Predicting Higher Education Enrollment Based on Parental Information and Drinking Patterns

Bidone Federico 892054
Passera Thomas 901685
Simeoni Matteo 918544

May 2024

Abstract

This project aims to predict whether students will pursue higher education after high school based on information about their parents and their drinking patterns. Using a dataset provided by Kaggle, the project implements various preprocessing and modeling techniques within the KNIME analytics platform. The analysis involves converting categorical variables, normalizing data, feature selection, and comparing multiple classification models' performance.

Keywords: Machine Learning, Higher Education Prediction, Data Preprocessing

1 Introduction

The decision of students to pursue higher education is influenced by a multitude of factors, including parental background and personal habits. Research has shown that parental choices and behaviors significantly impact children's educational outcomes and overall well-being.

One of the most prominent influences is parental substance use. Studies indicate that children of parents who consume alcohol heavily are more likely to adopt similar drinking patterns themselves, which can adversely affect their academic performance and future aspirations (Christoffersen and Soothill, 2003). The intergenerational transmission of drinking behaviors is well-documented, suggesting that parental habits can set a precedent for children's behaviors (Latendresse et al., 2008).

Furthermore, children who grow up in households where alcohol abuse is prevalent often face additional challenges, such as emotional neglect, inconsistent discipline, and financial instability, which can hinder their educational attainment (Johnson and Leff, 1999). The stress and turmoil in such environments can lead to lower academic achievement and a reduced likelihood of pursuing higher education.

Parental education also plays a crucial role in shaping children's educational trajectories. Parents with higher levels of education are more likely to value and prioritize academic success, providing a supportive environment that fosters learning and growth (Davis-Kean, 2005). Conversely, children whose parents have lower educational attainment may lack

the necessary support and encouragement to excel academically and pursue further education.

The proverb "the faults of the parents fall on the children" aptly summarizes the profound impact of parental behavior on children's futures. It underscores the importance of addressing and mitigating negative parental influences to break the cycle of disadvantage and promote better educational outcomes for the next generation.

In this project, we analyze a dataset containing student information from two secondary schools to predict their higher education enrollment based on parental background and personal habits, particularly focusing on drinking patterns. By employing machine learning techniques, we aim to provide insights into how these factors interact and influence students' educational paths.

2 Dataset and Preprocessing

The dataset used for this analysis provides comprehensive insights into the socioeducational factors influencing student behavior, particularly focusing on alcohol consumption and academic performance in high school. Collected through surveys conducted among high school students, the dataset includes various social, educational, and demographic features. This data was originally gathered by P. Cortez and A. Silva for the study "Using Data Mining to Predict High School Student Performance," presented at the 5th Future Business Technology Conference (FUBUTEC, 2008) in Porto, Portugal.

2.1 Data Exploration

The dataset comprises two CSV files, "student-mat.csv" and "student-por.csv", representing students' academic records from two Portuguese schools. During the exploring the data, we noticed that the mathematics dataset is contained in the por dataset. We therefore decided to use the por dataset

- **school**: Student's school (GP or MS)
- **sex**: Student's sex (F or M)
- **age**: Student's age (15 to 22)
- **address**: Type of student's residential address (U - urban or R - rural)
- **famsize**: Family size (LE3 - less than or equal to 3 or GT3 - greater than 3)
- **Pstatus**: Parents' cohabitation status (T - living together or A - separated)
- **Medu**: Mother's education level (0 - none, 1 - elementary, 2 - middle school, 3 - high school, 4 - higher education)
- **Fedu**: Father's education level (0 - none, 1 - elementary, 2 - middle school, 3 - high school, 4 - higher education)
- **Mjob**: Mother's job (teacher, health, services, at_home, other)
- **Fjob**: Father's job (teacher, health, services, at_home, other)
- **reason**: Reason for choosing this school (home, reputation, course, other)
- **guardian**: Student's guardian (mother, father, other)
- **traveltime**: Travel time from home to school ($1- <$ 15 min, $2-$ 15 to 30 min, $3-$ 30 min to 1 hour, $4 >$ 10hour)
- **studytime**: Weekly study time ($1- <$ 2hours, $2-$ 2 to 5 hours, $3-$ 5 to 10 hours, $4- >$ 10 hours)
- **schoolsup**: Extra educational support (yes or no)

