Fundamentos de Aprendizado de Máquina

Palavras-Chave: Aprendizado de Máquina , Regressão Linear, Classificação de Imagens

Autores(as):

Felipe A. Amaral de Barros, RA 220447 – FEEC

Prof. Dr. Romis Attux (orientador), DCA - FEEC

INTRODUÇÃO E OBJETIVOS:

O aprendizado de máquina (ML) é uma subárea da inteligência artificial que tem ganhado destaque significativo devido às suas capacidades de resolver problemas complexos em diversos domínios, como reconhecimento de padrões, previsão de séries temporais, processamento de linguagem natural e muito mais. O ML se baseia na construção de algoritmos que podem aprender e fazer previsões a partir de dados, o que é particularmente útil em contextos em que as relações entre as variáveis de entrada e saída são complexas e não-lineares.

Este projeto de iniciação científica tem como foco principal proporcionar uma formação sólida em aprendizado de máquina, com ênfase em problemas de regressão e classificação. A motivação para este estudo vem da necessidade crescente de profissionais capacitados em técnicas de ML para enfrentar os desafios do mercado de trabalho atual e futuro. Além disso, a aplicação prática dos conceitos teóricos aprendidos é essencial para a compreensão profunda e para o desenvolvimento de habilidades práticas que são altamente valorizadas na indústria e na academia.

Os dois projetos abordados neste estudo são representativos de problemas clássicos no aprendizado de máquina: a predição de séries temporais e a classificação de imagens. O primeiro projeto envolve a previsão de valores futuros de uma série temporal utilizando modelos de regressão linear, especificamente aplicado à série histórica do número de manchas solares. O segundo projeto explora a classificação de dígitos manuscritos usando um modelo de regressão logística e uma rede neural MLP (Multi-Layer Perceptron). Ambos os projetos utilizam datasets abertos amplamente reconhecidos e métodos de validação rigorosos para avaliar o desempenho dos modelos implementados.

A implementação prática dos algoritmos de aprendizado de máquina em Python, utilizando bibliotecas como Scikit-Learn, TensorFlow e Keras, é uma meta central do projeto. Através da prática, espera-se solidificar o conhecimento teórico e desenvolver habilidades de programação que são essenciais para a resolução de problemas reais.

METODOLOGIA:

Os estudos foram baseados em materiais de cursos e livros renomados na área de aprendizado de máquina. Reuniões semanais foram realizadas para discussão dos conceitos teóricos e acompanhamento das implementações práticas. Dois projetos principais foram desenvolvidos durante o período de iniciação científica: a predição de séries temporais usando regressão linear e a classificação de dígitos manuscritos utilizando regressão logística e uma rede neural MLP (Multi-Layer Perceptron).

A realização dos projetos tiveram por base o material do curso de pós-graduação IA048 – Aprendizado de Máquina, que vem sendo ministrado pelo orientador e pelo Prof. Levy Boccato, mas os livros [Bishop, 2006] e [Goodfellow et al., 2016] serviram de apoio. O livro de Géron [Géron, 2023], por sua vez, foi uma referência útil para dar suporte ao aluno em sua caminhada de programação em Python.

RESULTADOS E DISCUSSÃO:

Parte 1 – Predição de séries temporais utilizando regressão linear:

O primeiro projeto abordou a predição de séries temporais utilizando a série histórica do número de manchas solares. Este conjunto de dados contém leituras mensais de 1749 a 2019, totalizando 3252 amostras, como pode ser visualizado pela *Figura 1*, retirada ao plotar a base de dados na orma de um gráfico. A previsão de valores futuros desta série foi realizada utilizando um modelo de regressão linear que mapeia um vetor de entradas, formado por um subconjunto de amostras passadas, para uma estimativa do valor futuro.

