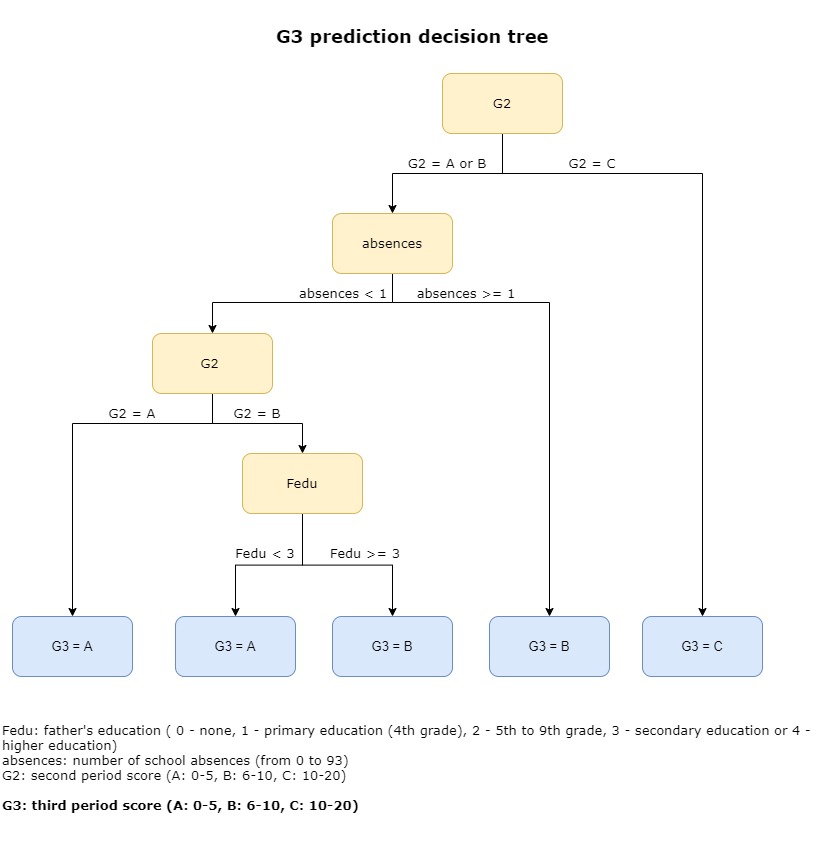
|  |  |  |
| --- | --- | --- |
| prediction | Important features for decision tree | Root node error |
| G1 | absences, failures, Fedu, Fjob, freetime, health, Medu, Mjob, Paid, reason, schoolsup | 0.50 |
| G2 | G1 | 0.49 |
| G3 | absence, Fedu, G2 | 0.49 |

Prediction for math

Predict grades into 3 categories: A, B, C

|  |  |  |
| --- | --- | --- |
| responses | Overall acc | Risk acc |
| G1 | 51.35% | 62.5% |
| G2 | 82.94% | 75.47% |
| G3 | 86.6% | 80.95% |



I will talk about the initial prototype and results.

This is the flow chart for our basic model.

As the aim of the project is to predict whether a student is potentially at risk with academic achievement, the model that we build will be focusing on students who are failing the subject, which is having grades under 10. We divided the student into 3 groups based on their grades: 0-5, 6-10, and more than 10.

We built decision trees to obtain features that are used in predicting a student’s grade. For example, for G3 prediction, the decision tree suggests that G2 is the most important feature, then is absence and father’s education level.

Then We split our dataset, 70% is used as training set and the rest to be the test set. The basic model is built using multinomial logistic regression with the important features obtained from the decision tree. The accuracy is evaluated with the test set and calculated from the confusion matrix.

from the bar chart we can see that, although the accuracy for G1 prediction is only 51%, the accuracy for G2 and G3 is quite high, this means that the information contained in G1 and G2 is highly relevant to G2 and G3 respectively.

We have also done a comparison with 4 categories grouping, which we further split students with grades more than 10 into 2 groups: 11-15 and 16-20.

The accuracy for 4 categories grouping is not as high, but we can still see the same trend of the high information content in G1 and G2.

This finding shows that for building our prediction models, we need to focus more on the prediction of G1. With a high accuracy of G1 prediction, the following prediction of G2 and G3 will be more reliable to indicate whether a student is going to face a potential risk academically.

Here is the timeline about what we have done and what we are going to do in the future, and in the next semester, we plan to merge several different datasets and build a system that provides support to the students at risk.

Comparing with 4-cate acc:

3-cate multinomial model has a better ability to predict student at risk.

And has a better performance in overall prediction