আন্তর্জাতিক ইসলামী বিশ্ববিদ্যালয় চট্টগ্রাম
الجامعة الإسلامية العالمية شيتاغونغ
International Islamic University Chittagong
Department of Computer Science and Engineering

Project Report

Student Performance Analysis using Machine Learning

Course Details	
Course Code	**Course Title**
CSE-3636	Artificial Intelligence Lab

Course Instructor
Sara Karim Adjunct Lecturer, Department of CSE, IIUC

Team Members			
ID	**Name**	**Section**	**Semester**
C223205	Syeda Afra Anam	6AF	6th, Spring 2025
C223215	Sumaiya Akter Emo	6AF	6th, Spring 2025

Submission Date
09/07/25

Table of Content

1. Project Title:

Student Performance Analysis using Machine Learning.

2. Objective:

This project analyzes a student performance dataset to:

i. Understand how various factors (gender, lunch type, parental education, etc.) affect student grades.
ii. Detect and handle outliers and impossible values.
iii. Perform Exploratory Data Analysis (EDA) for insights.
iv. Build a machine learning model to predict student grades.
v. Create an interactive Streamlit web app for predictions and visualization.

3. Technologies Used:

- Programming Language: Python
- Python Libraries: pandas, numpy, matplotlib, seaborn, sklearn, streamlit
- Machine Learning Model: Random Forest Classifier
- Frontend Framework: Streamlit
- Backend: Python (for data processing, model training, and prediction logic)
- Development Environment: Kaggle Notebook, Jupyter Notebook, Streamlit
- Software Tools: Visual Studio Code

4. Dataset Source:

StudentsPerformance.csv
https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

5. Data Cleaning and Preprocessing:

Data cleaning and preprocessing are essential to ensure the dataset is accurate, consistent, and ready for analysis. It helps remove errors, handle inconsistencies, and create meaningful features that improve model performance and insights.

Here we :

i. Checked for missing values and confirmed the dataset had none, ensuring data completeness.
ii. Converted column names to snake_case.
iii. Engineered features:
    - average_score, total_score
    - Binary flags: prep_completed, standard_lunch
iv. Created grade labels from average score using custom bins:
    - Grades: F, D, C, C+, B, B+, A-, A, A+

6. Handling Data Quality:

Handling data quality is necessary to ensure the dataset is reliable and accurate. Keeping valid outliers preserves important variations, while removing impossible values prevents errors that could mislead analysis and model training.

i. Outliers: Retained valid outliers since they may reflect real student performance variance.
ii. Impossible Values: Removed rows where any score is below 0 or above 100.

7. Data Description:

The dataset contains exam performance data for 1,000 students. It includes demographic information such as gender, race/ethnicity, parental level of education, lunch type, and test preparation course completion. The academic scores consist of three subject scores: math, reading, and writing, each ranging from 0 to 100. The data is clean, with no missing values, and serves as a basis to analyze factors affecting student grades and to build predictive models.

8. Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is performed to understand the underlying patterns, relationships, and distributions within the data. It helps identify trends, correlations, and factors that influence student performance, guiding better model building and interpretation.
Here we:

i. Analyzed effects of:
   - Gender
   - Lunch type
   - Test preparation course
   - Parental education level

ii. Visualized:
   - Score distributions
   - Grade distributions
   - Correlations between scores
   - Students per grade

iii. Key EDA Insights:

SL No.	Questions	Insight
a.	Does gender influence performance?	Female students tend to perform slightly better on average
b.	Does lunch type affect scores?	Students with standard lunch generally score higher than those with free/reduced lunch.
c.	Is test preparation effective?	Yes, students who completed the course perform better
d.	Parental education effect on scores?	Higher education level correlates with higher average scores
e.	Correlation between scores?	Strong positive correlation among math, reading, and writing
f.	Grade distribution	Most students are within B to A range
g.	Number of students per grade	The majority of students are concentrated in middle to higher grade bins.
h.	Correlations between scores	Strong positive correlations exist among math, reading, and writing scores.
i.	Score distribution by test preparation	Prepared students have higher median scores and less variability than unprepared ones.

