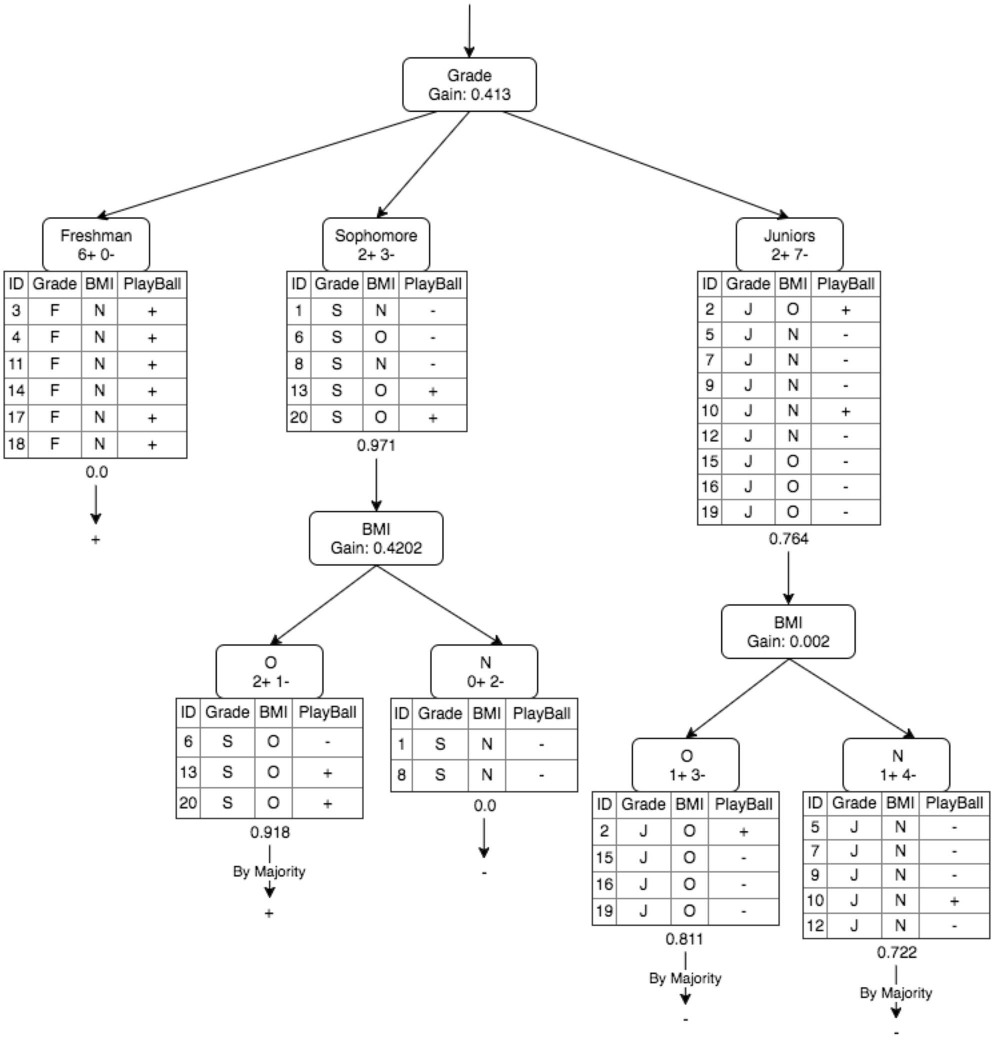
CS 440 - Homework 6

Matthew Liu

1.



ID	Grade	BMI	PlayBall
3	F	N	+
4	F	N	+
11	F	N	+
14	F	N	+
17	F	N	+
18	F	N	+

ID	Grade	BMI	PlayBall
1	S	N	-
6	S	O	-
8	S	N	-
13	S	O	+
20	S	O	+

ID	Grade	BMI	PlayBall
6	S	O	-
13	S	O	+
20	S	O	+

ID	Grade	BMI	PlayBall
1	S	N	-
8	S	N	-

ID	Grade	BMI	PlayBall
2	J	O	+
5	J	N	-
7	J	N	-
9	J	N	-
10	J	N	+
12	J	N	-
15	J	O	-
16	J	O	-
19	J	O	-

ID	Grade	BMI	PlayBall
2	J	O	+
15	J	O	-
16	J	O	-
19	J	O	-

ID	Grade	BMI	PlayBall
5	J	N	-
7	J	N	-
9	J	N	-
10	J	N	+
12	J	N	-

(a) Decision Tree is attached above

(b) Classifications:

- Student 21: +
- Student 22: +
- Student 23: -
- Student 24: +
- Student 25: -
- Student 26: -
- Student 27: +
- Student 28: +
- Student 29: -
- Student 30: +

