Data Mining Functionalities

1. **Class / Concept Description**
2. **Mining Frequent Patterns, Associations, and Correlations**
3. **Classification and Regression for Predictive Analysis**
4. **Cluster Analysis**
5. **Outlier Analysis**

(Kinds of Patterns that can be mined)

3. Classification and Regression for Predictive Analysis

Classification is the process of finding a model that describes and distinguishes data classes or data concepts

The model are derived based on the analysis of a set of training data. This model is used for predicting the class label of unknown dataset.

Ex:				
	80	80	80	Excellent
	70	70	70	Very Good
	60	60	60	Good
	90	80	90	Excellent
	75	70	60	Very Good
	60	62	61	Good
	65	61	62	Unknown?
	82	80	85	Unknown?

3. Classification and Regression for Predictive Analysis

Examples :

Emails classification based on types (Binary classification – two labels)

Designation of faculty – (Multi-label classification – more than one label)

Grade of students -??

Pass result of student - ??

3

Data Mining Functionalities

3. Classification and Regression for Predictive Analysis

Binary classification refers to predicting one of two classes

Multi-class classification involves predicting one of more than two classes.

"Classification refers to a <u>predictive modeling problem</u> where a class label is predicted for a given example of input data"

3. Classification and Regression for Predictive Analysi

Regression:

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor).

Ex: $y = mx+c$ (linear regression)

The main difference between them is that the output variable in **regression** is numerical (or continuous) while that for **classification** is categorical (or discrete).

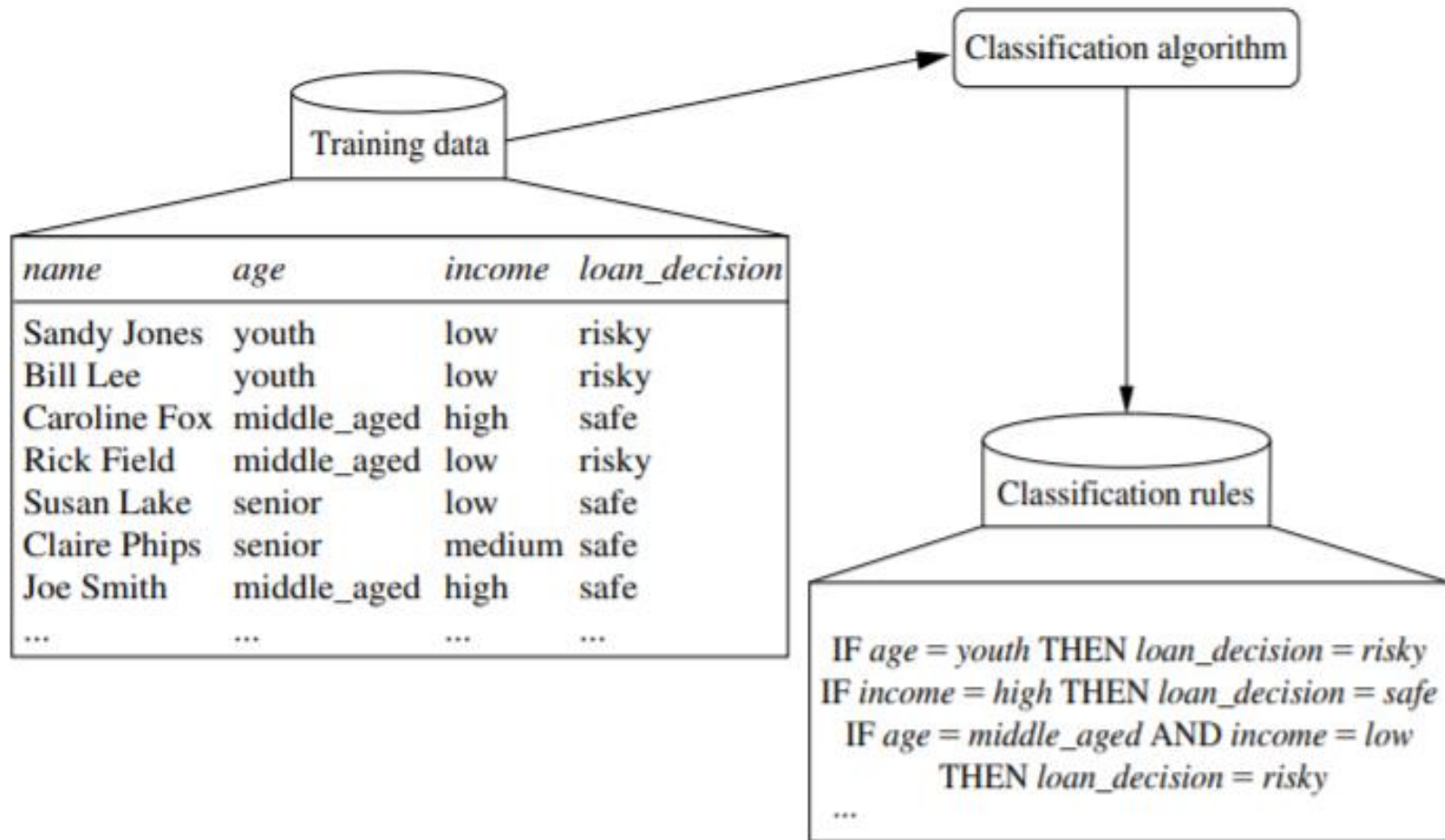## 3. Classification and Regression for Predictive Analysis

