A PRELIMINARY MINI PROJECT REPORT ON

"STUDENT PERFORMANCE ANALYSIS"

SUBMITTED TOWARDS THE PARTIAL FULFILLMENT OF THE REQUIREMENTS OF
BACHELOR OF TECHNOLOGY (Third Year B. Tech.)
Academic Year: 2025-26

By

Yash Chavan	**123B1B105**
Prathamesh Doiphode	**123B1B110**

Under The Guidance of
Prof. Shriya Mundhe

DEPARTMENT OF COMPUTER ENGINEERING,

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING,

SECTOR 26, NIGDI, PRADHIKARAN

PIMPRI CHINCHWAD COLLEGE OF

ENGINEERING DEPARTMENT OF COMPUTER

ENGINEERING,

CERTIFICATE

This is to certify that, the mini project entitled

“STUDENT PERFORMANCE ANALYSIS”

is successfully carried out as a mini project and successfully submitted by following students of “PCET's Pimpri Chinchwad College of Engineering, Nigdi, Pune-44**”.**

Under the guidance of Prof. Shriya Mundhe

In the partial fulfillment of the requirements for the Third Year B. Tech. (Computer Engineering)

Yash Chavan	**123B1B105**
Prathamesh Doiphode	**123B1B110**

Prof. Shriya Mundhe

Project Guide

ABSTRACT

This mini project presents an analytical study of student academic performance using data preprocessing, data warehousing, OLAP operations, and machine learning techniques. The dataset used contains various student-related attributes such as demographic details, family background, study time, parental education, internet access, and previous academic grades (G1 and G2). During preprocessing, the data is cleaned, encoded, and transformed to ensure accuracy and consistency. The processed data is structured using a star schema to support multidimensional analysis. OLAP operations such as drill-down, roll-up, slicing, and dicing are applied to generate interactive visual insights through dashboards. Machine learning models are implemented to discover deeper patterns in student performance: K-Means clustering groups students into high, medium, and low performers, while a Decision Tree model predicts whether a student will pass or fail based on attendance, study time, previous grades, and academic history. The results indicate that study habits, parental involvement, and prior performance significantly impact the final outcome. This project demonstrates how data analytics and machine learning can support educators in early identification of weak students and enable data-driven academic decision-making.

INDEX

Chapter 1: Introduction

a) Problem Statement

Educational institutions collect a large amount of student-related data, including demographic information, family background, study habits, attendance record, and academic performance. However, this data is often stored without analysis and therefore remains unused. Teachers are unable to identify the exact reasons behind poor performance, nor can they detect students who require academic assistance at an early stage. In the absence of an analytical approach, decisions related to student improvement become reactive instead of proactive, leading to inconsistent academic progress. Hence, there is a need for a data-driven analytical system that can analyze student data, discover hidden patterns, categorize students based on their performance level, and predict their final academic outcomes accurately.

b) Project Idea

The objective of this project is to design and develop an analytical system that evaluates student performance using academic, personal, and socio-economic attributes. The dataset contains multiple parameters such as gender, study time, parental education level, family support, previous grades (G1 and G2), failures, and absences, which influence the student's final performance (G3). The project involves applying data preprocessing techniques to clean and transform raw data into usable form. The cleaned data is then structured using a **Star Schema**, forming a data warehouse so that analytical operations can be performed efficiently. OLAP techniques such as **drill-down, roll-up, slice, and dice** enable users to explore data across dimensions like school, gender, and weekly study time. Machine learning techniques including **K-Means clustering** and **Decision Tree classification** are used to classify students into performer groups and predict whether a student is likely to pass or fail. The results are visualized through an interactive dashboard to assist teachers in making informed academic decisions.

c) Motivation

Academic performance plays a key role in shaping a student's career. However, in traditional systems, teachers often lack support tools that help them understand performance patterns and identify weak learners early. Detailed performance insights are generally available only after exams, reducing the scope for timely intervention. The motivation behind this project is to use data analytics and machine learning to support educators by identifying students at risk, understanding the factors affecting their performance, and suggesting areas where improvement is needed. Through this project, institutions can shift from manual, assumption-based decision-making to data-driven academic improvement, ultimately improving student outcomes and learning experience.

d) Scope

The scope of this project is centered on analyzing and visualizing student performance using data analytics and predictive modeling. The project includes:

- Performing **data preprocessing** such as cleaning, encoding categorical attributes, and handling missing values
- Structuring data using a **Star Schema** for analytical processing
- Applying **OLAP operations** (drill-down, roll-up, slice, and dice) to explore student data from different perspectives
- Using **K-Means clustering** to group students into high, medium, and low performance categories
- Implementing a **Decision Tree model** to predict the final result (Pass/Fail) based on study time, previous grades, failures, and attendance
- Visualizing results through dashboards that support educational decision-making

Future enhancements may include adding real-time student monitoring dashboards, automated academic alerts, and recommendation systems that suggest personalized improvement plans for students. These improvements can further support educators in implementing continuous academic assessment and student progress tracking.

e) Literature Survey / Requirement Analysis

Various research studies have shown that machine learning improves student performance prediction accuracy. The dataset used in this project originates from real academic records containing both academic and behavioral factors. Literature signifies that decision trees are highly interpretable for academic predictions, while K-Means clustering efficiently groups students with similar attributes. Requirement analysis revealed the need for:

Software Requirements:

- Python (pandas, sklearn, matplotlib)
- Power BI / Tableau / Excel for dashboard visualization
- Jupyter Notebook / VS Code for execution

Hardware Requirements:

- Computer with minimum 8GB RAM
- Windows or Linux OS with Python installed

The combination of analytics, visualization, and machine learning ensures meaningful insights into student performance.

