Projeto da Disciplina

Germano C. Vasconcelos
Centro de Informática - UFPE

Objetivo

Realizar um estudo experimental sobre a aplicação de modelos de redes neurais em um problema do mundo real

Motivações

- Possibilitar ao aluno uma visão prática do uso de redes neurais na solução de problemas
- Consolidar os conhecimentos teóricos apresentados em sala de aula
- Permitir o contato com ferramentas do Github, Keras, Scikit-learn na Linguagem Python

Projeto – Análise de Risco de Crédito

- Classificação de padrões
  - Base real de instituição que vende a crédito
  - Base em larga escala: +- 400 mil registros para treinamento e +-130 mil registros para teste
  - Problema: com base no perfil de clientes, decidir a quem conceder crédito (risco de inadimplência)

Descrição do Projeto

- Conjunto de classificadores disponíveis
  - Perceptron multicamadas (MLP) (obrigatório)
  - Máquina de Vetores de Suporte (obrigatório)
  - Ensemble de MLPs (obrigatório)
  - Random Forest (usado para comparação)
  - Gradient Boosting (usado para comparação)
  - Ensemble de Classificadores (usado para comparação)
- Investigar diferentes topologias da rede e diferentes valores dos parâmetros (básico)
  - Número de camadas
  - Número de unidades intermediárias
  - Influência da taxa de aprendizagem no treinamento
  - Função de ativação
  - Método de amostragem (SMOTE)

Descrição do Projeto

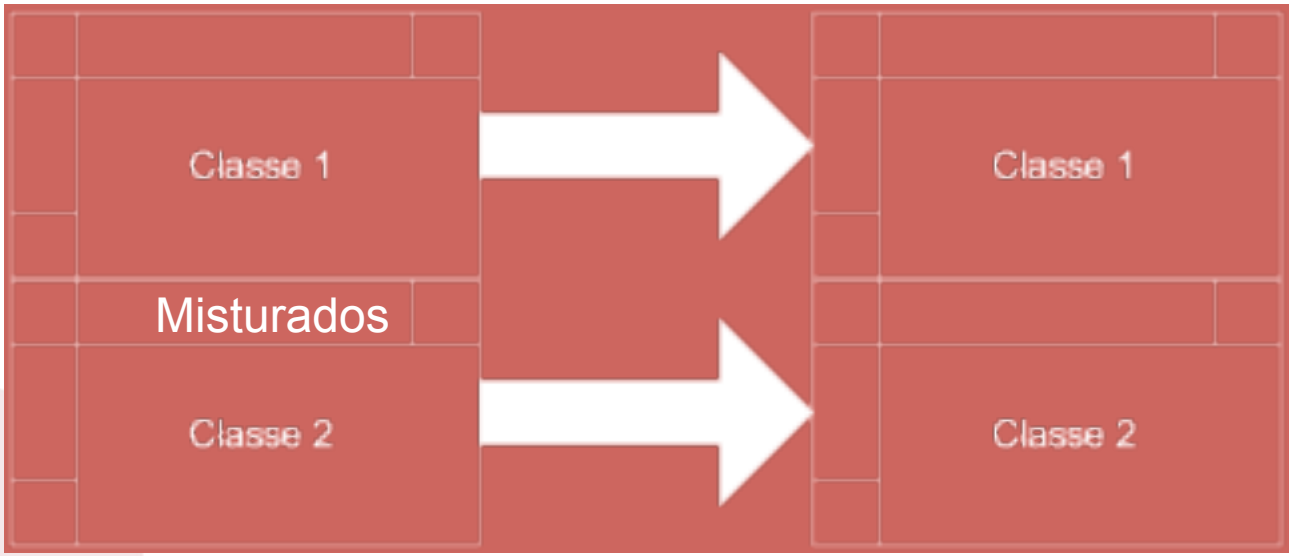
- Parâmetros adicionais que podem ser explorados
  - Algoritmo de aprendizagem
  - Taxa de aprendizagem adaptativa
  - SMOTE Adaptado
  - Outros

Preparação de Dados: (**divisão e balanceamento**)

- Conjuntos de dados independentes
  - Treinamento
  - Validação
  - Teste (já está separado)
- Estatisticamente representativos e independentes
  - Não pode haver sobreposição

