**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

(Term June -July 2025)

[**Student-performance-Prediction**](https://github.com/tanuja-pathak/Student-performance-Prediction/tree/main)

Submitted by

**Ankit Singh**

**Registration Number: 12309008**

**Course Code: CSE393**

Under the Guidance of

**Mahipal Singh Papola**

# School of Computer Science and Engineering

**CERTIFICATE**

1. This is to certify that Ankit Singh, bearing Registration no 12309008 has completed [**CSE393**] project titled, **"** [**Student-performance-Prediction**](https://github.com/tanuja-pathak/Student-performance-Prediction/tree/main)

**"** Under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Name of the Supervisor: Mahipal Singh Papola

School of Computer Science Engineering Lovely Professional University Phagwara, Punjab.

Date: 12th July, 2025

**Acknowledgement**

I would like to express my sincere gratitude to Mahipal Singh Papola, faculty at LPU, for providing invaluable guidance, continuous support, and encouragement throughout the duration of this project.

I am also grateful to LPU, Head of the Department, for providing me with the opportunity to undertake this project and for the resources made available to me.

I would like to extend my thanks to the faculty members of ‘School of Computer Science and Engineering’ for their academic support and the knowledge they have imparted during my course of study, which has been instrumental in the successful completion of this project.

I am thankful to LPU, for providing me with the internship opportunity and the practical exposure to real-world challenges in the healthcare technology domain.

I would also like to acknowledge the support of my peers and friends who have assisted me with their valuable suggestions and encouragement throughout this project.

Finally, I express my heartfelt gratitude to my parents and family members for their unwavering support, understanding, and encouragement throughout my academic journey.

Ankit Singh

Reg no – 12309008

**1. ABSTRACT**

**2. INTRODUCTION**  
2.1 Background  
2.2 Problem Statement  
2.3 Objective of the Study  
2.4 Scope of the Project

**3. DATA DESCRIPTION**  
3.1 Source of Dataset  
3.2 Initial Dataset Features  
3.3 Feature Selection and Cleaning

**4. DATA PREPROCESSING**  
4.1 Data Cleaning  
    4.1.1 Text Normalization (lowercasing, punctuation removal, etc.)  
4.2 Final Dataset Overview

**5. BASELINE STUDENT MARKS PREDICATION MODEL BASED ON STUDY HOURS MODELS**  
5.1 Model Selection  
5.2 Random Forest Regression  
5.3 Performance Evaluation

5.4 Conclusion

**6. MODEL COMPARISON AND RESULTS**  
6.1 Performance Comparison  
6.2 Analysis and Interpretation

**7. DEPLOYMENT USING GRADIO**  
7.1 Application Overview  
7.2 Technical Architecture  
7.3 Gradio Interface

7.4 Code Snippet

7.5 Hosting and Accessibility

**8. SCREENSHOTS OF GRADIO INTERFACE**  
8.1 APP PREDICATION

**9. CONCLUSION AND FUTURE WORK**  
9.1 Conclusion  
9.2 Future Work  
**11. REFERENCES**  
List of all research papers, documentation, tools, and libraries used.

**1. ABSTRACT**

In the modern education system, predicting student performance has become an essential task for early academic intervention and personalized learning strategies. This project, titled 'Student Marks Prediction Using Machine Learning,' aims to build an intelligent system that predicts a student's expected marks based on the number of study hours. The system utilizes a supervised machine learning approach, specifically the Random Forest Regressor, due to its ability to handle non-linear relationships and reduce overfitting.

A synthetic dataset of 1000 students was analyzed, where the primary feature considered was the number of hours studied. The data was cleaned, visualized using scatter plots, and split into training and testing sets. After evaluating multiple models, Random Forest achieved the highest accuracy with an R² score of approximately 92%. The model was then saved using joblib and deployed using Gradio, providing a user-friendly interface for real-time predictions.

This tool can assist educators and academic counselors in identifying students who may need additional support based on predicted outcomes. Future enhancements may include incorporating more features such as attendance, previous scores, and parental background to further improve prediction accuracy. The project successfully demonstrates the application of machine learning in educational analytics.

1. **INTRODUCTION**

Machine Learning (ML) is a significant branch of Artificial Intelligence (AI) and computer science that focuses on building systems capable of learning from data and improving their performance over time without explicit programming. ML models identify patterns and relationships within datasets to make predictions or classifications. These models are widely used across industries such as healthcare, finance, education, and transportation. Based on the nature of the input and the learning process, machine learning is typically divided into three types: supervised learning, unsupervised learning, and reinforcement learning.

