Mineração de Dados

Conceitos Básicos

Sumário

Conceitos Básicos

Aprendizado Supervisionado e Não Supervisionado

- Aprendizado Supervisionado
  - Classificação e regressão
  - Supervisão: os dados de treinamento são acompanhados por valores esperados
  - Os valores esperados podem vir de observações, medições, indicações de um especialista, etc
  - Novas instâncias são classificadas com o que se aprendeu sobre os dados de treinamento

- Aprendizado Não Supervisionado
  - Agrupamento e Regras de Associação
  - Não há um valor esperado nos dados de treinamento
  - Deve-se estabelecer relações entre elementos

Regressão e Classificação

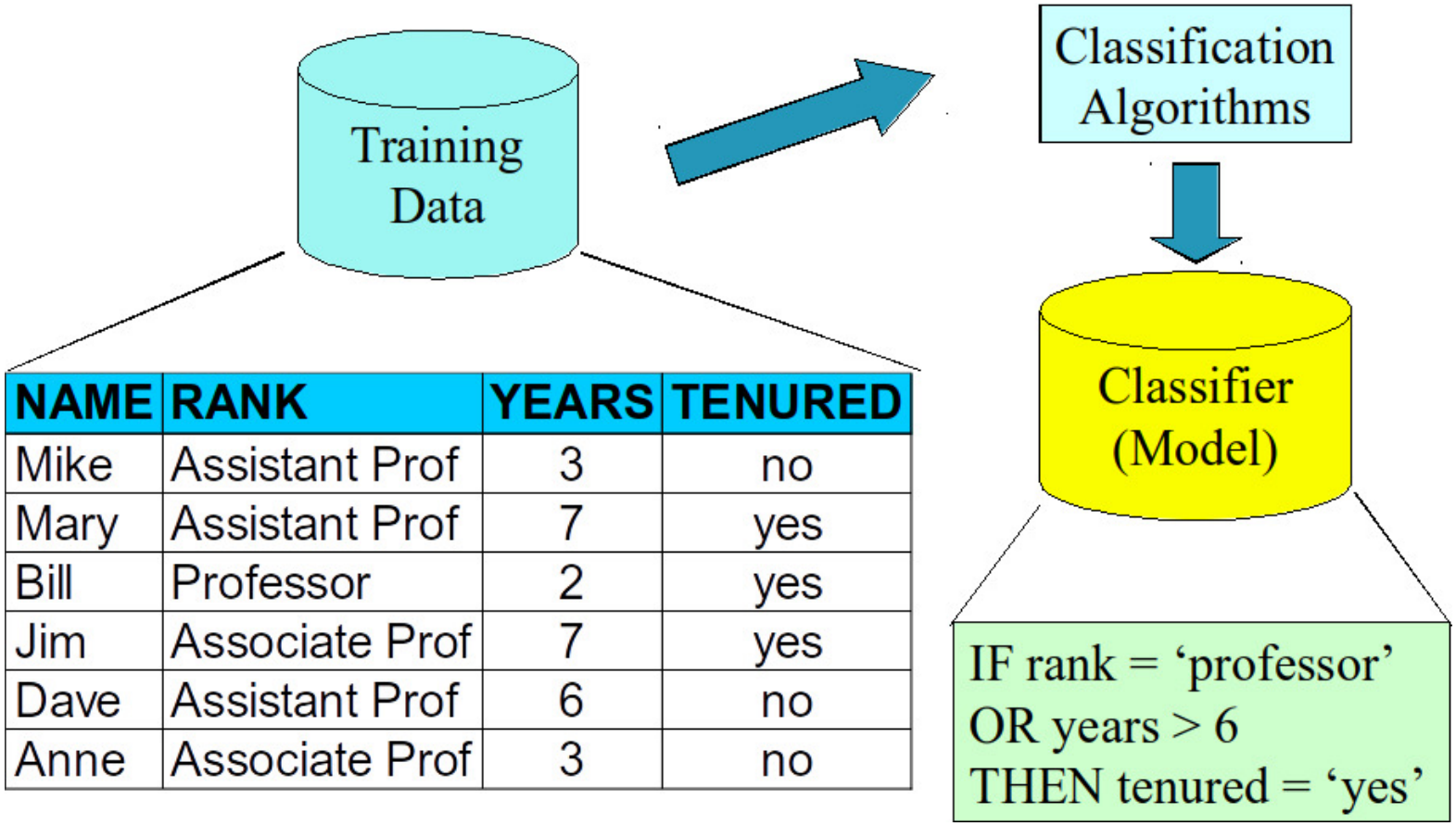
Regressão e Classificação

- Regressão
  - Predição numérica
  - Os modelos são funções que retornam valores contínuos
- Classificação
  - O modelo deve prever uma classe para um conjunto de valores de entrada
- Algumas aplicações de classificação
  - Aprovação de empréstimos
  - Diagnóstico médico
  - Detecção de fraude
  - Categorização de páginas

