بسمه تعالی

گروه انفورماتیک پزشکی

Data Mining: Classification

Dr O. Pournik ***MD, MPH, MSc, PhD***

pournik@gmail.com

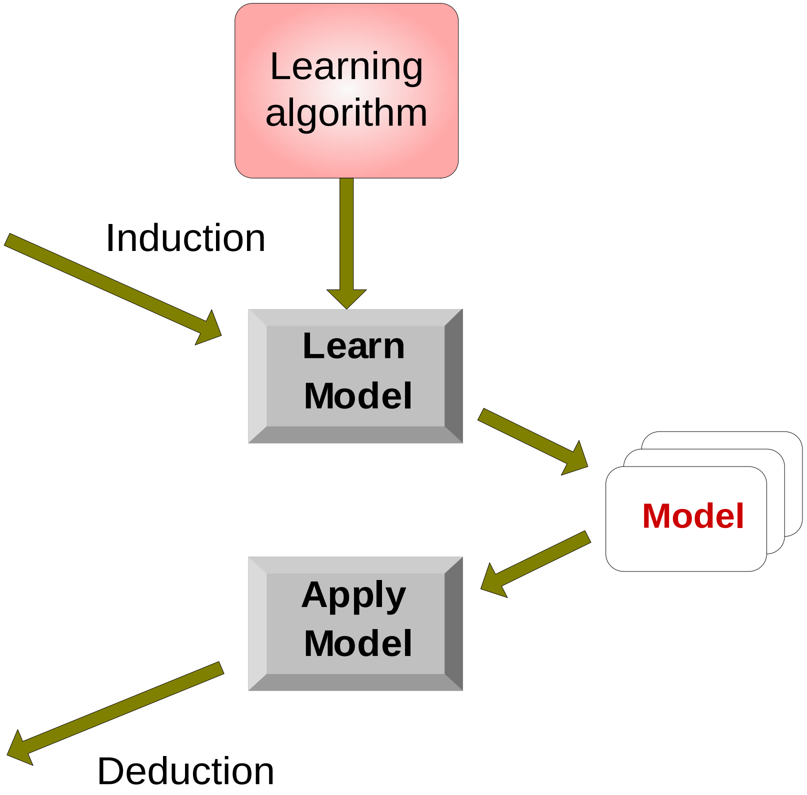
Classification: Definition

- *Given a collection of records (training set)*
  - *Each record contains a set of attributes, one of the attributes is the class.*
- *Find a model for class attribute as a function of the values of other attributes.*
- *Goal: previously unseen records should be assigned a class as accurately as possible.*
  - *A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.*

Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	**No**
2	No	Medium	100K	**No**
3	No	Small	70K	**No**
4	Yes	Medium	120K	**No**
5	No	Large	95K	**Yes**
6	No	Medium	60K	**No**
7	Yes	Large	220K	**No**
8	No	Small	85K	**Yes**
9	No	Medium	75K	**No**
10	No	Small	90K	**Yes**

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	**?**
12	Yes	Medium	80K	**?**
13	Yes	Large	110K	**?**
14	No	Small	95K	**?**
15	No	Large	67K	**?**

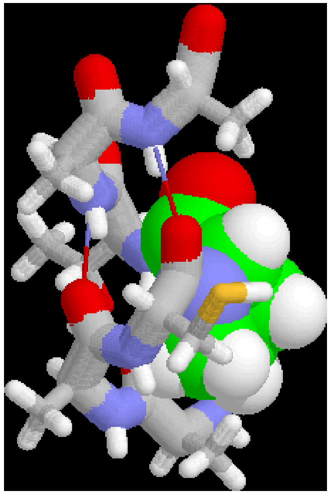
Test Set

Examples of Classification Task

- *Predicting tumor cells as benign or malignant.*

- *Classifying credit card transactions as legitimate or fraudulent.*

- *Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil.*



- *Categorizing news stories as finance,weather, entertainment, sports, etc.*

Classification Techniques

- *Decision Tree based Methods*
- *Rule-based Methods*
- *Memory based reasoning*
- *Neural Networks*
- *Naïve Bayes and Bayesian Belief Networks*
- *Support Vector Machines*

