# **Student Academic Performance Dataset Documentation**

## **Overview**

This dataset contains academic and administrative records for 905 students enrolled in Computer Engineering (CE) and CE-adjacent programs (e.g., Computer Science and Design (CSD), Artificial Intelligence and Machine Learning (AIML)) across three semesters. The data includes student demographics, branch and department assignments, division and roll numbers, mentor assignments, and academic performance (theory/practical marks, attendance) for core and non-core subjects. The dataset is designed to support predictive modeling tasks, such as identifying students likely to experience reduced academic performance (e.g., percentile rank decline) in key Semester 3 subjects.

* **Rows**: 905 (one per student).
* **Columns**: 56 (covering demographics, administrative details, and academic performance).
* **Semesters**: 3 (Semester 1, Semester 2, Semester 3).
* **Core Subjects**: High-priority subjects (e.g., Math, Java, Python) determining roll assignments and student focus.
* **Non-Core Subjects**: Secondary subjects (e.g., Environmental Science, Law) included for regulatory or workshop purposes.
* **Marks and Attendance Range**: All attendance percentages, theory marks, and practical marks are numerical values in the range 0–100 (inclusive).
* **Data Quality**: Clean, with no missing or null values. Marks and attendance are complete where applicable (e.g., no practical for Math, no attendance for workshops).

## **Data Collection and Privacy**

* **Source**: Academic records from a CE program, anonymized for privacy.
* **Privacy Measures**:
  + **Student ID**: Unique identifier replacing Enrollment Number to ensure anonymity.
  + **Name**: Excluded from the dataset to protect privacy.
  + **Gender and Religion**: Derived from student names using GPT-3.0, potentially introducing noise (e.g., misclassifications).
* **Grading**: Exams are graded anonymously, minimizing bias in marks. Anecdotal evidence suggests low teaching bias, but this is unverified quantitatively.

## **Column Descriptions**

The dataset includes 56 columns, grouped into demographics, administrative, and academic categories. Below is a detailed description of each column, including format, values, and nuances.

### **Demographics (4 columns)**

| **Column** | **Description** | **Format/Values** | **Notes** |
| --- | --- | --- | --- |
| Student ID | Unique identifier for each student, replacing Enrollment Number for privacy. | String (e.g., S001, S002, ..., S905) | Used for tracking, not modeling unless needed for grouping. |
| Gender | Gender inferred from student name using GPT-3.0. | Categorical: Male, Female, Other | Potential noise due to inference errors; validate predictive power. |
| Religion | Religion inferred from student name using GPT-3.0. | Categorical: Hindu, Muslim, Christian, etc. | Potential noise; test relevance before including in models. |
| Branch | Academic branch assigned based on school marks, with exceptions (e.g., management quotas). | Categorical: CE, CSD, AIML, etc. | Proxy for school performance; combine with Roll-1 for better accuracy. |

### **Administrative (9 columns)**

| **Column** | **Description** | **Format/Values** | **Notes** |
| --- | --- | --- | --- |
| Div-1 | Division in Semester 1 (e.g., A1, B2, C3), based on school merit, grouped by department. | Categorical: A1, A2, B1, B2, C1, C2, etc. | Reflects department (A, B, C) and merit rank; redundant for modeling. |
| Div-2 | Division in Semester 2, based on Semester 1 core subject theory marks, per department. | Categorical: A1, A2, B1, B2, C1, C2, etc. | Redundant; avoid in models to prevent leakage-like effects. |
| Div-3 | Division in Semester 3, based on Semester 2 core subject theory marks, per new departments (A, B, C, D). | Categorical: A1, A2, B1, B2, C1, C2, D1, etc. | Redundant; exclude from models. |
| Roll-1 | Roll number in Semester 1, based on school merit, unique within department but duplicated across departments. | Integer: 1, 2, ..., ~60 | Proxy for school performance; combine with Branch to resolve duplicates. |
| Roll-2 | Roll number in Semester 2, based on total Semester 1 core subject theory marks, per department. | Integer: 1, 2, ..., ~60 | Reflects Semester 1 performance; redundant for modeling. |
| Roll-3 | Roll number in Semester 3, based on total Semester 2 core subject theory marks, per department. | Integer: 1, 2, ..., ~60 | Reflects Semester 2 performance; redundant. |
| Mentor-1 | Mentor assigned randomly for Semester 1, responsible for parent communication. | Categorical: Mentor\_A, Mentor\_B, etc. | Anecdotal low impact on marks; test predictive power before inclusion. |
| Mentor-2 | Mentor assigned randomly for Semester 2. | Categorical: Mentor\_A, Mentor\_B, etc. | Test before inclusion. |
| Mentor-3 | Mentor assigned randomly for Semester 3. | Categorical: Mentor\_A, Mentor\_B, etc. | Test before inclusion. |

