# Data Mining

## Lecture 7

Ananya Jana
CS360

Fall 2024

# Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

# Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values

# Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

- Typically used for association analysis

- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

# Binarization

Conversion of a categorical attribute to three binary attributes

| Categorical Value | Integer Value | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| *awful* | 0 | 0 | 0 | 0 |
| *poor* | 1 | 0 | 0 | 1 |
| *OK* | 2 | 0 | 1 | 0 |
| *good* | 3 | 0 | 1 | 1 |
| *great* | 4 | 1 | 0 | 0 |

# Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: $x^k$, $\log(x)$, $e^x$, $|x|$
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

# Correlation

Correlation between variables: Correlation measure the linear relation between objects.

$$\mathrm{corr}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x}, \mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) * \mathrm{standard\_deviation}(\mathbf{y})}$$

X and Y are a set of n observations ($x_i$, $y_i$) where i = 1, 2,...n
Simply put, you can calculate correlation using three different sums of squares - sum of squares for variable X (denoted by $SS_{XX}$), sum of squares for variable Y (denoted by $SS_{YY}$) and the sum of the cross-products XY (denoted by $SS_{XY}$).

$$SS_{XX} = \sum (x_i - \bar{x})^2$$
$$SS_{YY} = \sum (y_i - \bar{y})^2$$
$$SS_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

Where $\bar{x}$ and $\bar{y}$ are the the sample means of X and Y.

# Correlation

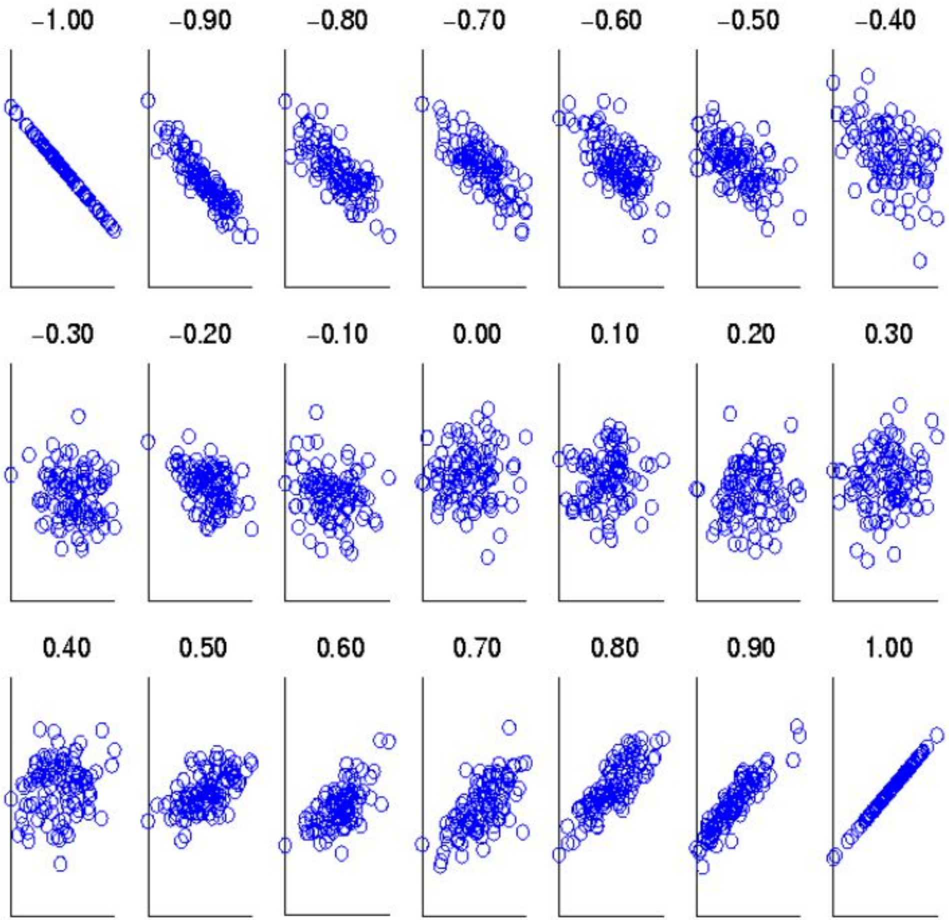
Then correlation is

$$r = \frac{SS_{XY}}{\sqrt{(SS_{XX})(SS_{YY})}}$$

The value of a correlation coefficient ranges between −1 and +1 .
The rough guidelines for correlation

