Credit Card Default Prediction

João E.S. Moreira

Universidade de Coimbra

18 de maio de 2024

Introdução

Os objetivos deste trabalho são

- Utilização de técnicas de redução e seleção de *features*
  - KW
  - AUC
  - PCA
  - LDA
- Classificação de amostras com diversos classificadores simples
  - MDC
  - Fisher LDA
  - Naive bayes
  - K-NN
  - SVM
  - Random Florest

Metodologia

- Pré-processamento dos dados
- Aplicação de técnicas de redução e seleção de features
- Divisão do *data set*
  - 80% treino
  - 20% teste
- Normalização do *data set*
- Aplicação dos classificadores
  - 30 execuções com *data sets* diferentes
- Avaliação dos classificadores com diferentes métricas
  - Exatidão
  - Sensibilidade
  - Especificidade

Exploração dos Dados

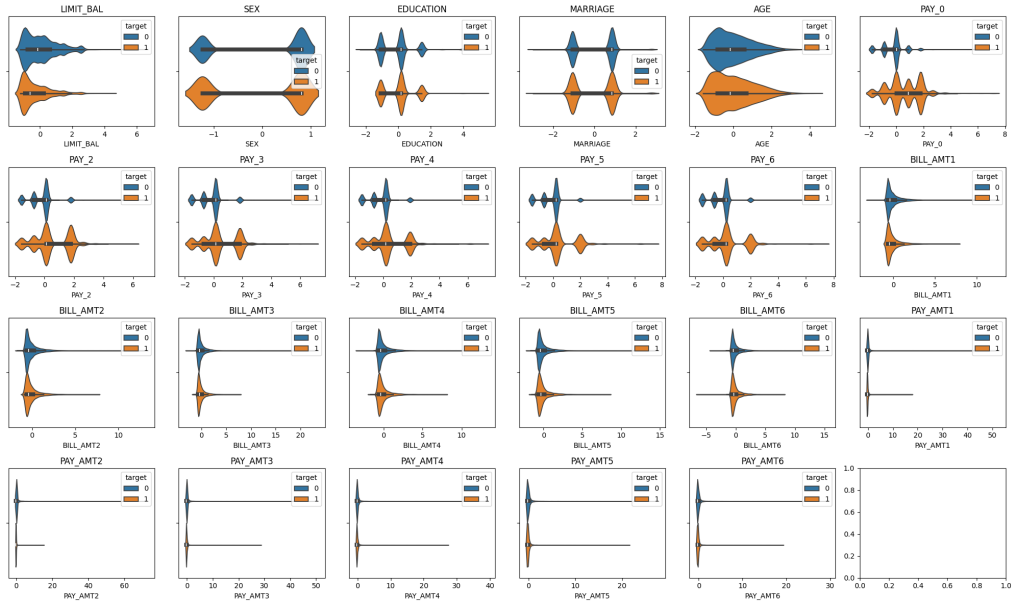


Figura: Distribuição dos dados cruz

- Muitas *features* categóricas
- Muitas *features* sobrepostas
- As melhores *features* aparentam ser as *PAY**

Exploração dos Dados

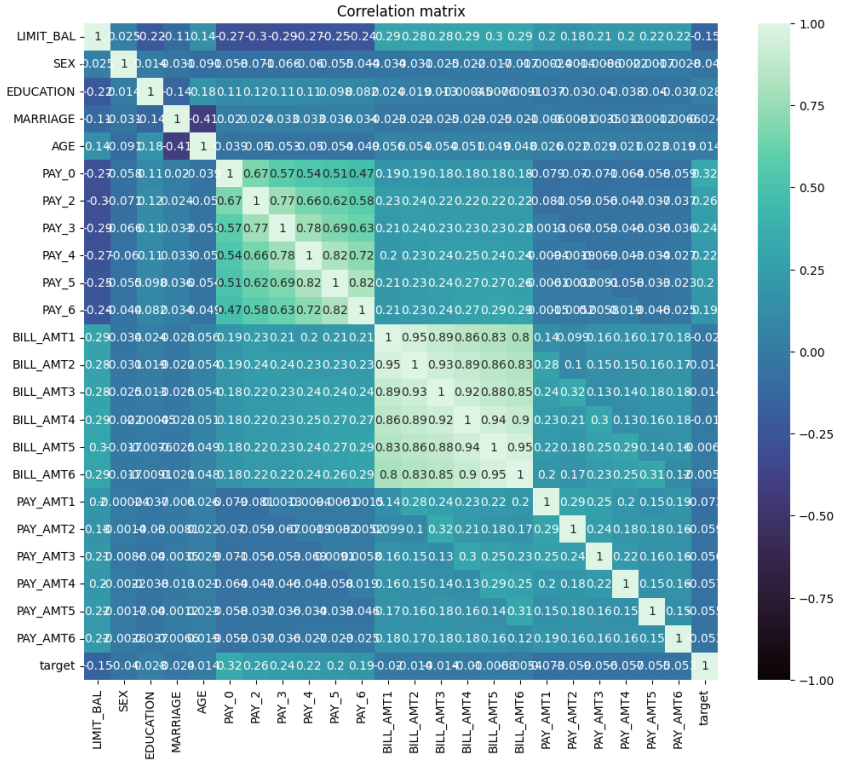


Figura: Distribuição dos dados cruz

- Os grupos *PAY**, *BILL_AMT** e *PAY_AMT** apresentam muita redundância
- As *features* que apresentam maior correlação com as classes *target* são as *PAY**

