Predicting Students' Final Grades Using Machine Learning

Benjamin Batte

University of the Cumberlands

ITS-836 Data Science & Big Data Analytics

Dr. Charles Edeki

October 5, 2025

Introduction

Predicting students' final grades is a critical domain of educational data science because it enables institutions to design proactive interventions and deliver personalized support mechanisms. Accurately identifying at-risk students has mitigated dropout rates and improved overall academic achievement (Al-Barrak & Al-Razgan, 2020). Empirical studies consistently demonstrate that predictors such as study time, attendance, family support, and prior academic performance exert significant influence on student outcomes (Shorfuzzaman et al., 2021). By systematically analyzing these factors, educators and policymakers can better enhance resource allocation and target support services more effectively. This project extends earlier investigations using traditional machine learning for student performance forecasting (Panigrahi et al., 2021) and ensemble learning frameworks for academic prediction tasks (Manhães et al., 2022).

The overarching objective is to apply advanced regression methodologies to predict final grades with higher accuracy while isolating the most influential variables driving student success. The central hypothesis guiding this research is that prior academic performance (G1 and G2) will emerge as the strongest predictor of final performance (G3), while socio-behavioral and demographic factors will exert comparatively weaker influence.

Methods

Data Source

The analysis employed the Student Performance dataset from the UCI Machine Learning Repository, aggregating academic records from two Portuguese secondary schools (Cortez & Silva, 2008). The dataset contains 649 observations and 33 attributes

for the Portuguese language course, and 395 observations with the same 33 attributes for the mathematics course. These features span demographic, social, and academic dimensions, including parental education, study time, alcohol consumption, and prior grades, providing a comprehensive basis for multi-factorial performance prediction.

Data Wrangling and Preprocessing

Data wrangling represented a critical component of the methodological framework, given that data cleaning and transformation processes can consume up to 80% of the data science lifecycle (McKinney, 2022). In this study, preprocessing followed a structured, multi-stage approach. Categorical variables were encoded to convert nominal and ordinal data into numerical formats suitable for machine learning models. Numerical variables were scaled to normalize distributions and reduce sensitivity to variance, thereby improving model stability and performance. The datasets were subsequently partitioned into stratified training and testing subsets to preserve class balance and ensure generalizability across unseen data. These transformations were systematically implemented through a Column Transformer pipeline within scikit-learn, ensuring reproducibility, modularity, and scalability throughout the experimental workflow (Kumari & Singh, 2023).

Analytical Techniques and Tools

Python served as the primary computational environment due to its extensive ecosystem of libraries optimized for data-intensive research. Data handling was performed using Pandas and NumPy (McKinney, 2022). Model development, training, and validation were conducted in scikit-learn (Kumari & Singh, 2023), while data

visualization and exploratory analysis were implemented using Matplotlib and Seaborn (Sasikala & Renuka Devi, 2023).

The predictive analysis was conducted using two supervised learning paradigms: Linear Regression, chosen for its interpretability, and Random Forest Regression, selected for its robustness in handling complex, non-linear data structures. As a baseline algorithm, Linear Regression provides interpretable coefficient-based insights into variable influence but is limited in handling nonlinearity and categorical complexity. Random Forest Regression, an ensemble method based on bootstrap aggregation of decision trees, was included to address these limitations and has demonstrated robust predictive accuracy in prior educational modeling studies (Shorfuzzaman et al., 2021). Complementary research employing ensemble learning further validates the suitability of tree-based models for student performance prediction (Manhães et al., 2022).

Results

Exploratory Data Analysis

Exploratory data analysis was conducted separately for the Mathematics and Portuguese datasets to examine grade distributions and assess potential predictors of student performance.

Figure 1 illustrates the distribution of absences in the mathematics dataset, which was highly right-skewed, indicating that while most students had relatively few absences, a small subset exhibited extreme absenteeism. This skew suggests that absenteeism may disproportionately affect a minority of students.

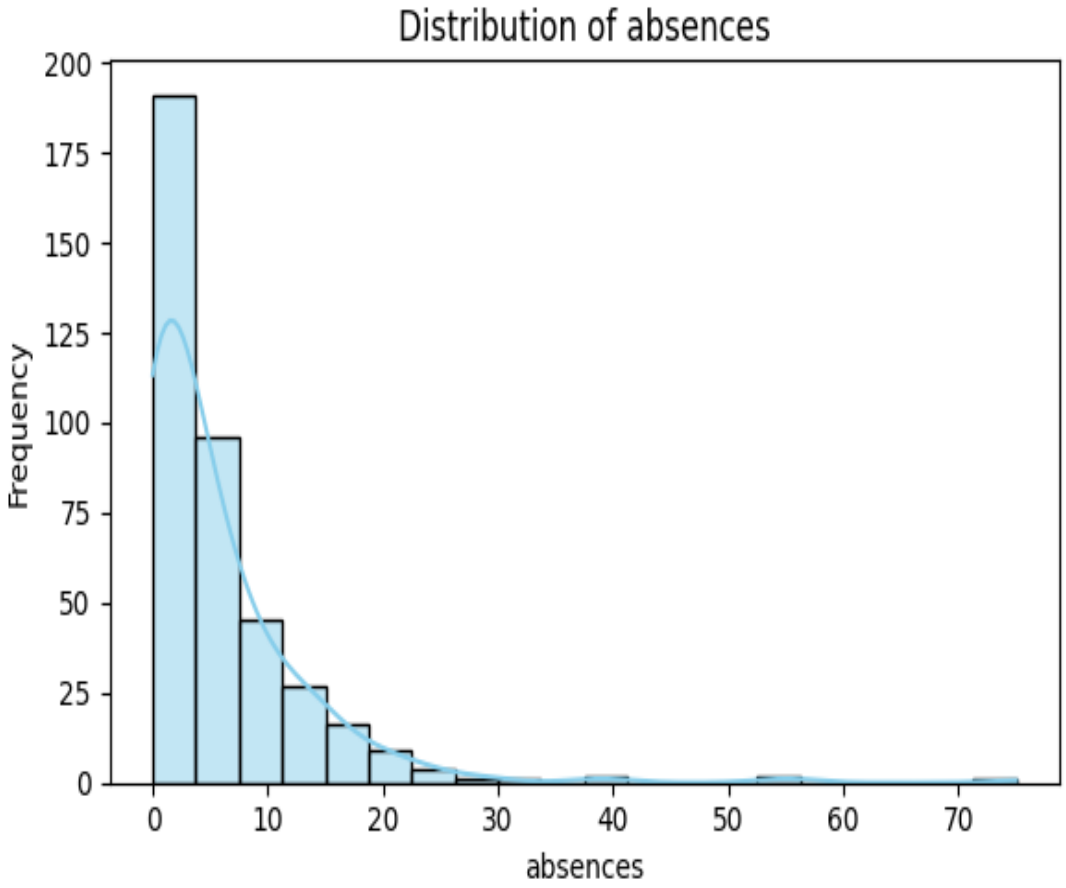

Figure 1. Distribution of absences in the mathematics dataset.

Figure 2 shows the Portuguese dataset's final grades (G3) distribution. Grades followed an approximately normal distribution, clustering 10–12 points out of 20. A non-trivial number of students scored zero, reflecting complete course failure and underscoring the necessity of early intervention strategies.

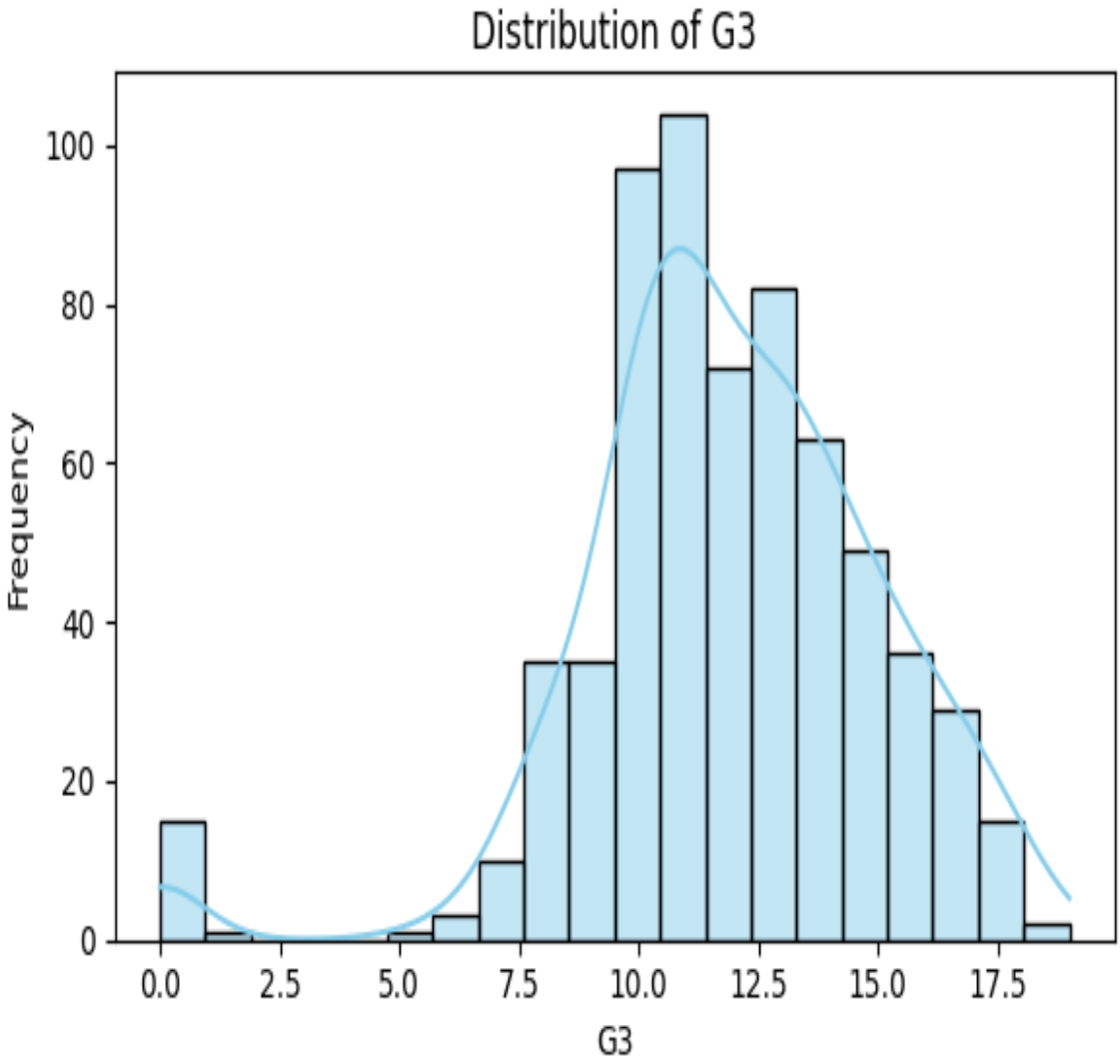

Figure 2. Distribution of final grades (G3) in the Portuguese dataset.

