Assignment - 8

Impurity Calculations using Gini & Entropy

HARSH SIDDHAPURA

1230169813

02/18/2024

Node Impurity Calculations

In the video, we discussed the Gini Index for computing Node Impurity. The book explains another impurity measure: Entropy. For each of the nodes below, compute the Gini Index & Entropy given their class distributions.

1. **Node 1 contains 20 data points**
    - **16 points belong to class C1**
    - **4 points belong to class C2**

	Node 1
C1	16
C2	4

- Gini Index = 1 - (16/20) ^2 – (4/20) ^2 = 0.32
- Entropy = - (16/20) * log2(16/20) – (4/20) * log2(4/20) = 0.722

2. **Node 2 contains 20 data points**
    - **11 points belong to class C1**
    - **9 points belong to class C2**

	Node 2
C1	11
C2	9

- Gini Index = 1 - (11/20) ^2 – (9/20) ^2 = 0.495
- Entropy = - (11/20) * log2(11/20) – (9/20) * log2(9/20) = 0.993

3. **Node 3 contains 20 data points**
    - **10 points belong to class C1**
    - **10 points belong to class C2**

	Node 3
C1	10
C2	10

- Gini Index = 1 - (10/20) ^2 – (10/20) ^2 = 0.5
- Entropy = - (11/20) * log2(11/20) – (9/20) * log2(9/20) = 1

- **How do the Gini values for the three nodes compare?**
  The Gini Index values for the three nodes are as follows:

  ➔ Node 1: 0.32

  ➔ Node 2: 0.495

  ➔ Node 3: 0.5

  The Gini Index measures the impurity of a node, with 0 representing a perfectly pure node (all instances belong to the same class) and 0.5 representing the highest impurity (instances are evenly split between two classes).

  Comparing these values, we can see that Node 1 has the lowest Gini Index, indicating that it is the most pure node with the majority of instances belonging to class C1. Node 3 has the highest Gini Index, indicating that it is the most impure node with an equal number of instances from both classes. Node 2 has a Gini Index value in between Node 1 and Node 3, indicating a relatively balanced mix of instances from both classes, but not as evenly distributed as in Node 3.

  In the context of decision tree models, lower Gini Index values are generally preferred as they indicate higher node purity. This can lead to more accurate classifications. However, the specific context and the overall structure of the tree can also influence the interpretation of these values.

- **Which node is the purest?**
  Node 1 is the purest node as it has the lowest Gini Index and Entropy values.

- **Which node has the highest level of impurity?**
  Node 3 has the highest level of impurity as it has the highest Gini Index and Entropy values.

- **Can you reach the same results from looking at the Entropy values?**
  Yes, we can reach the same results from looking at the Entropy values. Both Gini Index and Entropy are measures of impurity used in building decision trees, and lower values indicate purer nodes. So, the node with the lowest Entropy is the purest, and the node with the highest Entropy has the highest level of impurity. In this case, the results are consistent with those obtained from the Gini Index.

Split Impurity Calculations

(Gini Values)

	A	B	Class
1	F	T	C1
2	T	F	C1
3	T	F	C1
4	F	T	C1
5	F	F	C2
6	T	T	C2
7	F	T	C2
8	T	T	C2
9	T	T	C2

1. **What is the Gini Index for the 9 data points without splitting? You can compute it given that 4 data points belong to C1 & 5 belong to C2.**
   The Gini index without splitting is:
   Gini = 1 – (4/9) ^ 2 – (5/9) ^ 2 = 0.496

2. **What is the Gini Index if the data points are split based on attribute A?**
   - **Remember that you will need to split the 9 data points into 2 nodes, one contains all data points with A=T, and another node that contains all data points with A=F.**
   - **Then compute the Gini index for each of the two nodes.**
   - **Then combine the two Gini values using a weighted average to get the overall Gini Index for Split based on attribute A.**
   - **Review course materials & text if you are confused about how to compute the Gini Index for a split.**

First, we need to split the data points into two nodes based on whether attribute A is true (T) or false (F). From the given data, we can see that:

- Node 1 (A=T) contains 5 data points: 2 belong to class C1 and 3 belong to class C2.
- Node 2 (A=F) contains 4 data points: 2 belong to class C1 and 2 belong to class C2.