Processo de Classificação

- Construção e utilização do modelo
- Construção
  - Como relacionar os atributos com o valor esperado
  - Cada tupla é associada a uma classe
  - Um conjunto de dados de treinamento é usado para criar o modelo
  - Um modelo pode ser representado por superfícies separadoras, regras de classificação, árvores de decisão ou expressões aritméticas

Construção do Modelo

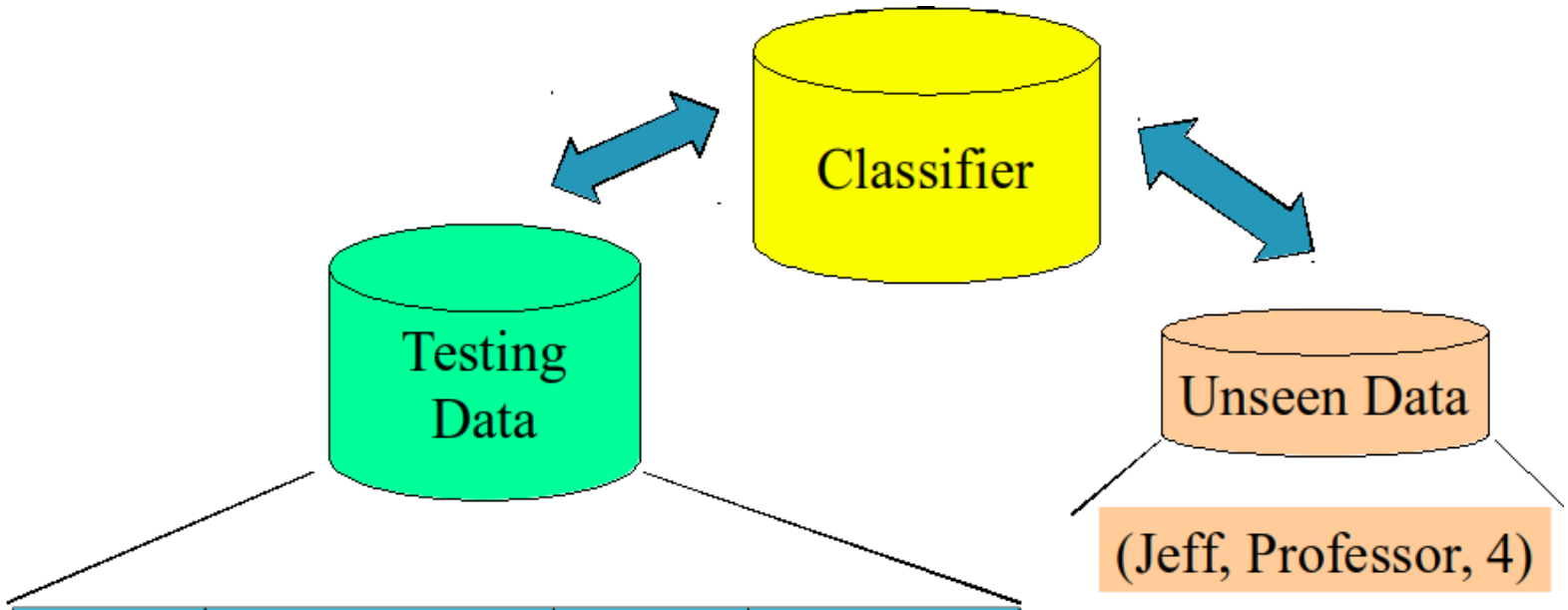


NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Processo de Classificação

- Construção e utilização do modelo
- Utilização do modelo
  - Estimar a acurácia do modelo
  - As classes indicadas pelo modelo para um conjunto de dados de teste podem ser usadas para determinar sua qualidade
  - Acurácia: porcentagem dos dados de teste que são corretamente classificados pelo modelo
  - O conjunto de dados de teste deve ser independente do conjunto de dados de treinamento
  - Se a acurácia (ou outra forma de avaliação de modelos) for aceitável, o modelo pode ser adotado para prever classes para novas instâncias
  - Um subconjunto de dados pode ser usado para selecionar modelos e/ou seus parâmetros e, neste caso, este é chamado de dados de validação

