

The background of the slide is a light gray gradient. It is decorated with several realistic water droplets of various sizes, some clustered in the top left and bottom right corners. A faint, concentric circular pattern is visible in the upper center of the image.

# UNIT-5

CLASSIFICATION AND PREDICTION

# INTRODUCTION

There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows –

- Classification
- Prediction

*Classification models predict categorical class labels; and prediction models predict continuous valued functions.*

For example, *we can build a classification model to categorize bank loan applications as either safe or risky, or*

*a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.*

# CLASSIFICATION

## **Classification:**


- [Classification](#) is the process of finding a good model that describes the data classes or concepts, and the purpose of classification is to predict the class of objects whose class label is unknown.
- In simple terms, we can think of classification as categorizing the incoming new data based on our current or past assumptions that we have made and the data that we already have with us.

# WHAT IS PREDICTION?

Following are the examples of cases where the data analysis task is prediction –

- *Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value.* Therefore the data analysis task is an example of numeric prediction.
- In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

***Note – Regression Analysis is a statistical methodology that is most often used for numeric prediction.***

- 
- **SUPERVISED LEARNING**
    - CLASSIFICATION AND REGRESSION
  - **UNSUPERVISED LEARNING**
    - CLUSTERING

# CLASSIFICATION

Following are the examples of cases where the data analysis task is classification –

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

# PREDICTION: CLASSIFICATION VS. REGRESSION

## ■ Classification

- Predicts categorical class labels (discrete or nominal)
- Classifies data (constructs a model) based on the training set and the values (**class labels**) in a classifying attribute and uses it in classifying new data

## ■ Regression

- Models continuous-valued functions, i.e., Predicts unknown or missing values

## ■ Typical applications

- Credit approval
- Target marketing
- Medical diagnosis
- Fraud detection

# CLASSIFICATION—A TWO-STEP PROCESS

- **Model construction:** describing a set of predetermined classes
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the **class label attribute**
  - The set of tuples used for model construction is **training set**
  - The model is represented as classification rules, decision trees, or mathematical formulae
- **Model usage:** for classifying future or unknown objects
  - **Estimate accuracy** of the model
    - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set, otherwise over-fitting will occur
  - If the accuracy is acceptable, use the model to **classify data** tuples whose class labels are not known



Sr.No.

Prediction

Classification

1.

Prediction is about predicting a missing/unknown element(continuous value) of a dataset

Classification is about determining a (categorical) class (or label) for an element in a dataset

2.

Eg. We can think of prediction as predicting the correct treatment for a particular disease for an individual person.

Eg. Whereas the grouping of patients based on their medical records can be considered classification.

3.

The model used to predict the unknown value is called a predictor.

The model used to classify the unknown value is called a classifier.

4.

The predictor is constructed from a training set and its accuracy refers to how well it can estimate the value of new data.

A classifier is also constructed from a training set composed of the records of databases and their corresponding class names

# CLASSIFICATION EXAMPLE

- Example training database
  - Two attributes:  
age and car-type (**s**port, **m**inivan  
and **t**ruck)
  - Age is ordered, car-type is  
categorical attribute
  - Class label indicates  
whether person bought  
product
  - Dependent attribute is *categorical*

