Student Performance Prediction: A Linear Regression Approach

This presentation explores the application of supervised learning techniques, specifically linear regression, to predict student performance. The study focuses on the impact of study hours, attendance, and past scores on exam scores. This research was conducted for the CSET211 course, Statistical Machine Learning, at the School of Computer Science and Engineering.

by Harsh (E23CSEU0517)

Abstract

This project investigates the use of machine learning for predicting student performance in an academic setting. The study utilizes a dataset containing features such as study hours, attendance, and past scores to predict exam scores.

A linear regression model was implemented and evaluated for its accuracy. The results indicate the feasibility of using machine learning techniques to estimate student performance effectively.

Introduction

Predicting Student Performance

Predicting student performance is crucial for educators and institutions to understand student needs and implement interventions effectively. This project investigates the ability of machine learning models to anticipate student performance.

Data-Driven Insights

By leveraging data analysis, institutions can identify at-risk students early on and provide them with appropriate support. This report details the process of building a linear regression model to predict exam scores based on various factors.

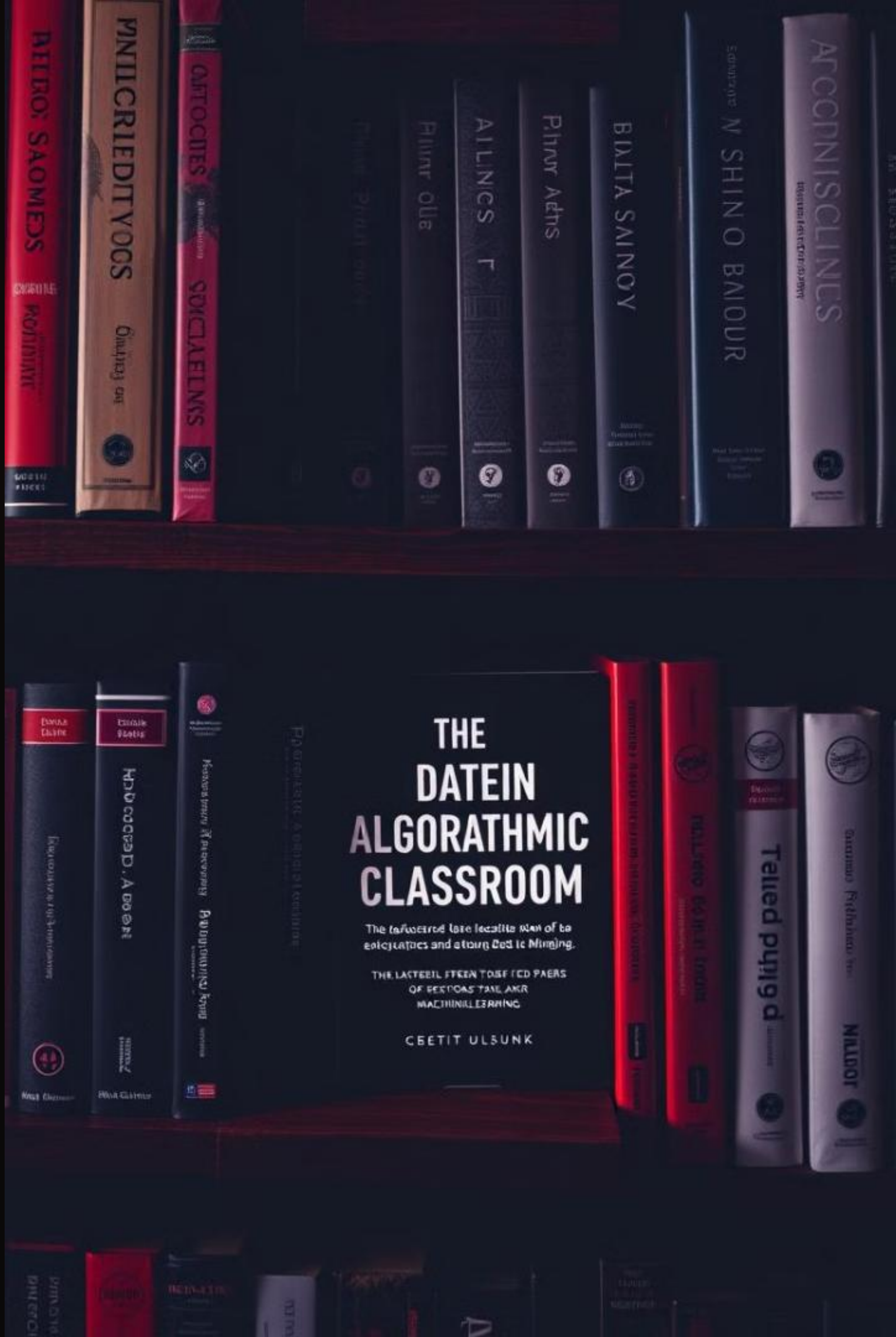
Related Work

Educational Data Mining

Research in educational data mining emphasizes the utilization of machine learning algorithms like regression, decision trees, and neural networks for analyzing student data and understanding patterns.

Linear Regression Focus

This project builds upon previous research by focusing on linear regression due to its interpretability and ability to provide clear insights into the relationship between variables.



Datasets and Preprocessing

Dataset Description

The dataset comprises 1000 observations of student performance, with features such as study hours, attendance percentage, past exam scores, and actual exam scores. The exam scores serve as the target variable for prediction.