9. Model Building:

Model building is necessary to create a system that can learn from existing student data and accurately predict future student grades. It enables identifying key factors influencing performance and provides a practical tool for automated grade prediction.

Here:

i. Algorithm used: Random Forest Classifier
ii. Target: Grade (categorical)
iii. Input Features:
    - Gender, race/ethnicity, parental education, lunch, test preparation, scores, flags
iv. Performance:
    - Accuracy: 0.9900
    - Evaluated with classification report and feature importance

10. Prediction System:

Four main tabs:

i. Dashboard: Data filtering by gender, lunch, parental education
ii. EDA: Interactive insights
iii. Model: Accuracy, feature importance, and classification report
iv. Predict: Enter a student's details and predict the grade

11. Model Evaluation:

i. Classifier Used: Random Forest Classifier
ii. Features Used: All numerical and encoded categorical fields
iii. Evaluation Metrics:
    - Accuracy Score
    - Classification Report
    - Feature Importance Chart

12. Top Features:

i. Math Score
ii. Reading Score
iii. Writing Score

13. Prediction Example:

A student with the following profile:

i. Gender: Female
ii. Parental Education: Bachelor's Degree
iii. Lunch: Standard
iv. Test Preparation: None
v. Math: 80, Reading: 81, Writing: 82

Predicted Grade: A

14. Result:

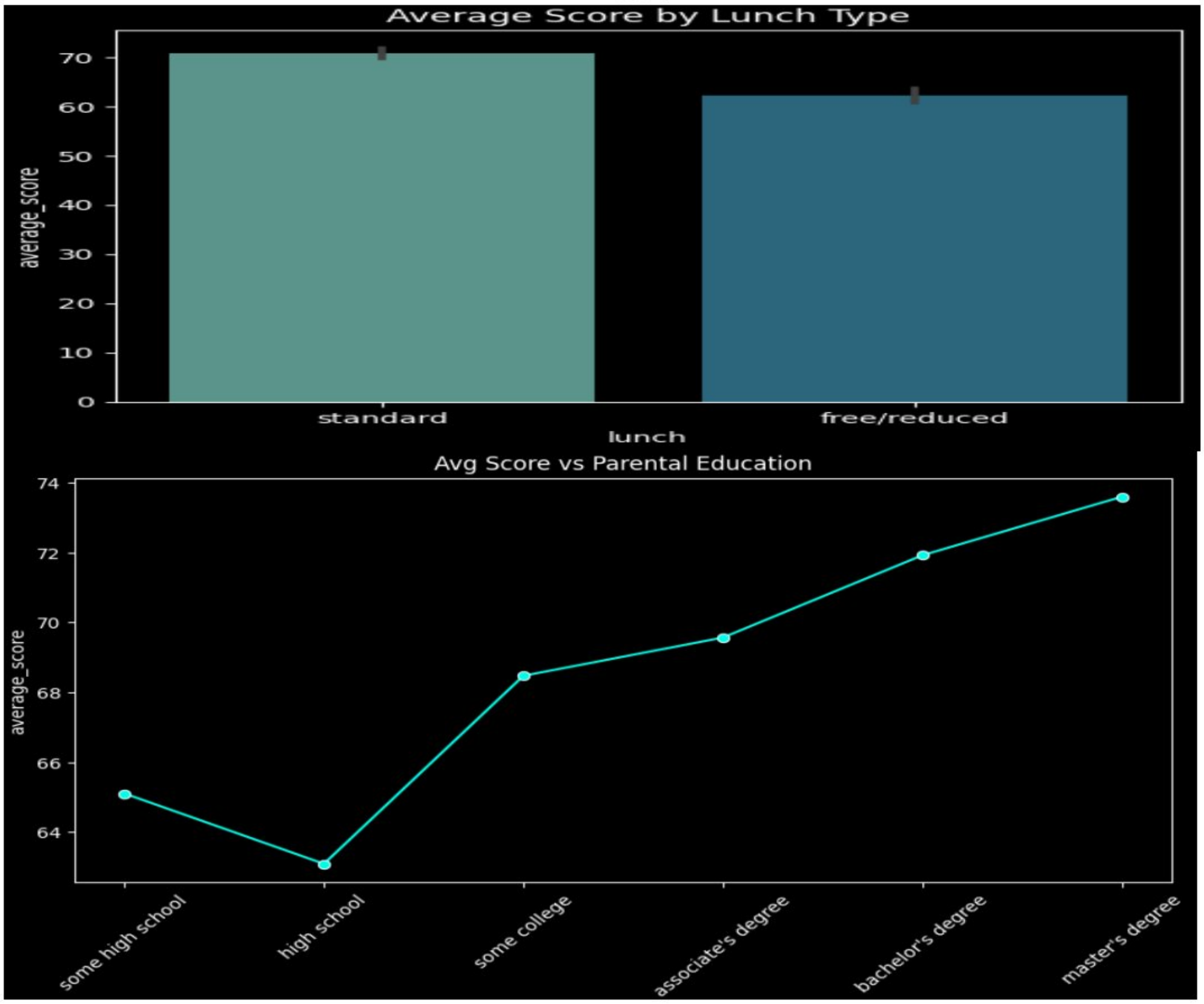
Dashboard

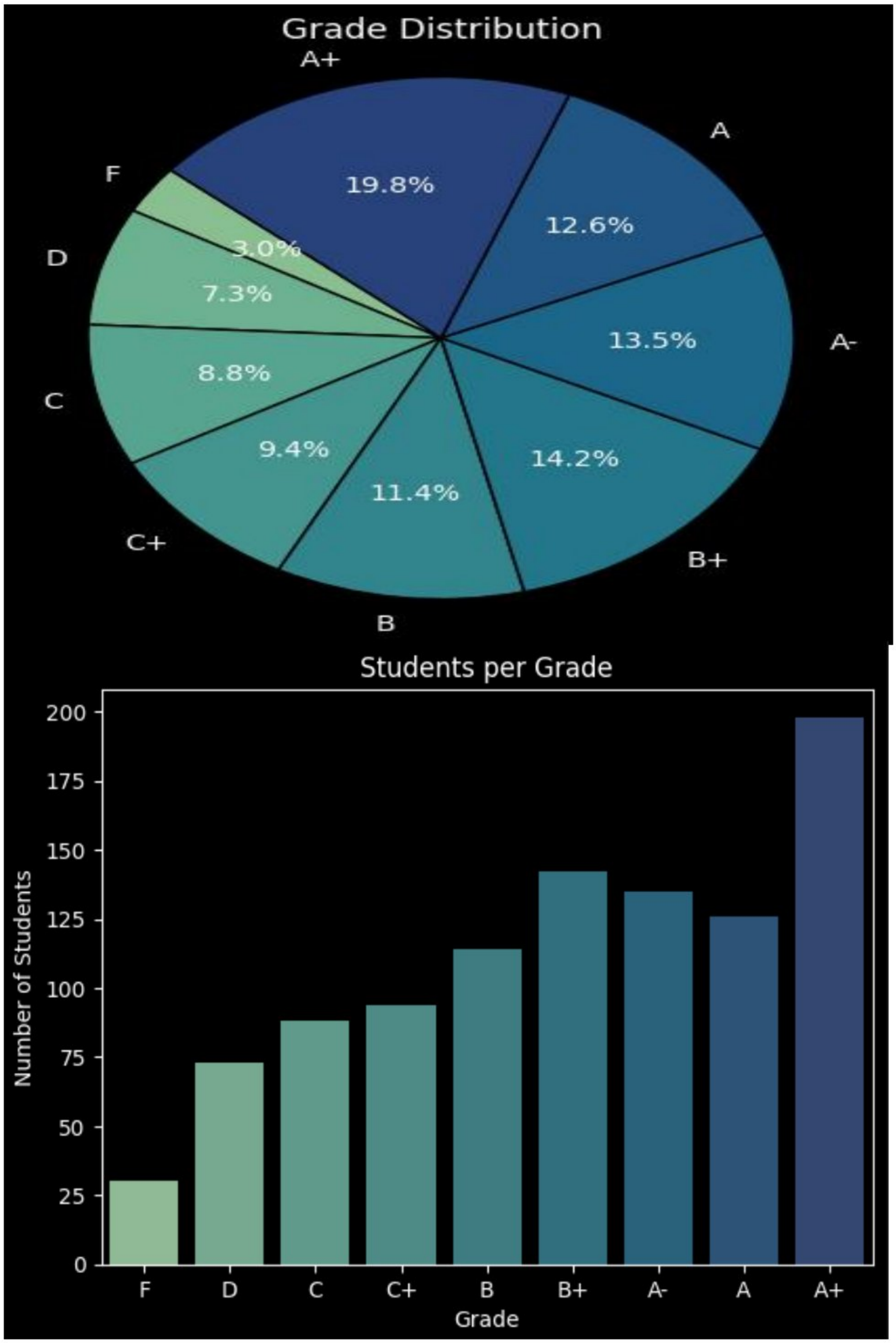
Student Performance Analysis using Machine Learning