Now, we can calculate the Gini Index for each node.

- Gini Index for Node 1 (A=T): 1 – (2/5) ^ 2 – (3/5) ^ 2 = 0.48
- Gini Index for Node 2 (A=F): 1 – (2/4) ^ 2 – (2/4) ^ 2 = 0.5

Finally, we can calculate the overall Gini Index for the split based on attribute A using a weighted average of the Gini Index values for the two nodes. The weights are the proportions of data points in each node.

Overall Gini Index for Split based on A:

Average Weighted Gini = (5/9) * 0.48 + (4/9) * 0.5 = 0.49

So, the overall Gini Index for the split based on attribute A is approximately 0.49. This value is a measure of the impurity of the nodes after the split. The lower the Gini Index, the better the split.

3. **What is the Gini Index if the data points are split based on attribute B?**
   - **You will just repeat what you did in Q2 but using attribute B instead of A.**

First, we need to split the data points into two nodes based on whether attribute B is true (T) or false (F).

From the given data, we can see that:

- Node 1 (B=T) contains 6 data points: 2 belong to class C1 and 4 belong to class C2.
- Node 2 (B=F) contains 3 data points: 2 belong to class C1 and 1 belongs to class C2.

Now, we can calculate the Gini Index for each node.

- Gini Index for Node 1 (B=T): 1 – (2/6) ^ 2 – (4/6) ^ 2 = 0.44
- Gini Index for Node 2 (B=F): 1 – (2/3) ^ 2 – (1/3) ^ 2 = 0.44

Finally, we can calculate the overall Gini Index for the split based on attribute B using a weighted average of the Gini Index values for the two nodes. The weights are the proportions of data points in each node.

Overall Gini Index for Split based on B:

Average Weighted Gini = (6/9) * 0.44 + (3/9) * 0.44 = 0.44

So, the overall Gini Index for the split based on attribute B is approximately 0.44. This value is a measure of the impurity of the nodes after the split. The lower the Gini Index, the better the split.

4. **Which attribute gives a purer split?**

The attribute that gives a purer split is the one with the lower Gini Index after the split. From our previous calculations:

- The overall Gini Index for the split based on attribute A is approximately 0.49.
- The overall Gini Index for the split based on attribute B is approximately 0.44.

Therefore, **attribute B gives a purer split** because it has a lower Gini Index after the split. This means that splitting on attribute B results in child nodes that are more pure (i.e., contain more instances of a single class) compared to splitting on attribute A. In the context of decision tree models, purer nodes are generally preferred as they lead to more accurate classifications.

5. **Repeat steps 1, 2 & 3 using Entropy instead of Gini.**

Attempted on next page:

Split Impurity Calculations

(Entropy Values)

	A	B	Class
1	F	T	C1
2	T	F	C1
3	T	F	C1
4	F	T	C1
5	F	F	C2
6	T	T	C2
7	F	T	C2
8	T	T	C2
9	T	T	C2

1. **What is the Entropy for the 9 data points without splitting? You can compute it given that 4 data points belong to C1 & 5 belong to C2.**

   The Entropy without splitting is:

   Entropy = - (4/9) * log2(4/9) – (5/9) * log2(5/9) = 0.991

2. **What is the Gini Index if the data points are split based on attribute A?**
   - **Remember that you will need to split the 9 data points into 2 nodes, one contains all data points with A=T, and another node that contains all data points with A=F.**
   - **Then compute the Gini index for each of the two nodes.**
   - **Then combine the two Gini values using a weighted average to get the overall Gini Index for Split based on attribute A.**
   - **Review course materials & text if you are confused about how to compute the Gini Index for a split.**

First, we need to split the data points into two nodes based on whether attribute A is true (T) or false (F). From the given data, we can see that:

- Node 1 (A=T) contains 5 data points: 2 belong to class C1 and 3 belong to class C2.
- Node 2 (A=F) contains 4 data points: 2 belong to class C1 and 2 belong to class C2.