Resultados da Seleção de Features

KS-Test

- Rejeitamos a hipótese nula
- Nenhuma *feature* apresenta uma distribuição normal dos dados

KW-Test

- As *features* mais promissoras foram *PAY_AMT**

AUC

- As *features* mais promissoras foram *PAY**

Resultados da Seleção de Features

PCA

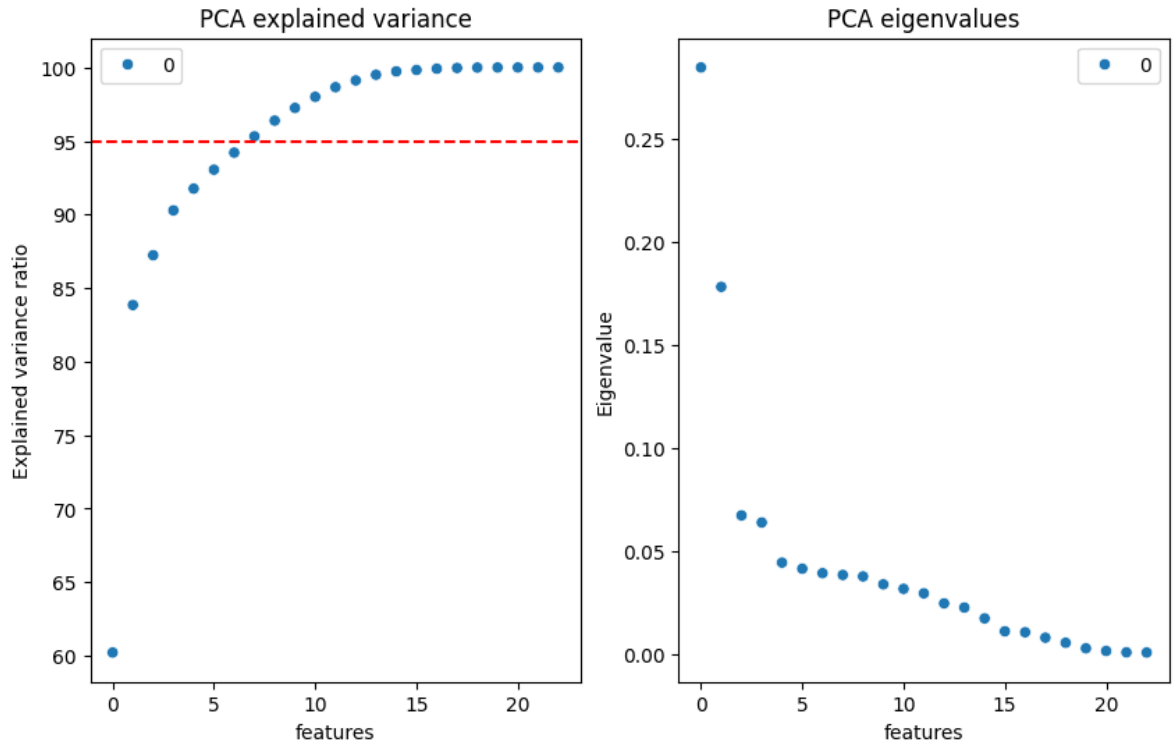


Figura: Resultados da PCA

- Decidi escolher os vetores que permitem uma preservação de 95% da informação

Resultados da Seleção de Features

LDA

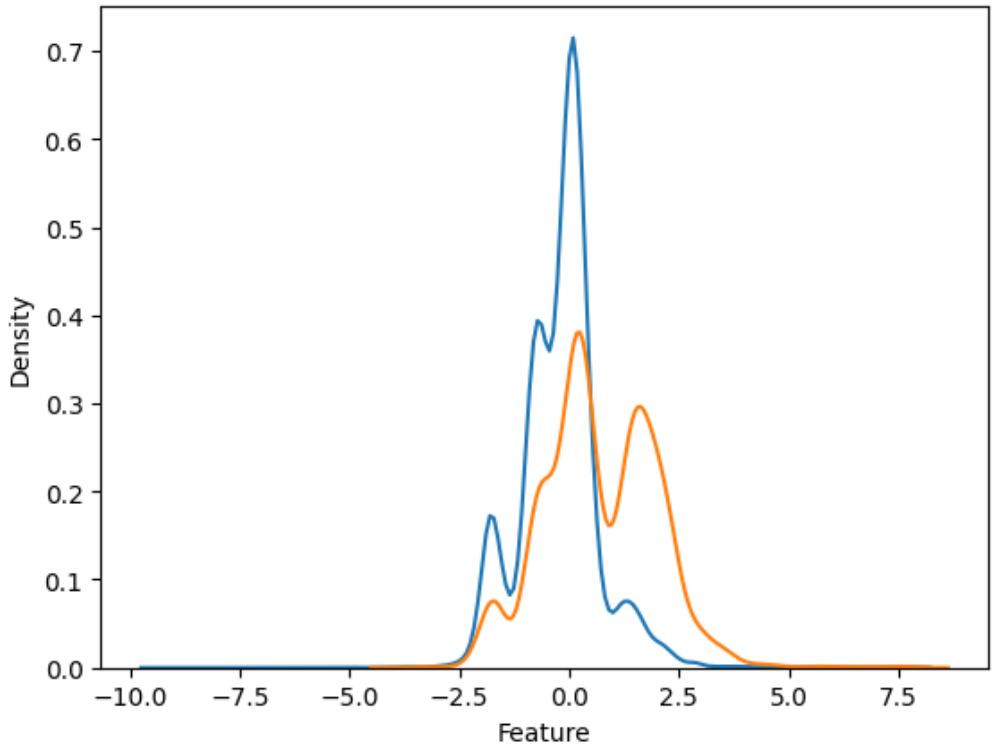
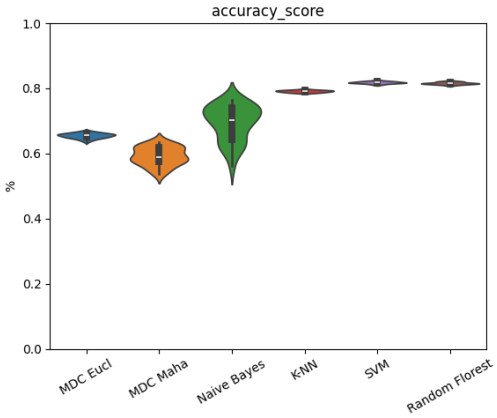


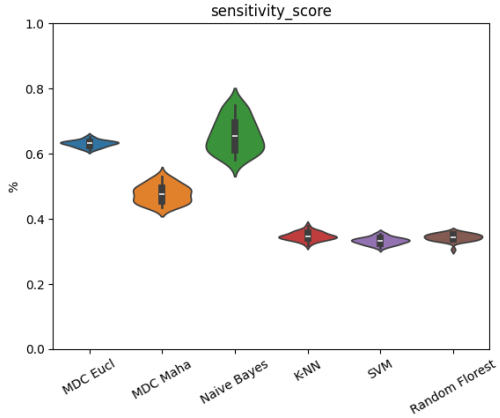