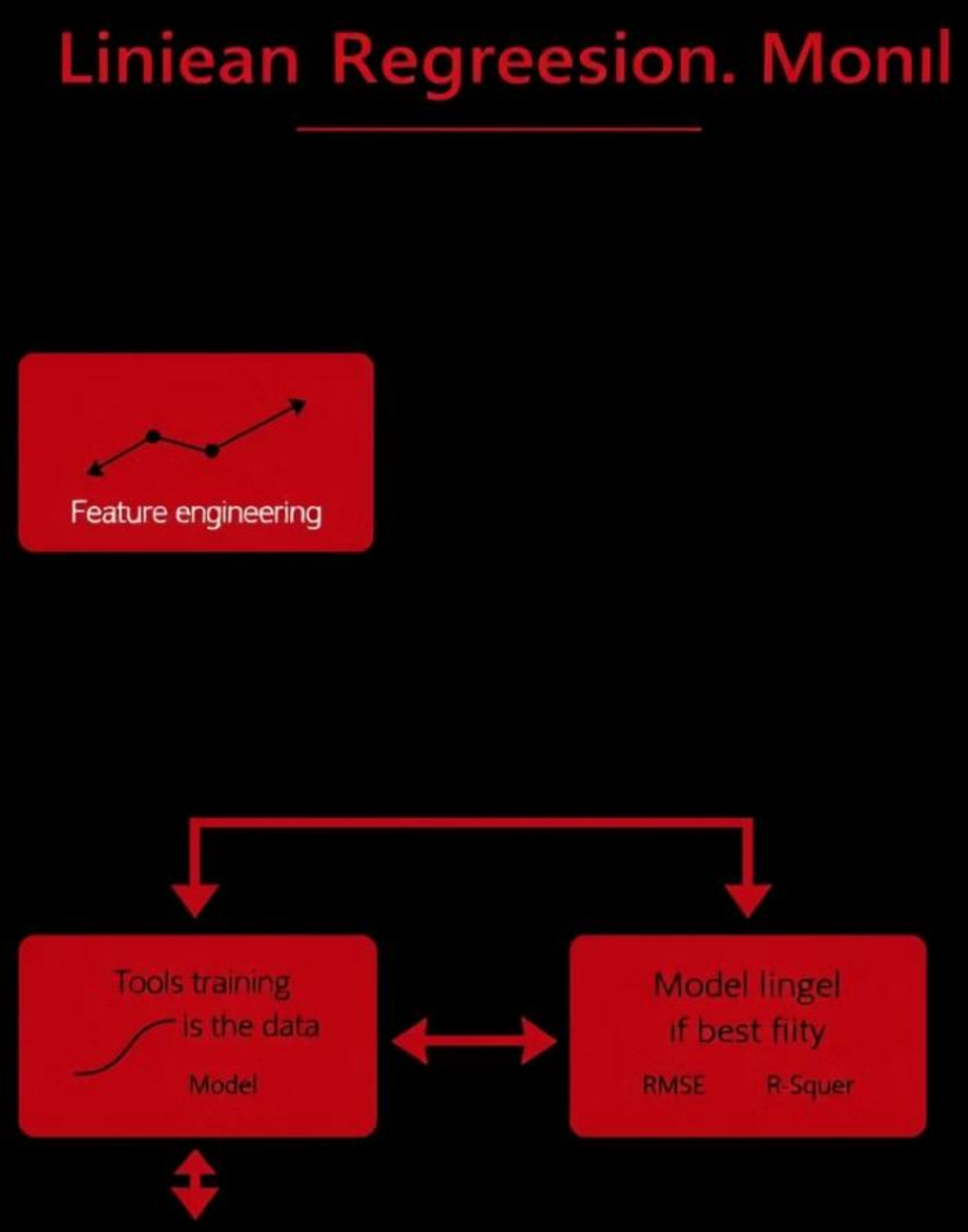
Data Preprocessing

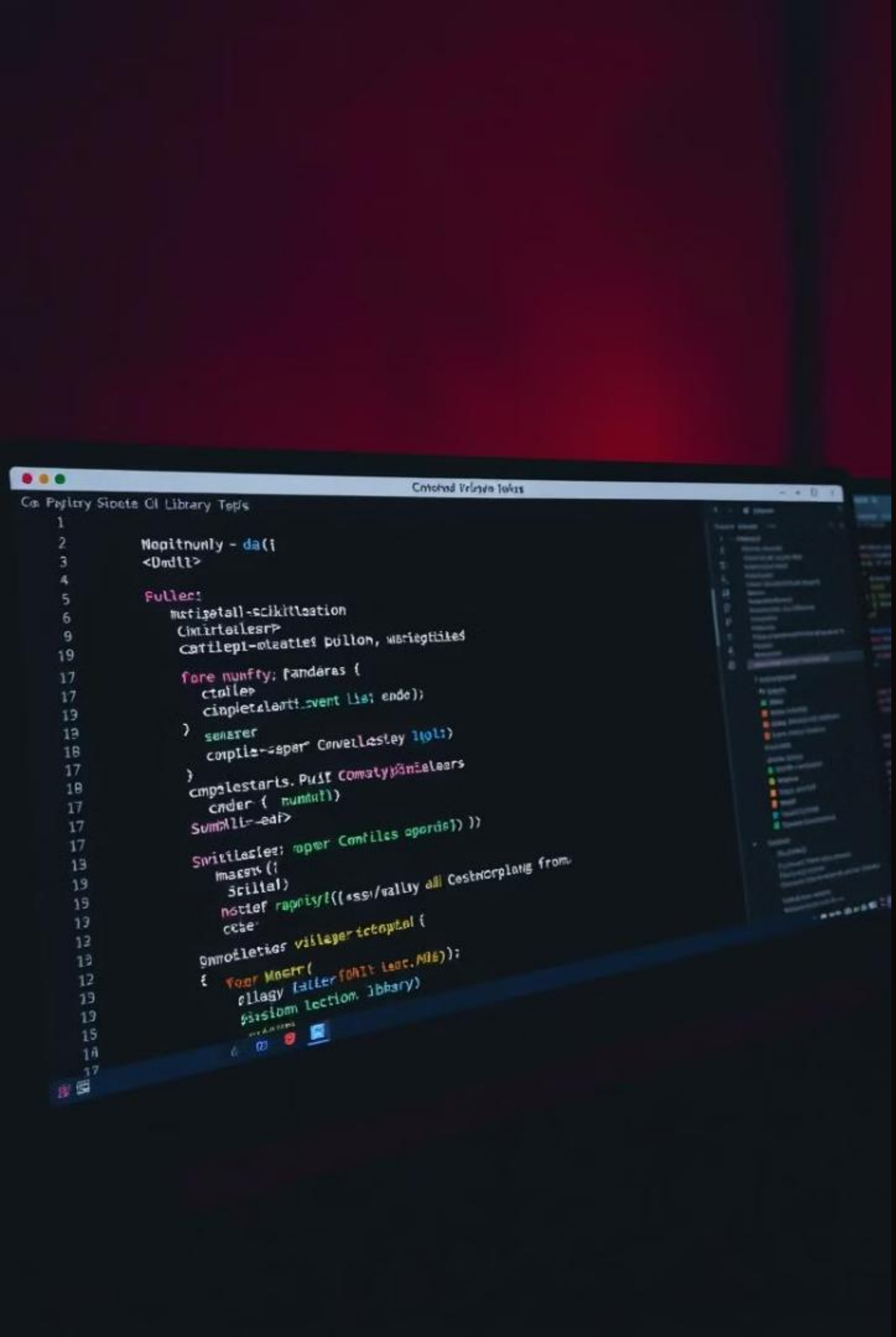
Data preprocessing is essential for ensuring data quality and consistency. This involved checking for missing values and conducting exploratory data analysis through pair plots and correlation heatmaps to visualize relationships between features.

