1. Briefly outline why a model that has high accuracy during training may not always

generalize well after it is deployed? Give 2 reasons.

One issue could be overfitting,The model might have learned to memorize the training data rather than capture underlying patterns, leading to poor performance on unseen data.

Data Mismatch: The training data might not fully represent the variability present in real-world data, causing the model to struggle when faced with new examples. The model itself could be too complex for causing a high amount of accuracy.

2. The following data visualizations are based on a dataset collected to predict tachycardia

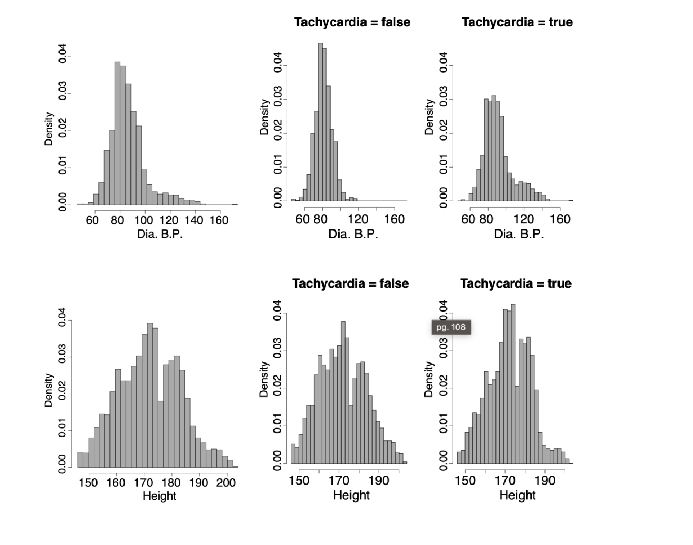
in adults (rapid heart rate).

a. The visualizations below illustrate the relationship between the target feature,

Tachycardia, and the continuous descriptive features Dia. B.P. (Diastolic Blood

Pressure) and Height. Based on the distributions, which descriptive feature has a

stronger relationship to the target and why?

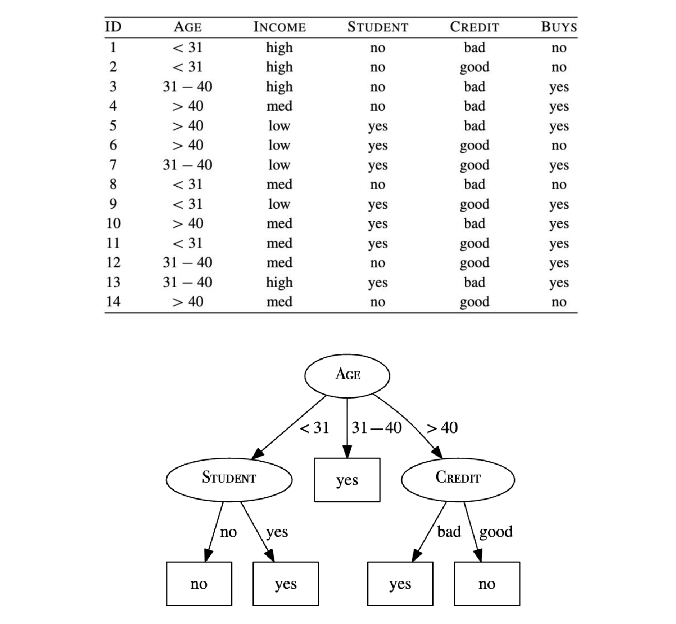


Between the visualizations, the Dia. B.P. seems to have more of a relationship to Tachycardia as they are all slightly skewed right compared to height which is more uniform overall making less of a relationship being prevalent in the second visualization.

3. The following dataset shows details of customers and whether they responded to a

special offer (BUYS). This dataset has been used to build a decision tree to predict

which customers will respond to future special offers.

