Preparing Data for Smarter Decisions

- How raw student data is transformed into reliable insights
- Audience : Business and Non-Technical Stakeholders

By Aman Verma

1753 FY AI-ML

Why Data Preparation Matters

- Machine learning systems depend entirely on data quality.
- Poor data leads to unreliable decisions.
- Clean and structured data leads to accurate and trustworthy outcomes.

Our Starting Point: Raw Student Data

Information collected includes:

- Student identifier
- Gender
- Academic marks
- Attendance records
- Study hours

Challenges observed:

- Missing values
- Different data formats
- Uneven value ranges

Step 1: Identifying Missing Information

- Some student records contain missing marks or attendance values.

- If left untreated, missing data can distort analysis and reduce confidence in results.

- The first step is to clearly identify these gaps.

Step 2: Filling Missing Values

- Instead of removing records, missing numerical values are filled logically.

Common approaches:

- Mean (average value)
- Median (middle value)

- This preserves data completeness and fairness.

Business Perspective on Missing Data

- Ensures no student record is unfairly excluded.
- Maintains full dataset size.
- Supports balanced and unbiased insights.

Step 3: Converting Text into Numbers

- Machine learning systems cannot interpret text values.

- Categorical information such as gender is converted into numeric form while preserving meaning.

- This enables automated analysis.

Step 4: Standardizing Numerical Values

- Different fields operate on different scales.

For example:

- Marks range higher
- Study hours range lower

- Scaling aligns all values to a comparable range, preventing bias.

Step 5: Handling Unusual Values

- Some values may be unusually high or low compared to the rest.

- These outliers can disproportionately influence results.

- They are identified and adjusted or controlled.

Step 6: Training and Testing the Model

The dataset is divided into two parts:

- Training data to build the model
- Testing data to evaluate accuracy

- This validates real-world performance.

Final Outcome

The resulting dataset is:

- Complete
- Consistent
- Balanced
- Ready for machine learning

- This ensures dependable predictions.

Business Impact

- Higher confidence in analytics.
- Improved decision-making.
- Reduced risk due to unreliable data.
- Scalable approach for future datasets.

Implementation Reference

- The full implementation code demonstrating this data preparation process is available at the following link:

https://colab.research.google.com/drive/1nPzATVESBFU_mAABK2eMlLbtkFzVhHVm?usp=sharing

- This link is provided for transparency and auditability.

Key Takeaway

- Data preparation is the foundation of reliable machine learning.

- Well-prepared data leads to trustworthy insights and better business outcomes.