Predicting Scores Based on Study Hours Using Linear Regression

A PROJECT REPORT

Submitted by

SAI NITHICK ROSHAAN S (2303811724321094)

in partial fulfillment of requirements for the award of the course

AGI1252-FUNDAMENTALS OF DATA SCIENCE USING R

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY

(An Autonomous Institution, affiliated to Anna University Chennai and Approved by AICTE, New Delhi)

SAMAYAPURAM – 621 112

JUNE- 2025

K. RAMAKRISHNAN COLLEGE OF TECHNOLOGY (AUTONOMOUS)

SAMAYAPURAM – 621 112

BONAFIDE CERTIFICATE

Certified that this project report on **“Predicting Scores Based on Study Hours Using Linear Regression”** is the bonafide work of **SAI NITHICK ROSHAAN S (2303811724321094)** who carried out the project work during the academic year 2024 - 2025 under my supervision.

SIGNATURE	**SIGNATURE**
Dr.T. AVUDAIAPPAN, M.E.,Ph.D.,	Mrs.S. MURUGAVALLI, M.E.,(Ph.D).,
HEAD OF THE DEPARTMENT	**SUPERVISOR**
ASSOCIATE PROFESSOR	ASSISTANT PROFESSOR
Department of Artificial Intelligence	Department of Artificial Intelligence
K.Ramakrishnan College of Technology (Autonomous)	K.Ramakrishnan College of Technology (Autonomous)
Samayapuram–621112.	Samayapuram–621112.

Submitted for the viva-voce examination held on …03.06.2025……………….

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

I declare that the project report on **“Predicting Scores Based on Study Hours Using Linear Regression”** is the result of original work done by us and best of our knowledge, similar work has not been submitted to **“ANNA UNIVERSITY CHENNAI”** for the requirement of Degree of **BACHELOR OF TECHNOLOGY**. This project report is submitted on the partial fulfilment of the requirement of the completion of the course **AGI1252 – FUNDAMENTALS OF DATA SCIENCE USING R.**

.

Signature

SAI NITHICK ROSHAAN S

Place: Samayapuram

Date: 02.06.2025

ACKNOWLEDGEMENT

It is with great pride that I express our gratitude and in-debt to our institution "**K.Ramakrishnan College of Technology (Autonomous)**", for providing us with the opportunity to do this project.

I glad to credit honourable chairman **Dr. K. RAMAKRISHNAN**, **B.E.,** for having provided for the facilities during the course of our study in college.

I would like to express our sincere thanks to our beloved Executive Director **Dr. S. KUPPUSAMY, MBA, Ph.D.,** for forwarding to our project and offering adequate duration in completing our project.

I would like to thank **Dr. N. VASUDEVAN, M.Tech., Ph.D.,** Principal, who gave opportunity to frame the project the full satisfaction.

I whole heartily thanks to **Dr. T. AVUDAIAPPAN, M.E.,Ph.D.,** Head of the department, **ARTIFICIAL INTELLIGENCE** for providing his encourage pursuing this project.

I express our deep expression and sincere gratitude to our project supervisor **Ms.S.Murugavalli., M.E.,(Ph.D).,** Department of **ARTIFICIAL INTELLIGENCE,** for her incalculable suggestions, creativity, assistance and patience which motivated us to carry out this project.

I render our sincere thanks to Course Coordinator and other staff members for providing valuable information during the course.

I wish to express our special thanks to the officials and Lab Technicians of our departments who rendered their help during the period of the work progress.

INSTITUTE

Vision:

- To serve the society by offering top-notch technical education on par with global standards.

Mission:

- Be a center of excellence for technical education in emerging technologies by exceeding the needs of industry and society.
- Be an institute with world class research facilities.
- Be an institute nurturing talent and enhancing competency of students to transform them as all – round personalities respecting moral and ethical values.

DEPARTMENT

Vision:

- To excel in education, innovation, and research in Artificial Intelligence and Data Science to fulfil industrial demands and societal expectations.

Mission

- To educate future engineers with solid fundamentals, continually improving teaching methods using modern tools.
- To collaborate with industry and offer top-notch facilities in a conducive learning environment.
- To foster skilled engineers and ethical innovation in AI and Data Science for global recognition and impactful research.
- To tackle the societal challenge of producing capable professionals by instilling employability skills and human values.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

- **PEO1:** Compete on a global scale for a professional career in Artificial Intelligence and Data Science.
- **PEO2:** Provide industry-specific solutions for the society with effective communication and ethics.
- **PEO3** Enhance their professional skills through research and lifelong learning initiatives.

