STATISTICS II

CA 1

Vanshika Sharma

How do socioeconomic, behavioral, and resource-related factors collectively predict academic performance as measured by exam scores

Vanshika Sharma
BSc(hons) in Data Science
School of Computing
Dublin, Ireland
x23198389@student.ncirl.ie

I. INTRODUCTION

The subject of this analysis is to identify the factors that influence students' exam scores. Academic performance is an important measure of success, and understanding these factors can help improve learning outcomes. This study focuses on analysing a dataset of 100 students to find the key variables that affect their exam scores.

The dataset contains various factors such as hours studied, attendance, access to resources, motivation level, and extracurricular activities. The goal is to determine which factors have the strongest impact on students' exam performance.

The research question for this analysis is: **"What are the main factors that significantly influence students' exam scores?"**

To answer this question, the study uses statistical methods like correlation analysis. The primary objective is to find the relationships between exam scores and other variables in the dataset. This analysis will focus on identifying the most important predictors of academic success. By understanding these key factors, we aim to provide recommendations for improving students' performance.

This report will present the findings in a structured way, showing the analysis results and the interpretation of the data. The focus is on clarity and simplicity to ensure that the results are easy to understand and apply.

II. BACKGROUND

Regression analysis and Principal Component Analysis (PCA) are key tools in data analytics. Regression analysis helps us understand relationships between variables. It allows us to predict one variable based on others. For example, we can predict exam scores using factors like hours studied and attendance.

PCA is a technique used to reduce the number of variables in a dataset. It transforms original variables into new ones called principal components. These components capture most of the information in the data. PCA helps simplify complex datasets while retaining important patterns.

Our dataset was downloaded from Kaggle[1]. It contains information about students' academic performance and factors that might affect it. Variables include hours studied, attendance, and exam scores.

In our analysis, we used regression to model the relationship between exam scores and other factors. We applied PCA to reduce the number of variables, focusing on the most significant ones. This approach improves prediction accuracy and makes the analysis more manageable.

Many studies have used regression and PCA in education research. Researchers have found that regression can identify key factors like study habits and attendance that predict academic success. PCA has been used to simplify complex educational data, highlighting the main influences on learning outcomes. By combining these methods, we can get a clearer picture of what affects exam scores.

In a study by Alshanqiti[2], a hybrid regression model and multi-label classifier was built to predict student performance and identify key factors. Using advanced methods, it improves accuracy and helps improve educational programs through targeted interventions.

Similarly, a journal by Ofori et al[3] highlights the importance of accurate predictions and analyzing socioeconomic factors to guide learning improvements, using machine learning to predict and improve student performance.

Moreover, in a paper published by Albreiki[4], the study reviews Educational Data Mining (EDM) from 2009 to 2021, focusing on machine learning methods to predict student dropouts and risks. These techniques help improve student performance and address challenges in education.

In relation to PCA, the model by Liu et el[5] aligns with the goal of identifying important features for analysis which integrates both feature types using a recurrent neural network with attention mechanisms, achieving 98% accuracy. PCA simplifies data by reducing dimensions, helping to focus on the most impactful features.

A study by Zhou[6] introduces a reliable method for high-dimensional analysis by combining kernel PCA (KPCA) for dimension reduction with Gaussian process regression (GPR). It uses active learning and Monte Carlo simulation for accurate reliability estimation.

III. DATA DESCRIPTION

The dataset for this analysis was sourced from Kaggle, originally containing a large amount of data with 6,000 entries. The description of the original data is as follows:

	Format	*Size*	*N(Records)*	*N(Features)*
Value	csv	642 kB	6,607	20

To make the dataset more manageable and focused for analysis, a Python script was used to create a random sample of 100 entries. This sampling process ensured the data was easier to work with while still retaining a representative portion of the original dataset. This smaller, cleaned dataset provided a strong foundation for exploring the relationships between variables and drawing meaningful insights. The str()

function in R was used to inspect the structure of the dataset, showing that it consisted of 20 variables. These variables were categorized as follows:

Variable Type	Example Variables	Count
Numeric	Hours_Studied, Attendance	7
Categorical	Gender, Internet_Access	8
Ordinal	Motivational_Level, Family_Income	5

Data cleaning methods were applied to ensure accuracy and consistency. Missing values were handled, numeric variables were checked for consistency, and formats were standardized using the scale() function. Standardization ensures that all variables contribute equally to the analysis by scaling them to have a mean of 0 and a standard deviation of 1. Outliers were inspected to prevent bias in the analysis. The dataset was then prepared for advanced analysis by standardizing numeric variables to a common scale, ensuring that all features contributed equally to the results. This cleaned dataset provided a solid basis for exploring relationships and drawing meaningful insights.

