
Tensor Flow	Reinforcement Learning	Deep Learning	R programming
Pattern Recognition	NOSQL Database	Data mining	Computer Vision and Image Processing
Apache Spark	Cloud Environment for ML: Google Colab	Machine Learning	Cyber Security
Data wrangling and Preproceesing	Big Data Analytics	Data Visualization Tableau	
Neural Networks	Cloud Environment for ML: Kaggle	Data Modelling	

Component 2: There are 14 items that loaded in component 2 and they are given in the Table 6. All these items pertain to health care operations. Hence, this component can be named as Health care functions.

Table 6: Component 2 Health Care Functions

Healthcare regulations	Functions of a health research facility	Decision making process	Drug Design and Development
Patient care operations	Health Information Management System	Financial matters	Genome Sequence Analysis
Hospital operations	Business operations	Medical Imaging	
Patient rights, privacy protection and legal matters	Human resource matters	Organizational Structure / Decision making process	

Component 3: Eight of the items loaded in component 3. These items are non-technical and focus on soft skills. Hence, this can be termed as soft skills and the individual items are given in Table 7

Table 7: Component 3 Soft skills

Lifelong learning	Decision making	Design Thinking	Interpersonal Skills
Team work	Team Building	Ethical practices	Communication Skills

Component 4: In component 4, ten items loaded and all of them pertain to mathematics and statistics. Hence, this component can be termed as Math and Statistics. The ten items in the component are given in Table 8.

Table 8: Component 4 Math and Statistics

Differential Equations	Sampling Methods	Hypothesis Testing and Decision making	Central tendencies and measures
Calculus	Probability and Statistical distributions	Time Series Analysis	
Linear Algebra	Regression and Correlation	Design of Experiments	

Component 5: There are 6 items that loaded in component 5. All these items focus on using Python for programming. Hence this component can be termed as Programming skill. They are listed in the Table 9.

Table 9: Component 5 Programming Skill

Python with Seaborn	Python with ScikitLearn	Python with Numpy
Python with MatPlotlib	Python with Pandas	Python with PyTorch

There are 5 items that do not fall under any of these 5 components and hence they are excluded.

The components with their names, number of items and mean score are given below in Table 10.

Table 10: Component wise PCA results

Component number	Component name	Number of items in the component	Mean Score (on a 5-point scale)
1	Artificial Intelligence and Data Analytics	22	2.13
2	Health Care Functions	14	2.3
3	Soft Skills	8	3.4
4	Math and Statistics	10	2.79
5	Programming Skill	6	2.14

Table 10 clearly indicates that there is skill shortage in Artificial Intelligence and Data Analytics, Health Care Functions, Math and Statistics, and Programming Skill.

5 Conclusion

The study has yielded five important components, namely, Artificial Intelligence and Data Analytics, Health Care Functions, Soft Skills, Math and Statistics, and Programming Skill that are essential for a graduate to perform efficiently in the health care industry. The mean score for these five components indicate that only soft skill has a mean score above 3 (out of 5). Hence, students have more than 60% proficiency in soft skills. Rest of the components have a mean score less than 3 which indicates that the students have less than 60% proficiency for these components. This clearly indicates that students have skill shortage in Artificial Intelligence and Data Analytics, Health Care Functions, Math and Statistics, and Programming Skill. The employment opportunities in health care is expected to grow 13% in the next 10 years resulting in 2 million new jobs (Occupational Outlook Handbook, 2022). It is the right time for the current graduates to tap in to this job market and take a good career path. Academic institutions need to focus on imparting these skills to their graduates. The institutions should offer courses that help to remove the skill shortage among students. This will provide an opportunity for the graduates to equip themselves with the right set of skills to work in the health care industry.

References:

1. Valenta AL, Berner ES, Boren SA, Deckard GJ, Eldredge C, Fridsma DB, Gadd C, Gong Y, Johnson T, Jones J, Manos EL, Phillips KT, Roderer NK, Rosendale D, Turner AM, Tusch G, Williamson JJ, Johnson SB. AMIA Board White Paper: AMIA 2017 core competencies for applied health informatics education at the master's degree level. J Am Med Inform Assoc. 2018 Dec 1;25(12):1657-1668. doi: 10.1093/jamia/ocy132. PMID: 30371862; PMCID: PMC7647152.
2. Kazuhisa Tsunoyama " Creating innovative medicines through collaboration between AI and humans" Science, Mar. 24, 2021, https://www.astellas.com/en/stories/science/aia-japan
3. Avishek Majumder and Priya Chetty, "Application of Applied Statistics in the Pharma Industry" (November 16, 2018), https://www.projectguru.in/pharma-industry-applied-statistics/
4. Sirajudeen, Mohamed Sherif & Siddik, Shaikhji. (2017). Knowledge of Computer Ergonomics among Computer Science Engineering and Information Technology Students in Karnataka, India. Asian Journal of Pharmaceutical Research and Health Care. 9. 64. 10.18311/ajprhc/2017/11023.
5. Sisodia, Swati & Agarwal, Neetima. (2017). Employability Skills Essential for Healthcare Industry. Procedia Computer Science. 122. 431-438. 10.1016/j.procs.2017.11.390.