| | |
|---|---|
| $0 < \|r\| < .3$ | weak correlation |
| $.3 < \|r\| < .7$ | moderate correlation |
| $\|r\| > 0.7$ | strong correlation |

# Visually Evaluating correlation




**Scatter plots showing the similarity from –1 to 1.**

# Drawbacks of correlation

- $\mathbf{x} = (-3, -2, -1, 0, 1, 2, 3)$
- $\mathbf{y} = (9, 4, 1, 0, 1, 4, 9)$

$$y_i = x_i^2$$

- $\text{mean}(\mathbf{x}) = 0$, $\text{mean}(\mathbf{y}) = 4$
- $\text{std}(\mathbf{x}) = 2.16$, $\text{std}(\mathbf{y}) = 3.74$

- $\text{corr} = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) \quad / ( 6 * 2.16 * 3.74 )$

  $= 0$

# Basic Classification

Classification: Given a collection of records (training set) each record is characterized by a tuple (x, y) where x is the attribute set and y is the class label

x :attribute, predictor, independent variable, input
y : class, response, dependent variable, output.

Task: Learn a model that maps each attribute set x into one of the predefined class labels y

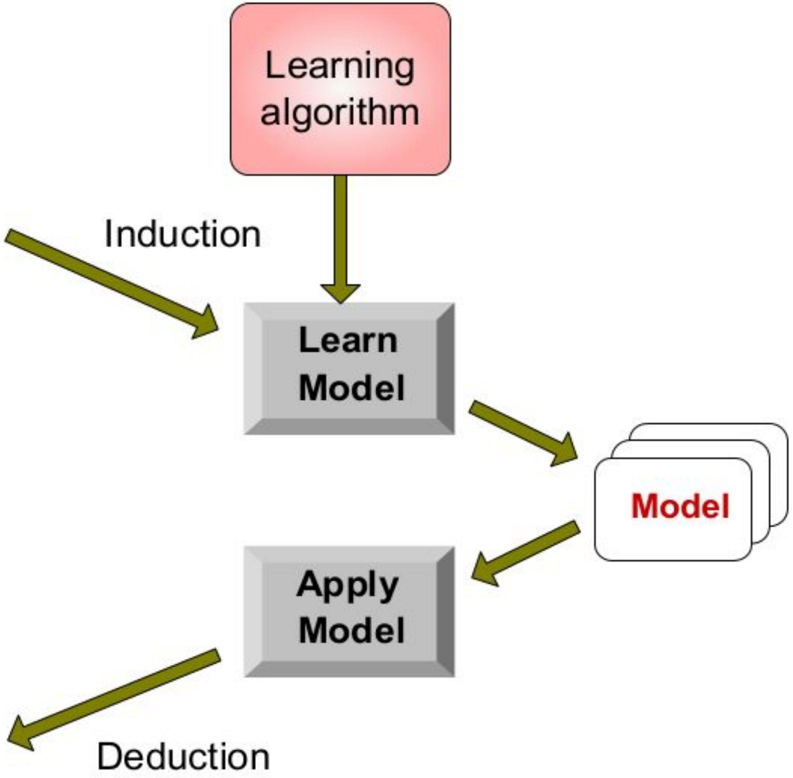
# Examples of Classification Task

| Task | Attribute set, $x$ | Class label, $y$ |
|---|---|---|
| Categorizing email messages | Features extracted from email message header and content | spam or non-spam |
| Identifying tumor cells | Features extracted from MRI scans | malignant or benign cells |
| Cataloging galaxies | Features extracted from telescope images | Elliptical, spiral, or irregular-shaped galaxies |