PROGRAM SPECIFIC OUTCOMES (PSOs)

- **PSO1:** Capable of finding the important factors in large datasets, simplify the data, and improve predictive model accuracy.
- **PSO2:** Capable of analyzing and providing a solution to a given real-world problem by designing an effective program.

PROGRAM OUTCOMES (POs)

Engineering students will be able to:

1. **Engineering knowledge:** Apply knowledge of mathematics, natural science, computing, engineering fundamentals, and an engineering specialization to develop solutions to complex engineering problems.

2. **Problem analysis:** Identify, formulate, review research literature and analyze complex engineering problems reaching substantiated conclusions with consideration for sustainable development.

3. **Design/development of solutions:** Design creative solutions for complex engineering problems and design/develop systems/components/processes to meet identified needs with consideration for the public health and safety, whole-life cost, net zero carbon, culture, society and environment as required.

4. **Conduct investigations of complex problems:** Conduct investigations of complex engineering problems using research-based knowledge including design of experiments, modelling, analysis & interpretation of data to provide valid conclusions.

5. **Engineering Tool Usage:** Create, select and apply appropriate techniques, resources and modern engineering & IT tools, including prediction and modelling recognizing their limitations to solve complex engineering problems.

6. **The Engineer and The World:** Analyze and evaluate societal and environmental aspects while solving complex engineering problems for its impact on sustainability with reference to economy, health, safety, legal framework, culture and environment.

7. **Ethics:** Apply ethical principles and commit to professional ethics, human values, diversity and inclusion; adhere to national & international laws.

8. **Individual and Collaborative Team work:** Function effectively as an individual, and as a member or leader in diverse/multi-disciplinary teams.

9. **Communication:** Communicate effectively and inclusively within the engineering community and society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations considering cultural, language, and learning differences.

10. **Project management and finance:** Apply knowledge and understanding of engineering management principles and economic decision-making and apply these to one's own work, as a member and leader in a team, and to manage projects and in multidisciplinary environments.

11. **Life-long learning:** Recognize the need for, and have the preparation and ability for i) independent and life-long learning ii) adaptability to new and emerging technologies and iii) critical thinking in the broadest context of technological change.

ABSTRACT

This project presents a data-driven approach to predicting students’ academic performance based on the number of study hours using linear regression in R programming. The objective is to explore the relationship between study time and scores, and to develop a predictive model capable of estimating academic outcomes. Using a dataset containing study hours and corresponding student scores, the project applies data preprocessing, exploratory data analysis with visualization tools like ggplot2, and model building using linear regression through base R and the caret package. Evaluation metrics such as R-squared, RMSE, and MAE were used to assess model performance. The results reveal a strong positive linear correlation, indicating that increased study time generally leads to better academic results. This project not only reinforces core concepts in machine learning and statistical analysis but also demonstrates the practical application of R in solving real-world educational problems. It highlights the importance of data-driven insights in academic planning and showcases how programming and analytics can be effectively utilized in the field of education.

ABSTRACT WITH POs AND PSOs MAPPING

CO 5 : BUILD R PROJECT FOR SOLVING REAL-TIME PROBLEMS.

ABSTRACT	POs MAPPED	PSOs MAPPED
This project uses linear regression in R to predict students’ academic performance based on study hours. By analyzing a dataset of study time and scores, it applies data preprocessing, exploratory visualization (using ggplot2), and model building with base R and the caret package. The model is evaluated using R-squared, RMSE, and MAE, revealing a strong positive correlation between study time and academic results. The project demonstrates the practical use of machine learning and statistical analysis in education, emphasizing the value of data-driven insights for academic planning.	**PO1 -3** **PO2 -3** **PO3 -3** **PO4 -2** **PO5 -3** **PO6 -1** **PO7 -1** **PO8 -1** **PO9 -2** **PO10 -2** **PO11-2**	**PSO1 -3** **PSO2 -3**

