

Data Preprocessing and Missing Data Handling for Predicting High School Academic Outcomes

Rinku Mani Kalita^{1*}, Dr Siddhartha Baruah²

¹PhD Scholar, Assam Science and Technology University(ASTU) Assam

Email:rinkumani@gmail.com, ORCID ID: 0009-0003-7790-872X

²Professor and HOD MCA Dept., Jorhat Engineering College,

Jorhat, Assam Email: sbaruain@gmail.com, ORCID ID: 0009-0005-0597-1783

Abstract—The challenge of accurately estimating student academic results based on complicated education records, including HSLC outcomes, where rigorous preprocessing and attentive handling of multi-field information, including demographics, institution characteristics, and topic-level marks, is frequently afflicted by non-existent values and discrepancies that undermine the independence of models (Hunt et al., 2021; Konda, 2022). It is a paper on an end-to-end preprocessing system harmonised to the educational marks data, comprising schema auditing, type enforcement, categorical harmonisation, the scale of numeric, outlier treatment, feature engineering, and imputation strategies which avoid information leakage on validation. The framework combines a mean imputation of statistical attributes, constant-token imputation of categorical variables with informative missing-indicators, with optional correlation elimination and a Principal Component Analysis (PCA) to provide a concise representation where necessary. Structured gaps (unused fields and systematically empty components) are given particular attention, and non-applicable practical are treated, wholly empty columns are not included. Our framework also dramatically increases the predictive algorithm performance and stability of predictive algorithms such as Random Forests, k-NN, and SVMs by making them work on clean and well-structured data. Moreover, it makes missingness consciously encoded, increasing the interpretability of the model, which is more informative to educational policy analysis. The work emphasises preparing an accurate, clean, and reliable dataset with regard to various approaches that are used as a subset of data preprocessing, which is considered paramount important in converting live, raw and messy large educational data into a trustworthy dataset to make decisions.

Keywords—Educational data mining, Data preprocessing, Missing data, Imputation, HSLC, Pipeline.

I. INTRODUCTION

Education datasets, including High School Leaving Certificate (HSLC) scores, tend to be heterogeneous, and they can comprise numerical scores, demographic categorical data, and other multi-dimensional information (Khelifi et al., 2024). This complexity makes data preprocessing a critical step to do any meaningful analysis (Zhen and Zhang, 2019). Cleaning and imputing missing data methods have a direct impact on the quality and fairness of the results (Hunt et al., 2021). Even shallow biases at this level may be magnified, and the results could be systematically biased (Konda, 2022). In this paper, the focus is limited to methodology of data preprocessing to come up with a standardised solution in dealing with educational data that is efficient, ethical, and can be implemented in various schools and in various groups of students.

II. IMPLEMENTING STEPS

A. Data Source and Initial Processing

The research is founded on the results of High School Leaving Certificate (HSLC), divided into identifiers, demographics, school data, subject scores, overall results, grades, and

remarks. The initial processing stage was a data audit to validate the entries against official regulations and missing data patterns in the catalogue (Çetin and Yildiz, 2023). The area that was totally blank was recorded and therefore not included in analysis to simplify the dataset.

B. Dataset Description

All the 72 fields are grouped in seven logical categories with elaborate specifications on data validation and processing in the comprehensive data dictionary.

Identifiers (11 fields)

These fields uniquely identify districts, schools, centres, and individual students within the examination system.

TABLE 1

Field Name	Format	Definition	Validation Notes
DST_CODE	Integer	Unique numeric identifier for the district	Standard field validation
DST_NAME	String	Full name of the district where school is located	Text validation required
CNT_CODE	Alphanumeric	Unique centre identification code	3 digit district code + 4 digit centre code
CNT_NAME	String	Full name of the centre where school is attached	Text validation required
SCH_CODE	Alphanumeric	Unique school identification code	7 digit
SCH_NAME	String	Full name of the school	Text validation required
CNAME	String	Full name of the student	Text validation required
FNAME	String	Full name of the student's Father	Text validation required
MNAME	String	Full name of the student's Mother	Text validation required
REGISTRATION NO	String	Student registration number	Format: District/Type/School/Number/Year
ROLLNO	Integer	Sequential roll number within school	Range: 1-17 in sample

Student Demographics (4 fields)

Personal information fields capturing student identity and background characteristics.