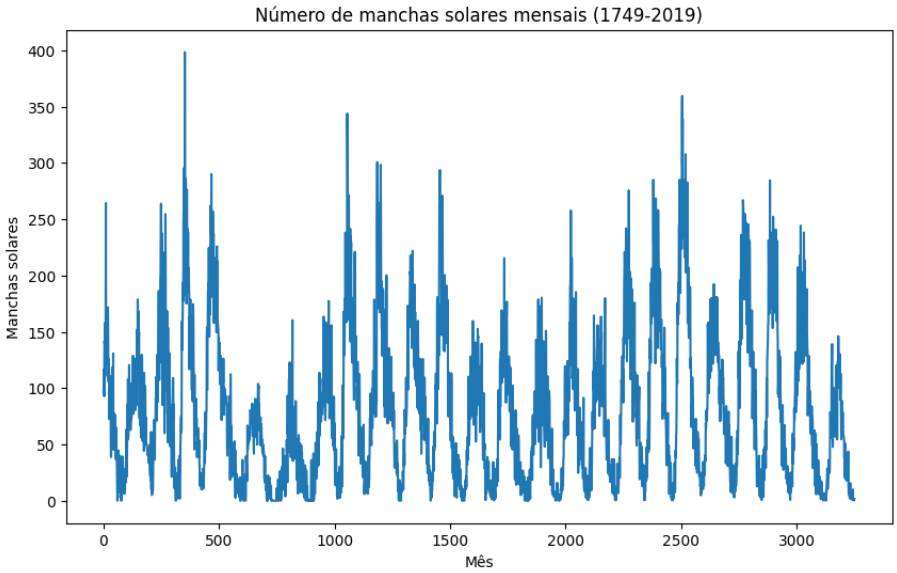

Figura 1: Manchas solares (1749-2019) – Confecção própria

Para desenvolver o modelo, dividimos os dados em conjuntos de treinamento e teste, reservando as amostras dos últimos dez anos (2010-2019) para o teste. O objetivo era garantir que o modelo fosse avaliado em dados não vistos anteriormente, proporcionando uma medida mais precisa de seu desempenho em previsões reais e evitando fenômeno de *Overfitting* estudados no decorrer do projeto.

Além disso, empregamos um esquema de validação cruzada do tipo *k-fold* para selecionar o melhor valor do hiperparâmetro *K*. Este hiperparâmetro representa o número de amostras passadas utilizadas para prever o valor futuro. Variamos *K* de 1 a 24 e calculamos a raiz quadrada do erro quadrático médio (RMSE) para cada valor de *K*. O RMSE é uma medida comum de erro em modelos de regressão, que quantifica a diferença entre os valores previstos pelo modelo e os valores reais.

