# Decision Trees

Mikołaj Nowak 151813

01-11-2024

## 1 Attribute Selection

The primary goal of this analysis is to predict whether a student will pass or fail a subject based on selected attributes. After reviewing the attributes I selected those that may have the most significant influence on academic outcomes.

- **Absences**: This attribute represents the total number of school absences a student has accumulated. Frequent absences can lead to lower academic performance due to missed instructional time and decreased engagement with course materials.
- **Higher Education Aspirations**: Students who intend to pursue higher education may be more motivated and disciplined in their studies, which can positively impact their performance.
- **Internet Access**: Internet access at home can provide students with additional learning resources, potentially improving their understanding of the subject and academic outcomes.
- **Family Educational Support**: Family support is often crucial to a student's success, as it may provide encouragement, motivation, and assistance with academic tasks.
- **Paid Classes**: Extra paid classes within the subject area (Math or Portuguese) could provide targeted support to improve academic outcomes, making this attribute relevant for predicting performance.
- **Weekly Study Time** : Study time is a direct measure of the amount of time a student dedicates to learning, which is often correlated with academic success. More study hours typically lead to better understanding and retention of course material.
- **Failures in Previous Classes**: A history of failures in previous classes can indicate academic challenges or gaps in understanding, making this attribute a potential indicator of future performance.
- **Mother's Education Level**: Parental education level, particularly the mother's education, has been shown to influence student performance. Higher educational levels may be associated with more academic support at home.
- **Father's Education Level**: Similar to maternal education, the father's education level can contribute to a supportive learning environment, potentially impacting the student's motivation and resources available for academic achievement.

## 2 Selection of Evaluation Metrics

To effectively evaluate the performance of the classifier on the student dataset, we need metrics that capture different aspects of classification accuracy. Given the nature of our task — determining if a student passes or fails based on several factors — I selected three key metrics: **accuracy**, **precision and recall**, and **F1 score**.

**Accuracy** provides a general measure of how well the model performs overall by calculating the proportion of correctly classified instances. While useful, accuracy alone can be misleading in cases of imbalanced classes, where one outcome (e.g., pass or fail) might dominate.

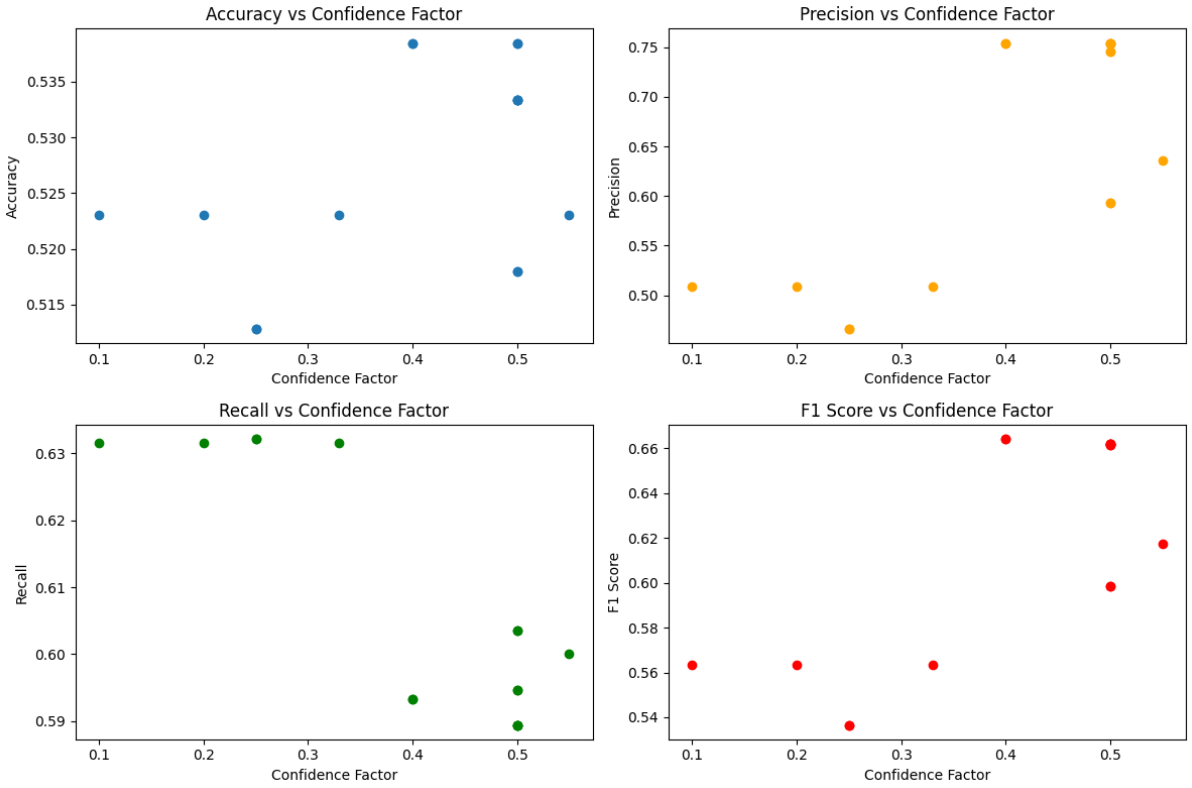
**Precision** and **Recall** are therefore necessary to understand the classifier's performance beyond simple accuracy. **Precision** tells us the proportion of positive predictions (e.g., predicting a student will pass) that were actually correct, while **Recall** measures how many actual positives were captured by the model. Since both metrics are relevant in evaluating correct identification without bias, their combination is particularly useful.