- **famsup**: Family educational support (yes or no)
- **paid**: Private classes (yes or no)
- **activities**: Extracurricular activities (yes or no)
- **nursery**: Attended daycare (yes or no)
- **higher**: Desire to pursue higher education (yes or no)
- **internet**: Internet access at home (yes or no)
- **romantic**: Romantic relationship status (yes or no)
- **famrel**: Quality of family relationships (1 - very bad to 5 - excellent)
- **freetime**: Free time after school (1 - very low to 5 - very high)
- **goout**: Time spent with friends (1 - very low to 5 - very high)
- **Dalc**: Workday alcohol consumption (1 - very low to 5 - very high)
- **Walc**: Weekend alcohol consumption (1 - very low to 5 - very high)
- **health**: Current health status (1 - very bad to 5 - very good)
- **absences**: Number of school absences (0 to 93)
- **G1**: First semester grade (0 to 20)
- **G2**: Second semester grade (0 to 20)

2.2 Data Transformation

The preprocessing steps included:

- **Converting categorical variables**: Transform binary variables such as 'yes' and 'no' into 1 and 0, respectively.
- **One-hot encoding**: Apply one-hot encoding to nominal variables to create dummy variables for each category.
- **Normalization**: Normalize numerical variables to a range of 0 to 1.
- **Conversion**: Convert string variables to integers where applicable.
- **Column removal**: Remove irrelevant columns to focus on the most impactful features.
- **Data formatting**: Ensure all variables are in double format to maintain consistency.
- **Dataset partitioning**: Partition the dataset into 80% training and 20% testing sets, using stratified sampling based on the target variable "higher".

2.3 Data Transformation

The preprocessing steps included:

- Converting categorical variables: "yes" to 1 and "no" to 0.
- Applying one-hot encoding to transform categorical variables into dummy variables.
- Normalizing numerical variables to a range of 0 to 1.
- Converting string variables to integers where applicable.
- Removing irrelevant columns.
- Ensuring all variables are in double format.
- Partitioning the dataset into 80% training and 20% testing sets, with stratified sampling on the target variable "higher".

3 Modeling

The project compares eight different classification models:

- **J48**: J48 is an implementation of the C4.5 algorithm in WEKA, a decision tree classifier that splits the dataset based on attribute values to create a tree structure, used for both classification and regression tasks. It is simple to understand and interpret but may be prone to overfitting.

- **Multilayer Perceptron (MLP)**: MLP is a type of artificial neural network with one or more hidden layers between the input and output layers. It uses backpropagation for training and is capable of capturing complex patterns in the data, making it suitable for various prediction tasks.

- **Random Forest**: Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is robust against overfitting and provides good accuracy.

- **NBTree (Naive Bayes Tree)**: NBTree is a hybrid algorithm that combines decision trees and Naive Bayes classifiers. It uses a decision tree structure where the leaves are Naive Bayes classifiers, allowing it to handle both numeric and categorical data effectively.

- **Logistic Regression**: Logistic Regression is a linear model for binary classification that estimates the probability of a binary outcome based on one or more predictor variables. It is easy to implement and interpret, making it a popular choice for many classification problems.

- **A1DE (Averaged One-Dependence Estimators)**: A1DE is an ensemble learning algorithm that averages over all One-Dependence Estimators (ODEs). Each ODE assumes that every attribute depends on exactly one other attribute in addition to the class. This model balances the trade-off between simplicity and flexibility.

- **Support Vector Machine (SVM)**: SVM is a powerful classification algorithm that finds the hyperplane that best separates the classes in the feature space. It is effective in high-dimensional spaces and is particularly useful for classification tasks with clear margin separation.

- **Naive Bayes**: Naive Bayes is a probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions between features. Despite its simplicity, it is surprisingly effective for many real-world problems, especially text classification.

Each model was evaluated with and without feature selection to determine the impact on performance.

4 Cross-Validation and Performance Evaluation

Cross-validation is a robust statistical method used to estimate the performance of machine learning models. It involves partitioning the dataset into multiple subsets, training the model on some subsets while validating it on the remaining ones. This process is repeated several times to ensure that each subset is used for both training and validation, thereby providing a comprehensive assessment of the model's performance.

In this project, we implemented a 10-fold cross-validation process to evaluate the performance of various classification models. The dataset was randomly divided into 10 equal-sized folds. For each iteration, 9 folds were used for training the model, and the remaining fold was used for validation. This process was repeated 10 times, with each fold being used exactly once as the validation set. The results from each iteration were averaged to produce a single performance metric for each model, reducing the variability due to random sampling.

For models with feature selection, we employed an attribute-selected classifier, which applies a wrapper method using the same model as the evaluator. This approach helps in selecting the most relevant features that contribute to the predictive power of the model. After selecting the features, the models were trained on the training set and evaluated on the test set, ensuring a fair comparison between models with and without feature selection.

This rigorous evaluation process allows us to assess the generalizability and robustness of each model, ensuring that the chosen models perform well on unseen data.

