ZENTHIC AI INTERNSHIP

Data Quality Report

-DEETHIKSHA R

1. Short context: dataset source & intent

The dataset consists of student demographic, academic, and administrative records collected for analytical and learning purposes. It is intended to simulate real-world educational data used in data science workflows. The primary objective of this dataset is to assess data quality issues, perform exploratory analysis, and apply data cleaning and preprocessing techniques to transform raw student information into a reliable, consistent, and model-ready dataset suitable for downstream machine learning and analytical tasks.

2. Top 5 issues found and fixed along with code

i) Placeholder values used instead of missing data

Several columns contained placeholder strings such as NA, N/A, unknown, -, and empty values, which masked true missingness in the dataset. These were standardized and replaced with proper NaN values to enable accurate handling of missing data.

```
df.replace(["NA", "N/A", "na", "-", "--", "unknown", "Unknown", ""], np.nan, inplace=True)
```

ii) Inconsistent text formatting in categorical columns

Categorical variables such as gender, city, and device type contained inconsistent casing and leading/trailing spaces, resulting in duplicate category labels. Text normalization was applied across all object-type columns.

```
for col in df.select_dtypes(include="object").columns:
df[col] = df[col].str.strip().str.lower()
```

iii) Mixed and invalid date formats in admission dates

The admission date column included multiple date formats and invalid strings, which prevented direct conversion. These were parsed using pandas datetime conversion with error coercion.

```
df["admission_date"] = pd.to_datetime(df["admission_date"], errors="coerce")
```

iv) Non-numeric and impossible values in numeric columns

Several numeric fields contained text values or unrealistic values such as GPA greater than 10 and attendance rate exceeding 100%. These were coerced to numeric types and corrected using business rules.

```
for col in ["age","study_hours_per_week","attendance_rate","prior_gpa_10pt","test_score",
"fee_paid_inr"]:
df[col] = pd.to_numeric(df[col], errors="coerce")


df.loc[df["prior_gpa_10pt"] > 10, "prior_gpa_10pt"] = np.nan
df.loc[df["attendance_rate"] > 100, "attendance_rate"] = np.nan
```

v) Missing values in key academic and categorical fields

Important variables such as GPA and attendance rate had missing values that could impact analysis. Median imputation was applied to numeric columns, and categorical missing values were labeled as "unknown" to preserve information about missingness.

```
df["prior_gpa_10pt"] = df["prior_gpa_10pt"].fillna(df["prior_gpa_10pt"].median())
df["attendance_rate"] = df["attendance_rate"].fillna(df["attendance_rate"].median())
for col in categorical_cols:
    df[col] = df[col].fillna("unknown")
```

3. Assumptions Made During Data Cleaning

- Placeholder values such as NA, N/A, unknown, and empty strings were assumed to represent missing data and were converted to NaN.
- GPA values greater than 10 and attendance rates above 100 were considered invalid based on standard academic rules and were treated as missing values.
- Median imputation was applied to numerical columns (e.g., GPA and attendance rate) under the assumption that missing values were randomly distributed.
- Missing categorical values were intentionally filled with "unknown" to preserve information about data absence rather than discarding records.

- Exact duplicate records were assumed to be unintentional repetitions and were removed without affecting data integrity.

4. Final Missingness Table

After completing all cleaning and imputation steps, the remaining missing values were assessed using the following code:

final_missing = df.isna().sum().sort_values(ascending=False).head(10)

final_missing

Column Name	Missing Count
comments	112
admission_date	12
age	0
gender	0
course_stream	0
device_type	0
parental_education	0
attendance_rate	0
prior_gpa_10pt	0
fee_paid_inr	0