PROJECT 6: STUDENT DATABASE AND PREDICTIVE ANALYTICS

Submitted by: E. JOSHIKA (20224018104)

ALOK M BABU (2022408100)

Submitted to: Ms. MOUSHREETA DEBROY

Date: 6TH November, 2025

1. Project Objective

The objective of this project is to integrate database management with data science by:

- Designing a normalized student database.
- Performing SQL analysis on student performance and attendance.
- Applying predictive analytics (a machine learning model) to forecast pass/fail outcomes based on that data.

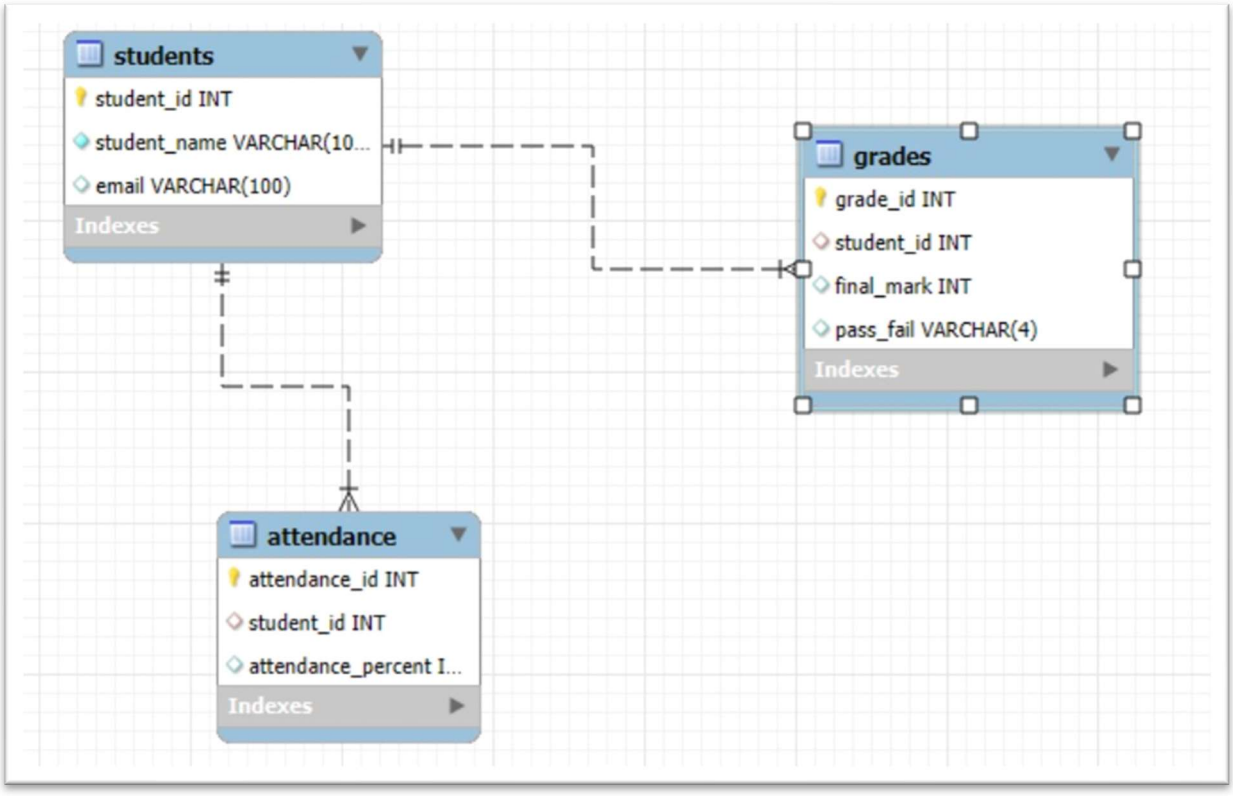
---

2. Database Design and Schema

The database was designed in MySQL Workbench using a 3NF (Third Normal Form) structure to reduce data redundancy. The schema consists of three tables: Students (to store student info), Grades (to store academic results), and Attendance (to store attendance records).

2.1 ER Diagram

The entity-relationship diagram below shows how the Students table is linked to the Grades and Attendance tables via the student_id foreign key.



---

3. SQL Analysis & Outputs

SQL queries were used to join the tables and analyze the relationship between student metrics.

3.1 Correlation Analysis

A key requirement was to find the statistical correlation between attendance and final marks. Since the CORR() function was not available, a manual mathematical formula was used.