Key steps for finding the classification model

1. Learn or train the data
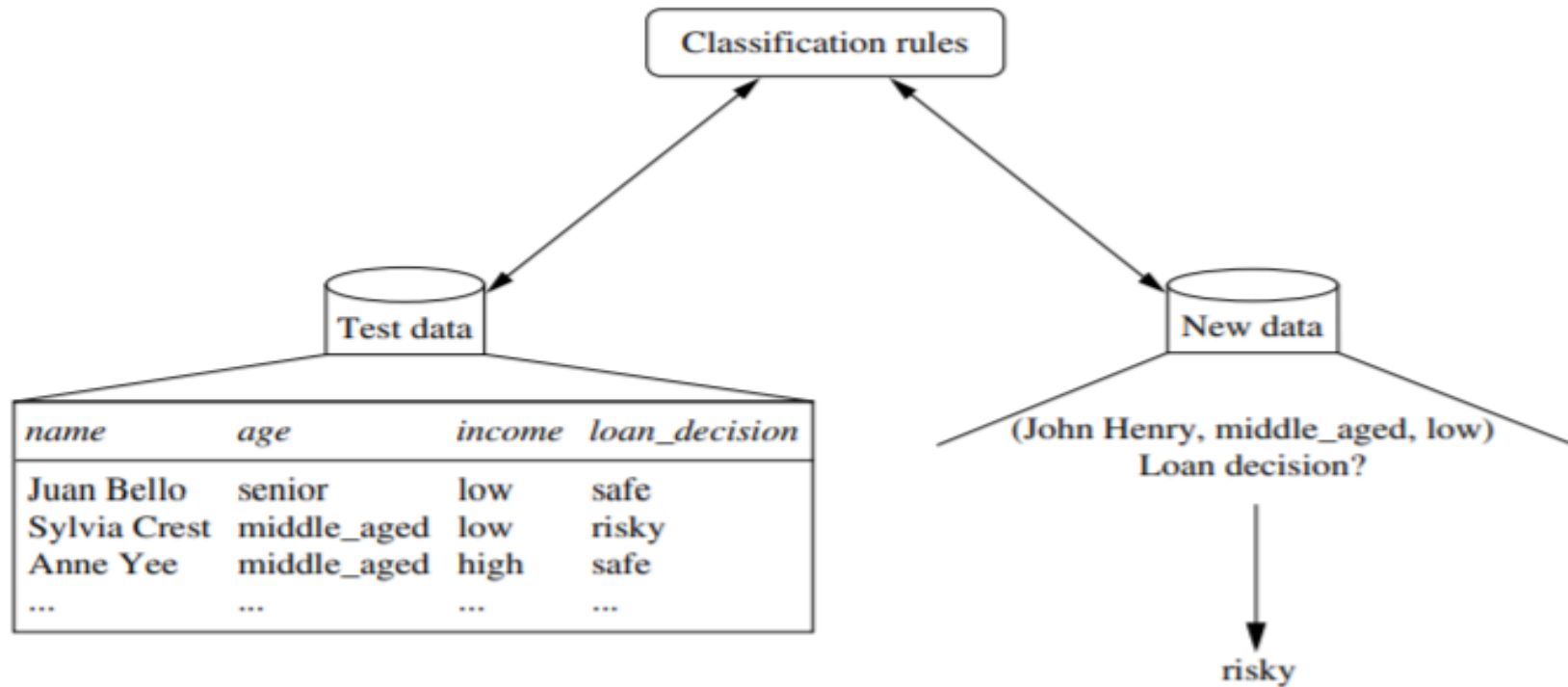2. Build the model
3. Classify the unknown test data

3. Classification and Regression for Predictive Analysis

Learning step:

Training data

name	age	income	loan_decision
Sandy Jones	youth	low	risky
Bill Lee	youth	low	risky
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risky
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
...	...	...	...