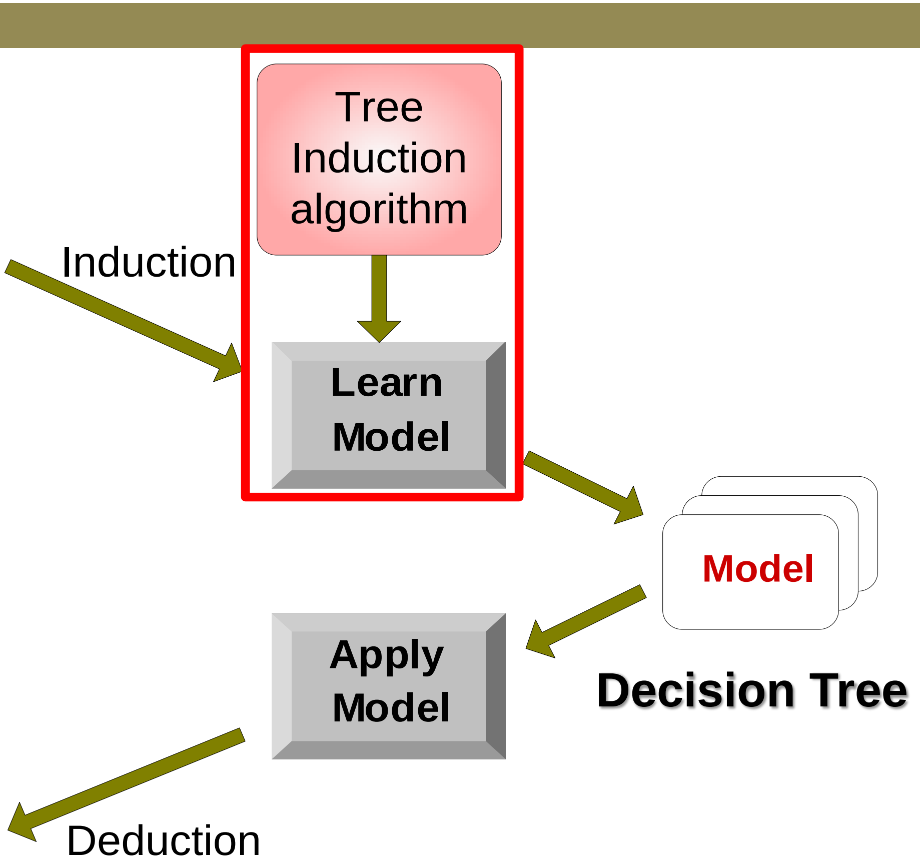
Decision Tree Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