The data cleaning process involved several steps to prepare the dataset for analysis. Categorical variables like Extracurricular_Activities, Internet_Access, and Learning_Disabilities were converted into numeric representations (e.g., "Yes" to 1 and "No" to 0). Ordered variables such as Motivation_Level and Access_to_Resources were encoded as numeric levels (e.g., Low = 1, Medium = 2, High = 3). Variables irrelevant to the analysis, such as Distance_from_Home, were excluded to focus on factors that directly impact exam scores. These steps ensured consistency and usability for correlation analysis.

The analysis was done using R, which has many tools for data analysis and visualization. The dplyr package was used to clean and organize the dataset. The psych package helped calculate descriptive statistics, such as mean and standard deviation, for variables like Hours_Studied, Attendance, and Exam_Score. The ggplot2 package was used to create visualizations to show patterns in the data. The caret package was used to standardize the data, so all variables had the same scale. The stats package was used to compute the correlation matrix and perform Principal Component Analysis (PCA). These tools ensured accurate analysis.

IV. METHODOLOGY AND CALCULATIONS

The analysis was conducted using the **Knowledge Discovery in Databases (KDD)** methodology, which consists of the following steps:

1. Data Selection: A random sample of 100 entries was selected from the original dataset to ensure it was manageable while remaining representative. The dataset includes variables like Hours_Studied, Attendance, Motivation_Level, Access_to_Resources, and Exam_Score. These variables were selected for their relevance to understanding factors that influence academic performance.

```
> head(data)
  Hours_Studied Attendance Parental_Involvement Access_to_Resources Extracurricular_Activities Sleep_Hours Previous_Scores Motivation_Level
1            20         71               Medium                 Low                         No           7              87             High
2            22         71               Medium                 Low                        Yes           7              98              Low
3            21         91                 High              Medium                        Yes           6              53             High
4            12         91               Medium                 Low                        Yes           8              81              Low
5            21         63                  Low                High                        Yes           8              95           Medium
6            21         79                  Low              Medium                        Yes           7              84              Low
  Internet_Access Tutoring_Sessions Family_Income Teacher_Quality School_Type Peer_Influence Physical_Activity Learning_Disabilities
1             Yes                 1        Medium          Medium      Public       Negative                 5                    No
2             Yes                 2           Low            High      Public        Neutral                 2                    No
3             Yes                 1        Medium          Medium      Public       Positive                 3                    No
4             Yes                 0           Low             Low      Public       Positive                 4                    No
5             Yes                 2          High          Medium      Public        Neutral                 5                    No
6             Yes                 2        Medium          Medium     Private        Neutral                 3                    No
  Parental_Education_Level Distance_from_Home Gender Exam_Score
1              High School               Near   Male         65
2              High School           Moderate Female         65
3              Postgraduate              Near Female         71
4              High School           Moderate   Male         64
5              High School               Near   Male         66
6              High School               Near   Male         66
```

Fig 1. Output 1

2. Data Preprocessing: The dataset was inspected using the str() function to check its structure and ensure it was ready for analysis. Several cleaning steps were performed:

- **Encoding Variables:** Categorical variables such as Extracurricular_Activities, Internet_Access, and Learning_Disabilities were converted to numeric values (e.g., "Yes" to 1, "No" to 0). Similarly, Gender was encoded as 1 for "Male" and 0 for "Female".
- **Ordinal Variables:** Ordered variables like Motivation_Level and Access_to_Resources were encoded as numeric levels (Low = 1, Medium = 2, High = 3).
- **Removing Irrelevant Variables:** Columns unrelated to the research question, such as Distance_from_Home, were not used to focus on relevant factors only.
- **Handling Missing Values:** Any missing or inconsistent data was deleted to ensure a clean dataset.