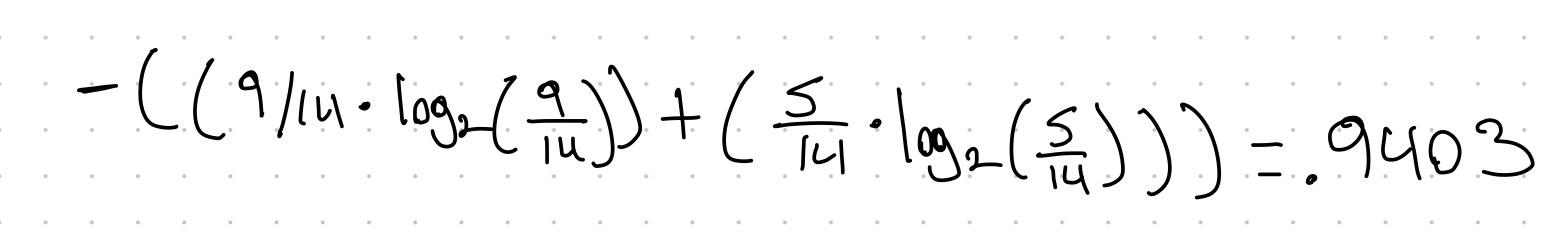


a. The information gain (calculated using entropy) of the feature AGE at the root

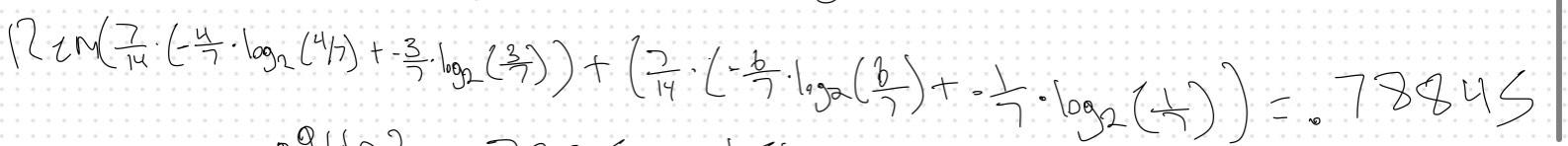
node of the tree is 0.247. A colleague has suggested that the STUDENT feature

would be better at the root node of the tree. Show that this is not the case

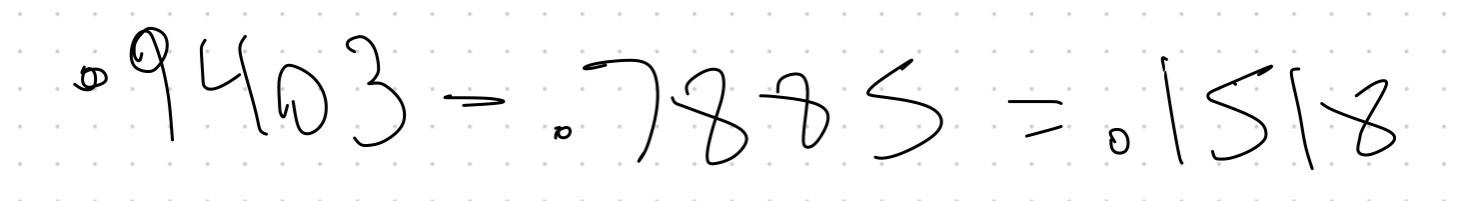
(calculate initial entropy, remaining entropy, and information gain).

Initial: 