4.1 Balancing the Dataset

Class imbalance occurs when the number of instances in one class is significantly higher than in another class. This can lead to biased model performance, where the model favors the majority class. To address this issue, two common techniques are undersampling and oversampling.

Undersampling involves reducing the number of instances in the majority class to balance the class distribution. While this method can lead to faster training times and less memory usage, it risks discarding potentially valuable information, which might negatively impact the model's performance.

Oversampling involves increasing the number of instances in the minority class by replicating existing instances or generating new ones. Techniques like SMOTE (Synthetic Minority Oversampling Technique) create synthetic examples to achieve a balanced dataset. Oversampling can help the model learn the minority class better but may lead to overfitting if not done carefully.

In this project, we applied oversampling to the minority class "no" in the "higher" column to address the class imbalance. This approach was chosen to ensure that the model has sufficient examples from both classes to learn from, improving its ability to predict the minority class accurately. By repeating the entire analysis with the balanced dataset, we aimed to achieve more reliable and unbiased performance metrics for the models.

5 Results

The performance of each model was evaluated using accuracy metrics. Charts were generated to compare the cross-validation accuracy with the test set accuracy for each model. The results showed the impact of feature selection and class balancing on the models' performance.

The box plots show the performance of various classification models on balanced and filtered data, balanced and unfiltered data, unbalanced and filtered data, and unbalanced and unfiltered data. Here is a detailed comment highlighting the differences between filtering and balancing on model accuracy:

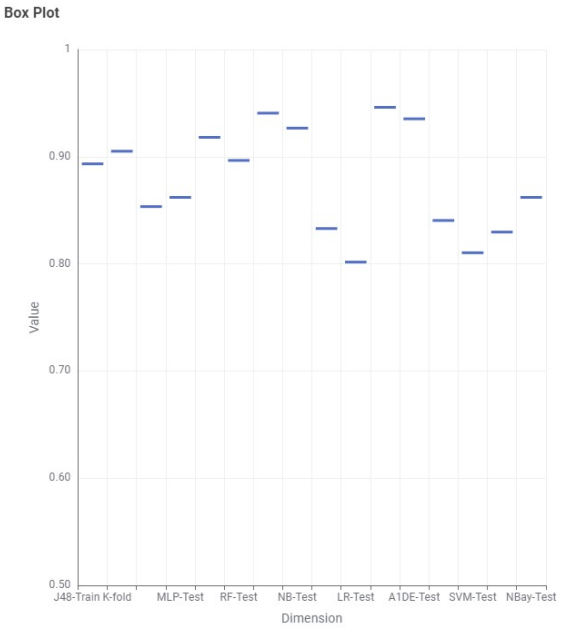

Figure 1: Box Plot balanced and filtered.

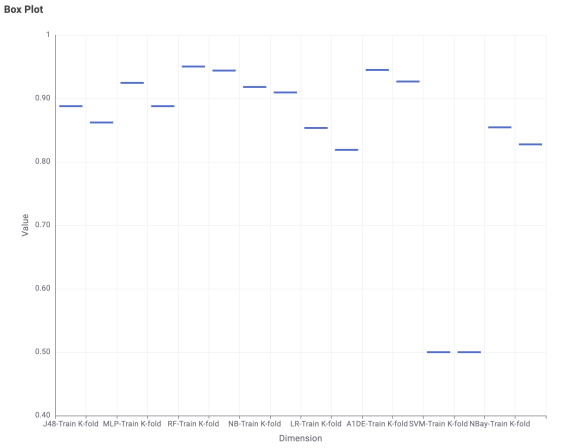

Figure 2: Box Plot balanced and Not filtered.

In general, comparing the two graphs. The filtered one and the unfiltered one, it seems that due to filtering the accuracy tends to be higher, especially in the SVM model which seems particularly sensitive. There are no significant differences in variability, except for the SVM model, which in the unfiltered case deviates greatly

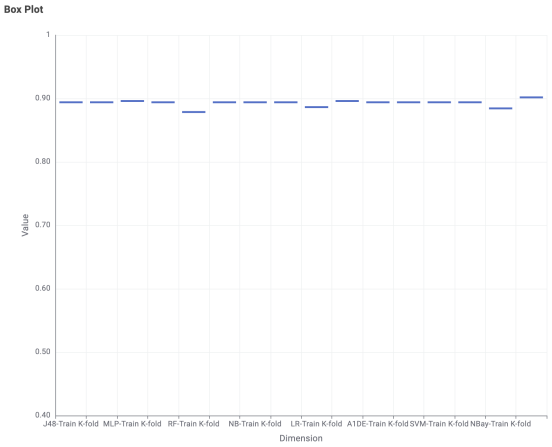

Figure 3: Box Plot Not balanced and filtered.

