Data mining and knowledge discovery Seminar

ICT2

November 14, 2023

1 Introduction

Santa Claus, as usual, creates his gifts by using many *ingredients*. Each ingredient consists of many possible aspects – values. Unfortunately, Santa forgot which ingredients he used to create the last batch of gifts. Help him find the appropriate ingredients to save the holidays!

2 Data

You can find the data of ingredients as *dataset.tsv*. With "Success" you can find marked the feature that tells you whether a gift could be created. With "IngredientX", where X is some number, you can find the features that represent ingredients. Values are always integers.

3 Tasks

1. Perform exploratory data analysis. Interpret the results – make sure you consider class imbalance in your interpretation!
2. Select the appropriate features - what is the target feature? What are the descriptive features? How many are there?
3. You realize that machine learning can help you figure out which ingredients are the most important. Create a decision tree and evaluate its predictive performance. Visualize the tree and comment on the results. What is your baseline?
4. Which features are the most important? Why? If you reduce the data set only to these features, what happens with performance?
5. Identify the best classifier – consider at least three other approaches and at least two different configurations for each of them. Present the final benchmark results in terms of AUC and Accuracy. Your classifier *must* perform better than the constant classifier.
6. BONUS: Santa remembered he used only two ingredients, can you find a rule that helps him pinpoint them? (hint: CN2 rule viewer; hint2: you can subsample the data for better performance)

Grading: All tasks weigh 20% each. Bonus task (6.) can bring you 10% extra (110% possible).