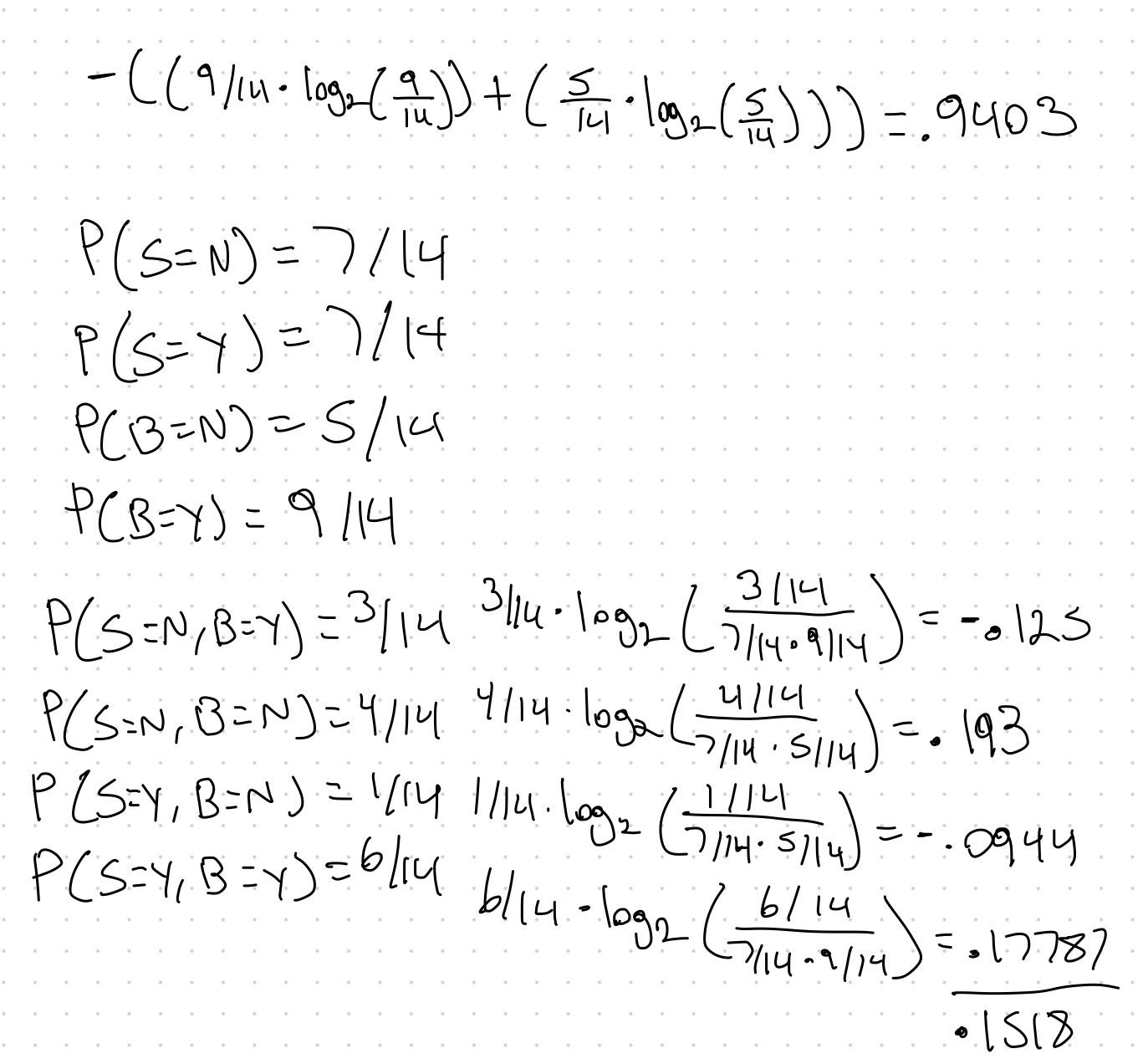
Remainder:



Information gain:



ID3 Way:

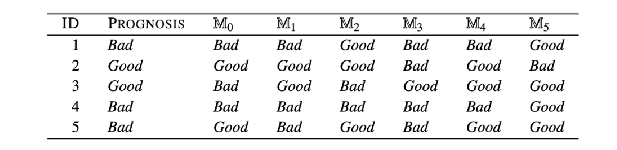


b. Another colleague has suggested that the ID feature would be more effective at

the root node of the tree. Would you agree with this suggestion? No, because that data is used as an identifier for the variables and holds no actual data related to the tree. It would be no different than selecting the whole tree of .9403.