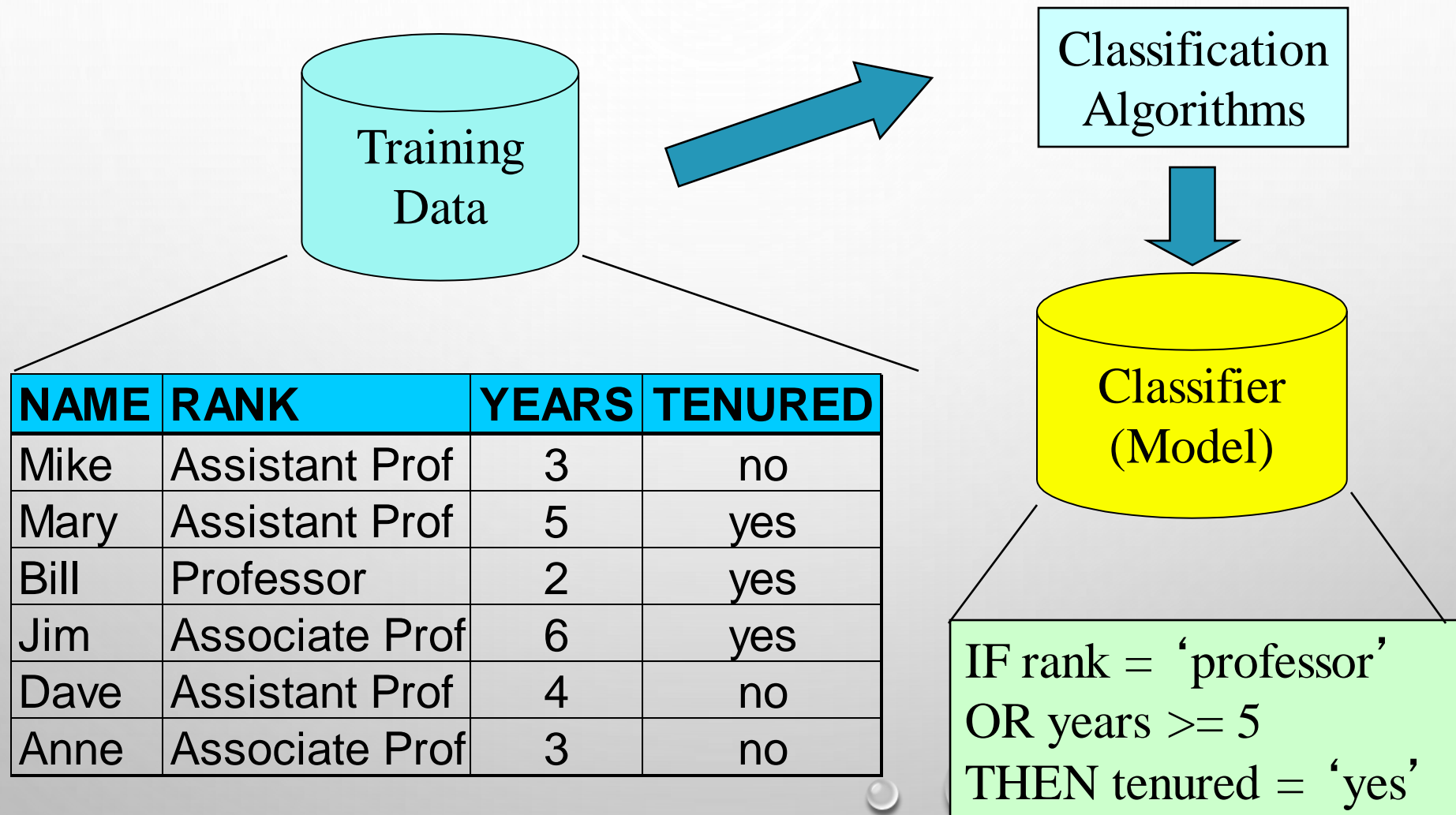
| Age | Car | Class |
|-----|-----|-------|
| 20  | M   | Yes   |
| 30  | M   | Yes   |
| 25  | T   | No    |
| 30  | S   | Yes   |
| 40  | S   | Yes   |
| 20  | T   | No    |
| 30  | M   | Yes   |
| 25  | M   | Yes   |
| 40  | M   | Yes   |
| 20  | S   | No    |