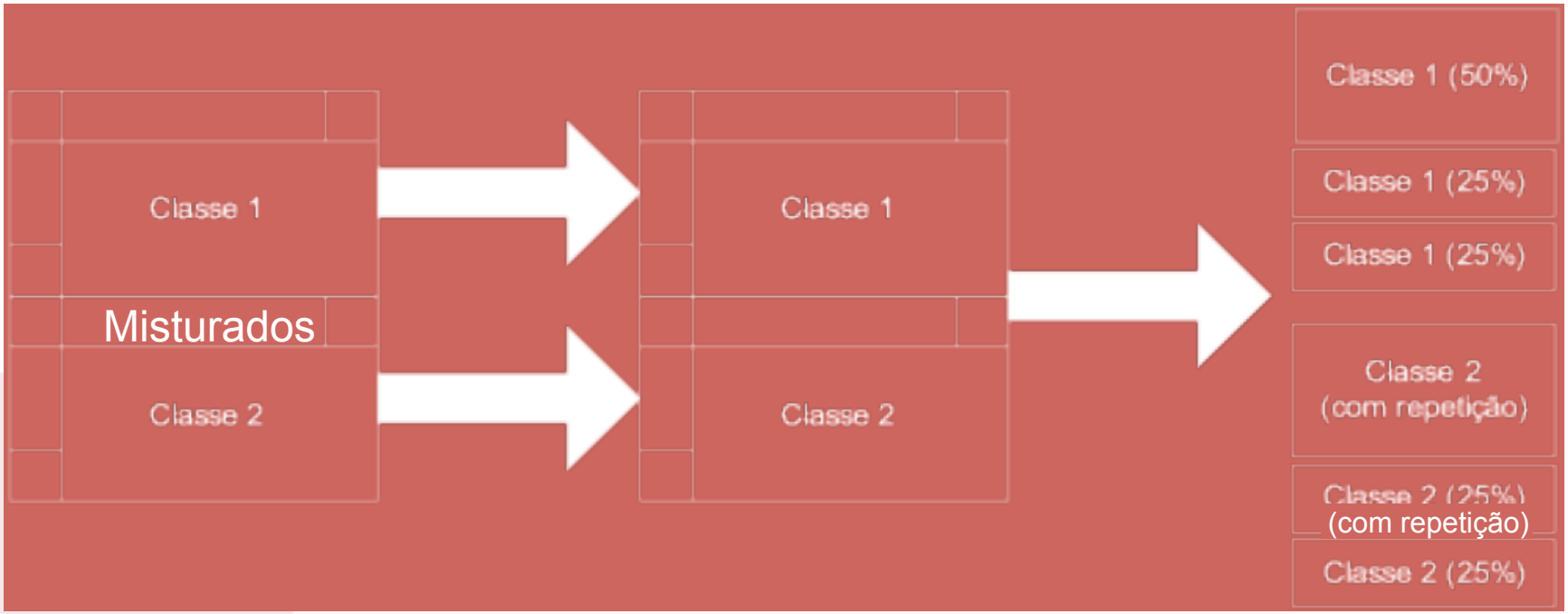
Preparação de Dados: (**divisão e balanceamento**)

Particionamento dos Dados – Primeira etapa



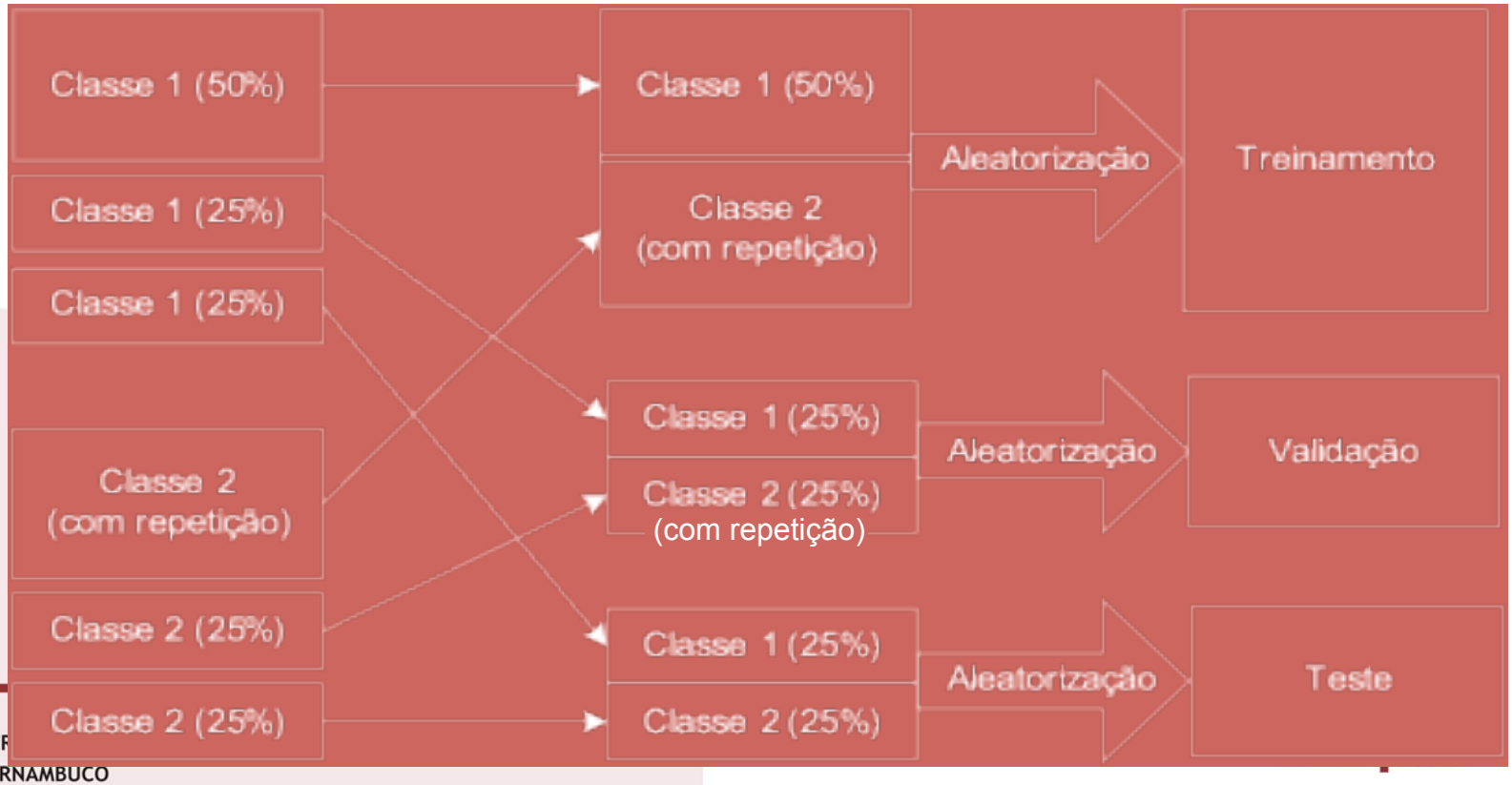
Preparação de Dados: (**divisão e balanceamento**)