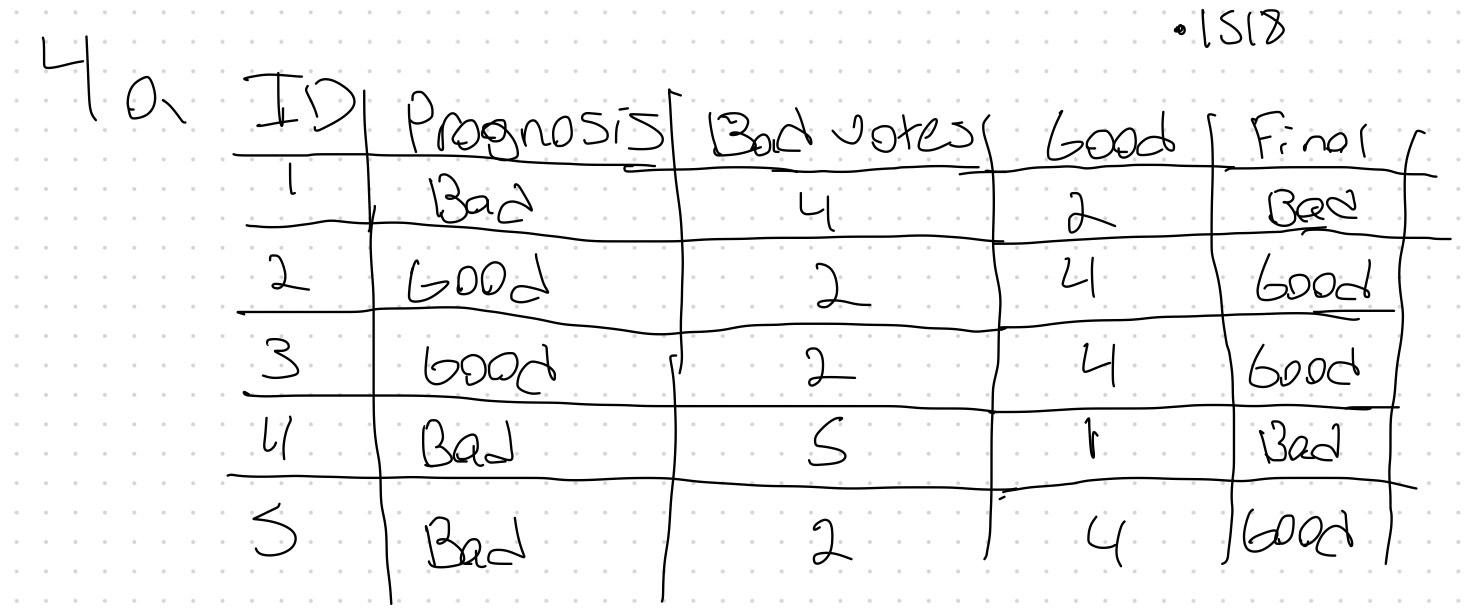
4. The following table shows a set of predictions made by six models in an ensemble and

the ground truth of the target feature (Prognosis).



