Data Mining Assignment Q1
Name: Akhlaq Ahmed
Student ID: 223195551

1. **Plot the decision tree that can illustrate the traffic conflict problem in Table 1.**

	traffic_volume_low	traffic_volume_medium	driver_age_Old	driver_age_Young	license_yes	visibility_fair	conflict_conflict_yes
0	0	0	0	1	0	1	0
1	0	0	0	1	0	0	0
2	0	0	0	0	0	1	1
3	0	0	0	0	1	1	1
4	0	1	0	1	0	1	0
5	0	1	1	0	1	1	1
6	0	1	0	1	1	0	1
7	0	1	0	0	0	0	1
8	0	1	1	0	0	0	0
9	0	1	1	0	0	1	1
10	1	0	1	0	1	1	1
11	1	0	1	0	1	0	0
12	1	0	0	0	1	0	1
13	1	0	0	1	1	1	1

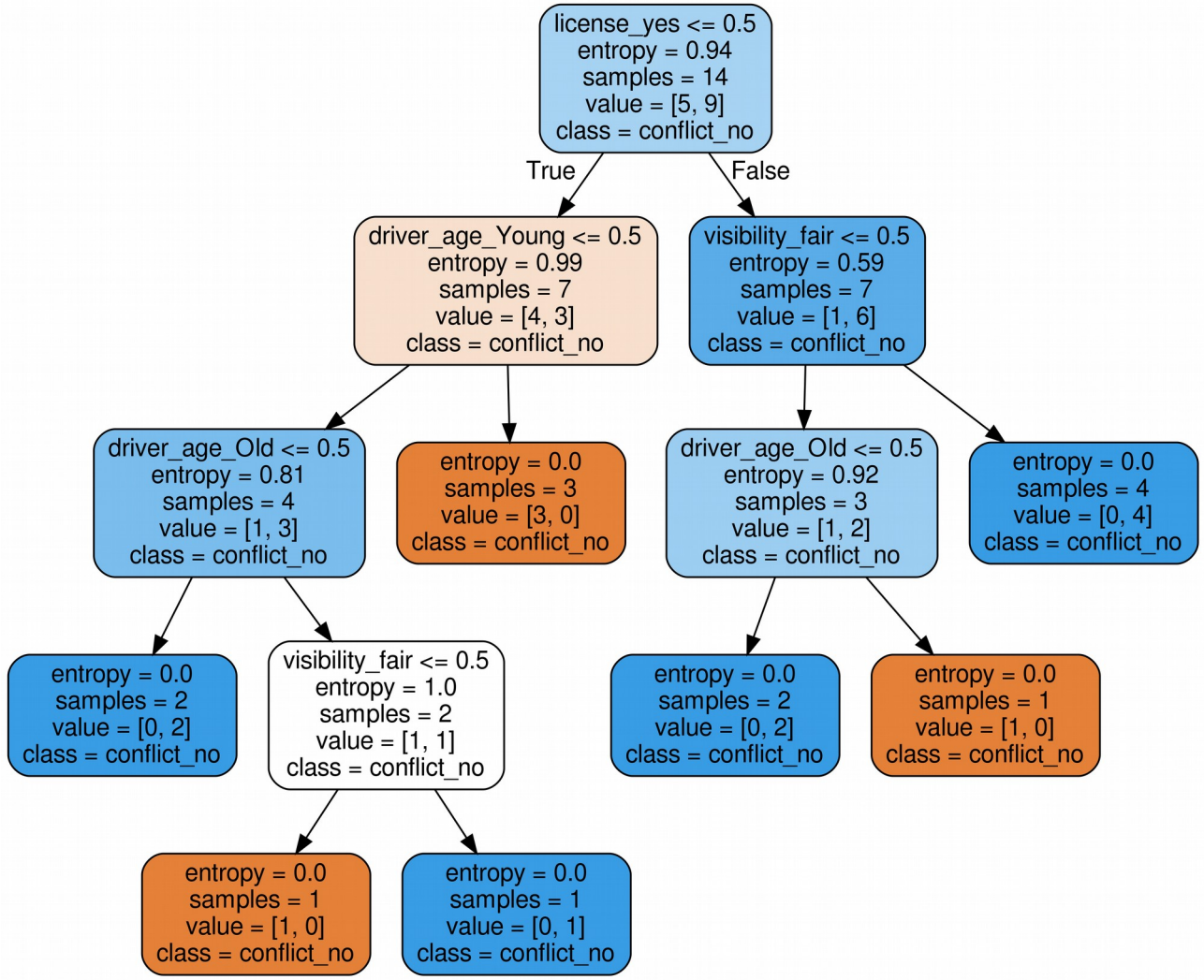
In [118]:

```
dtree=DecisionTreeClassifier(criterion='entropy')
dtree.fit(X,y)

from sklearn.tree import export_graphviz
# Export as dot file
export_graphviz(dtree, out_file='tree1.dot',
                feature_names = X.columns,
                class_names = df['conflict'],
                rounded = True, proportion = False,
                precision = 2, filled = True)

# Convert to png
from subprocess import call
call(['dot', '-Tpng', 'tree1.dot', '-o', 'tree1.png', '-Gdpi=600'])

# Display in python
import matplotlib.pyplot as plt
plt.figure(figsize = (14, 18))
plt.imshow(plt.imread('tree1.png'))
plt.axis('off');
plt.show();
```



2. Fill in Table 2 for the positive counts of conflict (p_i) and negative counts of conflict (n_i), and their totals.

Table 2. Positive and Negative Counts of Conflicts

Traffic Volume	p_i	n_i	Total
Low	3	1	4
Medium	4	2	6
High	2	2	4
Total	9	5	14

3. What is the probability of the class “conflict->yes” and the class “conflict -> no”?

```
In [86]: pivot_table
```

Out[86]:

	driver_age			license			visibility			sum	probability
traffic_volume	high	low	medium	high	low	medium	high	low	medium		
conflict											
conflict_no	2	1	2	2	1	2	2	1	2	15.0	0.357143
conflict_yes	2	3	4	2	3	4	2	3	4	27.0	0.642857

4. Calculate the information gained by branching on attribute “Traffic Volume” using ID3/C4.5 method.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

$$Gain(9,5) = -\frac{9}{14}\log_2 \frac{9}{14} - \frac{5}{14}\log_2 \frac{5}{14} = 0.94$$

$$Gain_{TrafficVolume}(D) = \frac{4}{14}I(3,1) + \frac{6}{14}I(4,2) + \frac{4}{14}I(2,2) = 0.91$$

$$Gain(Traffic - Volume) = 0.94 - 0.91 = 0.029$$