Correlation analysis was also performed to examine relationships among features. As shown in Figure 3, the Portuguese dataset's correlation heatmap confirmed that first- and second-period grades (G1 and G2) exhibited the strongest correlations with final grades (G3). At the same time, demographic and lifestyle features (e.g., alcohol consumption, family relations, travel time) displayed weak or negligible correlations. These findings are consistent with prior studies demonstrating the dominant role of earlier academic performance in predicting outcomes (Shorfuzzaman et al., 2021).

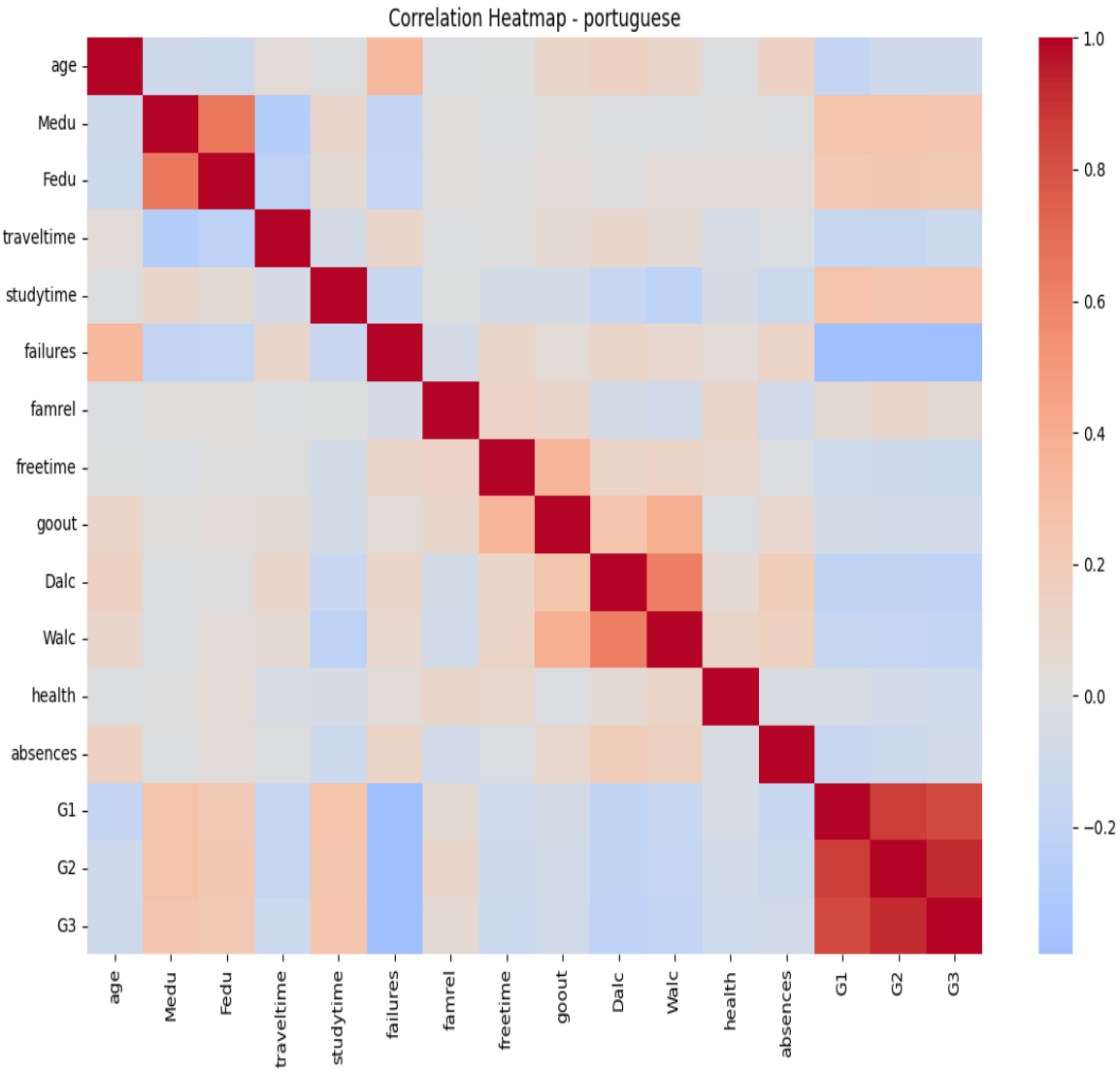

Figure 3. Correlation heatmap of the Portuguese dataset.

Model Evaluation

Two supervised learning algorithms, Linear Regression and Random Forest Regression, were applied to both datasets to assess predictive performance.

Mathematics Dataset.

As shown in Table 1, Linear Regression achieved an R^2 value of 0.034, indicating almost no explanatory power, alongside relatively high error values (RMSE = 4.91, MAE = 3.71). By contrast, Random Forest Regression achieved substantially stronger

performance with an R^2 of 0.672, RMSE of 2.15, and MAE of 1.65, reflecting superior ability to capture non-linear relationships.

Model	R^2	RMSE	MAE
Linear Regression	0.034	4.91	3.71
Random Forest	0.672	2.15	1.65

Table 1: Model evaluation metrics for the mathematics dataset.

Portuguese Dataset.

A similar pattern was observed in the Portuguese dataset (see Table 2). Linear Regression yielded a negative R^2 value (–0.016), suggesting the model performed worse than a mean-only baseline predictor. Random Forest Regression, however, achieved an R^2 of 0.701, with RMSE of 2.02 and MAE of 1.58, confirming its robustness for student performance prediction.

These findings reinforce the limitations of linear approaches when dealing with complex, categorical-heavy educational datasets (Panigrahi et al., 2021). Conversely, the strong performance of Random Forest models aligns with prior ensemble-based studies in educational data mining, where tree-based methods consistently outperformed linear baselines (Manhães et al., 2022).

Model	R^2	RMSE	MAE
Linear Regression	-0.016	4.82	3.69
Random Forest	0.701	2.02	1.58

Table 2: Model evaluation metrics for the Portuguese dataset.

Discussion

The comparative analysis clearly demonstrated that Random Forest Regression consistently outperformed Linear Regression in both the Mathematics and Portuguese datasets. This outcome underscores the advantage of ensemble learning methods for modeling heterogeneous and non-linear patterns inherent in educational datasets (Manhães et al., 2022). Most importantly, the results confirmed the study's hypothesis: prior academic performance, as captured by G1 and G2 scores, emerged as the strongest predictor of final grades (G3). This finding closely aligns with prior research emphasizing the predictive primacy of cumulative academic history in educational forecasting (Shorfuzzaman et al., 2021).

By contrast, lifestyle and demographic variables, including alcohol consumption, family relations, and health status, displayed only weak or negligible correlations with final grades. While these socio-behavioral attributes provide valuable contextual information about student life, their explanatory power in forecasting academic outcomes was limited. Such evidence is consistent with earlier studies that found non-academic factors to be secondary influences in grade prediction (Al-Barrak & Al-Razgan, 2020).

These findings reinforce the broader academic consensus that embedding advanced machine learning models into institutional risk assessment frameworks can substantially improve the early identification of at-risk students. Proactive detection allows for designing and implementing targeted interventions that can mitigate underperformance and reduce dropout rates (Panigrahi et al., 2021).

Nevertheless, certain limitations must be acknowledged. The study relied on data from Portuguese schools during a specific period, which may limit the generalizability of

results to other cultural or institutional contexts. Furthermore, the static nature of the dataset excludes dynamic variables, such as real-time engagement metrics or adaptive learning behaviors, that could enhance predictive performance. Therefore, Future research should expand to cross-institutional datasets, incorporate richer behavioral variables, and explore hybrid deep learning architectures to refine prediction accuracy further and extend applicability.

Conclusion

This study offered robust evidence that Random Forest Regression substantially outperforms Linear Regression in predicting student performance using the UCI Student Performance dataset (Cortez & Silva, 2008). Consistent with the study's central hypothesis, prior academic performance, specifically G1 and G2, emerged as the most influential determinant of final grades (G3). Additional variables such as absences and prior course failures contributed to prediction accuracy, though their impact was comparatively modest. These results underscore the capacity of ensemble models to capture the complex, non-linear relationships that characterize educational data and to yield actionable insights for institutional decision-making (Manhães et al., 2022).

The findings strongly advocate for integrating data-driven predictive methodologies within educational systems. Embedding machine learning models into institutional decision-making frameworks enables more strategic resource allocation, targeted intervention design, and improved student support mechanisms. Such proactive measures can potentially reduce dropout rates while enhancing the overall quality of educational delivery (Shorfuzzaman et al., 2021).

Ultimately, this project validates the hypothesis that machine learning can accurately forecast student performance and provides a replicable framework for extending predictive analytics to broader educational contexts. Future work should expand upon these findings by incorporating dynamic behavioral data, cross-institutional datasets, and advanced hybrid modeling techniques to enhance predictive accuracy and applicability further.

References

Al-Barrak, M. A., & Al-Razgan, M. (2020). Predicting students' performance through classification: A case study. *International Journal of Advanced Computer Science and Applications, 11*(1), 1–7. https://doi.org/10.14569/IJACSA.2020.0110101

Batte, B. (2025). *Student grade prediction [Computer software]*. GitHub. https://github.com/BenjaminBatte/student-grade-prediction

Cortez, P., & Silva, A. M. G. (2008). *Student performance dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/320/student+performance

Kumari, A., & Singh, V. (2023). Comparative analysis of machine learning techniques for student performance prediction. *International Journal of Emerging Technologies in Learning, 18*(4), 22–34. https://doi.org/10.3991/ijet.v18i04.35575

Manhães, L. M. B., Souza, L. A., Moreira, E. D., & Rocha, Á. R. (2022). Using ensemble learning to predict academic performance. *Education and Information Technologies, 27*(3), 3259–3278. https://doi.org/10.1007/s10639-021-10733-4

McKinney, W. (2022). *Python for data analysis* (3rd ed.). O'Reilly Media.

