COMP 331 – Final Project Report

Option 2: Data Quality & Bias Analysis

Dataset: UCI Student Performance

Bhuvraj Khangura
300217460

1. Introduction

This project analyzes the Student Performance datasets from the UCI Machine Learning Repository. The data includes demographic background, family context, academic support, attendance, lifestyle habits, and three grade outcomes (G1, G2, G3) for secondary school students enrolled in Mathematics and Portuguese courses. The information is provided in two separate CSV files. These datasets are frequently used for educational research and predictive modeling. Because model accuracy depends on the quality of the input data, this report evaluates the dataset using three key data quality dimensions: completeness, consistency/uniqueness, and representativeness.

GitHub Repository: https://github.com/BSK15/COMP331-Final-Project

2. Data Quality Analysis

2.1 Completeness

Individually, each file appears complete with no obvious missing values. However, a major issue emerges when combining the two subjects. The Math file contains 395 students, while the Portuguese file contains 649. After aligning records using shared demographic and family attributes, only **about 382 students** appear in both files. Any analysis requiring both subjects must rely on this smaller overlapping group, limiting the usefulness of cross-subject comparisons.

2.2 Consistency and Uniqueness

Each student should appear once per subject, so ensuring proper row alignment is important. If the datasets were merged incorrectly, duplicate or mismatched records could occur. The merge process avoids this by using multiple identifiers together, helping maintain one unique entry per student and reducing the risk of inconsistent or conflicting information.

2.3 Representativeness

The data comes from only two Portuguese schools. As a result, the sample does not represent broader student populations from other regions or school systems. Models trained on this narrow sample may not generalize well and could unintentionally favour overrepresented groups within the dataset.

3. Recommendations

- Keep subject-specific attributes separate when merging (e.g., absences_math and absences_port).
- Use only the overlapping subset when comparing Math and Portuguese outcomes.
- Interpret results carefully due to the dataset's limited and non-representative sample.

4. Conclusion

The dataset provides valuable student information but contains important quality limitations. Differences in file sizes reduce completeness, the merging process must be handled carefully to maintain uniqueness, and the narrow sampling limits how broadly findings can be applied. Addressing these issues strengthens the reliability of any analysis performed using this dataset.

References

Bache, K., & Lichman, M. (2013). *UCI Machine Learning Repository: Student Performance Data Set*. University of California, Irvine. https://archive.ics.uci.edu/dataset/320/student+performance

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Zhang, Y., & Yang, Q. (2015). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 29*(12), 2595–2613. https://doi.org/10.1109/TKDE.2016.2628028

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). *Data mining for business analytics: Concepts, techniques, and applications in Python* (2nd ed.). Wiley.

UFV COMP 331 Course Materials. *Data quality, warehousing concepts* [Lecture slides].

ChatGPT (GPT-5.1) was used to assist with generating R scripts and troubleshooting code for data quality checks in this project.