Particionamento dos Dados – Segunda etapa



Preparação de Dados: (**divisão e balanceamento**)

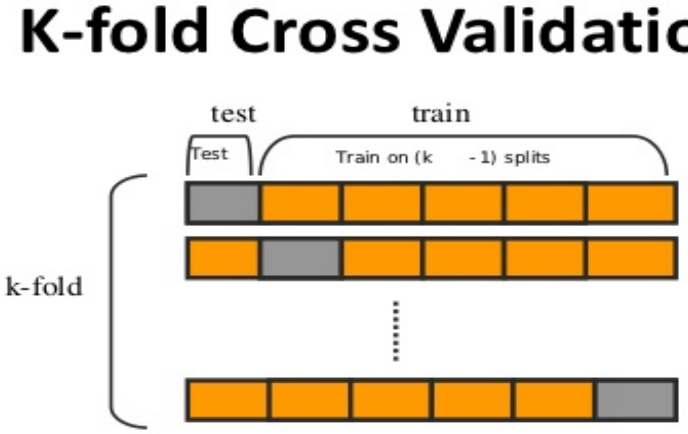
Particionamento dos Dados – Terceira etapa



Preparação de Dados: (**divisão e balanceamento**)

Particionamento dos Dados com K-folds

OBS: use 1 fold para validação também em cada rodada

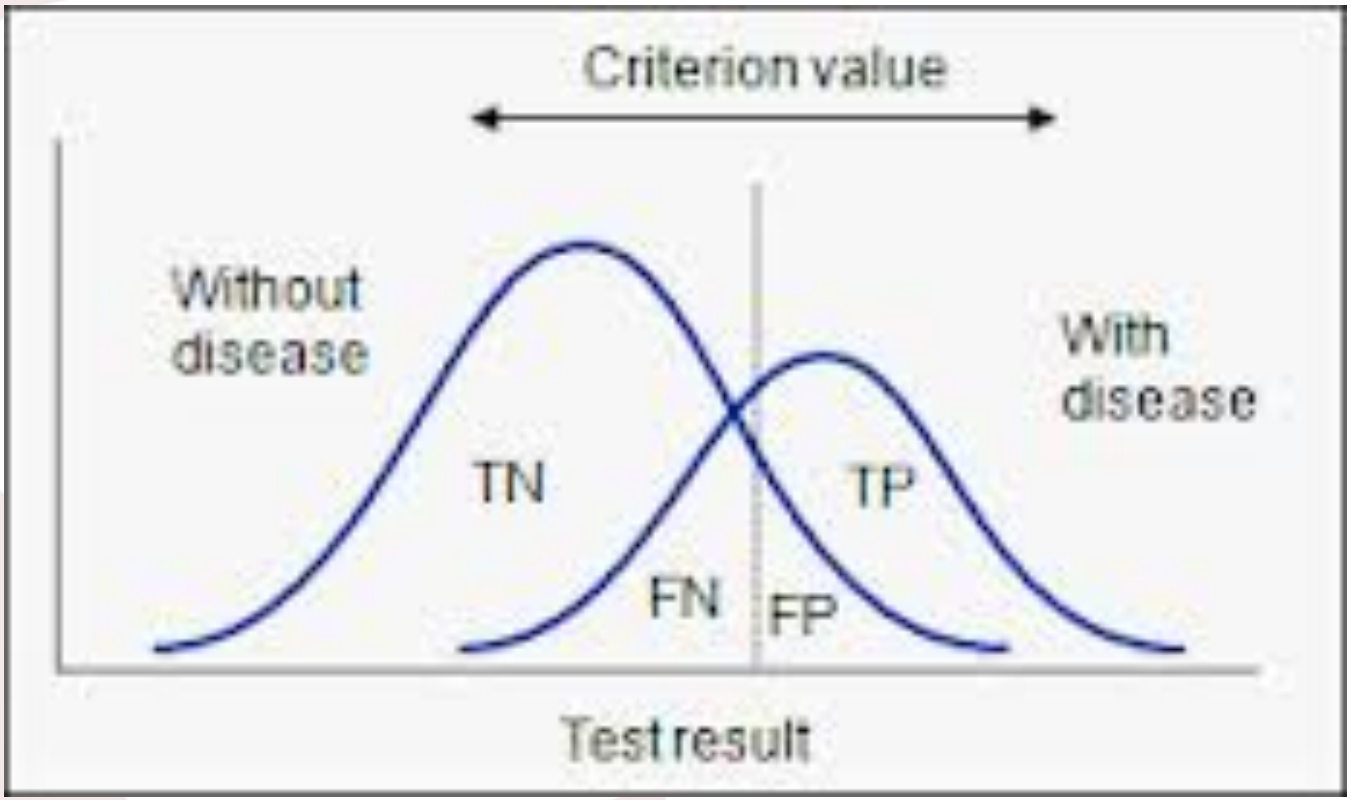


- Randomly divide your data into K pieces/folds