Note: 1- Low, 2-Medium, 3- High

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

In today's increasingly data-driven society, the integration of data analytics into education has opened new possibilities for enhancing academic outcomes. This project, titled **"Predicting Student Academic Performance Using Linear Regression in R,"** is a practical implementation of that concept, aimed at examining how the number of study hours affects a student's academic scores. The central goal of this project is to develop a simple yet effective predictive model using **linear regression**, one of the most fundamental and widely-used statistical techniques in the field of machine learning. By establishing a clear relationship between the amount of time a student dedicates to studying and their performance, the project offers meaningful insights that can aid in personalizing study plans and academic strategies. Various stages of the data science workflow are covered in this project, including data preprocessing, exploratory data analysis using the ggplot2 library for clear and intuitive visualizations, and model training and evaluation using both base R functions and the caret package. Model performance is assessed using evaluation metrics such as **R-squared (R^2)**, **Root Mean Square Error (RMSE)**, and **Mean Absolute Error (MAE)** to ensure the accuracy and reliability of predictions. The results of this analysis revealed a strong positive correlation between study hours and academic performance, confirming that increased study time generally leads to better results. This outcome supports the idea that data analytics can serve as a powerful tool in education, helping stakeholders understand patterns, make informed decisions, and ultimately improve learning experiences. Furthermore, this project serves as an excellent example of how statistical programming and predictive modeling can be applied to solve real-world problems, reinforcing the importance of **R programming skills** in modern data science and educational research.

1.1 OBJECTIVE

The primary objective of the Student Score Predictor using Study Hours using Linear Regression is to design and develop an interactive and intelligent data-driven application that accurately estimates a student's academic performance based on study time. The system aims to:

1. Automate academic score prediction: Apply linear regression techniques to forecast student scores from the number of study hours, minimizing the guesswork in academic planning.
2. Improve educational insight: Enable students and educators to explore the relationship between study time and performance through real-time, data-backed visualizations.
3. Showcase the power of R and Shiny: Utilize R programming and the Shiny framework to integrate statistical modeling, data visualization, and web-based interaction in a single cohesive platform.
4. Bridge academic concepts with practical implementation: Demonstrate how fundamental machine learning algorithms like linear regression can be used in real-world educational applications to support better decision-making and study planning.

1.2 OVERVIEW

The Student Score Predictor using Study Hours using Linear Regression comprises three main modules:

1. Data Upload Module:
    a. Enables users to upload a CSV file containing historical academic data with Study_Hours and Scores columns.

    b. Validates the uploaded file to ensure it follows the correct structure before processing.

c. Provides the foundational dataset required to train the prediction model.

2. Prediction Module:

   a. Accepts numeric input from the user for study hours, limited between 1 and 10 hours.

   b. Utilizes a linear regression model to predict the corresponding student score based on the input.

   c. Displays the predicted score dynamically, offering immediate feedback to the user.

3. Visualization Module:

   a. Acts Generates a scatter plot using the uploaded dataset to represent actual data points.

   b. Draws a regression line to visualize the model's interpretation of the relationship between study hours and scores.

   c. Highlights the user’s predicted point on the graph for easy comparison and clarity.

1.3 R PROGRAMMING CONCEPTS USED

The project leverages the following core R programming principles:

1. Reactive Programming (Shiny Framework):
    a. Uses reactive expressions and render functions to enable real-time data updates and dynamic UI behavior based on user input or uploaded data.
    b. Ensures required inputs like file uploads or numeric values are available before executing dependent computations.

2. Statistical Modeling:
    a. Applies linear regression using the lm() function to predict student scores from study hours.
    b. Uses the predict() function to generate score predictions based on new user input.
3. Data Handling:
    a. Allows CSV file input and reads data using read.csv().
    b. Manages tabular data using data frames for model training and prediction.
4. Control Flow:
    a. Uses conditional execution and validation to control how data and outputs are rendered based on user interaction.
    b. Validates required inputs using req() to prevent errors during model building or prediction.
    c. Handles user-defined numeric inputs and ensures they fall within predefined limits.
    d. Dynamically updates the plot and prediction output only when valid and complete data is available.