Classification algorithm

Classification rules

IF *age = youth* THEN *loan_decision = risky*
IF *income = high* THEN *loan_decision = safe*
IF *age = middle_aged* AND *income = low*
    THEN *loan_decision = risky*
...

3. Classification and Regression for Predictive Analysi
Classification Step:

3. Classification and Regression for Predictive Analysi

Classification Step:

Unknown tuples

(ramu, youth, high) = ?

(gopi, senior, high)=?

(karan, middle, low)=?

(bhavitha, middle, high)=?

(arun, senior, low)=?

9

3. Classification and Regression for Predictive Analysis (Contd..)

Forms of Classification
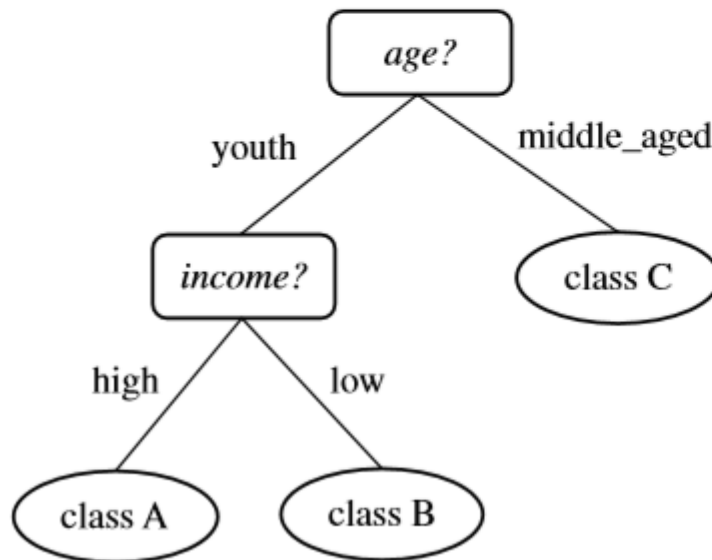
1. IF-THEN Rules
2. Decision Tree
3. A Neural Network

10

3. Classification and Regression for Predictive Analysis (Contd..)

age(X, "youth") AND income(X, "high") $\longrightarrow$ *class(X, "A")*
age(X, "youth") AND income(X, "low") $\longrightarrow$ *class(X, "B")*
age(X, "middle_aged") $\longrightarrow$ *class(X, "C")*
age(X, "senior") $\longrightarrow$ *class(X, "C")*

(a)

(b)
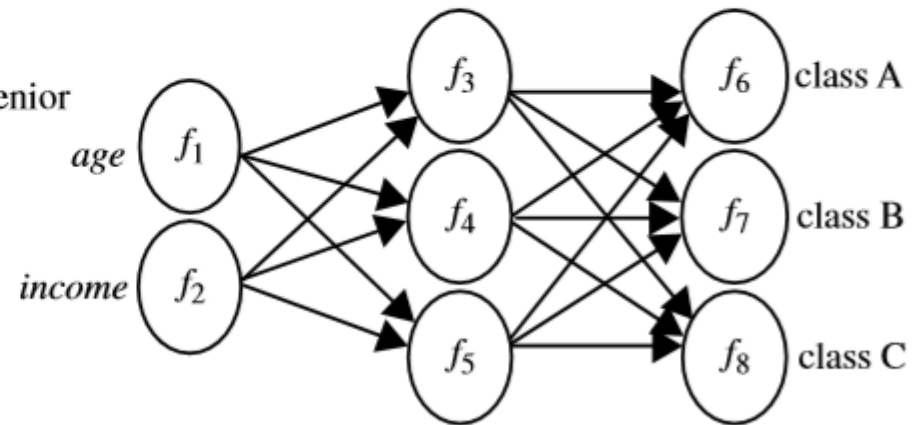
(c)

3. Classification and Regression for Predictive Analysis (Contd..)

Regression is used for predicting the any missing value or unavailable numerical data rather than class labels