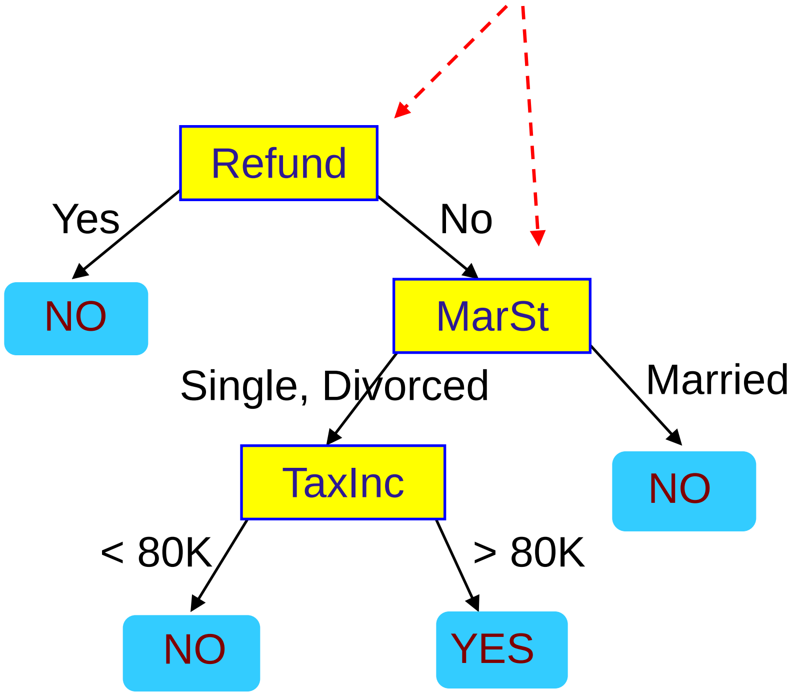
Example of a Decision Tree

categorical categorical continuous class

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Training Data

Splitting Attributes



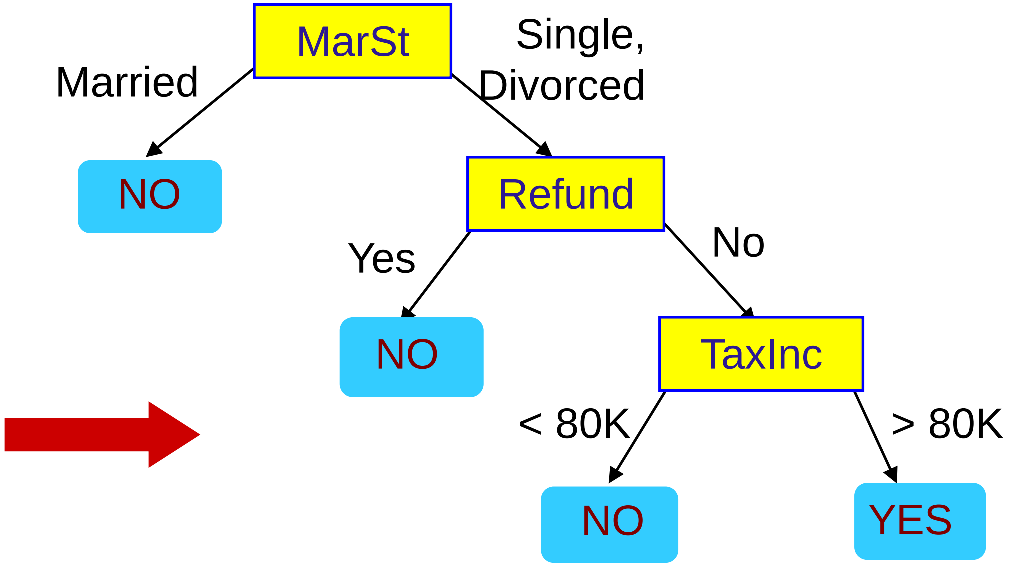
Model: Decision Tree

Another Example of Decision Tree

categorical categorical continuous class

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Training Data

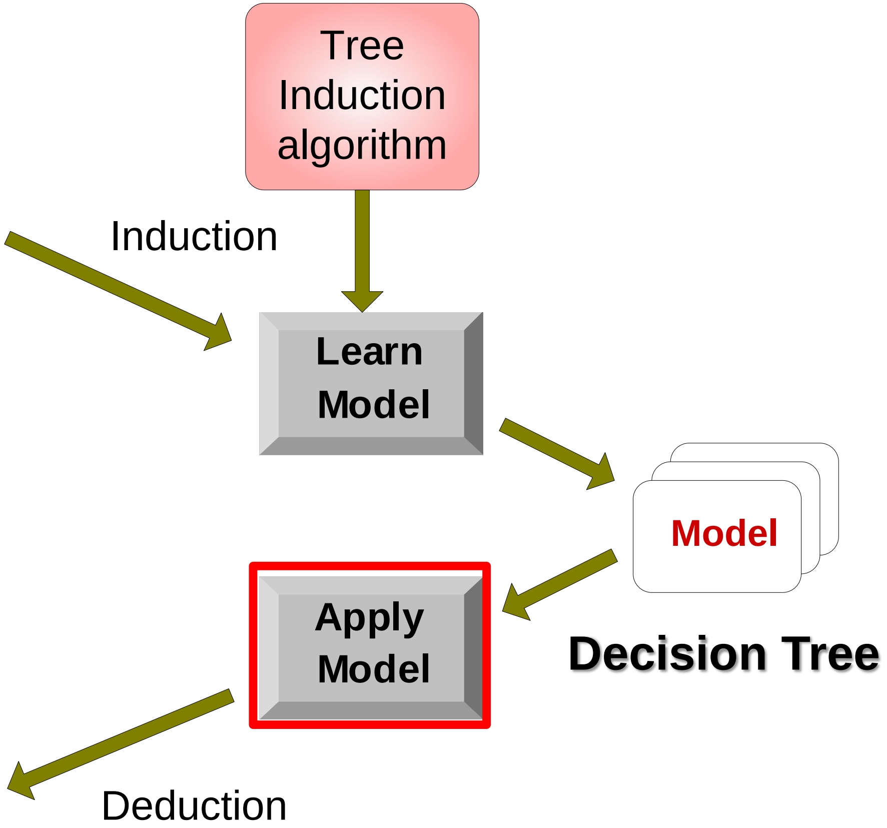


There could be more than one tree that fits the same data!