Now, we can calculate the Entropy for each node.

- Entropy for Node 1 (A=T): - (2/5) * log2(2/5) – (3/5) * log2(3/5) = 0.971
- Entropy for Node 2 (A=F): - (2/4) * log2(2/4) – (2/4) * log2(2/4) = 1

Finally, we can calculate the overall entropy for the split based on attribute A using a weighted average of the entropy values for the two nodes. The weights are the proportions of data points in each node.

Overall Entropy for Split based on A:

Average Weighted Entropy = (5/9) * 0.971 + (4/9) * 1 = 0.983

So, the overall entropy for the split based on attribute A is approximately 0.983. This value is a measure of the impurity of the nodes after the split. The lower the entropy, the better the split.

3. **What is the Entropy if the data points are split based on attribute B?**
    - **You will just repeat what you did in Q2 but using attribute B instead of A.**

First, we need to split the data points into two nodes based on whether attribute B is true (T) or false (F). From the given data, we can see that:

- Node 1 (B=T) contains 6 data points: 2 belong to class C1 and 4 belong to class C2.
- Node 2 (B=F) contains 3 data points: 2 belong to class C1 and 1 belongs to class C2.

Now, we can calculate the Entropy for each node.

- Entropy for Node 1 (B=T): - (2/6) * log2(2/6) – (4/6) * log2(4/6) = 0.918
- Entropy for Node 2 (B=F): - (2/3) * log2(2/3) – (1/3) * log2(1/3) = 0.918

Finally, we can calculate the overall entropy for the split based on attribute B using a weighted average of the entropy values for the two nodes. The weights are the proportions of data points in each node.

Overall Entropy for Split based on B:

Average Weighted Entropy = (6/9) * 0.918 + (3/9) * 0.918 = 0.918

So, the overall entropy for the split based on attribute B is approximately 0.918. This value is a measure of the impurity of the nodes after the split. The lower the entropy, the better the split.