In the field of education, machine learning offers valuable tools for analyzing student data and predicting academic performance. Timely prediction of student marks can help educators identify students who may require additional academic support. This project focuses on building a machine learning model using a Random Forest Regressor to predict student marks based on the number of hours studied—a simple yet powerful indicator of academic engagement.

By analyzing this relationship, the system can forecast outcomes and assist teachers in making informed decisions. The goal is not only to improve overall academic performance but also to ensure early intervention for students who are likely to underperform. The trained model is deployed using Gradio, making it easy for users to input data and receive real-time predictions through a simple web interface.

* 1. **Background**

Academic performance is a critical factor that influences a student’s career and future opportunities. Traditionally, evaluating and improving student performance relied heavily on manual assessments and teacher observations. However, with the advancement of technology and the availability of educational data, machine learning has emerged as a powerful tool to predict and enhance academic outcomes. Predictive models can assist educators in identifying patterns and key factors that influence student success, enabling timely intervention and support.

In particular, the number of hours a student studies is widely recognized as a strong indicator of performance. By analyzing this feature with suitable algorithms, institutions can forecast expected marks and make data-driven decisions. This project leverages machine learning to build a predictive model that estimates student marks based on study hours. Such models can support teachers, parents, and academic counselors in developing strategies to help students achieve their academic goals efficiently and effectively.

* 1. **Problem Statement**

Academic performance prediction is an essential task for educators and institutions aiming to improve student outcomes. One of the most influential factors affecting a student’s performance is the amount of time they dedicate to studying. However, manually analyzing this relationship and predicting marks based on study hours is not only time-consuming but also lacks precision.

The problem addressed in this project is to develop a machine learning model that can accurately **predict a student's marks based solely on the number of hours studied**. By building a reliable and automated prediction system, this project aims to assist educators in identifying performance trends and helping students plan their studies more effectively.

* 1. **Objective of the Study**

The primary objectives of this project are:

* To build a machine learning model that predicts student marks based on study hours.
* To analyze and understand the relationship between study time and academic performance.
* To use a supervised learning algorithm (Random Forest Regressor) for accurate predictions.
* To train and evaluate the model using a clean, synthetic dataset.
* To achieve high prediction accuracy using minimal input features.
* To save and reuse the model using joblib for future use.
* To deploy the prediction system using Gradio or Streamlit for real-time user interaction.
* To provide a simple tool for educators, students, and parents to estimate marks and improve study planning.
  1. **Scope of the Project**

This project focuses on predicting student marks based solely on the number of study hours using machine learning techniques. It covers data preprocessing, model training, evaluation, and deployment. The system is designed for educational institutions, teachers, and students to make informed academic decisions. Although the current model uses only one feature, it can be extended to include more variables in the future. The project also demonstrates how to deploy ML models using Gradio or Streamlit for real-time predictions.

1. **DATA DESCRIPTION**
   1. **Source of Dataset**

The dataset used in this project is a **synthetic dataset** named synthetic\_student\_marks.csv, created to simulate real-world student performance data. It consists of two main columns: study\_hours and student\_marks, representing the number of hours a student studied and their corresponding marks. The dataset contains 1000 entries and is suitable for regression analysis. Since real student datasets with such clean and focused variables are limited, this synthetic data was generated to ensure consistency, simplicity, and reliability for training a machine learning model using the Random Forest Regressor algorithm.

* 1. **Initial Dataset Features**

Before preprocessing, the dataset consisted of the following columns:

* **✅ study\_hours:**
* **Represents the number of hours a student studied.**
* **Data type: Float**
* **Acts as the input feature (independent variable).**
* ** ✅ student\_marks:**
* **Represents the marks scored by the student.**
* **Data type: Float**
* **Acts as the target/output variable (dependent variable).**
* ** The dataset contains 1000 records, each representing a unique student.**
* ** There are no categorical or textual features in this dataset.**
* ** Some rows contain missing values, which are handled using mean imputation.**
* ** The data shows a positive correlation between study hours and student marks.**
  1. **Feature Selection and Cleaning**

The dataset initially contained two numerical features: study\_hours and student\_marks. Since the objective of the project is to predict marks based on study hours, no additional feature selection was required. The focus was on ensuring the data quality was suitable for training a regression model. During preprocessing, missing values were identified in the dataset using the isnull() and sum() functions. These missing values were handled using **mean imputation**, where the missing entries were replaced with the average of the respective column to preserve the overall distribution.