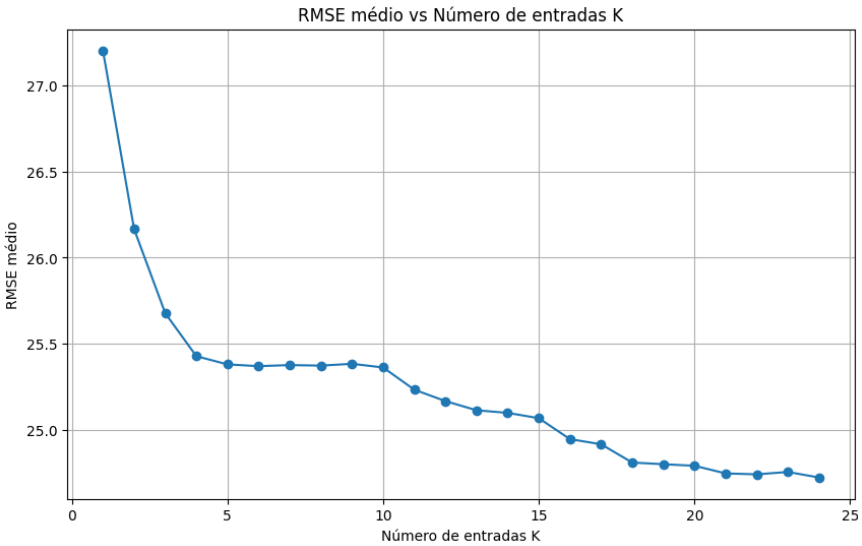

Figura 2: Erro médio quadrado vs número de entradas K – Confecção Própria

Os resultados mostraram a progressão do valor médio da RMSE em função do número de entradas *K* do preditor. A análise indicou que o valor ótimo de $K = 24$ minimiza o RMSE, demonstrando a capacidade do modelo de prever valores futuros da série temporal com boa precisão. A *Figura 2* mostra o gráfico da progressão do RMSE em função de *K*. Além disso, plotamos o gráfico das amostras de teste da série temporal e as respectivas estimativas geradas pelo melhor modelo preditivo (*Figura 3*), evidenciando a eficácia do modelo na predição de valores futuros.

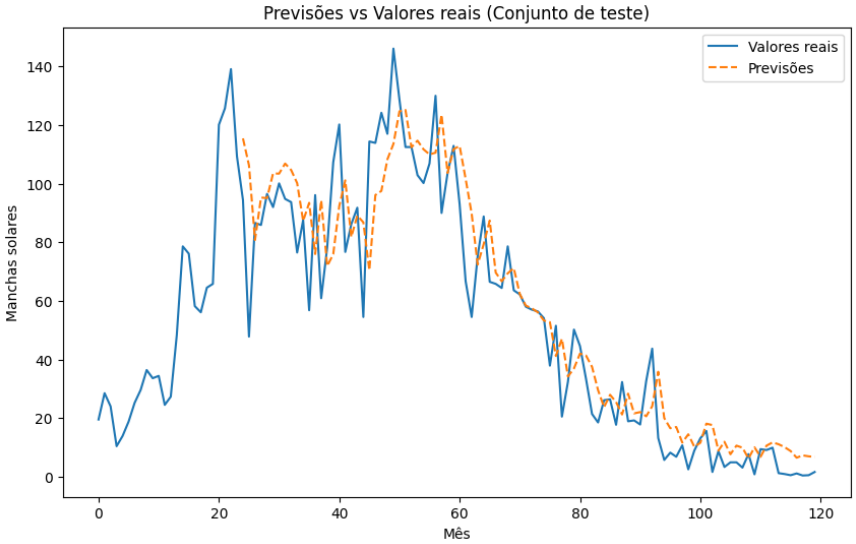

Figura 3: Predição do modelo comparada aos valores reais – Confecção Própria

Parte 2 – Classificação de Dígitos Manuscritos usando Regressão Logística e MLP:

O segundo projeto explorou a classificação de dígitos manuscritos utilizando o dataset MNIST. Este conjunto de dados é amplamente utilizado em estudos de aprendizado de máquina e contém 60.000 imagens de treinamento e 10.000 imagens de teste, cada uma representando um dígito de 0 a 9.

Primeiramente, implementamos um classificador de regressão logística. Este modelo é um método estatístico que, apesar do nome, é utilizado para tarefas de classificação. A regressão logística calcula a probabilidade de uma determinada entrada pertencer a uma classe específica e toma decisões de classificação baseadas nessas probabilidades. Treinamos o modelo com o conjunto de treinamento e avaliamos seu desempenho utilizando o conjunto de teste, onde o classificador obteve uma porcentagem

de acerto de 92,48%. Para visualizar seu desempenho, plotamos a matriz de confusão, que mostra a contagem de predições corretas e incorretas para cada classe, como podemos visualizar na *Figura 4*, bem como as 10 primeiras imagens do dataset e a previsão do algoritmo na *Figura 5*, onde observamos que ele falhou em identificar uma imagem do número 5.

