Lecture 14

Data mining and knowledge discovery

- Introduction, or or what is data mining?
- Data warehouse and query tools
- Decision trees
- Case study: Profiling people with high blood pressure
- Summary

© Negnevitsky, Pearson Education, 2002

What is data mining?

- **Data** is what we collect and store, and **knowledge** is what helps us to make informed decisions.
- The extraction of knowledge from data is called **data mining**.
- Data mining can also be defined as the exploration and analysis of *large* quantities of data in order to discover meaningful patterns and rules.
- The ultimate goal of data mining is to discover knowledge.

© Negnevitsky, Pearson Education, 2002

Data warehouse

- Modern organisations must respond quickly to any change in the market. This requires rapid access to current data normally stored in operational databases.
- However, an organisation must also determine which trends are relevant. This task is accomplished with access to historical data that are stored in large databases called **data warehouses**.

© Negnevitsky, Pearson Education, 2002

- The main characteristic of a data warehouse is its capacity. A data warehouse is really big – it includes millions, even billions, of data records.
- The data stored in a data warehouse is
  - **time dependent** – linked together by the times of recording – and
  - **integrated** – all relevant information from the operational databases is combined and structured in the warehouse.

© Negnevitsky, Pearson Education, 2002

Query tools

- A data warehouse is designed to support decision-making in the organisation. The information needed can be obtained with **query tools**.
- Query tools are **assumption-based** – a user must ask the *right* questions.

© Negnevitsky, Pearson Education, 2002

How is data mining applied in practice?

- Many companies use data mining today, but refuse to talk about it.
- In direct marketing, data mining is used for targeting people who are most likely to buy certain products and services.
- In trend analysis, it is used to determine trends in the marketplace, for example, to model the stock market. In fraud detection, data mining is used to identify insurance claims, cellular phone calls and credit card purchases that are most likely to be fraudulent.

© Negnevitsky, Pearson Education, 2002

Data mining tools

Data mining is based on intelligent technologies already discussed in this book. It often applies such tools as neural networks and neuro-fuzzy systems.

However, the most popular tool used for data mining is a **decision tree**.

© Negnevitsky, Pearson Education, 2002

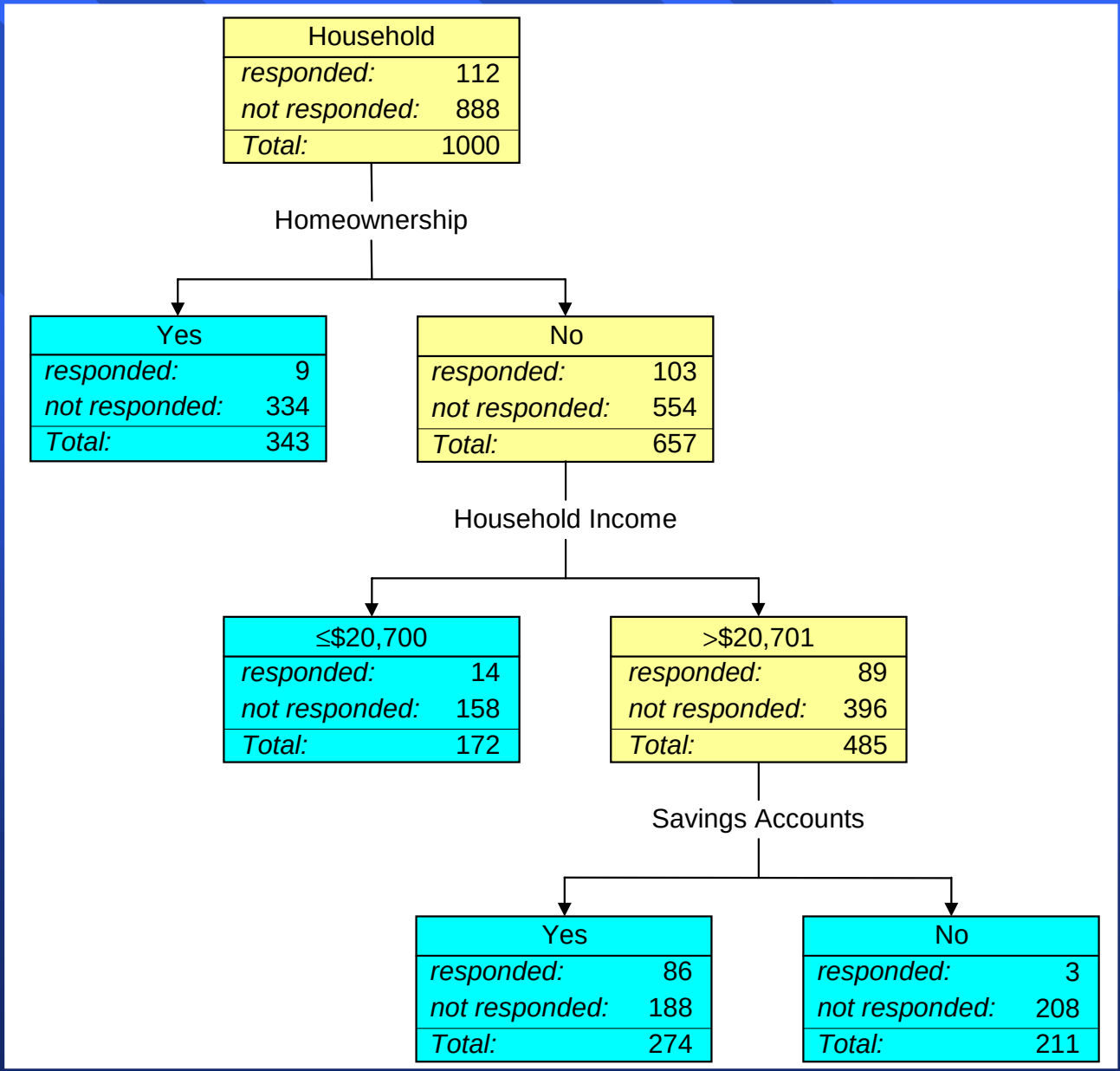
Decision trees

A decision tree can be defined as a map of the reasoning process. It describes a data set by a tree-like structure.