# REGRESSION EXAMPLE

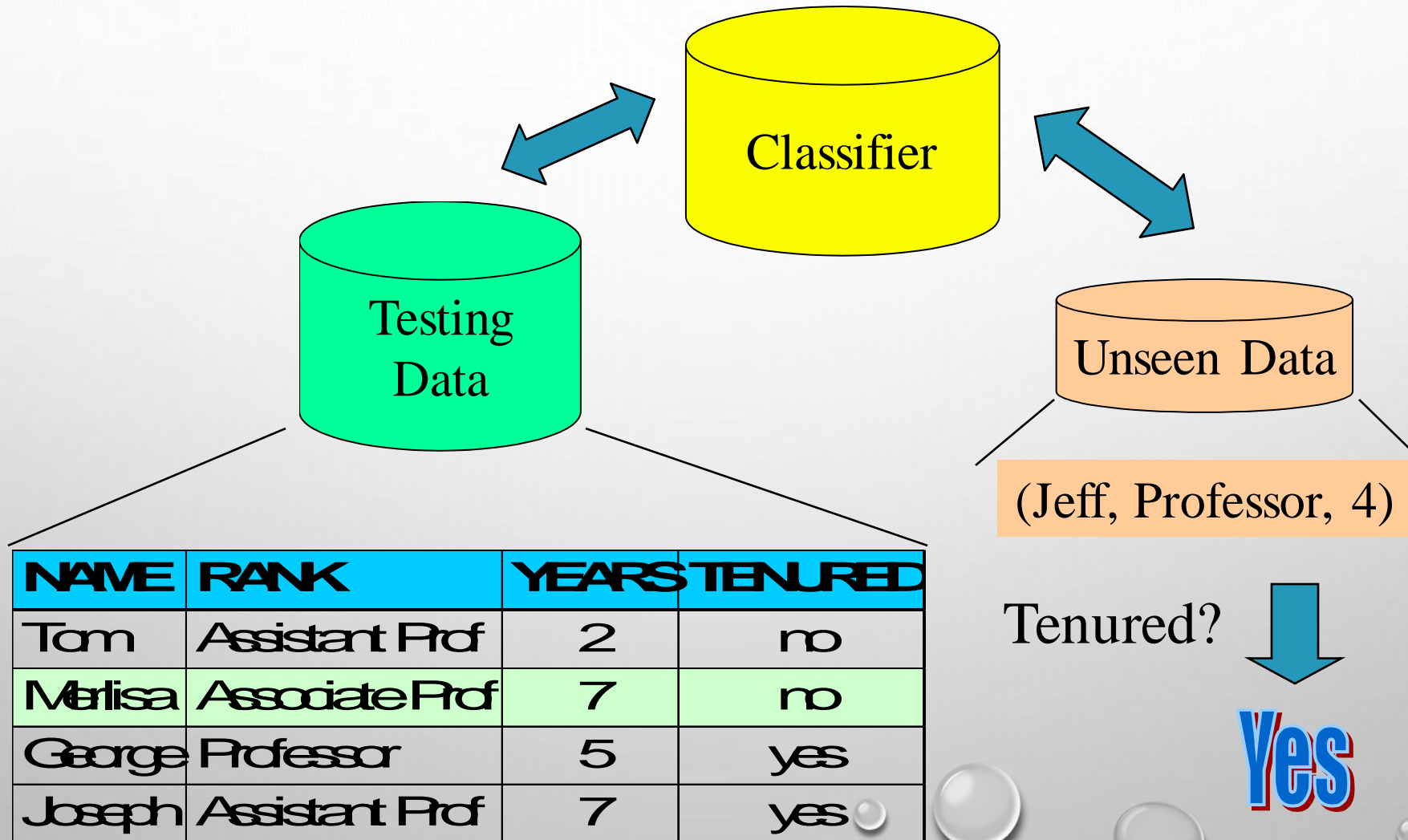
- Example training database
  - Two predictor attributes:  
age and car-type (**s**port, **m**inivan  
and **t**ruck)
  - Spent indicates how much person  
spent during a recent visit to the  
web site
  - Dependent attribute is *numerical*

| Age | Car | Spent |
|-----|-----|-------|
| 20  | M   | \$200 |
| 30  | M   | \$150 |
| 25  | T   | \$300 |
| 30  | S   | \$220 |
| 40  | S   | \$400 |
| 20  | T   | \$80  |
| 30  | M   | \$100 |
| 25  | M   | \$125 |
| 40  | M   | \$500 |
| 20  | S   | \$420 |

