Predicting Student Success Using Regression Analysis

Overview

This project explores the factors that influence student success and builds a predictive model using regression analysis to classify students as "successful" or "unsuccessful." By analyzing key predictors such as study hours, attendance, and parental involvement, the model achieved an accuracy of **82%**.

Objective

- **Identify Critical Factors**:
  Analyze the key features influencing student success, such as study hours, attendance, and parental involvement.
- **Develop a Predictive Model**:
  Build a regression-based model to categorize students as "successful" or "unsuccessful" based on their input features.
- **Make Predictions**:
  Predict student outcomes using the developed model by applying it to new or unseen data.

Dataset

Source: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors
The dataset includes the following features:

1. **Numeric Features**:
   - Hours Studied
   - Attendance (%)
   - Sleep Hours
   - Previous Scores
   - Physical Activity
   - Tutoring Sessions
2. **Categorical Features**:
   - Parental Involvement (Low, Medium, High)
   - Motivation Level (Low, Medium, High)
   - School Type (Public, Private)
   - Peer Influence (Positive, Neutral, Negative)
   - Internet Access (Yes, No)
   - Extracurricular Activities (Yes, No)
   - Learning Disabilities (Yes,No)

    - Teacher Quality (High, Medium, Low)
    - Distance from Home (Near, Moderate, Far)
    - Parental Education Level (High School, College, Postgraduate)
    - *Excluded Variables*: Certain variables, such as `Gender`, were not converted or included in the predictive model as they were not statistically significant predictors of exam scores and were deemed more relevant for demographic analysis rather than modeling purposes.

The dependent variable is **Exam Score**, which indicates student success. Success is defined as a score of **70 or higher**.

Methodology

Step 1: Data Cleaning

- **Handled Missing Values**:
    - Replaced missing numeric values with the column mean.
    - Filled missing categorical values with the most common value using pivot tables.
- **Standardized Numeric Data**:
    - Converted numeric columns to a standardized scale using:

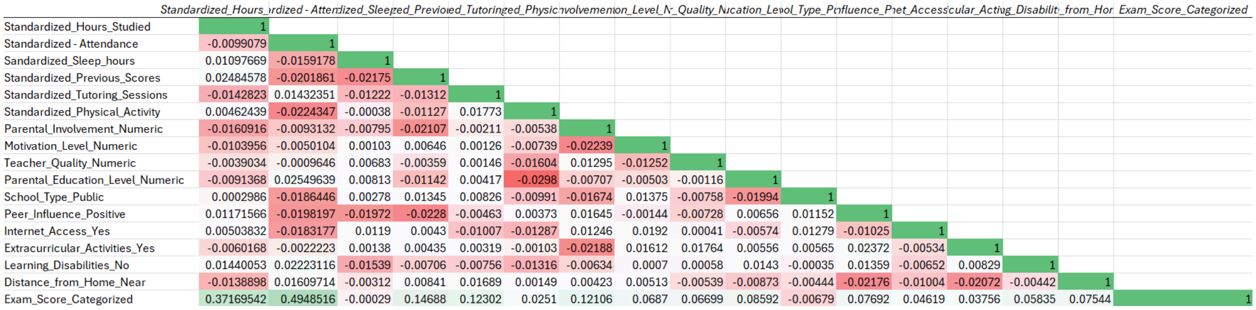
Standardized Value = (X−Mean) / Standard Deviation

Step 2: Encoding Categorical Data

- **Ordinal Encoding**: For variables with a natural order, such as:
    - Parental Involvement: Low = 1, Medium = 2, High = 3.
    - Motivation Level: Low = 1, Medium = 2, High = 3.
- **One-Hot Encoding**: For variables with no natural order, such as:
    - School Type: Public = 1, Private = 0.

Step 3: Correlation Analysis

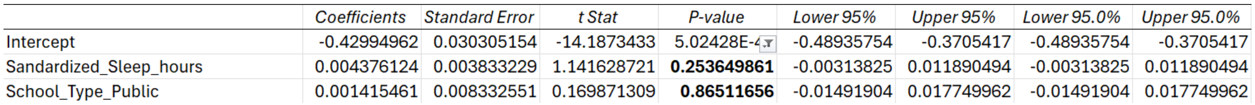
- Performed correlation analysis to identify the most impactful variables.
- Excluded features with weak correlations or multicollinearity issues.