Decision Tree Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tree Induction algorithm

Induction

Learn Model

Model

Decision Tree

Apply Model

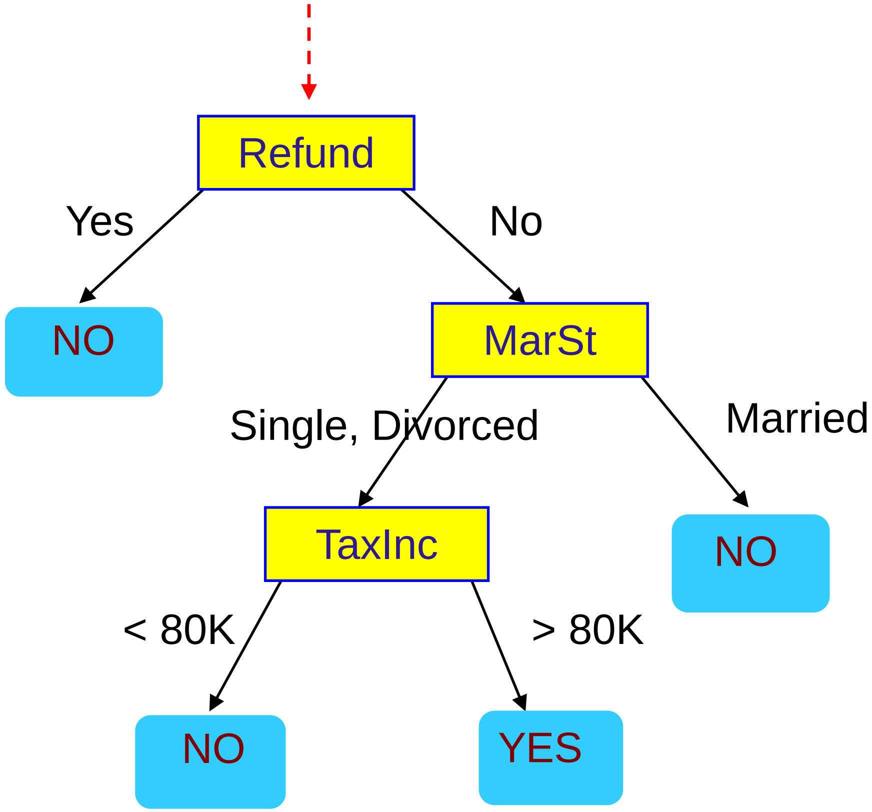
Deduction

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Apply Model to Test Data

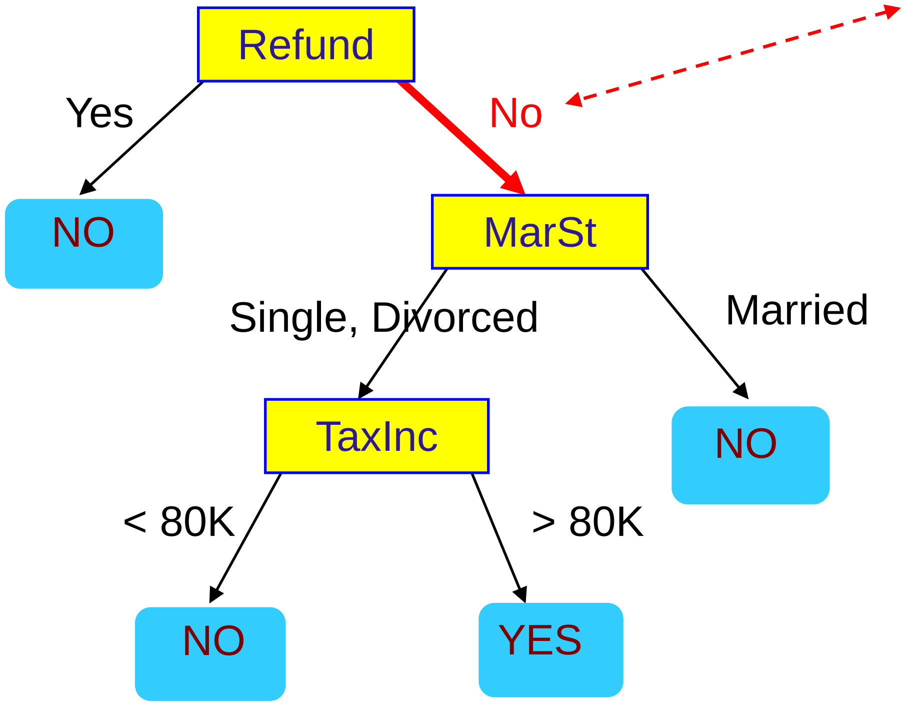
Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data