Decision trees are particularly good at solving classification problems.

© Negnevitsky, Pearson Education, 2002

An example of a decision tree



© Negnevitsky, Pearson Education, 2002

- A decision tree consists of **nodes**, **branches** and **leaves**.
- The top node is called the **root node**. The tree always starts from the root node and grows down by splitting the data at each level into new nodes. The root node contains the entire data set (all data records), and child nodes hold respective subsets of that set.
- All nodes are connected by **branches**.
- Nodes that are at the end of branches are called **terminal nodes**, or **leaves**.

© Negnevitsky, Pearson Education, 2002

How does a decision tree select splits?

- A split in a decision tree corresponds to the predictor with the maximum separating power. The best split does the best job in creating nodes where a single class dominates.
- One of the best known methods of calculating the predictor's power to separate data is based on the **Gini coefficient of inequality**.

© Negnevitsky, Pearson Education, 2002

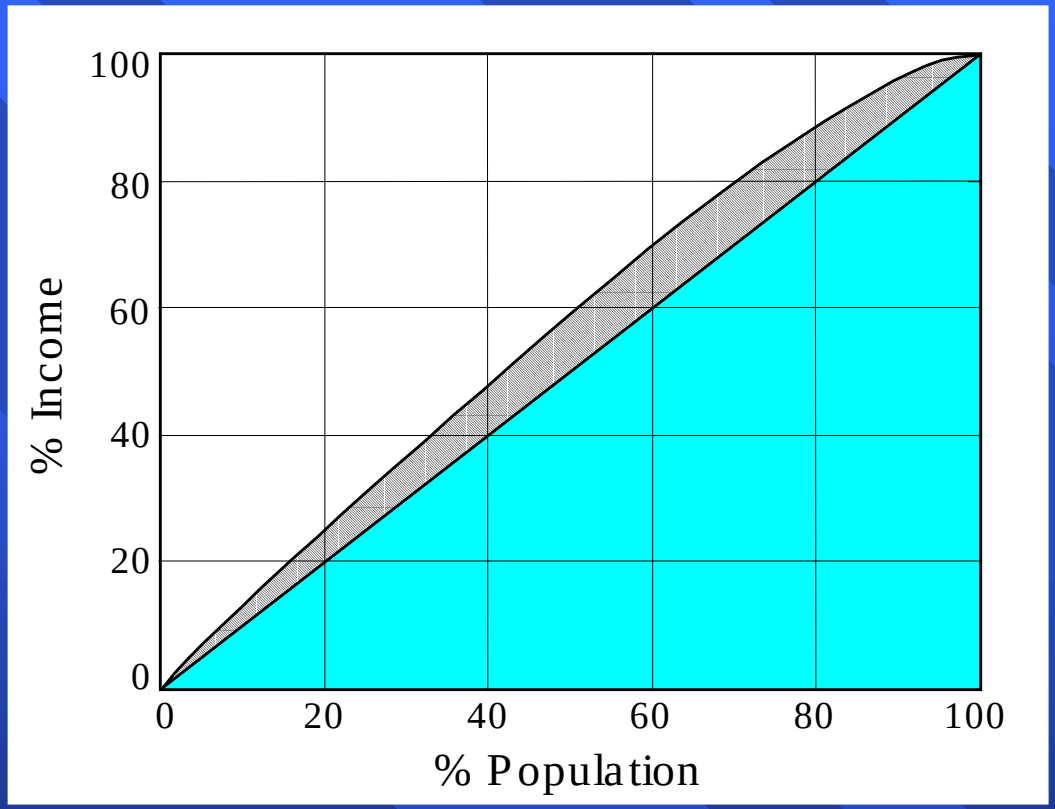
The Gini coefficient

The Gini coefficient is a measure of how well the predictor separates the classes contained in the parent node.

Gini, an Italian economist, introduced a rough measure of the amount of inequality in the income distribution in a country.