Utilização do Modelo



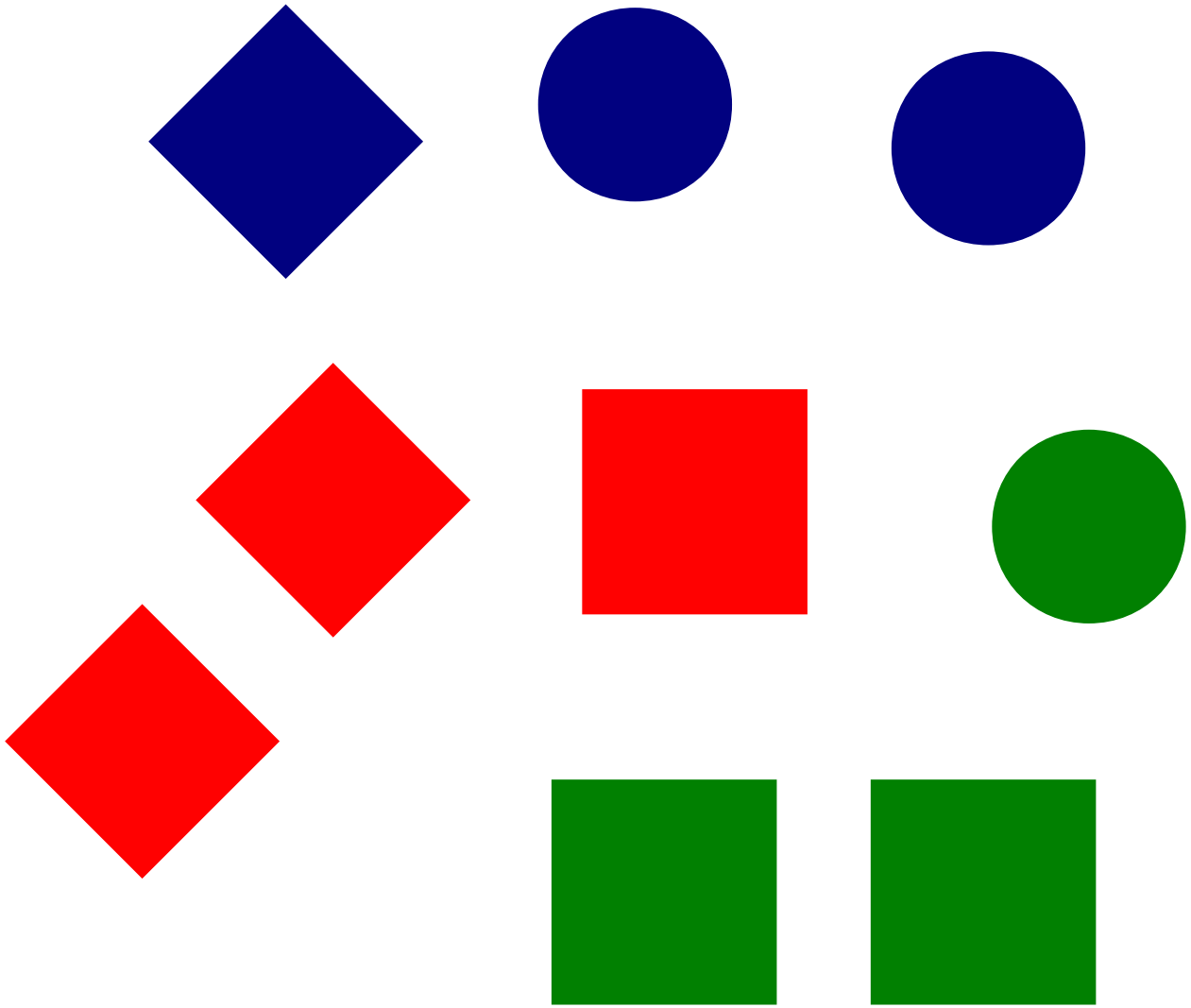
NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Agrupamento e Regras de Associação