- Treat 1st fold as the test dataset. Fit the model to the other folds (training data).
- Apply the model to the test data and repeat k times.
- Calculate statistics of model accuracy and fit from the test data only.

Análise de Desempenho

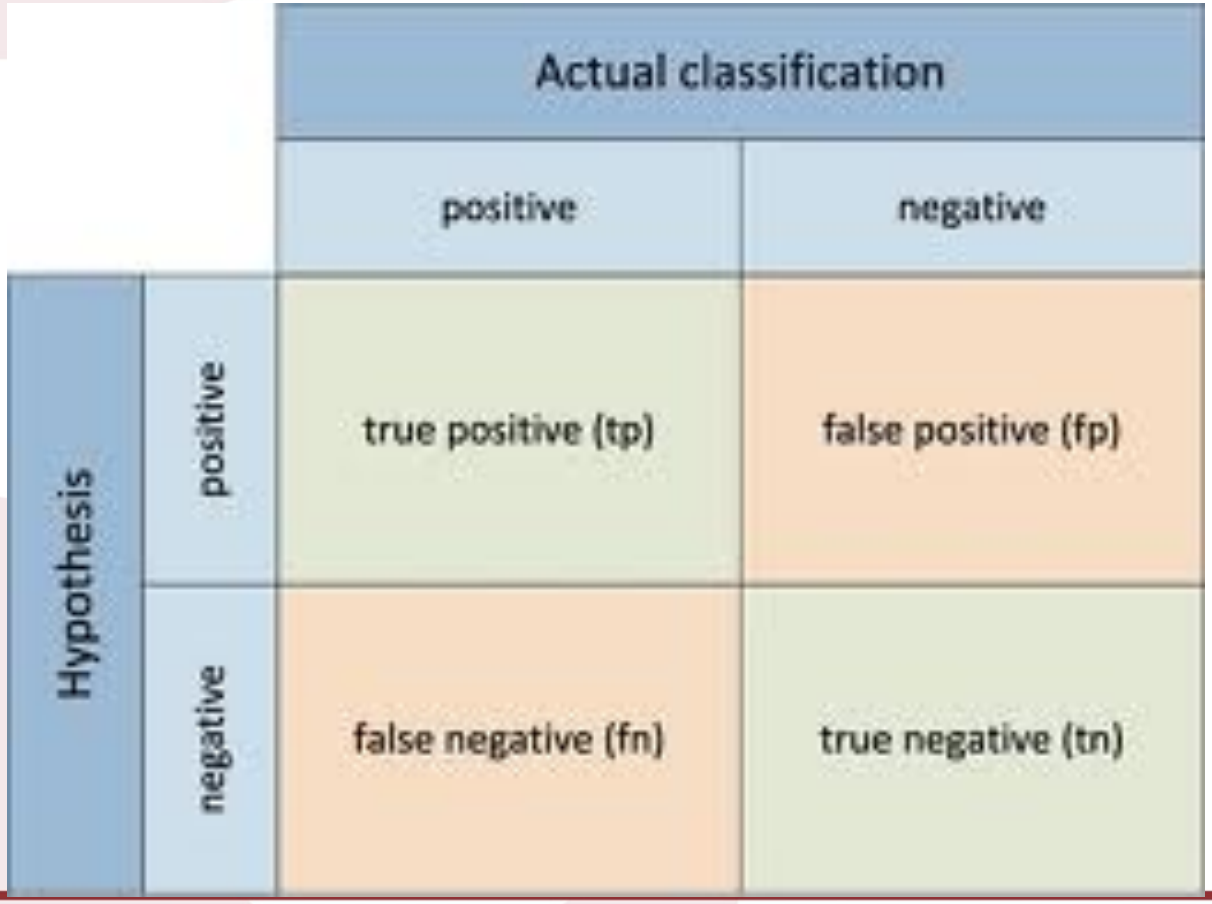
- Classificação
  - MSE (erro médio quadrado)
  - Teste estatístico Kolmogorov-Smirnov -KS (principal)
  - Matriz de confusão
  - Auroc (Área sob a Curva Roc)
  - Recall, Precision e F-Measure