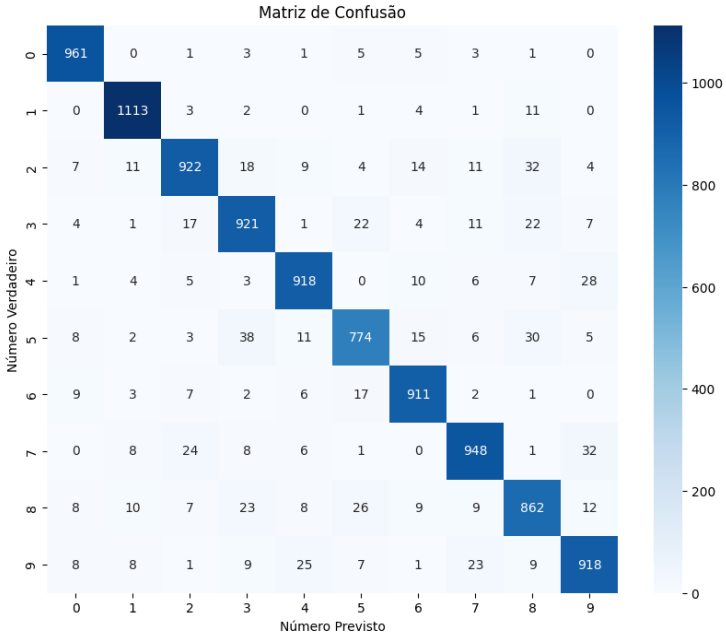

Figura 4: Matriz de Confusão do modelo de Regressão Logística – Confecção Própria

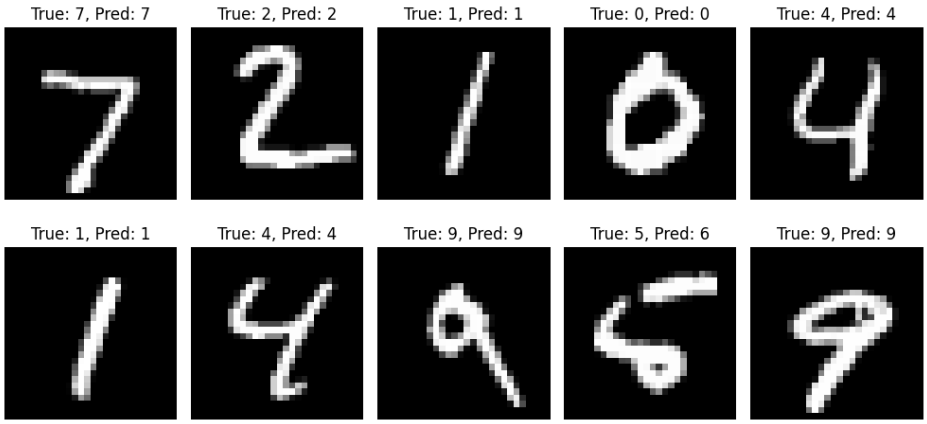

Figura 5: Previsão do algoritmo para as 10 primeiras imagens do dataset – Confecção Própria

Em seguida, implementamos uma rede neural MLP com uma camada oculta. As redes neurais MLP são modelos de aprendizado profundo que podem capturar padrões mais complexos nos dados devido à sua estrutura não linear. Utilizamos a mesma divisão de dados para treino e teste e avaliamos o desempenho do MLP. A arquitetura do MLP foi projetada para ter uma única camada oculta, permitindo uma comparação direta com o modelo de regressão logística em termos de complexidade e desempenho.

Os resultados mostraram que o MLP apresentou melhor desempenho em comparação à regressão logística, onde observamos uma porcentagem de acerto sob o conjunto de teste de 97,51% (Superior aos 92,48% do modelo de regressão logística). A capacidade do MLP de capturar padrões mais complexos nos dados por se tratar de sua estrutura não linear, permitiu um melhor ajuste aos dados de treinamento e, consequentemente, melhores previsões no conjunto de teste. Com isso, a *Figura 6* mostra a matriz de confusão para o MLP, onde observamos menores previsões incorretas que na regressão logística, bem como na *Figura 7*, em que as 10 primeiras imagens do dataset não se realizou nenhum erro de

classificação. Essas visualizações destacam as melhorias no desempenho alcançadas com o uso do MLP, evidenciando a importância de modelos mais complexos para tarefas de classificação quando necessário.