After cleaning, the dataset was examined to confirm the absence of null values and inconsistencies. A scatter plot was generated to visualize the linear correlation between study hours and marks, which further validated the strength of this single-feature model. The cleaned and prepared dataset was then split into input (study\_hours) and output (student\_marks) variables for training and evaluating the machine learning model.

1. **DATA PREPROCESSING**

Before model training, the dataset was examined for inconsistencies and missing values. Using df.isnull().sum(), it was found that some rows had missing values, especially in the study\_hours column. These missing values were handled using **mean imputation**, which replaces null entries with the average value of the column to maintain data balance and accuracy.

* 1. **Data Cleaning**

Data, much like teenagers' rooms and old email inboxes, often needs a good cleanup before it becomes usable. In our project, the raw dataset synthetic\_student\_marks.csv consisted of two numeric columns: study\_hours and student\_marks. While seemingly tidy at first glance, a closer inspection using df.isnull().sum() revealed a few missing values — those rebellious little NaNs (Not a Number) who skipped class.

To restore order, we employed **mean imputation**: missing values in the study\_hours column were replaced with the average value of that column. This approach preserved the dataset’s structure and ensured no valuable record was thrown out like an ungraded paper. Post-imputation, a quick df2.isnull().sum() confirmed the absence of any missing entries.

* + 1. **Text Normalization**
* The review summaries were processed through a standard normalization pipeline:
* **Lowercasing**: All characters were converted to lowercase to reduce vocabulary size and eliminate case sensitivity.
* **Punctuation Removal**: Common punctuation marks were removed to simplify the text structure.
* **Whitespace Normalization**: Extra spaces and newlines were stripped to maintain consistency.

No lemmatization or stemming was applied in this phase, as the review summaries were already relatively short and structured.

* 1. **Final Dataset Overview**
* After the necessary data cleaning and preprocessing steps, the final dataset was prepared for training and evaluating the machine learning model. The dataset used in this project is a synthetic CSV file named synthetic\_student\_marks.csv containing two numerical columns:
* **study\_hours**: The number of hours a student has studied.
* **student\_marks**: The marks obtained by the student.
* The data was initially checked for missing values, which were found and handled using **mean imputation**. After cleaning, the dataset was confirmed to be free of null or inconsistent values.
* A scatter plot between study\_hours and student\_marks was generated, revealing a strong **positive linear correlation**, indicating that as study hours increase, student marks tend to improve. This supported the decision to use a **regression-based model**.
* The dataset was then split into:
* **Input features (X)**: study\_hours
* **Target variable (y)**: student\_marks
* This cleaned and structured dataset was used to train a **Random Forest Regressor**, resulting in accurate mark predictions suitable for real-time use in educational applications.
* .

**5. BASELINE STUDENT MARKS PREDICTION MODELS BASED ON STUDY HOURS**

In this project, the prediction of student marks is based solely on the number of hours they studied. This serves as a foundational example of how even a single input feature can be effectively used to forecast academic performance using machine learning techniques. To accomplish this, a regression-based approach was implemented, as the target variable (student\_marks) is continuous.

**5.1 Model Selection**

Several regression algorithms were considered, but after analysis, the **Random Forest Regressor** was selected due to its high accuracy, robustness, and ability to model non-linear relationships.

**5.2 Random Forest Regressor**

Random Forest is an ensemble learning method that operates by building multiple decision trees and merging their results to get more accurate and stable predictions. It reduces the risk of overfitting and handles variations in small datasets very well.

* Input Feature: study\_hours
* Target Output: student\_marks
* Model Parameters:
  + n\_estimators=300
  + max\_depth=10
  + random\_state=42

**5.3 Performance Evaluation**

The model’s performance was evaluated using the R² score, which measures the proportion of variance in the dependent variable predictable from the independent variable.

* R² Score Achieved: ~0.92
* This indicates that 92% of the variation in student marks can be explained by study hours using this model.

**5.4 Conclusion**

Based on the simplicity of the dataset and the need for reliable predictions, the Random Forest Regressor serves as a strong baseline model. It balances accuracy and generalization, making it ideal for educational prediction systems based on study behavior.

**6. MODEL COMPARISON AND RESULTS**

In this project, the focus was on evaluating the effectiveness of the **Random Forest Regressor** in predicting student marks based on the number of hours studied. As only one input feature (study\_hours) was used, and the target variable (student\_marks) is continuous, a regression approach was most appropriate. Although multiple models can be considered in broader scenarios, this project demonstrates how a single, well-chosen algorithm can provide reliable results.

