Classificação de Dígitos MNIST com k-Nearest Neighbors e Classificadores Lineares

Eduardo Camargo Neves - GRR20196066

Julho de 2025

Abstract

Este relatório detalha uma série de experimentos realizados sobre o dataset MNIST para a tarefa de classificação de dígitos manuscritos. O código-fonte e todos os dados de experimentação estão disponíveis no repositório do GitHub. Foram empregados e comparados dois tipos de classificadores: k-Nearest Neighbors (kNN), com uma análise de múltiplas métricas de distância e pesos, e modelos lineares (Regressão Logística e SVM Linear). Adicionalmente, foi investigado o impacto da redução de dimensionalidade através da Análise de Componentes Principais (PCA) com diferentes níveis de variância retida. O desempenho dos modelos foi avaliado sistematicamente, culminando na identificação de uma configuração ótima que alcançou uma acurácia final de 97,54% no conjunto de teste.

1 Introdução

O reconhecimento de padrões em imagens é uma das áreas fundamentais da visão computacional, e o dataset MNIST representa um dos benchmarks mais clássicos e difundidos para essa tarefa. Composto por 70.000 imagens em escala de cinza de dígitos manuscritos (0 a 9), de dimensões 28x28 pixels, ele serve como um excelente ambiente para testar e validar algoritmos de classificação.

Neste trabalho, realizamos uma exploração sistemática com os seguintes objetivos:

- Comparar o desempenho do classificador por proximidade k-Nearest Neighbors (kNN) com um classificador linear (SGDClassifier).
- Avaliar o impacto de diferentes métricas de distância (Euclidean, Manhattan, Chebyshev) e da ponderação dos vizinhos (Uniform vs. Distance) no desempenho do kNN.
- Investigar o efeito da redução de dimensionalidade com a técnica PCA, analisando a relação entre a complexidade do modelo (número de features) e sua performance em acurácia e tempo de execução.
- Identificar a combinação de modelo, parâmetros e pré-processamento que oferece o melhor resultado para esta tarefa.

2 Metodologia

Os experimentos foram conduzidos seguindo um pipeline estruturado de pré-processamento, modelagem e avaliação.

2.1 Pré-processamento e Divisão dos Dados

As imagens do dataset MNIST foram inicialmente normalizadas, com os valores de seus pixels sendo reescalados do intervalo [0, 255] para [0, 1]. Subsequentemente, cada imagem 28x28 foi achatada (flattened) para um vetor de 784 features.

O conjunto de dados completo, com 70.000 amostras, foi dividido em três subconjuntos de forma estratificada:

- **Treinamento:** 49.000 amostras (70%)
- **Validação:** 10.500 amostras (15%)
- **Teste:** 10.500 amostras (15%)

O conjunto de validação foi utilizado para o ajuste de hiperparâmetros, enquanto o conjunto de teste foi reservado para a avaliação final do modelo campeão.

2.2 Redução de Dimensionalidade

A técnica de Análise de Componentes Principais (PCA) foi aplicada para reduzir a dimensionalidade do vetor de 784 features. Foram testadas duas configurações principais, buscando reter:

- 95% da variância original, o que resultou em 154 componentes.
- 90% da variância original, o que resultou em 87 componentes.

2.3 Modelos e Hiperparâmetros

Foram explorados dois tipos de classificadores, com uma gama de hiperparâmetros:

- **k-Nearest Neighbors (kNN):**
  - *Número de vizinhos (k):* {1, 3, 5, 9, 13, 15, 21}
  - *Métrica de distância:* {'euclidean', 'manhattan', 'chebyshev'}
  - *Peso dos vizinhos:* {'uniform', 'distance'}
- **Classificador Linear (SGDClassifier):**
  - *Função de custo (loss):* {'hinge' (SVM Linear), 'log_loss' (Regressão Logística)}
  - *Tipo de penalidade:* {'l2', 'l1', 'elasticnet'}
  - *Fator de regularização (alpha):* {0.0001, 0.001, 0.01, 0.1}

3 Resultados

A performance das diferentes abordagens foi registrada e é apresentada visualmente para permitir uma análise comparativa clara.

3.1 Análise de Acurácia e Desempenho Computacional

Para avaliar o trade-off entre precisão e custo computacional, a acurácia máxima na validação e a duração média dos treinos foram comparadas entre todas as abordagens.

