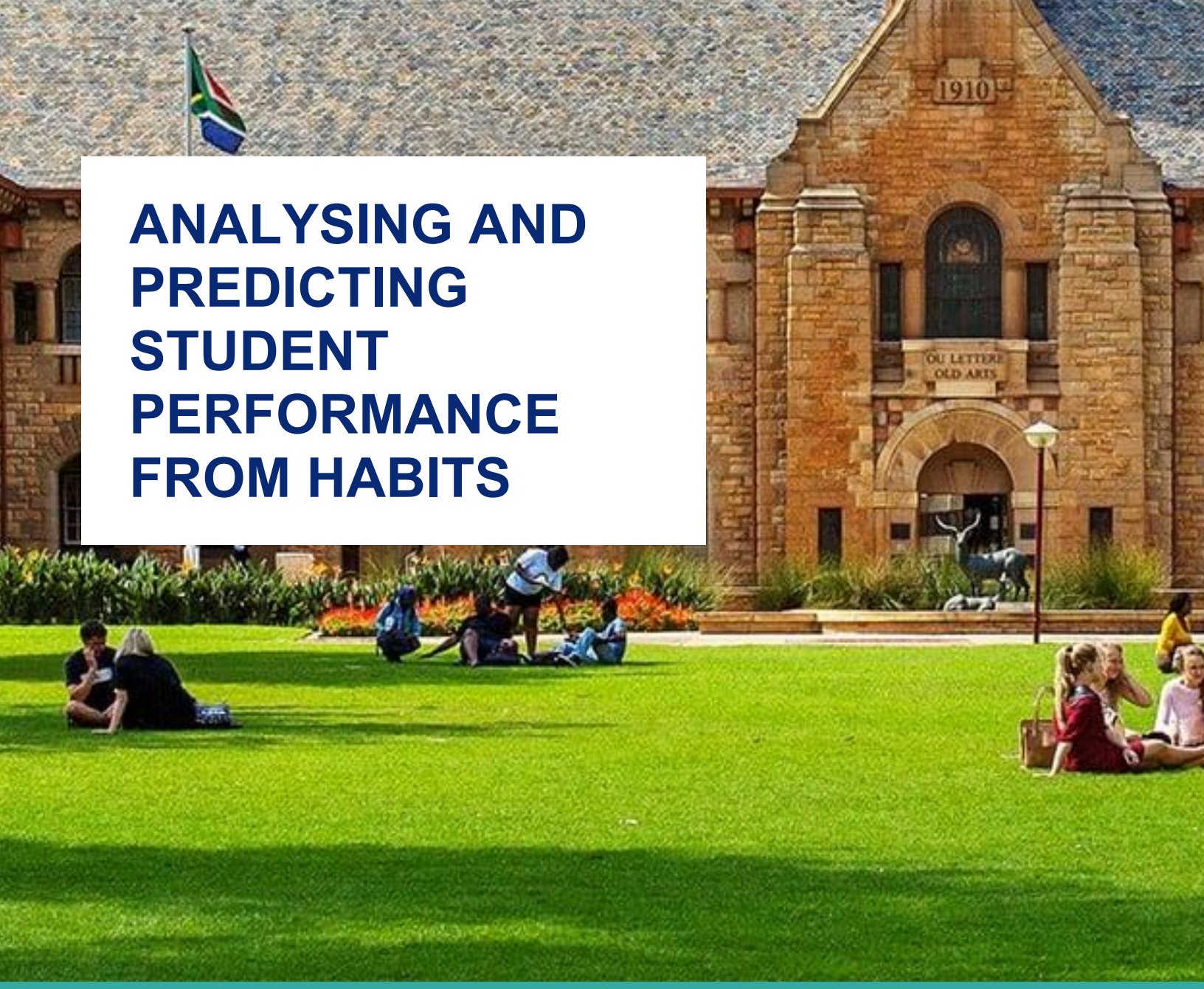


ANALYSING AND PREDICTING STUDENT PERFORMANCE FROM HABITS

2 NOVEMBER 2025

Zaio Institute of Technology
Authored by: Sesethu M. Bango

Logo Name

Student Habits Analysis and Performance Predictions

Analysing Student Habits and their effect on Academic Performance using Machine Learning Techniques