**Administrative Notes**:

* **Departments**:
  + Semester 1: A, B, B (grouping branches, e.g., CE in A, CSD in B), with varying student counts and merit variances due to branch admission criteria.
  + Semester 3: Reassigned to A, B, C, D after Semester 2, balancing branches.
* **Divisions**: Assigned per department (e.g., A1, A2 for department A), ranked by merit (school marks for Semester 1, prior semester core theory marks for Semesters 2 & 3). Class sizes ~30, but vary slightly.
* **Roll Numbers**: Assigned per department, with top ranks (e.g., Roll 1–30) in higher divisions (e.g., A1). Roll-1 duplicates across departments (e.g., multiple students with Roll-1 in A, B, C).

## **Branch to Department Mappings**

The following mappings list the branches assigned to each department for Semesters 1, 2, and 3, based on the Div-1, Div-2, and Div-3 columns, respectively. Departments are derived from the first character of division values (e.g., A1 indicates department A). Branches are sorted alphabetically within each department.

### **Semester 1**

* **A**: AIDS, AIML, CEA, CS&IT
* **B**: CSE, CST, RAI
* **D**: CE, CSD, IT

### **Semester 2**

* **A**: AIDS, AIML, CEA, CS&IT
* **B**: CSE, CST, RAI
* **D**: CE, CSD, IT

### **Semester 3**

* **A**: AIDS, CE
* **B**: CEA, CSE
* **C**: AIML, CSD, IT, RAI
* **D**: CS&IT, CST

### **Academic: Semester 1 (12 columns)**

| **Column** | **Description** | **Format/Values** | **Notes** |
| --- | --- | --- | --- |
| Math-1 Theory | Theory marks for Math-1 (core). | Integer: 0–100 | Core subject, used for Roll-2 assignment. |
| Math-1 Attendance | Attendance percentage for Math-1. | Float: 0–100 | Key predictor of performance. |
| Physics Theory | Theory marks for Physics (core). | Integer: 0–100 | Core subject, used for Roll-2. |
| Physics Practical | Practical marks for Physics (core). | Integer: 0–100 | Core subject. |
| Physics Attendance | Attendance percentage for Physics. | Float: 0–100 | Key predictor. |
| Java-1 Theory | Theory marks for Java-1 (core). | Integer: 0–100 | Core subject, used for Roll-2. |
| Java-1 Practical | Practical marks for Java-1 (core). | Integer: 0–100 | Core subject. |
| Java-1 Attendance | Attendance percentage for Java-1. | Float: 0–100 | Key predictor. |
| Software Engineering Theory | Theory marks for Software Engineering (core). | Integer: 0–100 | Core subject, used for Roll-2. |
| Software Engineering Practical | Practical marks for Software Engineering (core). | Integer: 0–100 | Core subject. |
| Software Engineering Attendance | Attendance percentage for Software Engineering. | Float: 0–100 | Key predictor. |
| Environmental Science Theory | Theory marks for Environmental Science (non-core). | Integer: 0–100 | Checkbox subject, less predictive. |
| Environmental Science Attendance | Attendance percentage for Environmental Science. | Float: 0–100 | Less predictive. |
| IOT Workshop Practical | Practical marks for IOT Workshop (non-core). | Integer: 0–100 | Workshop-based, no attendance. |
| Computer Workshop Practical | Practical marks for Computer Workshop (non-core). | Integer: 0–100 | Workshop-based, no attendance. |

**Semester 1 Notes**:

* **Core Subjects**: Math-1, Physics, Java-1, Software Engineering (theory marks summed for Roll-2).
* **Non-Core Subjects**: Environmental Science, IOT Workshop, Computer Workshop (less focus, not used for roll assignments).
* **Attendance**: Missing for workshops (IOT, Computer) by design.