CHAPTER 2

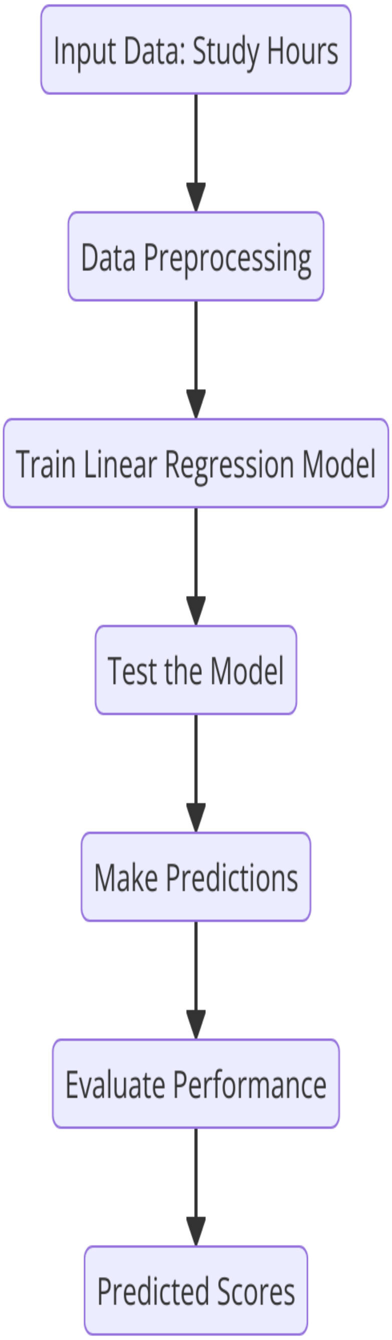
PROJECT METHEDOLOGY

2.1 PROPOSED WORK

The development methodology includes the following steps:

1. System Design:
   a. Design an intuitive, browser-based interface using R Shiny for interactive user engagement.
   b. Ensure modularity by organizing functionalities into separate components: data upload, prediction, and visualization.
2. Implementation:
   a. Use linear regression to model the relationship between study hours and student scores.
   b. Apply Shiny's reactive programming structure to dynamically handle user input and update outputs in real-time.
3. Testing:
   a. Test individual components such as file input validation, model accuracy, and numeric prediction functionality.
   b. Conduct end-to-end testing to ensure the system accurately predicts and visualizes scores based on user input.
4. Enhancement Planning:
   a. Outline Plan future improvements such as support for multiple regression models, downloadable reports, and database integration.
   b. Consider adding UI features like sliders, input validation alerts.

2.2 BLOCK DIAGRAM



CHAPTER 3

MODULE DESCRIPTION

3.1 Data Collection & Preprocessing Module

This module is designed to enable users to upload and prepare the dataset required for building the prediction model. It ensures the data is accurate and ready for analysis.

Functionalities:

1. Data Upload::
    a. Users upload a CSV file containing student data.
    b. The CSV must have specific columns: Study_Hours and Scores.
    c. The system verifies the file format and structure before processing.
2. Data Validation::
    a. Checks for missing or invalid values in the dataset.
    b. Ensures all study hours are numeric and within reasonable ranges.
    c. Removes or flags inconsistent records to maintain data quality.
3. Data Cleaning:
    a. Handles any preprocessing needed such as trimming spaces or correcting data types.
    b. Prepares the dataset in a format compatible with the linear regression model.
4. Data Storage:
    a. Stores the cleaned data in a reactive data frame for further use in modeling.
    b. Enables dynamic updates when new data is uploaded.

3.2 Model Building (Linear Regression) Module

This module focuses on creating the predictive model that estimates student scores based on study hours using linear regression.

Functionalities:

1. Model Training:
    a. Builds a linear regression model using the uploaded and preprocessed dataset.
    b. Uses the formula Scores ~ Study_Hours to establish a relationship between variables.
    c. The model is trained dynamically whenever new data is uploaded.
2. Parameter Estimation:
    a. Calculates coefficients (slope and intercept) that best fit the data.
    b. Uses least squares method to minimize prediction error.
3. Model Storage:
    a. Stores the trained model reactively for continuous prediction use.
    b. Ensures the model updates automatically if the dataset changes.
    c. Enables efficient retrieval of model parameters for real-time predictions.
    d. Supports easy integration with the prediction module for seamless user interaction.
4. Model Summary:
    a. Provides statistical details such as R-squared and residuals for evaluating model fit.
    b. Can be extended to display detailed regression diagnostics in future.

3.3 Model Evaluation & Prediction Module

This module integrates the prediction functionality with model assessment to provide accurate student score predictions based on study hours.

Functionalities:

1. Prediction Input Handling:
    a. Accepts user input for study hours within a predefined range(1to10hours).
    b. Validates input to ensure it meets the required criteria before making predictions.
    c. Restricts invalid or extreme inputs to maintain data integrity.
    d. Enhances usability through a numeric input control in the user interface.
2. Model Evaluation::
    a. Job Evaluates the linear regression model's performance using metrics like Mean Squared Error (MSE) and R-squared values.
    b. Continuously monitors model accuracy as new data is introduced.
    c. Helps identify underfitting or overfitting by comparing predicted and actual scores.
3. Score Prediction::
    a. Uses the trained linear regression model to predict student scores based on the input study hours.
    b. Provides real-time predictions that update dynamically as users change input values.
    c. Displays predicted points clearly on the regression plot for visual feedback.