Panigrahi, S., Srivastava, P. R., & Sharma, S. (2021). Predictive modeling in education: A review of trends, techniques, and applications. *Education and Information Technologies, 26*(1), 725–749. https://doi.org/10.1007/s10639-020-10254-6

Sasikala, S., & Renuka Devi, D. (2023). *Research practitioner's handbook on big data analytics*. Taylor & Francis.

Shorfuzzaman, M., Hossain, M. S., & Nazir, A. (2021). Performance prediction of students using machine learning algorithms: A case study. *Computers, 10*(8), 93. https://doi.org/10.3390/computers10080093

Appendix

This section provides additional exploratory data analysis (EDA) figures for the Mathematics and Portuguese datasets.

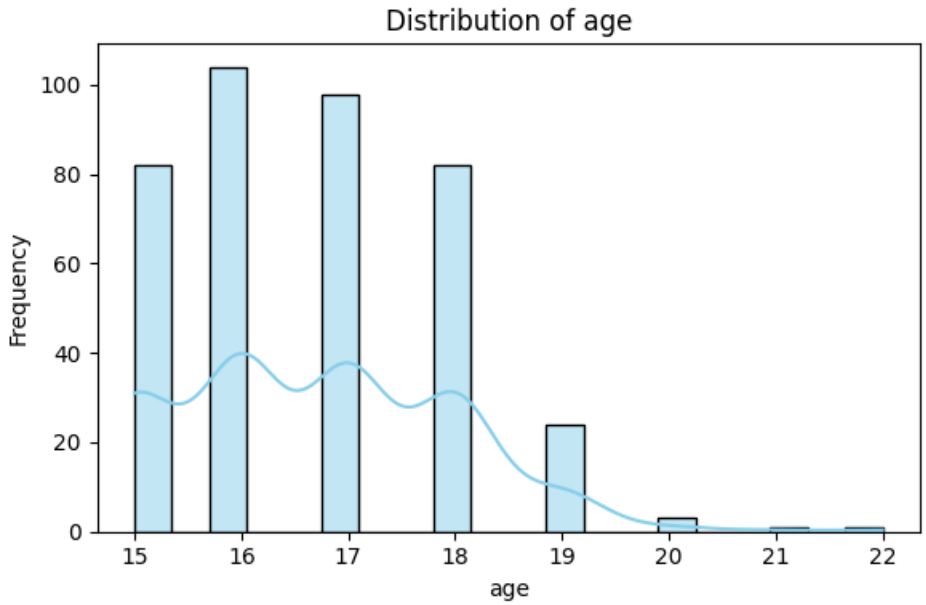

Figure A1—age distribution in the Mathematics dataset.

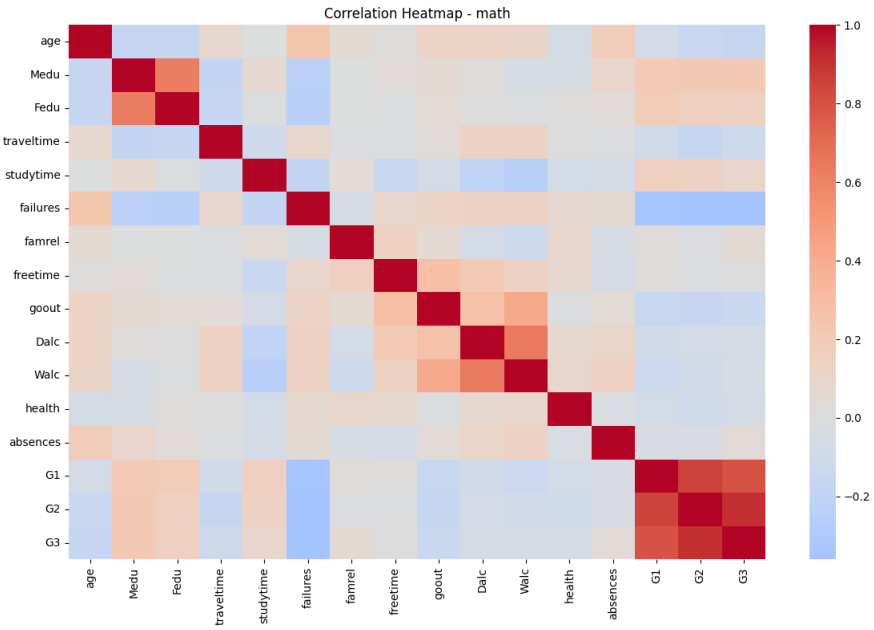

Figure A2. Correlation heatmap for the Mathematics dataset.

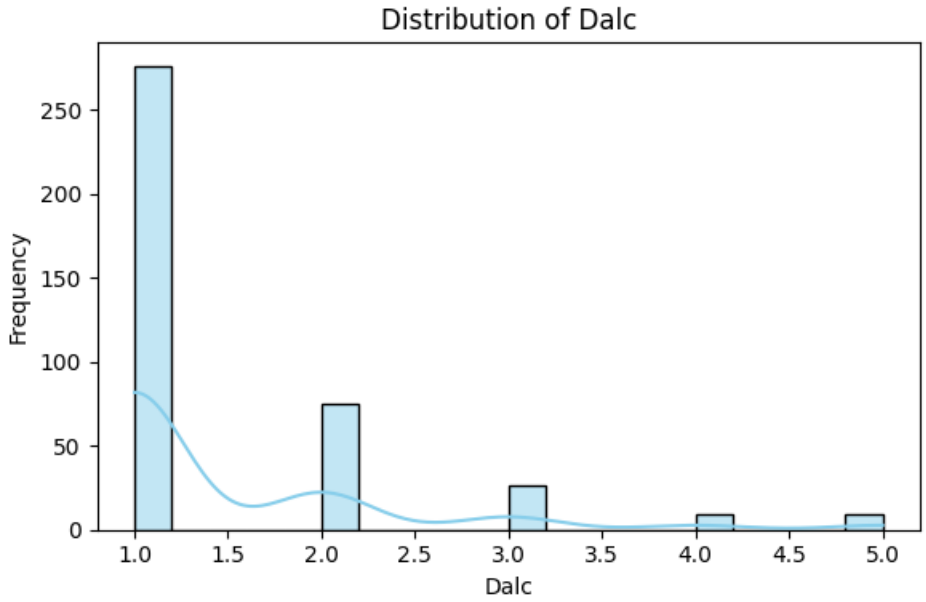

Figure A3. Weekday alcohol consumption distribution (Math).

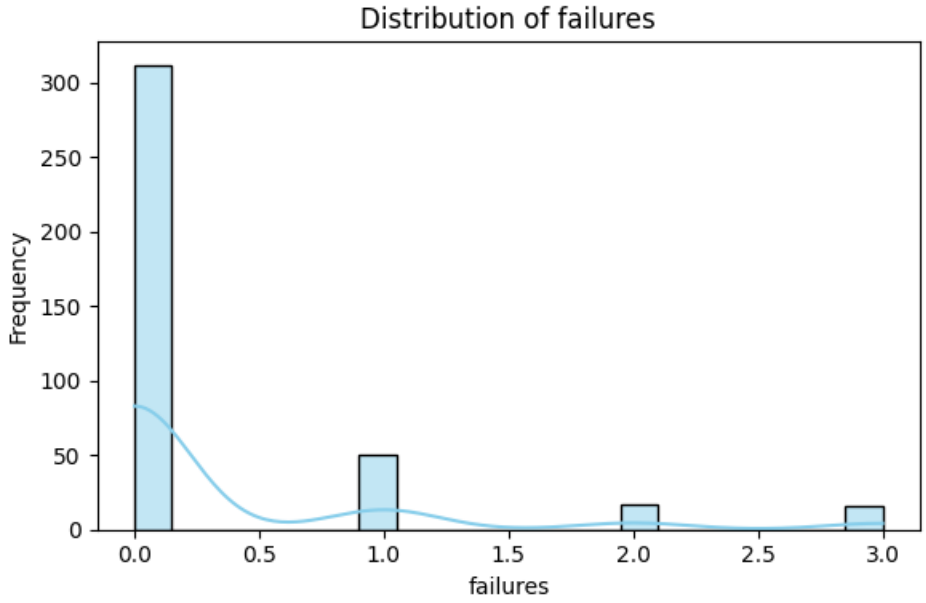

Figure A4. Distribution of prior class failures (Math).

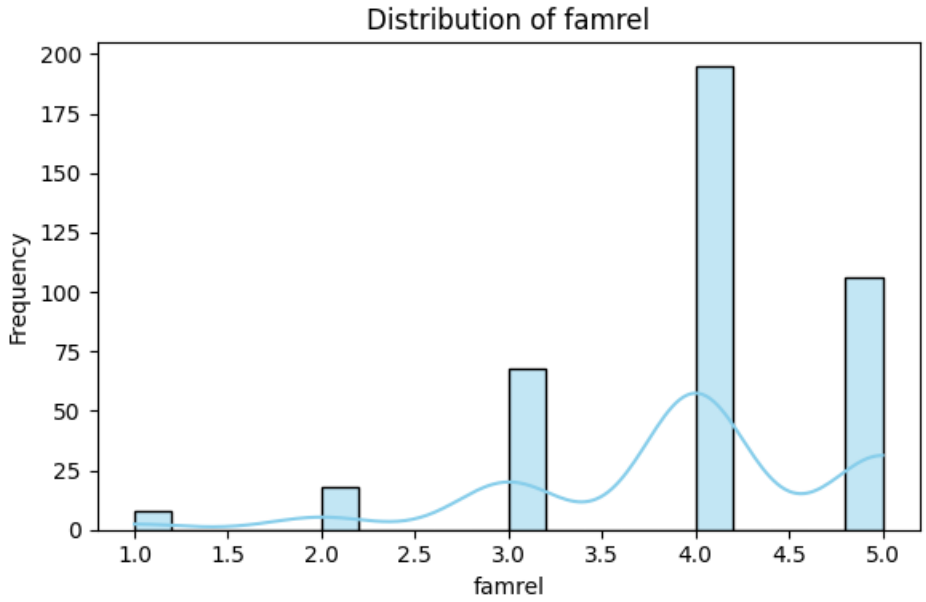

Figure A5. Family relationship quality distribution (Math).

