CS4445 Data Mining and Knowledge Discovery in Databases. B Term 2010

Exam 1 - November 19, 2010

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

NAME: Carolina Ruiz and Yutao Wang

Instructions:

- Show your work and justify your answers
- Use the space provided to write your answers
- Ask in case of doubt

Problem I. Data Mining [10 points]

What is the difference between the following data mining tasks? Explain decisively.

1. **[5 points]** Classification and Regression.

Classification and Regression deal with the construction of a model that is able to predict the target attribute. The difference is that classification is used when the target attribute is nominal, and regression when the target attribute is continuous.

2. **[5 points]** Clustering and Association Analysis.

Clustering is used to group data instances in such a way that instances in the same cluster are more similar to each other than to instances in other clusters. Association analysis is used to find relationships among data attributes.

Problem II. Data Preprocessing [10 points]

The following is the correlation matrix of the predicting attributes (not including the target attribute) of the diabetes dataset described in class:

	preg	plas	pres	skin	insu	mass	pedi	age
preg	1	0.129459	0.141282	-0.08167	-0.07353	0.017683	-0.03352	0.544341
plas		1	0.15259	0.057328	0.331357	0.221071	0.137337	0.263514
pres			1	0.207371	0.088933	0.281805	0.041265	0.239528
skin				1	0.436783	0.392573	0.183928	-0.11397
insu					1	0.197859	0.185071	-0.04216
mass						1	0.140647	0.036242
pedi							1	0.033561
age								1

1. **[5 points]** Based on this correlation matrix, would you remove any of the attributes from consideration? Explain your answer.

No, I wouldn't remove any attributes from consideration as no attribute is highly correlated with any other predicting attribute (note that this list contains only the predicting attributes, not the target attribute).

2. **[5 points]** If you had to eliminate 1 (and just 1) attribute from consideration, which one would you remove? Why?

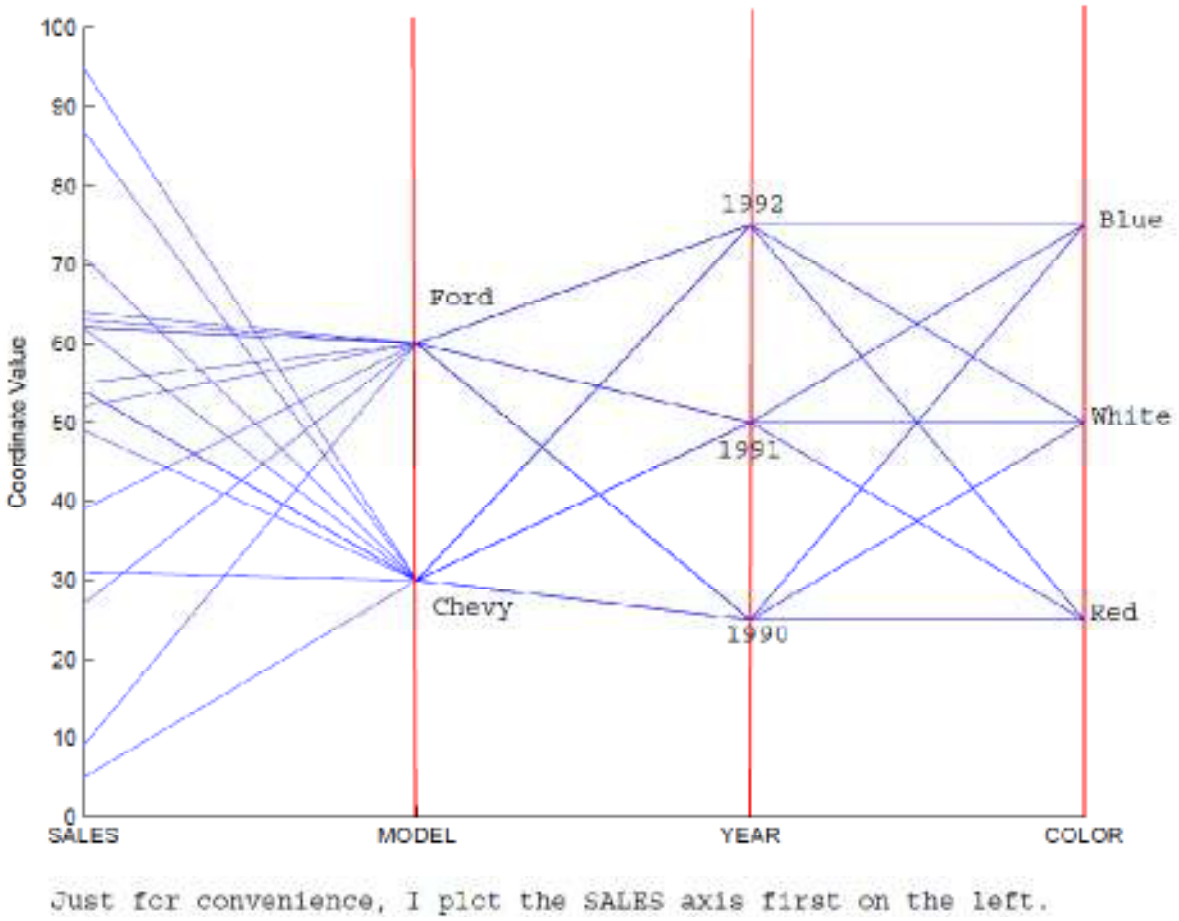
Based on just the correlation matrix above, I'd remove the *skin* attribute as it has the highest correlation values with other attributes overall (they are still low, though). Another possibility would be to eliminate the *age* attribute for similar reasons.