© Negnevitsky, Pearson Education, 2002

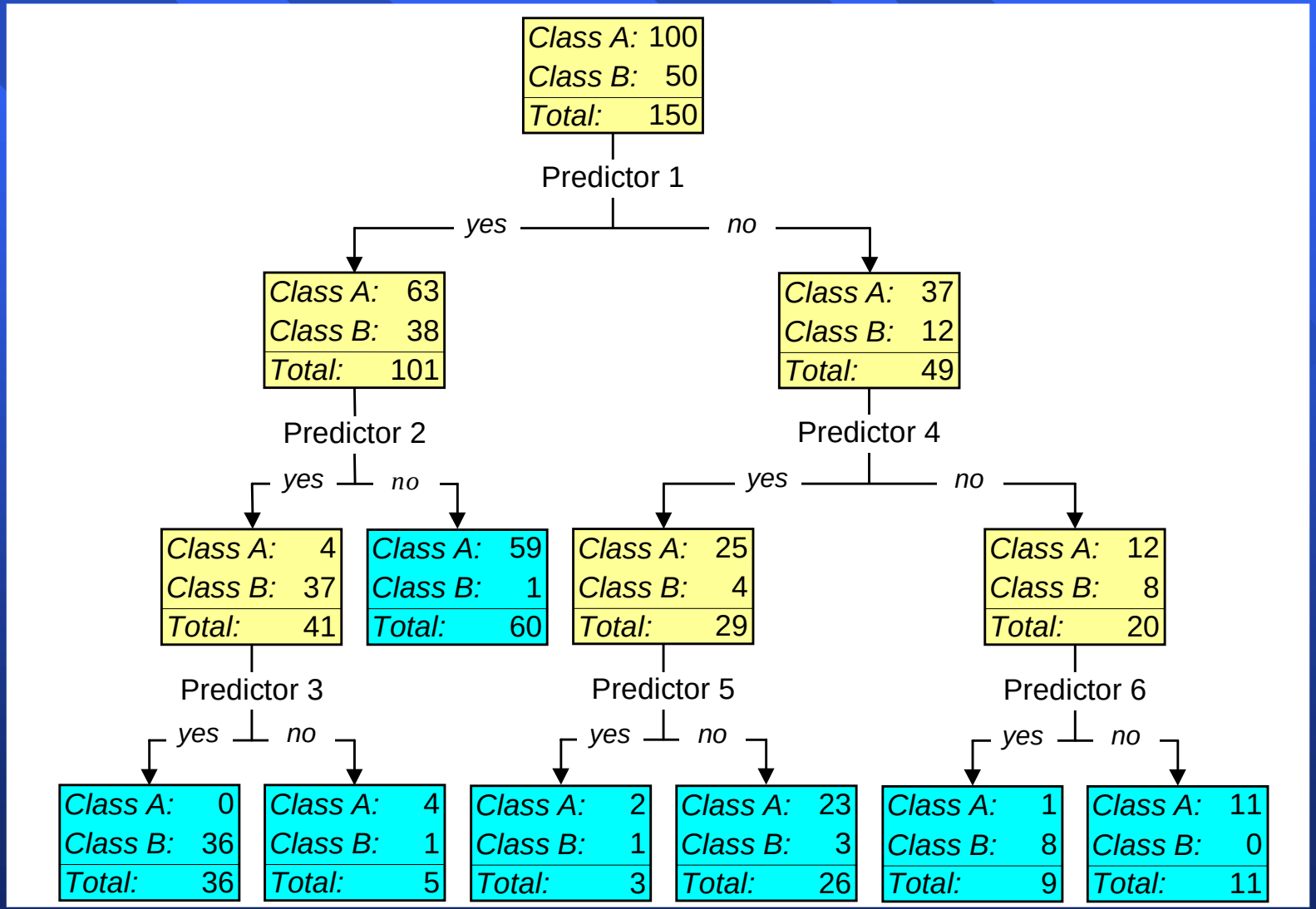
Computation of the Gini coefficient



The Gini coefficient is calculated as the area between the curve and the diagonal divided by the area below the diagonal. For a perfectly equal wealth distribution, the Gini coefficient is equal to zero.