Avaliação (Desempenho e Resultados)



Avaliação (Desempenho e Resultados)

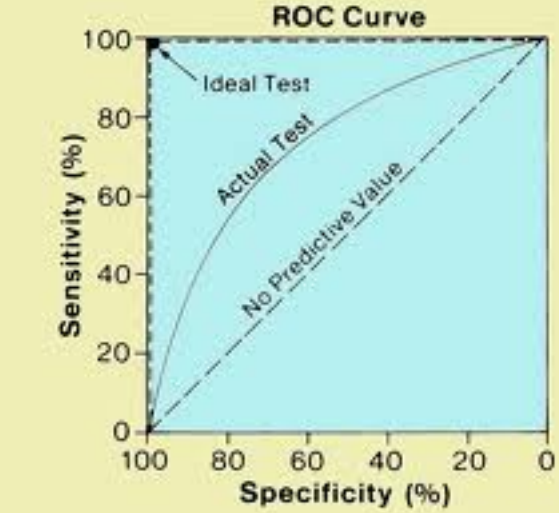
Matriz de Confusão

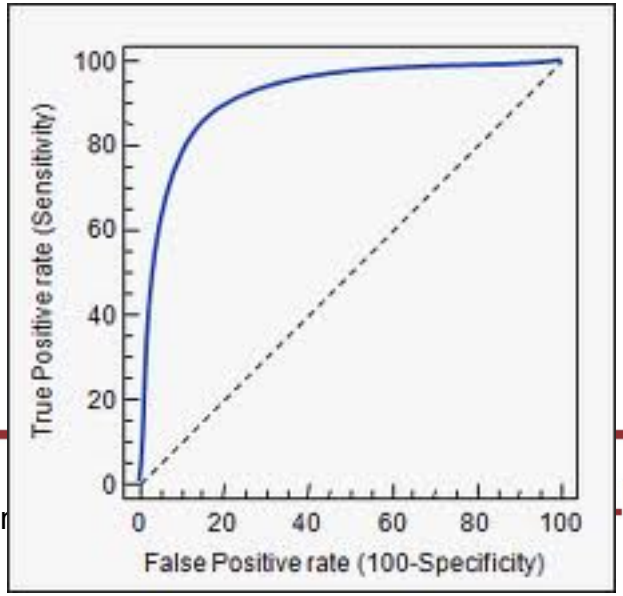


Avaliação (Desempenho e Resultados)

		condition +	condition -	
test outcome	+	True positive	False positive (type I error, p-value)	→ positive predictive value
test outcome	-	False negative (type II error)	True negative	→ negative predictive value
		↓ sensitivity	↓ specificity	