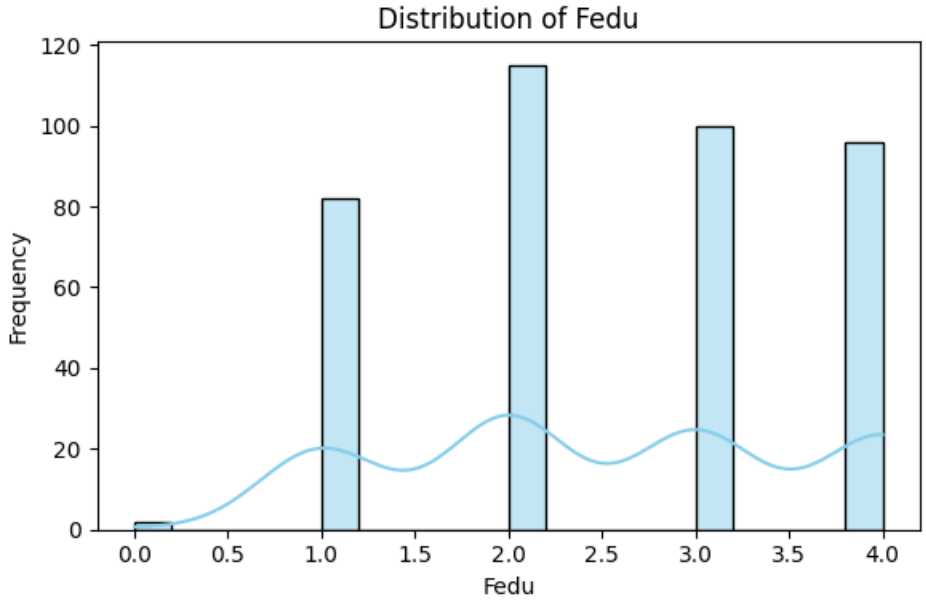

Figure A6. Father's education distribution (Math).

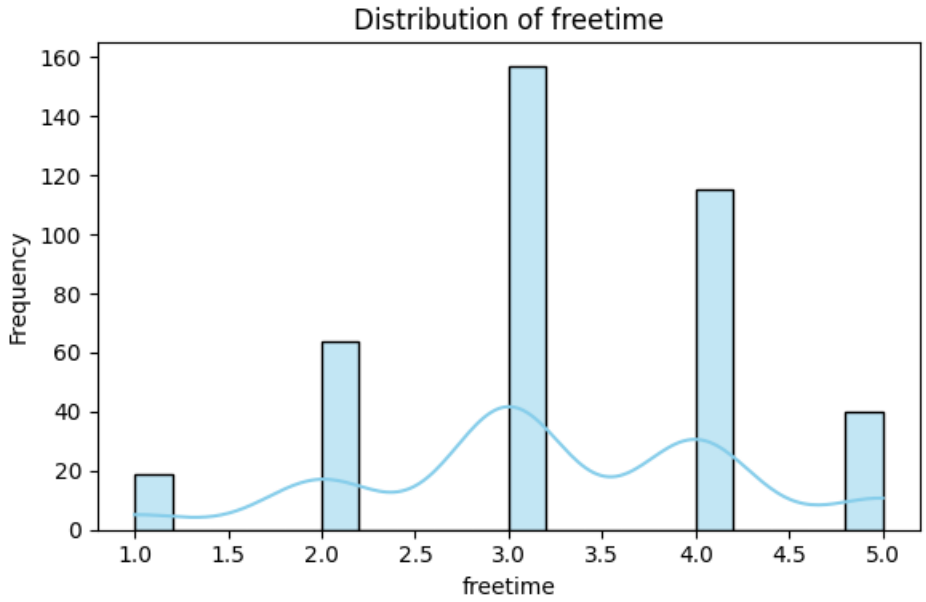

Figure A7. Free time distribution (Math).

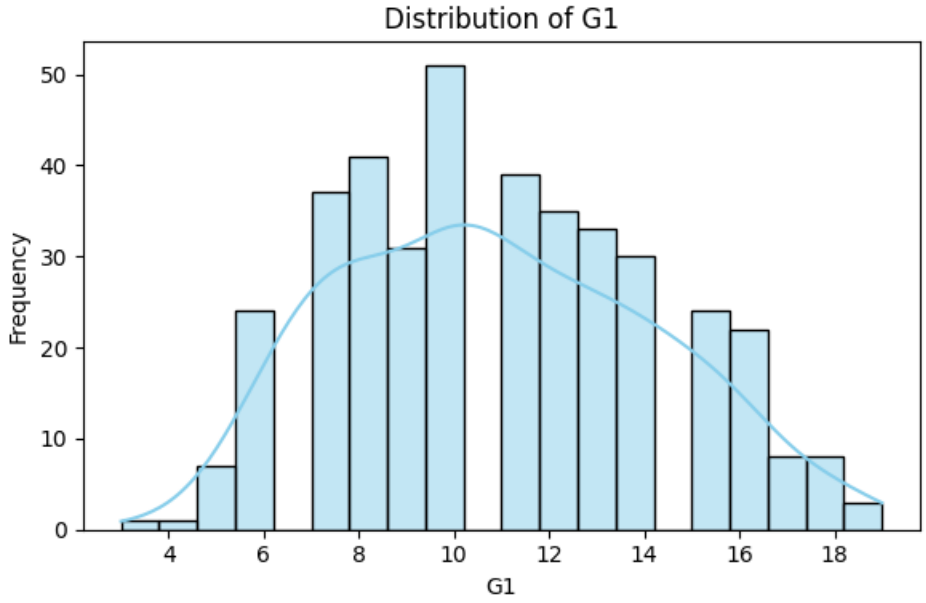

Figure A8. First period grade distribution (Math).

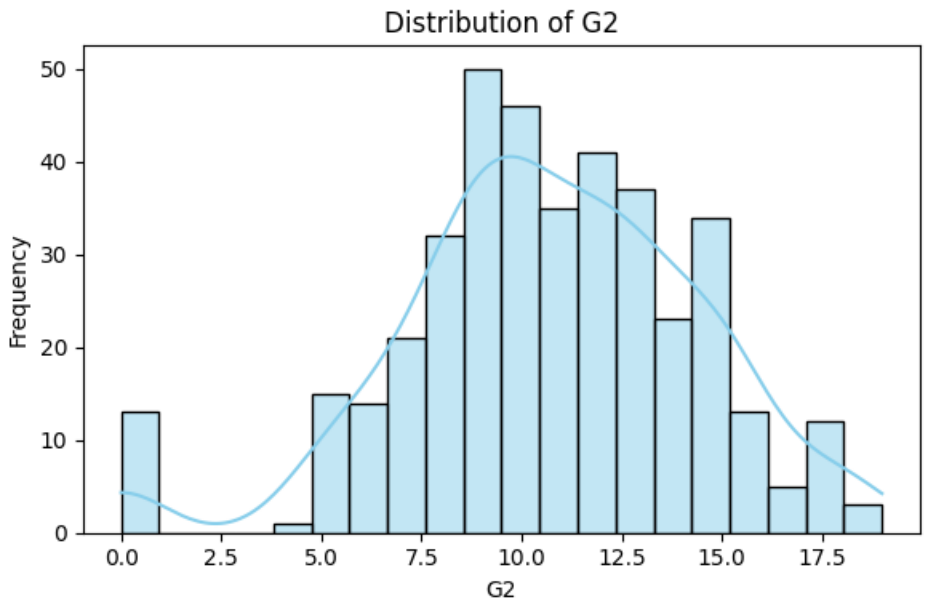

Figure A9. Second period grade distribution (Math).

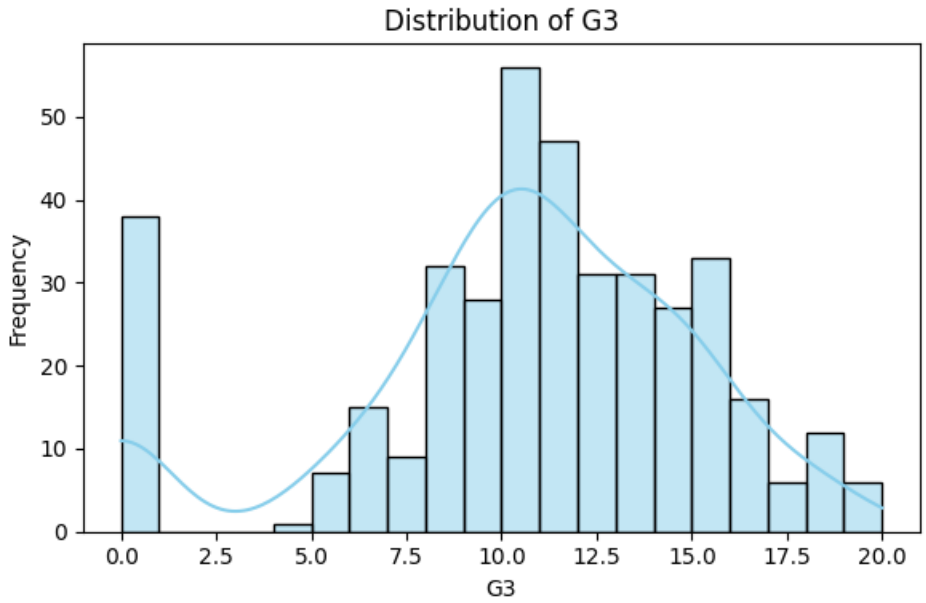

Figure A10. Final grade distribution (Math).

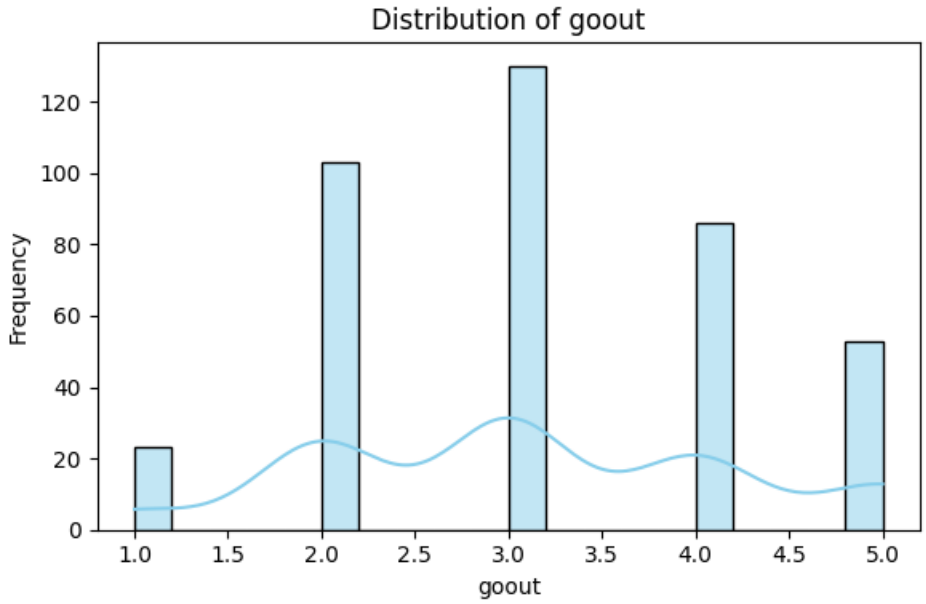

Figure A11. Going out with friends distribution (Math).