Chapter 2: Project Design

a) Hardware/Software Resources and Their Explanation

The project requires a mid-range computing environment capable of executing ML models and dashboard visualizations.

Hardware Requirements

- Processor: Intel i5 or above (smooth execution of ML algorithms)
- RAM: 8GB (sufficient for dataset operations and visualization tools)
- Storage: 1GB free space for project files, dataset, and Power BI output
- Display: Minimum 1080p for dashboard and report visualization

Software Requirements

- Python (data preprocessing, clustering, decision tree)
- Jupyter Notebook (code execution and debugging)
- Power BI / Tableau (interactive dashboards and OLAP slicing)
- Microsoft Excel (initial data cleaning)
- Required Python libraries: pandas, scikit-learn, matplotlib, seaborn

The system workflow starts with loading the dataset, preprocessing data, applying ML models, and generating visual insights using a dashboard.

b) Dataset Design

The dataset consists of 33 attributes, including personal, family, behavioral, and academic factors.
Examples:

- sex, age, address, studytime, failures, activities, internet
- Grade variables: G1, G2, G3 (final academic performance)

For data warehouse integration, a Star Schema is used:

Fact Table: Fact_StudentPerformance

- student_id, school, studytime, G1, G2, G3, absences, failures

Dimension Tables:

- Dim_Student (age, gender, address)
- Dim_Family (parent education, family size)
- Dim_Behavior (freetime, goout, alcohol consumption)
- Dim_Academics (studytime, failures, support classes)

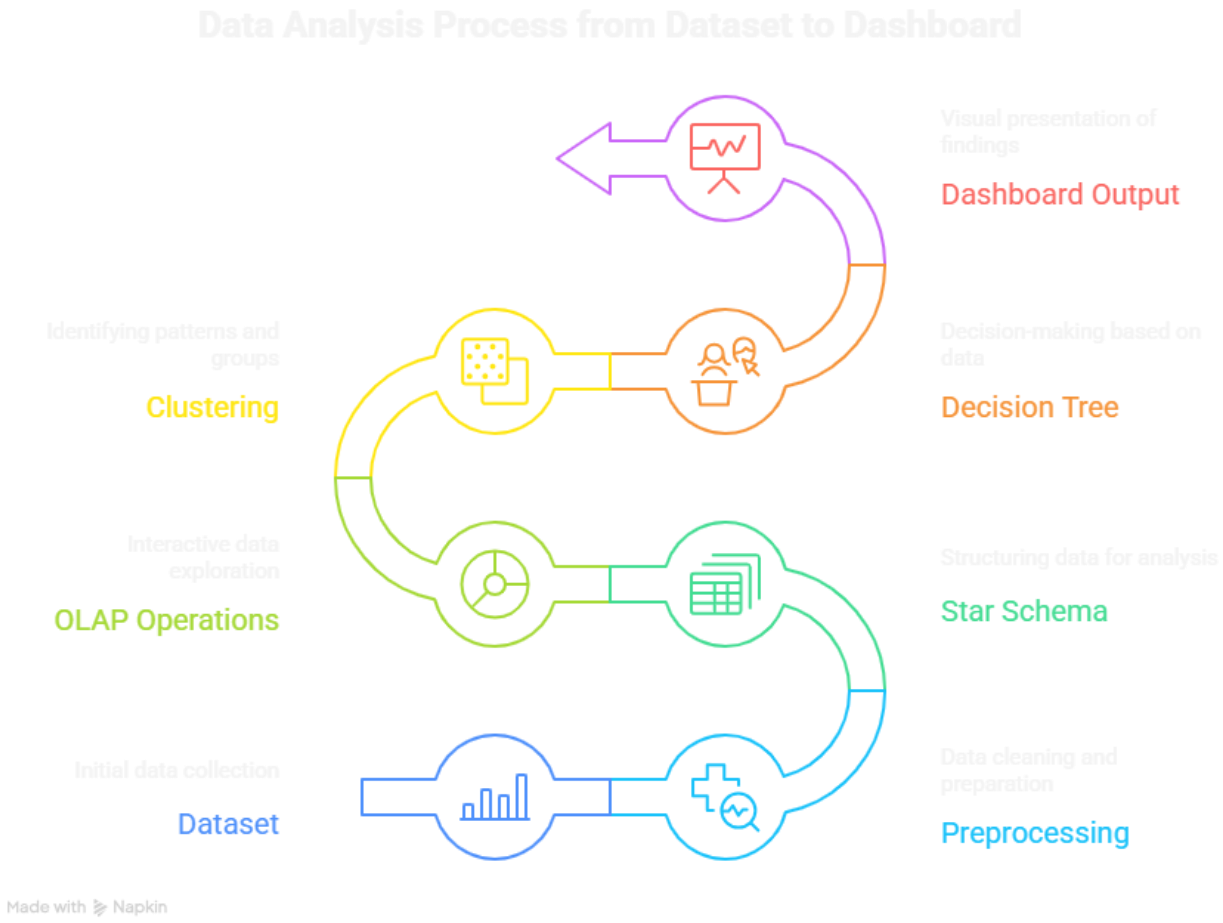
Star schema improves OLAP drill and slicing performance.