4. **Which attribute gives a purer split?**

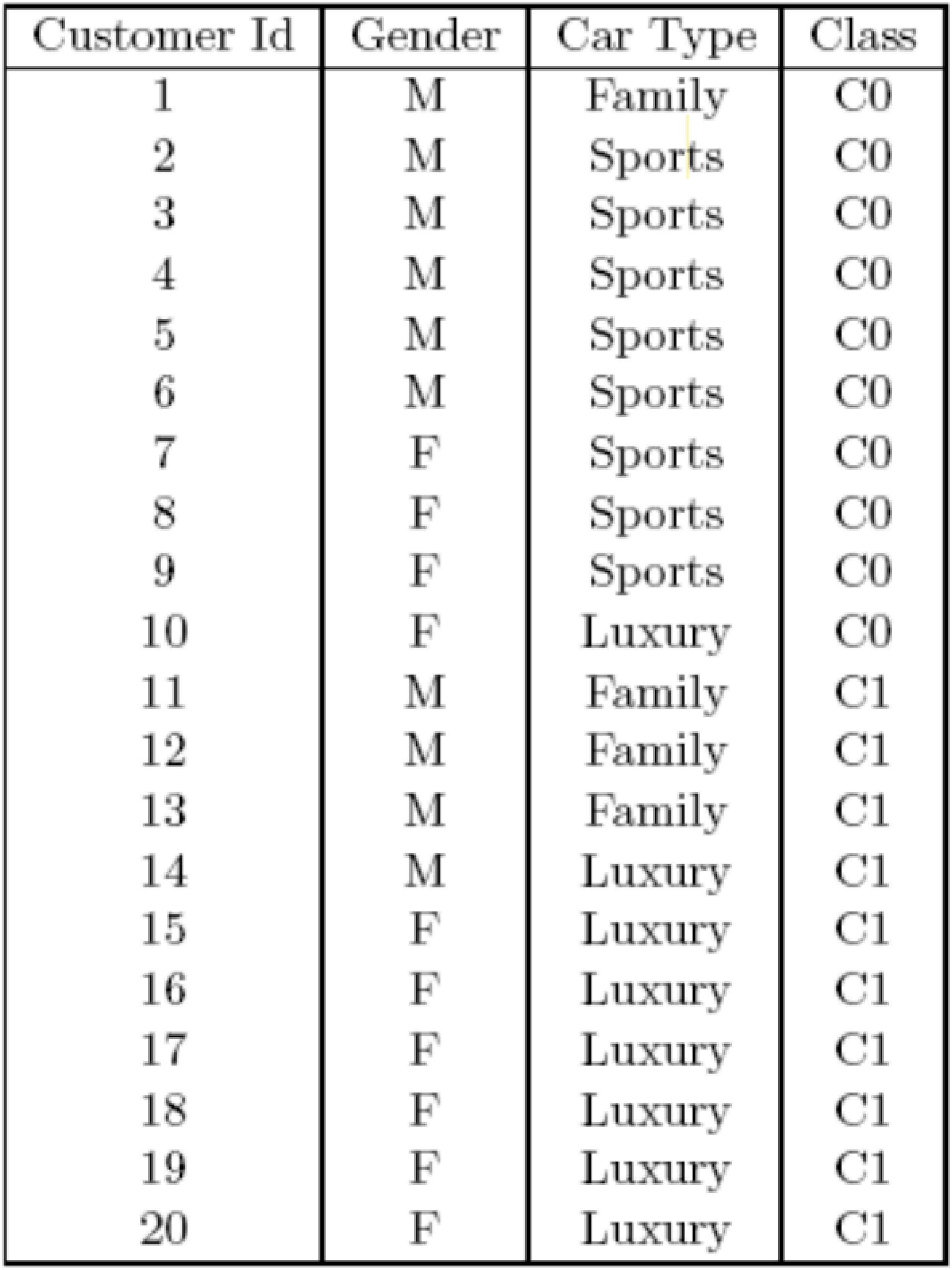
The attribute that gives a purer split is the one with the lower entropy after the split. From our previous calculations:

- The overall entropy for the split based on attribute A is approximately 0.983.
- The overall entropy for the split based on attribute B is approximately 0.918.

Therefore, **attribute B gives a purer split** because it has a lower entropy after the split. This means that splitting on attribute B results in child nodes that are more pure (i.e., contain more instances of a single class) compared to splitting on attribute A. In the context of decision tree models, purer nodes are generally preferred as they lead to more accurate classifications.

Building a Decision Tree for a Given Dataset

For the dataset below, show how the final decision tree will look like. To build the tree, calculate Gini values for all possible splits, select the split with the lowest Gini, repeat these steps for any impure nodes until you run out of attribute splits, or until all the leaf nodes are pure.



Customer Id	Gender	Car Type	Class
1	M	Family	C0
2	M	Sports	C0
3	M	Sports	C0
4	M	Sports	C0
5	M	Sports	C0
6	M	Sports	C0
7	F	Sports	C0
8	F	Sports	C0
9	F	Sports	C0
10	F	Luxury	C0
11	M	Family	C1
12	M	Family	C1
13	M	Family	C1
14	M	Luxury	C1
15	F	Luxury	C1
16	F	Luxury	C1
17	F	Luxury	C1
18	F	Luxury	C1
19	F	Luxury	C1
20	F	Luxury	C1

- **Split - 1: Gender (M, F)**

    - Node 1 (Gender = M) contains 10 data points:
        - 6 belong to class C0
        - 4 belong to class C1
        - Gini Index for Node 1 (Gender = M):
            - 1 – (6/10) ^ 2 – (4/10) ^ 2 = 0.48