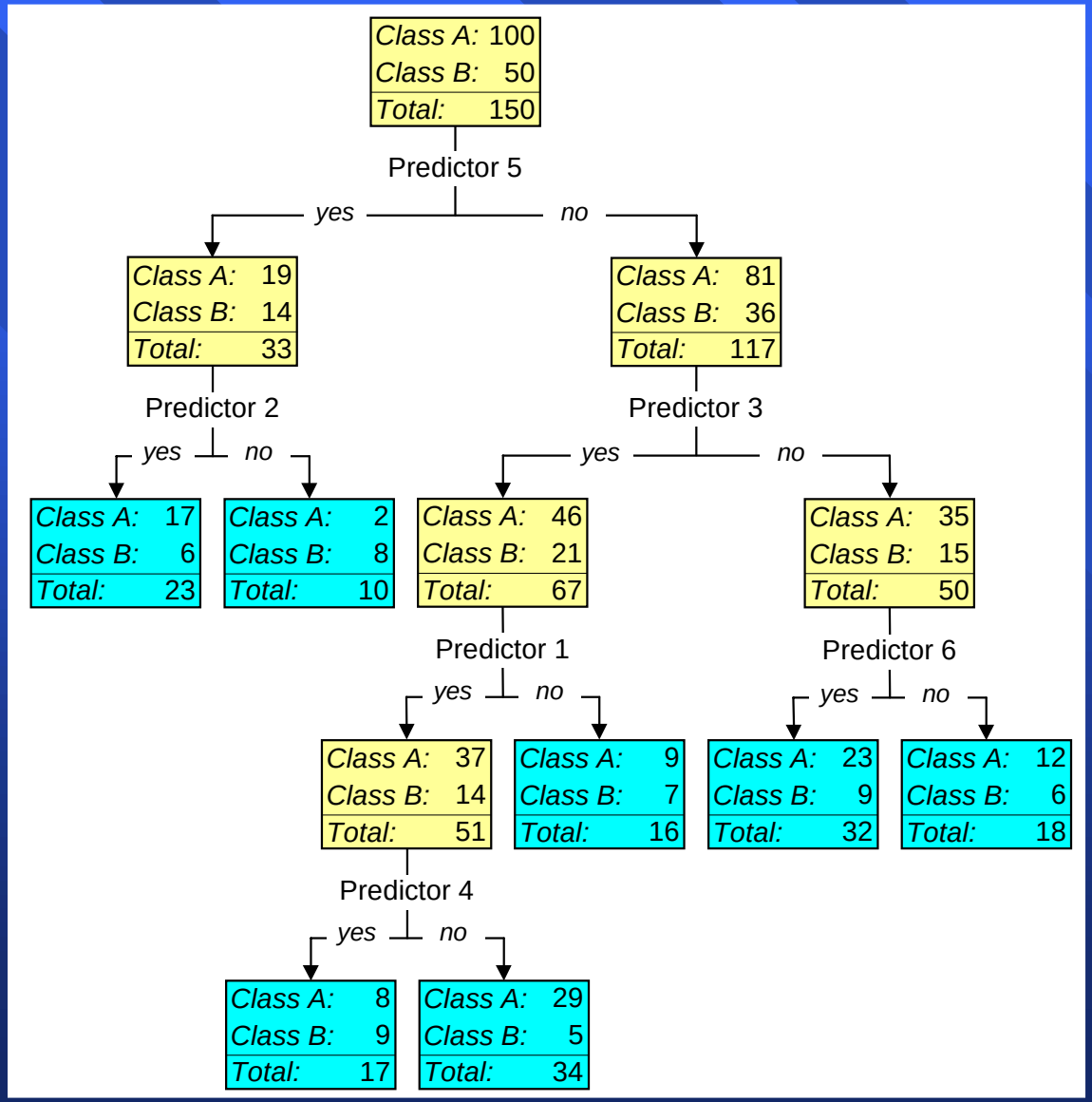
© Negnevitsky, Pearson Education, 2002

Selecting an optimal decision tree:
(*a*) Splits selected by Gini



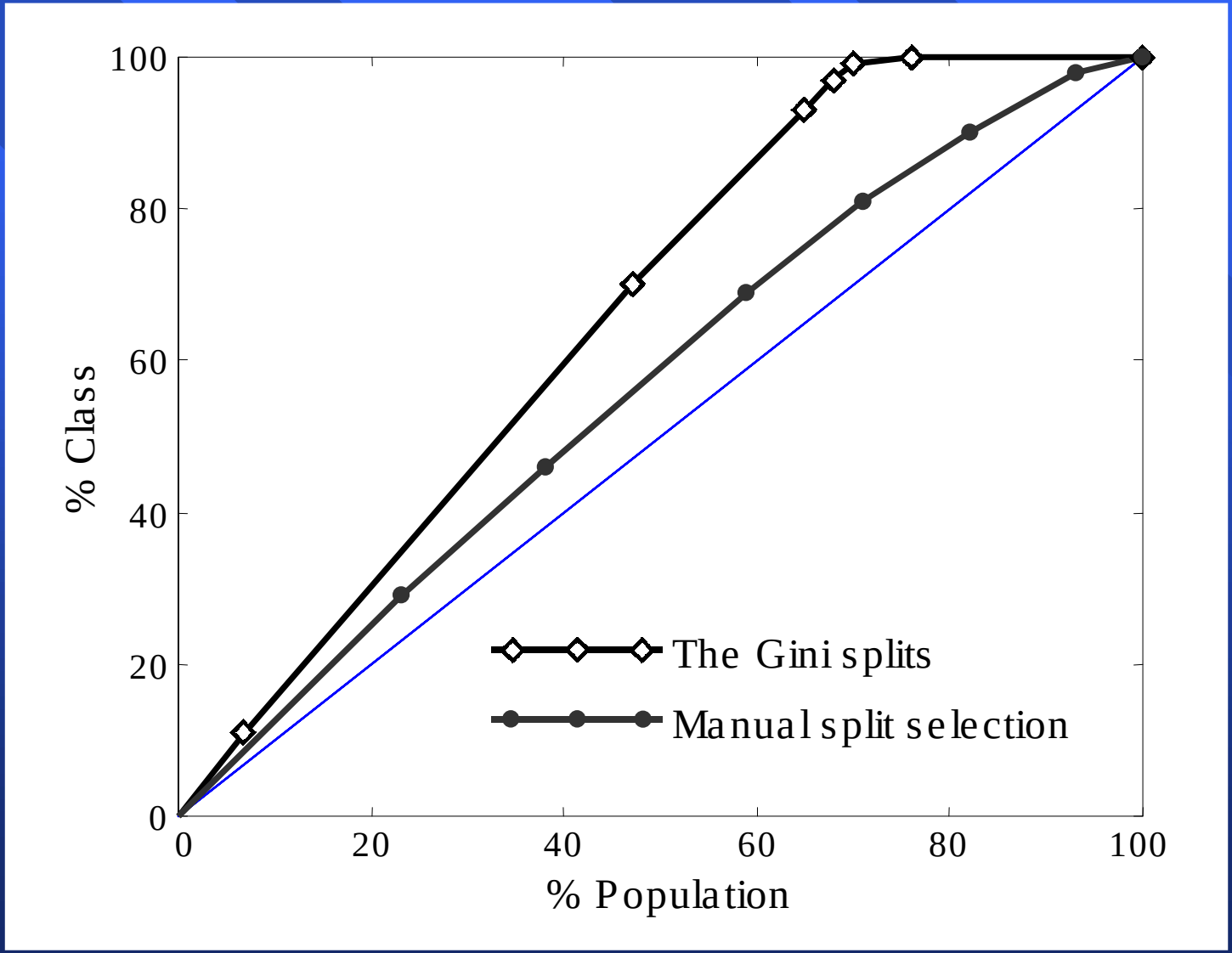
© Negnevitsky, Pearson Education, 2002

Selecting an optimal decision tree:
(*b*) Splits selected by duesswork



© Negnevitsky, Pearson Education, 2002

Gain chart of *Class A*



© Negnevitsky, Pearson Education, 2002

Can we extract rules from a decision tree?

The pass from the root node to the bottom leaf reveals a decision rule.

For example, a rule associated with the right bottom leaf in the figure that represents Gini splits can be represented as follows:

if (Predictor 1 = *no*)
and (Predictor 4 = *no*)
and (Predictor 6 = *no*)
then class = *Class A*

© Negnevitsky, Pearson Education, 2002

Case study:
Profiling people with high blood pressure

A typical task for decision trees is to determine conditions that may lead to certain outcomes.