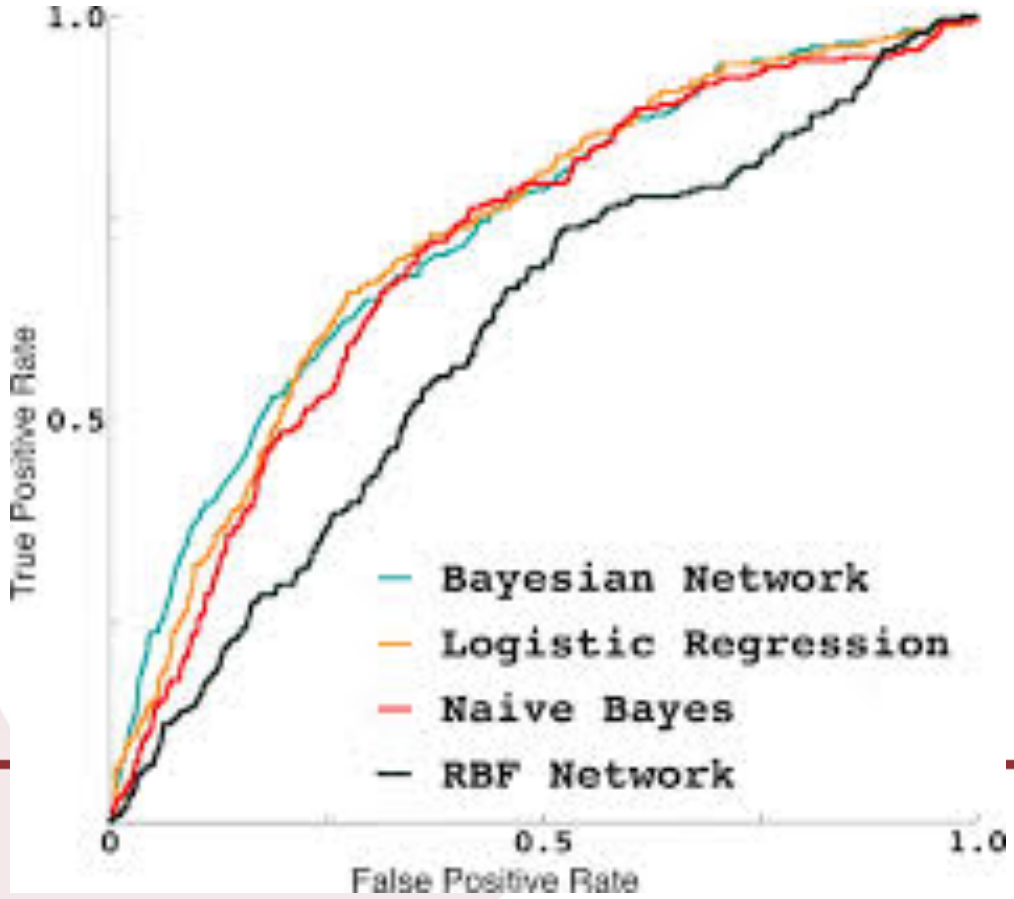
Curvas ROC





Avaliação (Desempenho e Resultados)

Curvas ROC: Exemplo

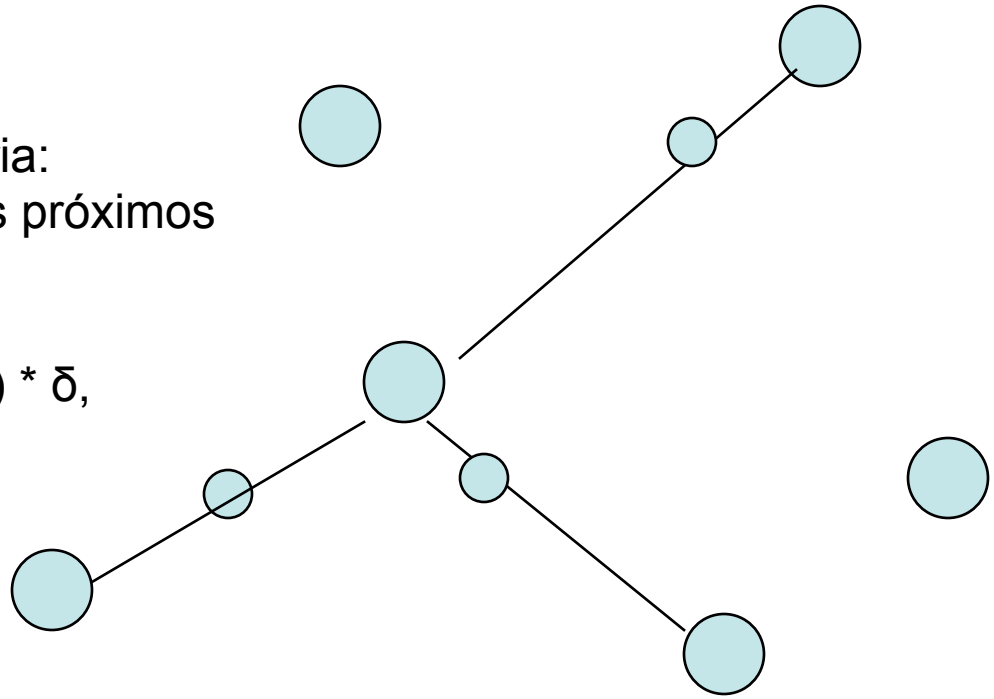


Experimentos Adicionais

- Replicação (*oversampling*) da Classe Minoritária com SMOTE
- Replicação (*oversampling*) da Classe Minoritária com SMOTE Adaptado

SMOTE (*Synthetic Minority Oversampling Technique* – Chawla et al)