    - Node 2 (Gender = F) contains 10 data points:
        - 4 belong to class C0
        - 6 belongs to class C1
        - Gini Index for Node 2 (Gender = F):
            - 1 – (4/10) ^ 2 – (6/10) ^ 2 = 0.48

    - Overall Gini Index for split Gender(M, F):
        - Average Weighted Gini = (10/20) * 0.48 + (10/20) * 0.48 = 0.48

- **Split - 2: Car Type (Family, Sports, Luxury)**

    - Node 1 (CarType = Family) contains 4 data points:
        - 1 belong to class C0
        - 3 belong to class C1
        - Gini Index for Node 1 (CarType = Family):
            - 1 – (1/4) ^ 2 – (3/4) ^ 2 = 0.375

    - Node 2 (CarType = Sports) contains 8 data points:
        - 8 belong to class C0
        - 0 belongs to class C1
        - Gini Index for Node 2 (CarType = Sports):
            - 1 – (8/8) ^ 2 – (0/8) ^ 2 = 0

    - Node 3 (CarType = Luxury) contains 8 data points:
        - 1 belong to class C1
        - 7 belong to class C2.
        - Gini Index for Node 3 (CarType = Luxury):
            - 1 – (1/8) ^ 2 – (7/8) ^ 2 = 0.218

    - Overall Gini Index for split CarType(Family, Sports, Luxury):

        - Average Weighted Gini = (4/20) * 0.375 + (8/20) * 0 + (8/20) * 218 = 0.163

- **Split - 3: Car Type (Family, Sports+Luxury)**

    - Node 1 (CarType = Family) contains 4 data points:
        - 1 belong to class C0
        - 3 belong to class C1
        - Gini Index for Node 1 (CarType = Family):
            - 1 – (1/4) ^ 2 – (3/4) ^ 2 = 0.375

    - Node 2 (CarType = Sports+Luxury) contains 16 data points:
        - 9 belong to class C0
        - 7 belongs to class C1
        - Gini Index for Node 2 (CarType = Sports+Luxury):
            - 1 – (9/16) ^ 2 – (7/16) ^ 2 = 0.492

    - Overall Gini Index for split CarType(Family, Sports+Luxury):
        - Average Weighted Gini = (4/20) * 0.375 + (16/20) * 0.492 = 0.468