3.4 VISUALIZATION MODULE

This module focuses on presenting the data and prediction results through graphical representations, making it easier for users to interpret relationships and outcomes.

Functionalities:

1. Regression Plot Generation:
    a. Generates a scatter plot using ggplot2, displaying the relationship between study hours and scores.
    b. Adds a regression line to visualize the linear pattern within the dataset.
    c. Clearly labels axes and provides an informative title for better user understanding.
2. Real-Time Prediction Display:
    a. Highlights the predicted score point based on the user's input of study hours.
    b. Dynamically updates the plot whenever the input changes, providing instant feedback.
    c. Shows how the predicted point aligns with the regression line for visual validation.
3. User Interface Integration::
    a. Seamlessly integrates the plot into the Shiny app's main panel.
    b. Ensures the plot automatically refreshes with changes in input or uploaded data.

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

CONCLUSION

The **"Predicting Scores Based on Study Hours Using Linear Regression"** project highlights the growing importance of machine learning in education by demonstrating how predictive modeling can help estimate student academic performance. Using R programming and the linear regression algorithm, the system models the relationship between study hours and exam scores with a strong positive correlation. This project proves that even with a basic dataset and a simple statistical method, valuable insights can be drawn to aid in learning and performance improvement.

A major feature of this project is its **interactive R Shiny application**, which allows users to upload their own datasets, input custom study hours, and receive instant predictions. Alongside this functionality, the inclusion of a regression plot provides a visual representation of the data and the prediction model. This not only enhances user understanding but also serves as a great learning aid for students new to machine learning and data visualization. The system's ability to provide on-the-fly predictions makes it a practical tool for both educators planning curriculum strategies and students optimizing their study schedules.

In conclusion, this project successfully integrates core concepts from data science, programming, and educational analysis into one accessible tool. It promotes the use of data-driven decisions in academic planning while demonstrating the practical application of linear regression in real-life scenarios. The modular and well-documented design of the system also makes it suitable for future upgrades, such as adding more variables, incorporating predictive intervals, or adapting it for larger educational institutions.

FUTURE SCOPE

While the current version of the project achieves its primary objective, there are several potential enhancements and extensions to improve its functionality and applicability:

1. **Incorporate Multiple Predictors**

   Extend the model to include additional factors affecting academic performance such as attendance, previous exam scores, study environment, and participation in extracurricular activities. This would allow for multivariate regression models, providing more accurate and comprehensive predictions.

2. **Advanced Machine Learning Models**

   Integrate more sophisticated algorithms such as decision trees, random forests, or support vector machines to capture complex, non-linear relationships between study habits and academic scores, improving prediction accuracy.

3. **Interactive Time and Score Filters**

   Add input options for users to filter data by date ranges, semesters, or specific exam types, enabling analysis of performance trends over different periods and conditions.

4. **Export & Reporting Features**

   Allow users to export prediction results, regression plots, and summary reports in various formats (CSV, PDF, PNG) for offline review, record-keeping, or sharing with educators and students.

5. **Real-Time Data Integration**

   Enable real-time data collection through integration with educational management systems or live input from students to provide dynamic, up-to-date performance predictions without manual dataset uploads.