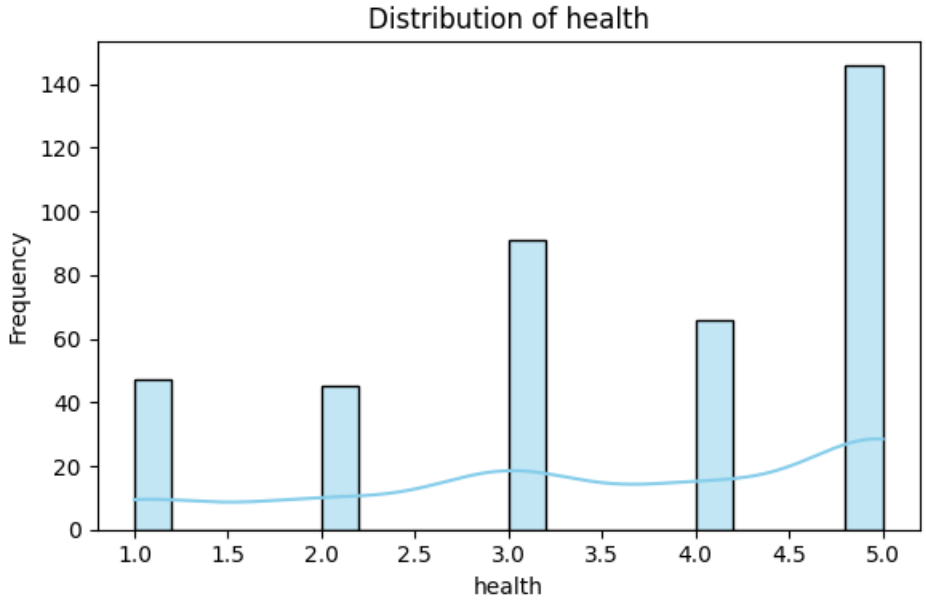

Figure A12. Health status distribution (Math).

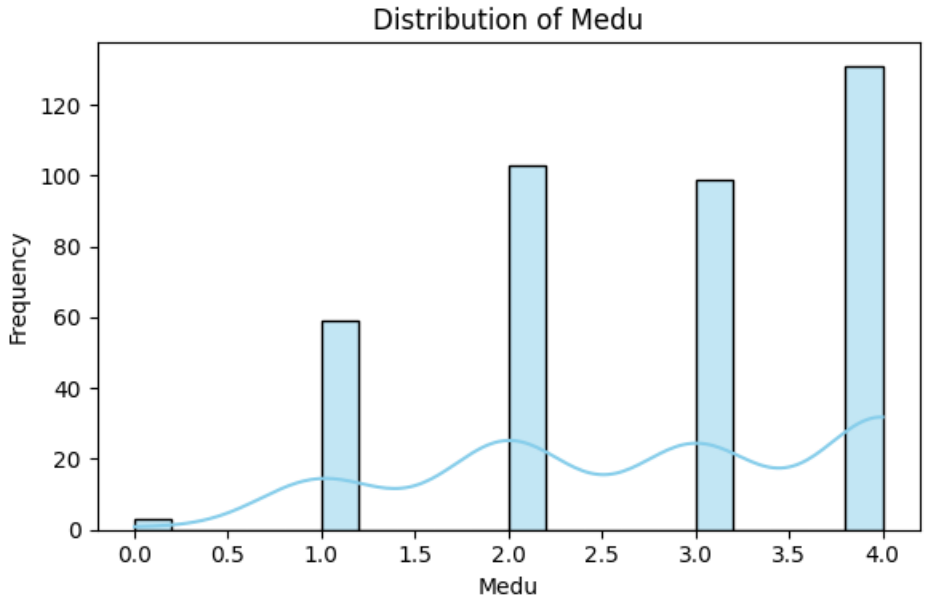

Figure A13. Mother's education distribution (Math).

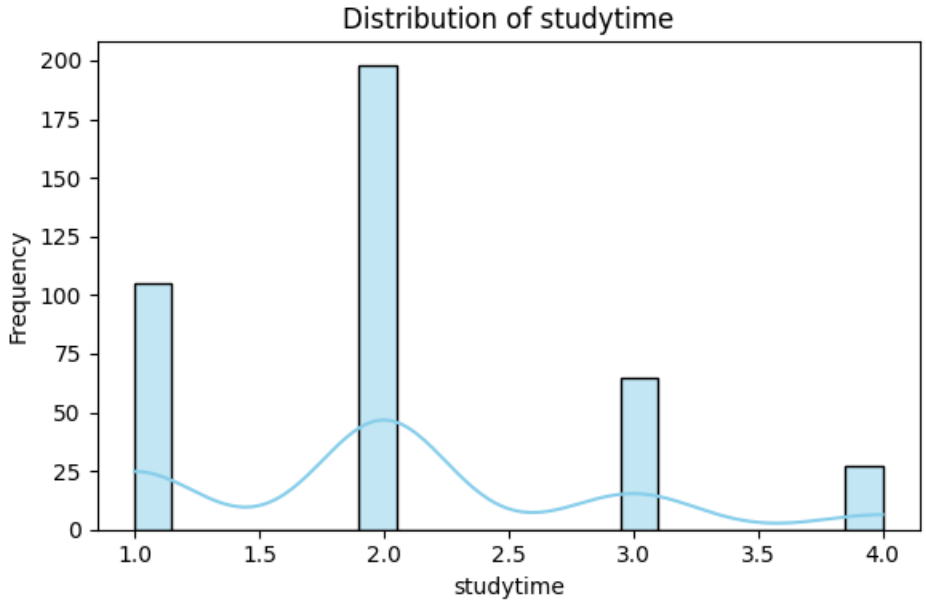

Figure A14. Study time distribution (Math).

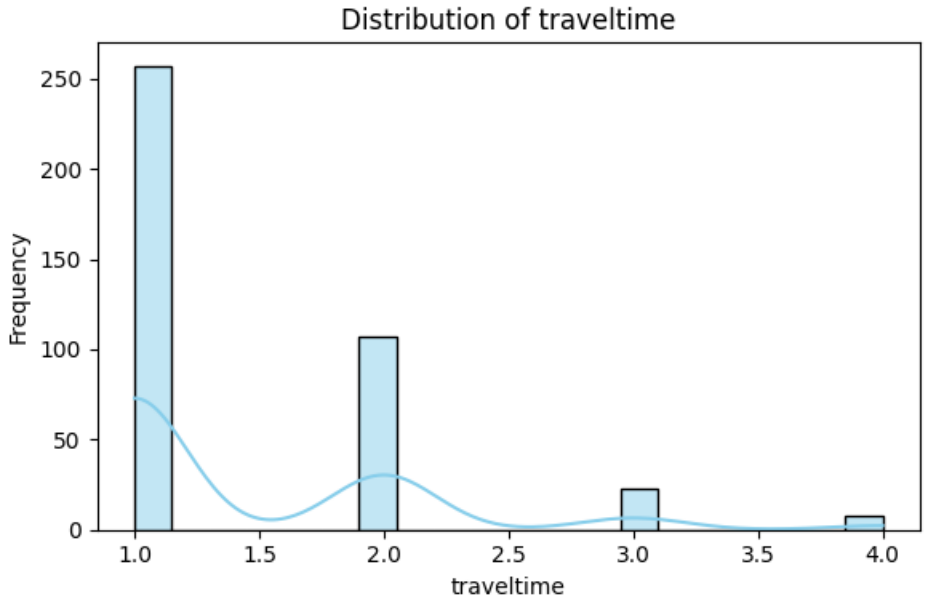

Figure A15. Travel time distribution (Math).

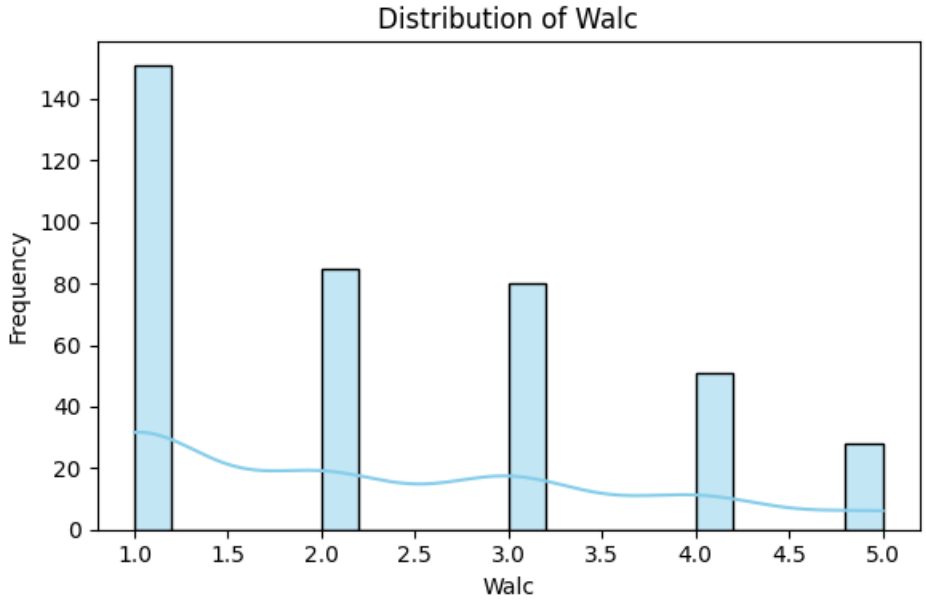

Figure A16. Weekend alcohol consumption distribution (Math).