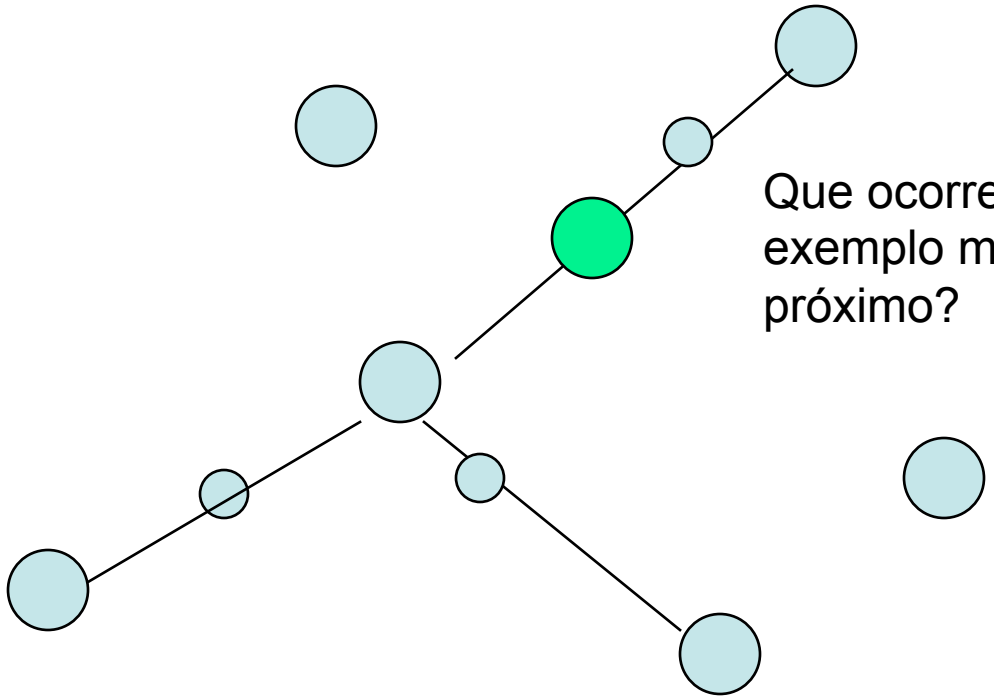
Para a classe minoritária:
Calcule k vizinhos mais próximos (da sua classe);
Escolha 1 deles (x');
Crie um $x_{new} = x + (x' - x) * \delta$,
Onde $\delta \in [0,1]$



: exemplo minoritário

: Novo dado sintético

SMOTE (*Synthetic Minority Oversampling Technique* – Chawla et al)



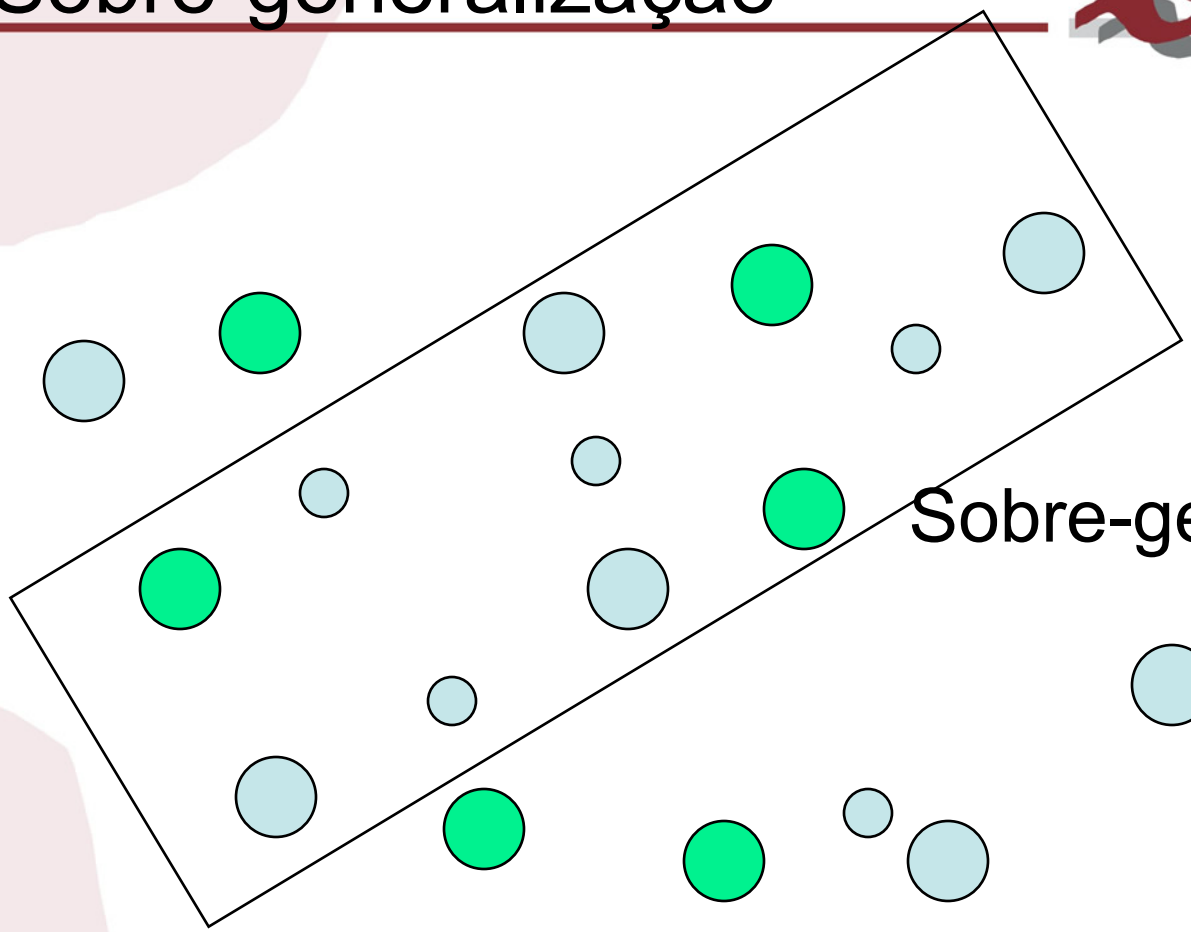
Que ocorre se há um exemplo majoritário próximo?

