Universidade Federal de Pernambuco

Centro de Informática

CIN0208 - Ciência de Dados

Luana Cristina de Carvalho Brito

Análise de Scaling e KNN

Dataset: Spambase (OpenML ID: 44)

Link do Notebook: https://github.com/lccbrito/knn_spambase

Recife, Pernambuco

2025

1. INTRODUÇÃO

1.1 Objetivo da Atividade

Esta atividade tem como objetivo explorar o impacto de diferentes técnicas de normalização e padronização (scaling) no desempenho do algoritmo K-Nearest Neighbors (KNN). O estudo visa compreender como a escala das features influencia algoritmos baseados em distância e identificar qual técnica de scaling proporciona os melhores resultados para o problema em questão.

1.2 Características do Dataset

O dataset escolhido foi o **Spambase**, disponível no repositório OpenML (ID: 44). Este conjunto de dados é amplamente utilizado para estudos de classificação e possui as seguintes características:

- **Número de instâncias:** 4.601 emails
- **Número de features:** 57 variáveis numéricas
- **Variável alvo:** Binária (0 = Não-Spam, 1 = Spam)
- **Distribuição das classes:** Não-Spam: 2.788 instâncias (60,6%) e Spam: 1.813 instâncias (39,4%)

As features representam:

- Frequências de palavras específicas (ex: "make", "address", "free", "money")
- Frequências de caracteres especiais (ex: "!", "$", "#")
- Estatísticas sobre letras maiúsculas (média de sequências, maior sequência, total)

Observação: Todas as 57 features são **numéricas contínuas**, portanto não houve necessidade de aplicar técnicas de encoding (One-Hot, Dummy ou Effect Coding).

1.3 Tipo de Problema

Trata-se de um problema de **classificação binária supervisionada**, onde o objetivo é predizer se um email é spam ou não-spam com base nas características extraídas do seu conteúdo. Este é um problema clássico de filtragem de emails, com aplicações práticas em sistemas de segurança e comunicação digital.

2. ENCODING E SCALING

2.1 Encoding de Variáveis Categóricas

O dataset Spambase não possui colunas nominais ou categóricas. Todas as 57 features são variáveis numéricas contínuas que representam frequências e estatísticas extraídas dos emails. Portanto, não foi necessário aplicar nenhuma técnica de encoding.

Esta característica do dataset permitiu focar exclusivamente no impacto das técnicas de scaling no desempenho do KNN.

2.2 Técnicas de Scaling Aplicadas

1. **Sem Scaling (Baseline)**: Dados brutos sem qualquer transformação.
2. **StandardScaler**: Padroniza features para média = 0 e desvio padrão = 1.
3. **MinMaxScaler**: Normaliza features para o intervalo [0, 1].
4. **MaxAbsScaler**: Escala pelo valor absoluto máximo, resultando em intervalo [-1, 1].
5. **RobustScaler**: Usa mediana e intervalo interquartil (IQR), robusto a outliers.
6. **QuantileTransformer (uniform)**: Transforma features para distribuição uniforme.
7. **QuantileTransformer (normal)**: Transforma features para seguir distribuição normal.

2.3 Resultados do KNN

Tabela 1: Acurácia do KNN por Técnica de Scaling

Dataset	Encoding	Scaling	Acurácia KNN
Spambase	N/A	Sem Scaling	0.79
Spambase	N/A	StandardScaler	0.90
Spambase	N/A	MinMaxScaler	0.90
Spambase	N/A	MaxAbsScaler	0.90
Spambase	N/A	RobustScaler	0.90
Spambase	N/A	QuantileTransformer (uniforme)	**0.94**
Spambase	N/A	QuantileTransformer (normal)	0.93

Tabela 2: F1-Score do KNN por Técnica de Scaling

Dataset	Encoding	Scaling	F1-Score KNN
Spambase	N/A	Sem Scaling	0.73
Spambase	N/A	StandardScaler	0.87
Spambase	N/A	MinMaxScaler	0.87
Spambase	N/A	MaxAbsScaler	0.87
Spambase	N/A	RobustScaler	0.88
Spambase	N/A	QuantileTransformer (uniforme)	**0.92**
Spambase	N/A	QuantileTransformer (normal)	0.91

2.4 Gráficos Comparativos

Os gráficos gerados (Figura 1) apresentam quatro análises complementares:

1. **Comparação de Acurácia entre Scalers:** Mostra claramente que o QuantileTransformer (uniforme) obteve a maior acurácia (93,99%), enquanto o modelo sem scaling teve apenas 79,15%
2. **Comparação de F1-Score entre Scalers:** Confirma a superioridade do QuantileTransformer (uniforme) também no F1-Score (92,34%)
3. **Comparação Conjunta**: Permite visualizar lado a lado ambas as métricas, evidenciando a consistência do QuantileTransformer
4. **Ganho de Performance em relação ao Baseline:** Demonstra que o QuantileTransformer proporcionou ganho de +14,8% na acurácia e +18,9% no F1-Score em relação ao modelo sem scaling