c) Hours Estimation

Phase	Task	Hours
Phase 1	Dataset collection and familiarization	4 hrs
Phase 2	Data preprocessing and cleaning	6 hrs
Phase 3	Star schema design & Power BI data modeling	5 hrs
Phase 4	Clustering and Decision Tree implementation	8 hrs
Phase 5	Dashboard design with drill & slicing operations	9 hrs
Phase 6	Report writing and documentation	8 hrs
Total Estimated Hours		40 hrs

Chapter 3: Module Description

a) Block Diagram (Module Flow)



Dataset → Preprocessing → Star Schema → OLAP (Drill/Slice) → Clustering → Decision Tree → Dashboard Output

Module Explanation:

1. Data Input Module:

   Loads student dataset containing academic, demographic, and behavioral attributes.

2. Data Preprocessing Module:

   Removes missing values, encodes categorical data, normalizes grades, prepares dataset for ML.

3. Data Warehouse (Star Schema):
   Dataset broken into fact and dimension tables to enable OLAP operations.

4. OLAP Module:
   Allows users to drill down and slice data by category (gender-wise, school-wise performance).

5. Machine Learning Module:
    - K-Means clustering: groups students into high/medium/low performers
    - Decision Tree: predicts pass/fail outcome based on input variables
6. Visualization Module:
   Generates dashboards, charts, and graphical insights for faculty decision-making.