Finally, I included the **F1 Score**, which balances Precision and Recall to provide a single metric that accounts for both false positives and false negatives. This balance is especially important in educational data, where misclassifications could lead to incorrect conclusions about a student's abilities.

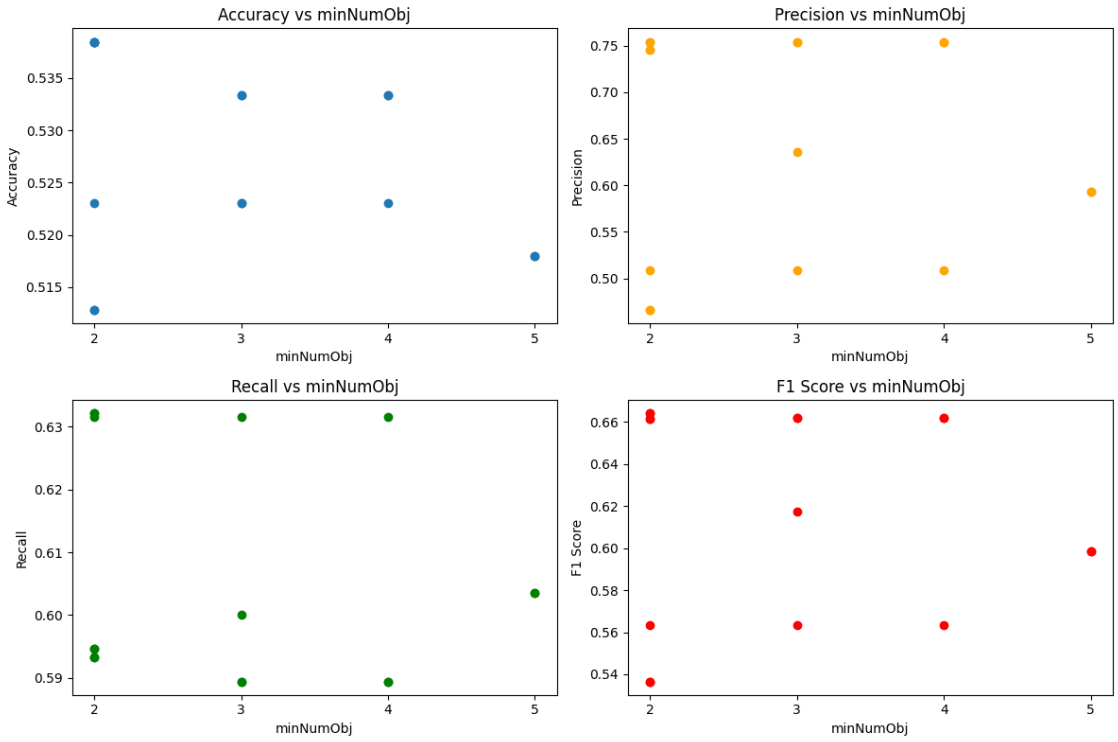
Together, **Accuracy, Precision, Recall, and F1 Score** should provide a well-rounded evaluation of the model's effectiveness for this dataset. These are among the most commonly used metrics in classification tasks, so they are expected to perform well on most datasets, including this one.