Problem III. Data Exploration [40 points]

MODEL	YEAR	COLOR	SALES
Chevy	1990	Red	5
Chevy	1990	white	87
Chevy	1990	Blue	62
Chevy	1991	Red	54
Chevy	1991	white	95
Chevy	1991	Blue	49
Chevy	1992	Red	31
Chevy	1992	white	54
Chevy	1992	Blue	71
Ford	1990	Red	64
Ford	1990	white	62
Ford	1990	Blue	63
Ford	1991	Red	52
Ford	1991	white	9
Ford	1991	Blue	55
Ford	1992	Red	27
Ford	1992	white	62
Ford	1992	Blue	39

1. **Visualization**
    a. **[5 points]** Construct a parallel coordinates visualization of the above dataset. Label the coordinates and the values on those coordinates. Use the space above on the right.

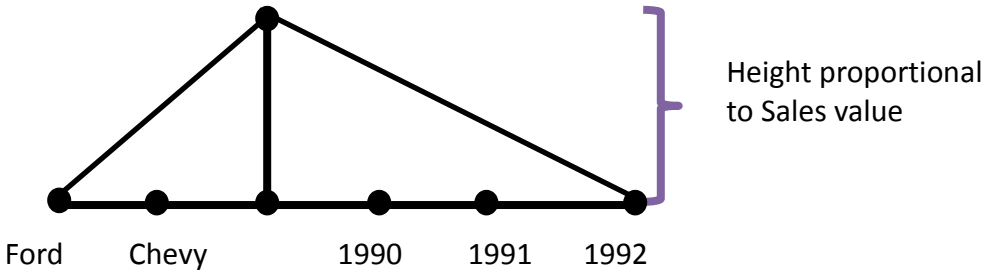
See my parallel coordinates visualization below.



Just for convenience, I plot the SALES axis first on the left.

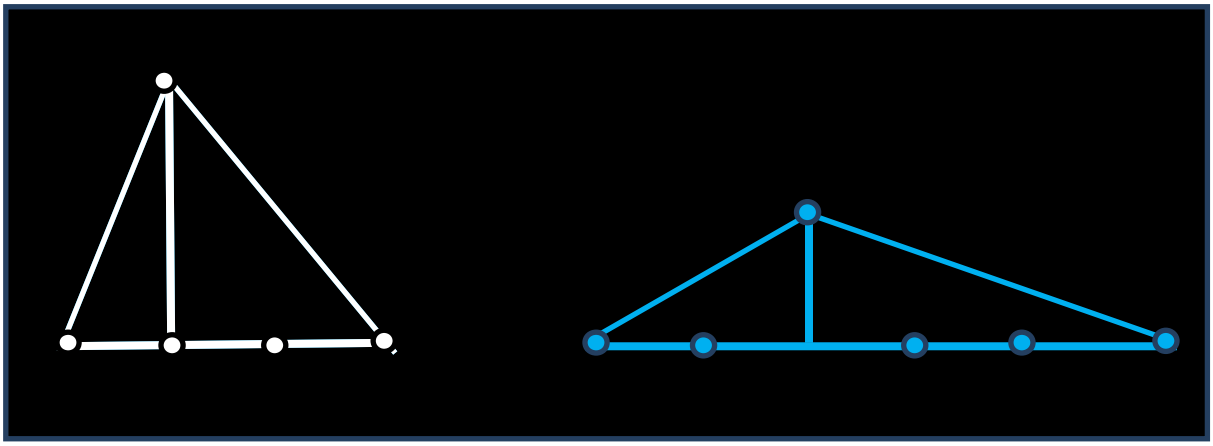
b. **[5 points]** Describe how you would design a star coordinates display of this dataset.