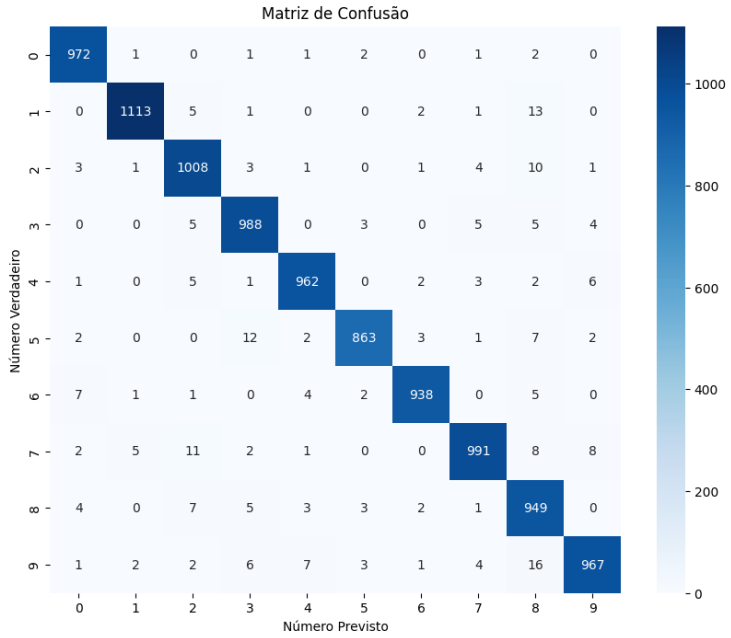

Figura 6: Matriz de Confusão do modelo MLP – Confecção Própria

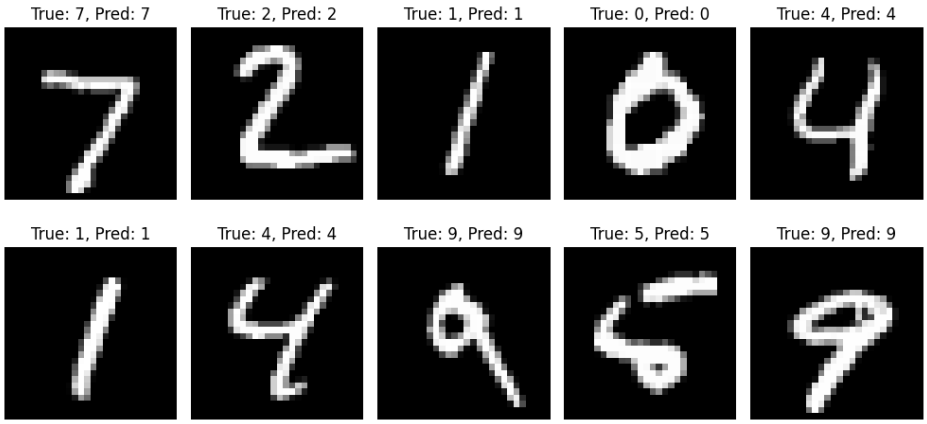

Figura 7: Previsão do algoritmo para as 10 primeiras imagens do dataset – Confecção Própria

CONCLUSÃO:

Os dois projetos demonstraram a eficácia de diferentes abordagens de aprendizado de máquina em problemas canônicos de regressão e classificação. O estudo comparativo entre os modelos reforçou a importância de selecionar o modelo apropriado para cada tipo de problema, considerando a complexidade e a natureza dos dados.

BIBLIOGRAFIA

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Boccato, L., & Attux, R. (2020). Notas de Aula do Curso IA048 – Aprendizado de Máquina, FEEC/UNICAMP.

Géron, A. (2023). Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow. O'Reilly. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521, 436-444.

Principe, J. C. (2010). Information Theoretic Learning: Rényi's Entropy and Kernel Perspectives. Springer.