# 3 Model Evaluation and Results

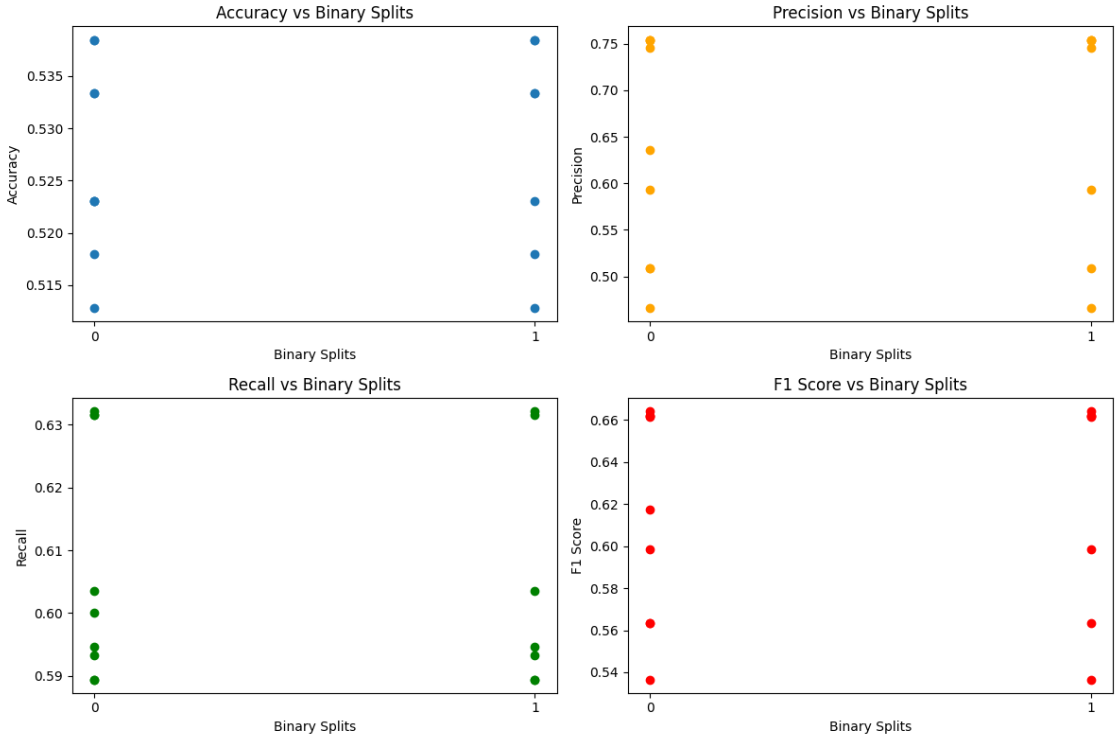
In this section, we evaluate the performance of our classification model using various metrics and parameters. We selected five attributes for the analysis: Medu, Fedu, studytime, failures, and absences. We tested multiple combinations of parameters, including confidence factor, minimum number of objects, and binary splits, to identify the optimal settings for our model. Below, we present the results for each parameter set, visualized through graphics.