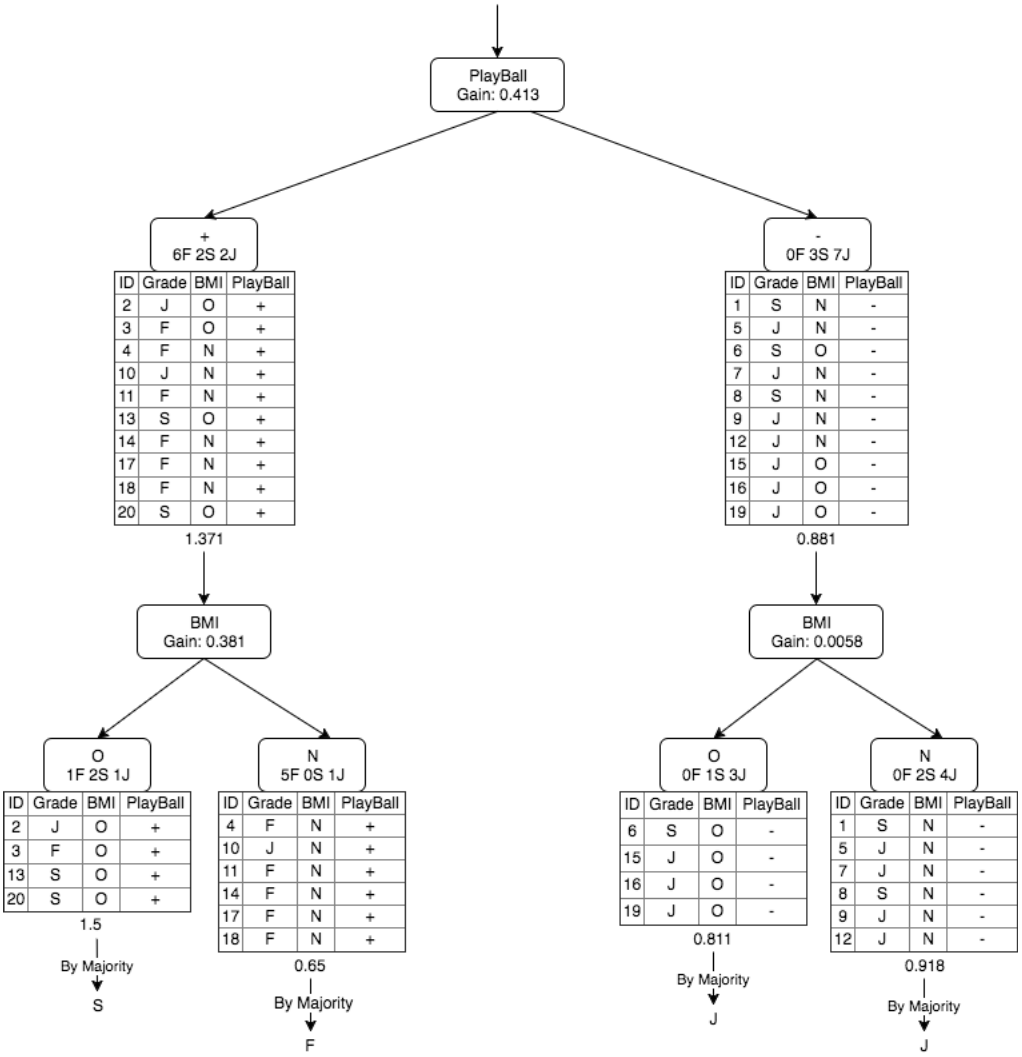
(c) Classifications:

- Student 11: +
- Student 12: -
- Student 13: +
- Student 14: +
- Student 15: -
- Student 16: -
- Student 17: +
- Student 18: +
- Student 19: -
- Student 20: +

(d) The results from **(b)** would be the better estimate as **(c)** was part of the training set for the classifier

(e) The classifier from Scenario 2 would be significantly more accurate then the classifier from Scenario 1, as the Scenario 1 does not have enough training data to produce meaningful results

(f) Decision Tree is attached below



ID	Grade	BMI	PlayBall
2	J	O	+
3	F	O	+
4	F	N	+
10	J	N	+
11	F	N	+
13	S	O	+
14	F	N	+
17	F	N	+
18	F	N	+
20	S	O	+

ID	Grade	BMI	PlayBall
1	S	N	-
5	J	N	-
6	S	O	-
7	J	N	-
8	S	N	-
9	J	N	-
12	J	N	-
15	J	O	-
16	J	O	-
19	J	O	-

ID	Grade	BMI	PlayBall
2	J	O	+
3	F	O	+
13	S	O	+
20	S	O	+

ID	Grade	BMI	PlayBall
4	F	N	+
10	J	N	+
11	F	N	+
14	F	N	+
17	F	N	+
18	F	N	+

ID	Grade	BMI	PlayBall
6	S	O	-
15	J	O	-
16	J	O	-
19	J	O	-

ID	Grade	BMI	PlayBall
1	S	N	-
5	J	N	-
7	J	N	-
8	S	N	-
9	J	N	-
12	J	N	-

2.

(a) The following are three possible approaches

- Limit the depth: limit how deep the decision tree can go
- Limit the minimum number of example used to select a split: Require a minimum amount of data to continue splitting the decision tree. This is similar to limiting the depth, but does so differently
- Cross Validation: Partition data into training set and testing set, and rotate data for both. sizes of both sets can vary.

(b) Cross Validation

(c) I would break up the data into sets of 5 (into a total of N sets), and rotate which N-1 sets will be used for training use the remaining set as a evaluation/testing set. This is better than limiting the depth because that forcefully prunes the tree, and in a tree like this one where it can't really go all that deep, all it does is create a poor classifier. Limiting the minimum number of examples used to select a split would work if there was a plethora of data, however this dataset does not have a wealth of data, hence this method would again limit the effectiveness of our classifier

(d) As this algorithm rotates through and considers all possible combinations of data subsets for training and evaluation, it will always produce the shortest decision tree because of it. Of course the cost is that depending on the size of the dataset, this algorithm can take a while to complete.