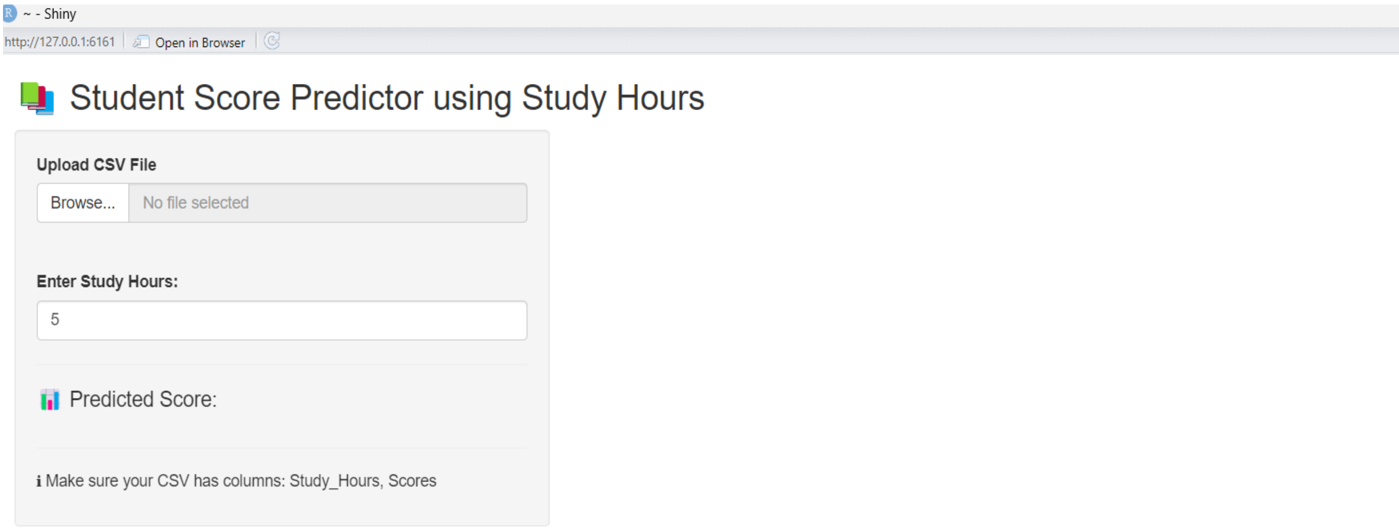
CHAPTER 5

APPENDIX A SOURCE CODE

```
library(shiny)
library(ggplot2)

ui <- fluidPage(
  titlePanel("📚 Student Score Predictor using Study Hours"),

  sidebarLayout(
    sidebarPanel(
      fileInput("file", "Upload CSV File", accept = ".csv"),
      numericInput("input_hours", "Enter Study Hours:",
                   value = 5, min = 1, max = 10, step = 0.25),
      hr(),
      h4("📊 Predicted Score:"),
      verbatimTextOutput("predicted_score"),
      hr(),
      h5("ℹ Make sure your CSV has columns: Study_Hours, Scores")
    ),

    mainPanel(
      plotOutput("regressionPlot")
    )
  )
)

server <- function(input, output) {
  data <- reactive({
    req(input$file)
    df <- read.csv(input$file$datapath)
    validate(
      need(all(c("Study_Hours", "Scores") %in% colnames(df)),
           "CSV must contain 'Study_Hours' and 'Scores' columns.")
    )
    df
  })

  model <- reactive({
    df <- data()
    lm(Scores ~ Study_Hours, data = df)
  })
  prediction <- reactive({
    req(data(), input$input_hours)
    predict(model(), newdata = data.frame(Study_Hours = input$input_hours))
  })

  output$predicted_score <- renderText({
    req(prediction())
    paste("Predicted score for", input$input_hours, "hours of study is:",
round(prediction(), 2))
  })

  output$regressionPlot <- renderPlot({
    req(data())
    df <- data()
    df$Predicted <- predict(model())

    ggplot(df, aes(x = Study_Hours, y = Scores)) +
      geom_point(color = "blue", size = 3) +
      geom_line(aes(y = Predicted), color = "red", size = 1.2) +
      geom_point(aes(x = input$input_hours, y = prediction()), color =
"darkgreen", size = 4) +
      labs(
        title = "📈 Study Hours vs Scores (with Prediction)",
        x = "Study Hours",
        y = "Score"
      ) +
      theme_minimal()
  })
}

shinyApp(ui = ui, server = server)
```

```
* In numericInput():
  min = 1, max = 10
```

