Project Report: Data Mining

1.Case Study Context:

Student Performance and Academic Decision Support

Educational institutions generate large volumes of data across departments, admissions, and academic performance. However, this data is often fragmented across systems, inconsistently maintained, and contains quality issues such as missing values, duplicates, and invalid records. These problems limit the institution's ability to make reliable, data-driven decisions.

This project presents an end to end analytics pipeline built on a student information dataset. The objective is to transform raw relational data into reliable insights by applying structured ETL processes, data quality validation, descriptive analytics, and interpretable predictive modeling.

The emphasis of this work is on correctness, transparency, and reproducibility rather than purely on model complexity.

This case study simulates a **university level student analytics system**, where administrators and academic planners are interested in answering questions such as:

- Are student records and departmental data internally consistent?
- How does student effort translate into academic performance?
- Can historical performance be used to reasonably predict future outcomes?
- What data quality issues could bias reporting or predictive models if left unresolved?

The dataset represents four core operational domains:

- Departments (organizational structure and history)

- Employees (staff linked to departments)
- Student counseling and admissions (student choices and placements)
- Student academic performance (marks and effort hours per course)

In a practical setting, such data would support:

- Academic advising and student support initiatives
- Departmental performance comparisons
- Early warning systems for underperforming students
- Policy decisions related to workload, curriculum, and resource allocation

However, before any meaningful analytics can be performed, the data must be validated, cleaned, and audited. Predictive models trained on flawed data can lead to incorrect conclusions, unfair evaluations, or poor institutional decisions. Key work areas addressed include:

- Duplicate identifiers that can break joins and inflate counts
- Invalid dates and impossible values that distort trends
- Missing admissions data that undermines departmental analysis
- Out of range marks and effort values that bias models

By explicitly identifying, reporting, and separating invalid records, the pipeline ensures that:

- Analytical results are defensible
- Decisions are based on credible evidence
- Stakeholders can trace every exclusion back to a documented reason

Once a clean and trustworthy dataset was established, the case study focused on two analytical goals:

1. Descriptive Understanding

To explore overall trends in student effort and academic outcomes, identify typical performance ranges, and validate assumptions about the effort and performance relationship.

2. Predictive Insight

To evaluate whether a simple, interpretable model (linear regression) can reasonably estimate future student performance based on historical effort, supporting decision making scenarios such as:

- Academic advising (“How much effort is needed to reach a target grade?”)
- Performance forecasting for upcoming assessments
- Identifying diminishing returns in study effort

The emphasis was placed on interpretability and transparency, reflecting how analytics is commonly deployed in educational and institutional environments.

2. Database Design and Implementation

The database design followed the given ER diagram and used four tables. Each table was connected through primary and foreign keys. The schema was created using SQL scripts and validated for referential integrity.

Table	**Primary Key**	**Foreign Key(s)**	**Description**
Department_Information	Department_ID	-	Department details, DOE checked for valid years
Employee_Information	Employee_ID	Department_ID	Employee data linked to department
Student_Counceling_Information	Student_ID, Department_Choices, DOA	Department_Choices, Department_Admission	Links students to chosen and admitted departments
Student_Performance_Data	Student_ID, Paper_ID, Semster_Name	Student_ID	Stores marks and effort hours

3. ETL Process and Data Cleansing

Data from all four entities was loaded into MariaDB, validated, and exported for analysis. SQL joins were used to combine related data, and invalid or missing entries were removed according to the project's quality criteria.

Table	Attribute	Issue	Action Taken
Department_Information	Department_ID	10 duplicates	Reported
Department_Information	DOE	3 values before 1900	Reported
Student_Counceling_Information	Department_Admission	1 missing, 1 invalid	Reported
Student_Performance_Data	Marks	5 out of range (>100)	Removed
Student_Performance_Data	Effort_Hours	1 negative value	Removed
Student_Performance_Data	Missing values	3 records	Removed

4. Descriptive Analytics

Descriptive statistics and visualizations were generated in Jupyter Notebook using Pandas, NumPy, Matplotlib, and Seaborn.

These results show the general trends and relationships between effort and marks:

Metric	Mean	Median	Std Dev
Marks	76.4	78	9.2
Effort Hours	8.7	9	2.1

Correlation between Marks and Effort: 0.72 (positive relationship).

Most students who studied for 8 to 10 hours scored above 75%.

There is an outlier of 270 hours in the Effort Hours attribute, which has been excluded from visualization.

5. Predictive Model: Linear Regression

A simple linear regression model was built using Scikit-learn to predict student marks based on effort hours. Each student's past performance was used to train the model, and predictions were made for 10 hours of effort.

Student ID	**Equation**	**R^2**	**Predicted (10h)**
SID20131151	6.3500x + 18.4154	0.9449	81.92
SID20149500	6.5647x + 16.6069	≈0.93	82.25
SID20182516	6.1540x + 20.6298	0.9323	82.17

Interpretation: The models show that each additional hour of effort increases marks by about six points. R^2 values above 0.9 indicate a very strong fit.

6. Key Insights

- Data quality validation identified **10 duplicate department identifiers**, **5 invalid mark values**, **3 missing performance records**, and **1 negative effort entry**. Removing or flagging these records prevented invalid data from influencing summary statistics and predictive models.

- After cleaning, the final analysis dataset showed a **mean mark of approximately 76** with most values concentrated between **70 and 85**, indicating a reasonably consistent academic performance range once outliers were removed.

- Effort hours demonstrated a strong positive association with marks. Students typically studying between **8 and 10 hours** achieved scores above **75 percent**, while lower effort levels were associated with noticeably lower marks.

- The relationship between effort hours and marks produced a **correlation of approximately 0.72**, confirming effort as a strong explanatory variable in the

dataset.

- Consistent ETL steps ensured reliable integration of data across four relational tables. Validating department admissions against department identifiers prevented incorrect joins and misleading departmental analysis.

- Linear regression models trained per student produced **R squared values above 0.9** for all target students, indicating an excellent model fit and strong explanatory power.

- The regression coefficients showed that **each additional hour of effort corresponded to an increase of roughly 6 marks**, providing a clear and interpretable quantitative insight.

- Exporting exception reports and cleaned datasets as CSV files improved transparency and made the analytical process reproducible and auditable.

7. Tools and Environment

Category	**Tool**
Database	MariaDB, SQLite
Programming Language	Python 3
Libraries	Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
IDE	Jupyter Notebook
OS	Windows 11

8. Conclusion

- This project successfully integrated database design, ETL processing, data quality validation, descriptive analytics, and predictive modeling into a single structured workflow.

- A total of **11 exception reports** were generated, documenting issues such as duplicate identifiers, invalid dates, missing values, and out of range performance

metrics. These reports provided a clear record of data quality concerns prior to analysis.

- Cleaning the performance data reduced noise and ensured that statistical summaries and models were based only on valid and complete records.

- Descriptive analysis confirmed that academic effort is strongly linked to performance, with higher effort consistently resulting in higher marks across the cleaned dataset.

- Predictive modeling using linear regression produced stable and interpretable predictions. For a fixed effort level of **10 hours**, predicted marks clustered around **82 percent** across all evaluated students.

- The project demonstrates that accurate analytics depend heavily on clean data and structured processing rather than complex modeling techniques.

- Overall, the workflow shows how disciplined data preparation and simple, well justified models can produce reliable insights from relational datasets.