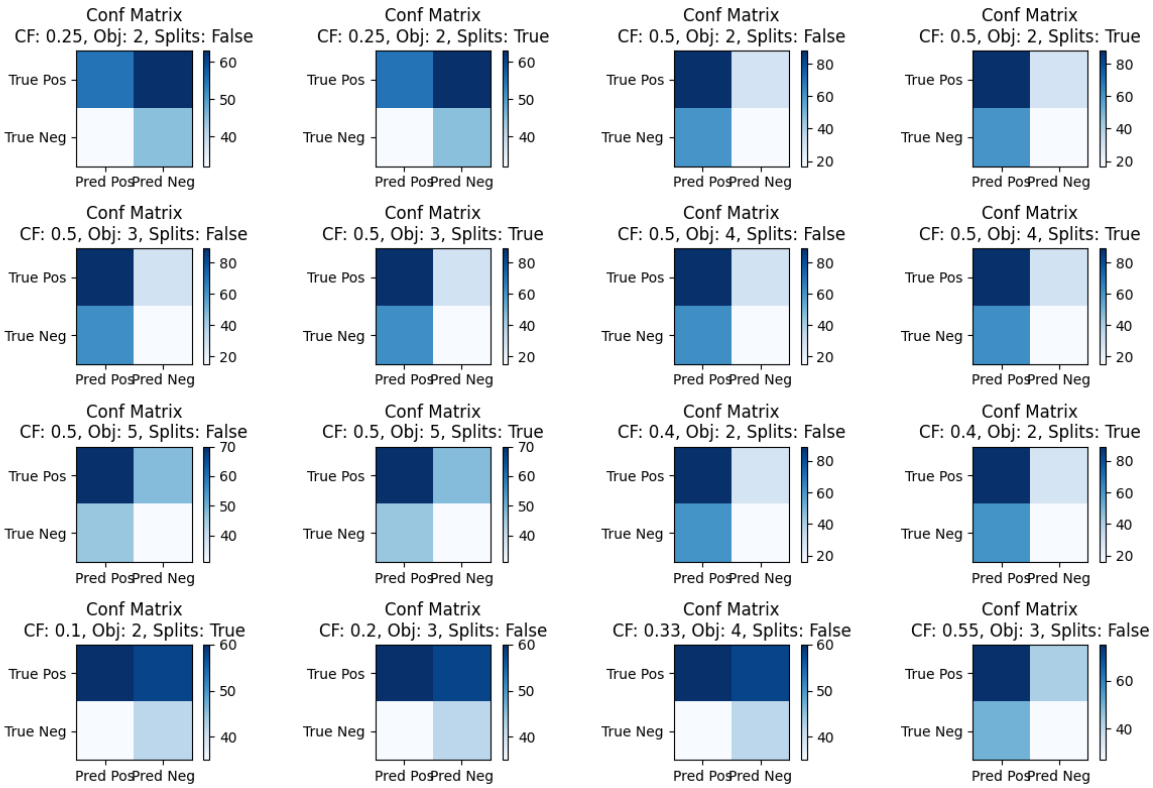

Confidence factor




minObjNum




Binary splits




Confusion matrices

Classifier output

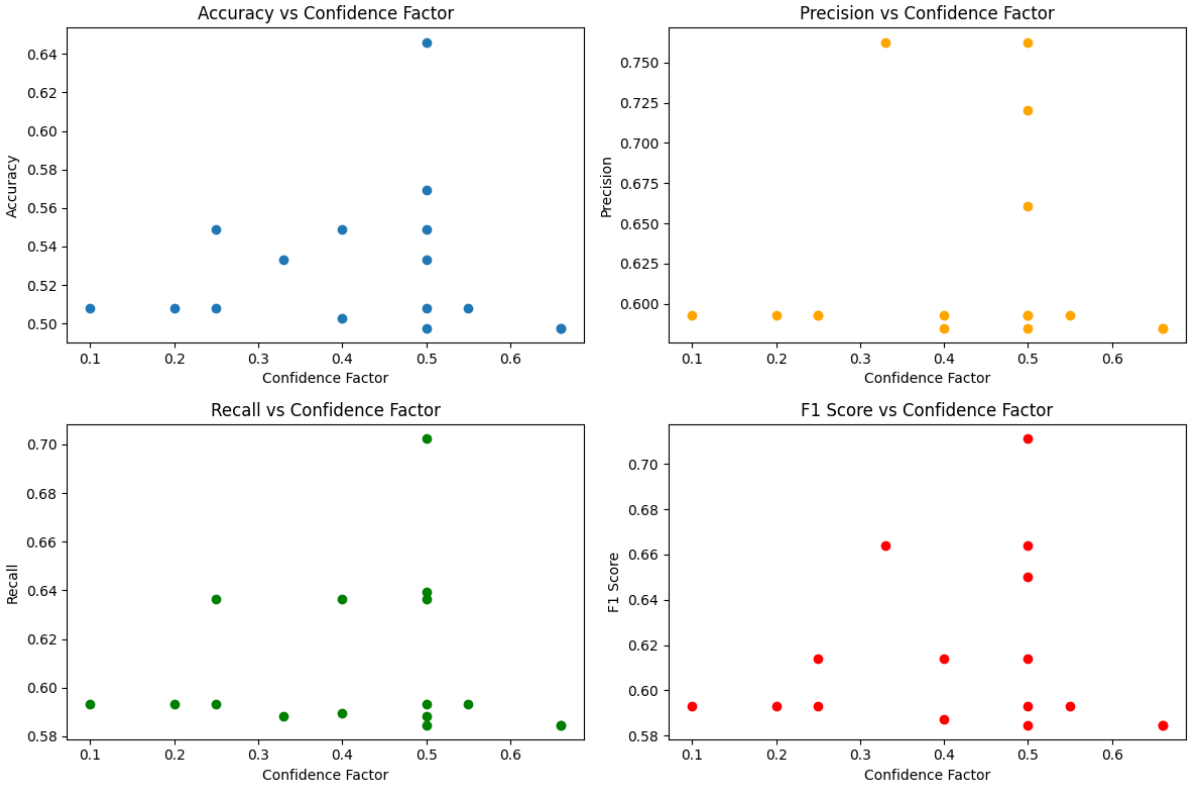
```
failures <= 1.0
|   studytime <= 1.0
|   |   Medu <= 1.0: '(-inf-11.5]' (7.0/2.0)
|   |   Medu > 1.0
|   |   |   famsup = no: '(11.5-inf)' (19.0/3.0)
|   |   |   famsup != no
|   |   |   |   internet = no: '(11.5-inf)' (4.0/1.0)
|   |   |   |   internet != no
|   |   |   |   |   paid = no
|   |   |   |   |   |   absences <= 0.0: '(-inf-11.5]' (2.0)
|   |   |   |   |   |   absences > 0.0: '(11.5-inf)' (9.0/2.0)
|   |   |   |   |   paid != no
|   |   |   |   |   |   absences <= 4.0: '(11.5-inf)' (6.0/2.0)
|   |   |   |   |   |   absences > 4.0: '(-inf-11.5]' (2.0)
|   studytime > 1.0
|   |   absences <= 7.0
|   |   |   internet = no
|   |   |   |   absences <= 2.0
