Project Report: Data Cleaning and Feature Identification

1. Introduction

The project explores the various methods of handling missing values and their implications on the success of various machine learning based prediction models. It also focuses on analyzing feature relationships in a structured dataset. Emphasis is placed on using strategies such as feature engineering and cross-validation techniques (like stratified K-folding) for improving the accuracy of the model.

NOTE: More details of each approach and experimentations are written in the Jupyter notebook.

2. Data Preprocessing

2.1 Missing Value Analysis

- Initial inspection revealed missing values in both categorical and numerical columns.
- Mean and KNN Imputer were used for numerical data, with similar outcomes due to limited missing values.
- Mode was used to fill missing values in categorical columns.

2.2 Categorical to Numerical Encoding

- One-hot encoding was avoided due to potential issues with distance calculations in KNN.(explained in Jupyter notebook)
- Used pd.get_dummies for binary encoding suitable for distance-based algorithms.

3. Feature Analysis

3.1 Correlation Heatmap of all features

- Used to identify relationships among features.
- Helped decode what Feature_1,2,3 were(tentatively)

3.2 Correlation heatmap of column encoded features with romantic_yes

3.3 Boxplots of Mother’s and Father’s education level with final grade

3.4 Violin plot of ages on the vertical axis, males to the right females to the left of the violin. And percentage of students of that certain age and gender plotted on the horizontal axis.

3.5 Violin plots linking parents’ education, final grade(G3) and wanting higher education

3.6 Boxplot linking final grade and involvement in extracurricular activites

3.7 Boxplot linking alcohol consumption and grades.

4. Prediction Algorithms

Three algorithms were used, namely; KNN, RandomForestClassifier and Logistic Regression.

All these gave similar accuracies.(around 62%)

Feature engineering bumped this up to around 66

K folding further increased it to around 73.

5. Tools Used

- Jupyter Notebook

- Python libraries: pandas, numpy, sklearn, seaborn, matplotlib

Appendix

- Full code available in `task1.ipynb` notebook.