INFO254 Lab 3

Yingyue Luan

February 13, 2018

1

There are 910 rows in the dataset for males and 1078 rows for females.

2

The Gini Index of this dataset, using males and females as the target classes is 0.496429.

3

The best split point of the 'height' feature is 68.5 inches.

4

The Gini Index of this best split is 0.26552881207029189.

5

This partitioning reduces the Gini Index over that of the overall dataset by 0.23090046783448887.

6

There are 905 'female' rows below the best split point and 142 'male' rows below the best split point.

7

There are 173 'female' rows above the best split point and 768 'male' rows above the best split point.

8

```
[['hazel'],
 ['brown'],
 ['hazel', 'brown'],
 ['green'],
 ['hazel', 'green'],
 ['brown', 'green'],
 ['hazel', 'brown', 'green'],
 ['blue'],
 ['hazel', 'blue'],
 ['brown', 'blue'],
 ['hazel', 'brown', 'blue'],
 ['green', 'blue'],
 ['hazel', 'green', 'blue'],
 ['brown', 'green', 'blue'],
 ['hazel', 'brown', 'green', 'blue'],
 ['other'],
 ['hazel', 'other'],
 ['brown', 'other'],
 ['hazel', 'brown', 'other'],
 ['green', 'other'],
 ['hazel', 'green', 'other'],
 ['brown', 'green', 'other'],
 ['hazel', 'brown', 'green', 'other'],
 ['blue', 'other'],
 ['hazel', 'blue', 'other'],
 ['brown', 'blue', 'other'],
 ['hazel', 'brown', 'blue', 'other'],
 ['green', 'blue', 'other'],
 ['hazel', 'green', 'blue', 'other'],
 ['brown', 'green', 'blue', 'other']]
```

9

The best split of eye color as measured by the Gini Index is ['green'].

10

The Gini Index of this best split is 0.49309157295097772.

11

This partitioning reduce the Gini Index over that of the overall data set by 0.0033377069538030324.

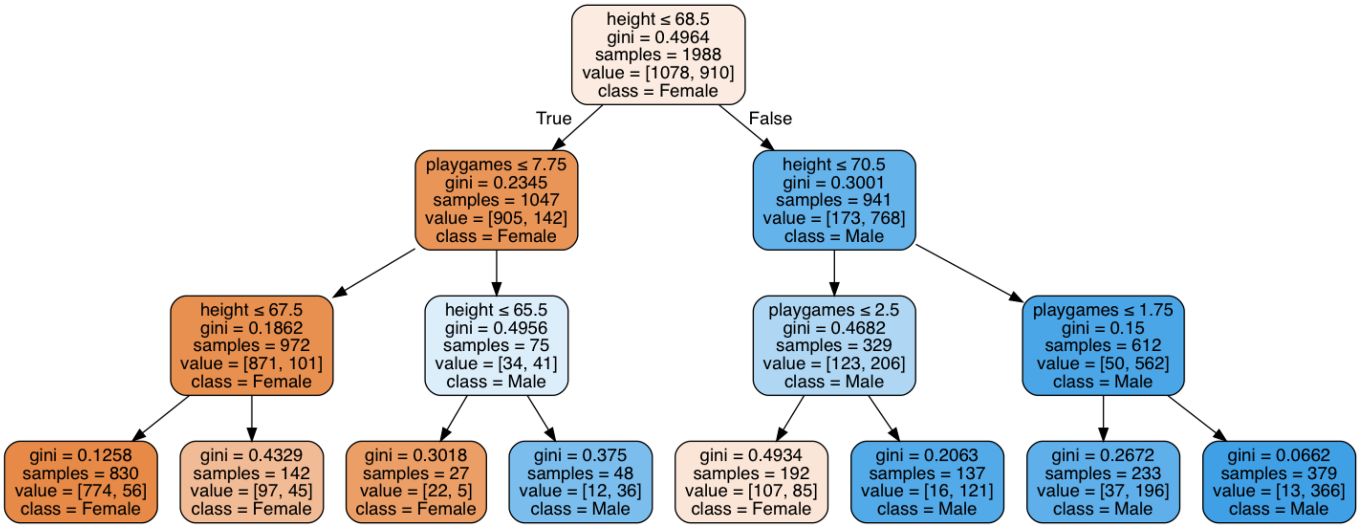
12

There are 190 'female' rows in the first partition and 107 'male' rows.
There are 888 'female' rows in the second partition and 803 'male' rows.

13

The accuracy score is 0.864688128773.

14



The primary split is conditioned on height, checking whether height is higher than 68.5 inches. This node contains 1988 samples and since it is the root node, it means the whole tree has 1988 samples. There are 1078 samples classified into 'female' and 910 samples into 'male'. This split gives us a gini impurity index of 0.4964.

- If the sample has height lower than or equal to 68.5 inches, the second split is whether the sample plays games more than 7.75 hours a week. This node contains 1047 samples and classifies 905 of them into 'female'. This split generates a gini impurity index of 0.2345.
  - If samples spend less than or equal to 7.75 hours playing games every week, they are further split by height of 67.5 inches. This node has 972 samples with 871 of them grouped as 'female'. The gini impurity index of this node is 0.1862.
    - If the samples are shorter than or equal to 67.5 inches, the split reaches a gini impurity index of 0.1258. This leaf node has 830 samples in total, 775 of them are tagged 'female' and 56 tagged 'male'.
    - For the 142 samples that are higher than 67.5 inches, 97 of them are classified into 'female' and 45 into 'male', with a gini impurity index of 0.4329.
  - If samples spend more than 7.75 hours playing games every week, they are further split by height of 66.5 inches. Only 75 samples are in this node and 41 of them are labeled 'male', with a gini impurity index of 0.4956.
    - If the samples are shorter than or equal to 66.5 inches, the split reaches a gini impurity index of 0.3018. This leaf node has 27 samples in total, 22 of them are tagged 'female' and 5 tagged 'male'.
    - For the 48 samples that are higher than 66.5 inches, 12 of them are classified into 'female' and 36 into 'male', with a gini impurity index of 0.4329.
- If the sample has height higher than 68.5 inches, we further check if the height is higher than 70.5 inches. There are 941 samples in this node and 173 of them are considered 'female' while 768 of them are considered 'male'. The gini impurity index of this node is 0.3001.
  - If samples are short than 70.5 inches, they are further split by whether they spend more than 2.5 hours playing games per week. This node has 329 samples with 206 of them grouped as 'male'. The gini impurity index of this node is 0.4682.
    - If the samples spend less than or equal to 2.5 hours playing games, the split reaches a gini impurity index of 0.4934. This leaf node has 192 samples in total, 107 of them are tagged 'female' and 85 tagged 'male'.

    - For the 137 samples that spend more than 2.5 hours playing games, 16 of them are classified into 'female' and 121 into 'male', with a gini impurity index of 0.2063.

- If samples are higher than 70.5 inches, they are further split by spending 1.75 hours on games. 612 samples are in this node and 562 of them are labeled 'male', with a gini impurity index of 0.15.
    - If the samples spend less than or equal to 1.75 hours on game per week, the split reaches a gini impurity index of 0.2672. This leaf node has 233 samples in total, 37 of them are tagged 'female' and 166 tagged 'male'.
    - For the 379 samples that spend more than 1.75 hours on games, 13 of them are classified into 'female' and 366 into 'male', with a gini impurity index of 0.0662.