3. Descriptive Statistics: Descriptive statistics were generated using the describe() function to summarize key metrics for the variables. Here are some highlights:

- Hours_Studied had a mean of 19.94 hours with a standard deviation of 6.44, indicating variation in study time among students.
- Attendance had a high mean of 80.14%, with most students maintaining good attendance.
- Exam_Score had a mean of 67.27, ranging from 55 to 89, with a slight positive skew, showing that most scores were on the higher side.

This analysis provided a clear understanding of the dataset's structure and distributions.

```
> #Generate descriptive statistics
> describe(data)
                          vars   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
Hours_Studied                1 100 19.94  6.44   20.0   19.86  5.19   3  39    36  0.04     0.90 0.64
Attendance                   2 100 80.14 12.67   79.0   80.25 17.79  60 100    40  0.00    -1.41 1.27
Parental_Involvement*        3 100  2.37  0.82    3.0    2.46  0.00   1   3     2 -0.76    -1.11 0.08
Access_to_Resources          4 100  2.16  0.69    2.0    2.20  0.74   1   3     2 -0.22    -0.94 0.07
Extracurricular_Activities   5 100  0.53  0.50    1.0    0.54  0.00   0   1     1 -0.12    -2.01 0.05
Sleep_Hours                  6 100  6.83  1.55    7.0    6.75  1.48   4  10     6  0.31    -0.57 0.16
Previous_Scores              7 100 76.46 14.66   78.5   76.89 19.27  51 100    49 -0.15    -1.27 1.47
Motivation_Level             8 100  1.90  0.73    2.0    1.88  1.48   1   3     2  0.15    -1.14 0.07
Internet_Access              9 100  0.92  0.27    1.0    1.00  0.00   0   1     1 -3.05     7.38 0.03
Tutoring_Sessions           10 100  1.27  1.02    1.0    1.19  1.48   0   4     4  0.57    -0.38 0.10
Family_Income               11 100  1.71  0.80    1.5    1.64  0.74   1   3     2  0.55    -1.22 0.08
Teacher_Quality*            12 100  3.10  0.97    4.0    3.14  0.00   1   4     3 -0.26    -1.75 0.10
School_Type                 13 100  0.76  0.43    1.0    0.82  0.00   0   1     1 -1.20    -0.57 0.04
Peer_Influence*             14 100  2.11  0.72    2.0    2.14  1.48   1   3     2 -0.16    -1.10 0.07
Physical_Activity           15 100  3.10  0.95    3.0    3.09  1.48   0   5     5 -0.13     0.03 0.09
Learning_Disabilities       16 100  0.12  0.33    0.0    0.03  0.00   0   1     1  2.30     3.34 0.03
Parental_Education_Level*   17 100  2.89  0.75    3.0    2.89  1.48   1   4     3 -0.11    -0.64 0.08
Distance_from_Home*         18 100  3.50  0.72    4.0    3.64  0.00   1   4     3 -1.38     1.49 0.07
Gender                      19 100  0.64  0.48    1.0    0.68  0.00   0   1     1 -0.57    -1.69 0.05
Exam_Score                  20 100 67.27  4.28   67.0   67.15  4.45  55  89    34  1.08     5.25 0.43
```

Fig 2. R Script's Output for Descriptive Statistics

4. Correlation Analysis: The correlation analysis focused on identifying factors that influence Exam_Score, as it represents student performance. The results showed that Attendance had the highest positive correlation with Exam_Score (0.558), followed by Hours_Studied (0.478). Motivation_Level (0.228) and Access_to_Resources (0.199) also showed positive correlations, though weaker. These findings indicate that regular attendance and consistent study habits are the most important factors for better performance, while motivation and access to resources have a smaller but positive impact. This analysis only included correlations with

Exam_Score to stay aligned with the research objective of understanding factors affecting academic performance.