Figura: Resultados da LDA

- Aplicação da LDA redução a uma dimensão
- Apresenta grande parte dos dados sobrepostos

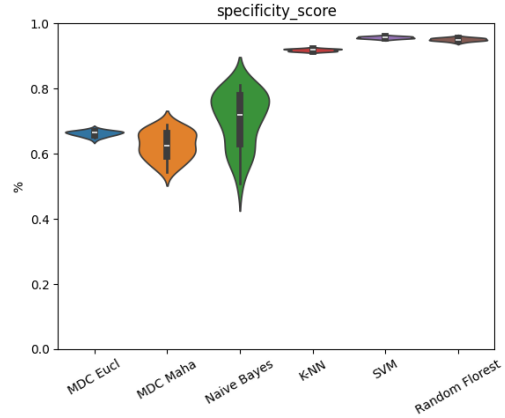
Classificação com Dados Cruz



(a) Exatidão dos classificadores



(b) Sensibilidade dos classificadores

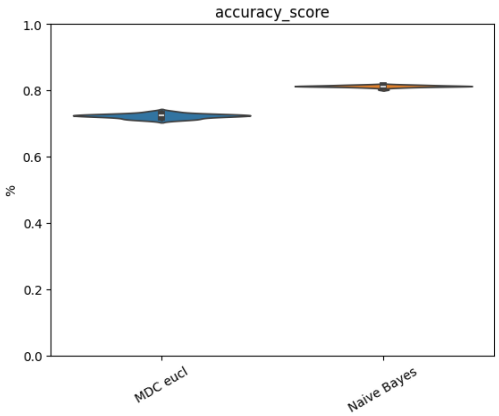


(c) Especificidade dos classificadores

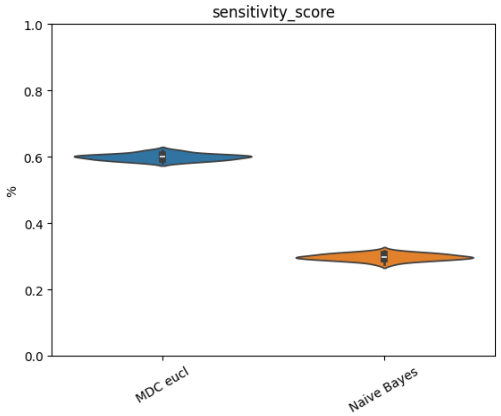
Figura: Métricas de performance dos classificadores com dados cruz

- Classificação feita com o intuito de ser a experiência controlo e verificar a evolução da seleção e redução de *features*

