1. (30 pts) Pattern recognition is a data analytics method that uses machine learning algorithms to automatically recognize patterns and regularities in data. The decision tree is one of the machine learning algorithms used in pattern recognition. Please look closely at the following synthetic dataset and see if you can recognize some patterns. It may be difficult for the human brain even if the dataset is small. Your task is to draw a binary decision tree (by hand, no coding please) which will help to recognize those patterns (classification). The class or the target is this question are Y meaning that one will buy a car and N meaning that he/she will not, based on the color, year, mileage, and model.

Color	Year	Mileage	Model	Class
Blue	2015	110000	Benz	N
Green	2018	70000	Honda	N
Blue	2016	120000	Honda	Y
Purple	2020	8000	Chevy	N
Blue	2020	100000	Nissan	Y
Glay	2016	100000	Honda	N
Bleu	2016	150000	Toyota	N
Red	2010	120000	Benz	Y
White	2016	100000	Honda	N
Bleu	2017	120000	Nissan	Y
Red	2020	50000	Toyota	N
Purple	2006	200000	Honda	N
Red	2014	90000	Benz	Y
Blue	2015	100000	Toyota	Y

In order to draw the binary tree, we first have to determine the root node for the tree, since there are four independent variables. The CART algorithm was selected, which employs the Gini impurity. The GIni index of the class variable is chosen to determine the best feature to split at each node of the tree.

GIni index is calculated as:

$Gini(D) = 1 - \sum_{j=1}^{n} p_j^2$, where p_j is the relative frequency of class j in D

$= 1 - [(8/14)^2 + (6/14)^2]$

= 0.4849

8 and 6 represent the frequencies of the target variable values N and Y respectively.
The average weighted Gini impurity of the independent variables as a function of the dependent variables, with corresponding tables, as follows:

G(x, y) = Gini(x, y) = $1 - [(x/(x + y))^2 + (y/(x + y))^2]$

Gini(D, mileage) = 1/14*G(0, 1) + 1/14*G(1, 0) + 3/14*G(3, 0) + 1/14*G(0, 1) + 4/14*G(2, 2) + 1/14*G(0, 1)
= **0.142857**

		buy		
		y	n	total
mileage	110000	0	1	1
	90000	1	0	1
	120000	3	0	3
	70000	0	1	1
	100000	2	2	4
	200000	0	1	1
	50000	0	1	1
	150000	0	1	1
	8000	0	1	1
				14

Gini(D, year) = 1/14*G(0, 1) + 1/14*G(1, 0) + 1/14*G(1, 0) + 2/14*G(1, 1)+ 1/14*G(1, 0) + 1/14*G(0, 1) + 3/14*G(3, 0) + 4/14*G(2, 2)
= 0.273809

		buy		
		y	n	total
year	2006	0	1	1
	2010	1	0	1
	2014	1	0	1
	2015	1	1	2
	2016	1	3	4
	2017	1	0	1
	2018	0	1	1
	2020	1	2	3
				14

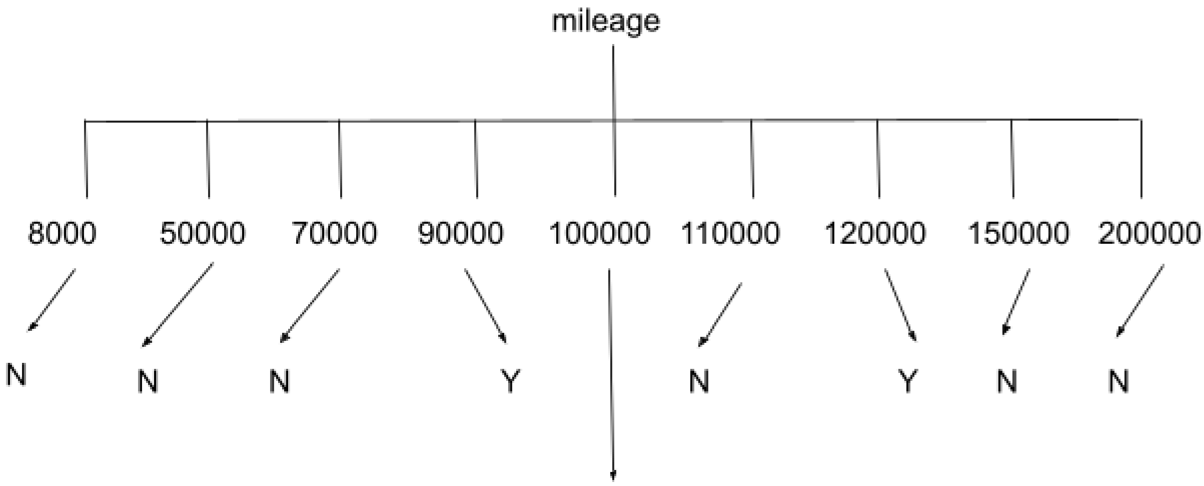
Gini(D, color) = 6/14*G(4, 2) + 3/14*G(2, 1) + 1/14*G(0, 1) + 1/14*G(0, 1) + 1/14*G(0, 1) + 2/14*G(0, 2)
= 0.285714

		buy		
		y	n	total
color	blue	4	2	6
	red	2	1	3
	Green	0	1	1
	Gray	0	1	1
	White	0	1	1
	purple	0	2	2
				14

