IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Statistical Analysis And Principal Component-Based Modeling Of Academic Performance Among Graduate Students In Madhya Pradesh.

AUTHOR-LOKESH PAYASI
CO-AUTHOR-DR GEETA TOMAR

1. ABSTRACT

Academic performance in higher education is shaped by a complex mix of academic and demographic factors. This study explores the underlying dimensions influencing student performance among undergraduate engineering students in South India through structured statistical techniques. A sample of 2,000 students was analyzed using Principal Component Analysis (PCA) and Factor Analysis to reduce dimensionality and uncover latent variables.

The Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin (KMO) measure confirmed the data's suitability for these methods. The analysis identified a small number of components—primarily related to academic engagement and past performance—that explain a significant proportion of the variance in student outcomes. Visual tools such as scree plots, biplots, and parallel analysis were used to aid interpretation and to guide the number of factors retained. The results offer valuable insights for academic institutions aiming to improve student support services and educational planning through evidence-based interventions.

2. DATA COLLECTION

The principal aims of this study are: (i) to construct various relationship models between the chosen features and students' academic performance, subsequently identifying the most effective model, (ii) to ascertain the attributes that most significantly influence performance outcomes, and (iii) to analyze the correlation between variables and establish the robust associations of marks with various factors and attributes. The study aims to develop and evaluate a prediction model that accurately estimates academic success, allowing institutions to proactively identify students in need of academic help and intervention.
A structured questionnaire was created to collect the necessary data, drawing from both established literature and newly proposed aspects pertinent to student success. Twenty-seven essential traits were determined, each converted into pertinent questions that were the foundation of the survey. The questionnaire comprised two primary sections: one detailing personal and demographic information, and the other assessing educational performance across various examination levels. Data were gathered from 2000 undergraduate engineering students. The questionnaire collected a combination of numerical, nominal, and ordinal data types. Examples encompass numerical inputs such as percentage scores, parental income, and age; nominal values including hobbies and parental employment; and ordinal data such as birth order, parental education level, and residence category. The questionnaire was designed to be adaptive, featuring semester-specific questions pertinent to students at various phases of their academic progression.

In addition to static feature analysis, the study examines the possible advantages of incorporating time series analysis to enhance forecast accuracy. Time series methods facilitate the discovery of evolving performance trends by monitoring students' academic records over time, assisting institutions in analyzing, interpreting, and predicting academic trajectories. This method can illuminate both enduring patterns of underachievement and trends of improvement, facilitating data-informed treatments customized to individual learning trajectories. Time series forecasting incorporates a temporal aspect into prediction, providing educators with enhanced insights to facilitate academic planning and individualized student engagement tactics.

3. FEATURE EXTRACTION

This work utilizes feature extraction approaches to tackle the issue of excessive dimensionality in the dataset and to prepare the data for later predictive modeling. Feature extraction is converting raw data into a condensed set of features that are more useful and manageable. Two notable techniques employed for this objective are Principal Component Analysis (PCA) and Factor Analysis. PCA is a statistical method that converts a collection of correlated variables into a reduced set of uncorrelated variables known as principal components, therefore maximizing the explained variance. Factor Analysis aims to elucidate the relationships among observed variables by finding underlying latent components, thus revealing the shared variation among them. PCA and Factor Analysis both simplify data complexity while preserving its fundamental information, hence enhancing model creation efficiency and interpretability.

3.1 Bartlett test of homogeneity of variances

The Bartlett test of homogeneity of variances was conducted to assess the suitability of the data for factor analysis and Principal Component Analysis (PCA).

```
        Bartlett test of homogeneity of variances

data:  predictors
Bartlett's K-squared = 5287.4, df = 25, p-value < 2.2e-16
```

Figure-1

This test assesses the null hypothesis that the variance-covariance matrix of the variables is an identity matrix, indicating that the variables are uncorrelated. Bartlett's K-squared test statistic is derived from the determinant of the correlation matrix. The Bartlett test produced a very significant outcome (Bartlett's K-squared = 5287.4, degrees of freedom (df) = 25, p-value < 2.2e-16). The exceedingly low p-value offers compelling evidence to dismiss the null hypothesis, suggesting that statistically significant correlations exist among the predictor variables. This discovery is essential for factor analysis and PCA, as these methods depend on inter-variable correlations to efficiently reduce dimensionality and uncover underlying structures.
The Bartlett test for homogeneity of variances was performed to evaluate the appropriateness of the data for factor analysis and Principal Component Analysis (PCA). This test assesses the null hypothesis that the variance-covariance matrix of the variables is an identity matrix, indicating that the variables are uncorrelated. The test statistic, Bartlett's K-squared, is derived from the determinant of the correlation matrix. The Bartlett test produced a very significant outcome (Bartlett's K-squared = 5287.4, degrees of freedom (df) = 25, p-value < 2.2e-16). The exceedingly low p-value offers compelling evidence to reject the null hypothesis, suggesting that statistically significant correlations exist among the predictor variables. This discovery is essential for factor analysis and PCA, as these methods depend on inter-variable correlations to efficiently reduce dimensionality and reveal underlying structures.

