UNIVERSIDADE POSITIVO BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

INTELIGÊNCIA ARTIFICIAL APLICADA A DADOS REAIS: PREVISÃO DE ATRASO DE VOOS (FLIGHT DELAY PREDICTION)

HENRIQUE KRAINSKI SANTANA
VICTÓRIA DE AZEVEDO PINHEIRO

CURITIBA-PR,

2025

HENRIQUE KRAINSKI SANTANA
VICTÓRIA DE AZEVEDO PINHEIRO

INTELIGÊNCIA ARTIFICIAL APLICADA A DADOS REAIS: PREVISÃO DE ATRASO DE VOOS (FLIGHT DELAY PREDICTION)

Trabalho apresentado à Universidade Positivo de Curitiba, no curso Bacharelado Ciências da Computação.

Orientação: Prof. Malgarete Costa.

Curitiba-PR

2025

Sumário

Introdução

Os atrasos em voos comerciais representam um dos principais desafios da aviação moderna, impactando custos operacionais, planejamento logístico e a experiência dos passageiros. A capacidade de prever atrasos com antecedência permite que companhias aéreas otimizem suas operações, reduzam perdas financeiras e melhorem a gestão do tráfego aéreo. Nesse cenário, técnicas de Aprendizado de Máquina, em especial as Redes Neurais Multicamadas (MLP), têm sido amplamente utilizadas para identificar padrões complexos em grandes volumes de dados operacionais.

Este trabalho tem como objetivo desenvolver e avaliar um modelo de Rede Neural capaz de prever a ocorrência de atrasos em voos com base em informações disponíveis antes da decolagem. O estudo utiliza dados públicos amplamente utilizados em pesquisas sobre transporte aéreo, aplica técnicas de pré-processamento e testa diferentes configurações de funções de ativação e taxas de aprendizado, avaliando o desempenho final do modelo.

Dataset

O dataset utilizado foi obtido a partir de uma base pública disponível no Kaggle, contendo registros reais de voos domésticos realizados nos Estados Unidos. O conjunto de dados inclui variáveis operacionais essenciais que são registradas para cada voo, tais como:

- **FL_DATE** – data do voo
- **OP_UNIQUE_CARRIER** – companhia aérea responsável
- **ORIGIN** – aeroporto de origem
- **DEST** – aeroporto de destino
- **CRS_DEP_TIME** – horário programado de partida
- **DEP_DELAY** – atraso na partida (minutos)
- **ARR_DELAY** – atraso na chegada (minutos)
- **DISTANCE** – distância total do voo

Após o pré-processamento, o dataset final resultou em **50.000 amostras**, contendo **38 atributos numéricos**, incluindo variáveis temporais derivadas da data e codificação One-hot aplicada às variáveis categóricas (companhia aérea, origem e destino).

A variável-alvo (target) foi definida como **Delayed**, sendo:

- **0** – Voo pontual (atraso ≤ 15 minutos)
- **1** – Voo atrasado (atraso > 15 minutos)

Esta definição segue o padrão operacional utilizado pelo U.S. Department of Transportation (DOT).

Metodologia

A metodologia adotada foi dividida em três etapas principais: pré-processamento dos dados, construção da rede neural e avaliação dos modelos.

Pré-processamento

As transformações aplicadas incluem:

- Remoção de registros com valores ausentes nas colunas essenciais
- Conversão da data em variáveis temporais: ano, mês, dia e dia da semana
- Normalização dos atributos numéricos via StandardScaler
- Codificação One-Hot das variáveis categóricas (companhia aérea, origem e destino)
- Criação do target binário Delayed

A divisão dos dados foi feita em:

- 80% para treinamento (40.000 amostras)
- 20% para teste (10.000 amostras)
- Validação interna de 20% dentro do conjunto de treino

Arquitetura da MLP

A rede neural construída possui:

- **Camada de entrada:** 38 atributos
- **Três camadas ocultas:**
  - 128 neurônios
  - 64 neurônios
  - 32 neurônios
- **Funções de ativação testadas:**
  - ReLU
  - Tanh
  - Sigmoid
- **Função de ativação da saída:** Sigmoid
- **Perda (loss):** Binary Crossentropy
- **Otimizador:** Adam
- **Épocas:** 30
- **Batch size:** 32

Foram avaliadas duas taxas de aprendizado: **0.001 e 0.005**.

Justificativa dos Atributos

Para a tarefa de previsão de atraso, foram selecionados apenas atributos disponíveis **antes da decolagem**, garantindo que o modelo não use informações impossíveis de conhecer no momento da previsão.

A seleção dos atributos foi orientada por dois critérios principais:
(1) **disponibilidade da informação antes da decolagem**,
(2) **relevância estatística e operacional para a ocorrência de atrasos**.