# General Approach for building classification model

| *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

# Classification Techniques

Base Classifiers
– Decision Tree based Methods
– Rule-based Methods
– Nearest-neighbor
– Neural Networks
– Deep Learning
– Naïve Bayes and Bayesian Belief Networks
– Support Vector Machines

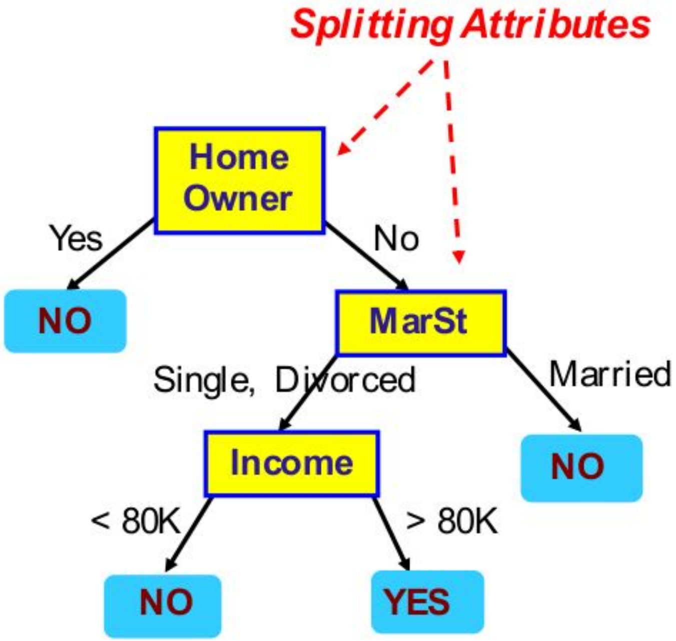
Ensemble Classifiers
– Boosting, Bagging, Random Forests

# Example of a Decision Tree

categorical categorical continuous class

| ID | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 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**