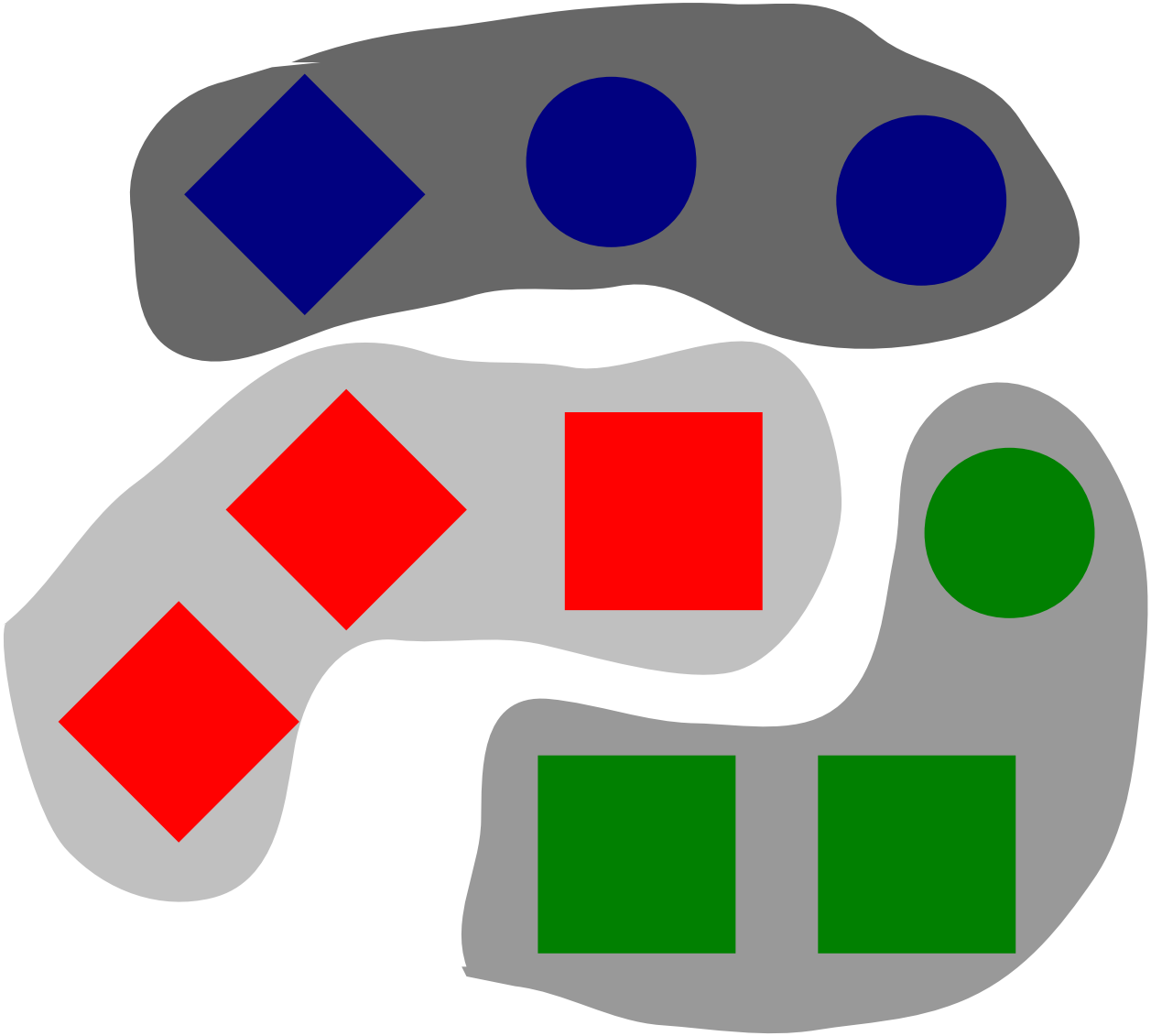
Agrupamento

- Particionar os dados em grupos baseando-se em similaridade
- Apenas uma representação dos grupos é necessária
  - Centroides e diâmetro, por exemplo
- Há vários tipos de métodos agrupamento