Blood pressure can be categorised as optimal, normal or high. Optimal pressure is below 120/80, normal is between 120/80 and 130/85, and a hypertension is diagnosed when blood pressure is over 140/90.

© Negnevitsky, Pearson Education, 2002

A data set for a hypertension study

Community Health Survey: Hypertension Study (California, U.S.A.)	
Gender	☑ Male ☐ Female
Age	☐ 18 – 34 years ☐ 35 – 50 years ☑ 51 – 64 years ☐ 65 or more years
Race	☑ Caucasian ☐ African American ☐ Hispanic ☐ Asian or Pacific Islander
Marital Status	☐ Married ☐ Separated ☑ Divorced ☐ Widowed ☐ Never Married
Household Income	☐ Less than $20,700 ☐ $20,701 – $45,000 ☑ $45,001 – $75,000 ☐ $75,001 and over

© Negnevitsky, Pearson Education, 2002

A data set for a hypertension study (continued)

Community Health Survey: Hypertension Study (California, U.S.A.)	
Alcohol Consumption	☐ Abstain from alcohol ☐ Occasional (a few drinks per month) ☑ Regular (one or two drinks per day) ☐ Heavy (three or more drinks per day)
Smoking	☐ Nonsmoker ☐ 1 – 10 cigarettes per day ☑ 11 – 20 cigarettes per day ☐ More than one pack per day
Caffeine Intake	☐ Abstain from coffee ☑ One or two cups per day ☐ Three or more cups per day
Salt Intake	☐ Low-salt diet ☑ Moderate-salt diet ☐ High-salt diet
Physical Activities	☐ None ☑ One or two times per week ☐ Three or more times per week
Weight 17 cm ***Height*** 93 kg	
Blood Pressure	☐ Optimal ☐ Normal ☑ High

© Negnevitsky, Pearson Education, 2002

Data cleaning

Decision trees are as good as the data they represent. Unlike neural networks and fuzzy systems, decision trees do not tolerate noisy and polluted data. Therefore, the data must be cleaned before we can start data mining.

We might find that such fields as *Alcohol Consumption* or *Smoking* have been left blank or contain incorrect information.

© Negnevitsky, Pearson Education, 2002

Data enriching

From such variables as *weight* and *height* we can easily derive a new variable, *obesity*. This variable is calculated with a body-mass index (BMI), that is, the weight in kilograms divided by the square of the height in metres. Men with BMIs of 27.8 or higher and women with BMIs of 27.3 or higher are classified as obese.

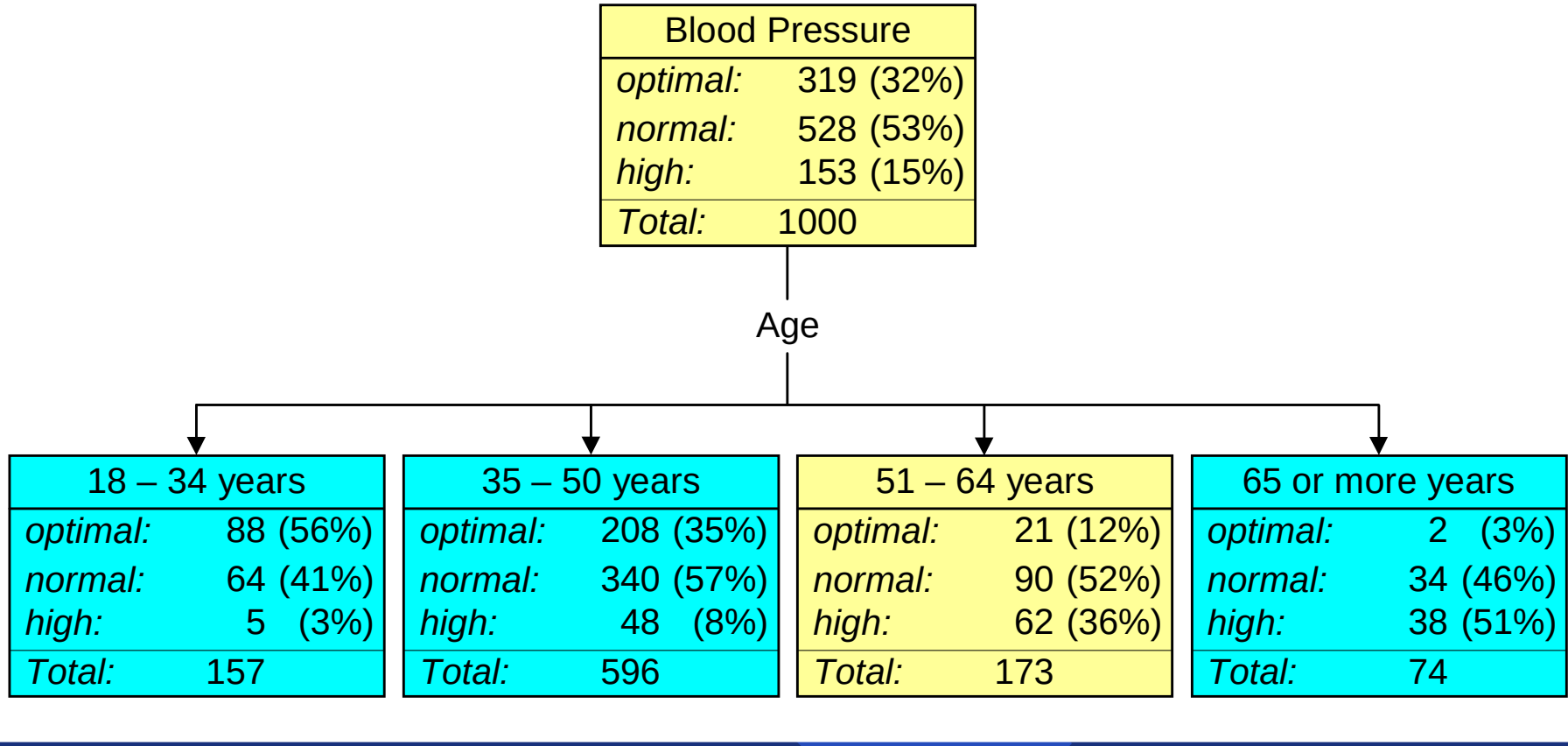
© Negnevitsky, Pearson Education, 2002

A data set for a hypertension study (continued)