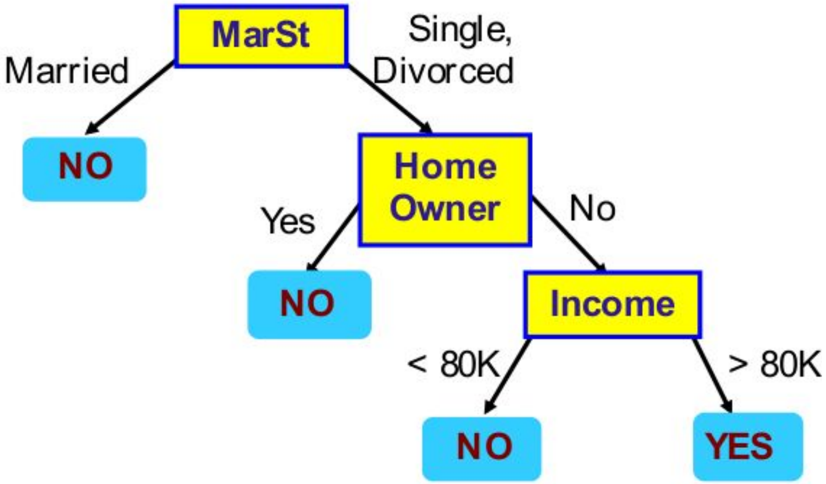



**Model: Decision Tree**

# Another Example of a Decision Tree

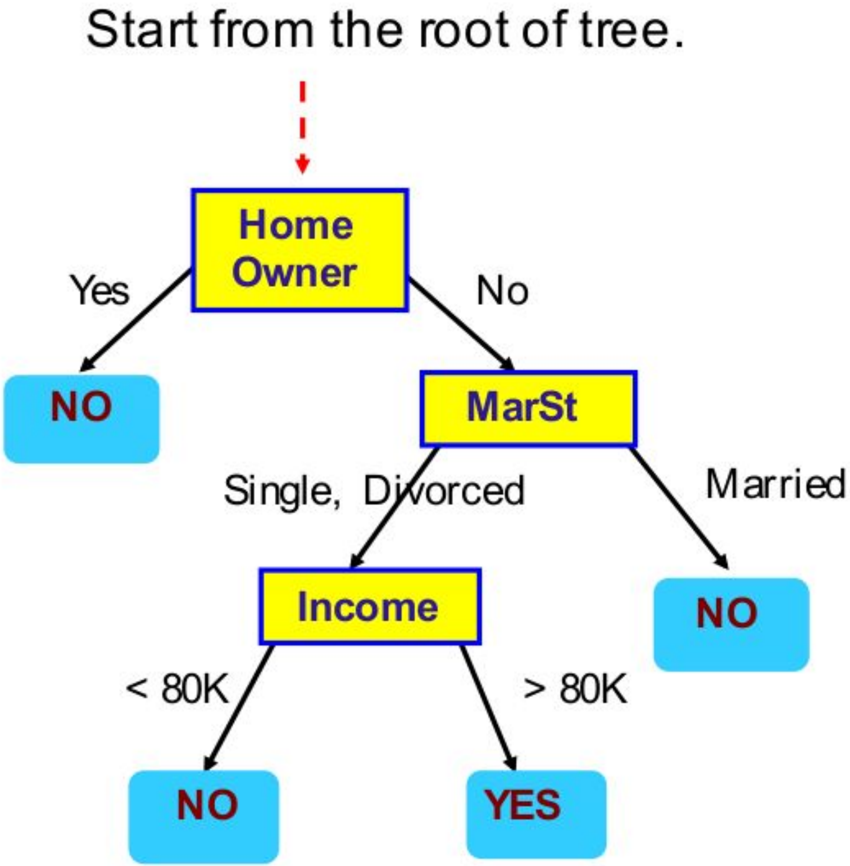
categorical categorical continuous class

| ID | Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|---|
| 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 |




