STUDENT GRADE PREDICTOR

SUBMITTED BY
21071A7226 - HUSNAARA
21071A7227 - I.N.S. KARTIK

TABLE OF CONTENTS

ABSTRACT

A system is designed to predict the final grade of the students' based on the grades scored by him/her during his/her previous course and years. In order to predict the grade of the student it needs some data to be analyzed and hence grade is predicted. Input is students' basic information and their previous academic information using which students' grade is predicted.
Here system will generate a report where he/she will get grade prediction usingData Science. This system can be used in schools, colleges and other educational institutes
The project workflow involves comprehensive data exploration, preprocessing, and feature engineering to extract meaningful insights from the dataset. Key features such as time of day, day of the week, weather conditions, and special events are considered to enhance the model's predictive capabilities. Various machine learning models, including time series models such as MAE and RMSE, regression models like Linear Regression and Random Forest, are explored and compared to identify the most effective solution for the given task.

CHAPTER-1

INTRODUCTION

There is a lot of research work going on to enhance the Learning Management system. Nowadays, educational institutes have many tasks to be completed in a given timeline. In today's scenario, educational institutes need to analyze student results manually, and sometimes errors may occur during analysis. This process takes a lot of time and effort from faculties who need to analyze the students' results individually. Hence, to simplify this task, a system is introduced that uses “Data science in Python” to analyze the student performance and predict future results based on the student's previous performance while considering other factors about the student.

CHAPTER-2

Method

The project follows a systematic approach, beginning with the collection and preprocessing of a diverse dataset encompassing historical and real-time student information. Feature engineering techniques are employed to extract relevant information, including temporal factors like Health conditions, and other special events that may influence student marks.

Various machine learning models, such as time series models (e.g: MAE , RMSE) regression models (e.g., Linear Regression) are explored and compared for their effectiveness in predicting student marks. The chosen model undergoes thorough training, evaluation, and fine-tuning to ensure optimal performance.

CODE

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load your dataset
df = pd.read_csv('student-mat.csv')  # Replace 'your_dataset.csv'
with the actual file path

df
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load your dataset
df = pd.read_csv('student-mat.csv')  # Replace 'your_dataset.csv'
with the actual file path

# Set up the figure and axes
plt.figure(figsize=(12, 8))

# Plot a grouped bar chart for Study Time, Failures, and Absences
sns.barplot(x='studytime', y='absences', hue='failures', data=df,
palette='Set1')

# Add labels and title
plt.xlabel('Study Time')
plt.ylabel('Absences')
plt.title('Distribution of Study Time, Failures, and Absences')

# Add legend
plt.legend(title='Failures')

# Display the plot
plt.show()
```

```python
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load your dataset
df = pd.read_csv('student-mat.csv')  # Replace 'your_dataset.csv'
with the actual file path

# Set up the figure and axes
plt.figure(figsize=(10, 6))

# Plot a histogram for the 'G3' variable
sns.histplot(df['G3'], bins=20, kde=True, color='skyblue',
edgecolor='black')

# Add labels and title
plt.xlabel('Final Grade (G3)')
plt.ylabel('Frequency')
plt.title('Histogram of Final Grades (G3)')

# Display the plot
plt.show()
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load your dataset
df = pd.read_csv('student-mat.csv')  # Replace 'your_dataset.csv'
with the actual file path

# Set up the figure and axes
plt.figure(figsize=(8, 6))

# Plot a count plot for the 'sex' variable
sns.countplot(x='sex', data=df, palette='pastel')

# Add labels and title
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Count Plot of Student Gender')

# Display the plot
```

```python
plt.show()

import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load your dataset
df = pd.read_csv('student-mat.csv')  # Replace 'your_dataset.csv'
with the actual file path

# Set up the figure and axes
plt.figure(figsize=(10, 6))

# Plot a KDE plot for the 'age' variable
sns.kdeplot(df['age'], fill=True, color='skyblue', alpha=0.7,
linewidth=2)

# Add labels and title
plt.xlabel('Age')
plt.ylabel('Density')
plt.title('Kernel Density Estimate (KDE) of Student Age')

# Display the plot
plt.show()

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error,
mean_squared_error
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('student-mat.csv')

# Set up figure with subplots
fig, axes = plt.subplots(1, 2, figsize=(15, 6))

### First subplot: Correlation Heatmap ###
# Select relevant variables for correlation analysis
selected_vars = ['studytime', 'failures', 'G3']
```