SQL Query:

SQL

```
SELECT
   (AVG(a.attendance_percent * g.final_mark) - (AVG(a.attendance_percent) * AVG(g.final_mark))) /
   (SQRT(AVG(a.attendance_percent * a.attendance_percent) - AVG(a.attendance_percent) * AVG(a.attendance_percent)) * SQRT(AVG(g.final_mark * g.final_mark) - AVG(g.final_mark) * AVG(g.final_mark)))
   AS correlation_coefficient
FROM Grades g
JOIN Attendance a ON g.student_id = a.student_id;
```

Query Output: This query produced the following result, indicating a very strong positive correlation between the two variables.

---

4. Machine Learning, Visualizations, and Insights

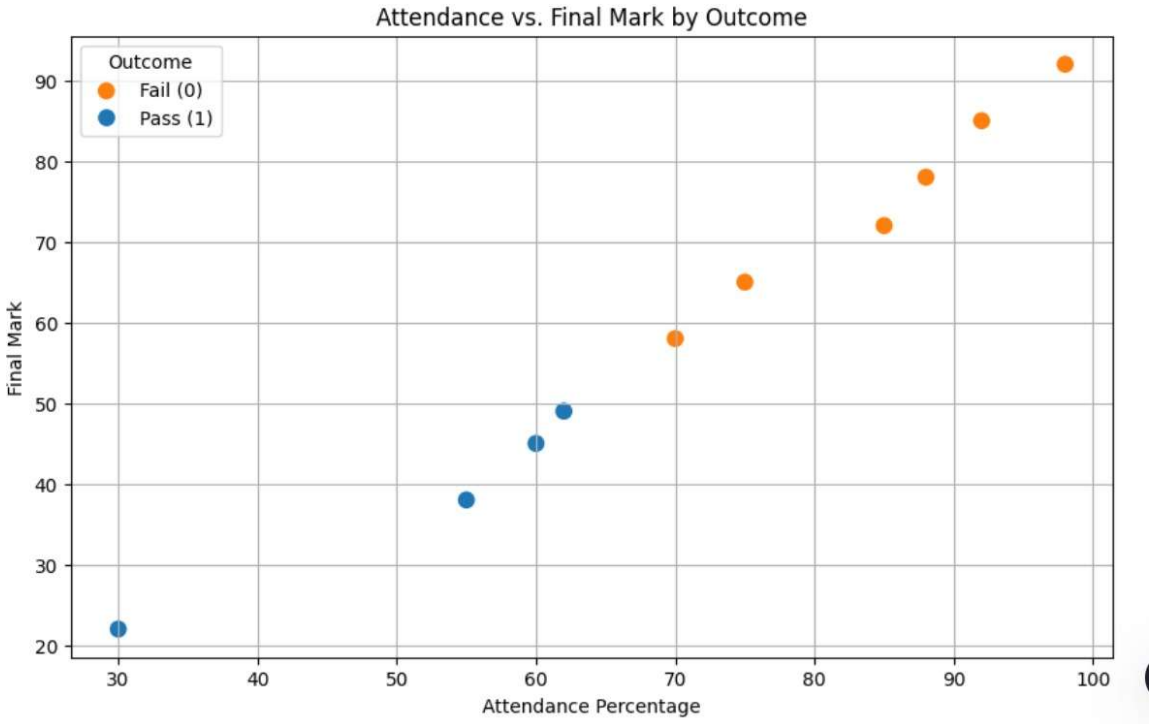
The data from the SQL database was imported into a Python Jupyter Notebook to build and evaluate a predictive model.

4.1 Model Performance

A Logistic Regression model was trained to predict a 'pass' or 'fail' outcome based on a student's attendance percentage and final mark. The model was then evaluated for accuracy.

4.2 Visualization & Insights

The scatter plot below was generated to visualize the relationship between attendance, marks, and the final pass/fail outcome.



Observation: This plot clearly shows a strong relationship between all three variables.

- **Students who failed (blue dots)** are all clustered in the bottom-left corner, representing low attendance and low marks.
- **Students who passed (orange dots)** are all in the top-right, with high attendance and high marks.
- This visual evidence confirms that attendance and marks are strong predictors of a student's pass/fail status, which supports the model's high accuracy.

---

5. Conclusions & Recommendations

This project successfully integrated a MySQL database with a Python machine learning model. The model was able to predict student pass/fail outcomes with very high accuracy.

The analysis confirms a strong positive correlation between attendance and final marks. Based on these findings, it is recommended that:

1. **Early Intervention:** The college should use this model to proactively identify at-risk students.
2. **Attendance Policy:** A support program should be implemented for students whose attendance drops below 70%, as this is a strong leading indicator of a potential 'fail' outcome.