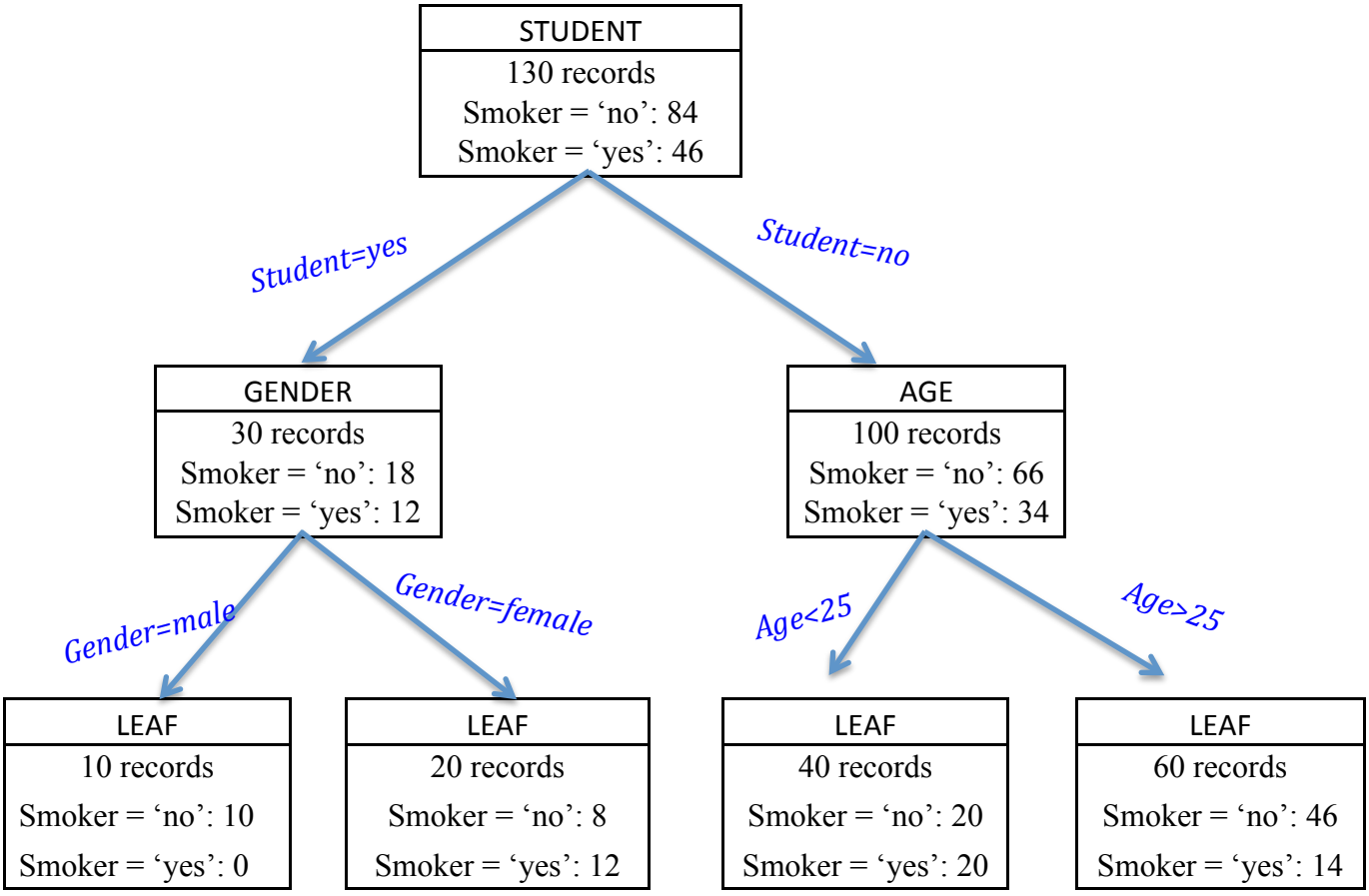

Taking the decision tree above, we are to decide if:

1. The branches check **Gender** should be kept or pruned
2. The branches checking **Age** should be kept or pruned

1: Does including the branches for gender = male / female reduce the error rate?

There are 30 students. 18 don't smoke and 12 do. So without the branch regarding gender, **18 students are labelled correctly** and 12 students are labelled incorrectly.

There are 10 male students, all non smokers. There are 20 female students, 12 smoke and 8 do not. So labelling all female students as smokers would label 12 student correctly and 8 student incorrectly.

Including the branches for gender results in **22 students being labelled correctly**, and 8 students being labelled incorrectly. Therefore the additional attribute of gender does reduce the error rate.

2: Does including age reduce the error rate?

There are 100 people who are not students. 66 do not smoke. So with the branches for age, **66 people are labelled correctly** and 34 people are labelled incorrectly.

Of the 100 non-students, 40 people are less than 25, and 50% of them smoke. So labelling under 25s as smokers would label **20 correctly** and 20 incorrectly.

60 non-students are over 25; 46 do not smoke and 14 do. So labelling over 25s as non-smokers would label **46 correctly** and 14 incorrectly.

So including age results in **66 people being labelled correctly**, and 34 students being labelled incorrectly. Therefore adding the two branches for age does not improve the error rate, and so age branches should be pruned.