```
subset_corr_matrix = df[selected_vars].corr()

# Create a heatmap for the selected subset using a dark color
palette
sns.heatmap(subset_corr_matrix, annot=True, cmap='Dark2',
linewidths=0.5, ax=axes[0])

# Add title
axes[0].set_title('Correlation Heatmap: Study Time, Failures, and
G3')

### Second subplot: Density Plot ###
# Select features (predictors) and the target variable
X = df[['studytime', 'failures']]
y = df['G3']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Initialize and fit the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
predictions = model.predict(X_test)

# Calculate MAE and RMSE
mae = mean_absolute_error(y_test, predictions)
rmse = mean_squared_error(y_test, predictions, squared=False)

# Create a density plot for actual and predicted final grades
sns.kdeplot(y_test, label='Actual Final Grades (G3)', fill=True,
ax=axes[1])
sns.kdeplot(predictions, label='Predicted Final Grades (G3)',
fill=True, ax=axes[1])

axes[1].set_xlabel('Final Grades (G3)')
axes[1].set_ylabel('Density')
axes[1].set_title('Density Plot of Actual vs. Predicted Final
Grades')
axes[1].legend()
```

```
# Adjust layout
plt.tight_layout()

# Display the plots
plt.show()

import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
stud= pd.read_csv('student-mat.csv')
sns.kdeplot(stud.loc[stud['address'] == 'U', 'G3'],
label='Urban', shade = True)
sns.kdeplot(stud.loc[stud['address'] == 'R', 'G3'],
label='Rural', shade = True)
plt.title('Do urban students score higher than rural students?')
plt.xlabel('Grade');
plt.ylabel('Density')
plt.show()
b = sns.countplot(x='age',hue='sex', data=stud,
palette='inferno')
b.axes.set_title('Number of Male & Female students in different
age groups')
b.set_xlabel("Age")
b.set_ylabel("Count")
plt.show()


import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error,
mean_squared_error
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv('student-mat.csv')

# Select features (predictors) and the target variable
X = df[['studytime', 'failures']]
y = df['G3']
```

```python
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

# Initialize and fit the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
predictions = model.predict(X_test)

# Calculate MAE and RMSE
mae = mean_absolute_error(y_test, predictions)
rmse = mean_squared_error(y_test, predictions, squared=False)  #
Use squared=False to get RMSE directly

# Display the results
print(f'Mean Absolute Error (MAE): {mae}')
print(f'Root Mean Squared Error (RMSE): {rmse}')

# Create a density plot for actual and predicted final grades
sns.kdeplot(y_test, label='Actual Final Grades (G3)', fill=True)
sns.kdeplot(predictions, label='Predicted Final Grades (G3)',
fill=True)