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

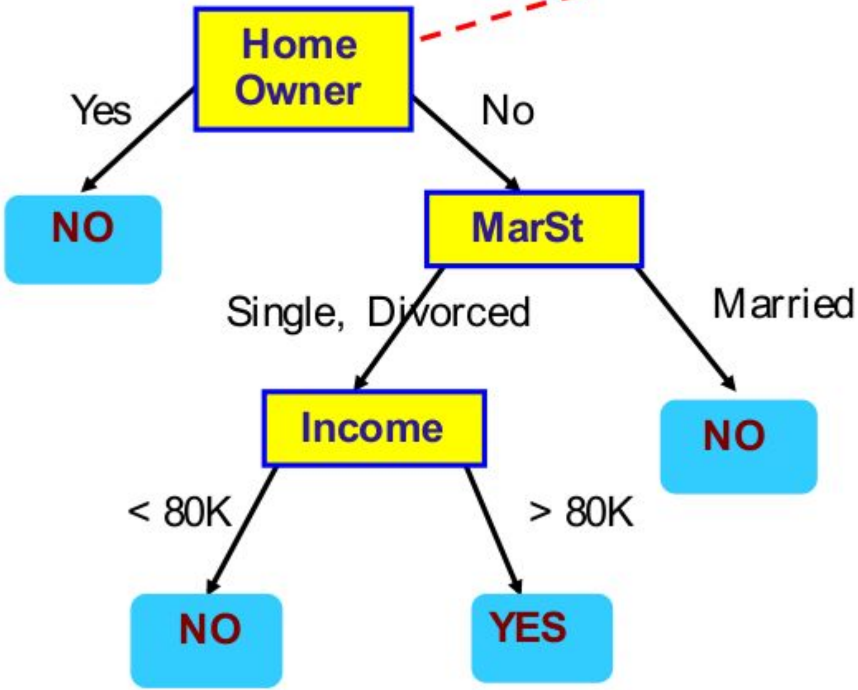
# Apply Model to test data




**Test Data**

| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| No | Married | 80K | ? |

# Apply Model to test data



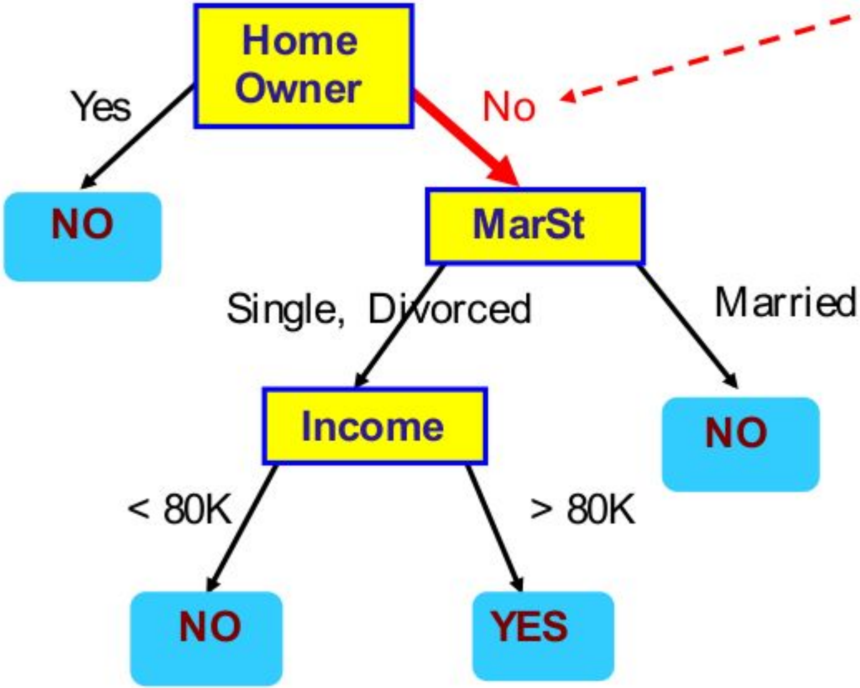

**Test Data**

| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| No | Married | 80K | ? |

# Apply Model to test data

**Test Data**

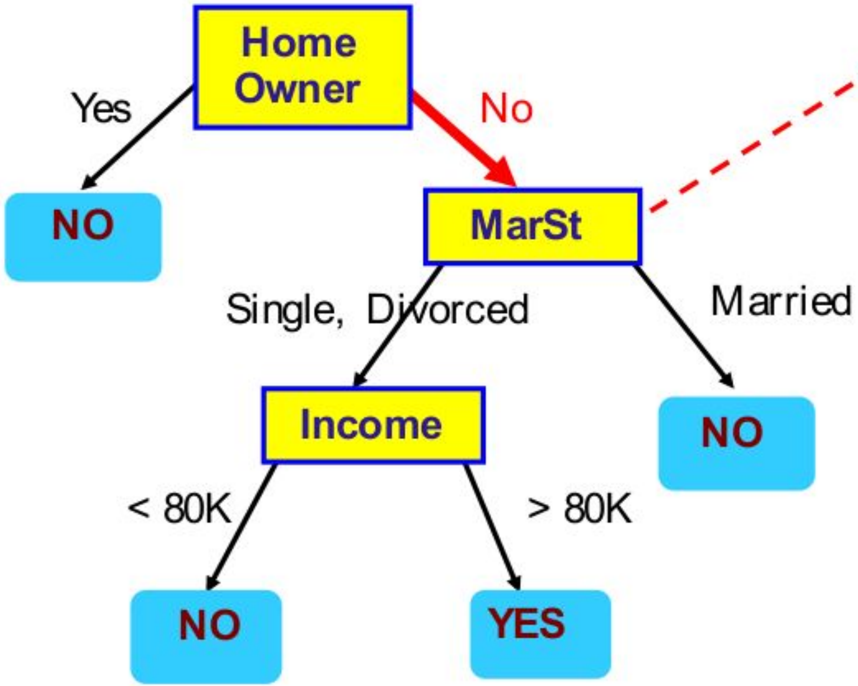
| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| No | Married | 80K | ? |