In my design, I use colored triangles to represent a data instance. The triangle color can be red, white, or blue according with the car COLOR. The other 3 attributes are represented as shown in the figure below. The big dots are used to convey to you the lengths of the axes for attributes MODEL and YEAR according to their nominal values.



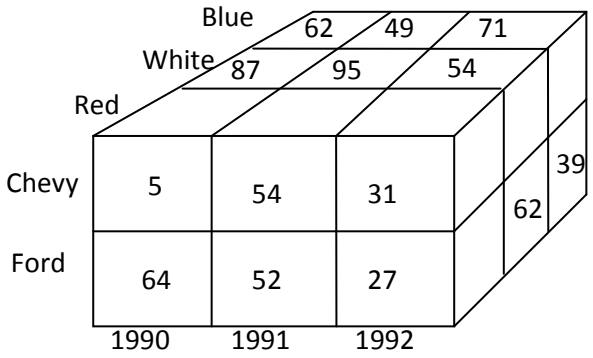
Apply your design to the following two data instances:

Chevy	1991	white	95
Ford	1992	Blue	39



2. **Data warehouses and OLAP.**

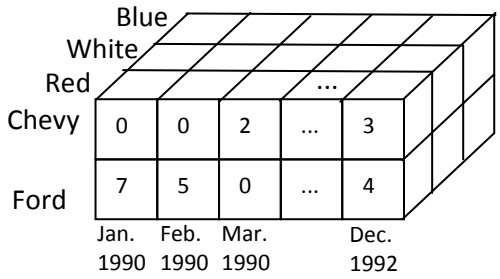
   a. **[5 points]** Depict the 18 instances in the dataset above using a multidimensional cuboid. Use MODEL, YEAR, COLOR as the dimensions, and SALES as the target quantity (that is, the values stored in the cells of the cuboid).



   b. **[5 points]** Illustrate the result of rolling-up MODEL from individual models to **all** (that is, aggregating Chevy and Ford into one).

	1990	1991	1992
Blue	125	104	110
White	149	104	116
Red	69	106	58

   c. **[5 points]** Starting from the cuboid in part (a), depict the result of drilling-down time from YEAR to month. (Make up some SALE values)



d. **[5 points]** Starting from the cuboid in part (a), depict the result of slicing for MODEL=Chevy.

	1990	1991	1992
Blue	62	49	71
White	87	95	54
Red	5	54	31

e. **[5 points]** Starting from the cuboid in part (a), depict the result of dicing for MODEL=Chevy and YEAR=1991.

Red	White	Blue
54	95	49

f. **[5 points]** Starting from the cuboid in part (a), use specific OLAP operations to write an algorithm to obtain the total number of red cars sold.

Slice COLOR = 'Red'
Roll-up on MODEL to all
Roll-up on YEAR to all

Problem III. Decision Trees [30 points]

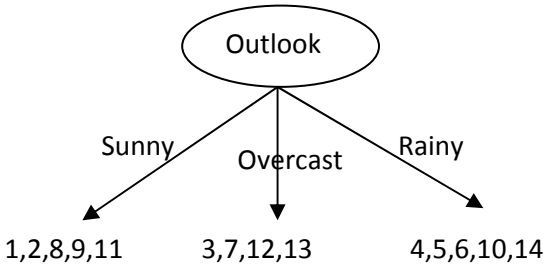
Consider the Weather dataset.

ATTRIBUTES:	POSSIBLE VALUES:
outlook	{sunny, overcast, rainy}
temperature	{hot, mild, cool}
humidity	{high, normal}
windy	{TRUE, FALSE}
play	{no, yes}

Id#	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Consider the ID3 algorithm to construct a decision tree for predicting the attribute **Play**.

1. Assume that root node of the ID3 tree is **Outlook** (you don't need to do entropy calculations to determine the root node. We are telling you that it is **Outlook**).
   [2 points] Depict the root node of the tree with the 3 associated branches.
   [3 points] Include in each branch the **id#s** of the instances that reach that branch.