	Standardized_Hours	*ırdized - Atten*	*dized_Sleep*	*zed_Previo*	*ed_Tutoring*	*zed_Physic*	*nvolvemen*	*on_Level_N*	*_Quality_N*	*ucation_Lev*	*ol_Type_P*	*nfluence_P*	*net_Access*	*cular_Activ*	*g_Disabilit*	*_from_Hor*	*Exam_Score_Categorized*
Standardized_Hours_Studied	1																
Standardized - Attendance	-0.0099079	1															
Sandardized_Sleep_hours	0.01097669	-0.0159178	1														
Standardized_Previous_Scores	0.02484578	-0.0201861	-0.02175	1													
Standardized_Tutoring_Sessions	-0.0142823	0.01432351	-0.01222	-0.01312	1												
Standardized_Physical_Activity	0.00462439	-0.0224347	-0.00038	-0.01127	0.01773	1											
Parental_Involvement_Numeric	-0.0160916	-0.0093132	-0.00795	-0.02107	-0.00211	-0.00538	1										
Motivation_Level_Numeric	-0.0103956	-0.0050104	0.00103	0.00646	0.00126	-0.00739	-0.02239	1									
Teacher_Quality_Numeric	-0.0039034	-0.0009646	0.00683	-0.00359	0.00146	-0.01604	0.01295	-0.01252	1								
Parental_Education_Level_Numeric	-0.0091368	0.02549639	0.00813	-0.01142	0.00417	-0.0298	-0.00707	-0.00503	-0.00116	1							
School_Type_Public	0.0002986	-0.0186446	0.00278	0.01345	0.00826	-0.00991	-0.01674	0.01375	-0.00758	-0.01994	1						
Peer_Influence_Positive	0.01171566	-0.0198197	-0.01972	-0.0228	-0.00463	0.00373	0.01645	-0.00144	-0.00728	0.00656	0.01152	1					
Internet_Access_Yes	0.00503832	-0.0183177	0.0119	0.0043	-0.01007	-0.01287	0.01246	0.0192	0.00041	-0.00574	0.01279	-0.01025	1				
Extracurricular_Activities_Yes	-0.0060168	-0.0022223	0.00138	0.00435	0.00319	-0.00103	-0.02188	0.01612	0.01764	0.00556	0.00565	0.02372	-0.00534	1			
Learning_Disabilities_No	0.01440053	0.02223116	-0.01539	-0.00706	-0.00756	-0.01316	-0.00634	0.0007	0.00058	0.0143	-0.00035	0.01359	-0.00652	0.00829	1		
Distance_from_Home_Near	-0.0138898	0.01609714	-0.00312	0.00841	0.01689	0.00149	0.00423	0.00513	-0.00539	-0.00873	-0.00444	-0.02176	-0.01004	-0.02072	-0.00442	1	
Exam_Score_Categorized	0.37169542	0.4948516	-0.00029	0.14688	0.12302	0.0251	0.12106	0.0687	0.06699	0.08592	-0.00679	0.07692	0.04619	0.03756	0.05835	0.07544	1