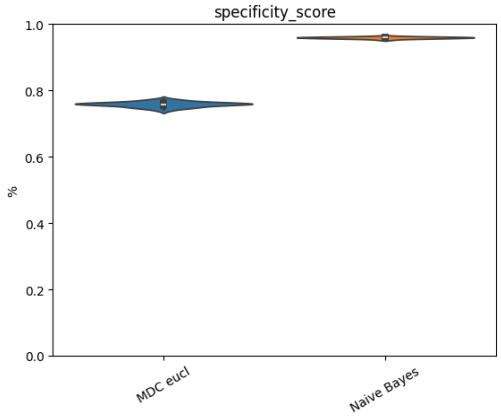
Classificação com LDA



(a) Exatidão dos classificadores



(b) Sensibilidade dos classificadores



(c) Especificidade dos classificadores

Figura: Métricas de performance dos classificadores com LDA

- Fisher LDA tem um ganho especificidade e exatidão, com uma pequena perda de sensibilidade
- Naive Bayes tem uma perda de sensibilidade e um ganho de especificidade

Classificação com PCA

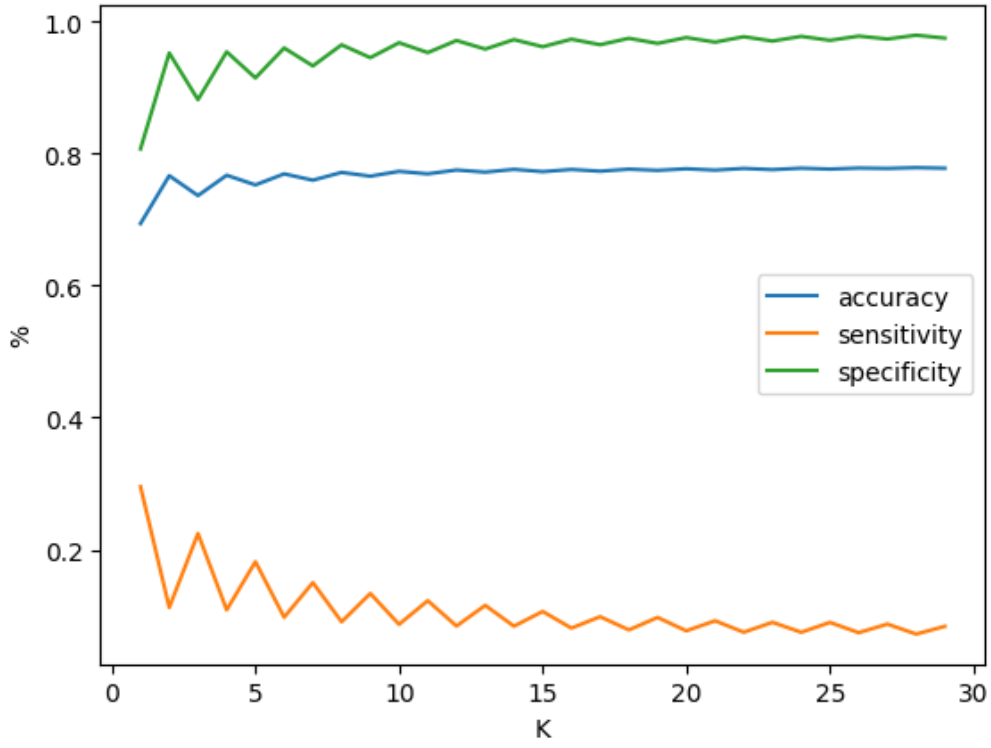


Figura: Estimação do melhor K para o classificador K-NN com PCA

- Resultados indicam dificuldade em manter boa sensibilidade
- Fixação de $K = 1$ para maximizar sensibilidade