|   |   |   |   |   Medu <= 3.0
|   |   |   |   |   |   Medu <= 2.0
|   |   |   |   |   |   |   paid = no: '(11.5-inf)' (2.0)
|   |   |   |   |   |   |   paid != no: '(-inf-11.5]' (2.0)
|   |   |   |   |   |   Medu > 2.0: '(-inf-11.5]' (2.0)
|   |   |   |   |   Medu > 3.0: '(11.5-inf)' (2.0)
|   |   |   |   absences > 2.0: '(-inf-11.5]' (9.0/1.0)
|   |   |   internet != no
|   |   |   |   Medu <= 1.0: '(-inf-11.5]' (10.0/2.0)
|   |   |   |   Medu > 1.0
|   |   |   |   |   paid = no
|   |   |   |   |   |   famsup = no: '(-inf-11.5]' (12.0/6.0)
|   |   |   |   |   |   famsup != no
|   |   |   |   |   |   |   failures <= 0.0: '(11.5-inf)' (21.0/5.0)
|   |   |   |   |   |   |   failures > 0.0: '(-inf-11.5]' (2.0)
|   |   |   |   |   paid != no: '(-inf-11.5]' (41.0/20.0)
|   |   absences > 7.0: '(-inf-11.5]' (25.0/4.0)
failures > 1.0: '(-inf-11.5]' (23.0)
```