2. Starting from this root node, construct the FULL decision tree for this dataset USING THE ID3 ALGORITHM. For your convenience, the logarithms in base 2 of selected values are provided.
   [20 points] Show all the steps of the entropy calculations.
   [5 points] Depict the tree at each stage. At each node state what instances (use **id#s**) are included in the node, what attributes are available to split the node, and which attribute is selected based on entropy.

x	1/2	1/3	2/3	1/4	3/4	1/5	2/5	3/5	1/6	5/6	1/7	2/7	3/7	4/7	1
$\log_2(x)$	-1	-1.5	-0.6	-2	-0.4	-2.3	-1.3	-0.7	-2.5	-0.2	-2.8	-1.8	-1.2	-0.8	0

Step 1. Choose an attribute when "Outlook = Sunny":

Id#	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes

The entropy of this node is:

$$Entropy = \underbrace{-\frac{3}{5} * log_2\left(\frac{3}{5}\right)}_{no} \underbrace{-\frac{2}{5} * log_2\left(\frac{2}{5}\right)}_{yes} = 0.94$$

$$Entropy(Temperature) = \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{2}{2} log_2\left(\frac{2}{2}\right)}_{no}\right]}_{Temperature=hot} + \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{1}{2} log_2\left(\frac{1}{2}\right)}_{no} \underbrace{-\frac{1}{2} log_2\left(\frac{1}{2}\right)}_{yes}\right]}_{Temperature=mild} + \underbrace{\frac{1}{5} * \left[\underbrace{-log_2(1)}_{no}\right]}_{Temperature=cool} =$$

0.4

$$Gain(Temperature) = Entropy - Entropy(Temperature) = 0.54$$

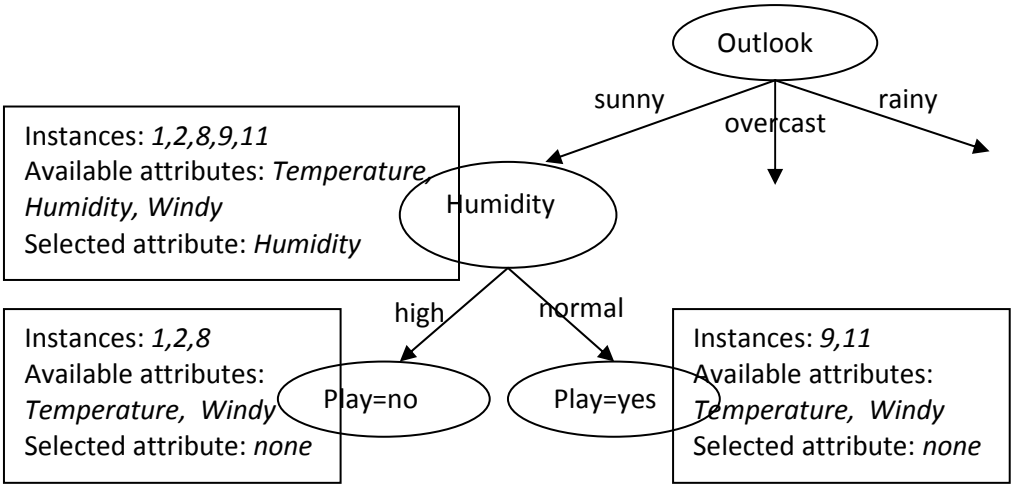
$$Entropy(Humidity) = \underbrace{\frac{3}{5} * \left[\underbrace{-\frac{3}{3} log_2\left(\frac{3}{3}\right)}_{no}\right]}_{Humidity=high} + \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{2}{2} log_2\left(\frac{2}{2}\right)}_{yes}\right]}_{Humidity=normal} = 0$$

$$Gain(Humidity) = Entropy - Entropy(Humidity) = 0.94$$

$$Entropy(Windy) = \underbrace{\frac{3}{5} * \left[\underbrace{-\frac{2}{3}log_2\left(\frac{2}{3}\right)}_{no}\underbrace{-\frac{1}{3}log_2\left(\frac{1}{3}\right)}_{yes}\right]}_{Windy=FALSE} + \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{1}{2}log_2\left(\frac{1}{2}\right)}_{no}\underbrace{-\frac{1}{2}log_2\left(\frac{1}{2}\right)}_{yes}\right]}_{Windy=TRUE} = 0.94$$