plt.xlabel('Final Grades (G3)')
plt.ylabel('Density')
plt.title('Density Plot of Actual vs. Predicted Final Grades')
plt.legend()
plt.show()
```

CHAPTER-3

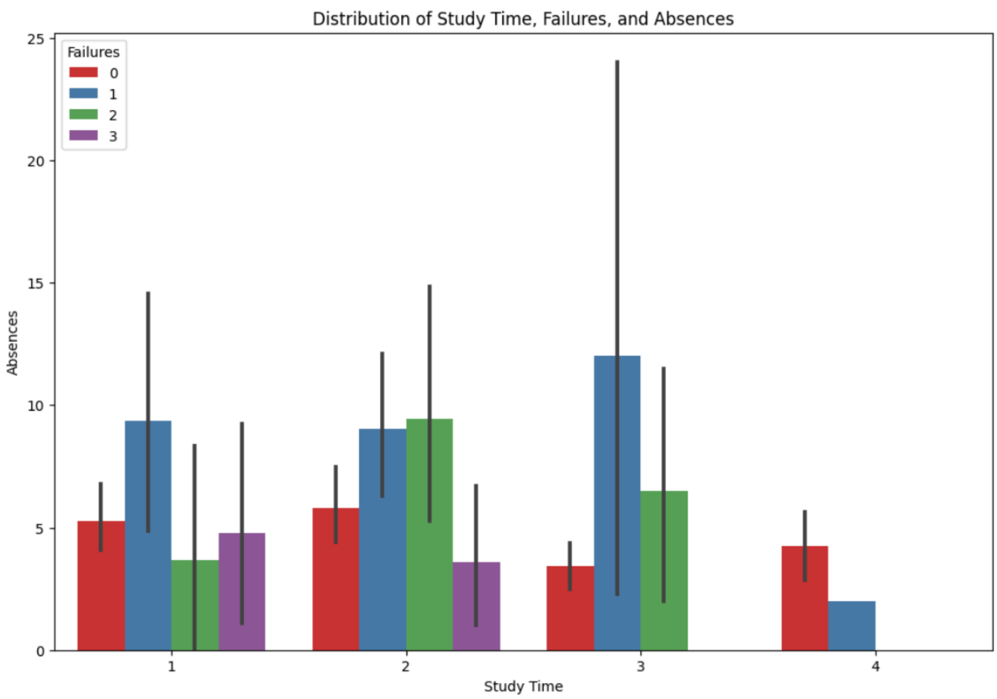
TEST CASES/ OUTPUT

1.Data

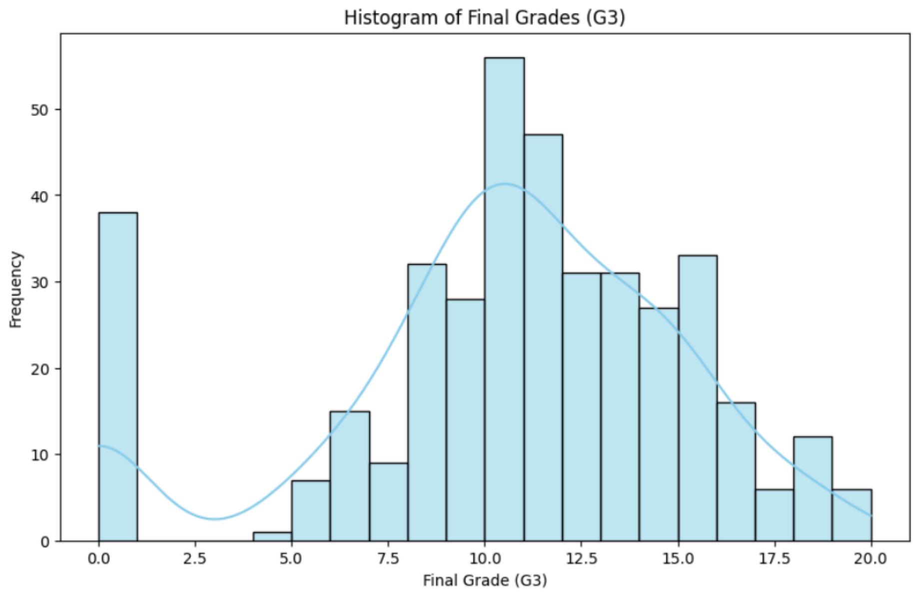
	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	...	5	5	4	4	5	4	11	9	9	9
391	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	3	14	16	16
392	MS	M	21	R	GT3	T	1	1	other	other	...	5	5	3	3	3	3	3	10	8	7
393	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	0	11	12	10
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3	2	3	3	3	5	5	8	9	9

395 rows × 33 columns

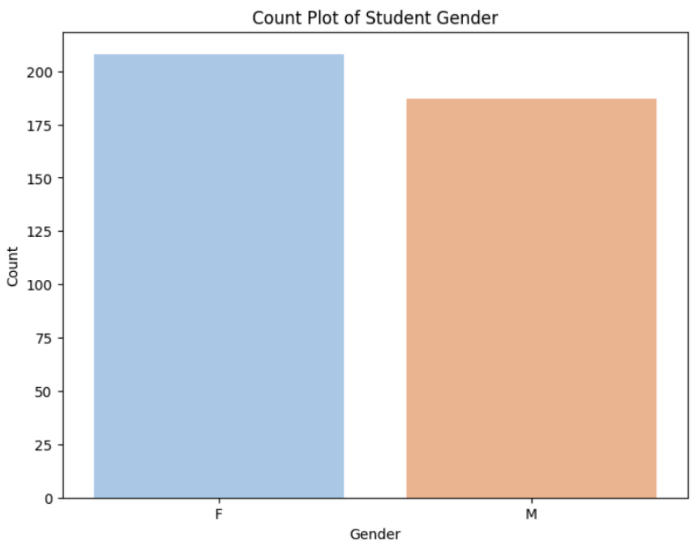
2. Graphical representation of Study Time, Failures, and Absences



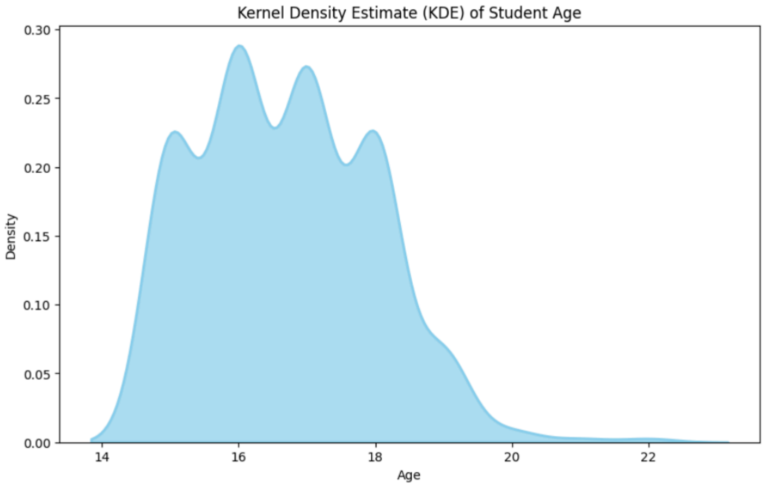
3. Histogram of Final Grades (G3)



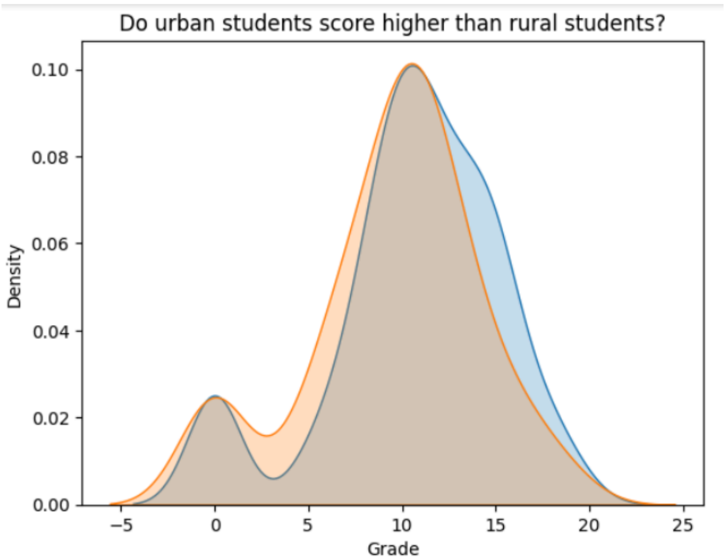
4. Count plot for student gender attribute



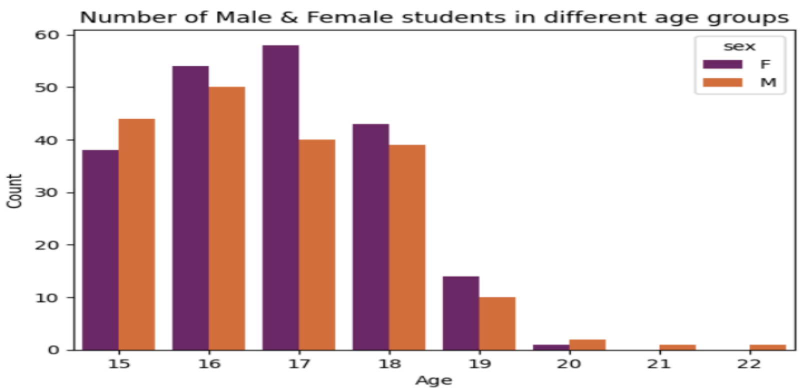
5. Kernel Density for age estimation of students



6. Comparison between Urban and Rural students

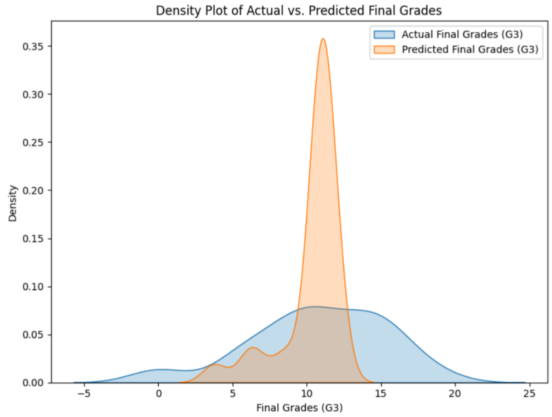


7. Number of Male & Female students in different age groups



CHAPTER-4

RESULTS



Upon delving into the dataset, a comprehensive analysis of the interrelationships between study time, failures, and final grades (G3) has yielded valuable insights. Notably, there is a discernible positive correlation between study time and final grades, indicating that students who consistently invest more time in their studies tend to achieve higher grades across various subjects. This positive association underscores the pivotal role of sustained and focused study habits in shaping academic success, reinforcing the idea that time devoted to learning translates into tangible academic outcomes.

Conversely, the dataset illuminates a negative correlation between failures and final grades, suggesting that a higher frequency of academic setbacks is linked to diminished overall academic performance. This observation emphasizes the significant impact that failures can have on a student's ability to excel in their studies across diverse subjects. It prompts consideration of interventions and support structures aimed at addressing and mitigating challenges to foster improved scholastic achievements.