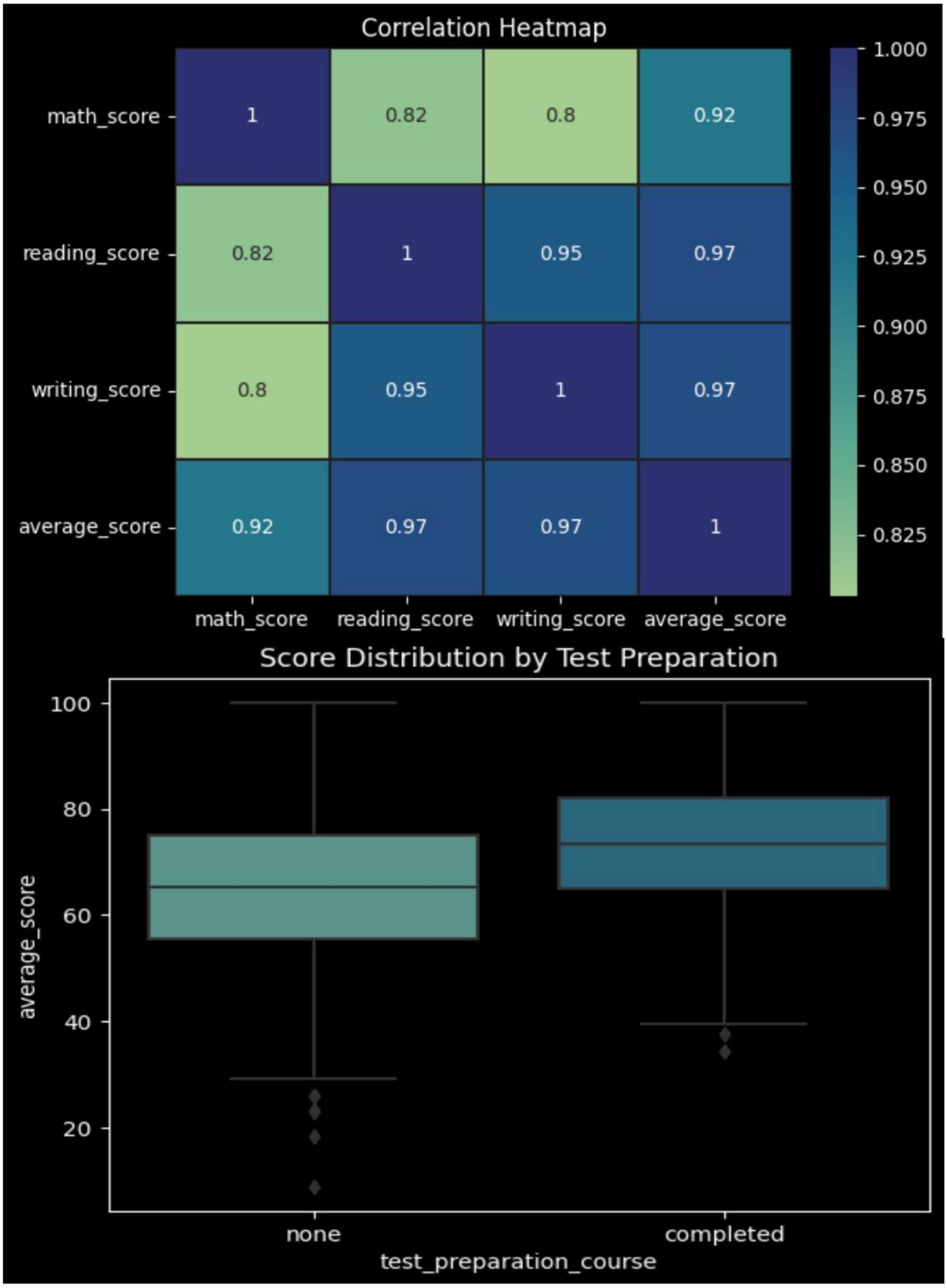
Student Information (1000)