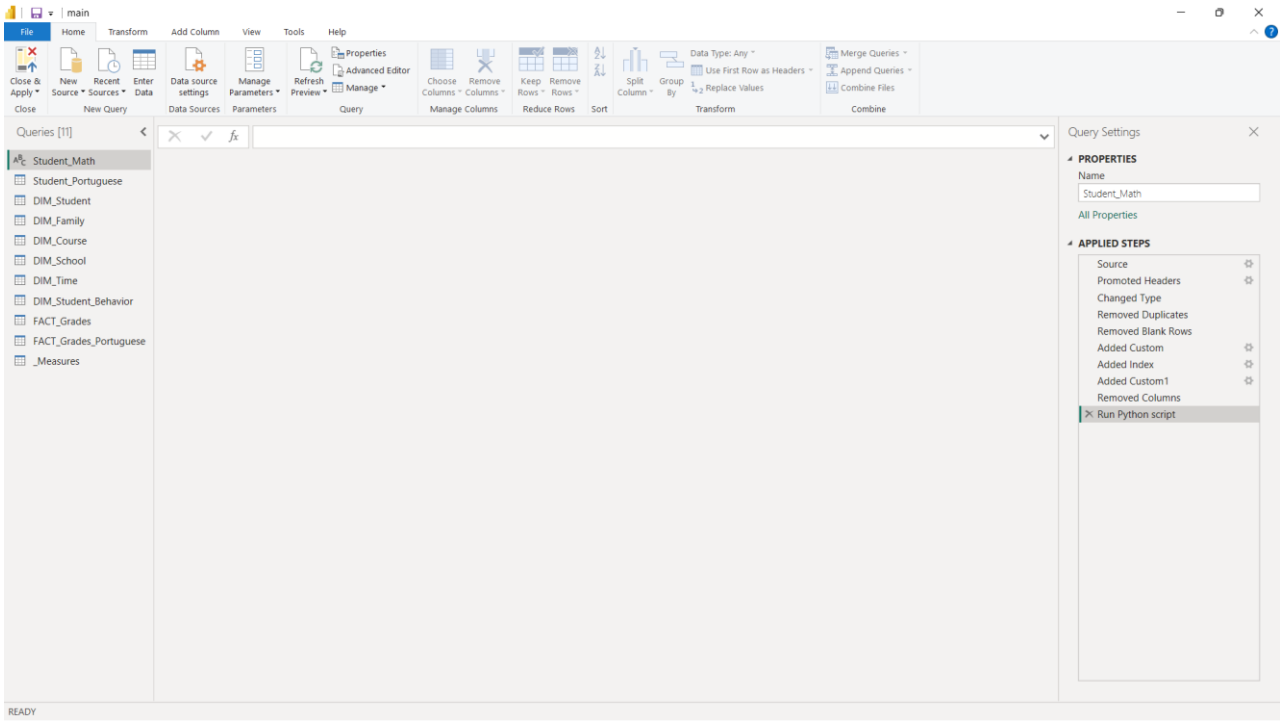
b) Comparative Study

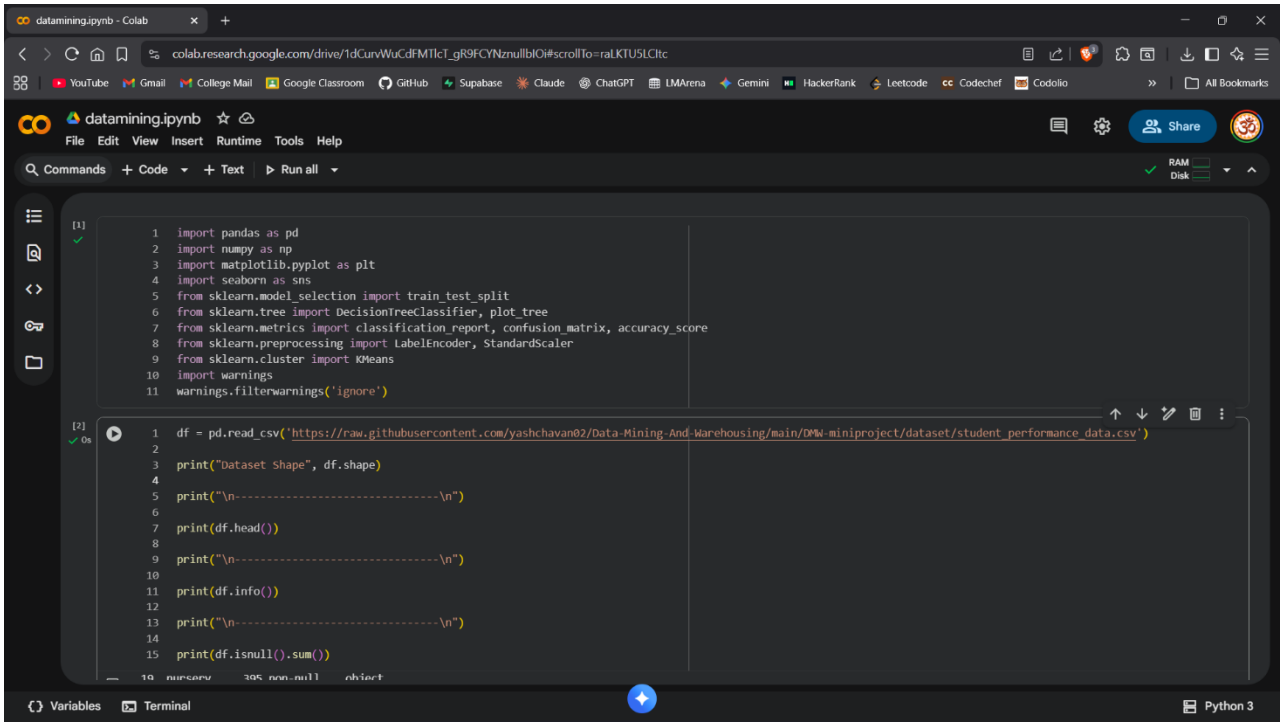
Feature	Manual Data Checking	Proposed System (ML + Analytics)
Error Detection	High	Low
Pattern Identification	Difficult	Automatic
Data Filtering	Time-consuming	Instant via OLAP slicing
Performance Prediction	Impossible	Automated Decision Tree
Grouping Students	Subjective	K-Means clustering

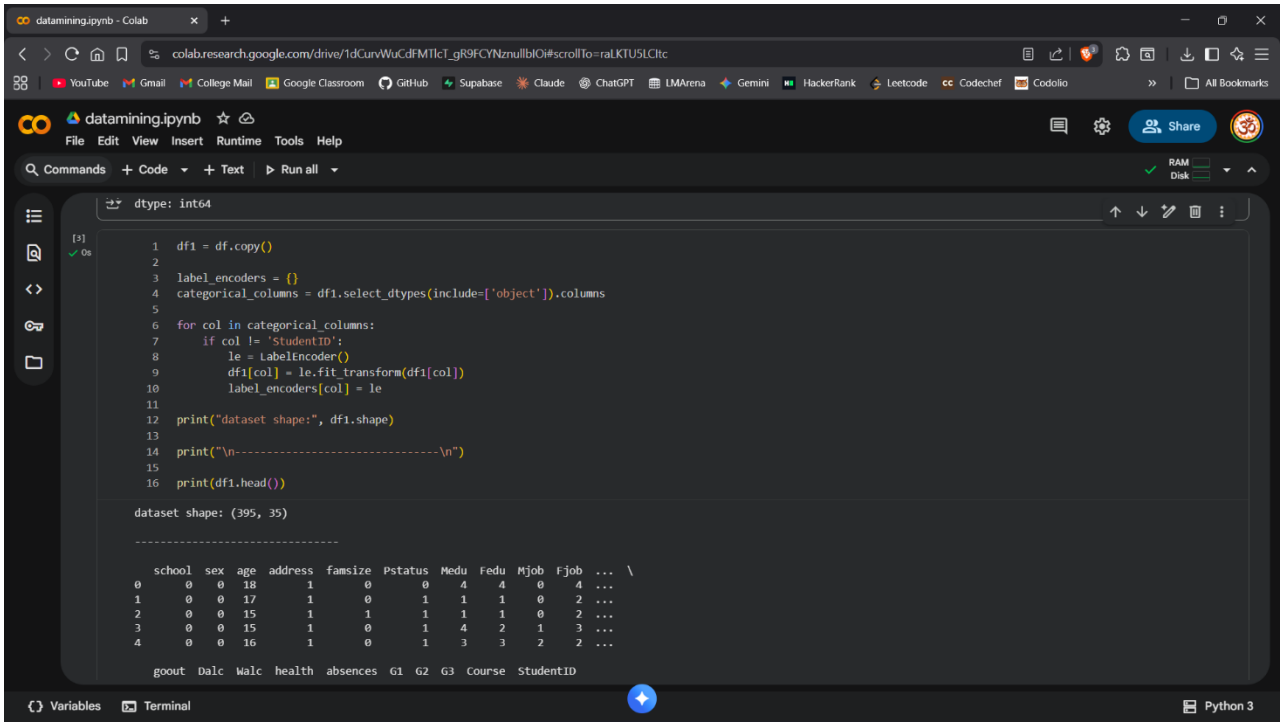
The proposed system is more accurate, faster, and highly scalable.