APPENDIX B SCREENSHOTS



~ - Shiny

http://127.0.0.1:6161 Open in Browser

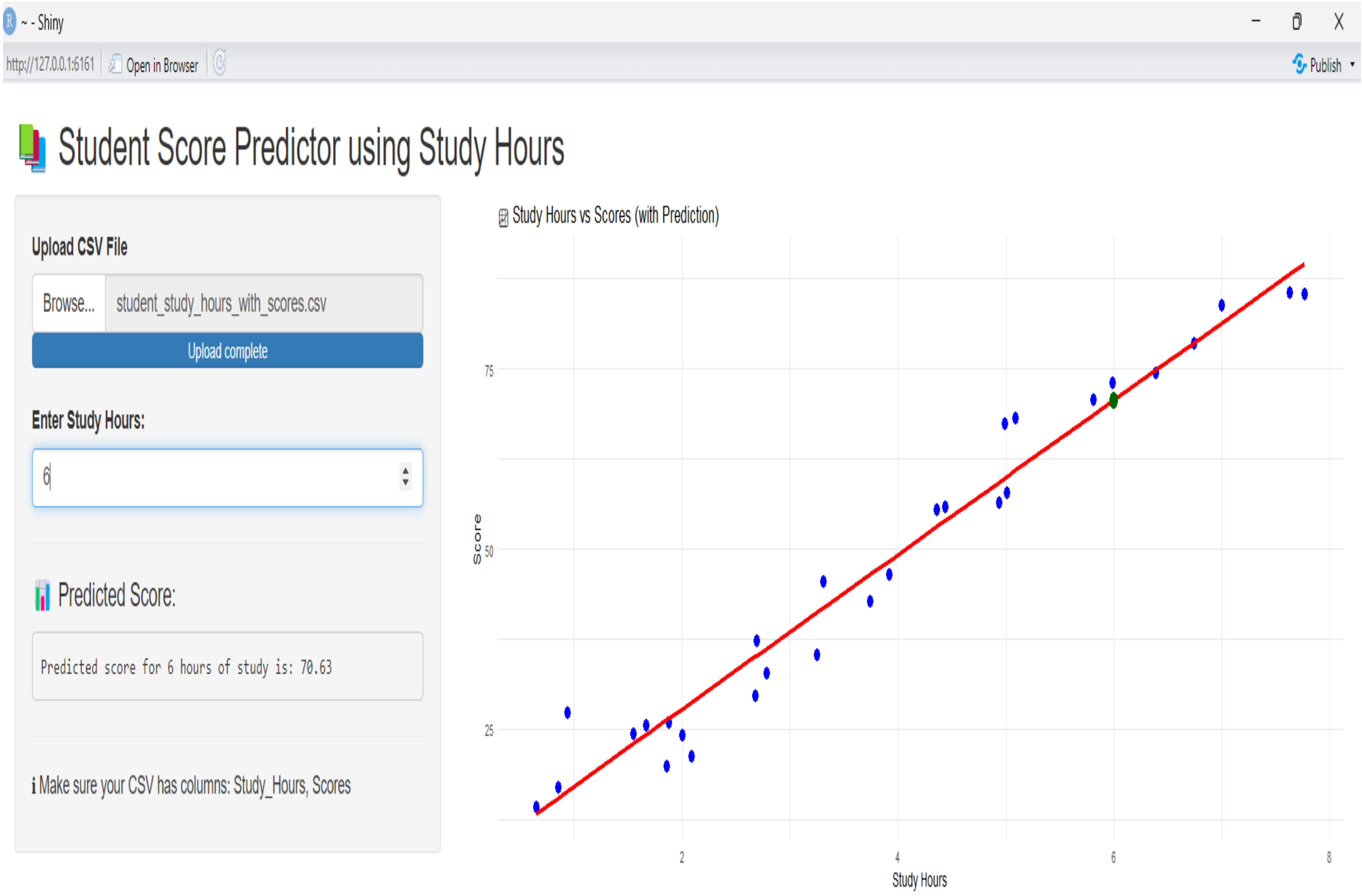
Student Score Predictor using Study Hours

Upload CSV File

Enter Study Hours: 5

Predicted Score:

ℹ Make sure your CSV has columns: Study_Hours, Scores



Student Score Predictor using Study Hours

Upload CSV File

Browse... student_study_hours_with_scores.csv

Upload complete

Enter Study Hours:

6

Predicted Score:

```
Predicted score for 6 hours of study is: 70.63
```

ℹ Make sure your CSV has columns: Study_Hours, Scores

REFERENCES

1. Books

An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani Comprehensive guide on statistical learning methods, including linear regression.

R for Data Science by Hadley Wickham and Garrett Grolemund Practical book on data manipulation, visualization, and modeling using R.

2. Documentation

R Documentation:https://www.r-project.org/Official documentation for the R language and its statistical packages.

Shiny Package Reference:https://shiny.rstudio.com/reference/shiny/latestDetailed API and examples for building Shiny web apps.

3. Videos

Data School – Linear Regression with R (YouTube):https://www.youtube.com/watch?v=nk2CQITm_eoClear explanations and practical examples of linear regression in R.