1. Background

The analysis investigates how students' lifestyle habits — including study time, sleep patterns, social media usage, diet, and mental well-being — relate to their academic performance. The study forms part of a broader initiative to understand the behavioural and situational factors that influence students' outcomes.
Through data analysis and predictive modelling, the goal is to identify actionable insights that can guide in the understanding of what drives academic success, and inversely the lack thereof.

2. Summary of Findings

At the end of the preparation stage:

- All 16 columns were validated, cleaned, and encoded.
- The dataset was free from missing values and duplicates.
- Categorical and numerical variables were properly prepared for model input.
- A standardized training and testing framework (80/20 split) was established.

Exploratory Data Analysis revealed that:

- **Study time is the single most influential factor** of student performance.
- **Lifestyle factors** (mental health, diet, exercise, media use) provide secondary yet significant insights.
- The dataset is **clean, balanced, and reliable**, with only minor outliers.
- Visual analysis supports and strengthens the statistical findings.

From model building and evaluation:

- Logistic Regression and Decision Tree models were trained.
- Both models performed **almost identically.**
- The **Logistic Regression** model was selected as the **preferred option** due to its simplicity, interpretability, and slightly stronger recall–precision balance.

The explainability section revealed:

- SHAP and LIME provide global and local explainability of feature influence on the model's decision-making process
- Based on SHAP — **higher study hours** strongly i**ncrease** the **likelihood of passing**, followed by a **strong mental health rating**.
- **Low mental_health_rating** and low **study_hours_per_day** had the **largest *negative* impacts**, making the student less likely to pass.

3. Reflection

A few key learnings and reflections emerged during development:

- **Correlation-driven feature selection** was an effective refinement step, improving model performance slightly while reducing noise.

- It would be beneficial to have an interactive dashboard, enabling stakeholders the ability to interact with the findings of this analysis.

4. Data Collection and Preparation

4.1 Data Source and Structure

The dataset, *student_habits_performance.csv*, was validated and loaded using the custom **DataLoaderIO_nException** class, which ensures proper CSV formatting and read permissions before proceeding.

Once imported, the dataset contained **1,000 entries** and **16 columns**, each column representing a key attribute of student behaviour or background.

The **df.info()** summary revealed **no null or missing values** post-cleaning, ensuring a fully complete dataset for analysis.

4.2 Data Cleaning

Data cleaning was handled through the **DataCleaner** class, which implemented several integrity checks and corrections:

- **Empty and duplicate checks** confirmed that the dataset was valid and free of redundant rows.
- **Missing values** were identified in the *parental_education_level* column and replaced with the value "No Education" via the **parentalNulls_fix()** method.
- **Negative values** were checked across all numeric fields to validate data ranges, ensuring no invalid entries.
- **Data types** were confirmed to align correctly (float64 for continuous variables, object for categorical ones).

This process produced a **clean, reliable base dataset,** ready for encoding and modelling.

4.3 Encoding of Categorical Variables

Categorical variables were transformed into numerical representations using the custom **ScorePredictor.string_to_Int()** function, which applied controlled mappings for interpretability and consistency.

The following columns were encoded:

Column	Encoding Mapping
gender	Female → 1, Male → 2, Other → 3
part_time_job	No → 1, Yes → 2, Other → 3
diet_quality	Good → 1, Fair → 2, Poor → 3, Other → 4
parental_education_level	Master → 1, High School → 2, Bachelor → 3, No Education → 4
internet_quality	Average → 1, Poor → 2, Good → 3, Other → 4
extracurricular_participation	No → 1, Yes → 2, Other → 3

This systematic encoding enabled the inclusion of all features in downstream machine learning models.