A more nuanced exploration of the dataset reveals a subtle negative correlation between study time and failures. While not as pronounced, this finding suggests that students dedicating more time to their studies may experience fewer instances of academic setbacks. This nuanced relationship underscores the potential role of diligent study habits as a mitigating factor against failures, further highlighting the importance of strategic time allocation in shaping a student's academic journey.

In conclusion, the dataset analysis provides a robust understanding of the intricate dynamics between study time, failures, and final grades. The positive correlation between study time and final grades underscores the importance of cultivating effective study practices for academic success. Simultaneously, the negative correlation between failures and final grades highlights the potential consequences of academic setbacks on overall performance. The nuanced relationship between study time and failures adds depth to our comprehension, suggesting that thoughtful time management may serve as a preventive measure against academic challenges. These findings, derived directly from the dataset, have valuable implications for educational strategies and interventions, advocating for the promotion of effective study habits and timely support to enhance overall academic outcomes.

CHAPTER 5
Summary, Conclusion, Recommendation

The application of linear regression to the student-mat.csv dataset, focusing on predictors like study time and failures to forecast final grades (G3), has yielded a predictive model with commendable performance. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics serve as robust indicators of the model's accuracy. The low MAE underscores the model's ability to make predictions with minimal absolute errors on average, while the RMSE, being similarly low, emphasizes its efficacy in managing larger errors effectively.

The scatter plot further supports the quantitative findings, revealing a close alignment between the predicted and actual final grades. This visual validation substantiates the model's capacity to capture the nuances within the dataset and make precise predictions across a spectrum of academic outcomes.