Ex:
GDP in 'current year' will have effect on vehicle sales 'next year'

X=GDP of the previous year
Y= Sales of Current Year

Predicting the sales of current year based on GDP of previous year
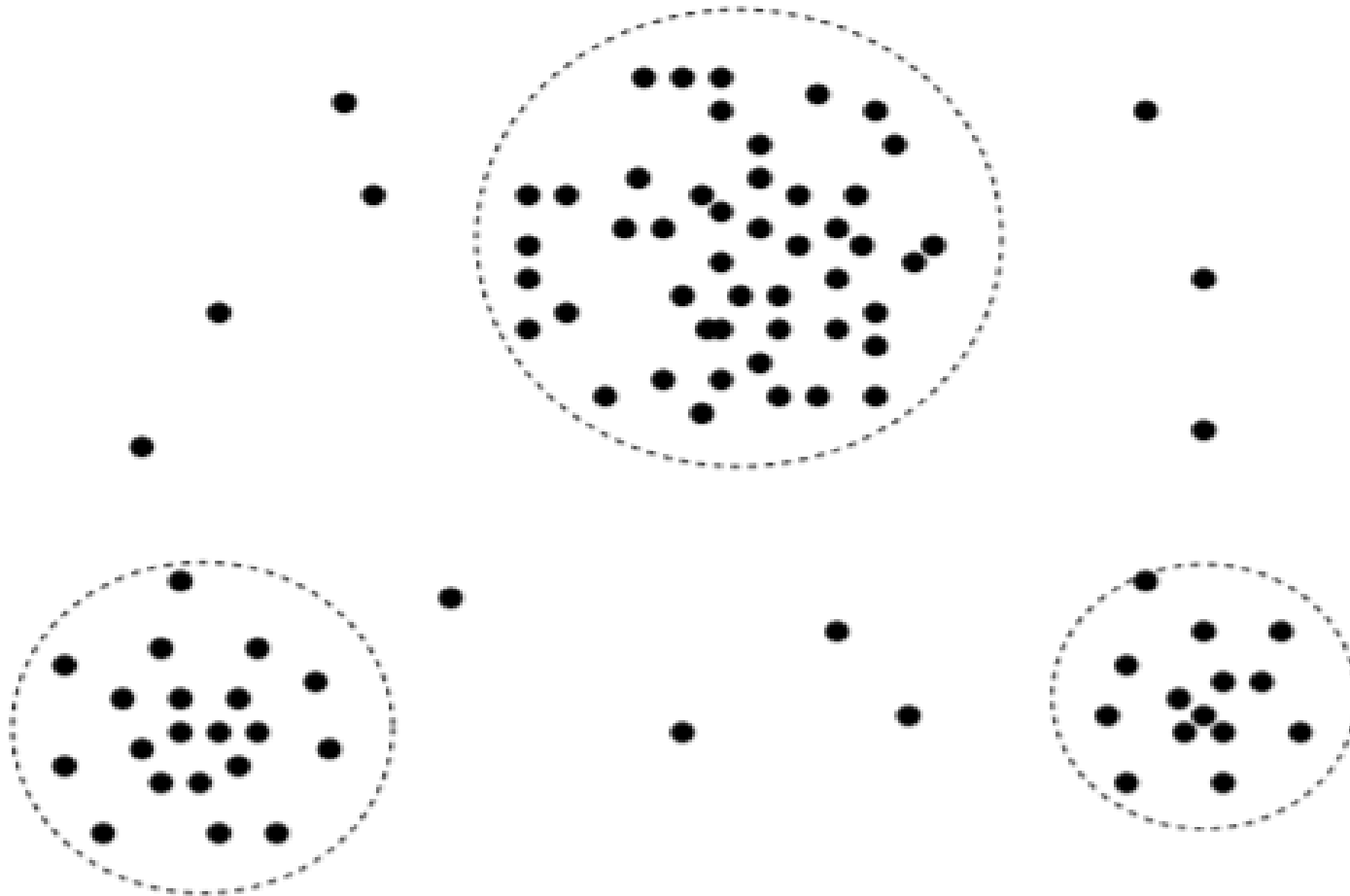**Y=mX+c ,**

In model creation of regression, regression coefficients 'm' and 'c' are derived

12

Cluster Analysis

**Clustering analyzes data objects without consulting class labels.**

"The objects are clustered or grouped based on the principle of <u>maximizing the intraclass similarity</u> and <u>minimizing the interclass similarity</u>**"**

