1) The answers to question 1 (a, b, c) can be found in the link provided below:

https://colab.research.google.com/drive/1PgS7IdojuZ1_fuke0vLvGGXchoGav3Hs#scrollTo=oZuFirNZYp1E&line=38&uniqifier=1

Note: I have also attached the .py file to the assignment.

2) A) Compute the Gini index for the entire training set and each attribute (i.e., Car Type, Shirt Size, Employed). For categorical variables with more than two categories, please compute the Gini index using multiway split.

Gini index for the entire training set:

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.5)² - (0.5)² = 1 - 0.25 - 0.25 = 0.5**

—----------

Gini index for each attribute:
Car Type:
P(Family) = 3 positive , 1 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.75)² - (0.25)² = 1 - 0.56 - 0.06 = 0.38**

P(Sports) = 0 positive , 8 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0)² - (1)² = 1 - 0 - 1 = 0**

P(Luxury) = 7 positive , 1 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.875)² - (0.125)² = 1 - 0.77 - 0.02 = 0.22**

Overall Gini for cart type is (0.2 * 0.38 + 0.4*0 + 0.4 * 0.22 = 0.16)

—----------

Shirt Size:

P(S) = 2 positive , 3 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.4)² - (0.6)² = 1 - 0.16 - 0.36 = 0.48**

P(M) = 4 positive , 3 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.57)² - (0.43)² = 1 - 0.32 - 0.18 = 0.5**

P(L) = 2 positive , 2 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.5)² - (0.5)² = 1 - 0.25 - 0.25 = 0.5**

P(XL) = 2 positive , 2 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.5)² - (0.5)² = 1 - 0.25 - 0.25 = 0.5**

Overall Gini for shirt size is (0.25 * 0.48 + 0.35*0.5 + 0.2 * 0.5 + 0.2 *0.5 = 0.495)

—----------

Employed:

P(Yes) = 4 positive , 6 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.4)² - (0.6)² = 1 - 0.16 - 0.36 = 0.48**

P(No) = 4 positive , 6 negative

Gini Index = 1- $\sum_{i=0}^{c-1} p_i(t)^2$ **= 1- (0.4)² - (0.6)² = 1 - 0.16 - 0.36 = 0.48**

Overall Gini for employed is (0.48 * 0.5 + 0.48 *0.5 = 0.48)

—----------

B) Using Gini index, which attribute is better to be used for splitting (to build a decision tree) and why?
Car Type is the best because it has the lowest gini among the three attributes.

C) Build a two-level decision tree using the greedy approach and the Gini index as the splitting criterion. Please provide the details of computations conducted for building the tree.

First-level Split Selection: Car Type has the lowest Gini index (0.16), so I choose Car Type as the first-level split.
Second-level Split: I choose the best split for each branch: family, sport, and luxury. So, I will split the dataset into three branches.

Branch: Family (3 positive, 1 negative)
Shirt Size for family:

Small: 0 positive, 1 negative
Medium: 1 positive, 0 negative
Large: 1 positive, 0 negative
XL: 1 positive, 0 negative

Since all subsets except Small are pure, the Gini index for Shirt Size in the Family branch is: **Gini = (0.25 * 0 + 0.25 * 0 + 0.25 * 0 + 0.25 * 0 = 0)**

Employed for family:

Yes: 3 positive, 1 negative
No: No data and no split

Gini yes = 1 - $(0.75)^2$ - $(0.25)^2$ = 1 - 0.56 - 0.06 = 0.38

Since Shirt Size has a lower Gini index (0) than Employed (0.38), Shirt Size is the better second-level split for the Family branch.

—----------

Branch: Sports (0 positive, 8 negative):
All labels are negative, so no further split is needed. The branch is pure.

—----------

Branch: Luxury (7 positive, 1 negative):
Shirt Size for Luxury:

S: 2 positive, 0 negative
Gini(S) = 1 - $(1)^2$ - $(0)^2$ = 0
M: 3 positive, 0 negative
Gini(M) = 1 - $(1)^2$ - $(0)^2$ = 0
L: 1 positive, 1 negative
Gini(L) = 1 - $(0.5)^2$ - $(0.5)^2$ = 0.5
XL: 1 positive, 0 negative
Gini(XL) = 1 - $(1)^2$ - $(0)^2$ = 0
Overall Gini(Luxury,Shirt Size) = (0.29 * 0 + 0.43 * 0 + 0.29 * 0.5 + 0.14 * 0 = 0.145)

Employed for Luxury:

Yes: 1 positive, 0 negative
Gini(Yes) = 1 - $(1)^2$ - $(0)^2$ = 0
No: 6 positive, 1 negative
Gini(No) = 1 - $(6/7)^2$ - $(1/7)^2$ = 1 - 0.73 - 0.02 = 0.25

Overall Gini(Luxury,Employed) = (0.125 * 0 + 0.875 * 0.25 = 0.22)

The overall Gini index for the Employed attribute in the Luxury branch is 0.22. Since the Gini index for Shirt Size is still lower (0.145), splitting on Shirt Size remains the better option for the Luxury branch.

The best first-level split is Car Type, and the best second-level split for both the Family and Luxury branches is Shirt Size. The Sports branch is pure, so no split is required.

D) Based on the decision tree built in part (c), compute the confusion matrix, accuracy, precision, recall, and F1 for the training set.

Car type	Shirt Size	Actual Class Label	Predicted label
Family	S	-	-
Sports	M	-	-
Sports	M	-	-
Sports	L	-	-
Sports	XL	-	-
Sports	XL	-	-
Sports	S	-	-
Sports	S	-	-
Sports	M	-	-
Luxury	L	-	+
Family	L	+	+
Family	XL	+	+
Family	M	+	+
Luxury	XL	+	+
Luxury	S	+	+
Luxury	S	+	+
Luxury	M	+	+
Luxury	M	+	+
Luxury	M	+	+
Luxury	L	+	-

Predictions for the Family branch:

- Small: Negative (-)
- Medium: Positive (+)
- Large: Positive (+)
- XL: Positive (+)

Predictions for the Sports branch: All predictions: Negative (-)

Predictions for the Luxury branch:

- Small: Positive (+)
- Medium: Positive (+)
- Large: Negative (-)
- XL: Positive (+)

Confusion Matrix:

Predicted\Actual	Positive(Actual +)	Negative (Actual -)
Positive	9	1
Negative	1	9

Accuracy =
$(TP + TN)/(TP + TN + FP + FN) = (9 + 9)/(9 + 9 + 1 + 1) = (18/20) = 0.9$
Accuracy is 90%
Precision = $(TP/(TP/FP) = 9/(9 + 1) = 0.9$
Recall = $(TP/(TP/FP) = 9/(9 + 1) = 0.9$
F1-score = 2*((Precision*Recall)/(Precision+Recall)) = 2*(0.81/1.8) = 0.9
F1-score = 90%