Best tree

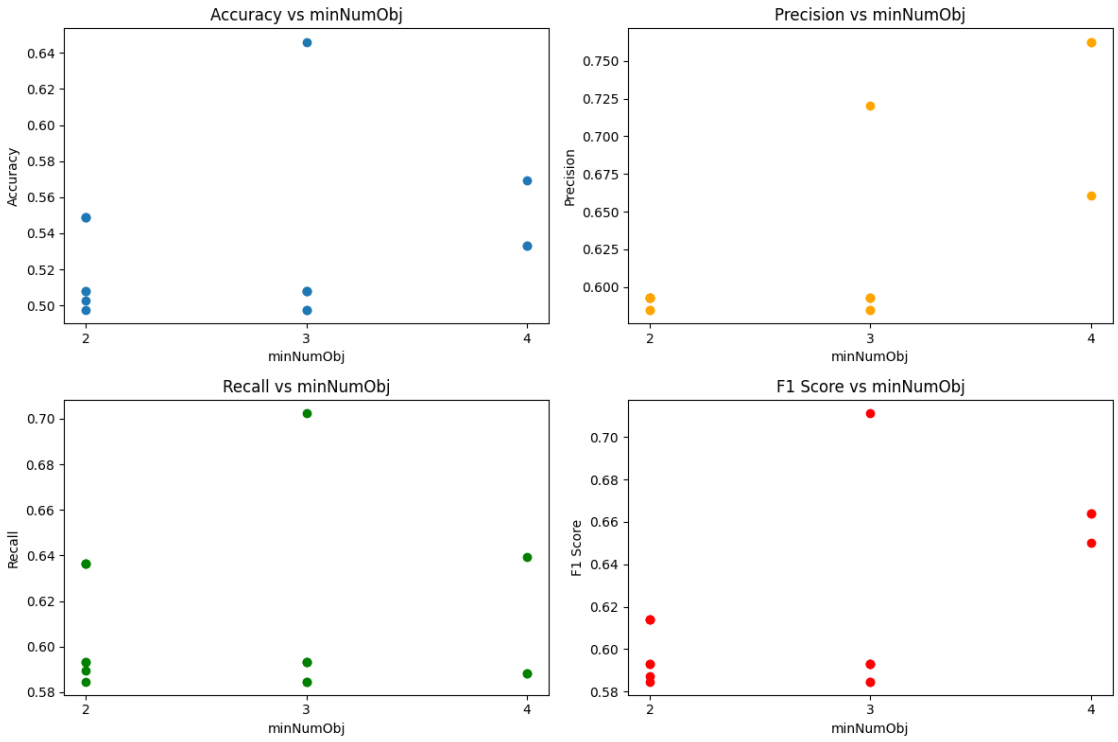
The best results were obtained with the parameters set to Confidence Factor = 0.5, minNumObj = 2, and Binary Splits = False, yielding a confusion matrix of (88, 30, 60, 17). However, overall performance remains suboptimal, likely due to the selection of inappropriate attributes for this classification task.

# 4 Comprehensive Model Evaluation and Results

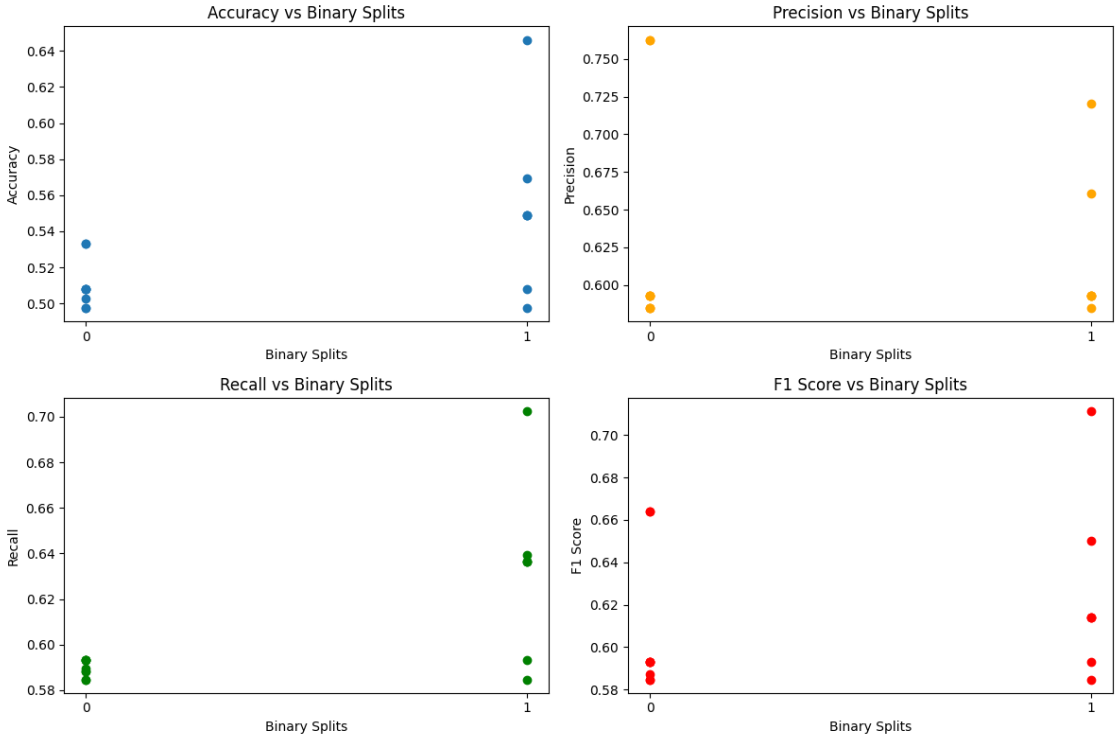
Let’s expand the analysis to include all available attributes in the classification model.