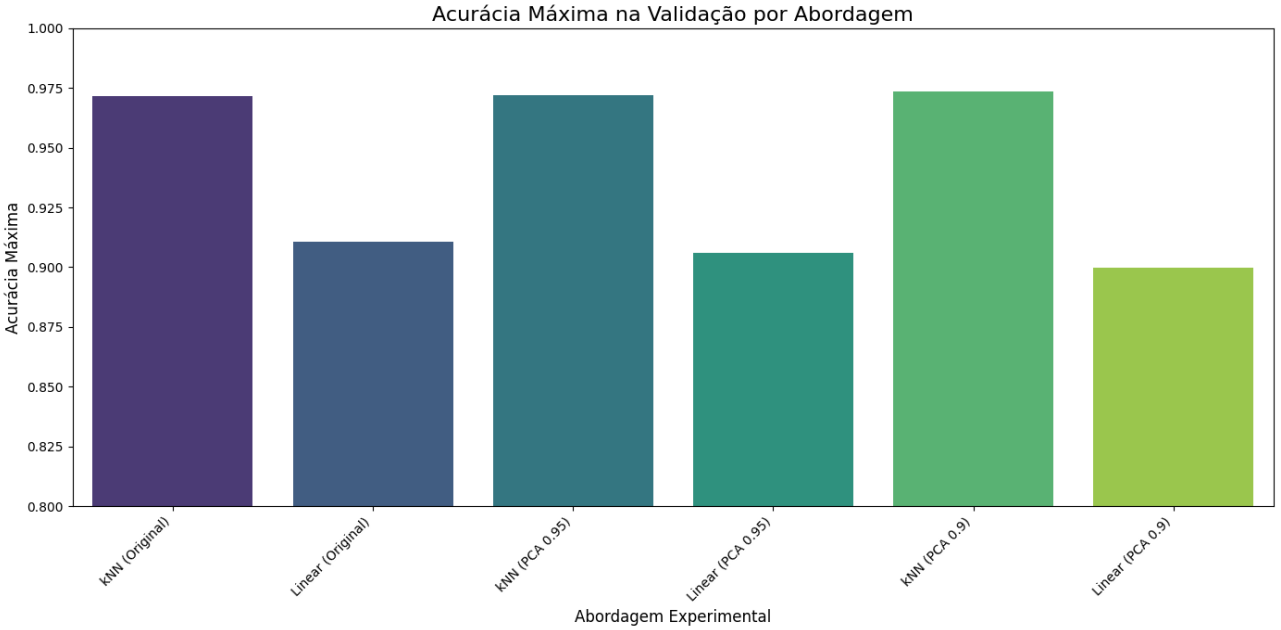

Figure 1: Comparação da acurácia máxima obtida no conjunto de validação para cada abordagem experimental. Observa-se que as abordagens com kNN, tanto no dataset original quanto nos reduzidos por PCA, alcançaram as maiores acurácias.

A Figura 1 demonstra que o modelo kNN consistentemente superou o modelo Linear em termos de acurácia máxima. Notavelmente, a aplicação do PCA não degradou a performance do kNN, mantendo uma acurácia

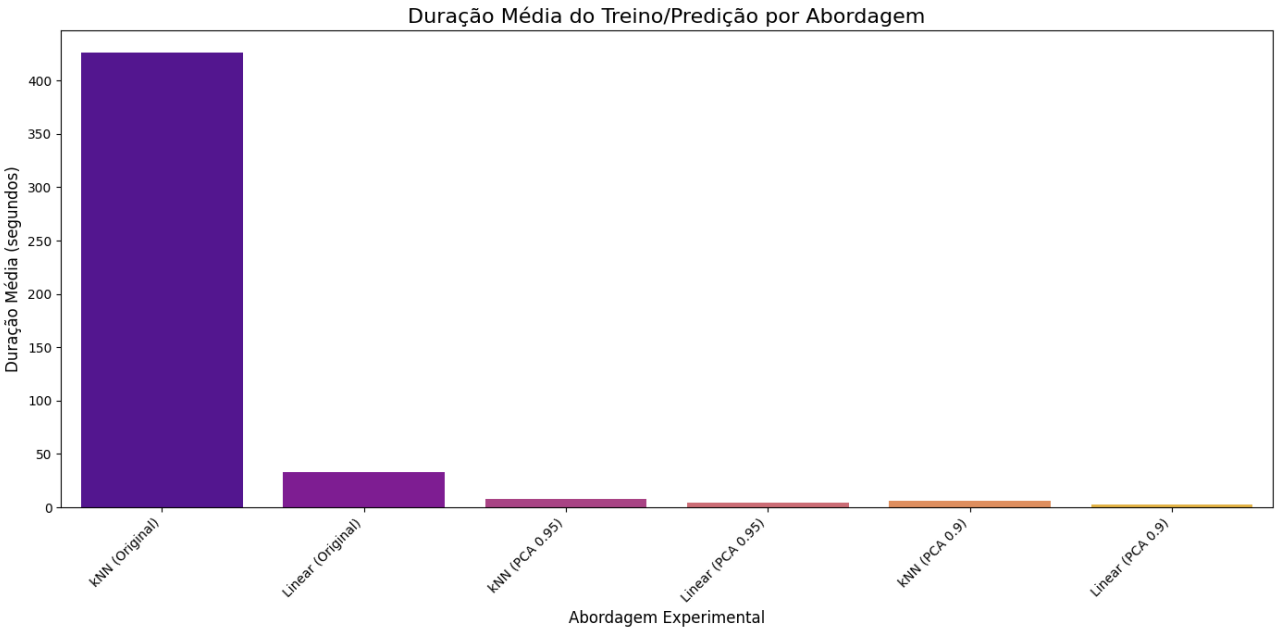

Figure 2: Comparação da duração média de treino e predição (em segundos) para cada abordagem. A diferença de escala entre o kNN no dataset original e as demais abordagens é proeminente.

superior a 97%. Por outro lado, o modelo Linear apresentou uma queda de performance mais acentuada com a redução de dimensionalidade.