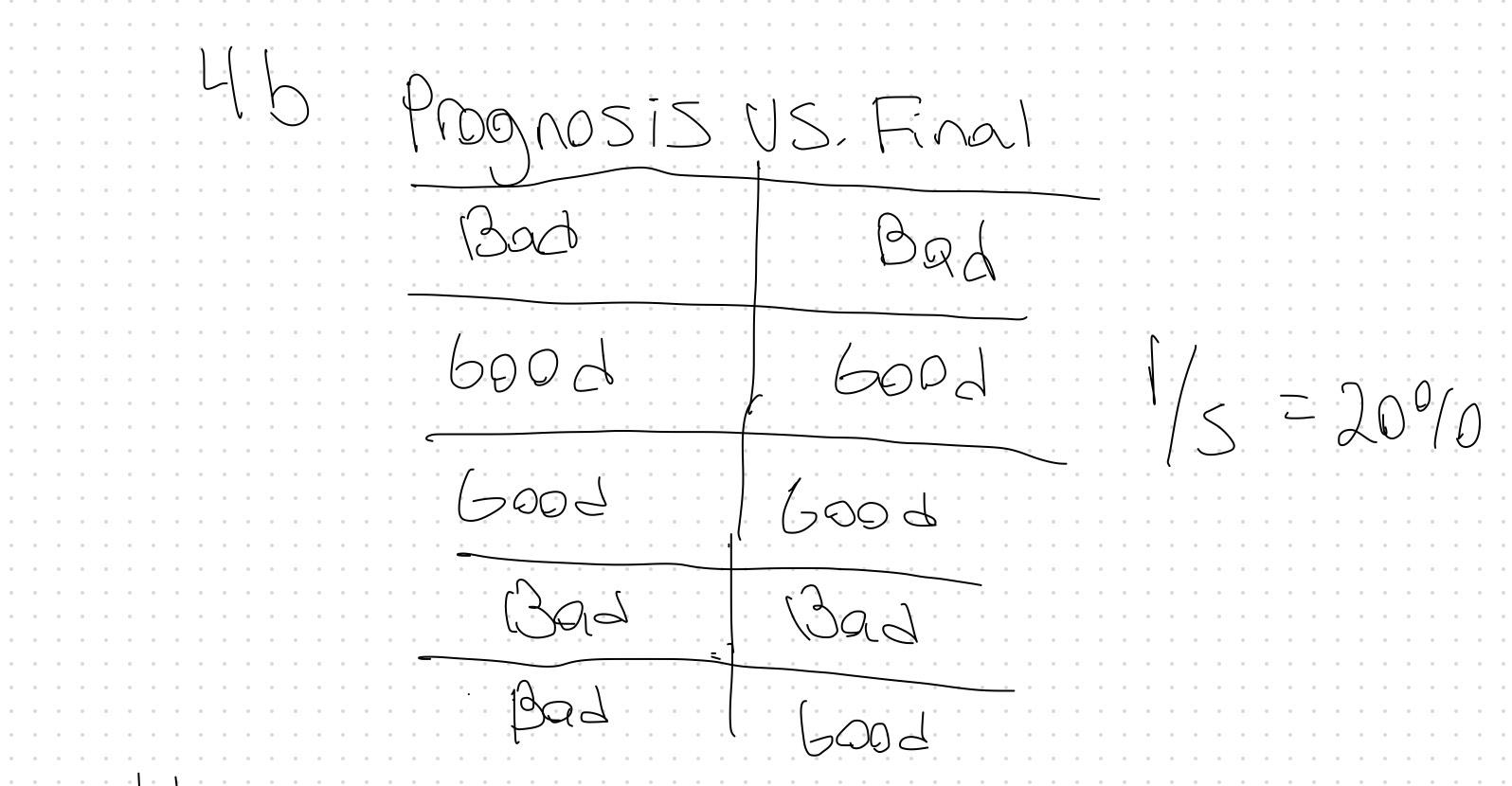
a. Assuming that these models are part of an ensemble training using bagging,

calculate the overall output of the ensemble for each instance in the test dataset.



b. Measure the performance of this bagged ensemble using misclassification rate

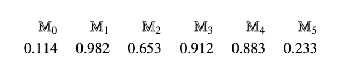
(percentage of instances in the test dataset that a model has incorrectly