4.4 Data Splitting and Scaling

Before modelling, the *exam_score* variable was separated as the target (y), while the remaining variables formed the features (X).

The dataset was split into training and testing sets using **train_test_split**, with an **80/20 ratio** (*test_size* = 0.2), ensuring robust model evaluation.

Numerical features were then standardized using StandardScaler, applied to both the training and test sets. This normalization ensures that variables whose scale is an order/s of magnitude larger than the rest, contributed equally to model training. Which prevents scale-driven bias.

5. Data Exploration

5.1 Correlation Analysis

A correlation matrix was computed to identify the most influential behavioural and lifestyle factors affecting academic performance (*exam_score*).

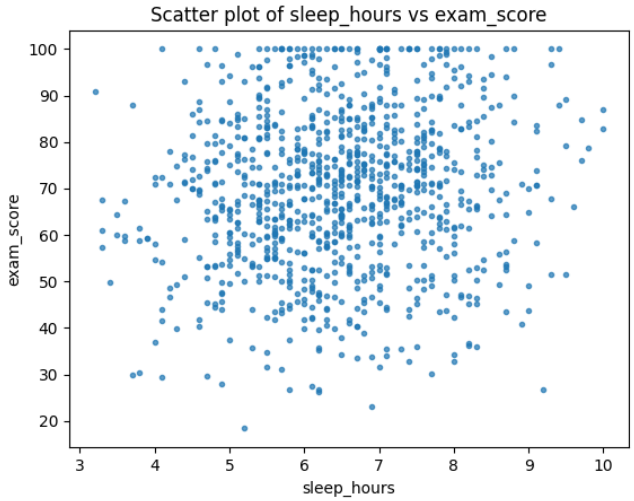
The top six relationships observed were:

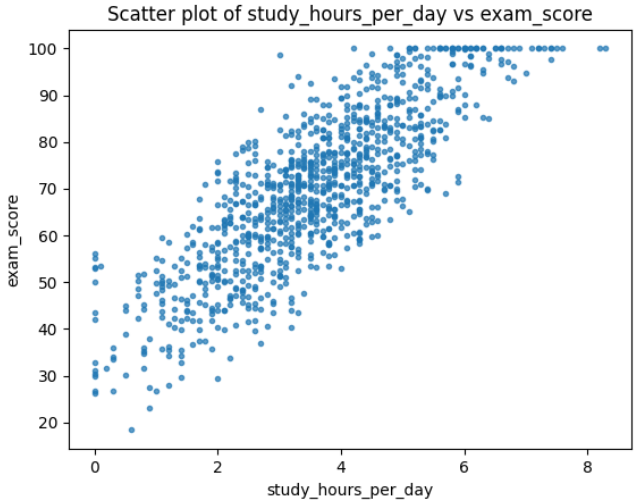
Feature Pair	Correlation Coefficient	Interpretation
study_hours_per_day ↔ *exam_score*	**0.83**	Strong positive correlation. Students who study longer tend to achieve significantly higher exam scores.
mental_health_rating ↔ *exam_score*	**0.32**	Moderate positive relationship. Better mental health is generally associated with improved performance.
social_media_hours ↔ *exam_score*	**-0.17**	Weak negative relationship. Increased time on social media may slightly hinder performance.
netflix_hours ↔ *exam_score*	**-0.17**	Weak negative correlation, indicating that more streaming time corresponds to lower exam results.
exercise_frequency ↔ *exam_score*	**0.16**	Slight positive trend, suggesting physical activity may play a small supportive role in academic achievement.
sleep_hours ↔ *exam_score*	**0.12**	Weak positive association, implying marginal benefit of adequate sleep on scores.