### **Academic: Semester 2 (11 columns)**

| **Column** | **Description** | **Format/Values** | **Notes** |
| --- | --- | --- | --- |
| Math-2 Theory | Theory marks for Math-2 (core). | Integer: 0–100 | Core subject, used for Roll-3. |
| Math-2 Attendance | Attendance percentage for Math-2. | Float: 0–100 | Key predictor. |
| Data Structures using Java Theory | Theory marks for Data Structures (core). | Integer: 0–100 | Core subject, used for Roll-3. |
| Data Structures using Java Practical | Practical marks for Data Structures (core). | Integer: 0–100 | Core subject. |
| Data Structures using Java Attendance | Attendance percentage for Data Structures. | Float: 0–100 | Key predictor. |
| DBMS Theory | Theory marks for DBMS (core). | Integer: 0–100 | Core subject, used for Roll-3. |
| DBMS Practical | Practical marks for DBMS (core). | Integer: 0–100 | Core subject. |
| DBMS Attendance | Attendance percentage for DBMS. | Float: 0–100 | Key predictor. |
| Fundamentals of Electronics Theory | Theory marks for Fundamentals of Electronics (core). | Integer: 0–100 | Core subject, used for Roll-3. |
| Fundamentals of Electronics Practical | Practical marks for Fundamentals of Electronics (core). | Integer: 0–100 | Core subject. |
| Fundamentals of Electronics Attendance | Attendance percentage for Fundamentals of Electronics. | Float: 0–100 | Key predictor. |
| Java-2 Theory | Theory marks for Java-2 (core). | Integer: 0–100 | Core subject, used for Roll-3. |
| Java-2 Practical | Practical marks for Java-2 (core). | Integer: 0–100 | Core subject. |
| Java-2 Attendance | Attendance percentage for Java-2. | Float: 0–100 | Key predictor. |

**Semester 2 Notes**:

* All subjects are core (Math-2, Data Structures, DBMS, Fundamentals of Electronics, Java-2).
* Theory marks summed for Roll-3 assignment.
* Full attendance data available.

### **Academic: Semester 3 (9 columns)**

| **Column** | **Description** | **Format/Values** | **Notes** |
| --- | --- | --- | --- |
| Math-3 Theory | Theory marks for Math-3 (core). | Integer: 0–100 | Core subject, key target for prediction. |
| DE Theory | Theory marks for Digital Electronics (core). | Integer: 0–100 | Core subject, key target. |
| DE Practical | Practical marks for Digital Electronics (core). | Integer: 0–100 | Core subject, key target. |
| FSD Theory | Theory marks for Full Stack Development (core). | Integer: 0–100 | Core subject, key target. |
| FSD Practical | Practical marks for Full Stack Development (core). | Integer: 0–100 | Core subject, key target. |
| Python Theory | Theory marks for Python (core). | Integer: 0–100 | Core subject, key target. |
| Python Practical | Practical marks for Python (core). | Integer: 0–100 | Core subject, key target. |
| Communication Theory | Theory marks for Communication (non-core). | Integer: 0–100 | Checkbox subject, less predictive. |
| Law Theory | Theory marks for Law (non-core). | Integer: 0–100 | Checkbox subject, less predictive. |

**Semester 3 Notes**:

* **Core Subjects**: Math-3, Digital Electronics (DE), Full Stack Development (FSD), Python (primary targets for prediction).
* **Non-Core Subjects**: Communication, Law (less focus, exclude from primary models).
* No attendance data, by design.