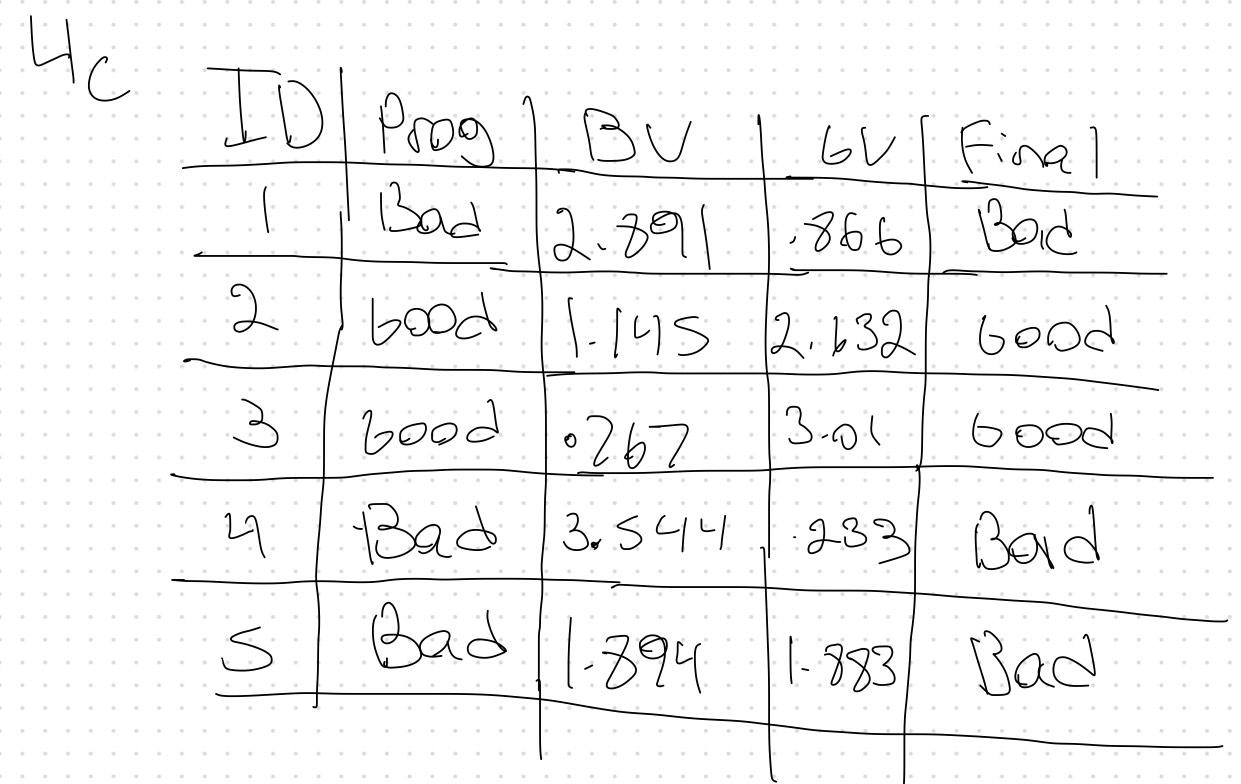
classified).