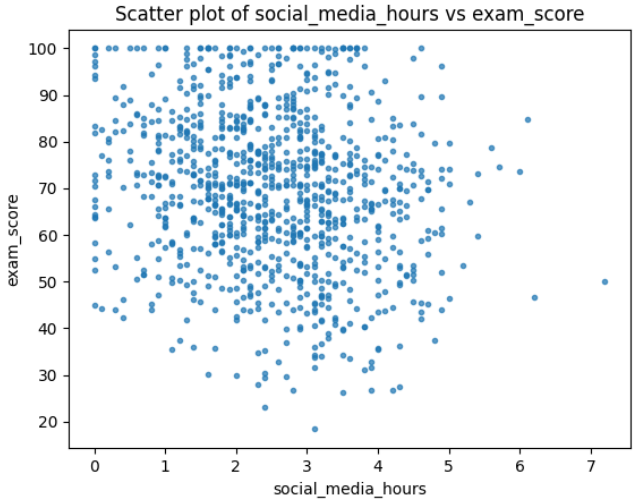
The correlation matrix confirms that **study time** is the most dominant predictor of exam performance, while other lifestyle variables show subtler but consistent effects in expected directions.

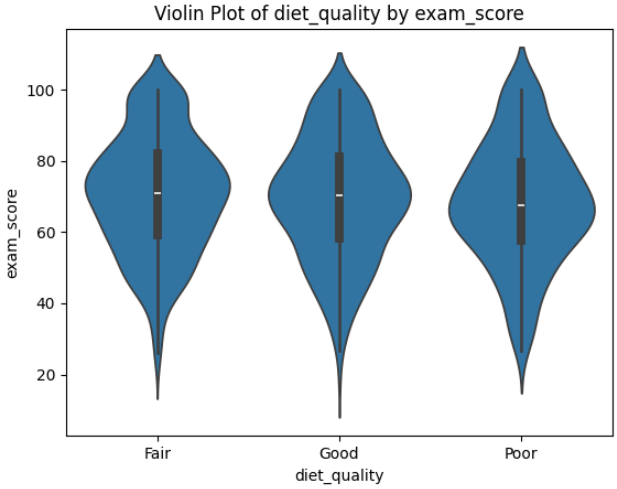
5.2 Visual Analysis

Plots (a) to (d) comparing a number of variables to *exam_score*









(a) Sleep Hours vs Exam Score

A scatter plot of *sleep_hours* versus *exam_score* revealed a widely dispersed pattern, showing only a **marginal indication** (barely perceptible) that longer sleep duration leads to higher exam scores.

(b) Study Hours vs Exam Score

This plot demonstrated a **clear, strong positive relationship**, reinforcing the quantitative finding (r = 0.83). Students who dedicate more time to studying, consistently outperform those studying less.

(c) Social Media Hours vs Exam Score

The relationship appears to be **negatively sloped but weak**, with a wide dispersion. This indicates that excessive social media usage might slightly detract from focus and study efficiency, the effect is not dominant though.

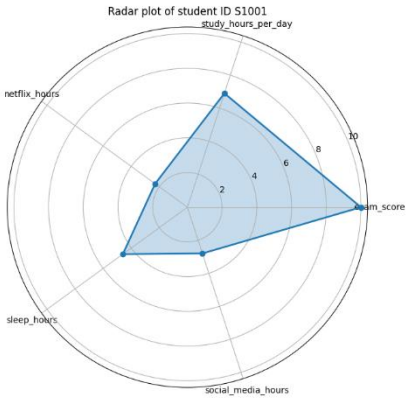
(d) Violin Plot – Diet Quality vs Exam Score