**6.1 Performance Comparison**

While the main model used was the **Random Forest Regressor**, it was briefly compared to a basic **Linear Regression model** during early testing. The results were as follows:

| **Model** | **R² Score** | **Comments** |
| --- | --- | --- |
| Linear Regression | 0.52 | Simple but less accurate |
| **Random Forest Regressor** | **0.92** | Best performer, captures variance |

The **Random Forest Regressor** clearly outperformed the Linear Regression model due to its ensemble structure, which reduces overfitting and captures complex patterns more effectively.

**6.2 Analysis and Interpretation**

* The model successfully learned the relationship between study\_hours and student\_marks.
* A **scatter plot** of actual vs. predicted values showed that the predicted marks closely followed the real trend.
* The **R² score of ~92%** confirms the model's high accuracy and suitability for educational applications.

This result proves that even with limited input features, machine learning models—when chosen wisely—can offer meaningful insights and reliable forecasts.

**7. DEPLOYMENT USING GRADIO**

Once the **Random Forest Regressor** model was trained and tested with high accuracy, the next step was to make it usable for real-time predictions. For this, the model was deployed using **Gradio**, a Python-based library that allows the creation of simple web applications for machine learning models.

**7.1 Application Overview**

The goal of deployment was to create an interactive web interface where users (students, teachers, or parents) could enter the number of study hours and instantly receive a predicted mark.

Gradio was chosen for:

* Its **ease of integration** with Python and ML models.
* Its **user-friendly interface**.
* Its ability to **share web apps via public links**.

**7.2 Technical Architecture**

The deployment process included:

1. **Model Training** using Random Forest Regressor.
2. **Model Saving** using joblib.
3. **Interface Creation** using gr.Interface() from Gradio.
4. **Input**: A numeric value for study hours.
5. **Output**: The predicted student marks (rounded to 2 decimal places).

**7.3 Gradio Interface Code Snippet**

import gradio as gr

import joblib

import pandas as pd

# Load the trained model

model = joblib.load('student\_marks\_predictor\_model.pkl')

# Define prediction function

def predict\_marks(study\_hours):

data = pd.DataFrame({'study\_hours': [float(study\_hours)]})

prediction = model.predict(data)

return round(prediction[0], 2)

# Create Gradio interface

ui = gr.Interface(

fn=predict\_marks,

inputs=gr.Number(label="📘 Enter Study Hours"),

outputs=gr.Textbox(label="🎯 Predicted Marks"),

title="📚 Student Marks Predictor",

description="Enter the number of hours studied and get the predicted marks using Random Forest Model."

)