3.2 Kaiser-Meyer-Olkin (KMO) test

The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy is a widely used diagnostic tool in multivariate statistics, particularly in factor analysis and principal component analysis (PCA), to evaluate the suitability of the data for structure detection. It specifically assesses whether the partial correlations among variables are small, which is a desirable property for factor extraction. Conceptually, the KMO statistic compares the magnitude of observed correlation coefficients to the magnitude of partial correlation coefficients.

The KMO value ranges from 0 to 1. A value closer to 1 indicates that a large proportion of variance in the variables can be attributed to common underlying factors, suggesting that factor analysis is appropriate and is likely to yield distinct and reliable factors. On the other hand, a KMO value closer to 0 implies that most of the variance is unique (specific to individual variables) or random (error variance), making factor analysis less effective or even inappropriate (Kaiser, 1974).

According to commonly accepted thresholds:

- KMO ≥ 0.90 is considered superb,
- 0.80–0.89 is great,
- 0.70–0.79 is good,
- 0.60–0.69 is mediocre,
- 0.50–0.59 is poor, and
- below 0.50 is unacceptable for factor analysis.

In practice, both the overall KMO for the entire model and the individual KMO values for each variable are examined. If the individual KMO values are low, even if the overall KMO is acceptable, it may be advisable to remove or revise those variables to improve model adequacy.

The test is typically reported alongside Bartlett's Test of Sphericity, which assesses whether the correlation matrix significantly differs from an identity matrix (i.e., where variables are uncorrelated). Together, these tests help determine whether factor analysis is statistically justified and methodologically sound.