## **Data Nuances**

* **Branch and Department**:
  + Branches (e.g., CE, CSD, AIML) are assigned based on school marks, with exceptions (e.g., management quotas).
  + Semester 1 departments (A, B, C) group branches with varying merit requirements and student counts.
  + Semester 3 departments (A, B, C, D) are reassigned post-Semester 2, balancing branches.
* **Division and Roll**:
  + Divisions (e.g., A1, A2) rank students by merit within departments (school marks for Semester 1, prior semester core theory marks for Semesters 2 & 3).
  + Roll numbers reflect merit (e.g., Roll-1 = top rank); Roll-1 duplicates across departments but can be resolved with Branch or Student ID.
  + Class sizes ~30, varying slightly.
* **Mentors**:
  + Randomly assigned each semester, responsible for parent communication (e.g., low marks, attendance).
  + Anecdotal evidence suggests low impact on marks; validate with feature importance.
* **Core vs. Non-Core Subjects**:
  + Core subjects (e.g., Math, Java, Python) receive maximum student and faculty focus, determine roll assignments, and are primary targets for predictive models.
  + Non-core subjects (e.g., Environmental Science, Law) are secondary, with lower predictive relevance.
* **Attendance**:
  + Available for Semester 1 & 2 core and some non-core subjects (except workshops).
  + Missing for Semester 3 and workshop subjects (IOT, Computer Workshop) by design.

## **Potential Biases and Limitations**

* **Derived Demographics**:
  + Gender and Religion inferred via GPT-3.0 may introduce errors (e.g., misgendering, incorrect religion), reducing reliability for modeling.
  + **Mitigation**: Test predictive power (e.g., correlation with marks) before inclusion; consider excluding if noisy.
* **Branch as School Marks Proxy**:
  + Branch assignment based on school marks is imperfect due to exceptions (e.g., management quotas).
  + Roll-1 is a better proxy but has duplicates across departments; combining with Branch improves accuracy.
* **Teaching Bias**:
  + Anecdotal evidence suggests low teaching bias, but lack of quantitative data (e.g., student feedback) limits validation.
  + **Mitigation**: Analyze mark distributions across divisions/departments for consistency.
* **Mentor Impact**:
  + Random mentor assignment suggests minimal impact, but unverified without statistical analysis.
  + **Mitigation**: Include Mentor columns initially, drop if feature importance is low.
* **Subject Hardness**:
  + Core subjects vary in difficulty (e.g., Math vs. Python), affecting mark distributions.
  + **Mitigation**: Use percentile ranks to normalize marks for predictive tasks.
* **Attendance Absence in Semester 3**:
  + Limits features for predicting Semester 3 outcomes, relying heavily on Semester 1 & 2 data.
  + **Mitigation**: Derive features like average Semester 1 & 2 attendance to capture engagement.
* **Class Imbalance**:
  + Reduced performance (e.g., percentile drop) may be rare, leading to imbalanced classification.
  + **Mitigation**: Use SMOTE, class weights, or focus on recall in evaluation.

## **Usage Notes**

* **Predictive Modeling**:
  + Suitable for tasks like predicting reduced academic performance (e.g., percentile rank drop in core Semester 3 subjects: Math-3, Python, DE, FSD).
  + Recommended features: Semester 1 & 2 core subject marks, attendance, derived features (e.g., GPA, percentile ranks), Gender, Religion, Branch (with validation).
  + Exclude: Student ID, Name, Division, Roll (redundant), Mentor (unless validated), non-core subjects (unless relevant).
* **Preprocessing**:
  + Standardize numerical features (marks, attendance).
  + One-hot encode categorical features (Gender, Religion, Branch, Mentor).
  + Compute percentile ranks for marks to normalize subject hardness.
* **Evaluation**:
  + Use F1-score, recall, and ROC-AUC for classification tasks (e.g., predicting reduced performance).
  + Visualize feature importance, ROC curves, and percentile trends for interpretability.
* **Applications**:
  + Identify students likely to underperform in core subjects for targeted interventions (e.g., tutoring).
  + Analyze predictors of academic decline (e.g., low attendance, prior marks).

## **Contact**

For questions or additional details, contact me at linkedin: https://www.linkedin.com/in/shail-k-patel/