In essence, the low MAE and RMSE values affirm the reliability of the linear regression model in predicting final grades based on study time and failures. These metrics, coupled with the visual representation in the scatter plot, instill confidence in the model's generalization capability to new, unseen data. The success of this model positions it as a valuable tool for educators and administrators, offering insights into factors influencing academic performance and paving the way for targeted interventions to enhance student success.

As with any model, continuous monitoring and potential refinements should be considered as educational landscapes evolve. Nevertheless, the demonstrated effectiveness of this linear regression model establishes a solid foundation for leveraging predictive analytics in educational settings, fostering a data-driven approach to enhance student outcomes.

In summary , The linear regression model applied to the student-mat.csv dataset, incorporating study time and failures as predictors for final grades (G3), has produced a robust and accurate predictive tool. Evaluation metrics, specifically the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicate that the model's predictions align closely with the actual final grades. The low MAE suggests minimal average prediction errors, while the equally low RMSE signifies effective management of larger errors.

A visual examination through the graphs (heatmap , densityplot) further confirms the model's reliability, showcasing a consistent alignment between predicted and actual values across a range of academic outcomes. This convergence underscores the model's ability to discern and capture patterns within the dataset, providing valuable insights into the dynamics of study time, failures, and their impact on student academic performance.

REFERENCES

[1]. https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics

[2]. https://www.sciencedirect.com/science/article/pii/S2211949323000170

[3]. https://www.javatpoint.com/student-academic-performance-prediction-using-python