Classificação com PCA

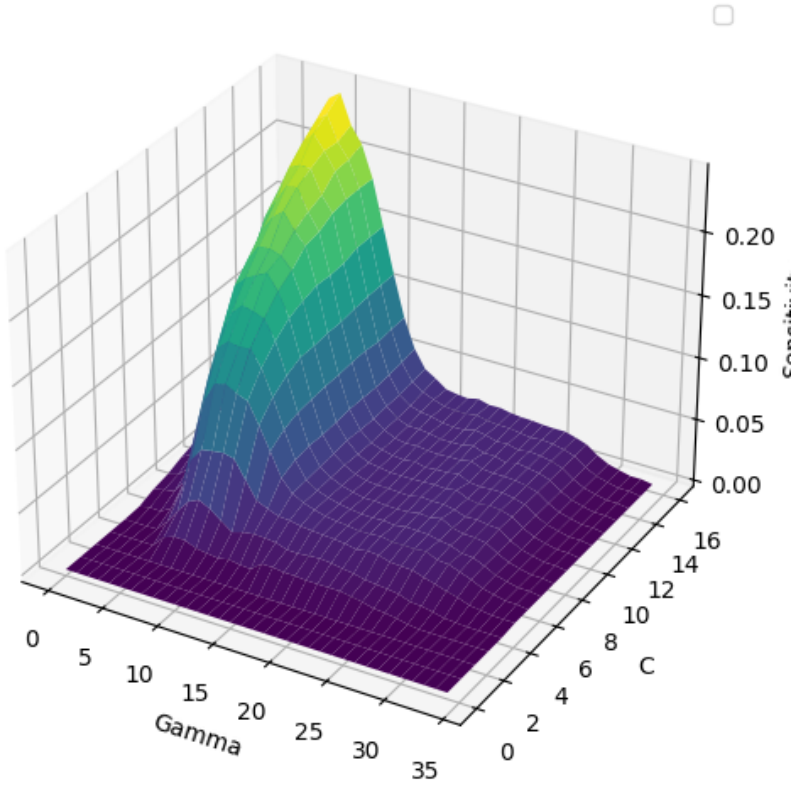
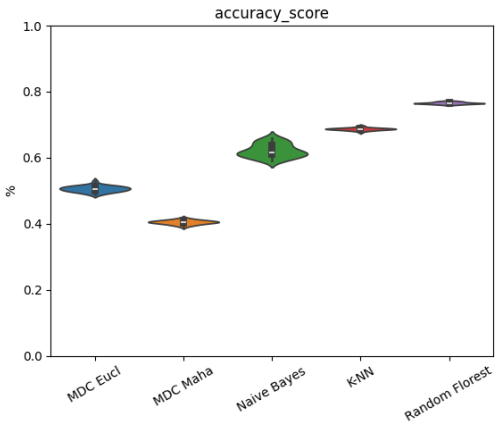


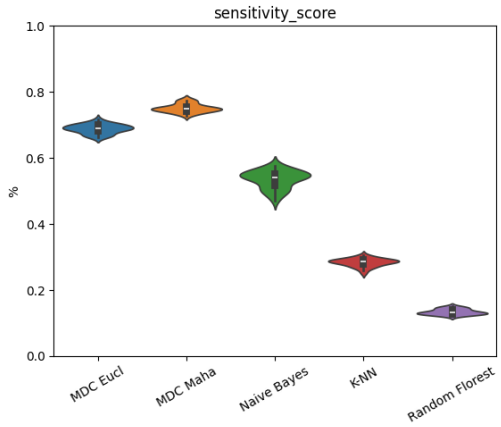
Figura: Estimação do melhor C e γ para a SVM com PCA para a sensibilidade

- Resultados indicam dificuldade em atingir uma boa sensibilidade
- Decidi não usar SVM com PCA

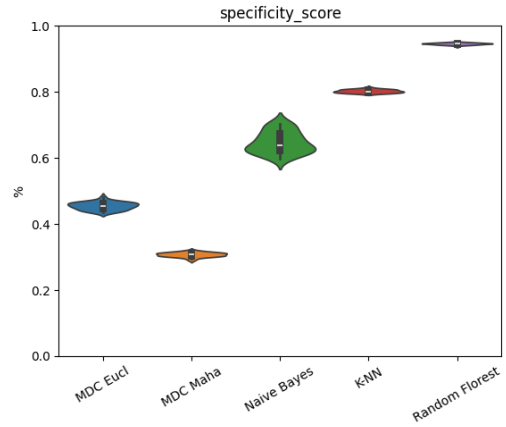
Classificação com PCA



(a) Exatidão dos classificadores



(b) Sensibilidade dos classificadores



(c) Especificidade dos classificadores

Figura: Métricas de performance dos classificadores com PCA

Em relação aos dados cruz

- As MDC ganham mais sensibilidade e perdem especificidade
- Naive Bayes mantém os resultados mas com menor desvio padrão
- K-NN perde sensibilidade e especificidade

