Karen Tafa

CA03

Answers

**Q.1.1 Why does it makes sense to discretize columns for this problem?**

It makes it easier to understand the data. A lot of the data was continuous instead of discrete, therefore not binning the data would mean that it might be harder to predict the target variable than it would be with binned columns. This is because according to a medium article from Towards Data Science, continuous features have a smaller chance of correlating with the target variable due to infinite degrees of freedom. Also, entropy-based methods may not work with continuous data, and we used these in our project.

**Q.1.2 What might be the issues (if any) if we DID NOT discretize the columns.**

Our accuracy when it comes to predicting might be low because it’s harder for the algorithm to tell the people apart. Therefore, it might have lower performance values because it has to account for so many different rows instead of bins that are easier to predict.

**Q.8.1 How long was your total run time to train the model?**

Total time: 0.03493022918701172 seconds to train the model

**Q.8.2 Did you find the BEST TREE?**

Yes, Tree 5 using Gini Impurity, minimum sample split 5, minimum sample leaf 10 and Maximum Depth 10

**Q.8.4 What makes it the best tree?**

it has slightly better values than the rest although not much different than Decision Tree #4. It's got a higher accuracy and precision score

**Q.10.1 What is the probability that your prediction for this person is accurate?**

0.7083614295347269