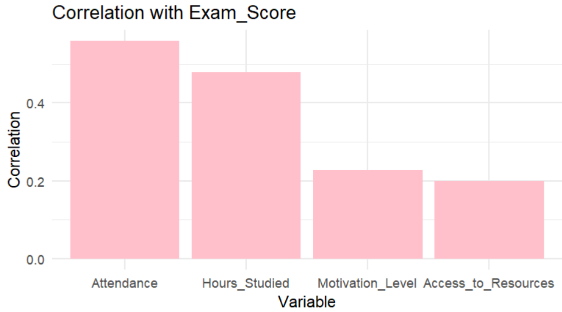

Fig 3. Correlation vs Exam Score

A correlation matrix was computed using the cor() function to examine the relationships between variables. These results helped identify the most important factors influencing exam performance. Multicollinearity was not checked as the correlation analysis showed weak relationships between the predictors.

Fig 4. R's Output for Correlation

```
> # Step 1: Select only numeric columns for correlation analysis
> numeric_data <- data[sapply(data, is.numeric)]
>
> # Step 2: Compute the correlation matrix
> cor_matrix <- cor(numeric_data)
>
> # Step 3: Extract correlations with Exam_Score
> exam_score_correlations <- cor_matrix[,"Exam_Score"]
>
> # Step 4: Print significant correlations
> print(exam_score_correlations)
      Hours_Studied          Attendance Access_to_Resources Extracurricular_Activities          Sleep_Hours
        0.478062979         0.557958055         0.199827597               -0.024954298         -0.096384375
    Previous_Scores    Motivation_Level     Internet_Access          Tutoring_Sessions        Family_Income
       -0.059069773         0.227781884         0.018674653                0.172070381          0.032108032
        School_Type   Physical_Activity Learning_Disabilities                   Gender           Exam_Score
       -0.046791275        -0.009199081         0.019921169               -0.020914151          1.000000000
```

5. Data Transformation: Standardization was applied to all numerical variables using the scale() function. This step ensured that all variables were on the same scale, with a mean of 0 and a standard deviation of 1. Standardization prevented variables with larger magnitudes (e.g., Hours_Studied) from dominating smaller-scale variables (e.g., Motivation_Level). This transformation was necessary for accurate Principal Component Analysis (PCA).

6. Data Mining: Principal Component Analysis (PCA) was performed using the prcomp() function to reduce dimensionality and identify key components that explain most of the variance in the data. The PCA results showed that the first three principal components captured most of the variance. These components were used for further regression analysis.

Regression analysis was conducted to predict Exam_Score based on the principal components. The model explained a significant portion of the variance, with an R-squared value indicating its effectiveness.

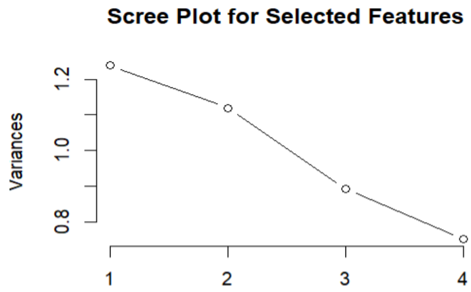

Fig 5. Screeplot

7. Evaluation: The regression model's accuracy was evaluated by plotting actual vs. predicted exam scores using ggplot2. The results showed a strong alignment between the actual and predicted values, confirming the reliability of the model. The residual sum of squares (RSS) and total sum of squares (TSS) were calculated to validate the model's performance, with an R-squared value providing a clear measure of fit.

V. RESULTS

The correlation heatmap visually represents the relationships between various factors and Exam_Score, with red indicating positive correlations and blue showing negative ones.