TABLE 2

Field Name	Format	Allowed Values	Missing Data Handling
DOB	DD/MM/YYYY	Valid date format	No missing values
GENDER	Single Character	M, F	No missing values in sample
RELIGION	String	Religious affiliation	All Hindu in sample
CASTE	String	GENERAL, OBC, MOBC, SC, ST	Four categories present

Institutional characteristics defining the educational context.

- COURSE: Course type - High School Leaving Certificate (HSLC)
- MEDIUM: Language medium of instruction (Assamese in sample)
- SCHOOL STATUS: School type (PR=Private, GV=Government)

Subject Performance (42 fields)

Detailed academic assessment data across six subjects with multiple components per subject.

Subject Structure

- S1: Assamese (Compulsory)
- S3: English (Compulsory)
- S4: General Mathematics (Compulsory)
- S5: General Science (Compulsory)
- S6: Social Science (Compulsory)
- S7: Elective subject (Hindi/Geography/Advanced Mathematics)

Performance Components per Subject

TABLE 3

Component	Definition	Format	Range	Remarks
_TH_FM	Theory Full Marks	Integer	0-100	Field for S1 Subject only
_TH_OM	Theory Obtained Marks	Integer	0-100	Field for all Subject only
_PR_OM	Practical Obtained Marks	Integer	0-100	Field for all Subject only
_FM	Total Full Marks	Integer	0-100	Field for all Subject only
_OM	Total Obtained Marks	Integer	0-100	Field for all Subject only
_GRS	Grace marks	Text	+1 to +5	Field for all Subject only
_FLAG	Performance Flag	String	F=Fail, L=Letter Grade	Field for all Subject only

Subject Description (6 fields)

TABLE 4

Field Name	Format	Definition	Validation Notes
S1_NAME	String	Full name subject	Text validation required
S3_NAME	String	Full name subject	Text validation required
S4_NAME	String	Full name subject	Text validation required
S5_NAME	String	Full name subject	Text validation required
S6_NAME	String	Full name subject	Text validation required
S7_NAME	String	Full name subject	Text validation required

Aggregated performance metrics and final classification.

- GTOTAL_FM: Grand total full marks (600 for all students)
- GTOTAL_PM: Grand total practical marks (180 for all students)
- GTOTAL_OM: Grand total obtained marks (Range: 172-442)
- DIVISION: Final result classification (1st, 2nd, 3rd, Fail, Withheld, Fully Absent)
- DIVESC: Star marks (*), Distinction (D)

REMARKS: Comment field

III. DATA PREPROCESSING

A. Handling of Missing Data

Educational datasets typically present three types of missing data:

- Missing Records: Entire student entries are absent from the dataset.
- Missing Values: Specific data fields within otherwise complete records are empty.
- Structural Absence: Fields are empty because they are not applicable to certain subjects or are from deprecated systems no longer in use (Hunt et al., 2021).

The dataset contain 12211 rows and 72 Columns.

Missing records (entire student entries absent) - This type is not directly visible inside the dataset because if a student row is missing, it simply isn't there to count after checking dataset for gaps in ROLLNO within the same school (a gap can indicate a missing student row, or it can also be "roll number not issued").

Table A. Missing records (row-level absence)

Category	What would be seen	Can it be Verified from the file ?	Practical check	Recommended handling
Missing records	The student row is not in the dataset at all	No	Compare against external roster; check roll number continuity within schools	Reconcile before calculating participation rates, pass rates, or school-level totals

Missing values (fields empty inside otherwise present records) - These are classic nulls like missing demographic fields. From dataset it is found that one student record has RELIGION missing while name, DOB, gender, caste, and medium are present. To handle demographic fields like RELIGION, it is treated as missing as its own category, such as Unknown, rather than dropping (avoids bias). As the religion isn't needed for the analysis, it leaves as null.

Table B. Missing values (field-level blanks)

Category	Typical example in dataset	Reasons	Recommended handling
Missing values	RELIGION is blank while identifiers and other demographics are present	Not recorded, data entry omission, privacy/no response	Keep as missing (null); report missing percentage; avoid filling with "0"; impute only if your study requires it and you can justify the method

Structural Absence. Structural absence is widespread in optional-subject fields: columns tied to S6 and S7 are frequently empty because these subjects are not applicable to many students. For those entries as "Not Applicable" and avoid inputting zeros, to prevent misclassification of non-participation as poor performance.