Classificação com AUC

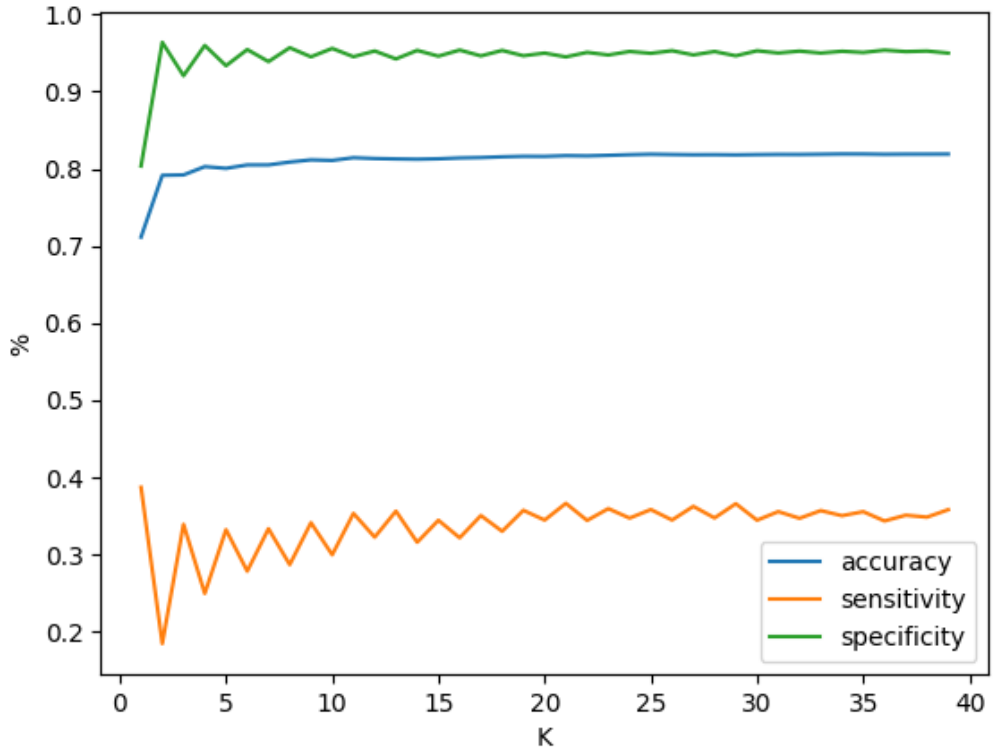


Figura: Estimação do melhor K para o classificador K-NN com AUC

- Fixação de $K = 20$ para melhor equilíbrio entre métricas

Classificação com AUC

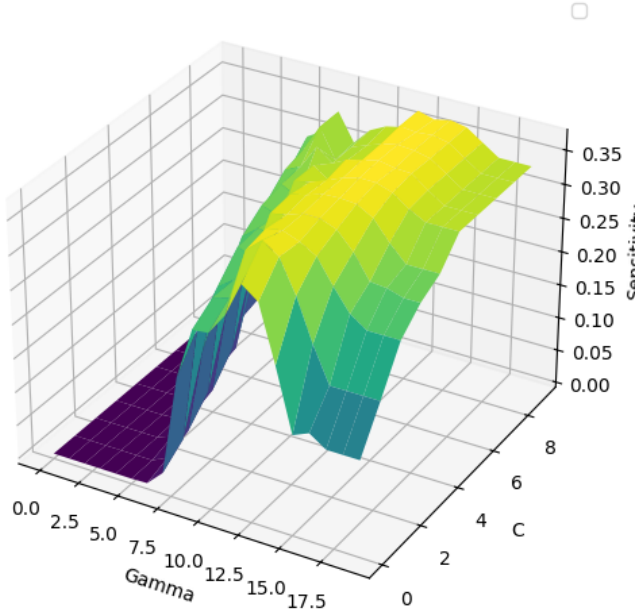
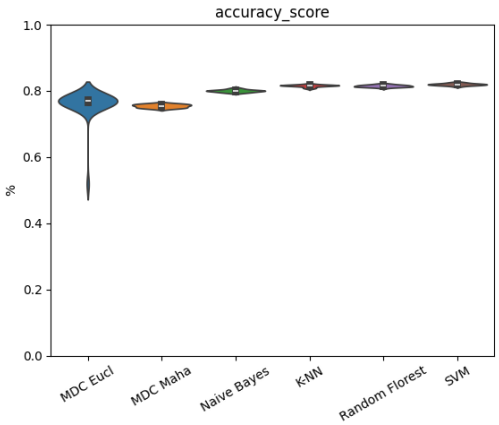


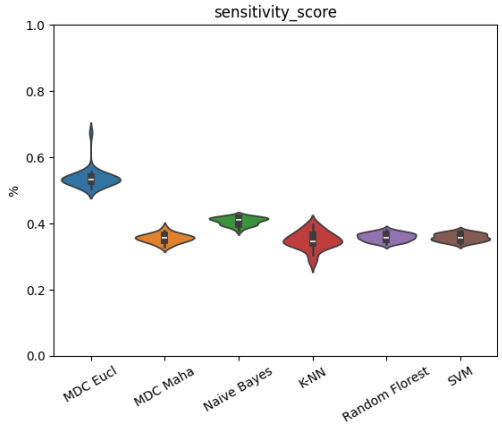
Figura: Estimação do melhor C e γ para a SVM com PCA para a sensibilidade

- Resultados indicam dificuldade em atingir uma boa sensibilidade
- Melhores parâmetros
  - $C = 2$
  - $\gamma = 0.25$