: exemplo minoritário

: exemplo majoritário

: Novo dado sintético

SMOTE – Tendência de Sobre-generalização



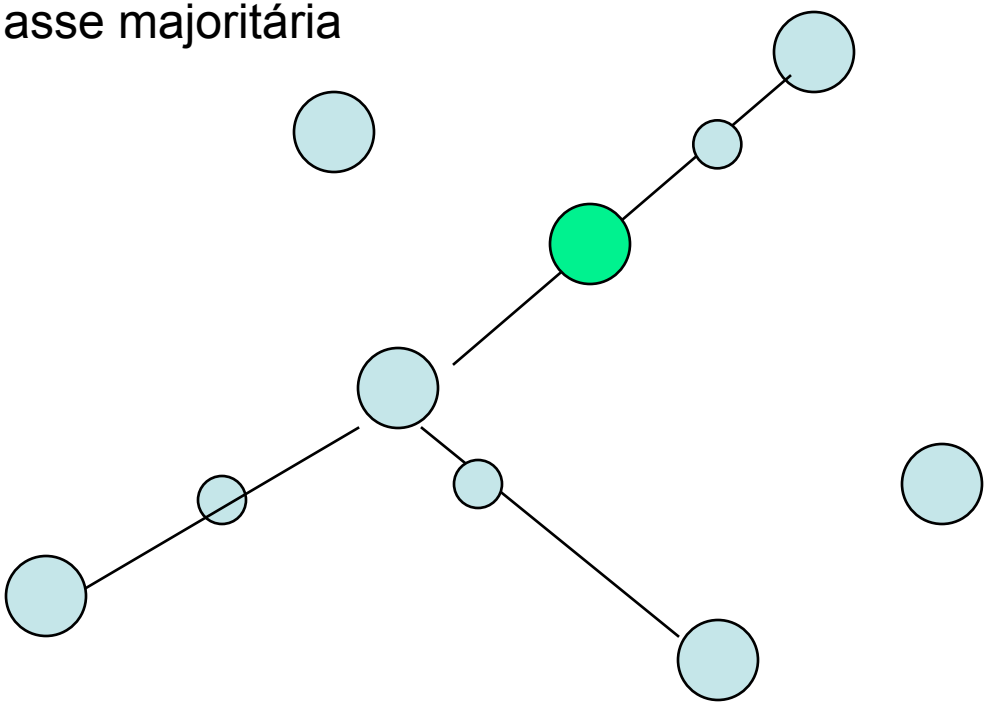
: exemplo minoritário

: exemplo majoritário

: exemplo sintético

Para a classe minoritária:
Calcule k vizinhos mais próximos
(de qualquer classe);
Escolha 1 deles (x');
Crie um $x_{new}=x + (x'-x) * \delta$,
Onde $\delta \varepsilon [0,1]$ se x'for da classe minoritária
e $\delta \varepsilon [0,0.5]$ se x'for da classe majoritária

SMOTE (Adaptado)



: exemplo minoritário

: exemplo majoritário

: Novo dado sintético

Ferramentas para o Projeto

- Código em Python
  - https://github.com/RomeroBarata/IF702-redes-neurais
- Conjuntos de dados do problema
  - http://www.cin.ufpe.br/~gcv/web_lci/intro.html

Resultados do Projeto

- Apresentação com todos do grupo com descrição do problema, divisão dos dados, estrutura experimental e interpretação dos resultados
- Entrega no final do semestre