Methodology

1 The project followed a structured approach to build and evaluate a linear regression model. This included data loading and inspection, visualizing relationships between features and the target variable, and splitting the data into training and testing sets.

2 A linear regression model was trained using the scikit-learn library in Python, followed by evaluation using metrics such as MAE, MSE, RMSE, and R² Score. The model's performance was further analyzed by visualizing actual versus predicted scores.





Hardware and Software Requirements

Hardware

The project was run on a computer with an Intel i7 processor and 16GB of RAM, sufficient for the analysis and model training.

Software

Python 3.8+ was used as the programming language, along with libraries like NumPy, pandas, Matplotlib, Seaborn, and scikit-learn for data manipulation, visualization, and machine learning.

[redacted]

kew performaanegregeeiandiatrs

	(MAE)	(NS%)	(RMSE)
	MSE	MSE	MSE
	216	1650	2326
	376	2531	3851
	138	1352	4374
	134	3405	3642
	168	1504	4342
	138	1360	6664
	158	1300	3733
	191	1200	3309
MSE	153	2371	1973
R1	165	2109	4346
MSE	136	1100	1927
R2	165	1700	2299
R2	434	1606	1344
R2	126	2600	1977
H3	338	3882	4847
R1	242	1501	1541
	308	2510	1942

Performance Metrics

2.62

MAE

Mean Absolute Error, measuring the average absolute difference between predicted and actual scores.

11.36

MSE

Mean Squared Error, measuring the average squared difference between predicted and actual scores.

