**Student Performance & Behavior Analysis using Apache PySpark**

1. **Dataset Description**

**1.1 Source**

Data collected from an educational dataset on **Student Performance & Behavior**, comprising multiple academic and behavioral indicators such as test scores, attendance, study hours, stress levels, and lifestyle habits.  
**Source link:** https://www.kaggle.com/datasets/mahmoudelhemaly/students-grading-dataset

**1.2 Columns**

* **Student\_ID** – Unique identifier for each student.
* **Gender** – Male/Female category for demographic analysis.
* **Age** – Student’s age in years.
* **Department** – Academic department (e.g., CSE, ECE, MECH, etc.).
* **Attendance (%)** – Average attendance percentage.
* **Study\_Hours\_per\_Week** – Weekly average study duration.
* **Sleep\_Hours\_per\_Night** – Average hours of sleep per day.
* **Stress\_Level (1–10)** – Self-reported stress level on a scale of 1–10.
* **Assignments\_Avg**, **Quizzes\_Avg**, **Projects\_Score**, **Participation\_Score**, **Midterm\_Score**, **Final\_Score** – Academic performance components.
* **Total\_Score** – Computed overall performance metric.
* **Extracurricular\_Activities** – Indicates participation in sports or clubs.
* **Internet\_Access\_at\_Home**, **Parent\_Education\_Level**, **Family\_Income\_Level** – Socio-economic background indicators.

**1.3 Data Quality**

* Contains **~1,000 student records** with 15+ attributes.
* Missing values were identified and imputed using **mean for numeric fields** and **‘Unknown’ for categorical fields**.
* All numeric columns were typecast properly for analysis.
* Duplicates based on *Student\_ID* were removed.
* Dataset standardized and ready for large-scale analytical operations.

**2. Operations Performed**

**2.1 Data Cleaning & Exploration**

* Loaded dataset into PySpark DataFrame and validated schema integrity.
* Renamed columns to remove spaces and special characters.
* Imputed missing numeric data using mean aggregation with pyspark.sql.functions.mean.
* Replaced nulls in categorical fields with placeholder "Unknown".
* Converted relevant columns into appropriate numeric types using cast().
* Removed duplicate student entries using dropDuplicates().
* Filtered outliers in attributes like attendance, stress level, and study hours to ensure consistency.

**2.2 Descriptive Analytics**

* Computed key aggregates (mean, median, mode) for attendance, stress, and total score.
* Visualized **Total Score distribution** using histograms.
* Compared **Department-wise average performance** through bar plots.
* Analyzed **Gender-based performance differences** using grouped bar charts.
* Created **scatter plots** between Study Hours vs Total Score and Stress Level vs Total Score.
* Developed **heatmaps** showing correlations between all numeric academic and behavioral metrics.

**2.3 Relationship Analysis**

* **Attendance vs Performance:** Found strong positive correlation — students with higher attendance had better scores.
* **Stress vs Total Score:** Identified moderate negative correlation — higher stress correlated with lower grades.
* **Study Hours vs Scores:** Non-linear trend — moderate study hours (10–20/week) linked to best performance.
* **Sleep Hours vs Stress:** Students with better sleep schedules had significantly lower stress levels.

**3. Key Insights**

**3.1 Performance Overview**

*  Students with **attendance above 85%** consistently achieved high total scores.
*  Around **70% of students** maintain study hours between 10–20 per week — an optimal performance range.
*  **Sleep deficiency** (less than 6 hours per night) corresponds with higher stress and lower grades.

**3.2 Department-Wise Highlights**

* Computer Science and Electronics departments displayed the highest total performance averages.
* Mechanical and Civil departments showed greater variability, likely due to workload differences.
* Departments with higher extracurricular participation scored better overall in participation and projects

**3.3 Behavioral Patterns**

* Students engaged in extracurricular activities had slightly higher average stress but better participation and social performance.
* Internet access at home and parental education positively influenced quiz and assignment averages.
* Family income levels showed minimal impact compared to study habits and attendance.

**3.4 Trends Identified**

* Academic success is **multi-dimensional** — not solely based on study hours but a balance of rest, consistency, and engagement.
* **High stress levels** negatively impact final exam scores even for academically strong students.
* **Female students** slightly outperformed male students in participation and assignments, while male students scored higher in quizzes and projects.

**4. Recommendations**

**4.1 Academic Strategy**

* Institutions should promote consistent attendance and provide flexible learning support for absentees.
* Encourage balanced study routines — long hours do not always yield better outcomes.
* Implement stress-management workshops and counseling programs to improve mental well-being.

**4.2 Behavioral and Lifestyle Support**

* Promote sleep hygiene awareness as sleep patterns significantly affect stress and academic efficiency.
* Encourage extracurricular involvement for holistic development and teamwork enhancement.
* Provide equitable internet and learning resources access, especially for lower-income groups.

**4.3 Data-Driven Decision Making**

* Build predictive analytics models in Spark MLlib to forecast academic risk levels.
* Develop a student dashboard for monitoring attendance, stress, and performance trends in real time.
* Correlate data with external metrics like teacher feedback or mental health surveys for deeper insight.

**4.4 Future Analytical Scope**

* Extend the dataset with **multi-semester or multi-year data** for longitudinal analysis.
* Apply **clustering (K-Means)** to segment students based on performance profiles (high-achievers, moderate, at-risk).
* Integrate **sentiment analysis** on student feedback or engagement data to complement behavioral metrics.
* Implement **Spark Streaming** for real-time analytics on classroom performance and attendance systems.

**Project Repository:** **https://github.com/jukantijyoshna/BigDataAnalytics**