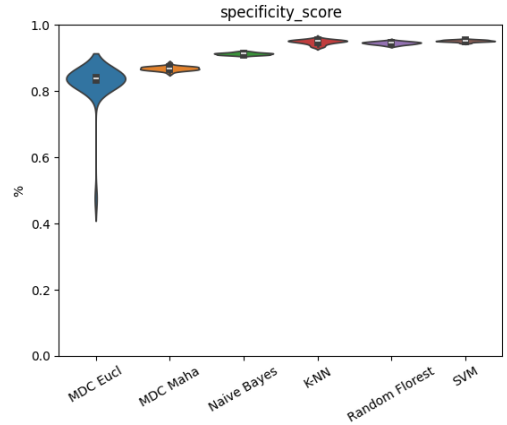
Classificação com AUC



(a) Exatidão dos classificadores



(b) Sensibilidade dos classificadores



(c) Especificidade dos classificadores

Figura: Métricas de performance dos classificadores com AUC

Os classificadores apresentam

- Muita especificidade e exatidão
- Pouca sensibilidade

Classificação com KW

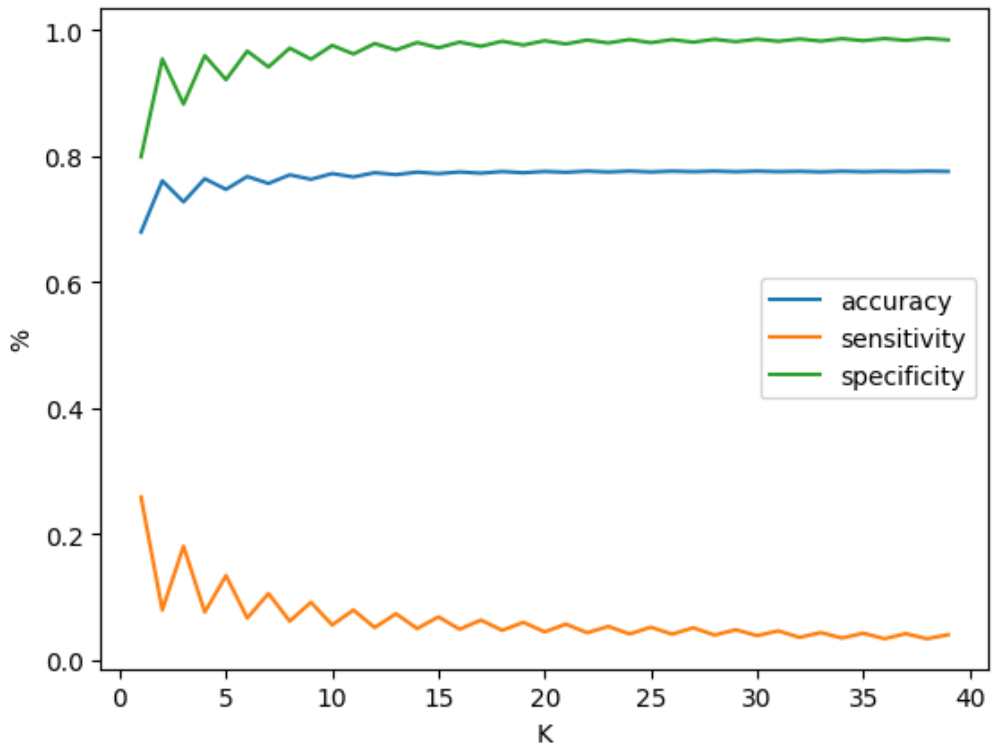


Figura: Estimação do melhor K para o classificador K-NN com KW

- Fixação de $K = 1$ devido a resultados semelhantes ao PCA

Classificação com KW

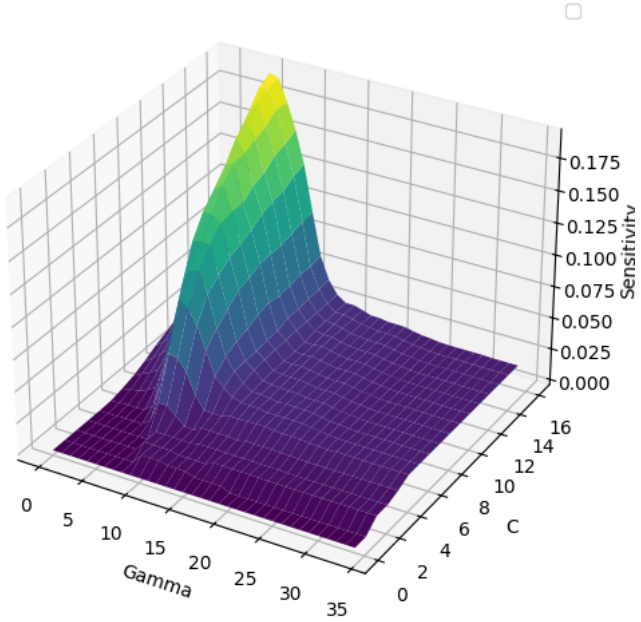
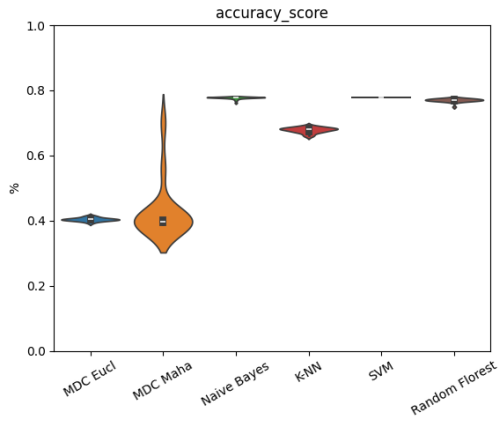