In summary, a high KMO value supports the feasibility of factor analysis by confirming the presence of underlying latent structures in the data. It ensures that the variables are sufficiently interrelated, thereby facilitating meaningful factor extraction and dimensionality reduction.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = predictors)
Overall MSA =  0.8
MSA for each item =
   Gender       Age Education    Income        A1        A2        A3
     0.53      0.54      0.46      0.50      0.64      0.60      0.82
       A4        A5        B1        B2        B3        B4        B5
     0.71      0.45      0.83      0.83      0.75      0.81      0.89
       C1        C2        C3        C4        C5        D1        D2
     0.83      0.82      0.84      0.87      0.82      0.87      0.86
       D3        D4        E1        E2        E3
     0.87      0.85      0.83      0.86      0.77
```

Figure-2

The Kaiser-Meyer-Olkin (KMO) test was carried out to evaluate the adequacy of the sample for conducting factor analysis and Principal Component Analysis (PCA). The overall KMO value obtained was 0.80, which falls into the *"meritorious"* category as per Kaiser's classification (Kaiser, 1974). This indicates that the dataset possesses a sufficient amount of shared variance among variables, making it appropriate for dimensionality reduction techniques.

A closer look at the individual Measures of Sampling Adequacy (MSA) revealed variability across the variables. Variables such as C4, D1, D2, D3, and E2 demonstrated high MSA values, suggesting they are well predicted by other variables in the dataset and are thus well-suited for inclusion in factor extraction. However, Gender, Age, Education, Income, and A5 recorded relatively lower MSA values, indicating that these variables contribute less to the shared variance and may not align as strongly with the underlying factor structure.

Although these lower-scoring variables should be interpreted with caution in the context of factor analysis, the overall KMO value supports the continuation of PCA or factor analysis for this dataset. These findings align with established guidelines for factor analysis, which recommend KMO values of 0.6 or above as acceptable, and 0.8 or above as ideal for robust analysis (Field, 2013; Hair et al., 2010).

3.3 Principal Component Analysis

To uncover latent structures within the data and reduce multicollinearity among predictors, Principal Component Analysis (PCA) was performed on the dataset. The PCA yielded 26 principal components, each representing a linear combination of the original variables. The standard deviation associated with each component reflects the amount of variance explained. Notably, the first principal component (PC1) had the highest standard deviation (2.411), suggesting it captures the largest proportion of variance across the dataset. The standard deviations of subsequent components decreased progressively, indicating a diminishing contribution to overall variance—consistent with the common behavior observed in dimensionality reduction (Jolliffe & Cadima, 2016).

This pattern validates the assumption that a smaller subset of principal components can retain most of the informative structure of the data. The number of components to retain for further analysis can be guided by the Kaiser criterion (eigenvalues > 1), cumulative variance threshold (typically 70–80%), and visual inspection of the scree plot for the 'elbow' point where the explained variance levels off (Abdi & Williams, 2010).

An examination of the component loading matrix provided additional insights. High and moderate loadings on PC1 were observed for variables such as A3 (equal opportunities), B5 (support from teachers/peers), C3 (goal-setting), and D2 (age appropriateness), indicating that PC1 may represent a broad construct related to student engagement, support, and readiness. Conversely, PC2 exhibited strong loadings from Gender and Age, highlighting its relevance to demographic structure within the data. The remaining components, while contributing less to the total variance, captured more specific and isolated relationships among subsets of variables.

Overall, the PCA not only simplified the data structure but also offered interpretable dimensions for downstream modeling. This approach aligns with the principle of parsimony, balancing dimensionality

reduction with information preservation. The derived components can be used to enhance the robustness and interpretability of machine learning or regression-based prediction models.

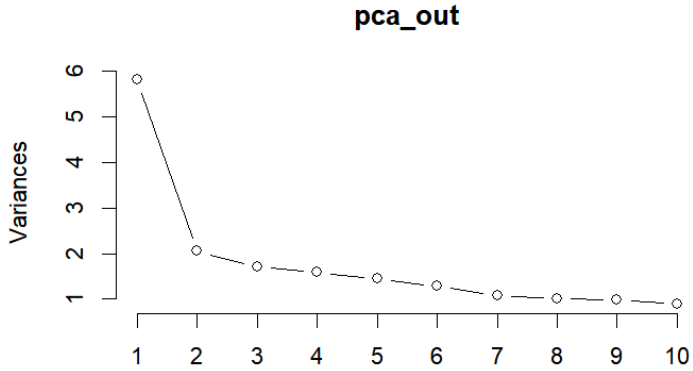

Figure-3:Scree Plot

3.4 Scree Plot interpretation