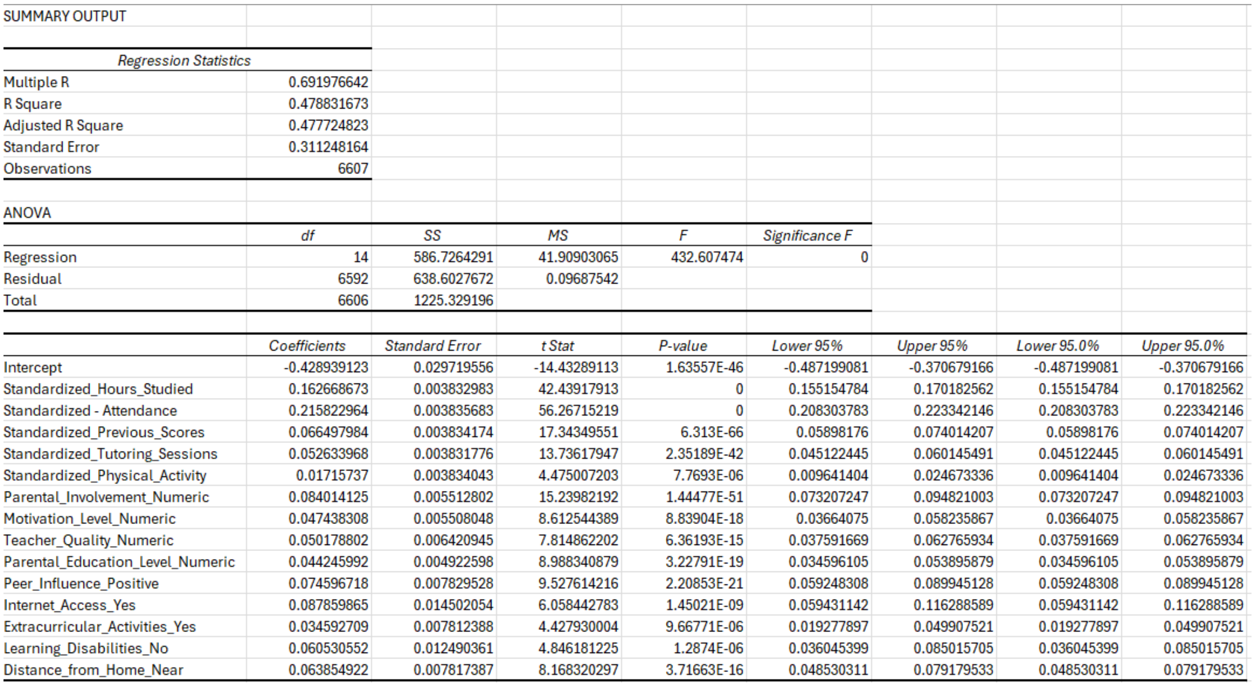
Step 4: Building the Regression Model

- Used the **Excel Data Analysis ToolPak** to perform regression analysis.
- Selected statistically significant variables by excluding those with p-values > 0.05. *Excluded variables:* Sandardized_Sleep_hours, School_Type_Public.



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.42994962	0.030305154	-14.1873433	5.02428E-4	-0.48935754	-0.3705417	-0.48935754	-0.3705417
Sandardized_Sleep_hours	0.004376124	0.003833229	1.141628721	**0.253649861**	-0.00313825	0.011890494	-0.00313825	0.011890494
School_Type_Public	0.001415461	0.008332551	0.169871309	**0.86511656**	-0.01491904	0.017749962	-0.01491904	0.017749962

Final regression table:



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.691976642
R Square	0.478831673
Adjusted R Square	0.477724823
Standard Error	0.311248164
Observations	6607

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	586.7264291	41.90903065	432.607474	0
Residual	6592	638.6027672	0.09687542		
Total	6606	1225.329196			

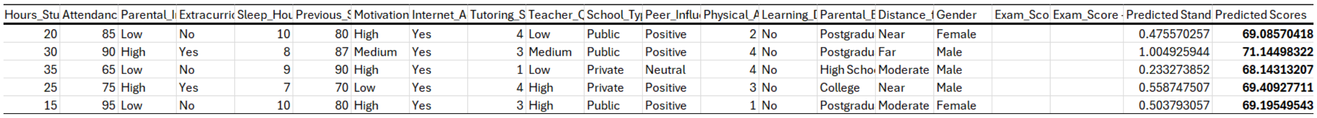
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.428939123	0.029719556	-14.43289113	1.63557E-46	-0.487199081	-0.370679166	-0.487199081	-0.370679166
Standardized_Hours_Studied	0.162668673	0.003832983	42.43917913	0	0.155154784	0.170182562	0.155154784	0.170182562
Standardized - Attendance	0.215822964	0.003835683	56.26715219	0	0.208303783	0.223342146	0.208303783	0.223342146
Standardized_Previous_Scores	0.066497984	0.003834174	17.34349551	6.313E-66	0.05898176	0.074014207	0.05898176	0.074014207
Standardized_Tutoring_Sessions	0.052633968	0.003831776	13.73617947	2.35189E-42	0.045122445	0.060145491	0.045122445	0.060145491
Standardized_Physical_Activity	0.01715737	0.003834043	4.475007203	7.7693E-06	0.009641404	0.024673336	0.009641404	0.024673336
Parental_Involvement_Numeric	0.084014125	0.005512802	15.23982192	1.44477E-51	0.073207247	0.094821003	0.073207247	0.094821003
Motivation_Level_Numeric	0.047438308	0.005508048	8.612544389	8.83904E-18	0.03664075	0.058235867	0.03664075	0.058235867
Teacher_Quality_Numeric	0.050178802	0.006420945	7.814862202	6.36193E-15	0.037591669	0.062765934	0.037591669	0.062765934
Parental_Education_Level_Numeric	0.044245992	0.004922598	8.988340879	3.22791E-19	0.034596105	0.053895879	0.034596105	0.053895879
Peer_Influence_Positive	0.074596718	0.007829528	9.527614216	2.20853E-21	0.059248308	0.089945128	0.059248308	0.089945128
Internet_Access_Yes	0.087859865	0.014502054	6.058442783	1.45021E-09	0.059431142	0.116288589	0.059431142	0.116288589
Extracurricular_Activities_Yes	0.034592709	0.007812388	4.427930004	9.66771E-06	0.019277897	0.049907521	0.019277897	0.049907521
Learning_Disabilities_No	0.060530552	0.012490361	4.846181225	1.2874E-06	0.036045399	0.085015705	0.036045399	0.085015705
Distance_from_Home_Near	0.063854922	0.007817387	8.168320297	3.71663E-16	0.048530311	0.079179533	0.048530311	0.079179533