$$Gain(Windy) = Entropy - Entropy(Windy) = 0$$

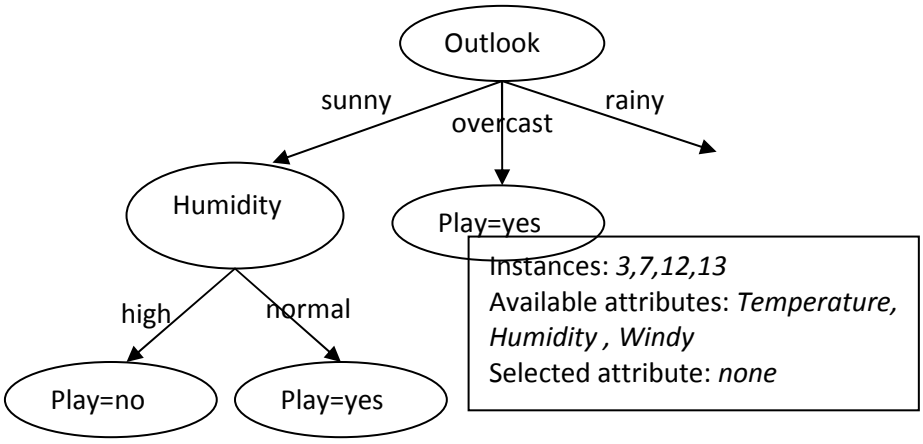
After this step, current decision tree is:



Step 2. Choose an attribute when "Outlook=overcast"

Id#	Outlook	Temperature	Humidity	Windy	Play
3	overcast	hot	High	FALSE	yes
7	overcast	cool	Normal	TRUE	yes
12	overcast	mild	High	TRUE	yes
13	overcast	hot	Normal	FALSE	yes

All the instances in this branch belong to class "yes", so the tree do not need to split at this node, current decision tree is:



Step 3. Choose an attribute when "Outlook=rainy"

Id#	Outlook	Temperature	Humidity	Windy	Play
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
10	rainy	mild	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

The entropy of this node is:

$$Entropy = \underbrace{-\frac{3}{5} * log_2\left(\frac{3}{5}\right)}_{yes}\underbrace{-\frac{2}{5} * log_2\left(\frac{2}{5}\right)}_{no} = 0.94$$

$$Entropy(Temperature) = \underbrace{\frac{3}{5} * \left[\underbrace{-\frac{2}{3}log_2\left(\frac{2}{3}\right)}_{yes}\underbrace{-\frac{1}{3}log_2\left(\frac{1}{3}\right)}_{no}\right]}_{Temperature=mild} + \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{1}{2}log_2\left(\frac{1}{2}\right)}_{yes}\underbrace{-\frac{1}{2}log_2\left(\frac{1}{2}\right)}_{no}\right]}_{Temperature=cool} = 0.94$$