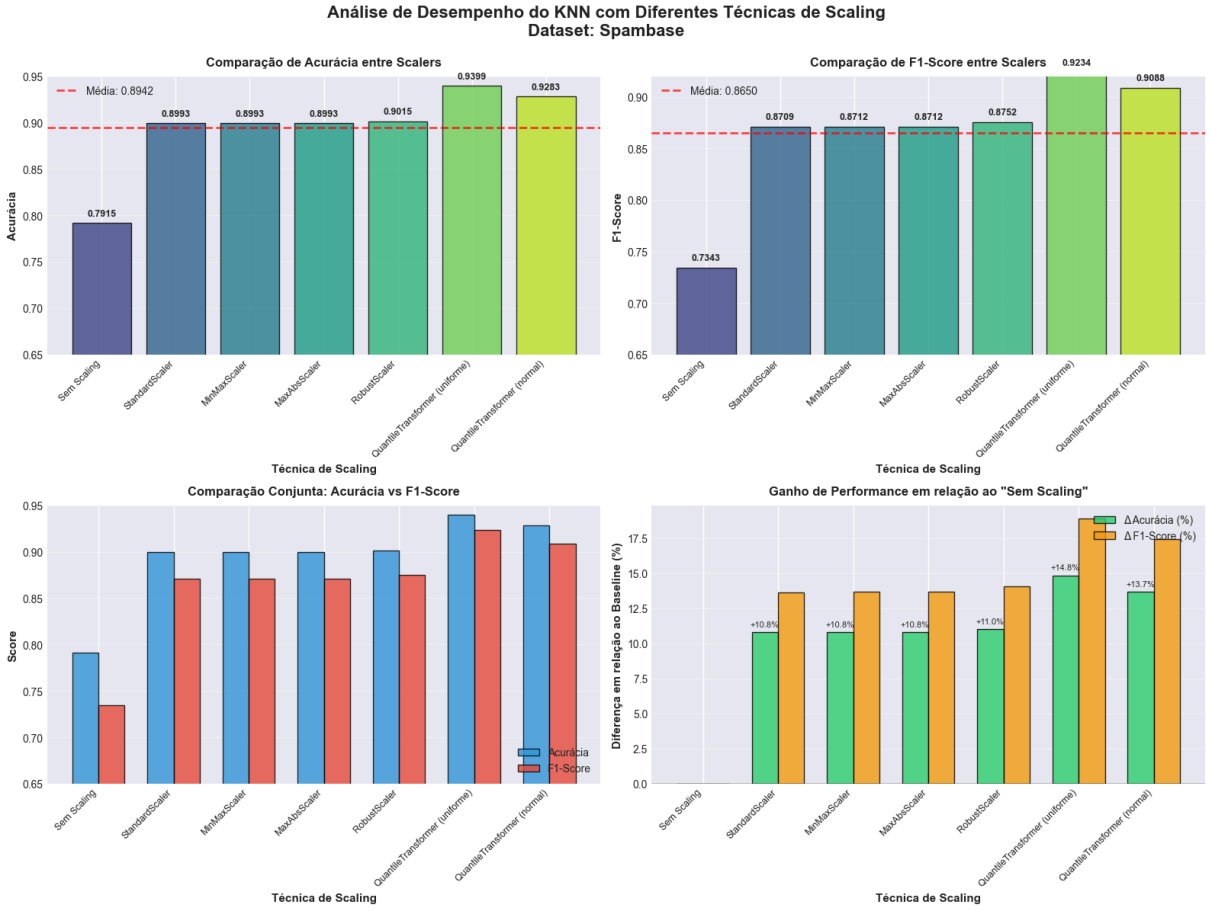

Figura 1: Comparação de desempenho do KNN com diferentes técnicas de scaling

2.5 Análise do Melhor Modelo

O modelo com **QuantileTransformer (uniforme)** apresentou os melhores resultados:

Métricas de Performance: Acurácia: 93,99%; F1-Score: 92,34%; Precision (Não-Spam): 95%, Recall (Não-Spam): 95%; Precision (Spam): 93% e Recall (Spam): 92%

Matriz de Confusão: Apenas 6% dos emails foram classificados incorretamente (83 de 1.381).