Em contrapartida, a análise de desempenho computacional (Figura 2) revela o alto custo do kNN quando aplicado sobre o dataset original de 784 dimensões, com uma duração média superior a 400 segundos. A aplicação do PCA resultou em uma redução drástica neste tempo, tornando o modelo computacionalmente viável e muito mais rápido que o modelo Linear no dataset original.

3.2 Seleção do Modelo e Avaliação Final

Com base nos experimentos de validação, a configuração que alcançou a maior acurácia foi o kNN aplicado sobre os dados reduzidos pelo PCA para 90% da variância (87 features), atingindo 97,35% de acurácia na validação. Este foi selecionado como o modelo campeão.

O modelo foi então retreinado utilizando os dados de treino e validação combinados, e sua performance final foi medida no conjunto de teste. A acurácia final obtida foi de 97,54%. A matriz de confusão (Figura 3) detalha o desempenho do modelo para cada classe.

A análise da matriz de confusão indica uma performance excelente, com a maioria das predições concentradas na diagonal principal. Os erros mais notáveis, embora poucos, ocorrem entre dígitos com caligrafia similar, como a confusão entre os dígitos 4 e 9 (23 erros) e os dígitos 8 e 5 (17 erros).

4 Conclusão

Este trabalho demonstrou uma exploração eficaz de modelos de classificação para o dataset MNIST. Os experimentos confirmaram a alta performance do kNN para esta tarefa, alcançando acurácias superiores a 97%. Foi evidenciado, no entanto, seu alto custo computacional em datasets de alta dimensionalidade.

A aplicação da técnica de redução de dimensionalidade PCA provou ser extremamente benéfica, não apenas reduzindo o tempo de execução do kNN em mais de 95%, mas também mantendo e até mesmo melhorando levemente sua acurácia. A configuração campeã, kNN com k=3 e dados reduzidos para 87 dimensões pelo PCA, alcançou uma robusta acurácia final de 97,54%, provando ser a abordagem com o melhor custo-benefício entre as testadas.

Os resultados reforçam a importância da etapa de pré-processamento e seleção de features, mostrando que é possível obter modelos de alta performance e computacionalmente eficientes através da combinação inteligente de técnicas.

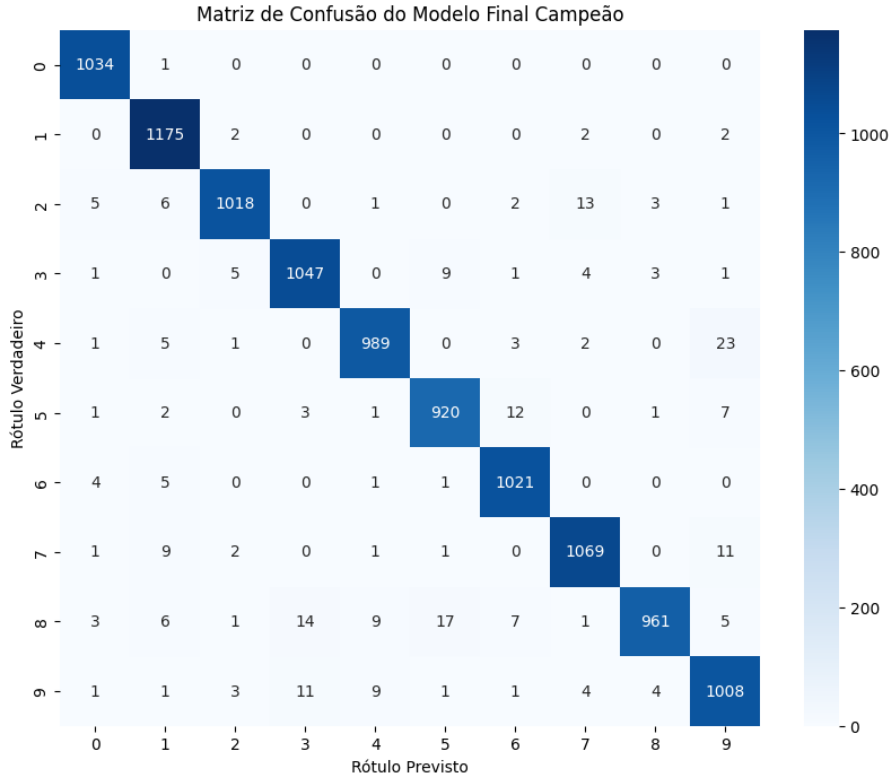

Figure 3: Matriz de confusão do modelo campeão no conjunto de teste. Os valores na diagonal principal representam as classificações corretas.