Community Health Survey: Hypertension Study (California, U.S.A.)	
Obesity	☑ Obese ☐ Not Obese

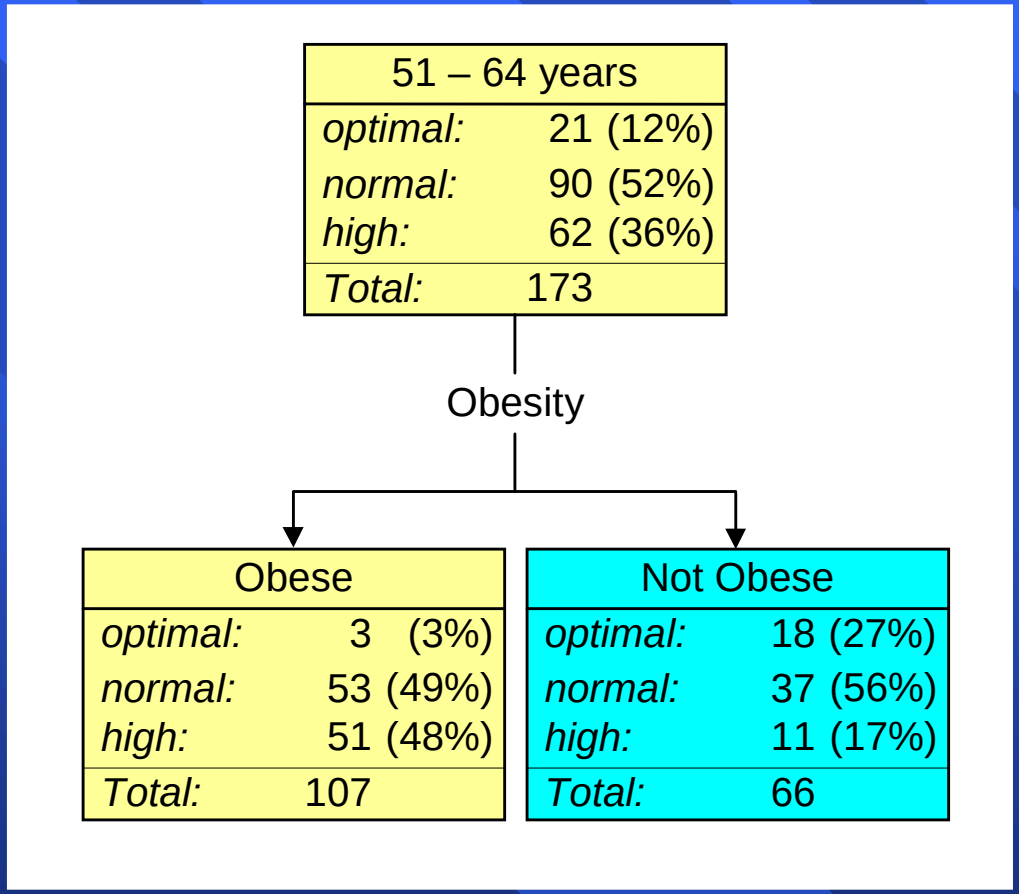
© Negnevitsky, Pearson Education, 2002

Growing a decision tree



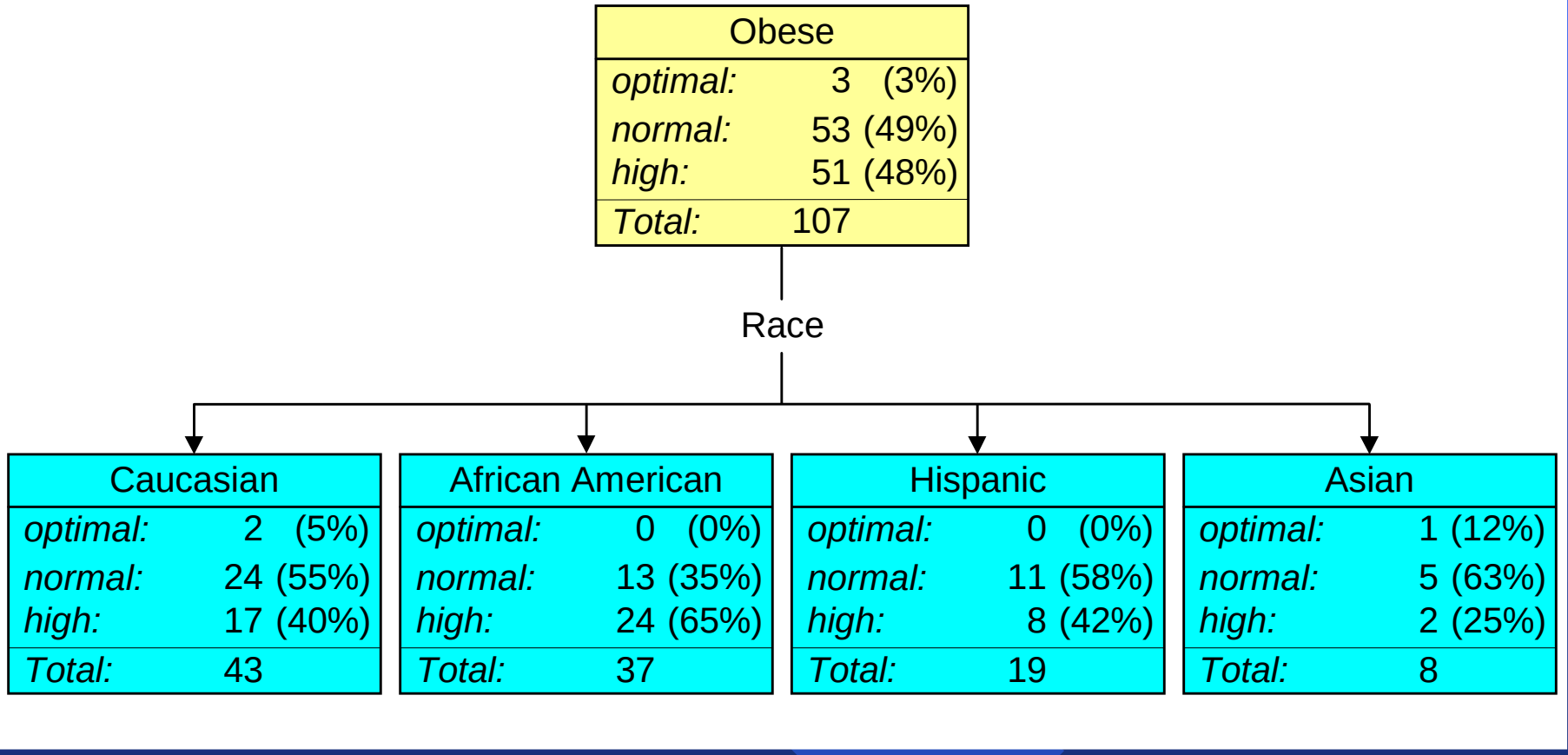
© Negnevitsky, Pearson Education, 2002

Growing a decision tree (continued)



© Negnevitsky, Pearson Education, 2002