gender	race/ethnicity	parental_level_of_education	lunch	test_preparation_course	math_score	reading_score	writing_sc
female	group B	bachelor's degree	standard	none	72	72	
female	group C	some college	standard	completed	69	90	
female	group B	master's degree	standard	none	90	95	
male	group A	associate's degree	free/reduced	none	47	57	
male	group C	some college	standard	none	76	78	
female	group B	associate's degree	standard	none	71	83	
female	group B	some college	standard	completed	88	95	
male	group B	some college	free/reduced	none	40	43	
male	group D	high school	free/reduced	completed	64	64	
female	group B	high school	free/reduced	none	38	60	

EDA Result

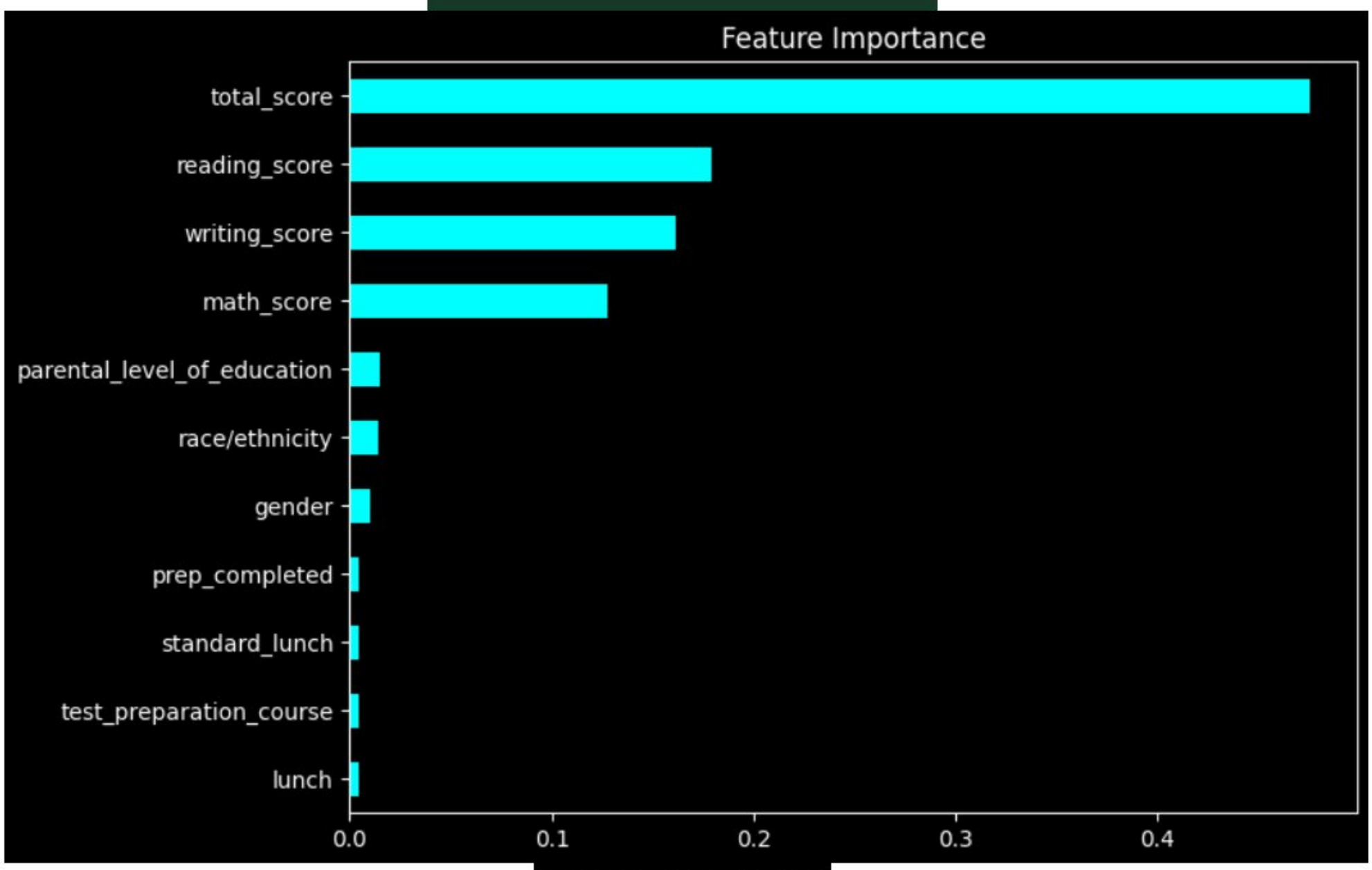



Grade Distribution
A+
19.8%
A
12.6%
A-
13.5%
B+
14.2%
B
11.4%
C+
9.4%
C
8.8%
D
7.3%
F
3.0%
Students per Grade
Number of Students
0
25
50
75
100
125
150
175
200
F
D
C
C+
B
B+
A-
A
A+
Grade


Correlation Heatmap
math_score
reading_score
writing_score
average_score
1
0.82
0.8
0.92
0.82
1
0.95
0.97
0.8
0.95
1
0.97
0.92
0.97
0.97
1
1.000
0.975
0.950
0.925
0.900
0.875
0.850
0.825
math_score
reading_score
writing_score
average_score
Score Distribution by Test Preparation
100
80
60
40
20
average_score
none
completed
test_preparation_course