Gini(D, model) = 3/14*G(2, 1) + 5/14*G(1, 4) + 3/14*G(1, 2)+ 2/14*G(2, 0)+ 1/14*G(0, 1)
= 0.3047617

		buy		
		y	n	total
model	Benz	2	1	3
	Honda	1	4	5
	Toyota	1	2	3
	Nissan	2	0	2
	Chevy	0	1	1
				14

From the Gini index values, mileage is the best feature to represent the root node, as it has the lowest GIni impurity. GIven mileage as the root node, the tree is drawn, and the next node is determined by repeating the process. The following shows the initial tree, and the second node determination.



Color	Year	Model	Class
Glay	2016	Honda	**N**
White	2016	Honda	**N**
Blue	2015	Toyota	Y
Blue	2020	Nissan	Y

$$Gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

$$= 1 - [(2/4)^2 + (2/4)^2]$$

= 0.5

Gini(T, model) = 2/4*G(0, 2) + 1/4*G(1, 0) + 1/14*G(1, 0) = **0**

		buy		
		y	n	total
model	Honda	0	2	2
	Toyota	1	0	1
	Nissan	1	0	1
				4

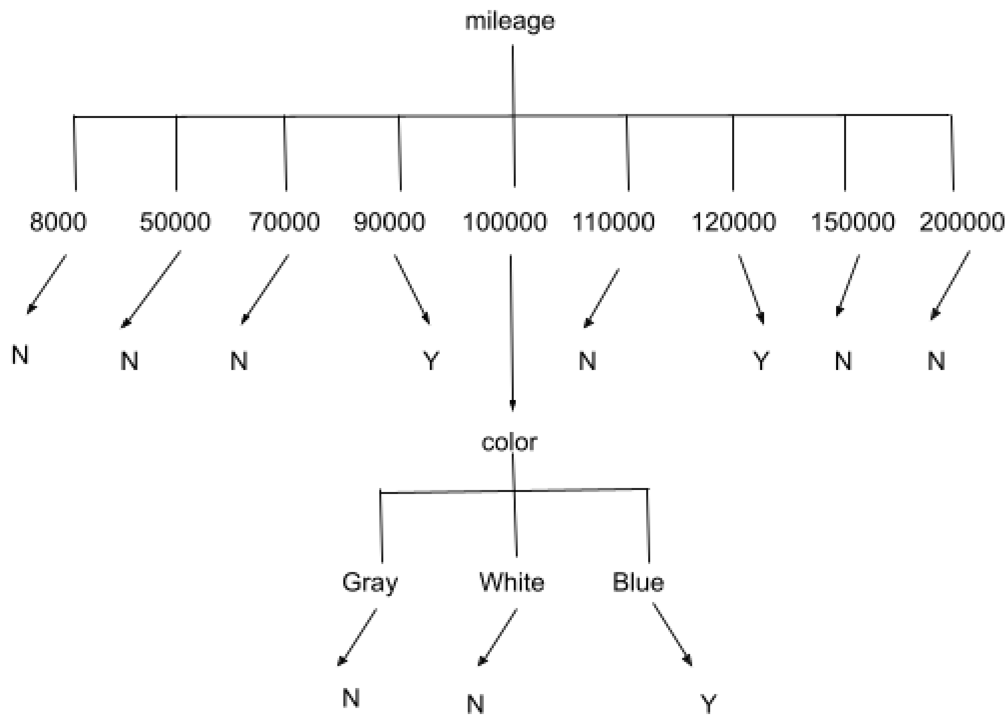
Gini(T, color) = 1/4*G(0, 1) + 1/14*G(0, 1) + 2/4*G(2, 0) = **0**

		buy		
		y	n	total
color	Glay	0	1	1
	White	0	1	1
	Blue	2	0	2
				4

Gini(T, year) = 1/4*G(1, 0) +2/4*G(0, 2) + 1/14*G(1, 0) = **0**

		buy		
		y	n	total
year	2015	1	0	1
	2016	0	2	2
	2020	1	0	1
				4

The features have the same Gini impurity values, hence they have the same importance. The final tree is shown below


mileage
8000
50000
70000
90000
100000
110000
120000
150000
200000
N
N
N
Y
N
Y
N
N
color
Gray
White
Blue
N
N
Y

2. The following is a balanced portion of IRIS dataset. Use the KNN algorithm and Manhattan distance to predict the variety of the new example (in red). Do not use any programming language, this is a hand calculation question, and show all your calculations.