$$Gain(Temperature) = Entropy - Entropy(Temperature) = 0$$

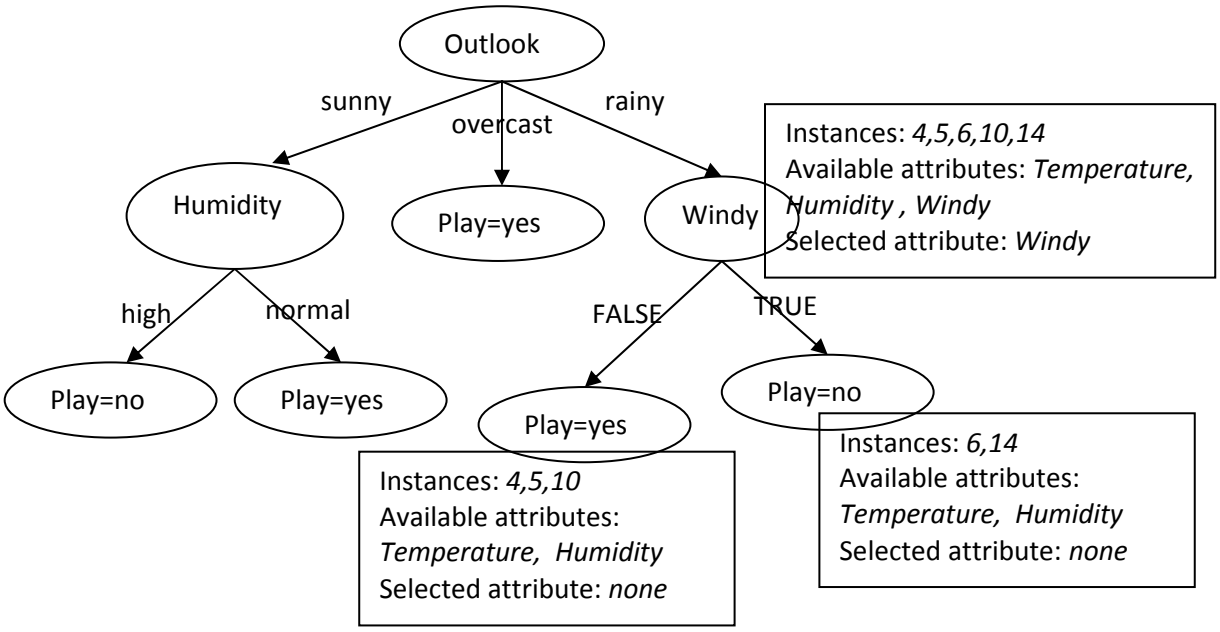
$$Entropy(Humidity) = \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{1}{2}log_2\left(\frac{1}{2}\right)}_{yes}\underbrace{-\frac{1}{2}log_2\left(\frac{1}{2}\right)}_{no}\right]}_{Humidity=high} + \underbrace{\frac{3}{5} * \left[\underbrace{-\frac{2}{3}log_2\left(\frac{2}{3}\right)}_{yes}\underbrace{-\frac{1}{3}log_2\left(\frac{1}{3}\right)}_{no}\right]}_{Humidity=normal} = 0.94$$

$$Gain(Humidity) = Entropy - Entropy(Humidity) = 0$$

$$Entropy(Windy) = \underbrace{\frac{3}{5} * \left[\underbrace{-\frac{3}{3}log_2\left(\frac{3}{3}\right)}_{yes}\right]}_{Windy=FALSE} + \underbrace{\frac{2}{5} * \left[\underbrace{-\frac{2}{2}log_2\left(\frac{2}{2}\right)}_{no}\right]}_{Windy=TRUE} = 0$$

$$Gain(Windy) = Entropy - Entropy(Windy) = 0.94$$

After this step, the final decision tree is:



Problem V. Model Evaluation [20 points].

1. Assume that we are testing a model over a test set that contains 100 instances and that the target class has two possible vales: **yes** and **no**. Assume also that the Confusion Matrix that results from the testing is the following:

   === Confusion Matrix ===

a	b	<-- classified as
60 **TP**	20 **FN**	| a = **yes**
10 **FP**	10 **TN**	| b = **no**

   1. **[3 points]** Calculate the classification accuracy of the model over the test set. Explain your work.

   Accuracy = number of correct predictions / total number of predictions made
   = 60 + 10 / 100 = 70%

   2. **[2 points]** Consider the following notions:
      True positives (TP): the number of test instances correctly classified as **yes**.
      True negatives (TN): the number of test instances correctly classified as **no**.
      False positives (FP): the number of test instances incorrectly classified as **yes**.
      False negatives (FN): the number of test instances incorrectly classified as **no**.

   Label the corresponding entries in the confusion matrix above with TP, TN, FP, and FP.

   3. **[5 points]** Provide a general formula for calculating the classification accuracy of a model in terms of only TP, TN, FP, and FN.

   Accuracy = $\frac{(TP + TN)}{(TP + TN + FP + FN)}$

2. **[10 Points]** Define validation set (not to be confused with cross-validation). Explain what the differences between validation set and training and test sets are. Give an example of when a validation set is used. Explain for example.

Validation set is a subset of the training set that is not used for training, but it is reserved for evaluating/testing the model during its construction. For example, during the construction of a decision tree, a validation set can be used for pruning (either to stop the construction of the tree when splitting a heterogeneous node does not increase the classification accuracy over the validation set; or for post-pruning subtrees of the constructed tree when eliminating those subtrees does not lower the classification accuracy over the validation set). Validation set is different from training set as the data instances in the validation set are not used to grow the model, but only to evaluate it during its construction. Validation set is different from test set in that instances in the validation set are used to evaluate the model during its construction (before it is outputted), and instances in the test set are used to evaluate the model after its construction has been completed (after the model has been outputted).