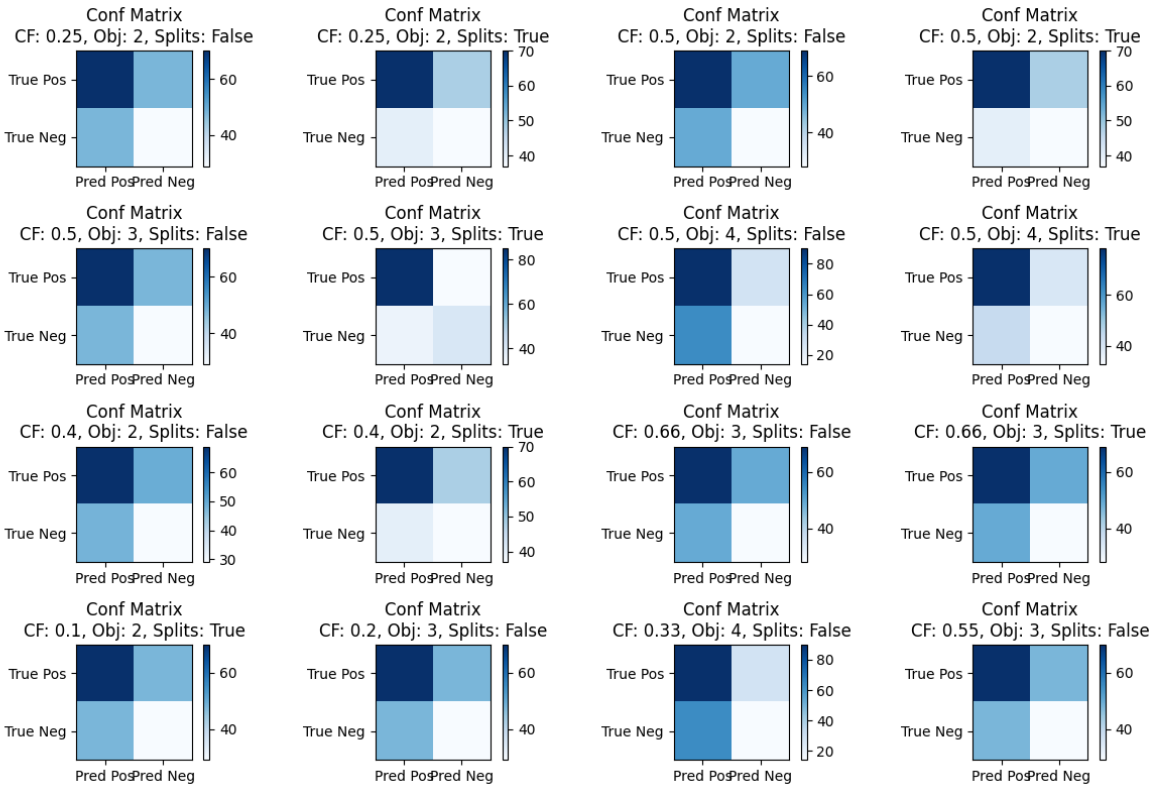

Confidence factor




minObjNum




Binary splits




Confusion matrices

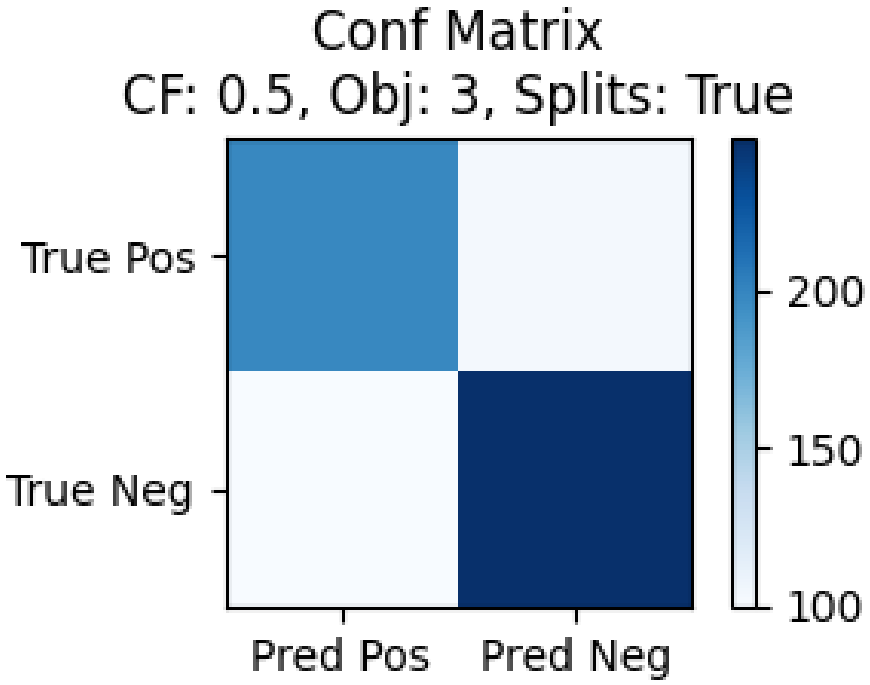
```
failures <= 1.0
|   schoolsup = yes
|   |   studytime <= 1.0
|   |   |   Mjob = services
|   |   |   |   reason = course: '(-inf-11.5]' (3.0)
|   |   |   |   reason != course: '(11.5-inf)' (3.0/1.0)
|   |   |   Mjob != services: '(11.5-inf)' (3.0)
|   |   studytime > 1.0: '(-inf-11.5]' (28.0/1.0)
|   schoolsup != yes
|   |   Mjob = at_home
|   |   |   famsup = no
|   |   |   |   age <= 16.0: '(11.5-inf)' (4.0)
|   |   |   |   age > 16.0: '(-inf-11.5]' (5.0)
|   |   |   famsup != no: '(-inf-11.5]' (6.0)
|   |   Mjob != at_home
|   |   |   Mjob = services
|   |   |   |   Walc <= 3.0
|   |   |   |   |   reason = other: '(-inf-11.5]' (3.0/1.0)
|   |   |   |   |   reason != other: '(11.5-inf)' (27.0/3.0)
|   |   |   |   Walc > 3.0: '(-inf-11.5]' (3.0)
|   |   |   Mjob != services
|   |   |   |   failures <= 0.0
|   |   |   |   |   Mjob = health
|   |   |   |   |   |   absences <= 4.0: '(11.5-inf)' (12.0/1.0)
|   |   |   |   |   |   absences > 4.0: '(-inf-11.5]' (4.0/1.0)
|   |   |   |   |   Mjob != health
|   |   |   |   |   |   paid = no
|   |   |   |   |   |   |   freetime <= 3.0
|   |   |   |   |   |   |   |   absences <= 8.0
|   |   |   |   |   |   |   |   |   internet = no: '(11.5-inf)' (3.0)
|   |   |   |   |   |   |   |   |   internet != no
|   |   |   |   |   |   |   |   |   |   famrel <= 4.0
|   |   |   |   |   |   |   |   |   |   |   Fjob = other: '(-inf-11.5]' (4.0)
|   |   |   |   |   |   |   |   |   |   |   Fjob != other: '(11.5-inf)' (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   famrel > 4.0: '(11.5-inf)' (5.0/1.0)
|   |   |   |   |   |   |   |   absences > 8.0: '(-inf-11.5]' (4.0)
|   |   |   |   |   |   |   freetime > 3.0: '(11.5-inf)' (20.0/3.0)
|   |   |   |   |   |   paid != no
|   |   |   |   |   |   |   Fjob = services: '(-inf-11.5]' (3.0)
|   |   |   |   |   |   |   Fjob != services
|   |   |   |   |   |   |   |   health <= 1.0: '(11.5-inf)' (3.0)
|   |   |   |   |   |   |   |   health > 1.0
|   |   |   |   |   |   |   |   |   sex = F: '(-inf-11.5]' (11.0/1.0)
|   |   |   |   |   |   |   |   |   sex != F
|   |   |   |   |   |   |   |   |   |   Walc <= 1.0: '(11.5-inf)' (4.0)
|   |   |   |   |   |   |   |   |   |   Walc > 1.0
|   |   |   |   |   |   |   |   |   |   |   famsup = no: '(11.5-inf)' (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   famsup != no: '(-inf-11.5]' (8.0/1.0)
|   |   |   |   failures > 0.0: '(-inf-11.5]' (5.0/1.0)
failures > 1.0: '(-inf-11.5]' (23.0)
```