Growing a decision tree (continued)



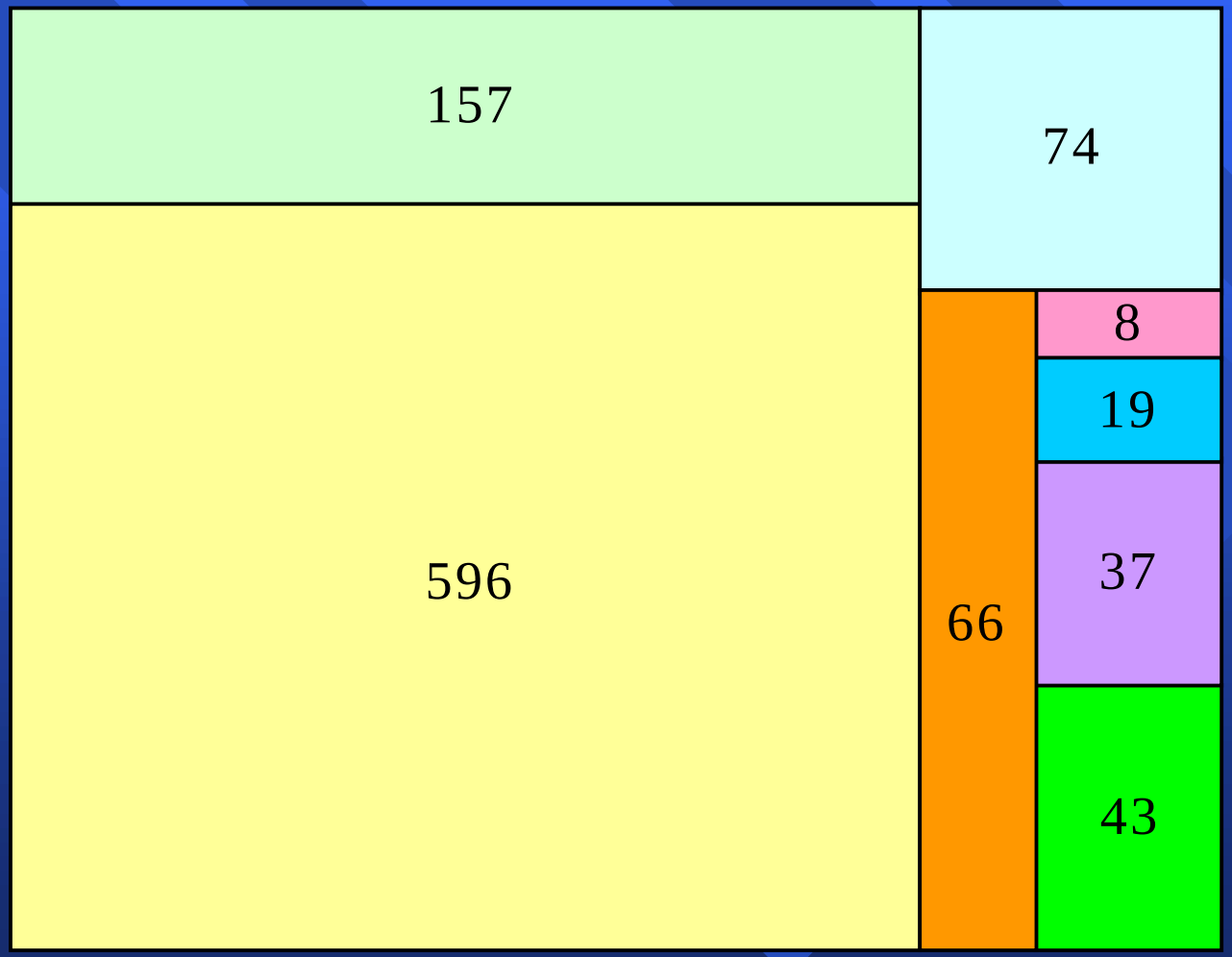
© Negnevitsky, Pearson Education, 2002

Solution space of the hypertension study

The solution space is first divided into four rectangles by *age*, then age group 51-64 is further divided into those who are overweight and those who are not. And finally, the group of obese people is divided by *race*.