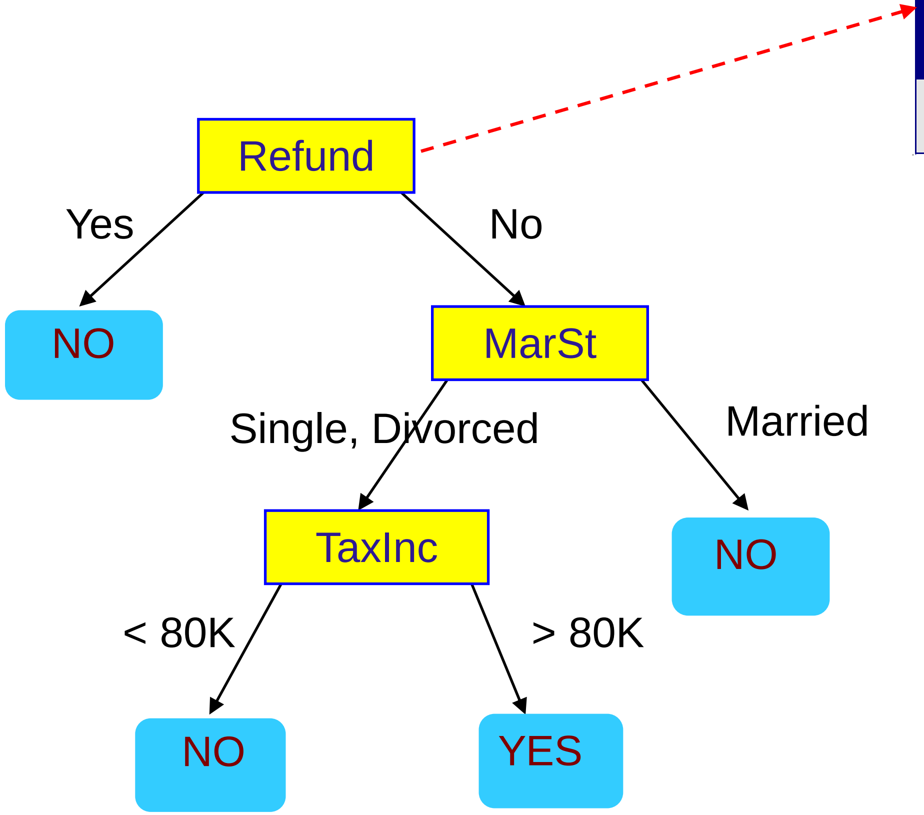
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Apply Model to Test Data

