Project Report

Student Performance Analytics and Forecasting Platform
Preferred Internship Program – Data Analytics
Team – 15

Internship Project Report

Project Title: Student Performance Analytics and Forecasting – DataVista Pro
Internship Domain: Data Analytics, Machine Learning, Business Intelligence

---

Project Overview

DataVista Pro is an enterprise-grade, real-time data analytics platform built to analyze and forecast student academic performance using a variety of influential factors. The project combines **data engineering, AI/ML, and visual analytics** to assist educational institutions in identifying key drivers of student success, predicting outcomes, and improving learning strategies.

Objectives

- To analyze key factors influencing student academic performance.
- To develop real-time and batch data pipelines for ingestion, processing, and analytics.
- To apply machine learning for performance prediction and NLP for sentiment analysis of feedback.
- To generate automated dashboards and reports for administrators and educators.

Tools & Technologies Used

Category	**Technologies**
Data Ingestion	Apache Kafka, Pandas, CSV Loader
Processing & ETL	Apache Spark, Apache Airflow
Storage	SQLite, MongoDB, AWS S3
AI/ML & Forecasting	Scikit-learn, XGBoost, Prophet
NLP Analysis	SpaCy, BERT

Category	**Technologies**
Visualization	Apache Superset, Matplotlib, Seaborn
Reporting	ReportLab, ExcelWriter
Backend/Notebook	Jupyter Notebook
Deployment	Docker

Project Components

1. Data Ingestion & Processing

- Real-time data ingestion simulated via Kafka for student metrics like attendance, test scores, and feedback.
- Apache Spark used for batch processing of StudentPerformanceFactors.csv.
- Airflow used to automate ETL jobs and scheduling workflows.

2. Storage & Management

- Structured academic and performance data stored in **SQLite**.
- Unstructured feedback, logs, and NLP outputs managed in **MongoDB**.
- Historical data archived to **AWS S3** for backup and audit purposes.

3. AI & Predictive Analytics

- **Regression models** used to predict Exam_Score based on features like Hours_Studied, Attendance, etc.
- **XGBoost** applied for performance classification.
- **Prophet** used for time-based forecasting of student success rates.

4. NLP Analysis

- **SpaCy** used to extract keywords and sentiments from student feedback.
- **BERT** model fine-tuned to assess review sentiments and emotion classification.

5. Visualization & Reporting

- **Apache Superset** dashboards highlight performance trends, student engagement, and risk factors.
- **Automated PDF/Excel reports** generated monthly using **ReportLab** and pandas. Excel Writer.

6. Security & Access

- Basic access control implemented via Docker network isolation and Jupyter password protection.
- Data files securely managed with role-specific access in Airflow and Superset.

7. Deployment & Scalability

- Modular containers created via **Docker Compose** for Kafka, Spark, Airflow, Jupyter, and Superset.
- Lightweight deployment using shared volumes and minimal cloud footprint.

Business Analysis Insights

- Identified key drivers of academic success: Motivation_Level, Parental_Involvement, and Previous_Scores.
- Forecasted academic performance trends across semesters.
- Sentiment analysis of feedback indicated positive correlation between Teacher_Quality and performance.
- Flagged students at academic risk based on predictive scoring.

Dashboards & Reporting Highlights

Live Dashboards (Superset)

- **Performance Overview**: Real-time analysis of class and student-level metrics.
- **Subject-wise Trends**: Performance grouped by subjects and activities.
- **Risk Prediction**: Visualization of students below performance threshold.

Reports

- **PDF Reports**: Monthly academic reports for each class section.
- **Excel Reports**: Aggregated performance analysis and department comparison.

Development Workflow

Phases:

1. Dataset Cleaning & Ingestion
2. Batch ETL with Apache Spark

3. Model Training & NLP Analysis
4. Dashboard Creation & Reporting
5. Docker Deployment
6. Final Testing and CI Simulation

Final Deliverables

- Fully operational data analytics platform for student performance.
- Jupyter notebooks with clean, well-documented code.
- Superset dashboards and downloadable reports.
- Docker-based environment for future scalability and deployment.

Key Outcomes

- Enabled data-driven decision-making in academic planning.
- Forecasted student success rates with up to **87% accuracy** using ML.
- Demonstrated the potential of open-source tools for education analytics.
- Delivered a scalable and modular architecture for real-time and batch analytics.