Table C. Structural absence (not applicable fields)

Category	Example in the dataset	Meaning of the blank entry	Prohibited Actions	Recommended handling
Structural absence	S6_* and S7_* (optional subject marks/grades) are blank	Student did not take that optional subject (not applicable)	Do not impute 0 or treat blank as failure	Label as Not Applicable; compute subject statistics only among students who attempted that subject

B. Python Implementation

In Python, the Pandas library is used to handle missing data in this HSLC dataset of 12,211 rows and 72 columns. To detect possible missing student records, it is sorted by school and roll number, then gaps are seen in roll numbers, which may indicate absent rows. For missing values inside records, such as one student with RELIGION missing but other details present, it is kept as null because religion is not required for the analysis, avoiding unnecessary changes. For structural absence, like empty optional-subject fields (S6 and S7), it is labelled as "Not Applicable" instead of zero, so it should not be confused with "not taken" or "low marks."

Step-by-step: how it is done in Python

Step 1: Load the dataset. It read the Excel file into a pandas dataframe called df

```
import pandas as pd
df = pd.read_excel("HSLC_R~4.XLS")
```

Step 2: Find possible missing student records by roll number gaps. It sorts by school and roll number, then checks gaps; big gaps may mean missing student rows.

```
df_sorted = df.sort_values(["SCHOOL_CODE", "ROLLNO"])
df_sorted["ROLL_GAP"] = df_sorted.groupby("SCHOOL_CODE")["ROLLNO"].diff()
```

```
possible_missing_students=df_sorted[df_sorted["ROLL_GAP"] > 1]
```

Step 3: Look at missing RELIGION values (and leave them null). It finds the rows where RELIGION is empty and keeps them as null, because religion is not needed.

```
missing_religion = df[df["RELIGION"].isna()]
```

Step 4: Mark structural absence in optional subjects (S6, S7). If S6 or S7 marks are empty, it will label them as "Not Applicable" instead of putting zero.

```
optional_cols = ["S6_MARKS", "S7_MARKS"]
```

```
for col in optional_cols: df[col] = df[col].where(df[col].notna(), "Not Applicable")
```

IV. RESULTS AND DISCUSSION

The High School Leaving Certificate (HSLC) dataset revealed that what initially looked like a "missing data" problem is actually much more complex. It is found that the dataset doesn't suffer from a single type of error instead, it reflects the real-world structure of the examination system itself. The most significant finding was that the vast majority of blank cells are not mistakes at all, they are instances of structural absence. This is most obvious in the columns for optional subjects and their corresponding grades. When a student's record shows a blank for an elective like Advanced Mathematics or Computer Science, it simply means they didn't choose that subject, not that the data was lost or unrecorded.

Due to this distinction, standard statistical fixes like mean imputation, where blanks are filled with an average score, would be completely wrong for this situation. If those optional subject blanks were filled with average scores, then data would be invented for exams that never took place. Therefore, the primary results show that the most appropriate method for handling these blanks is to reclassify them as "Not Applicable" rather than treating them as missing values or as zeros. A zero implies failure, whereas a blank in this instance implies non-participation, which is a crucial difference for accurate reporting in real world.

Demographic fields such as religion or caste were also examined, where missing values were found to be real but few in numbers. In these instances, the data is sufficiently clean and complex algorithms are not required to estimate missing values, simply reporting that a small percentage is unknown does not compromise the overall analysis. However, a significant limitation must be acknowledged the analysis is restricted to the available records. If an entire student's record is absent from the database, this omission cannot be detected from the table alone, and external enrolment lists would be required to verify total coverage.

V. CONCLUSION

The HSLC dataset is robust and ready for analysis. The "missingness" here is largely a feature, not an error by treating optional subject blanks as structural essentially ignoring them when calculating subject averages and future enhancements. It is ensured that performance metrics are fair and accurate. It is recommended that future reports focus on "attempt-based" statistics, meaning it only calculates pass percentages for students who actually sat for a specific paper. As long as analysts avoid the trap of treating every blank cell as an error to be fixed, this data can be reliably used for descriptive statistics and educational modelling without heavy modification. After implementing the above-mentioned preprocessing, the dataset will be ready for future analysis by comparing different algorithms and finding out which algorithm will give better results. It can be further used to find the hybrid model which will be used to find the best result.

REFERENCES

- [1] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-17.
- [2] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- [3] Costa, E. B., et al. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*, 73, 247-256.
- [4] Fix, E., & Hodges Jr, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.
- [5] Romero, C., et al. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68, 458-472.
- [6] Vapnik, V. (1995). The nature of statistical learning theory.