Model Result

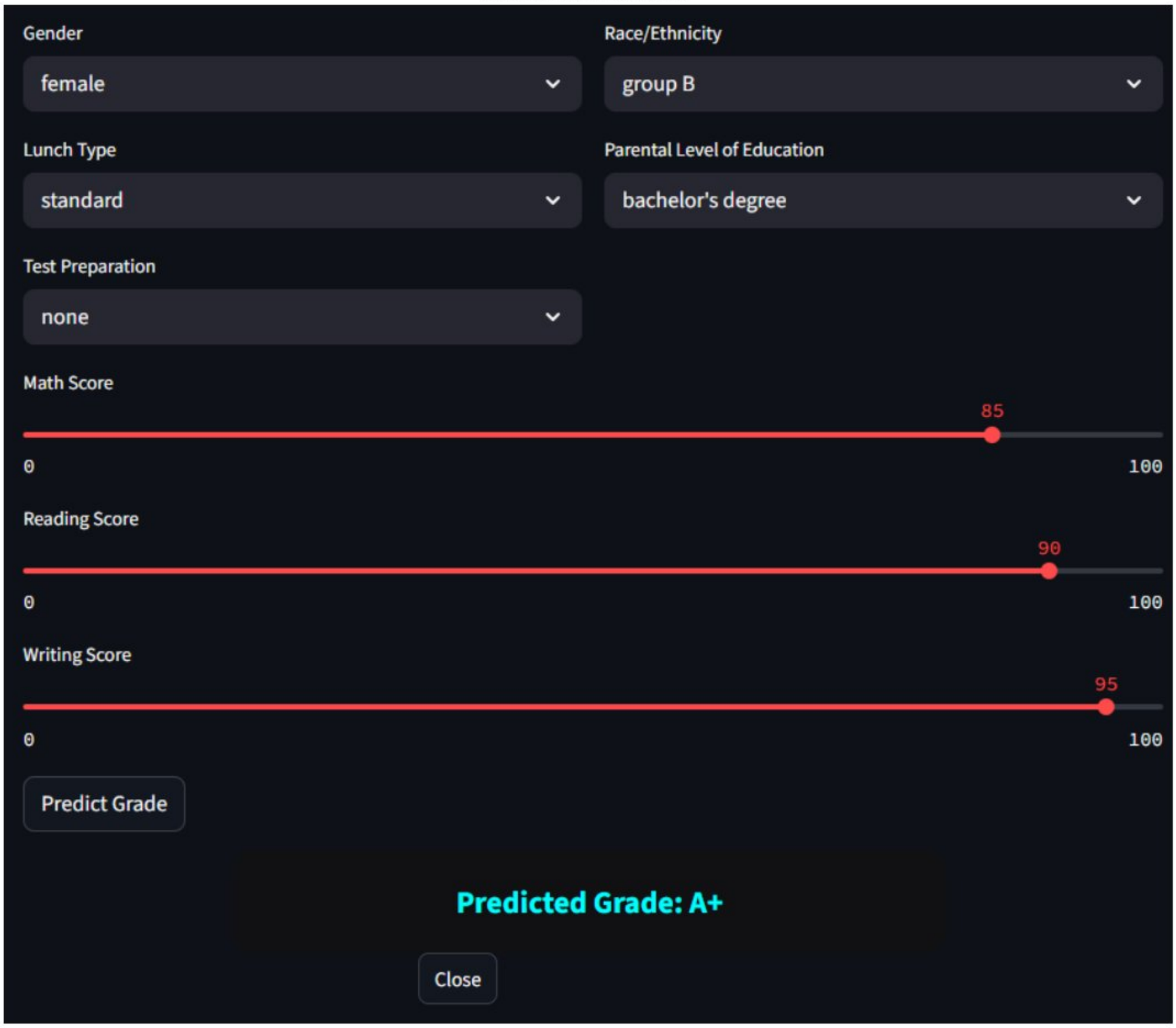
Model Accuracy: **0.9900**



Classification Report

	precision	recall	f1-score	support
A	1	1	1	33
A+	1	1	1	31
A-	1	1	1	26
B	0.96	1	0.98	22
B+	1	0.96	0.98	26
C	0.95	1	0.97	18
C+	1	0.94	0.97	17
D	1	1	1	17
F	1	1	1	10
accuracy	0.99	0.99	0.99	0.99
macro avg	0.99	0.99	0.99	200
weighted avg	0.99	0.99	0.99	200

Prediction Result



15. Conclusion:

The project successfully developed a reliable machine learning model to predict student grades. The analysis revealed that factors such as gender, lunch type, and test preparation have a significant impact on student performance. Additionally, an interactive application was created to allow users to explore the data and predict grades effectively.