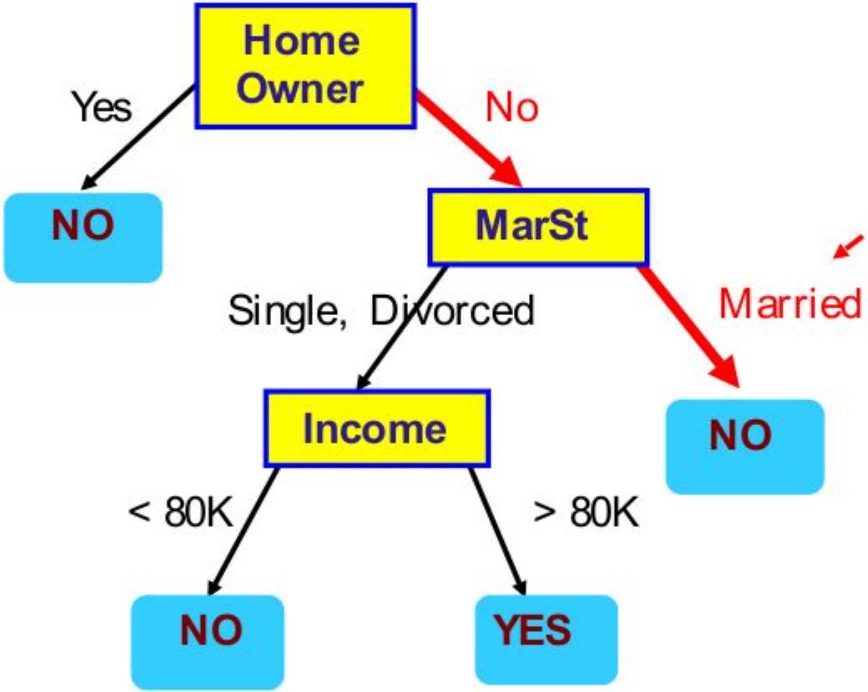
# Apply Model to test data

## Test Data

| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| No | Married | 80K | ? |



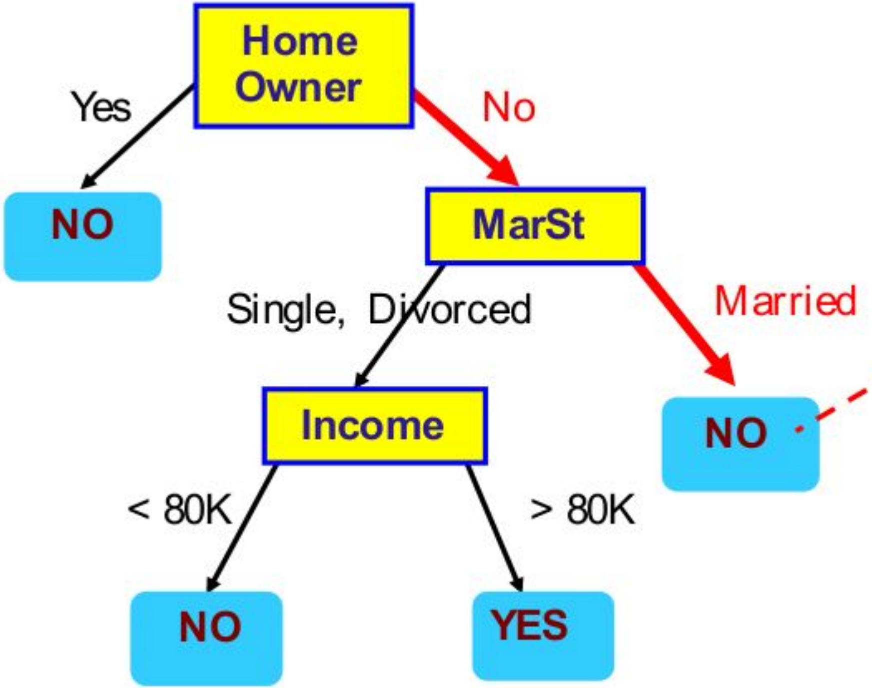
# Apply Model to test data




**Test Data**

| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| No | Married | 80K | ? |

# Apply Model to test data




**Test Data**

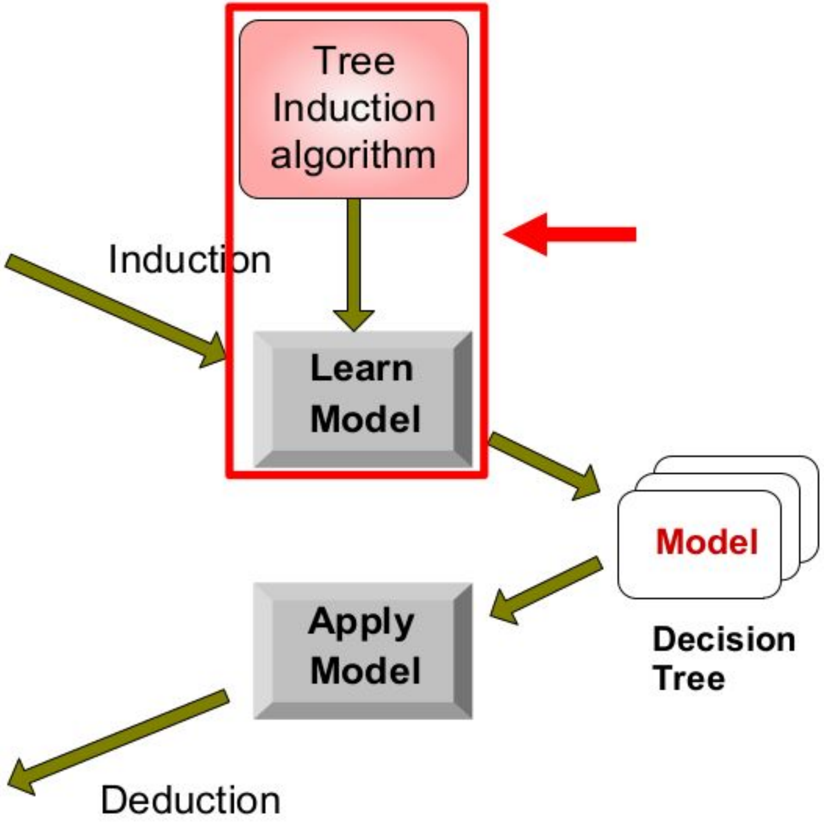
| Home Owner | Marital Status | Annual Income | Defaulted Borrower |
|---|---|---|---|
| No | Married | 80K | ? |

Assign Defaulted to "No"

# 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 | ? |

# Decision Tree Induction

- Many Algorithms:
  - Hunt’s Algorithm (one of the earliest)
  - CART
  - ID3, C4.5
  - SLIQ,SPRINT

# Task

| Area | % of people below poverty line | % of bike riders wearing helmet |
|---|---|---|
| Fair Oaks | 50 | 22.1 |
| Strandwood | 11 | 35.9 |
| Walnut Acres | 2 | 57.9 |
| Discov. Bay | 19 | 22.2 |

Calculate the correlation between the two columns.

# Curse of Dimensionality

Python code for showing curse of dimensionality (ipynb file on Blackboard)

```
import numpy as np
import matplotlib.pyplot as plt
import os
import math
```

```
values = []
for N in range(2,50):
    # Generate 1000 random points in N dimensions.
    P = [np.random.randint(-100, 100, N) for _ in range(1000)]
    # Generate 1 random point P2 in N dimensions.
    P2 = np.random.randint(-100,100,N)
    # calculate the difference between the set of points P and the random point P2
    diffs = [np.linalg.norm(p-P2) for p in P]
    max_d = max(diffs)
    min_d = min(diffs)
    value = math.log10(max_d-min_d)/min_d
    values.append( value )
```

```
plt.plot(range(2,50),values)
plt.xlabel('Number of dimensions')
plt.ylabel('Values')
plt.show()
```