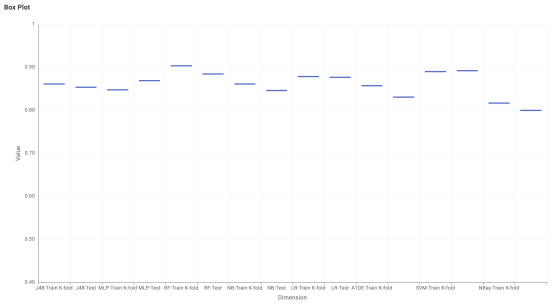

Figure 4: Box Plot Not balanced and Not filtered.

In the graphs presented, both involving unbalanced data, there are significant differences between the case with feature filtering and the case without filtering. In the first graph, with filtered data, model accuracies are generally very high and close to each other, with average values around 0.90. The variance between model performance is minimal, indicating very stable and predictable behavior. Feature filtering appears to have improved the consistency of performance across models, reducing the variance.

In the second graph, with unfiltered data, greater variability in model performance is observed. Some models, such as NBtree, A1DE and Naive Bayes show a significant reduction in accuracy, approaching 0.80.

The stability of performance is lower, with higher variance than in the filtered graph.

6 Conclusions

The project successfully demonstrated the use of KNIME for predicting higher education enrollment based on parental and personal information. The comparative analysis of multiple models highlighted that both oversampling (data balancing) and feature selection are essential for optimizing the performance of machine learning models.

Regarding the oversampling and filtering operations, we can say that in general both filtered and unfiltered balanced data have slightly higher accuracies. While regarding the filtering operation, it is clear how it goes to reduce the variability of model performance in unbalanced models. Feature filtering reduces variability in accuracies by removing noise from the data, preventing overfitting and simplifying models. This improves the quality of predictions by making models more stable and consistent in their performance.

7 Future Improvements

Although the work is complete, there are definitely improvements that can be made in the future. Definitely include the values of indicators such as precision, recall, and F1 score.

Precision is a measure of the performance of a classification model, indicating the proportion of true positives to the total number of positive results detected.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Recall is a measure of the performance of a classification model, indicating the proportion of true positives to the total number of true positive results detected.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

F1 score is a measure of the performance of a classification model, representing the harmonic mean between precision and recall.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

In addition, one could analyze other graphs such as the ROC curve (Receiver Operating Characteristic curve), which is a graph representing the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at various decision threshold levels of a classification model. It can also be used to compare models; one can compare the area under the ROC curve (AUC, Area Under the Curve).

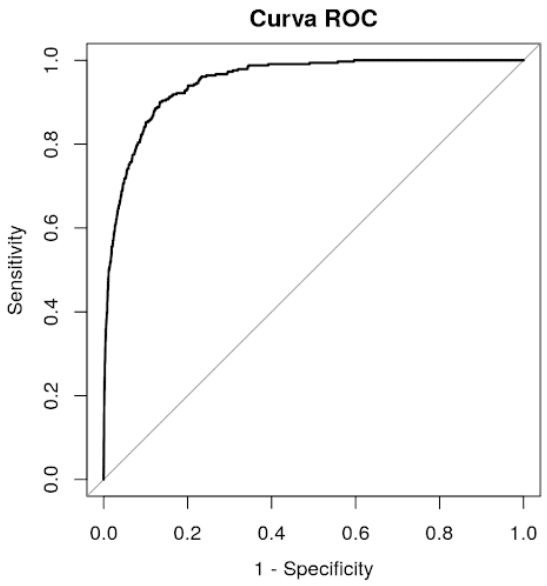

Figure 5: ROC curve

Finally, analyze box plots measuring

average accuracy validation, and average accuracy validation vs accuracy of each fold of the validation. To understand how the variance varies in the different models.

References

[1] Christoffersen, M. N., & Soothill, K. (2003). The long-term consequences of parental alcohol abuse: A cohort study of children in Denmark. *Journal of Substance Abuse Treatment*, 25(2), 107-116.

[2] Latendresse, S. J., Rose, R. J., Viken, R. J., Pulkkinen, L., Kaprio, J., & Dick, D. M. (2008). Parenting mechanisms in links between parents' and adolescents' alcohol use behaviors. *Alcoholism: Clinical and Experimental Research*, 32(2), 322-330.

[3] Johnson, J. L., & Leff, M. (1999). Children of substance abusers: Overview of research findings. *Pediatrics*, 103(Supplement 2), 1085-1099.

[4] Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, 19(2), 294-304.