Chapter 4: Results & Discussion

a) preprocessing



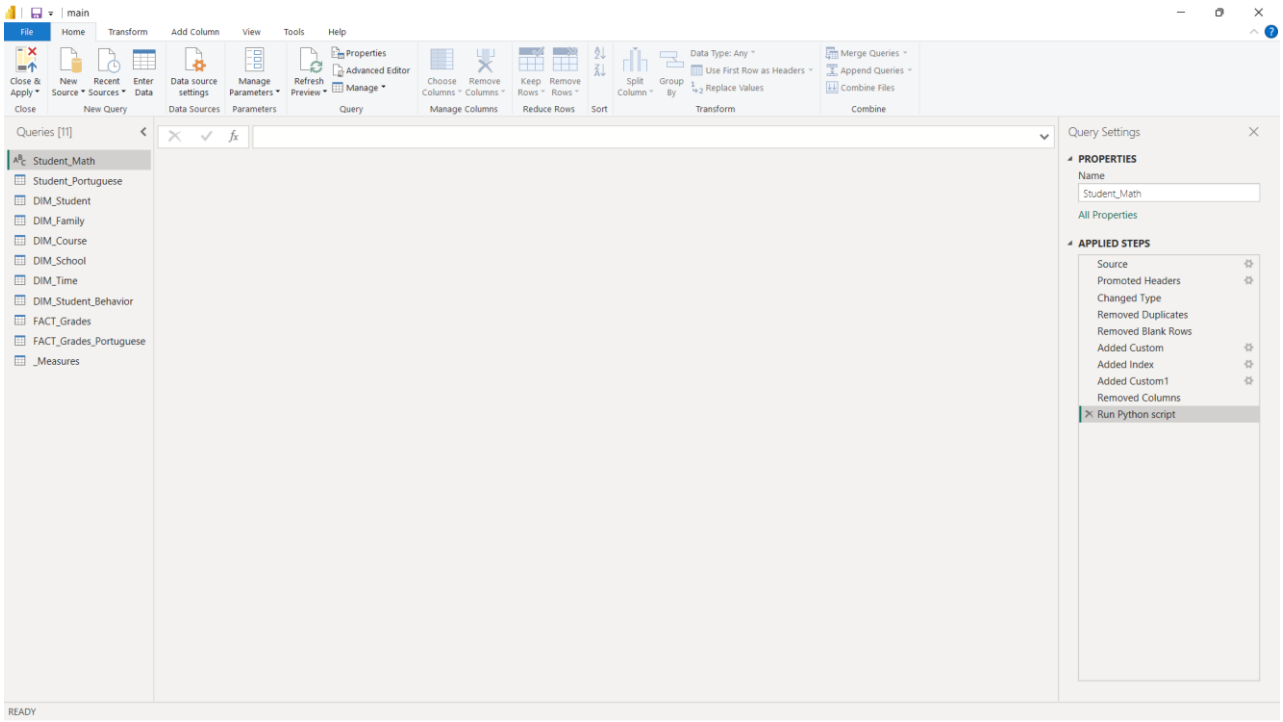




