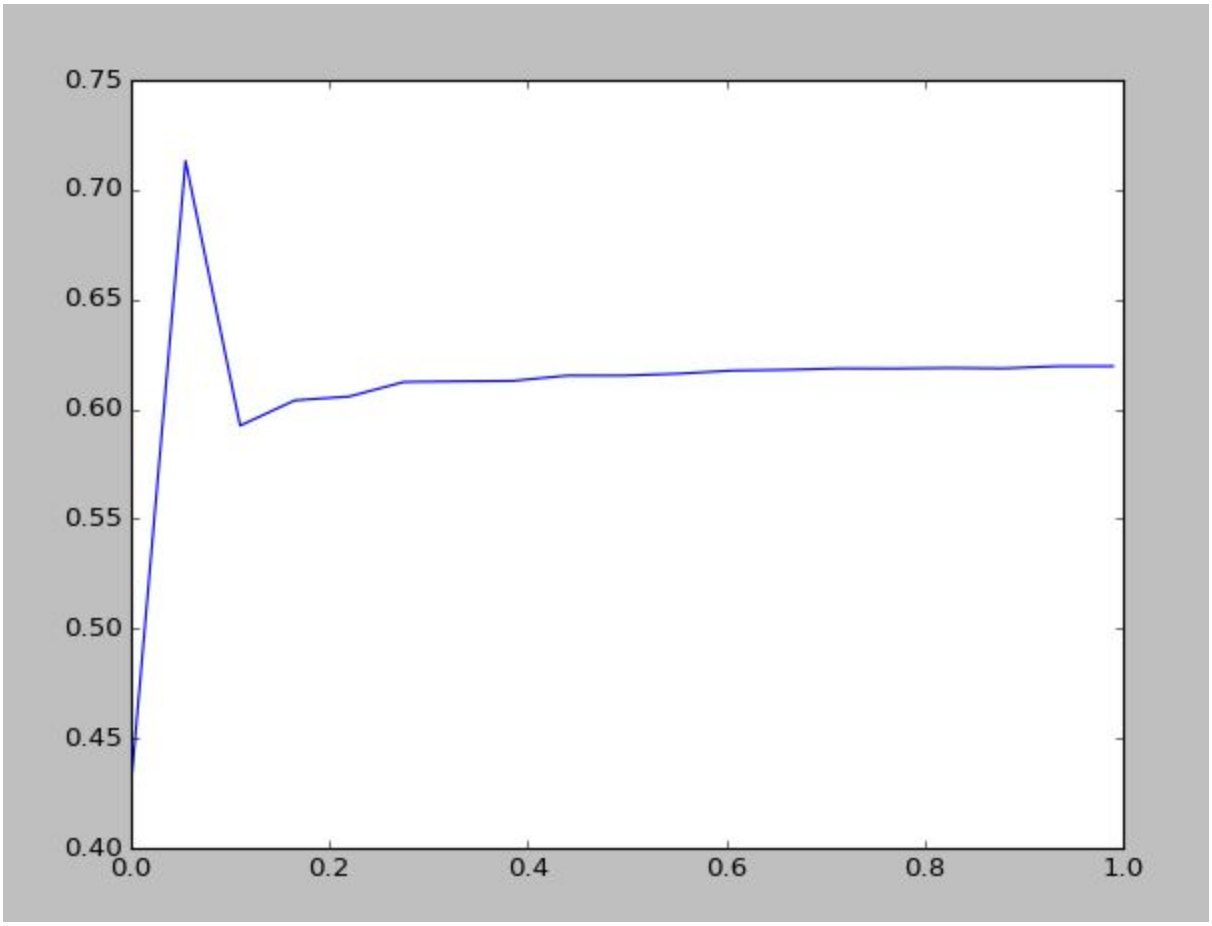
Ryan Miller (rsm408)
Rene Romo (rgr355)
Amanuel Alemu (aea592)

1) For nominal attributes we replaced the missing attribute with the mode of that attribute among all instances. For numeric attributes we averaged the values of that attribute across all instances
2)

```
if (oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured<1.0) AND (oppwinningpercent>=0.138129760285)
 OR(oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured<1.0) AND (rundifferential>=66.0)
 OR(oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured<1.0)
 OR(oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured<1.0)
 OR(oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured>=1.0) AND (opprundifferential<9.0) AND
(rundifferential<24.0)
 OR(oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured>=1.0) AND (opprundifferential<9.0) AND
(rundifferential>=24.0)
 OR(oppnuminjured<3.0) AND (numinjured<1.0) AND
(oppnuminjured>=1.0) AND (opprundifferential>=9.0) AND
(opprundifferential<26.0)
 OR(oppnuminjured<3.0) AND (numinjured>=1.0) AND
(oppwinningpercent<0.179993143003) AND (oppwinningpercent<-
0.0140945713339) AND (rundifferential>=9.0)
```

3) Rule 1 :if (oppnuminjured<3.0) AND (numinjured<1.0) AND (oppnuminjured<1.0) AND (oppwinningpercent>=0.138129760285) .
   If the opponent has less than 1 injured players and we have less than 1 injured player and the opponent's winning percentage is greater than 13 % we win
4) <no pruning>
5) <no pruning>
6) <no pruning>

7) Unpruned accuracy: 86.26% The accuracy of a pruned tree will be higher, since pruning removes noisy nodes and bad generalizations; it helps minimize over-fitting and helps removes errors
8)




Unpruned (Accuracy vs Percentage of Dataset)

9) The pruned tree will run better on the data set, since the pruned tree has minimized overfitting, which can result in bad predictions. We ran our unpruned tree on the test set, see our file PS2.csv
10) We pair programmed throughout all of the assignment and split up questions equally
11) Accuracy on Validation set using just IG: 84.81% This is lower than with using IGR, so using IGR instead of IG was a good model decision. By testing all split points on numeric attributes, we would guarantee that we would have the maximum information gain on that split this would lead to overfitting to our training data. The opposite side of things would be testing very few splits which would likely lead to a lower information gain and not be selecting a good splitting value. Limiting the step side for numeric attributes allows us to ensure a good splitting value that can generalize instead of overfitting.