To determine the optimal number of principal components to retain, a scree plot was examined (Figure 3). The scree plot visualizes the variance (eigenvalue) associated with each principal component, plotted in descending order. As illustrated in Figure 3, a sharp decline in variance is observed from the first to the second component, indicating that the first principal component explains a substantial portion of the total variance. Beyond the second or third component, the slope of the curve flattens, forming a distinct 'elbow.' This elbow suggests a point of diminishing returns, where subsequent components contribute relatively little to the overall variance explained. Based on the scree plot, retaining the first two or three principal components appears to be a reasonable strategy, effectively balancing dimensionality reduction with the preservation of a significant amount of information from the original variables. This decision aligns with the principle of parsimony, aiming to achieve a simplified representation of the data while minimizing information loss.

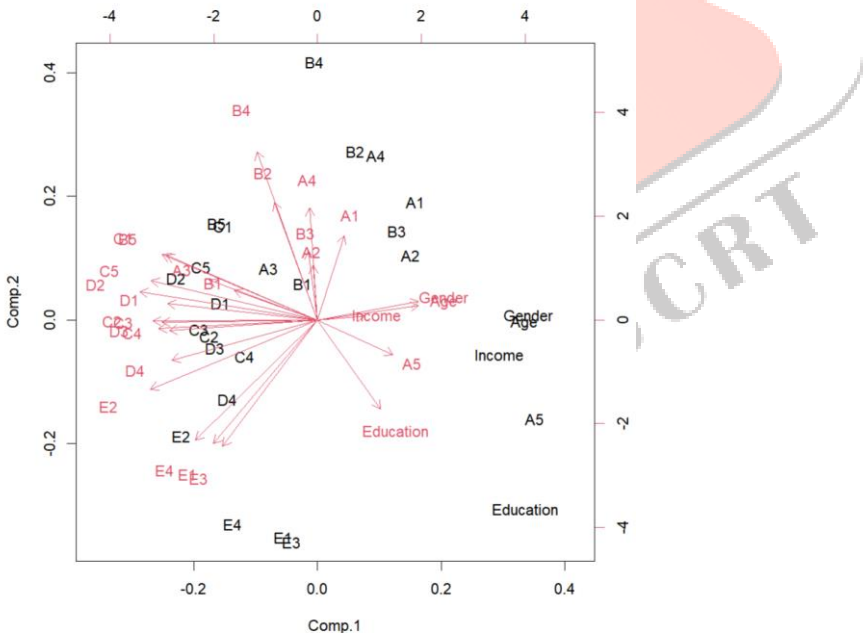

Figure-4: Biplot

3.5 Biplot interpretation

Principal Component Analysis (PCA) was applied in this study to explore the relationships between academic variables (such as attendance, 10th and 12th grade marks, and graduation performance) and demographic factors (including age, gender, income, and education level). The PCA biplot reveals that graduation performance (E4) is positively aligned with academic factors like 12th grade marks (E3), 10th grade marks (E2), and attendance (E1), suggesting that these variables significantly influence graduation outcomes. In contrast, demographic variables such as education level, income, and age are oriented in the opposite direction, indicating a potential inverse relationship with academic achievement. The length and direction of vectors in the biplot further highlight that academic variables contribute more strongly to the first two principal components, implying that they account for a larger proportion of variance in graduation performance compared to demographic variables. This analysis supports the conclusion that students' prior academic records and class engagement are more predictive of graduation success than socio-demographic characteristics in the observed sample.

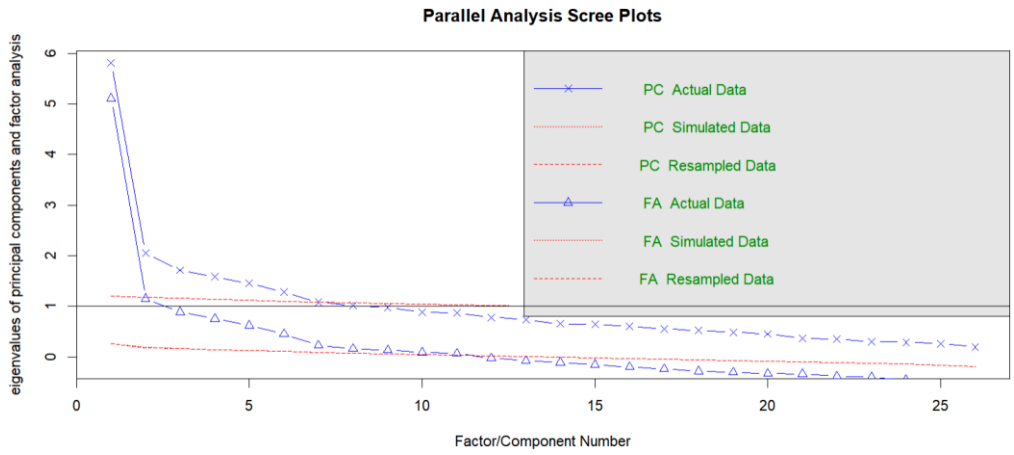

Figure-5: Parallel Analysis

3.6 Parallel Plot Analysis

A parallel analysis was conducted to determine the optimal number of components to retain in the principal component and factor analysis models. Parallel analysis suggests that the number of factors = 11 and the number of components = 6. The scree plot clearly indicates that the first three components/factors have eigenvalues exceeding those derived from simulated and resampled data, confirming their significance. Beyond the third component, the eigenvalues fall below the random data threshold, indicating that additional components do not explain meaningful variance. Therefore, a three-factor structure is appropriate for interpreting the underlying dimensions in the dataset. This conclusion supports the findings from the initial PCA and strengthens the dimensionality reduction strategy adopted in the study.

4. CONCLUSION

This research employed Principal Component Analysis (PCA) and Factor Analysis to analyze factors influencing the academic performance of engineering students. Statistical tests including Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) measure indicated strong inter-variable correlations and adequate sampling, supporting the use of these dimensionality reduction techniques.

The PCA revealed that a few principal components, particularly those associated with academic variables like attendance and prior academic scores, explain a large portion of the overall variance in graduation performance. Visualizations like scree plots, biplots, and parallel analysis confirmed that retaining two to three components provided a meaningful and simplified structure of the data.

This study contributes to educational research by emphasizing the importance of reducing data complexity while preserving interpretability. The insights gained can help institutions better understand which academic and personal variables most significantly affect student performance. Importantly, no predictive machine learning models were used; the focus remained on statistical understanding and data simplification through PCA and Factor Analysis, offering a foundational tool for further educational research and policy development.

5. REFERENCE

- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. https://doi.org/10.1007/BF02291575
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. https://doi.org/10.1007/BF02291575
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage Publications.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Pearson Education.
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. https://doi.org/10.1002/wics.101
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75(3), 417–453. https://doi.org/10.3102/00346543075003417

- Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53(1), 371–399. https://doi.org/10.1146/annurev.psych.53.100901.135233
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis* (7th ed.). Pearson Education.
- Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Sage Publications.
- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), 433–459. https://doi.org/10.1002/wics.101
- Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202