c. Assuming that these models are part of an ensemble trained using boosting and

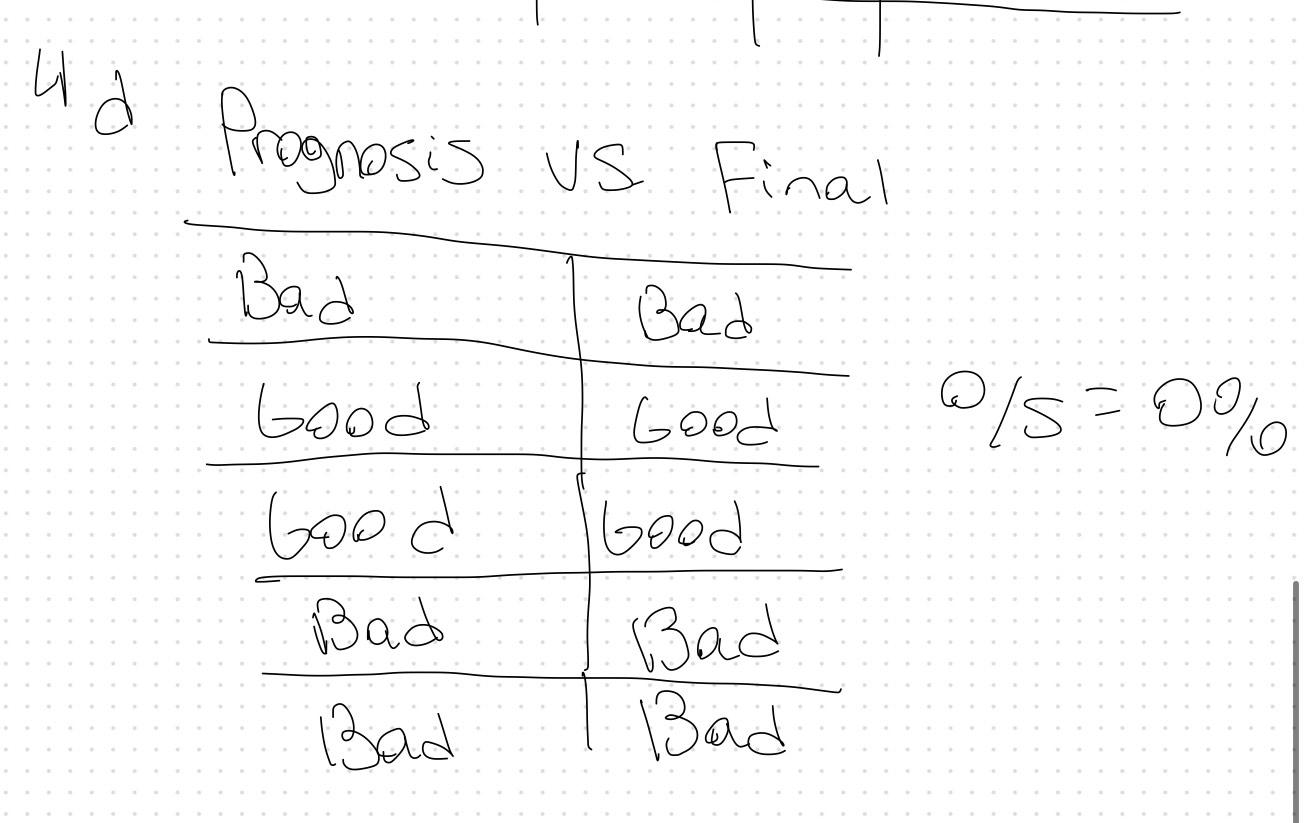
that the confidence factors, α, for the models are as follows, calculate the overall

output of the ensemble for each instance in the test dataset.





d. Measure the performance of this boosted ensemble using misclassification rate.



5. In constructing a random forest model:

a. What are the 2 main architectural components (i.e., what makes it a “random

forest” of trees)?

The main two components are the random trees that are created from bootstrap aggregating or “bagging” so that each sample that is used unique for each tree.

It also has random featuring to help make the less likely to overfit by letting not taking a a single dominant feature.

b. Why would a random forest technique reduce variance (overfitting) when

compared to training a single decision tree? By taking multiple trees predctions from multiple different subsets of the data, it reduces the chance of overfitting with the greater number of trees in the forest leads to higher accuracy

6. Fundamentally, how are boosting (AdaBoost) vs gradient boosting (XGBoost) ensemble

models different?

Adaboosting trains a seqeunce of weak learners, with each iteration giving more weight for the method to learn from the misclassified examples. Once all the weak learners have been determined, it combines it into one final model with the predictions.

Gradient boosting is a stagewise addition method, where multiple weak learning algorithms are trained and a strong learner algorithm is used as a final model from the addition of multiple weak learning algorithms trained on the same dataset, XGbossting is a higher version of gradient boosting using a regularization technique in it to find the best parameters.

7. The plots below display the test and train accuracy of a decision tree across tree depths.

a. What insights can be drawn from this plot?

The model complexity has only a max depth of 20 and after depth of 4, it fluctuates on both the train and test data. The training data goes beyond the test data and gets close to a near perfect accuracy, indicating that the train data has learned too much as the max depth increases. The gap between train and test make it seem that the model thats bering ran is too complex.

b. What steps would you take to address and improve the situation?

I would prune the tree all the way to where it would best be optimized or after its been fully grown I would post prune it back by removing some subtrees. Limiting the tree depth will lessen the opportunity for it to overfit. Early stopping could be another option to make sure to make sure it doesn’t keep going after a specific threshold.