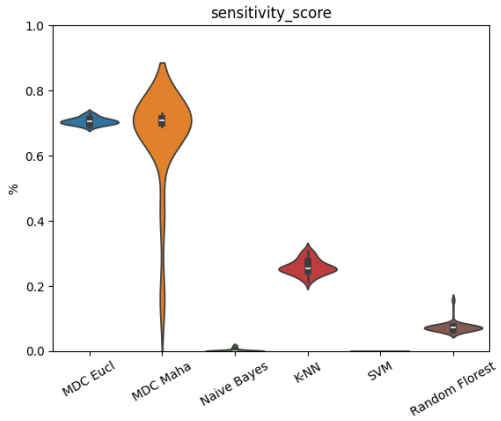
Figura: Estimação do melhor C e γ para a SVM com KW para a sensibilidade

- Resultados indicam dificuldade em atingir uma boa sensibilidade
- Melhores parâmetros
  - $C = 2048$
  - $\gamma = 2.98 \times 10^{-8}$

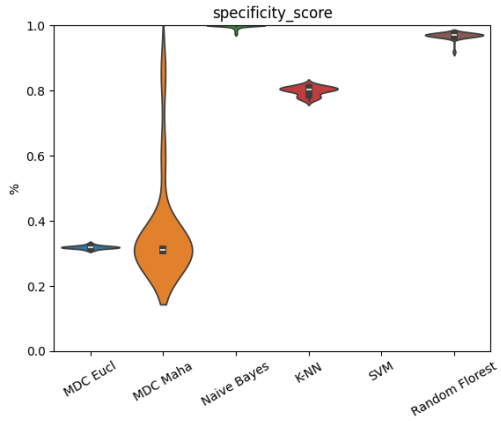
Classificação com KW



(a) Exatidão dos classificadores



(b) Sensibilidade dos classificadores



(c) Especificidade dos classificadores

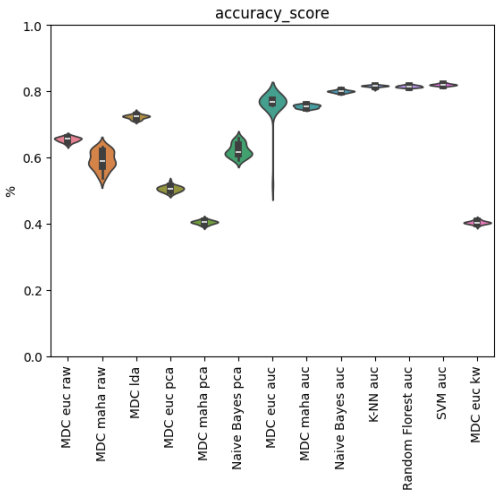
Figura: Métricas de performance dos classificadores com KW

- SVM e Naive Bayes perderam completamente a sensibilidade e ganharam sensibilidade total
- As MDC ganharam alguma sensibilidade mas perderam exatidão e especificidade

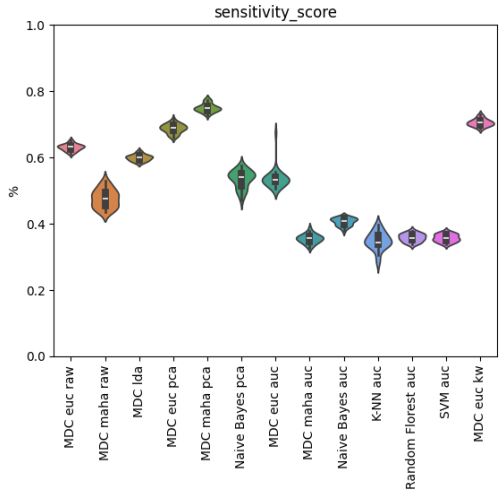
Conclusão

- A escolha da técnica de redução de dimensionalidade e das features é crucial
- Classificadores equilibrados: MDC com distância euclidiana, Naive Bayes com PCA, Fisher LDA
- Classificadores com alta especificidade: MDC com distância euclidiana, K-NN com AUC, Naive Bayes com AUC, Random Forest com AUC, SVM com AUC
- Classificadores com alta sensibilidade: MDC com distância euclidiana com KW, MDC com distância de Mahalanobis com PCA
- A escolha do melhor classificador depende dos objetivos específicos do problema

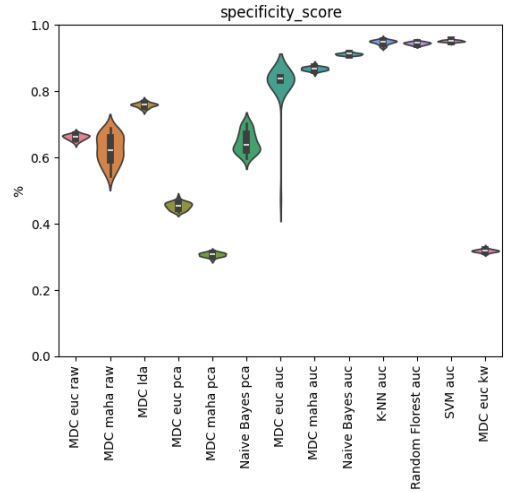
Métricas de Performance



(a) Exatidão dos classificadores



(b) Sensibilidade dos classificadores



(c) Especificidade dos classificadores

Figura: Métricas de performance dos classificadores mais relevantes