# PROCESS (1): MODEL CONSTRUCTION



## PROCESS (2): USING THE MODEL IN PREDICTION



# SUPERVISED VS. UNSUPERVISED LEARNING

## ■ Supervised learning (classification)

- Supervision: the training data (observations, measurements, etc.) Are accompanied by labels indicating the class of the observations
- New data is classified based on the model built on training set

## ■ Unsupervised learning (clustering)

- The class labels of training data is unknown
- Given a set of measurements, observations, etc. With the aim of establishing the existence of classes or clusters in the data

# ISSUES: DATA PREPARATION

- Data cleaning

- Preprocess data in order to reduce noise and handle missing values

- Relevance analysis (feature selection)

- Remove the irrelevant or redundant attributes

- Data transformation

- Generalize and/or normalize data



# CLASSIFICATION AND PREDICTION ISSUES

*The major issue is preparing the data for classification and prediction. Preparing the data involves the following activities –*

- **Data cleaning** – data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute. Although various classification algorithms have some structures for managing noisy or missing information, this step can support reducing confusion during learning.
- **Relevance analysis** – database may also have the irrelevant attributes. Correlation analysis is used to know Whether any two given attributes are related. Hence, relevance analysis, in the form of correlation analysis and attribute subset selection, can be used to detect attributes that do not contribute to the classification or prediction task.

*Such analysis can help improve classification efficiency and scalability.*



# CONTINUED...

- **Data transformation and reduction** – the data can be transformed by any of the following methods.
  - **Normalization** – the data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.
  - **Generalization** – the data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

*Note – data can also be reduced by some other methods such as wavelet transformation, binning, histogram analysis, and clustering.*

# ISSUES: EVALUATING CLASSIFICATION METHODS

## ■ Accuracy

- Classifier accuracy: predicting class label

- Regression accuracy: guessing value of predicted attributes

## ■ Speed

- Time to construct the model (training time)

- Time to use the model (classification/prediction time)