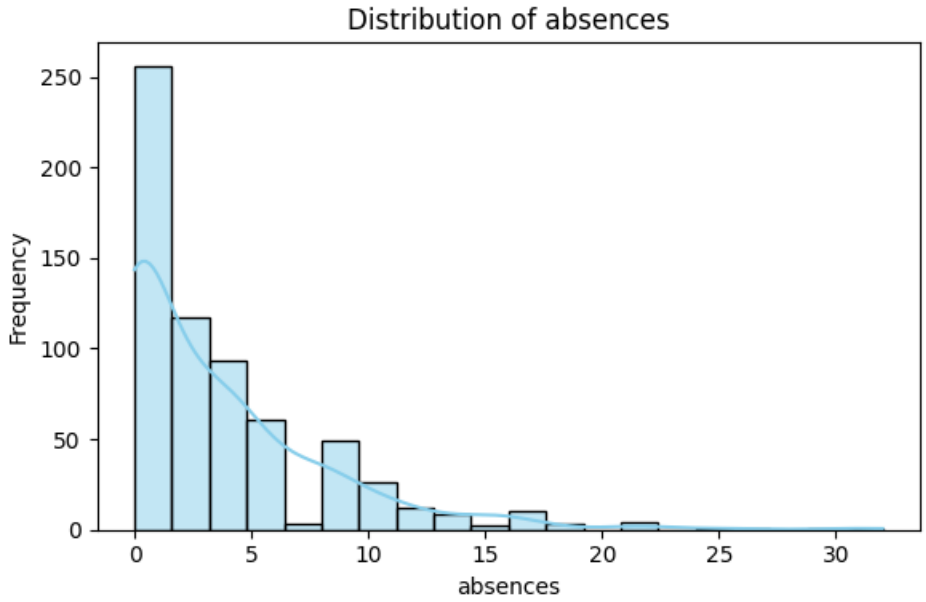

Figure A17. Absences distribution (Portuguese).

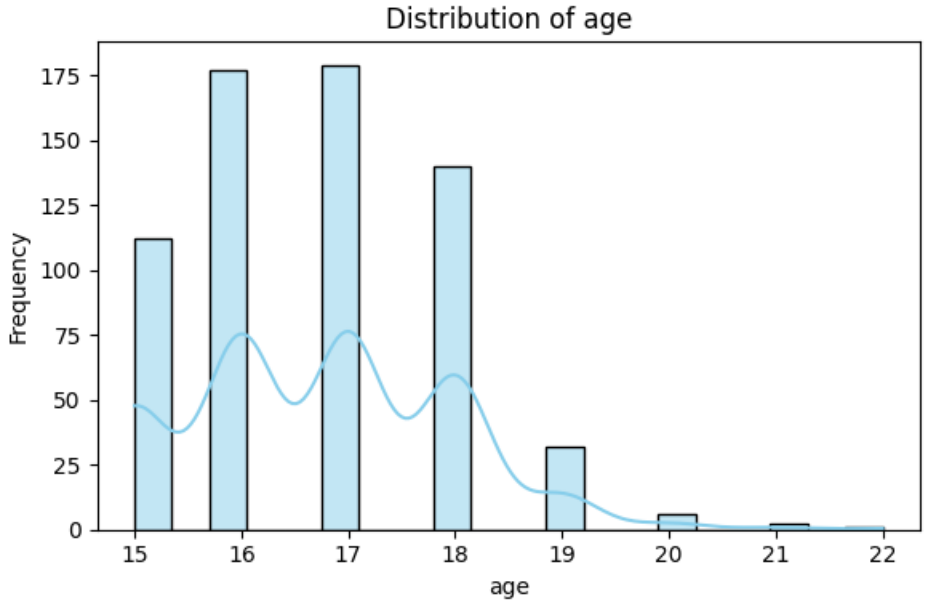

Figure A18. Age distribution (Portuguese).

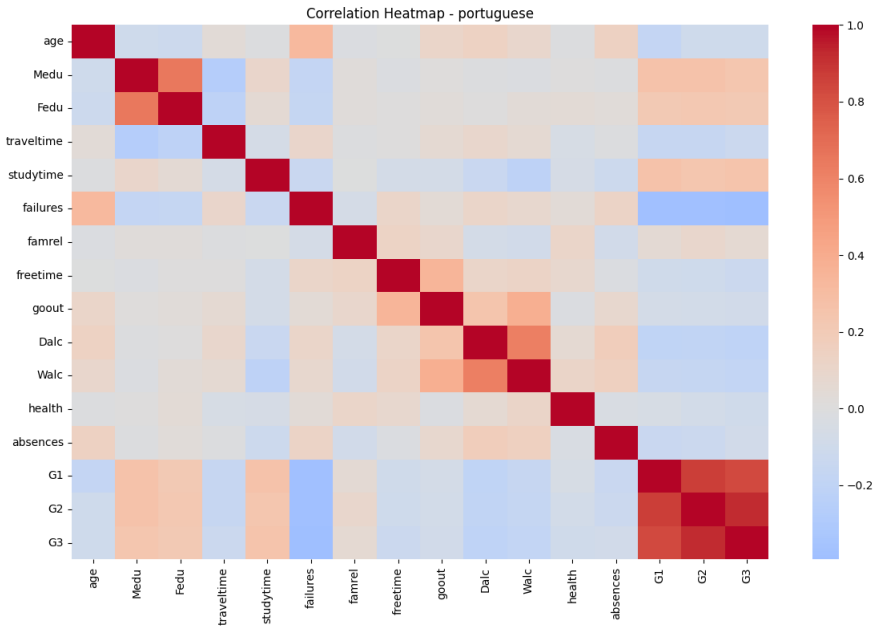

Figure A19. Correlation heatmap (Portuguese).

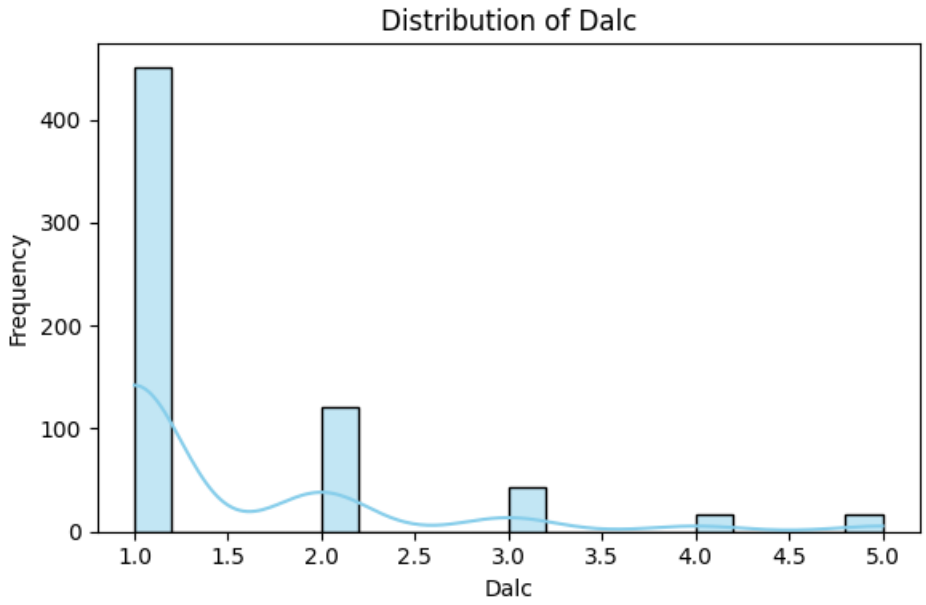

Figure A20. Weekday alcohol consumption distribution (Portuguese).

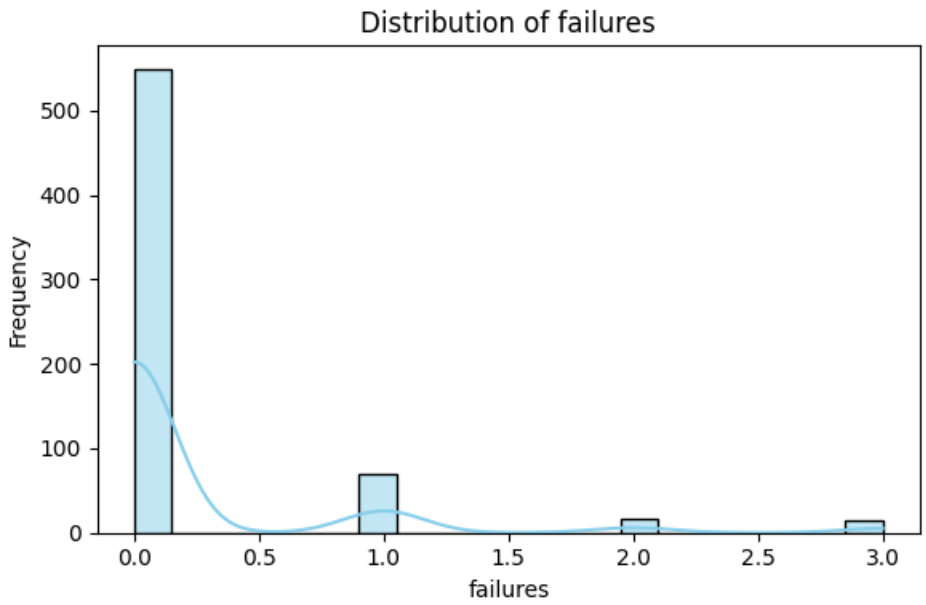

Figure A21. Distribution of prior class failures (Portuguese).

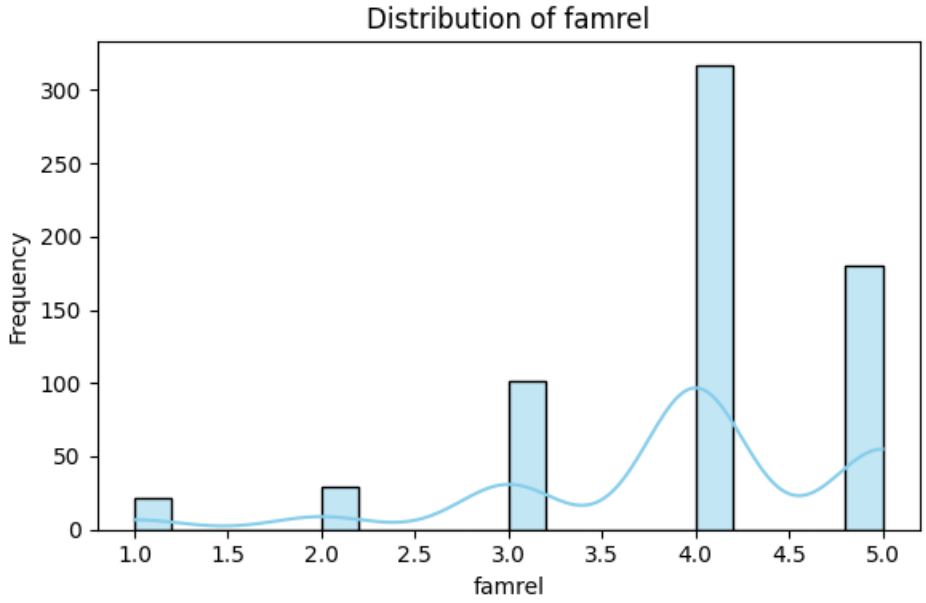

Figure A22. Family relationship quality distribution (Portuguese).