Sepal length	Sepal width	Petal length	Petal width	variety
7.9	3.8	6.4	2	Virginica
5.1	3.5	1.4	0.2	Setosa
6.3	2.8	5.1	1.5	Virginica
6.1	2.6	5.6	1.4	Virginica
4.9	3	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
6.4	2.8	5.6	2.2	Virginica
4.6	3.1	1.5	0.2	Setosa
5	3.6	1.4	0.2	Setosa
7.4	2.8	6.1	1.9	Virginica
6.6	2.9	4.6	1.3	?

Sepal length	Sepal width	Petal length	Petal width	Distance
7.9	3.8	6.4	2	|7.9-6.6+3.8-2.9+6.4-4.6+2-1.3| = **4.7**
5.1	3.5	1.4	0.2	|5.1-6.6+3.5-2.9+1.4-4.6+0.2-1.3| = **5.2**
6.3	2.8	5.1	1.5	|6.3-6.6+2.8-2.9+5.1-4.6+1.5-1.3| = **0.3**
6.1	2.6	5.6	1.4	|6.1-6.6+2.6-2.9+5.6-4.6+1.4-1.3| = **0.3**
4.9	3	1.4	0.2	|4.9-6.6+3-2.9+1.4-4.6+0.2-1.3| = **5.9**
4.7	3.2	1.3	0.2	|4.7-6.6+3.2-2.9+1.3-4.6+0.2-1.3| = **6**
6.4	2.8	5.6	2.2	|6.4-6.6+2.8-2.9+5.6-4.6+2.2-1.3| = **1.6**
4.6	3.1	1.5	0.2	|4.6-6.6+3.1-2.9+1.5-4.6+0.2-1.3| = **6**
5	3.6	1.4	0.2	|5-6.6+3.6-2.9+1.4-4.6+0.2-1.3| = **5.2**
7.4	2.8	6.1	1.9	|7.4-6.6+2.8-2.9+6.1-4.6+1.9-1.3| = **2.8**
6.6	2.9	4.6	1.3	?

Sepal length	Sepal width	Petal length	Petal width	Distance	variety	rank
6.3	2.8	5.1	1.5	**0.3**	Virginica	1
6.1	2.6	5.6	1.4	**0.3**	Virginica	2
6.4	2.8	5.6	2.2	**1.6**	Virginica	3
7.4	2.8	6.1	1.9	**2.8**	Virginica	4
7.9	3.8	6.4	2	**4.7**	Virginica	5
5.1	3.5	1.4	0.2	**5.2**	Setosa	
5	3.6	1.4	0.2	**5.2**	Setosa	
4.9	3	1.4	0.2	**5.9**	Setosa	
4.7	3.2	1.3	0.2	**6**	Setosa	
4.6	3.1	1.5	0.2	**6**	Setosa	
6.6	2.9	4.6	1.3		?	

a. From **three** nearest neighbors, the variety of the new sample is class **Virginica**
b. From **five** nearest neighbors, the variety of the new example is class **Virginica**

3. (10 pts) Choosing the optimal value of k in the k-NN (k-Nearest Neighbors) algorithm is an important task, as it can have a significant impact on the performance of the model. Which of the following methods are not used to choose the best k for KNN classifier? This question is tricky, you must choose all and only the right answers.
Ans: **K-means, K-fold cross validation**

4. (5 pts) Ensemble learning: a ML engineer trained a logistic regression, a KNN, and a Naive Bayesian model on a dataset, and then combined the models using a neural network, by voting. What kind of ensemble learning is this?
Ans: **Bagging**

5. (10 pts) The process of transforming raw data into informative attributes, or the original features into a new set of features that are more informative, and compact is called (Only one is correct):
Ans: **Feature Extraction**

6. (5 pts) Is this true or false: SMOTE is Synthetic Majority Undersampling Technique used to address imbalanced datasets.
Ans: **False**

Color	Year	Mileage	Model	Class
Blue	2015	110000	Benz	**N**
Red	2014	90000	Benz	Y
Red	2010	120000	Benz	Y
Green	2018	70000	Honda	**N**
Glay	2016	100000	Honda	**N**
White	2016	100000	Honda	**N**
Purple	2006	200000	Honda	**N**
Blue	2016	120000	Honda	Y
Red	2020	50000	Toyota	**N**
Bleu	2016	150000	Toyota	**N**
Blue	2015	100000	Toyota	Y
Blue	2020	100000	Nissan	Y
Bleu	2017	120000	Nissan	Y
Purple	2020	8000	Chevy	**N**