- Robustness: handling noise and missing values

- Scalability: efficiency on large databases

## ■ Interpretability

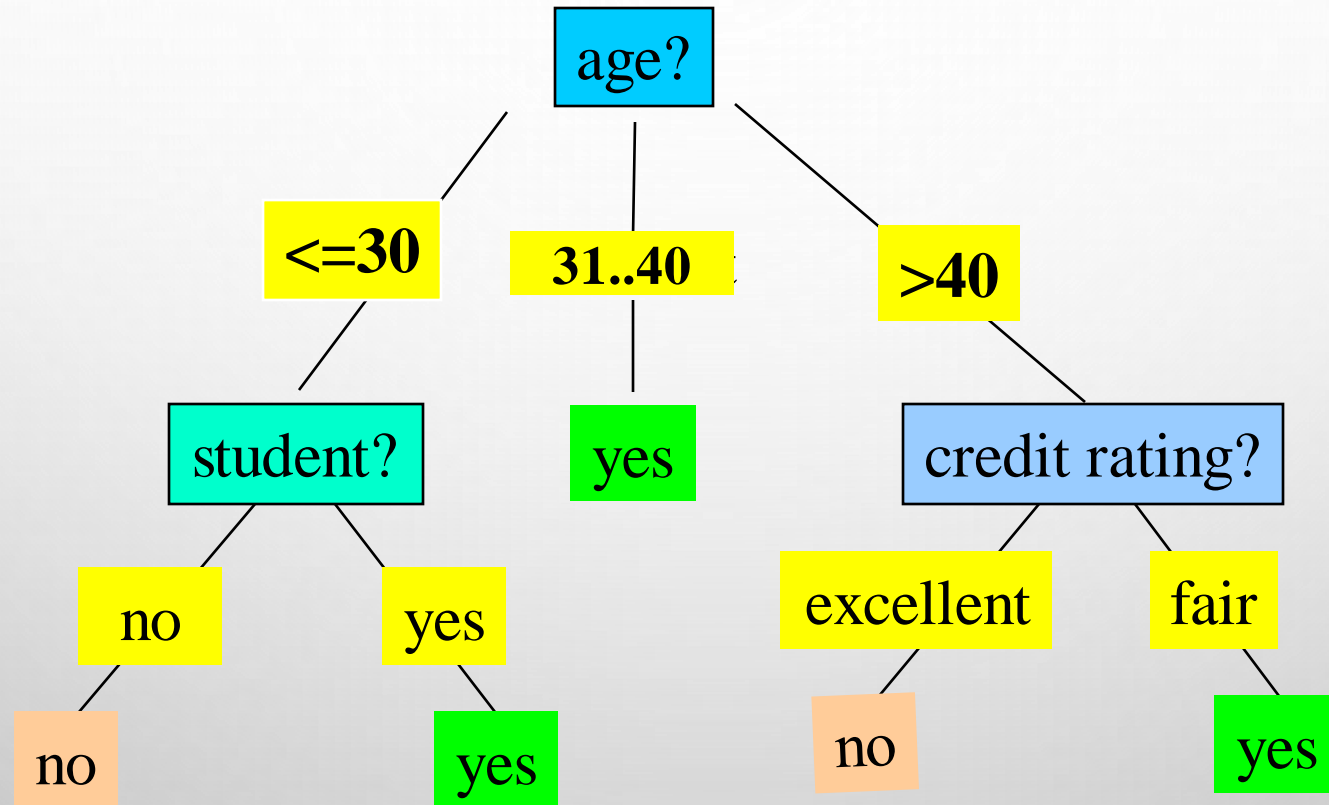
- Understanding and insight provided by the model