Cluster Analysis

Outlier Analysis : **Objects do not comply with the general behavior of the system or existing clusters- those objects are referred to as outlier**
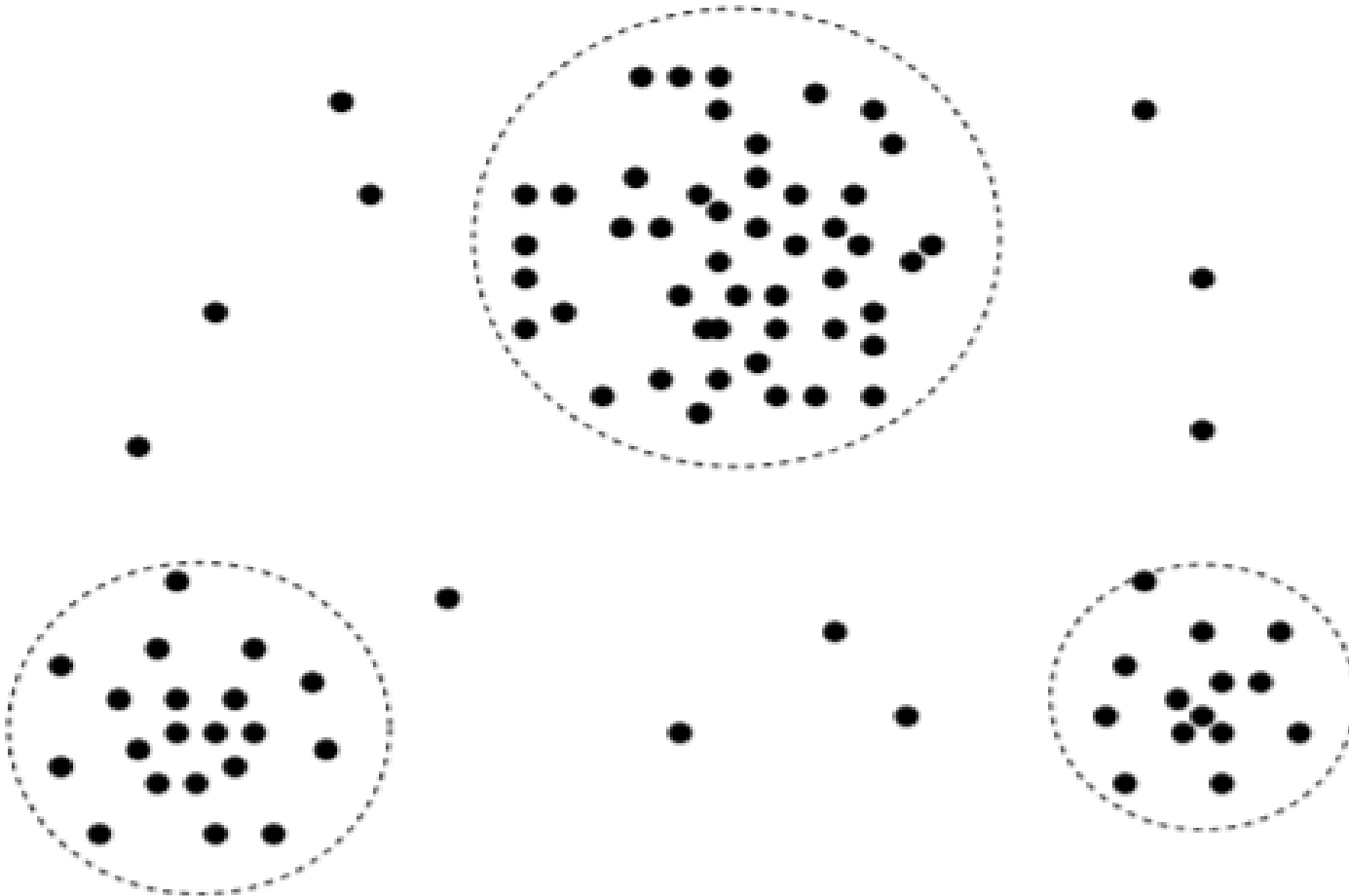
1. Define classification.
2. Distinguish between classification and regression.
3. What are the key steps of classification process.
4. Write the sample examples of classification
5. Write an example of regression
6. Define clustering process
7. What are the important principles of clustering?
8. How the clustering process is different from classification process?
9. Describe the clustering process with an example.
10. Describe outlier analysis

Thank You