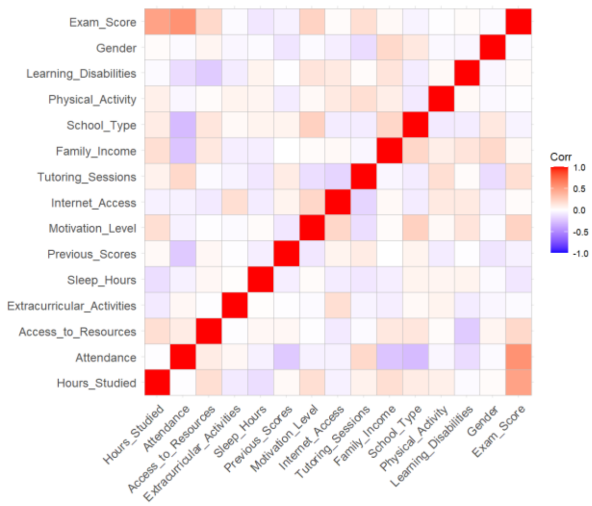

Fig 6. Heatmap

The analysis highlights that Attendance has the strongest positive correlation with Exam_Score (0.558), followed by Hours_Studied (0.478), emphasizing the importance of regular attendance and consistent study habits for better academic performance. Additionally, Motivation_Level (0.228) and Access_to_Resources (0.199) show weaker positive correlations, suggesting they contribute to performance but are less impactful. Factors like Learning_Disabilities and Gender show little to no correlation with Exam_Score, indicating they have minimal influence on student outcomes. This heatmap provides a clear visual summary of the key factors affecting exam performance, aligning with the research goal of identifying critical influences on academic success. The dataset used in this project contains various factors that may influence exam scores. The first few rows of the data are shown in **figure**. It includes both numerical and categorical variables such as Hours_Studied, Attendance, Previous_Scores, and

Motivation_Level. These features will help in understanding how different factors affect student performance.

PCA reduced the four features to three principal components, explaining 81.23% of the data's variance. PC1 explains 30.96%, PC2 explains 27.98%, and PC3 explains 22.29%. The fourth component was excluded as it explains only 18.77%. The first three components were used for further analysis.

```
> # Step 1: Standardize the data (excluding Exam_Score)
> data_scaled <- scale(selected_features %>% select(-Exam_Score))
> # Step 2: Perform PCA
> pca_result <- prcomp(data_scaled, center = TRUE, scale. = TRUE)
> # Step 3: PCA Summary
> summary(pca_result)  # Show variance explained by components
Importance of components:
                          PC1    PC2    PC3    PC4
Standard deviation     1.1128 1.0580 0.9442 0.8665
Proportion of Variance 0.3096 0.2798 0.2229 0.1877
Cumulative Proportion  0.3096 0.5894 0.8123 1.0000
```

Fig 7. PCA Summary

The initial model had an adjusted R-squared of 0.5212 and was inaccurate due to an outlier. After removing the outlier, the adjusted R-squared improved to 0.7945, increasing accuracy.

```
> # Fit a linear regression model using the principal components
> model <- lm(Exam_Score ~ ., data = regression_data)
> # Display the summary of the regression model
> summary(model)

Call:
lm(formula = Exam_Score ~ ., data = regression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7066 -1.2327 -0.2088  0.8288 23.2381

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.2700     0.2965 226.879  < 2e-16 ***
PC1           2.0044     0.2678   7.485 3.44e-11 ***
PC2          -1.3150     0.2817  -4.669 9.84e-06 ***
PC3           1.8113     0.3156   5.739 1.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.965 on 96 degrees of freedom
Multiple R-squared:  0.5357,	Adjusted R-squared:  0.5212
F-statistic: 36.92 on 3 and 96 DF,  p-value: 5.903e-16
```

Fig 8. Summary of Old Linear Model

```
> summary(new_model)

Call:
lm(formula = Exam_Score ~ ., data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9597 -1.0951  0.2027  0.9308  3.5251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.0217     0.1685 397.639  < 2e-16 ***
PC1           2.0551     0.1515  13.568  < 2e-16 ***
PC2          -1.7980     0.1628 -11.045  < 2e-16 ***
PC3           1.5274     0.1796   8.506 2.55e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.677 on 95 degrees of freedom
Multiple R-squared:  0.8008,	Adjusted R-squared:  0.7945
F-statistic: 127.3 on 3 and 95 DF,  p-value: < 2.2e-16
```

Fig 9. Summary of New Linear Model

The linear regression model was evaluated using diagnostic plots. The Residuals vs Fitted plot showed random spread, confirming the linear relationship. The Q-Q plot indicated that residuals follow a normal distribution. The Scale-Location plot showed consistent variance, meeting the homoscedasticity assumption. The Residuals vs Leverage plot revealed no influential points affecting the model. These results confirm that the model assumptions are satisfied, making it reliable for predicting Exam_Score.

