DATA ANALYSIS AND VISUALIZATION

Prepared by: 23BCE194(SARTH NAROLA) / 23BCE195(NEVIL ANGHAN)

Course / Course code: Data Analysis and Visualisation / 3CS103ME24

PRACTICAL : 1

***Educational Data Analysis for Student Status Prediction***

**Domain Overview – Education**

The education domain deals with data related to students, academic performance, institutional processes, and learning behavior. Analyzing such data helps institutions improve teaching quality, reduce dropout rates, and personalize education.

**Dataset Exploration**

Total Records: 4,424 student entries

Total Features: 37 input features + 1 target variable

Target Variable: Target – Student academic status:

* Dropout
* Graduate
* Enrolled

**Attribute Identification (with Feature Type Classification)**

| **Feature Name** | **Description** | **Type** | **Scale** |
| --- | --- | --- | --- |
| **Marital Status** | Marital status of student | Nominal | Categorical |
| **Application mode** | How application was submitted | Nominal | Categorical |
| **Application order** | Priority of the course applied for | Ordinal | Discrete |
| **Course** | Course identifier | Nominal | Categorical |
| **Daytime/evening attendance** | Attendance mode | Binary | Categorical |
| **Previous qualification** | Type of qualification held | Nominal | Categorical |
| **Previous qualification (grade)** | Grade achieved in previous qualification | Numeric | Ratio |
| **Nacionality** | Nationality of student | Nominal | Categorical |
| **Mother's qualification** | Mother's education level | Ordinal | Categorical |
| **Father's qualification** | Father's education level | Ordinal | Categorical |
| **Mother's occupation** | Mother's occupation type | Nominal | Categorical |
| **Father's occupation** | Father's occupation type | Nominal | Categorical |
| **Admission grade** | Grade on admission | Numeric | Ratio |
| **Displaced** | Is the student displaced (yes/no) | Binary | Nominal |
| **Educational special needs** | Any educational special needs | Binary | Nominal |
| **Debtor** | Student in debt (yes/no) | Binary | Nominal |
| **Tuition fees up to date** | Fee payment status | Binary | Nominal |
| **Gender** | Gender of the student | Binary | Nominal |
| **Scholarship holder** | Whether on scholarship | Binary | Nominal |
| **Age at enrollment** | Age when enrolled | Numeric | Ratio |
| **International** | Is the student international | Binary | Nominal |
| **Curricular units 1st sem (credited)** | Units credited in 1st semester | Numeric | Discrete |
| **Curricular units 1st sem (enrolled)** | Units enrolled in 1st semester | Numeric | Discrete |
| **Curricular units 1st sem (evaluations)** | Evaluations taken in 1st sem | Numeric | Discrete |
| **Curricular units 1st sem (approved)** | Approved units in 1st sem | Numeric | Discrete |
| **Curricular units 1st sem (grade)** | Grade average in 1st sem | Numeric | Ratio |
| **Curricular units 1st sem (without evaluations)** | Subjects skipped evaluations | Numeric | Discrete |
| **Curricular units 2nd sem (credited)** | Units credited in 2nd semester | Numeric | Discrete |
| **Curricular units 2nd sem (enrolled)** | Units enrolled in 2nd semester | Numeric | Discrete |
| **Curricular units 2nd sem (evaluations)** | Evaluations taken in 2nd sem | Numeric | Discrete |
| **Curricular units 2nd sem (approved)** | Approved units in 2nd sem | Numeric | Discrete |
| **Curricular units 2nd sem (grade)** | Grade average in 2nd sem | Numeric | Ratio |
| **Curricular units 2nd sem (without evaluations)** | Subjects skipped evaluations | Numeric | Discrete |
| **Unemployment rate** | National unemployment at time of study | Numeric | Interval |
| **Inflation rate** | National inflation at time of study | Numeric | Interval |
| **GDP** | GDP value at time of study | Numeric | Ratio |
| **Target** | Final status: Graduate, Dropout, Enrolled | Nominal (Target Variable) | Categorical |

**Applications of the Dataset**

1. **Machine Learning Tasks**

* **Classification**:
  + Predict final status (Graduate, Dropout, Enrolled)
  + Identify risk groups based on features
* **Regression**:
  + Predict continuous scores (e.g., grades, average performance)

1. **Practical Use-Cases**

* **Dropout Risk Prediction** – Early detection of high-risk students
* **Scholarship Eligibility** – Based on financial and academic data
* **Performance Monitoring** – Using grade-related features
* **Admission Planning** – Via insights from Application mode and Order
* **Curriculum Optimization** – Through evaluation/approval rates

**Challenges in the Dataset**

| **Challenge** | **Description** |
| --- | --- |
| **Missing Data** | Some features may have null values that must be imputed |
| **Imbalanced Target Classes** | ‘Graduate’ has more examples than ‘Dropout’ |
| **Interpretability** | Some features (like Application mode) need encoding |
| **Outliers** | Extreme values in grades, evaluations may skew results |
| **Static Data** | Dataset lacks time-series dimension for behavior tracking |

**Stakeholders & Their Use Cases**

| **Stakeholder** | **Use** |
| --- | --- |
| **Institutions** | Identify students at risk, optimize scholarships & resources |
| **Teachers** | Adjust teaching based on performance data |
| **Parents** | Monitor children’s performance and risk alerts |
| **Policy Makers** | Understand trends, allocate funding, policy design |

**Key Takeaways**

* The dataset enables critical analysis of **student academic lifecycle**.
* Proper preprocessing (handling missing values, encoding) is essential.
* Predictive modeling can aid **early interventions** for dropout prevention.
* Classification and regression tasks both have strong applicability