<Wine>

Kevin McKee

Executive Summary

We are given a data set that contains information on 13 different attributes of wine and then classifies each wine as either type 1, type 2 or type 3. The objective is to come up with a formula that will be able to classify any unknown wine as the correct type with the highest possible accuracy. I found that many of the attributes were similar for certain types and devised a way to classify any random wine sample with 100% accuracy. I call it the Coin-Sorter Method. It works like a coin sorting machine, where if it fits in the first slot it stops there, but if it doesn't fit there it slides right over that slot to the next one, continuing until it finds a place where it fits. I made a table where the top of the table is the first set of conditions. If the wine meets all of those conditions, then it should be classified as that type. It is very important that the wine meets all the conditions, and if it does then it should be classified as that type and not move on to the next set of conditions. If it does not meet the set of conditions, then it moves to the next set. If it meets that set then it is classified and removed; if it doesn't meet that set it keeps going down. Each wine keeps moving down until it is classified. This method will classify each wine with 100% accuracy as long as it is done correctly. The table is shown in the results section.

Problem Description

We are given a set of data on different wine attributes. The goal is to come up with a way to classify the wine as type one, two or three based on the differences in a number of attributes. Once we have this classification, we will be able to take an unknown sample of wine and plug the attributes into the formula and it should tell us which type of wine it is. The goal is to come up with a method that can categorize the wine with the highest percent accuracy possible.

Analysis Technique

First I found the correlation between each of the columns and the types of wines. Once I did this, I used the high correlations and sorted the data while analyzing which column I should start with. The first thing I noticed was if I sorted by proline, most of the type ones ended up at the bottom. Then I took all the data from the lowest type 1 (which included some type 2's and 3's) and looked for a way I could sort that data to separate the ones from the others. I found that if I sorted by Flavanoids, most of the type ones were at the bottom with only two type 2's left. Then I sorted by Color Intensity and it weeded out the rest of the type 2's, leaving me with only type 1's. I used these three rules to make the first step of the table. If the Proline is greater than 735 and the Flavanoids are greater than 2.19 and the Color Intensity is greater than 3.52, then you will get all of the type ones. Since the wines that fell into those categories had already been grouped as type one, I could take them out and just work with the rest. Next I found that if I sorted by Color Intensity, most of the type 2's were at the top and most of the type 3's were at the bottom. I made rules that if the Color Intensity was less than 3.8, it must be type 2 and if it is greater than 6.62 then it is type 3. By these rules, it would incorrectly classify some of the type ones as type 2's or 3's, but since the type 1's were already classified at the beginning, they will never get to this step. This classified most of the data with only a small gray area, which was reconciled by placing stipulations on Total Phenols and Hue.

Assumptions

I made the following assumptions:

1. The data I have is from an appropriate sample that represents these different wines.
2. The data I have was collected accurately

Results

I found a way to group the wines with 100 percent accuracy. I call this method the Coin-Sorter Method. It works like a coin sorter where all the coins pass over holes. The first hole the coins come to is the right size for only dimes, and all the dimes fall in that hole while the rest keep going. Next is pennies where pennies and dimes would fit, but since all the dimes were caught by the dime hole, you get only pennies. Next are Nickels and then Quarters working the same way. In classifying these wines, if the wine fits in the first category, it stops there and the rest of the categories do not matter. If it does not fit there, it goes down to the next one where it can either fit there or move down to the next until its attributes are consistent with what is declared in the table. The table only works if you work from top to bottom, but if you do then it works with 100% accuracy. This table correctly classified all 153 wines into their proper types.

Attributes	Classification	Number
Proline >= 735 and Flavanoids >= 2.19 and Color Intensity >= 3.52	Type 1	47/47
Color Intensity <= 3.8	Type 2	51/51
Color Intensity >= 6.62	Type 3	28/28
Total Phenols <= 2.32 and Hue <= .96	Type 3	17/17
Any left over	Type 2	9/9

Issues

The only issue I have with this method is the fact that some of the attributes overlap, so it is imperative that the table be followed from top to bottom. It only works when the type ones are taken out of the equation at the top.