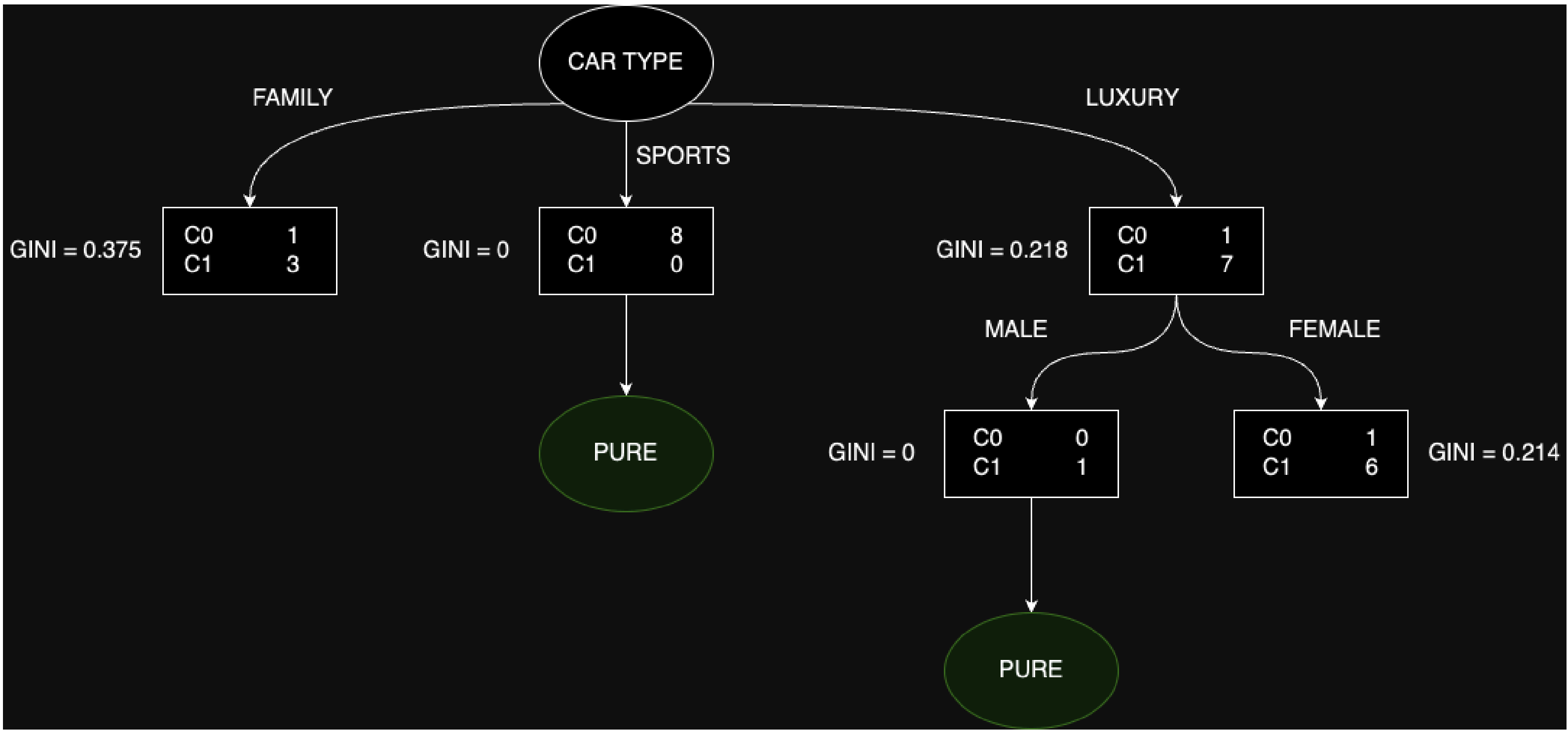
- **Split - 4: Car Type (Sports, Family+Luxury)**

    - Node 1 (CarType = Sports) contains 8 data points:
        - 8 belong to class C0
        - 0 belong to class C1
        - Gini Index for Node 1 (CarType = Sports):
            - 1 – (8/8) ^ 2 – (0/8) ^ 2 = 0

    - Node 2 (CarType = Family+Luxury) contains 12 data points:
        - 2 belong to class C0
        - 10 belongs to class C1
        - Gini Index for Node 2 (CarType = Family+Luxury):
            - 1 – (2/12) ^ 2 – (10/12) ^ 2 = 0.277

    - Overall Gini Index for split CarType(Sports, Family+Luxury):
        - Average Weighted Gini = (8/20) * 0 + (12/20) * 0.277 = 0.166

- **Split - 5: Car Type (Luxury, Family+Sports)**
    - Node 1 (CarType = Luxury) contains 8 data points:
        - 1 belong to class C0
        - 7 belong to class C1
        - Gini Index for Node 1 (CarType = Luxury):
            - 1 – (1/8) ^ 2 – (7/8) ^ 2 = 0.218
    - Node 2 (CarType = Family+Sports) contains 12 data points:
        - 9 belong to class C0
        - 3 belongs to class C1
        - Gini Index for Node 2 (CarType = Family+Sports):
            - 1 – (9/12) ^ 2 – (3/12) ^ 2 = 0.375
    - Overall Gini Index for split CarType(Luxury, Family+Sports):
        - Average Weighted Gini = (8/20) * 0.218 + (12/20) * 0.375 = 0.312


CAR TYPE
FAMILY
SPORTS
LUXURY
GINI = 0.375
C0 1
C1 3
GINI = 0
C0 8
C1 0
GINI = 0.218
C0 1
C1 7
MALE
FEMALE
PURE
GINI = 0
C0 0
C1 1
C0 1
C1 6
GINI = 0.214
PURE