Best tree

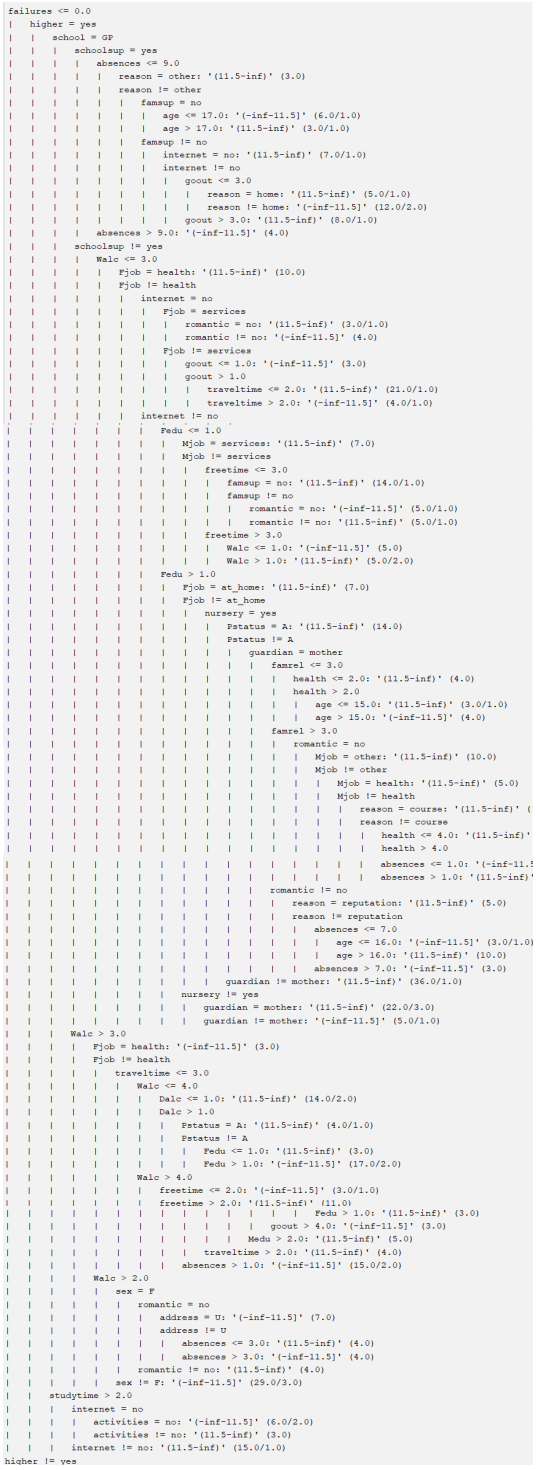
Now that we've considered all the attributes, the results have improved, but they're still not perfect. The model shows better performance, especially with the number of failures, which turned out to be an important factor. However, the test and train sets might be quite different, which could explain why the results aren't as good as we'd like. There's definitely room for improvement, and it is possible to keep experimenting with different attribute combinations to get better results.

## 5 Comparison of Decision Trees from Portuguese and Math Datasets

In this part, we compare the decision trees generated from both the Portuguese and math datasets. Even though the number of failures still stands out as the best attribute in both cases, the trees themselves look completely different! This shows that while some attributes are important, how they relate to student performance can vary a lot between subjects.



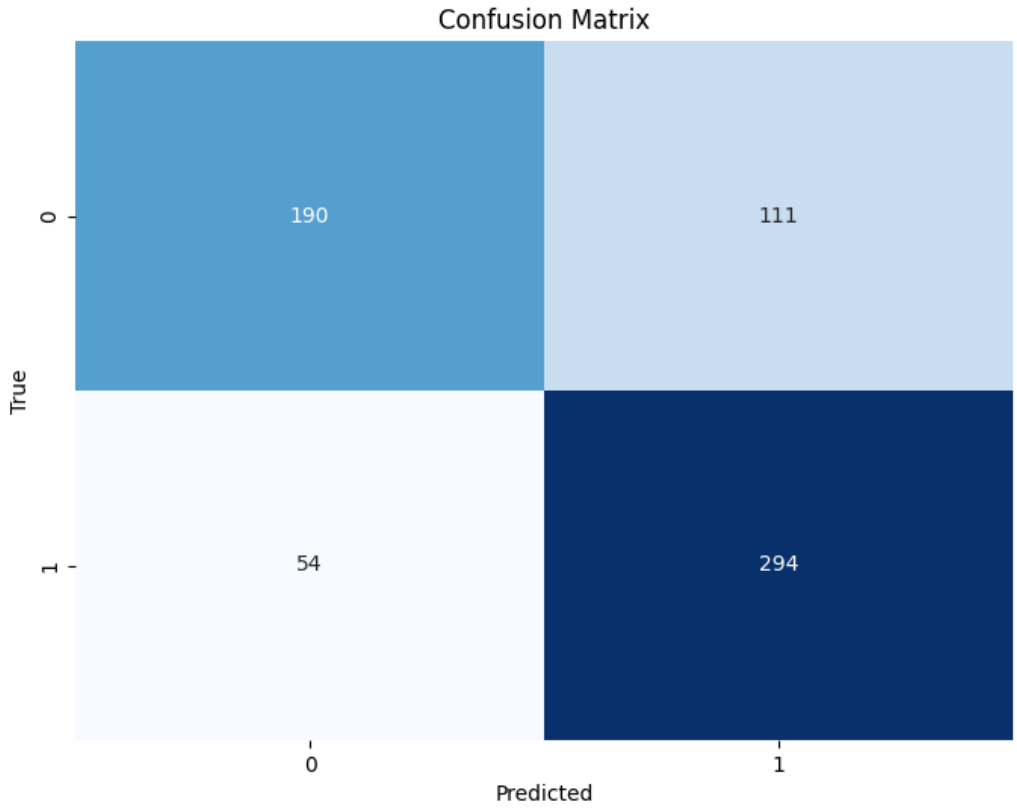

Confusion matrix




Best tree

These differences suggest that other factors might interact with the main attributes in ways I didn't expect. Student attentiveness is influenced by many things, and just looking at the number of failures won't tell the whole story.

# 6 Naive Bayes




Both algorithms overall chose similar attributes.