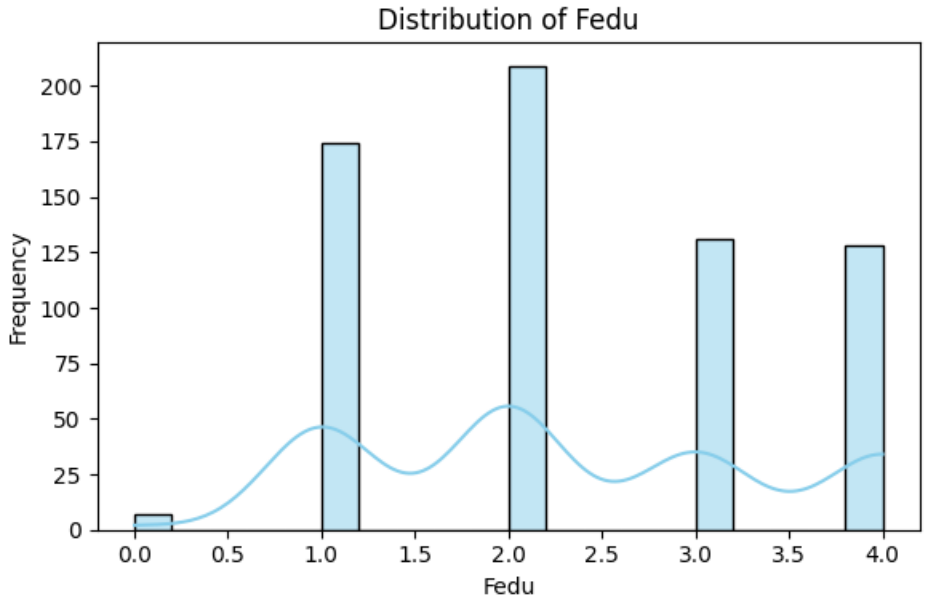

Figure A23. Father's education distribution (Portuguese).

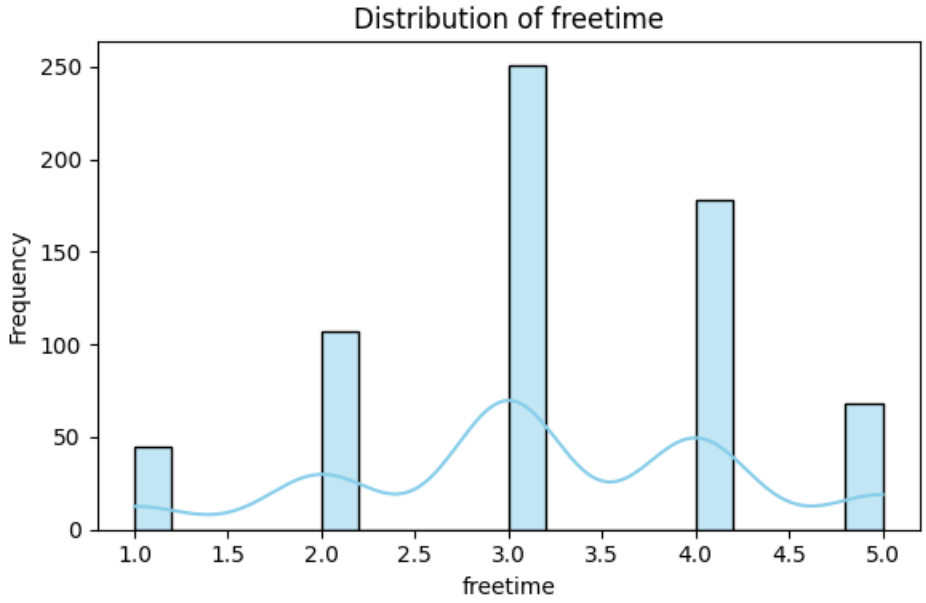

Figure A24. Free time distribution (Portuguese).

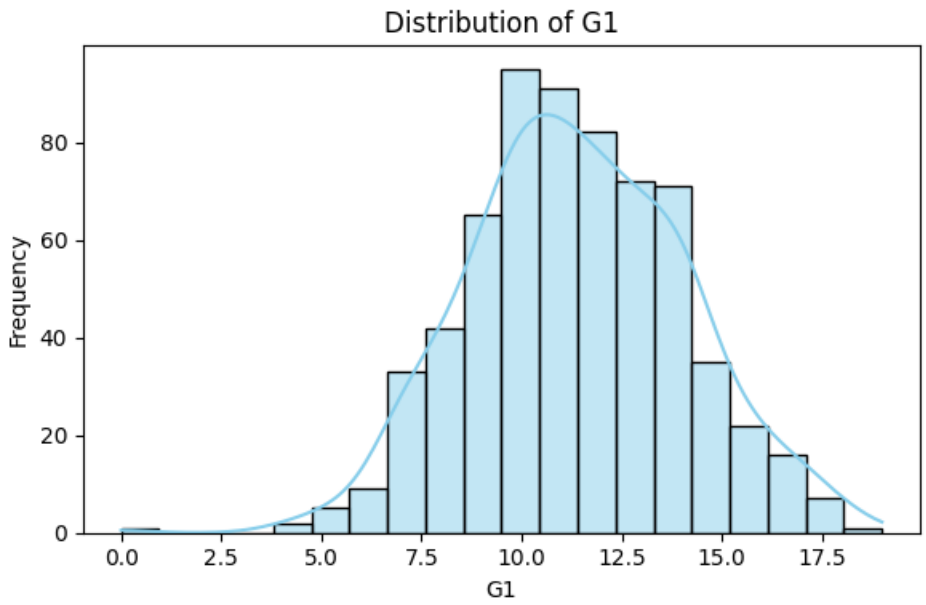

Figure A25. First period grade distribution (Portuguese).

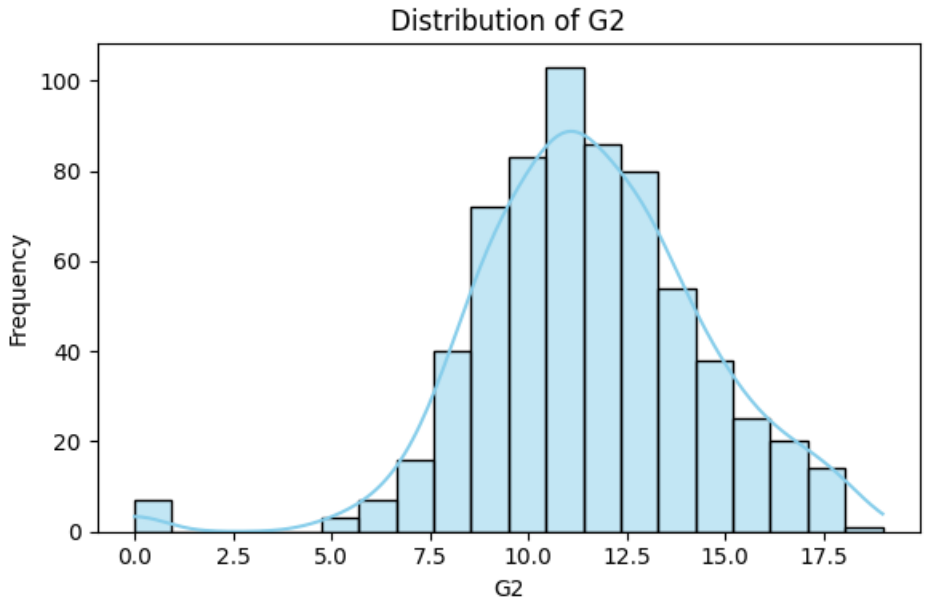

Figure A26. Second period grade distribution (Portuguese).

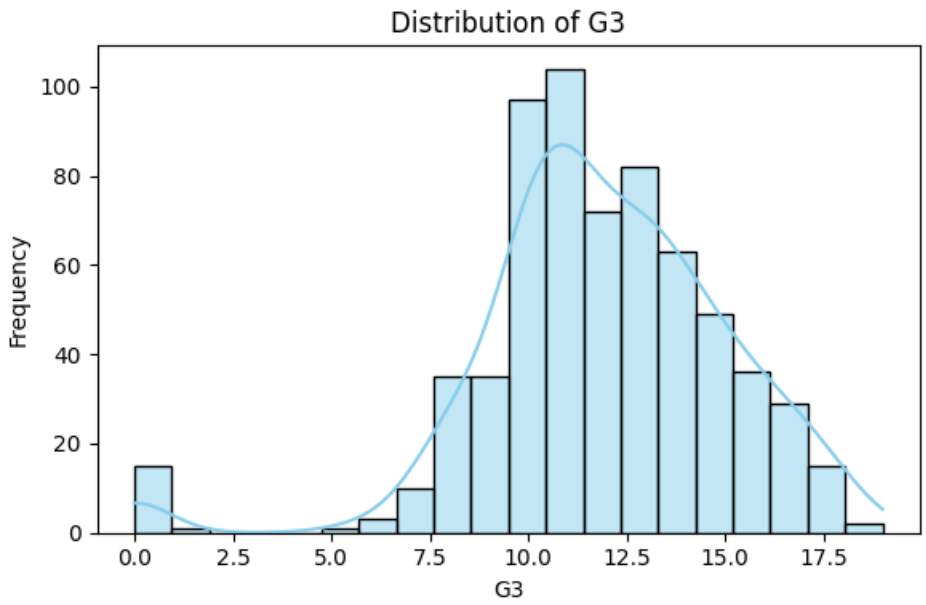

Figure A27. Final grade distribution (Portuguese).