VI. Discussion and Analysis of Results

The final regression model effectively explains the relationship between the principal components and the target variable, Exam_Score.

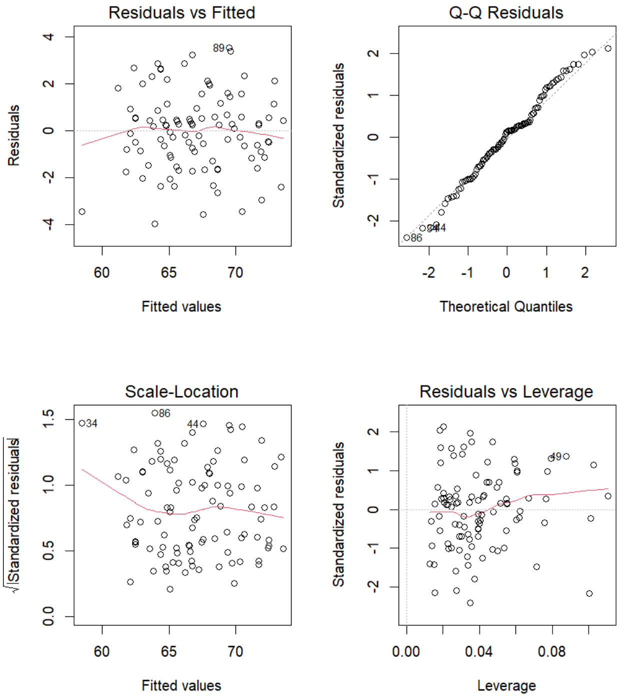

Fig 10. Plot

The adjusted R-squared value of **0.7945** indicates that the model explains approximately **79.45%** of the variation in Exam_Score, showing strong predictive power.

The coefficients provide important insights:

- **PC1 (2.0551)**: Positively influences Exam_Score, indicating that higher values in this component, which may represent attendance and study habits, lead to better exam performance.
- **PC2 (-1.7980)**: Negatively impacts Exam_Score, suggesting that this component may capture factors that hinder performance, such as limited resources or low motivation.
- **PC3 (1.5274)**: Has a positive effect, highlighting the role of additional contributing factors in improving scores.

The significant p-values (<0.001) for all components confirm that these relationships are statistically significant.

This model aligns well with the research objective of identifying factors affecting student performance. The results show that improving attendance and study habits can significantly boost performance, while addressing negative influences like lack of resources is equally important. These findings can guide interventions to help students achieve better academic outcomes. The model successfully forecasts Exam_Score based on key contributing factors.

REFERENCES

[1] "Find Open Datasets and Machine Learning Projects | Kaggle." Accessed: Nov. 13, 2024. [Online]. Available: https://www.kaggle.com/datasets

[2] A. Alshanqiti and A. Namoun, "Predicting Student Performance and Its Influential Factors Using Hybrid Regression and Multi-Label Classification," *IEEE Access*, vol. 8, pp. 203827–203844, 2020, doi: 10.1109/ACCESS.2020.3036572.

[3] F. Ofori, D. E. Maina, and D. Rhoda, "Students' Performance and Improve Learning," vol. 4, no. 1.

[4] B. Albreiki, N. Zaki, and H. Alashwal, "A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques," *Educ. Sci.*, vol. 11, no. 9, Art. no. 9, Sep. 2021, doi: 10.3390/educsci11090552.

[5] D. Liu, Y. Zhang, J. Zhang, Q. Li, C. Zhang, and Y. Yin, "Multiple Features Fusion Attention Mechanism Enhanced Deep Knowledge Tracing for Student Performance Prediction," *IEEE Access*, vol. 8, pp. 194894–194903, 2020, doi: 10.1109/ACCESS.2020.3033200.

[6] T. Zhou and Y. Peng, "Kernel principal component analysis-based Gaussian process regression modelling for high-dimensional reliability analysis," *Comput. Struct.*, vol. 241, p. 106358, Dec. 2020, doi: 10.1016/j.compstruc.2020.106358.