Test Data

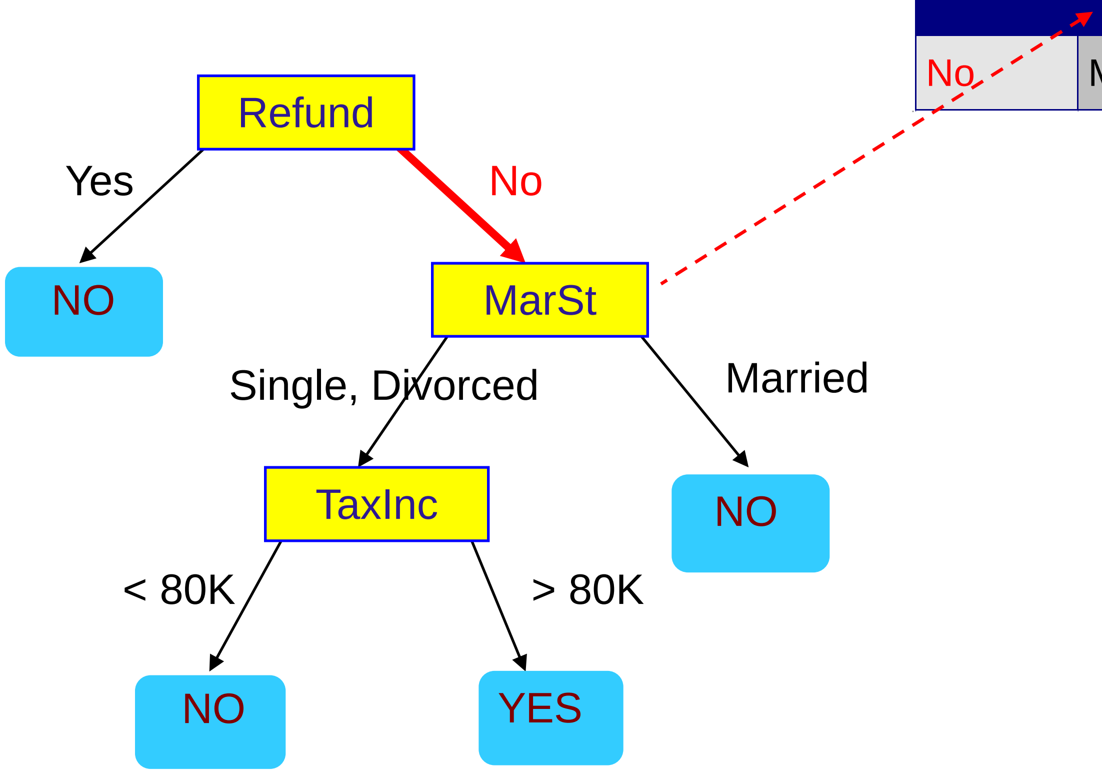
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

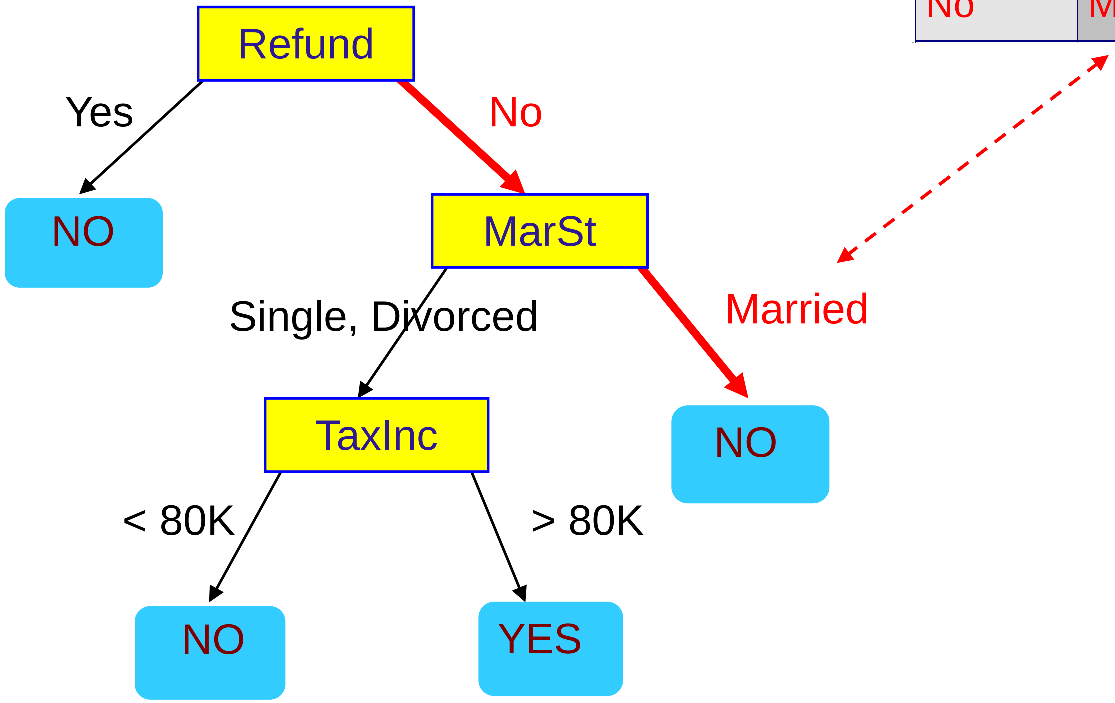
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