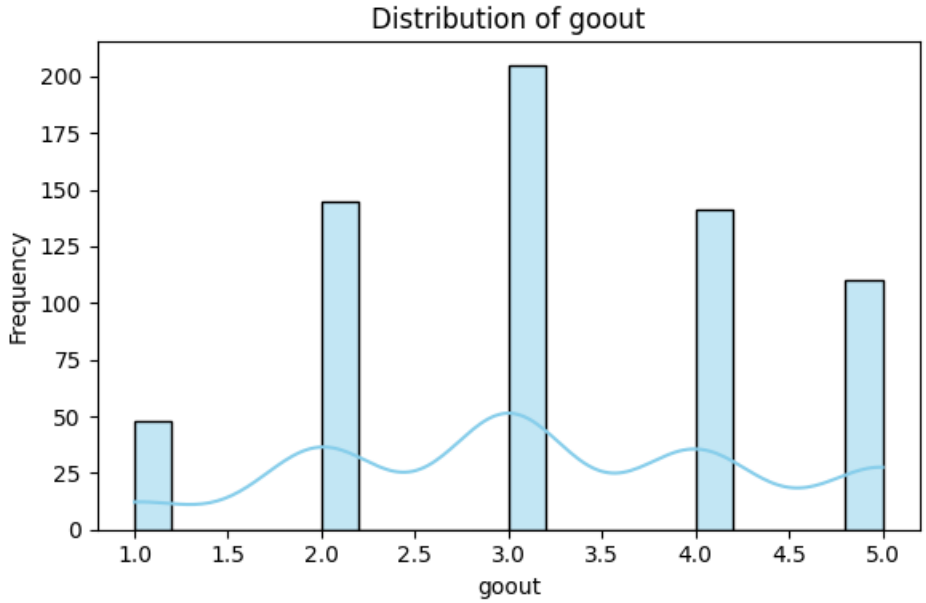

Figure A28. Going out with friends distribution (Portuguese).

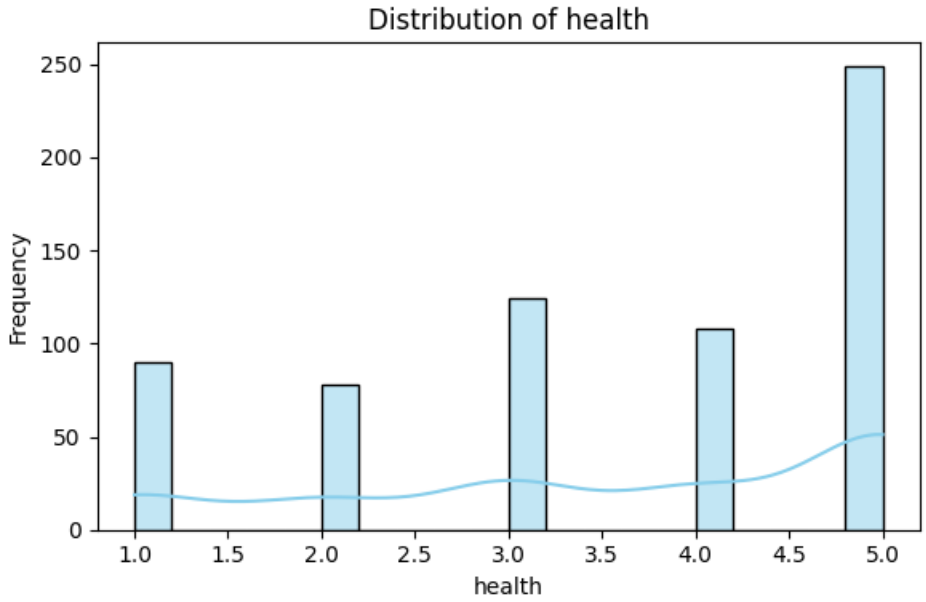

Figure A29. Health status distribution (Portuguese).

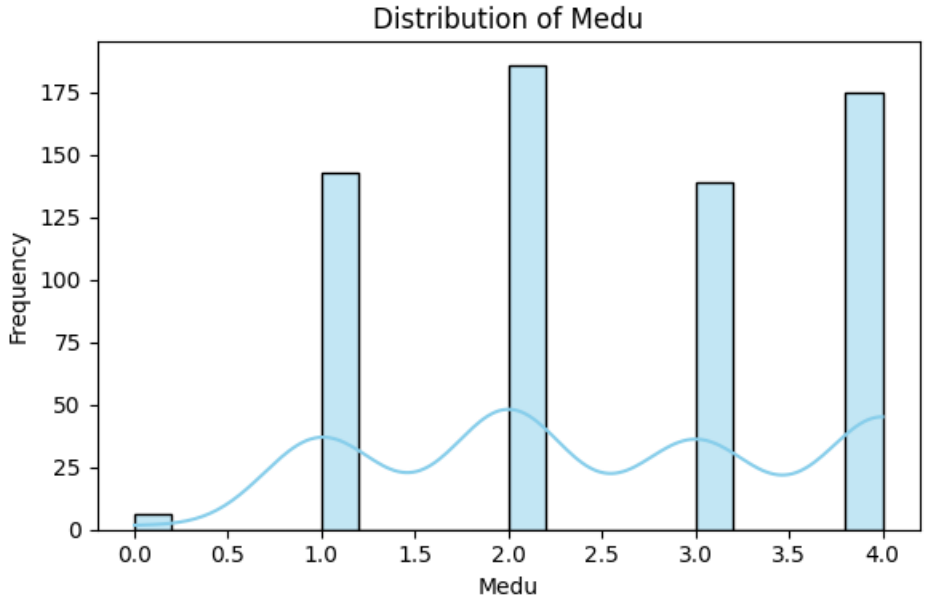

Figure A30. Mother's education distribution (Portuguese).

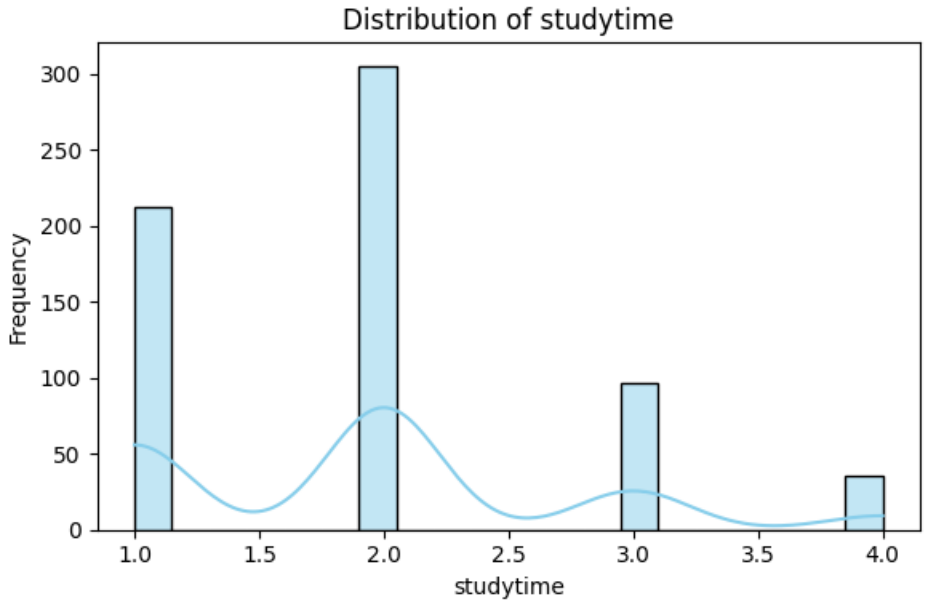

Figure A31. Study time distribution (Portuguese).

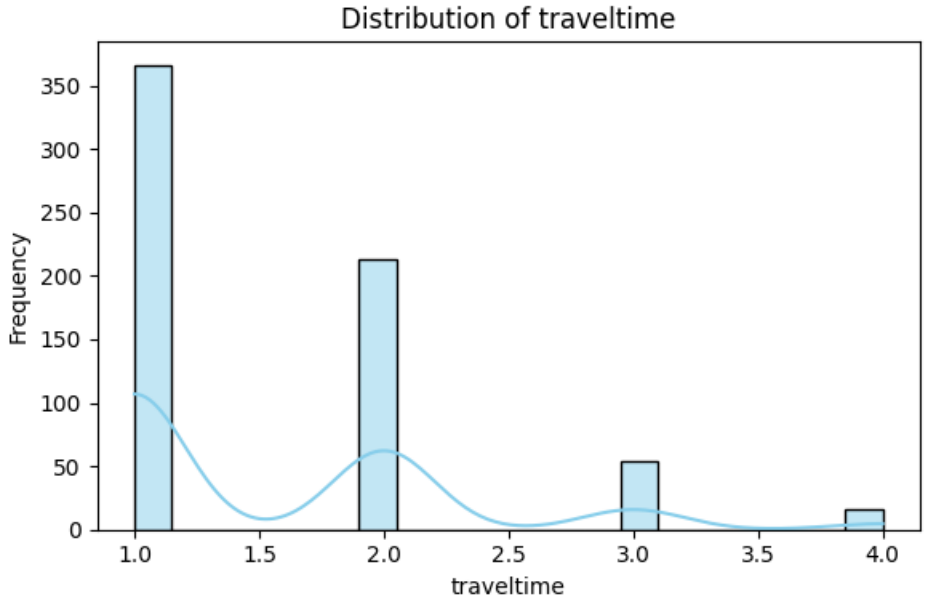

Figure A32. Travel time distribution (Portuguese).

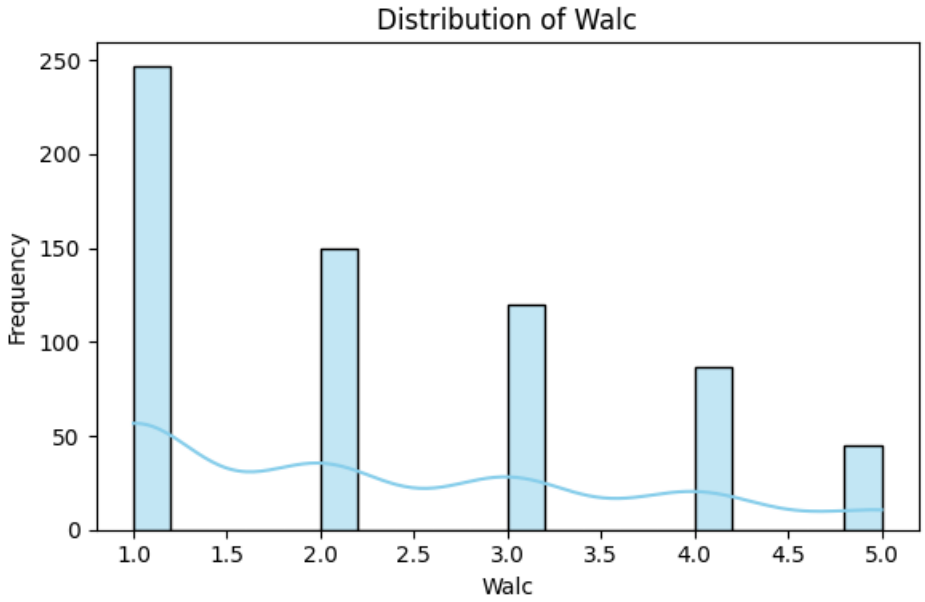

Figure A33. Weekend alcohol consumption distribution (Portuguese).