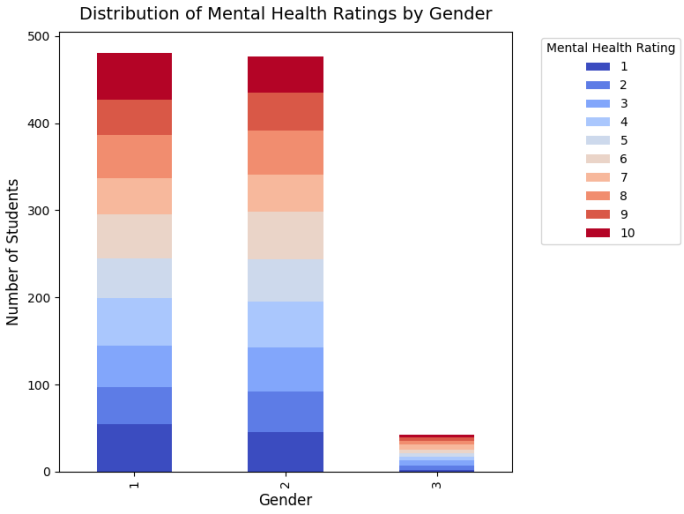
The violin plot compared score distributions across diet categories (Good, Fair, Poor). **Good** and **Fair** diets displayed nearly identical performance distributions. **Poor** diet quality showed a **slightly lower median** and a marginally more concentrated distribution around the median, hinting that nutrition may play a minor but consistent role in performance.

(e) and (f) Other Visual Tools

- The **radar plot** provides individualized student insight and comparison; while useful for personalized profiling, it does not contribute to the general population-level trends.
- The **stacked bar plot**, comparing gender with mental health ratings, revealed no gender imbalances.

Plots (e) and (f) Radar plot and Stacked Bar plot





5.3 Outlier Detection

It was found that there are very few outliers in the overall dataset. This low number of outliers indicates that there is no need for aggressive trimming or transformation. The outliers that do exist likely represent meaningful behavioural extremes rather than data errors.

Results from data exploration provide a **solid foundation for model development**, where predictions can then be made, and later their accuracy evaluated.

6. Model Building and Evaluation

6.1 Model Overview

Two predictive models were implemented to classify students' exam performance (Pass or Fail) based on their behavioural and demographic attributes. The **Pass threshold** was set to a score of **65+**.

These models are:

- **Logistic Regression** — A linear classification model that estimates the probability of passing or failing based on weighted combinations of input features.

- **Decision Tree Classifier** — A non-linear, tree-based model capable of capturing complex feature interactions and decision thresholds.

Hyperparameter tuning was performed for the decision tree classifier, as well as a cross-validation process. This ensures the chosen decision tree model is generalizable, not just optimized for one random train/test split.

Feature selection was refined to include all variables with a **correlation coefficient ≥ 0.05** with respect to *exam_score*. This threshold optimized model performance slightly compared to stricter or looser thresholds, Allowing for the best possible predictive capabilities for the trained model.

6.2 Model Training and Validation

Both models were trained using the same standardized datasets:

- **Training set:** 80% of data
- **Testing set:** 20% of data
- **Scaling:** Standardized using StandardScaler
- **Target Variable:** *exam_score* converted to categorical (0 = Fail, 1 = Pass) with Pass threshold 65+

6.3 Model Performance Comparison

Metric	Logistic Regression	Decision Tree
Accuracy	0.89	0.885
Precision	0.913	0.919
Recall	0.913	0.897
F1-Score	0.91	0.91

Both models performed **almost identically**, each achieving an F1-score of 0.91, signifying balanced precision and recall.