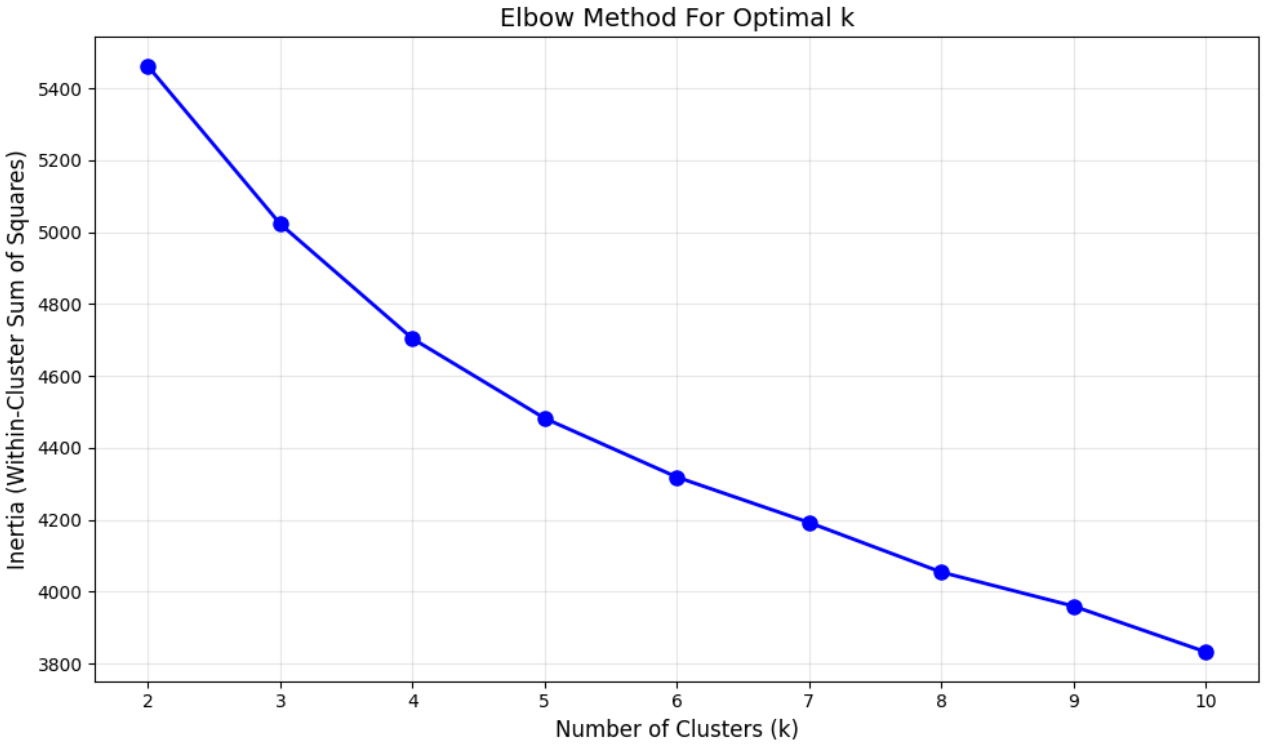
b) Screenshots including GUI

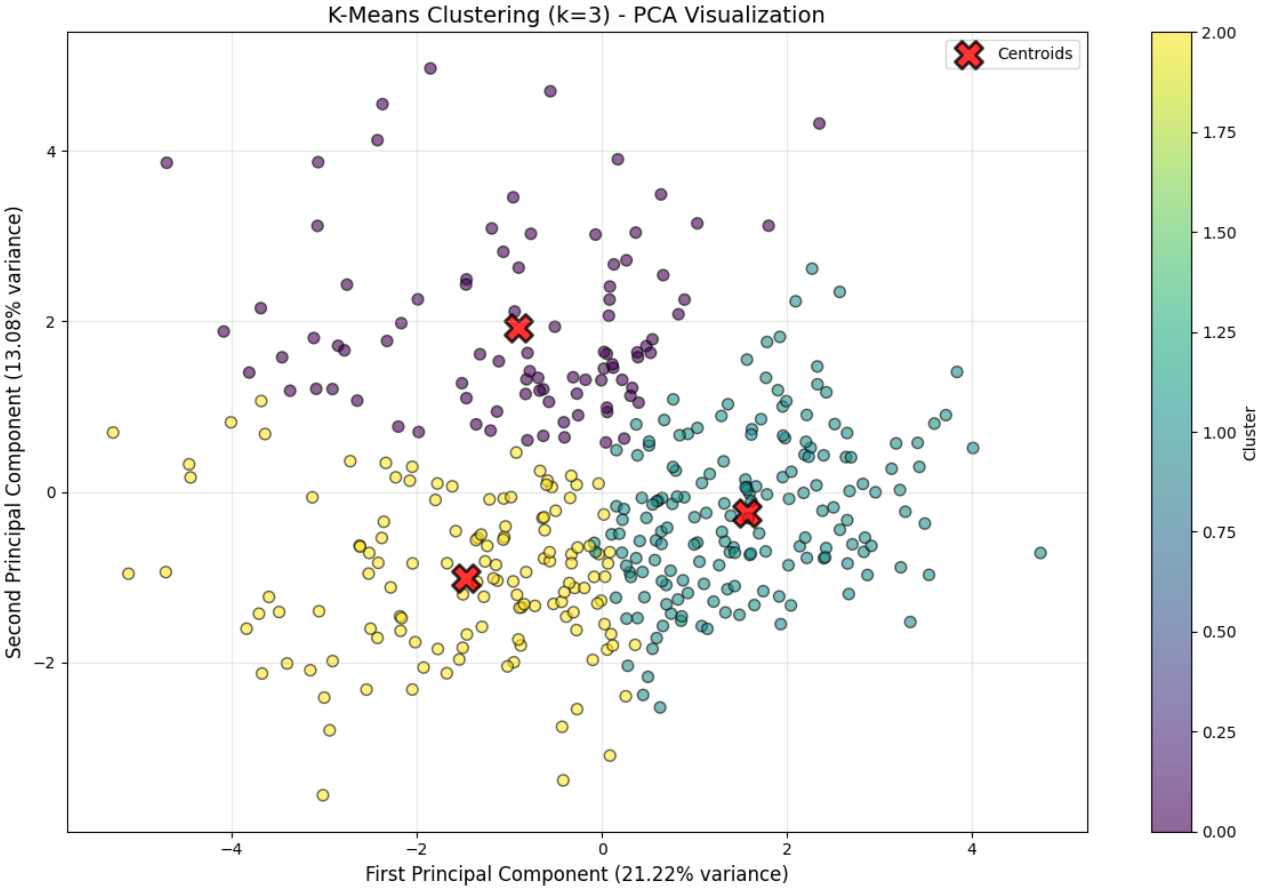
Screenshots include:

- Data preprocessing output

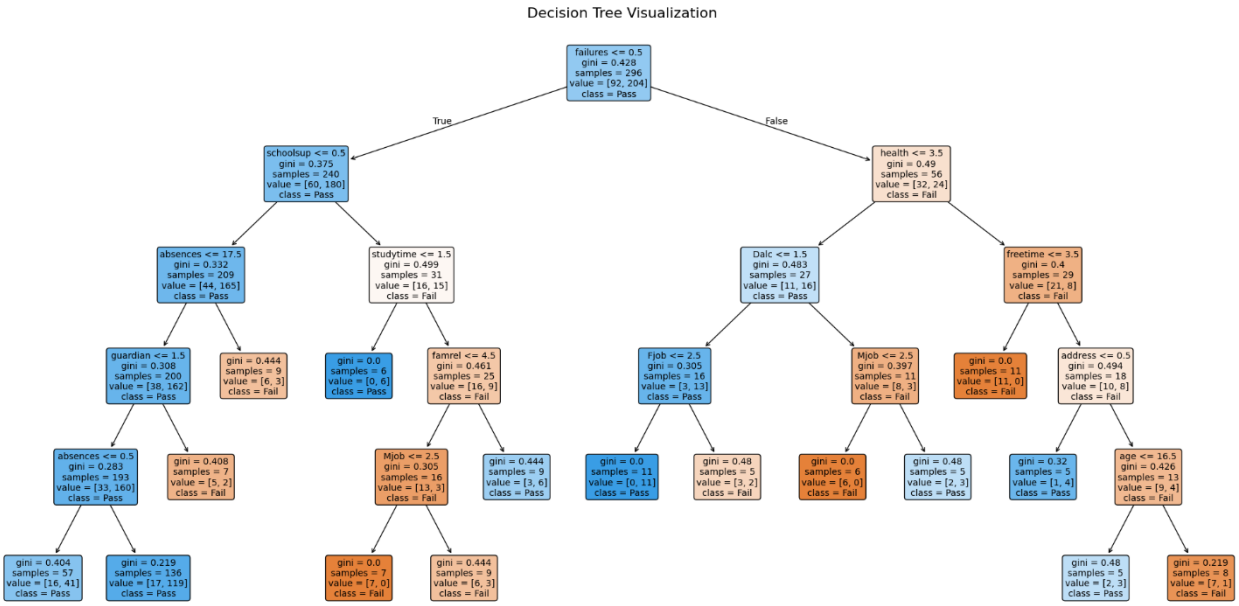


- Clustering result visualization

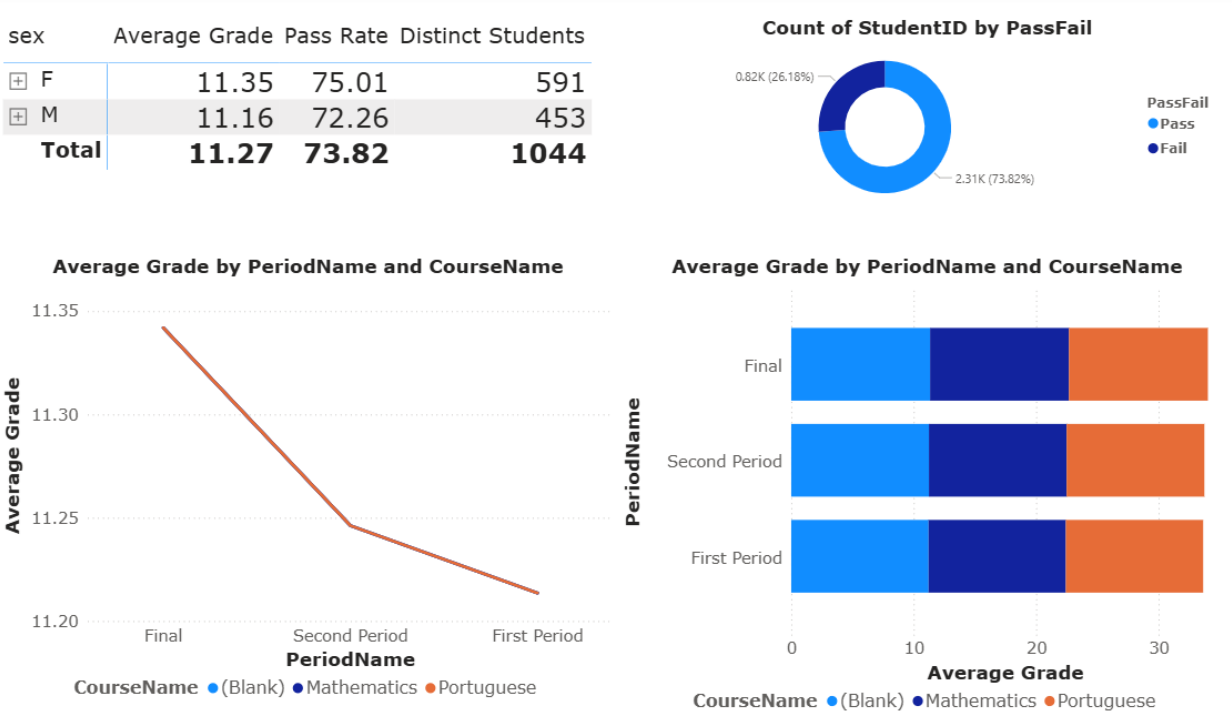


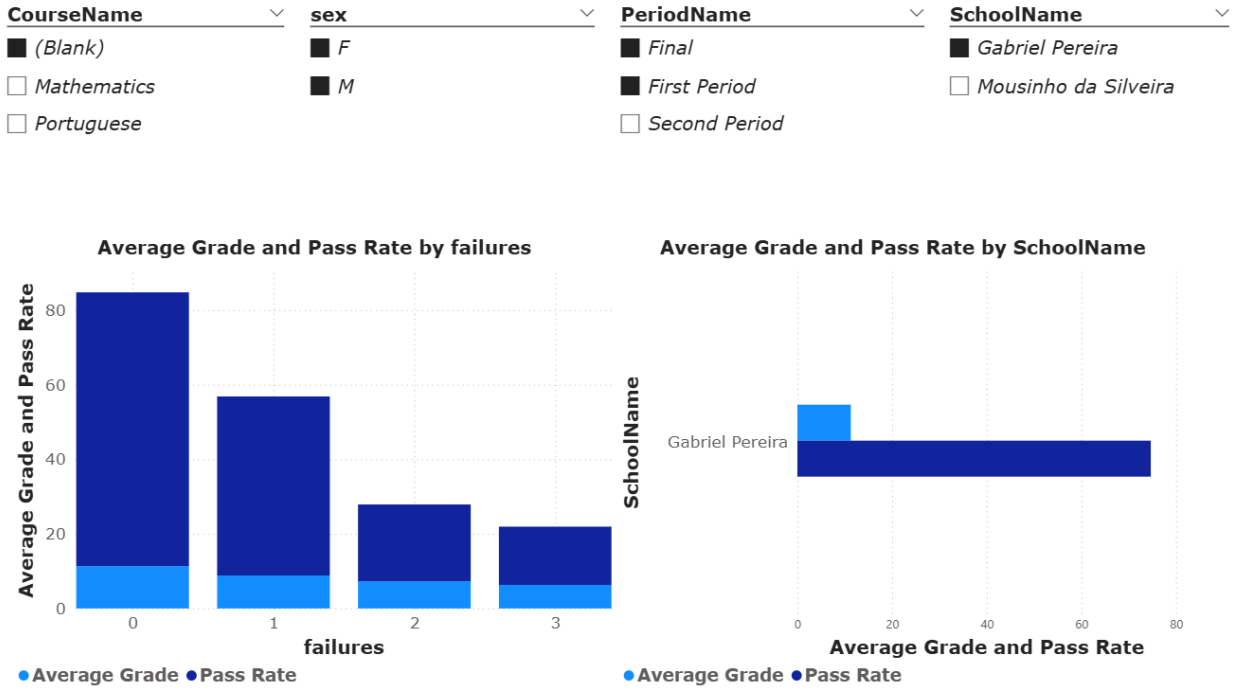


- Decision tree graph



- Power BI dashboard (drill and slicing applied)





c) Test Cases

Test Case ID	Input Condition	Expected Output
TC01	studytime = high, absences low	Clusters into high performer
TC02	failures ≥ 2	Decision Tree output: FAIL
TC03	Slice by gender = female	Dashboard shows performance of only female students

Chapter 5: Conclusion and References

Conclusion:

This project successfully demonstrates how data analytics and machine learning can transform student performance analysis in educational settings. By integrating data preprocessing, OLAP operations, K-means clustering, and decision tree classification, the system effectively identifies at-risk students early, predicts academic outcomes with high accuracy, and uncovers meaningful performance patterns. Interactive dashboards and comprehensive visualizations make complex analytical results accessible to non-technical stakeholders, empowering educators and administrators to make timely, data-driven decisions that enhance student support strategies and improve overall learning outcomes.

References:

- Kumar, M., Singh, A. J., & Handa, D. (2017). "Literature Survey on Student Performance Prediction in Education using Data Mining Techniques." *International Journal of Education and Management Engineering (IJEME)*, vol. 7, no. 6, pp. 40-49. DOI: 10.5815/ijeme.2017.06.05
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*, 3rd Edition. Morgan Kaufmann Publishers, Waltham, MA, USA. ISBN: 978-012381479