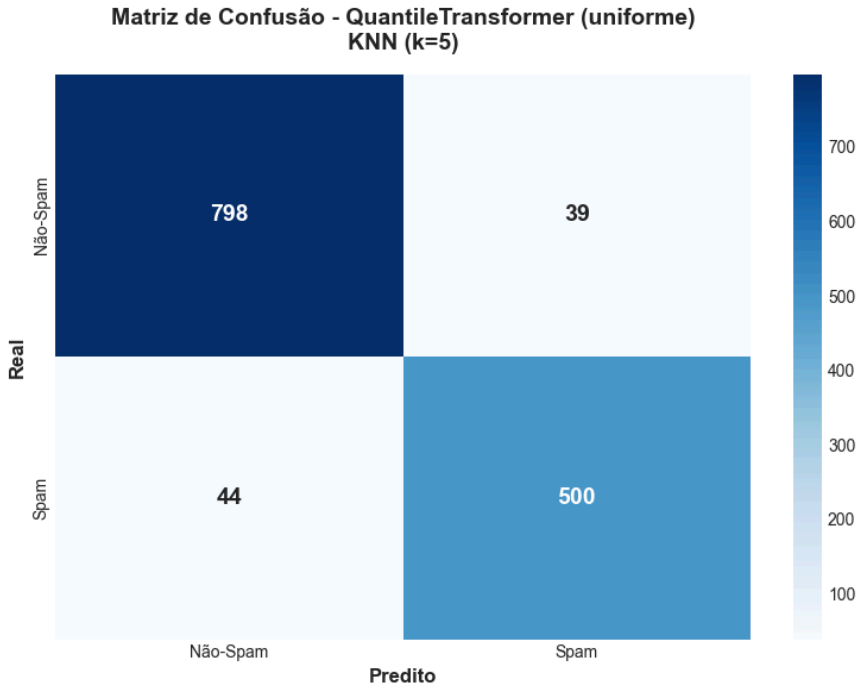

Figura 2: Matriz de confusão do melhor modelo (QuantileTransformer uniforme)

2.6 Discussão e Observações

2.6.1 Por que o Scaling é Crítico para o KNN?

O algoritmo KNN utiliza distância euclidiana para determinar a similaridade entre instâncias. Quando as features têm escalas muito diferentes, aquelas com valores maiores dominam o cálculo da distância, distorcendo os resultados.

2.6.2 Comparação entre os Scalers:

QuantileTransformer (uniforme) - MELHOR RESULTADO (93,99%)

- **Vantagem:** Transforma dados para distribuição uniforme, lidando perfeitamente com distribuições assimétricas e outliers
- **Por que funcionou melhor:** O Spambase possui features com distribuições muito assimétricas (muitos zeros e valores extremos)

StandardScaler, MinMaxScaler, MaxAbsScaler (89,93%)

- Resultados praticamente idênticos (~90%)
- Bom desempenho geral, mas sensíveis a outliers
- Assumem que os dados seguem distribuições razoavelmente normais

RobustScaler (90,15%)

- Ligeiramente superior aos anteriores
- Usa mediana e IQR, mais robusto a outliers
- Boa escolha quando há valores extremos conhecidos

QuantileTransformer (normal) (92,83%)

- Excelente resultado, mas inferior ao uniforme
- Transforma para distribuição gaussiana

- Útil quando algoritmos assumem normalidade

Sem Scaling (79,15%) - BASELINE

- Performance significativamente inferior
- Demonstra claramente a importância do pré-processamento
- Features com maior magnitude dominaram as predições

3. CONCLUSÃO

3.1 Resumo dos Principais Achados

- A diferença entre usar e não usar scaling foi de 14,8 pontos percentuais na acurácia (79,15% → 93,99%)
- Tanto acurácia quanto F1-Score mostram o mesmo padrão, indicando que os resultados são consistentes
- O F1-Score de 92,34% indica bom equilíbrio entre precision e recall, importante para problemas de spam onde tanto falsos positivos quanto falsos negativos são custosos
- O QuantileTransformer com distribuição uniforme superou todas as outras técnicas, alcançando: 93,99% de acurácia, 92,34% de F1-Score e apenas 6% de taxa de erro.
- **Ranking de performance:**
    - 1º QuantileTransformer (uniforme) - 93,99%
    - 2º QuantileTransformer (normal) - 92,83%
    - 3º RobustScaler - 90,15%
    - 4º StandardScaler/MinMaxScaler/MaxAbsScaler - 89,93%
    - 5º Sem Scaling - 79,15%
- **Encoding não aplicável:** O dataset Spambase não possui variáveis categóricas, evidenciando que nem todos os datasets requerem encoding

3.2 Lições Aprendidas

- Sobre Encoding:
    - Nem todo problema de ciência de dados requer encoding
    - É fundamental analisar o tipo de dados antes de aplicar transformações
    - Datasets com features numéricas (como frequências) dispensam encoding
- Sobre Scaling:
    - Essencial para algoritmos baseados em distância (KNN, SVM, clustering)
    - A escolha do scaler deve considerar a distribuição dos dados.
- Sobre o KNN:
    - É extremamente sensível à escala das features
    - A performance pode variar drasticamente (até 15 pontos percentuais) dependendo do scaling
    - Ideal para datasets de tamanho moderado (o Spambase tem ~4.600 instâncias)