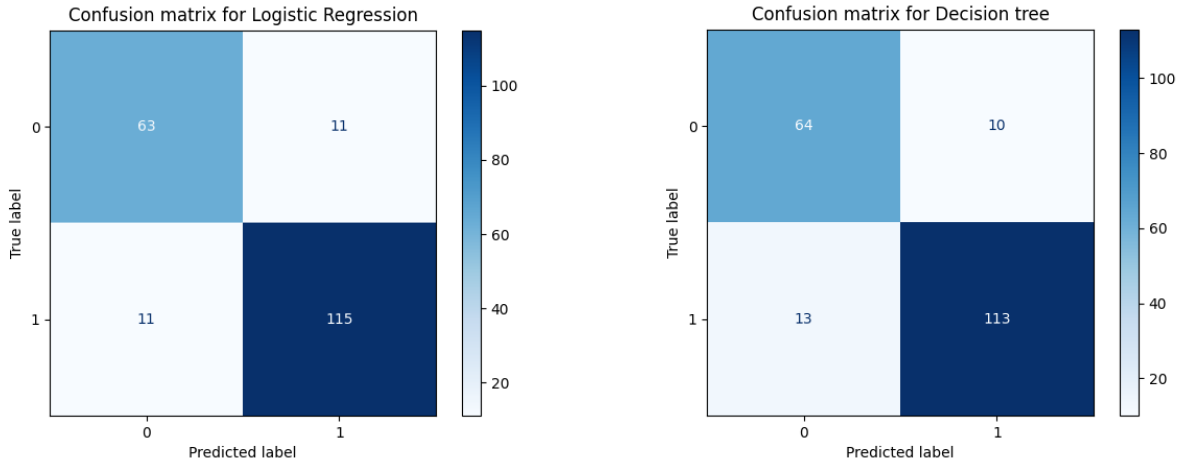
- Logistic Regression slightly outperformed on **recall**, indicating stronger ability to correctly identify passing students.
- Decision Tree had a marginally higher **precision**, meaning fewer false positives (students predicted to pass who actually failed).

Overall, the **Logistic Regression** model was selected as the **preferred option** due to its simplicity, interpretability, and slightly stronger recall–precision balance.

6.4 Evaluation and Visualisation

Evaluation was done using the following tools. They help to affirm the model's reliability and discriminative ability:

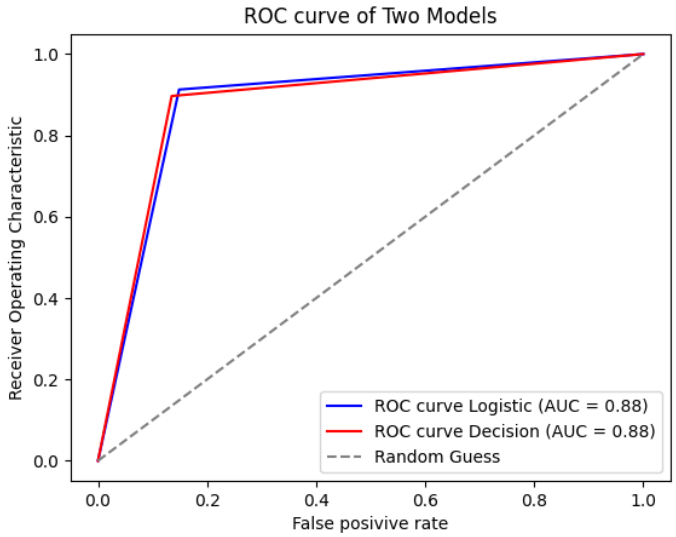
Plots (h) and (i) Confusion matrices



- **Confusion Matrices:**
  Provided a clear visualisation of correct (Dark blues) vs. incorrect (Bluey white) classifications, confirming balanced precision and recall across both models.
- **ROC Curves & AUC Scores**:
  The ROC (Receiver Operating Characteristic) curves for both models, shown below, overlap almost perfectly, each achieving an **AUC of 0.88**.

This means both models perform significantly better than random guessing (represented by the diagonal grey line). The near-identical shape of the red (Decision Tree) and blue (Logistic Regression) curves supports earlier findings where both models achieve high true positive rates with minimal false positives.

Plot (j) ROC curves



Overall, the ROC–AUC analysis **confirms the consistency and reliability** of both models and validates the selection of Logistic Regression as the preferred option.