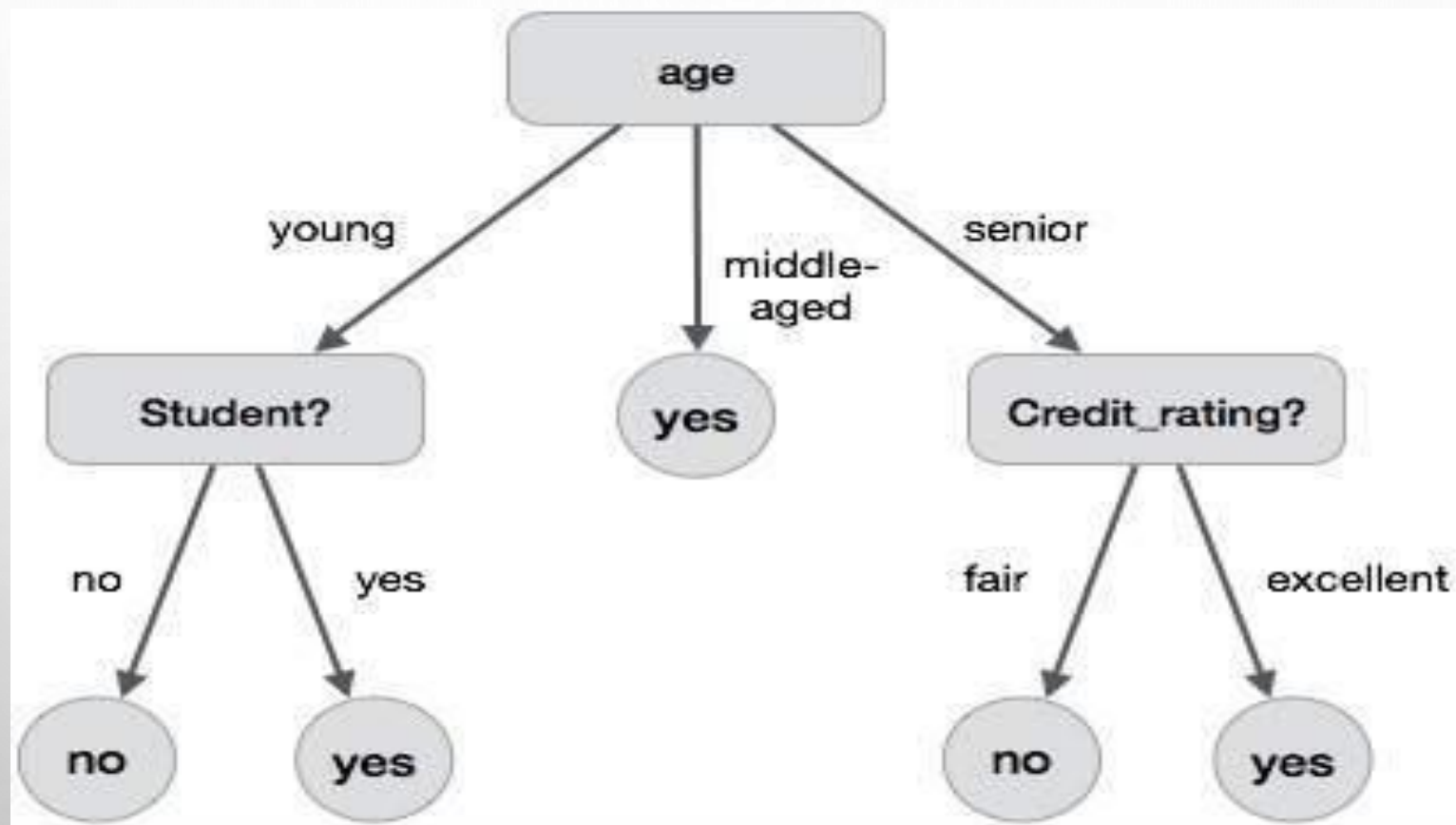
- Other measures, e.G., Goodness of rules, such as decision tree size or compactness of classification rules

# DECISION TREE INDUCTION: TRAINING DATASET

| age     | income | student | credit_rating | buys_computer |
|---------|--------|---------|---------------|---------------|
| <=30    | high   | no      | fair          | no            |
| <=30    | high   | no      | excellent     | no            |
| 31...40 | high   | no      | fair          | yes           |
| >40     | medium | no      | fair          | yes           |
| >40     | low    | yes     | fair          | yes           |
| >40     | low    | yes     | excellent     | no            |
| 31...40 | low    | yes     | excellent     | yes           |
| <=30    | medium | no      | fair          | no            |
| <=30    | low    | yes     | fair          | yes           |
| >40     | medium | yes     | fair          | yes           |
| <=30    | medium | yes     | excellent     | yes           |
| 31...40 | medium | no      | excellent     | yes           |
| 31...40 | high   | yes     | fair          | yes           |
| >40     | medium | no      | excellent     | no            |

## OUTPUT: A DECISION TREE FOR “*BUYS\_COMPUTER*”





# ALGORITHM FOR DECISION TREE INDUCTION

- Basic algorithm (a greedy algorithm)
  - Tree is constructed in a **top-down recursive divide-and-conquer manner**
  - At start, all the training examples are at the root
  - Attributes are categorical (if continuous-valued, they are discretized in advance)
  - Examples are partitioned recursively based on selected attributes
  - Test attributes are selected on the basis of a heuristic or statistical measure (e.G., **Information gain**)
- Conditions for stopping partitioning
  - All (or most) samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning – **majority voting** is employed for classifying the leaf