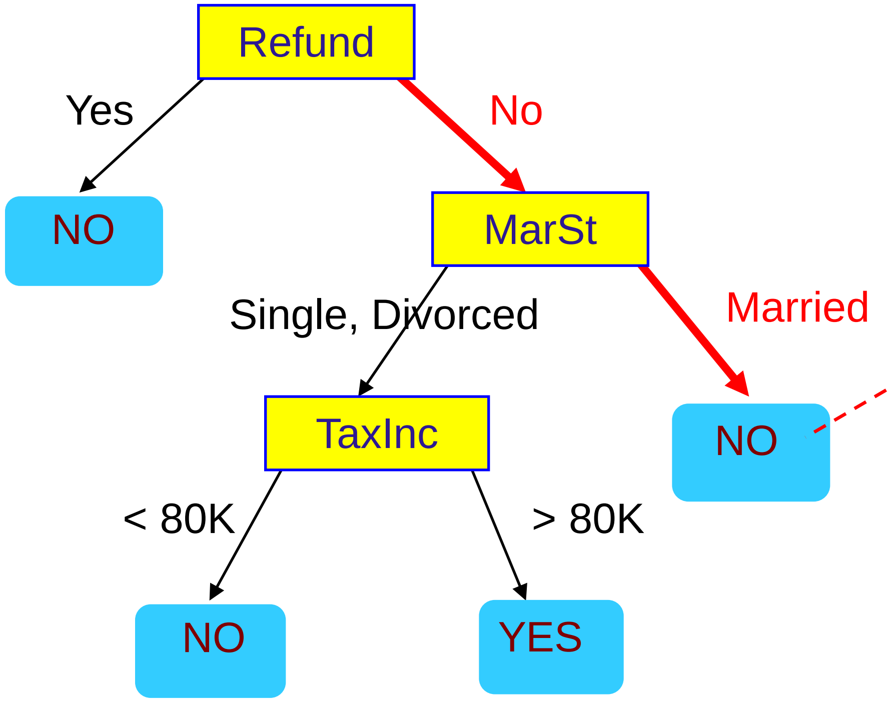
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to “No”

Decision Tree Induction Algorithm

Many Algorithms:

- *Hunt’s Algorithm (one of the earliest)*
- *CART*
- *ID3, C4.5*
- *C5*
- *CHAID*
- *SLIQ,SPRINT*

Classification: Application 1

Direct Marketing

- ***Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell-phone product.***
- ***Approach:***
  - ***Use the data for a similar product introduced before.***
  - ***We know which customers decided to buy and which decided otherwise. This {buy, don't buy} decision forms the class attribute.***
  - ***Collect various demographic, lifestyle, and company interaction related information about all such customers.***
  - ***Type of business, where they stay, how much they earn, etc.***
  - ***Use this information as input attributes to learn a classifier model.***

Classification: Application 2

Fraud Detection

- ***Goal: Predict fraudulent cases in credit card transactions.***
- ***Approach:***
  - ***Use credit card transactions and the information on its account-holder as attributes.***
    - ***When does a customer buy, what does he buy, how often he pays on time, etc.***
  - ***Label past transactions as fraud or fair transactions. This forms the class attribute.***
  - ***Learn a model for the class of the transactions.***
  - ***Use this model to detect fraud by observing credit card transactions on an account.***

Classification: Application 3

Customer Attrition/Churn:

- ***Goal: To predict whether a customer is likely to be lost to a competitor.***
- ***Approach:***
  - ***Use detailed record of transactions with each of the past and present customers, to find attributes.***
    - ***How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.***
  - ***Label the customers as loyal or disloyal.***
  - ***Find a model for loyalty.***

Any Questions?