7. Explainability

7.1 Purpose of Explainability

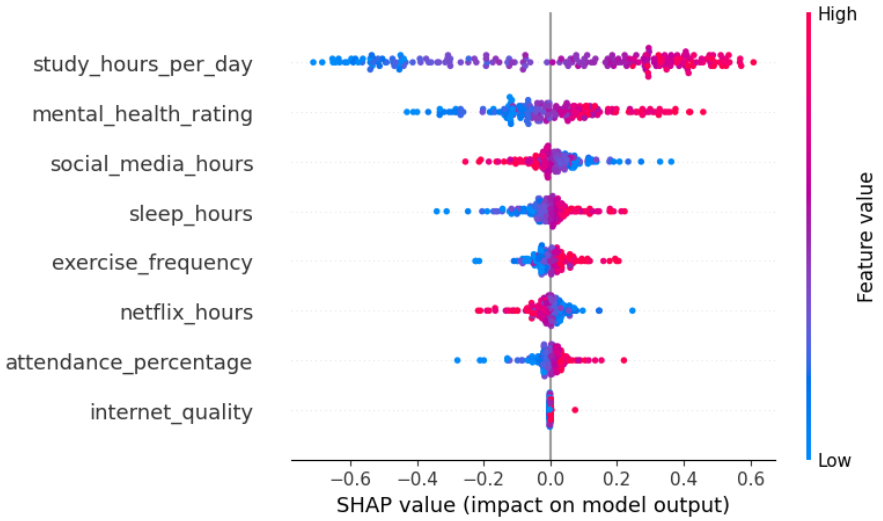
While model accuracy provides a measure for predictive performance, it does not reveal *why* certain predictions are made.

Explainability methods are therefore applied to uncover how each input feature influences the model's decision-making process, providing transparency and trustworthiness to the analysis.

Two complementary frameworks were used:

- **SHAP (SHapley Additive exPlanations):** to assess *global* feature importance across all predictions.
- **LIME (Local Interpretable Model-agnostic Explanations):** to explain *individual* predictions.

Plot (k) SHAP Summary Plot



7.2 SHAP Summary Plot

The SHAP summary plot above visualizes how each feature contributes to the predicted probability of a student passing or failing.

The colour gradient represents the feature's magnitude (red = high value, blue = low value), while horizontal dispersion shows its impact on the model's output.

Key Insights from SHAP:

1. **Study Hours per Day**
    - The most influential feature overall.
    - Higher study hours (red points on the right) strongly increase the likelihood of passing.
2. **Mental Health Rating**
    - Consistently positive influence; higher ratings (better mental health) push predictions toward "Pass."
3. **Social Media Hours** and **Netflix Hours**
    - Both show negative SHAP values for higher usage (red points on the left), confirming that increased screen time slightly lowers predicted performance.
4. **Sleep Hours** and **Exercise Frequency**
    - Moderate positive effects — sufficient rest and physical activity correlate with higher exam scores.

5. **Attendance Percentage** and **Internet Quality**
    - Low to moderate influence, indicating that while important, they are less decisive than study and mental health habits.

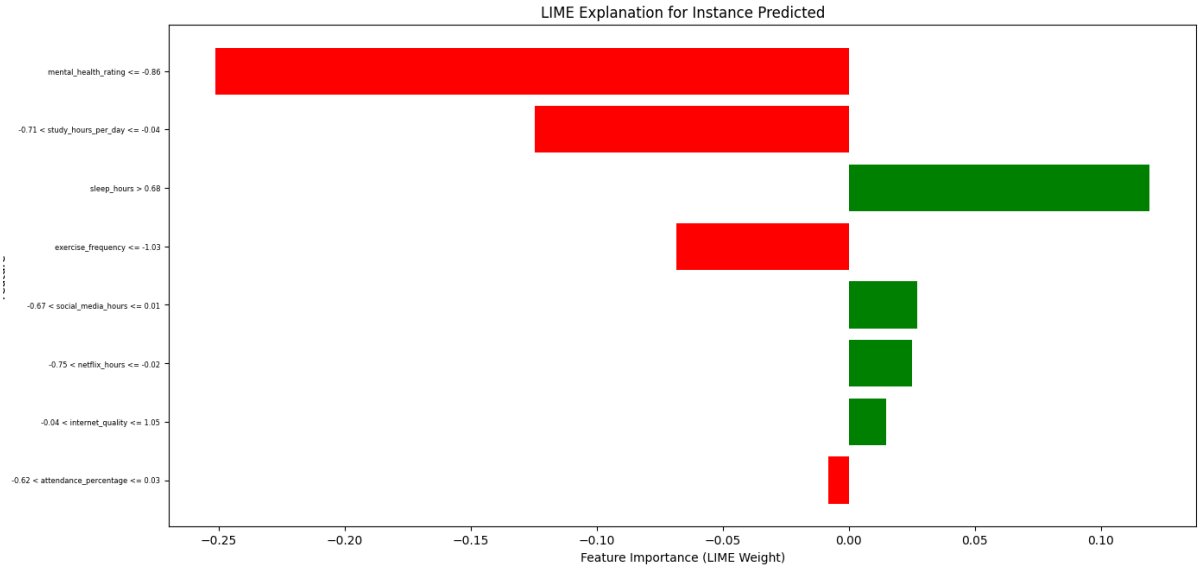
Overall, SHAP results reinforce the statistical correlations identified earlier: consistent study time and good mental health are the strongest drivers of academic success.