6. Chetty, P., "Indian drug market drivers that motivate expansion" (4 Nov. 2021)., https://www.projectguru.in/indian-drug-market-drivers-that-motivate-expansion/

7. António Pesqueira, Maria José Sousa, and Álvaro Rocha, "Big Data Skills Sustainable Development in Healthcare and Pharmaceuticals", J Med Syst. (2020); 44(11): 197,

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7544557/

8 Quanticate Team, "Machine Learning in the Pharmaceutical Industry", (May 21, 2019), https://www.quanticate.com/blog/machine-learning-in-the-pharmaceutical-industry

9. Benjamin Obi Tayo, "Data Science Minimum: 10 Essential Skills You Need to Know to Start Doing Data Science". https://www.kdnuggets.com/2020/10/data-science-minimum-10-essential-skills.html

10. Comrey, A. L., & Lee, H. B. (1992). A first course in factor analysis (2nd ed.). Hillsdale, NJ: Erlbaum

11. Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99* (3), 432–442. https://doi.org/10.1037/0033-2909.99.3.432

12. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 141–151. http://dx.doi.org/10.1177/001316446002000116

13. Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245–276. http://dx.doi.org/10.1207/s15327906mbr0102_10

14. Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179 –185. http://dx.doi.org/10.1007/ BF02289447

15. Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. Personnel Psychology, 39, 291–314. http://dx.doi.org/10.1111/ j.1744-6570.1986.tb00583.x

16. Occupational Outlook Handbook (2022), https://www.bls.gov/ooh/healthcare/home.htm

Appendix :

Table 2 Principal Component Analysis

Item	Loading					
	1	2	3	4	5	6
Predictive analytics	0.809					
Tensor Flow	0.806					
Pattern Recognition	0.788					
Apache Spark	0.784					
Data wrangling and Preproceesing	0.780					
Neural Networks	0.765					
Keras	0.764					
Reinforcement Learning	0.751					
NOSQL Database	0.750					
Cloud Environment for ML: Google Colab	0.749					
Big Data Analytics	0.739					
Cloud Environment for ML: Kaggle	0.738					
Cloud Environment for ML: Github	0.720					
Deep Learning	0.708					
Data mining	0.703					
Machine Learning	0.699					
Data Visualization Tableau	0.680					
Data Modelling	0.583					
Block Chain Technology	0.582					

R programming	0.522					
Computer Vision and Image Processing	0.505					
Cyber Security	0.501					
Healthcare regulations		0.781				
Patient care operations		0.771				
Hospital operations		0.766				
Patient rights, privacy protection and legal matters		0.743				
Functions of a health research facility		0.738				
Health Information Management System		0.727				
Business operations		0.718				
Human resource matters		0.710				
Decision making process		0.710				
Financial matters		0.695				
Medical Imaging		0.685				
Organizational Structure / Decision making process		0.682				
Drug Design and Development		0.663				
Genome Sequence Analysis		0.622				
Lifelong learning			0.871			
Team work			0.869			
Decision making			0.855			
Team Building			0.852			
Design Thinking			0.813			
Ethical practices			0.801			
Interpersonal Skills			0.783			
Communication Skills			0.721			
Differential Equations				0.773		
Calculus				0.756		
Linear Algebra				0.748		
Sampling Methods				0.684		
Probability and Statistical distributions				0.660		
Regression and Correlation				0.596		
Hypothesis Testing and Decision making				0.583		
Time Series Analysis				0.565		
Design of Experiments				0.552		
Central tendencies and measures				0.522		
Python with Seaborn					0.695	
Python with MatPlotlib					0.682	
Python with ScikitLearn					0.680	
Python with Pandas					0.677	
Python with Numpy					0.610	
Python with PyTorch					0.598	
DBMS SQL						0.764
Data Structures and Algorithms						0.709
Web Programming and Development						0.603
Cloud Computing						0.545
IoT						0.446

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

Rotation converged in 9 iterations.