- Constructed the regression equation:

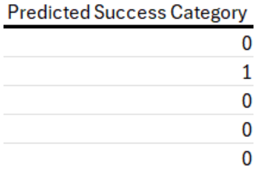
Predicted Exam Score (Standardized) = −0.429 + (0.163×Hours Studied) + (0.215×Attendance) + …

Step 5: Prediction

- **Raw Values Input**: Predicted scores for new students were calculated by first inputting their raw (actual) values (e.g., Hours Studied, Attendance) into the regression equation.
- **Standardization Post-Prediction**: A separate column was used to standardize these values based on the mean and standard deviation of the original dataset to ensure consistency with the model's scale.
- Classified students as **successful** or **unsuccessful** based on a threshold (≥ 70 corresponds to success).



Hours_Stu	Attendanc	Parental_I	Extracurric	Sleep_Hou	Previous_S	Motivation	Internet_A	Tutoring_S	Teacher_Q	School_Ty	Peer_Influ	Physical_A	Learning_D	Parental_E	Distance_f	Gender	Exam_Sco	Exam_Score -	Predicted Stand	Predicted Scores
20	85	Low	No	10	80	High	Yes	4	Low	Public	Positive	2	No	Postgradu	Near	Female			0.475570257	**69.08570418**
30	90	High	Yes	8	87	Medium	Yes	3	Medium	Public	Positive	4	No	Postgradu	Far	Male			1.004925944	**71.14498322**
35	65	Low	No	9	90	High	Yes	1	Low	Private	Neutral	4	No	High Scho	Moderate	Male			0.233273852	**68.14313207**
25	75	High	Yes	7	70	Low	Yes	4	High	Private	Positive	3	No	College	Near	Male			0.558747507	**69.40927711**
15	95	Low	No	10	80	High	Yes	3	High	Public	Positive	1	No	Postgradu	Moderate	Female			0.503793057	**69.19549543**



Predicted Success Category
0
1
0
0
0

Step 6: Model Validation

- **Method to Calculate Accuracy**:
  To validate the model, a new column, **"Actual = Predicted?"**, was created:
  - **1**: If the predicted outcome matches the actual outcome.
  - **0**: If the predicted outcome differs from the actual outcome.
- Accuracy was then calculated by taking the average of this column and multiplying it by 100:

  Accuracy = AVERAGE("Actual = Predicted?" Column) × 100
- **Result**:
  The model achieved an accuracy of **82%**, indicating strong predictive performance.

Key Results

- **Critical Factors**:
  - **Hours Studied** and **Attendance** had the strongest positive correlations with exam scores.
  - Categorical factors like **Parental Involvement** and **Motivation Level** also contributed significantly.
- **Model Performance**:
  - R-squared: **0.4788**
  - Adjusted R-squared: **0.4777**
  - Accuracy: **82%**

Usage Instructions

Step 1: Running the Model

- **Open the Excel File**:
  - Locate and open the project Excel file “StudentPerformanceFactors”.
- **Add a New Student Record**:
  - Add a new row to the sheet and input the **raw data** for the following variables:
    - `Hours Studied`
    - `Attendance (%)`
    - `Sleep Hours`
    - `Tutoring Sessions`

        - Other relevant features (Parental Involvement, Motivation Level, etc.).
- **Standardize the Raw Data**:
    - The standardized columns will automatically calculate values based on the **mean** and **standard deviation** of the original dataset.
    - **Important**:
        - Ensure the formulas are copied into the new row by dragging the previous cell's edge downward / upward (fill handle).
        - This will apply the standardization formula to the new input values.

Step 2: Interpreting Results

- The regression model will predict the **exam score** for the new student in a separate column.
- The following threshold is used to classify the results:
    - **Scores ≥ 70** → Successful (1)
    - **Scores < 70** → Unsuccessful (0)

Conclusion

This project demonstrates how data preprocessing, correlation analysis, and regression modeling can provide actionable insights into student success. With an **82% accuracy**, the model effectively identifies students who are likely to succeed, showcasing skills in data analysis, feature engineering, and predictive modeling.

Additionally, the model was tested on data collected from **five real students** based on survey responses. Their performance was predicted using the developed regression model, providing practical validation of the model's applicability to real-world scenarios. The predictions align well with observed patterns, reinforcing the model's reliability.

Contact

Feel free to reach out with questions or suggestions:

- **Name**: Gaurav Raghunand
- **Email**: graghun2@asu.edu | gauravraghunand@gmail.com
- **LinkedIn**: https://www.linkedin.com/in/gauravraghunand/
- **GitHub**: https://github.com/gauravraghunand