7.3 LIME Explanation for an Individual Prediction

The LIME explanation below visualizes which specific factors most affected a single student's predicted outcome. Each bar represents the *weight* (positive or negative) that a feature contributed to the prediction:

- **Red bars** = negative influence (lower chance of passing)
- **Green bars** = positive influence (higher chance of passing)

Plot (l) LIME explanation



Interpretation of this plot:

- Low ***mental_health_rating*** and ***study_hours_per_day*** had the largest *negative* impacts, making the student less likely to pass.
- High ***sleep_hours*** exerted a strong *positive* influence, slightly offsetting the negatives.
- Other features (e.g., ***exercise_frequency, diet_quality***) had smaller effects but followed the expected trends, healthier habits generally improve predicted performance.

LIME thus provides an explanation of how a single prediction was formed, complementing SHAP's broader, dataset-level view.

It's clear that consistent study, a balanced lifestyle, and strong mental health are the key predictors of academic success.

8. Conclusion

This analysis set out to understand how various lifestyle and behavioural habits influence student academic performance, and whether these insights could be modelled predictively.

Through a rigorous, structured process — encompassing data validation, cleaning, exploratory analysis, modelling, and interpretability, the project achieved its goal of producing both **accurate and explainable** predictions.

Key takeaways include:

1. **Study hours per day** emerged as the most powerful predictor of exam success, demonstrating a strong positive correlation (r = 0.83) and dominant SHAP contribution.
2. **Mental health**, **sleep quality**, and **exercise frequency** showed consistent but smaller positive effects, affirming the multifaceted nature of contributions to student performance.
3. **Screen time** represented by social media and streaming activity, exhibited weak negative correlations, suggesting a minor but discernible influence on academic outcomes.
4. The **Logistic Regression model**, while statistically simpler, delivered slightly stronger precision and accuracy than the tuned Decision Tree, making it the preferred choice for deployment.
5. Explainability through **SHAP** and **LIME** validated the model's internal logic, bridging technical accuracy with human interpretability.

Ultimately, this study demonstrates that **behavioural analytics can meaningfully forecast academic performance**.

References

1. **Dataset:** *Student Habits and Academic Performance* — dataset provided by Zaio Intitute of Technology (student_habits_performance.csv).
2. **Libraries:**
    - pandas, numpy, matplotlib, seaborn — Data manipulation and visualization
    - scikit-learn — Model building, training, and evaluation
    - shap and lime — Model interpretability
3. **Custom Modules:**
    - student_analysis.py (classes: *DataCleaner*, *VisualisationEngine*, *MakeModel*, etc.)
    - main_script.py (data pipeline and orchestrator)
4. **Stock Images:**
    - Cover image
      ishools (2025). "University of Pretoria". Ishools. Available at: https://www.ischools.org/ischools-members/university-of-pretoria. Accessed: 31 October 2025.