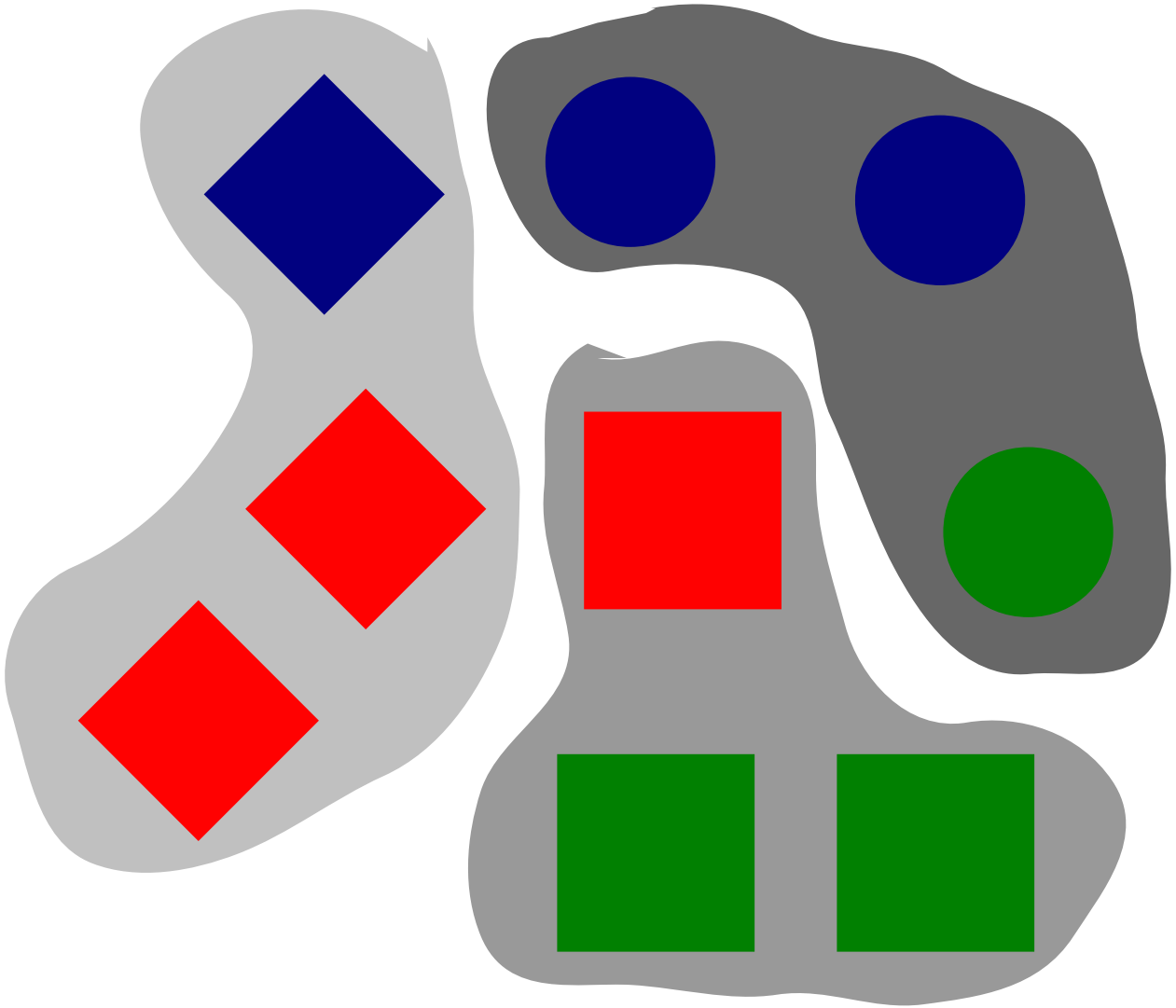
Agrupamento



Agrupamento



Agrupamento



Regras de Associação

- Descobrir itens que ocorram em comum
- Formular regras que associem esses itens que ocorrem com conjuntamente com frequência
  - fralda $\rightarrow$ cerveja
- Bases transacionais

Regras de Associação

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

(a) Binary database

t	$\mathbf{i}(t)$
1	$ABDE$
2	BCE
3	$ABDE$
4	$ABCE$
5	$ABCDE$
6	BCD

(b) Transaction database

x	A	B	C	D	E
$\mathbf{t}(x)$	1	1	2	1	1
	3	2	4	3	2
	4	3	5	5	3
	5	4	6	6	4
		5			5
		6			

(c) Vertical database

sup	itemsets
6	B
5	E, BE
4	$A, C, D, AB, AE, BC, BD, ABE$
3	$AD, CE, DE, ABD, ADE, BCE, BDE, ABDE$

Regra: $BC \rightarrow E$

Regras de Associação

ID	Pão	Leite	Fralda	Cerveja	Ovo	Café
1	1	1	1	1	1	0
2	0	1	0	0	0	1
3	1	0	1	1	1	0
4	0	0	1	1	0	1
5	1	1	1	0	0	1

Regra: fralda $\rightarrow$ cerveja