3.37

RMSE

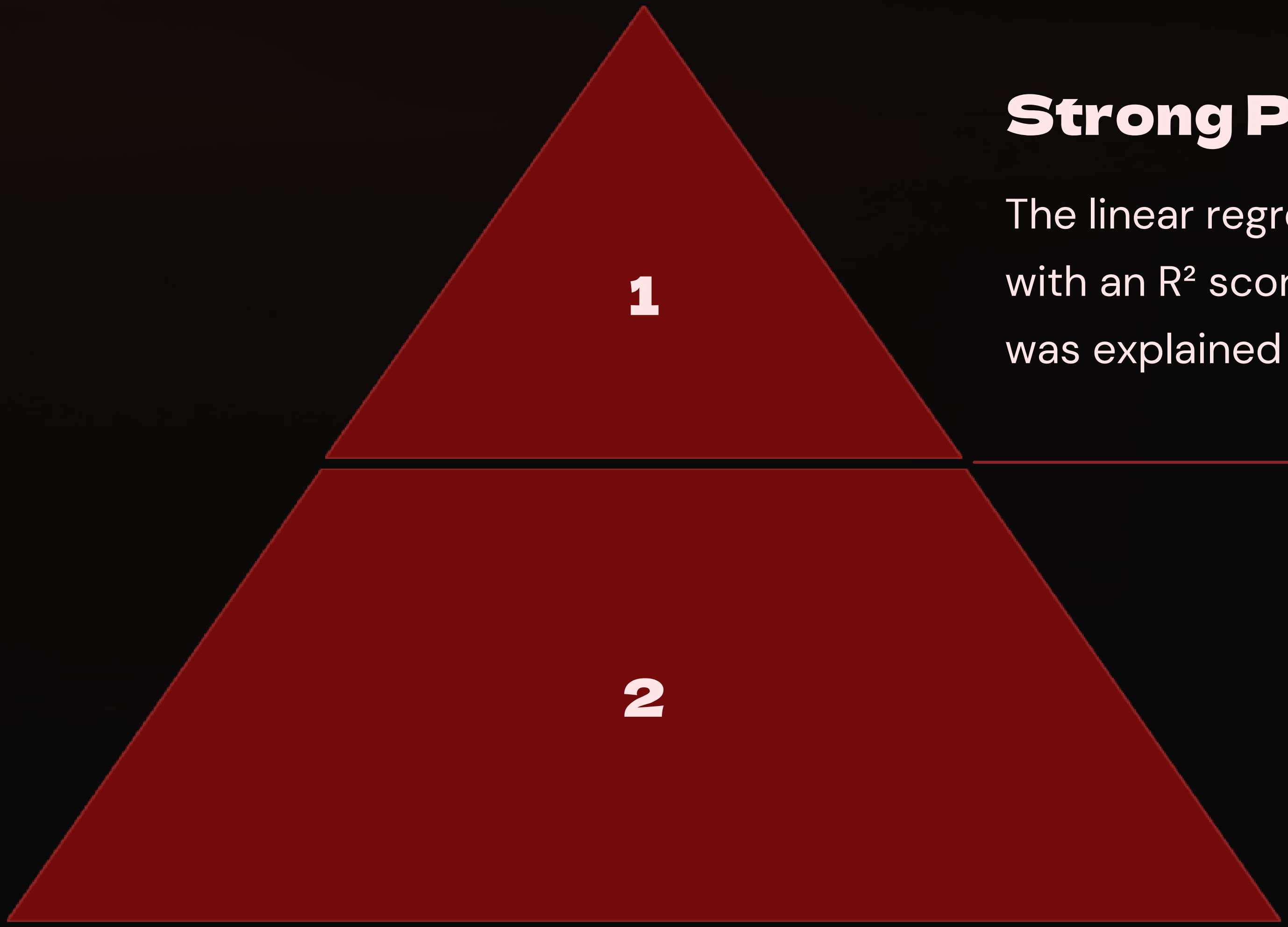
Root Mean Squared Error, measuring the square root of the MSE, providing a more interpretable measure of error.

0.92

R^2 Score

Coefficient of Determination, indicating the proportion of variance in exam scores explained by the model.

Results and Analysis



Strong Performance

The linear regression model demonstrated strong predictive performance with an R^2 score of 0.92, suggesting that 92% of the variance in exam scores was explained by the model's variables.

Visual Analysis

The scatter plot comparing actual and predicted values showed a strong linear alignment, indicating that the model captured the relationship between features and exam scores effectively.

[redacted]

Conclusions and Future Works

1

Applicability of Linear Regression

This project demonstrates the successful application of linear regression in predicting student performance, highlighting its potential for understanding factors influencing academic outcomes.

2

Future Directions

Future research could involve incorporating additional features like participation in extracurricular activities, exploring non-linear models for improved accuracy, and examining the impact of different learning styles on performance.