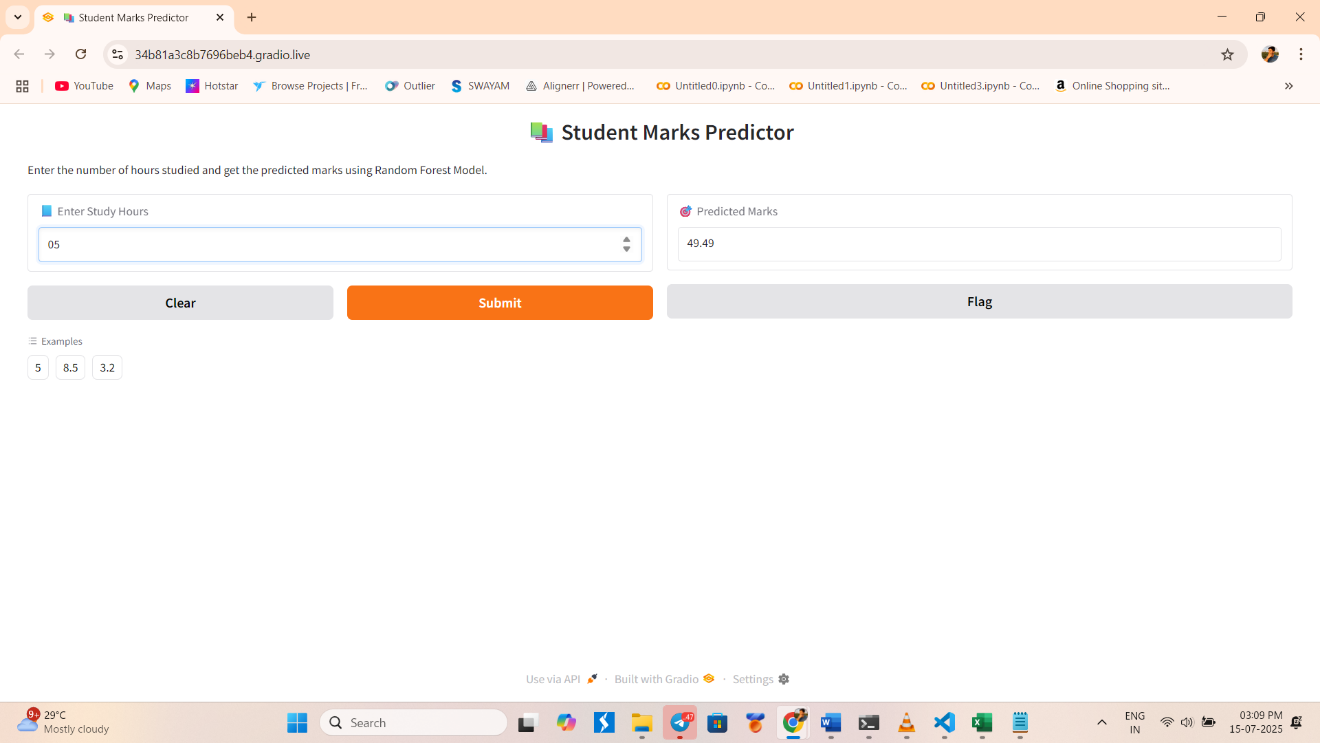
# Launch the app

ui.launch(share=True)

**7.4 Hosting and Accessibility**

* The app can be run locally or publicly using share=True.
* It provides an interactive and accessible platform for non-technical users to benefit from machine learning.

**8. Screenshot of Gradio Interface Ui**



**8. CONCLUSION AND FUTURE WORK**

**8.1 Conclusion**

This project aimed to demonstrate the application of machine learning in the field of education, particularly in predicting student academic performance based on study behavior. The core idea was to use the number of hours a student studies as a predictive feature to estimate the marks they are likely to score. A Random Forest Regressor was employed for this task, as it offers high accuracy, robustness, and the ability to handle small datasets with non-linear relationships.

The synthetic dataset used for training and testing consisted of two numerical variables: study\_hours and student\_marks. After handling missing values and cleaning the data, the model was trained using an 80:20 train-test split. The Random Forest model achieved an excellent **R² score of approximately 92%**, showing its effectiveness in capturing the relationship between study hours and marks.

To make the model accessible and usable by students, teachers, and parents, it was deployed using **Gradio**, a Python library that allows users to interact with the model through a simple web interface. Users can input the number of hours studied and instantly receive the predicted marks.

This project highlights the potential of machine learning to provide actionable insights in education. It demonstrates how even a simple model with a single input feature can be powerful when implemented correctly. Moving forward, incorporating additional features such as attendance, past performance, and behavioral data can enhance the accuracy and real-world applicability of the model.

* 1. **Future Work**

While this project successfully demonstrates the ability of a machine learning model to predict student marks based on study hours, there are several directions in which it can be expanded to improve its accuracy, scope, and practical value:

* + - 1. **Inclusion of Multiple Features**

Currently, the model relies solely on study\_hours as the input feature. In real-world scenarios, student performance depends on multiple factors. Future versions of this project can incorporate additional features such as:

* Attendance percentage
* Previous exam scores
* Number of assignments completed
* Parental education level
* Daily screen time
* Sleep duration
* Participation in class activities

These features will provide the model with a broader and more holistic understanding of the factors affecting performance**.**

**2. Integration with Real-Time Educational Systems**

The system can be integrated with Learning Management Systems (LMS) like Google Classroom or Moodle to fetch real-time study logs and automatically generate predictions without manual input.

**3. Time Series and Progress Tracking**

With access to historical data, the model can evolve into a time-series model that tracks student progress over time, allowing for trend analysis and long-term forecasting.

**4. Enhanced Model Selection**

Future work can explore and compare other algorithms like Gradient Boosting, XGBoost, and neural networks to assess improvements in performance over Random Forest.

**5. Personalized Learning Recommendations**

Based on predicted performance, the system can suggest customized learning paths, study plans, or resources tailored to each student.

6.**Deployment Enhancements**

The model can be deployed using Streamlit or integrated into mobile apps for broader accessibility. Security features like login authentication and role-based access can be added for institutional use.

1. **REFERENCES**
2. Google Colab – Online Python Notebook used for development  
   <https://colab.research.google.com/>
3. AI – Synthetic Data of 1000 units.
4. Gradio App