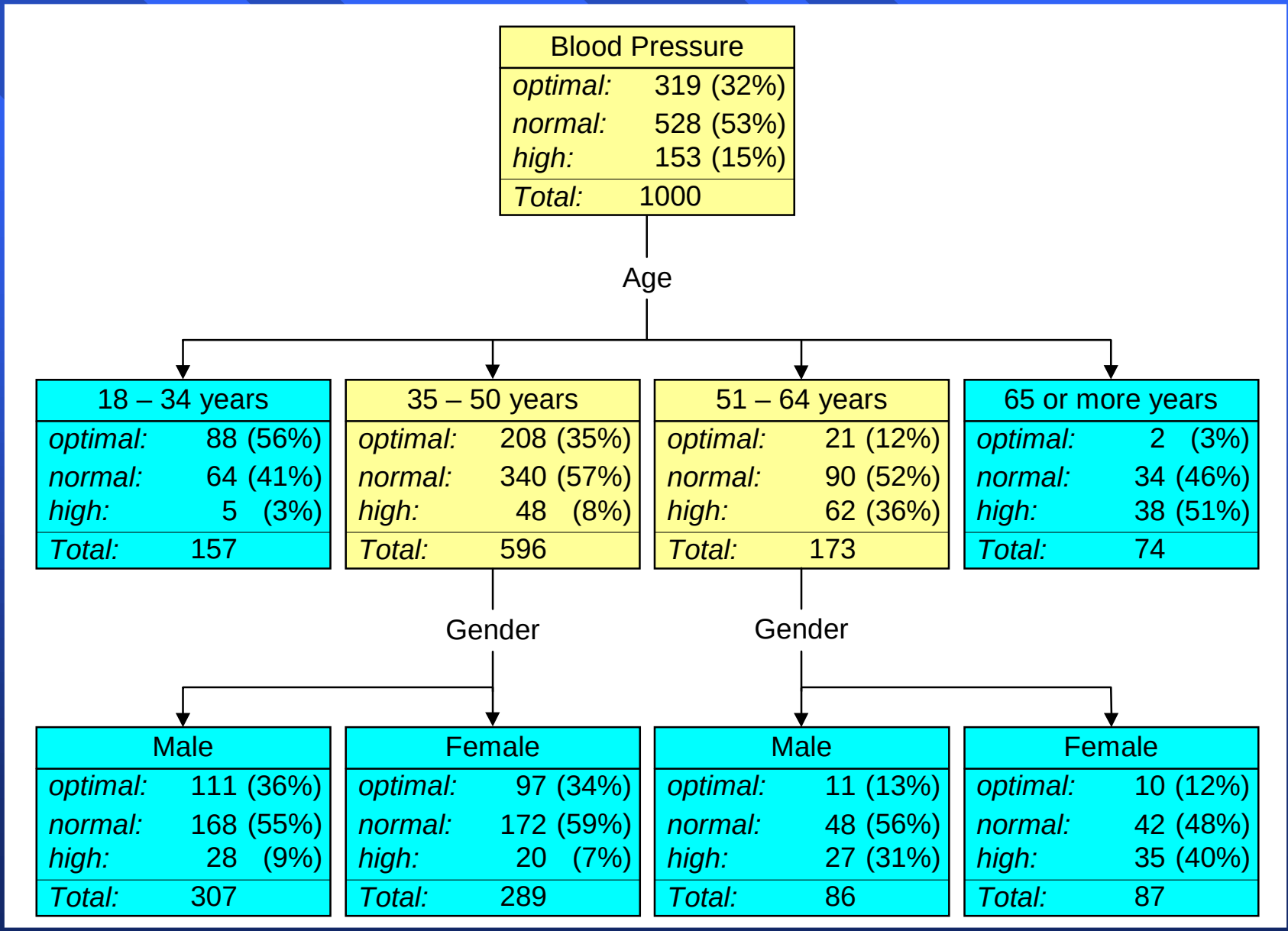
© Negnevitsky, Pearson Education, 2002

Solution space of the hypertension study



© Negnevitsky, Pearson Education, 2002

Hypertension study: forcing a split



© Negnevitsky, Pearson Education, 2002

Advantages of decision trees

- The main advantage of the decision-tree approach to data mining is it visualises the solution; it is easy to follow any path through the tree.
- Relationships discovered by a decision tree can be expressed as a set of rules, which can then be used in developing an expert system.

© Negnevitsky, Pearson Education, 2002

Drawbacks of decision trees

- Continuous data, such as age or income, have to be grouped into ranges, which can unwittingly hide important patterns.
- Handling of missing and inconsistent data – decision trees can produce reliable outcomes only when they deal with “clean” data.
- Inability to examine more than one variable at a time. This confines trees to only the problems that can be solved by dividing the solution space into several successive rectangles.

© Negnevitsky, Pearson Education, 2002

In spite of all these limitations, decision trees have become the most successful technology used for data mining.

An ability to produce clear sets of rules make decision trees particularly attractive to business professionals.

© Negnevitsky, Pearson Education, 2002