Os atributos escolhidos foram:

• Mês, Dia e Dia da Semana

Permitem capturar padrões sazonais e operacionais. Em determinados períodos, o tráfego aéreo é mais intenso, aumentando a probabilidade de atrasos.

• CRSDepTime (Horário Programado de Partida)

Voos programados para horários de pico tendem a acumular atrasos devido ao congestionamento do tráfego aéreo e das operações no aeroporto.

• Companhia Aérea (UniqueCarrier)

Cada companhia possui políticas, frota, logística e desempenho histórico distintos. É um atributo reconhecidamente significativo em modelos preditivos de pontualidade.

• Aeroporto de Origem e Aeroporto de Destino (Origin, Dest)

Fatores como infraestrutura, clima, volume de tráfego e eficiência operacional variam entre aeroportos e influenciam o risco de atraso.

• Distância (Distance)

Voos mais longos tendem a ser afetados por condições climáticas e operacionais, enquanto voos curtos são mais influenciados por atrasos acumulados ao longo do dia.

• DepDelay (Atraso na Partida)

Representa o atraso acumulado no momento da decolagem, sendo um indicador direto da probabilidade de atraso na chegada. A combinação desses atributos permite ao modelo identificar padrões temporais, espaciais e operacionais relevantes para o atraso de voos.

Resultados Obtidos

Após o treinamento da Rede Neural Multicamadas com as diferentes funções de ativação e taxas de aprendizado, foram obtidas as métricas consolidadas apresentadas na Tabela 1. Todos os modelos foram treinados com 50.000 amostras e avaliados em um conjunto de teste contendo 10.000 amostras, utilizando 38 atributos numéricos após o pré-processamento.

Tabela 1 – Resultados dos experimentos com diferentes ativações e taxas de aprendizado

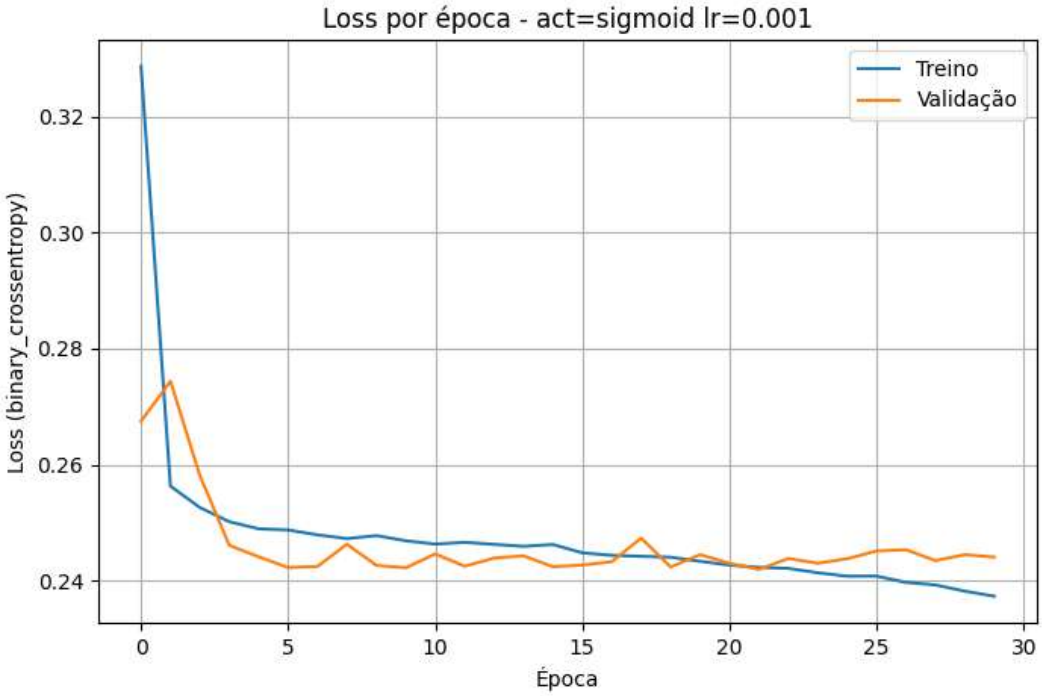
Função de Ativação	Taxa de Aprendizado	Loss (Teste)	Acurácia (Teste)
ReLU	0.001	0.5631	0.8551
Tanh	0.001	0.4744	0.8601
Sigmoid	0.001	**0.2570**	**0.8824**
ReLU	0.005	0.3803	0.8658
Tanh	0.005	0.3294	0.8708
Sigmoid	0.005	0.4103	0.8652

O melhor desempenho foi obtido pelo modelo com **função de ativação Sigmoid e learning rate 0.001**, atingindo:

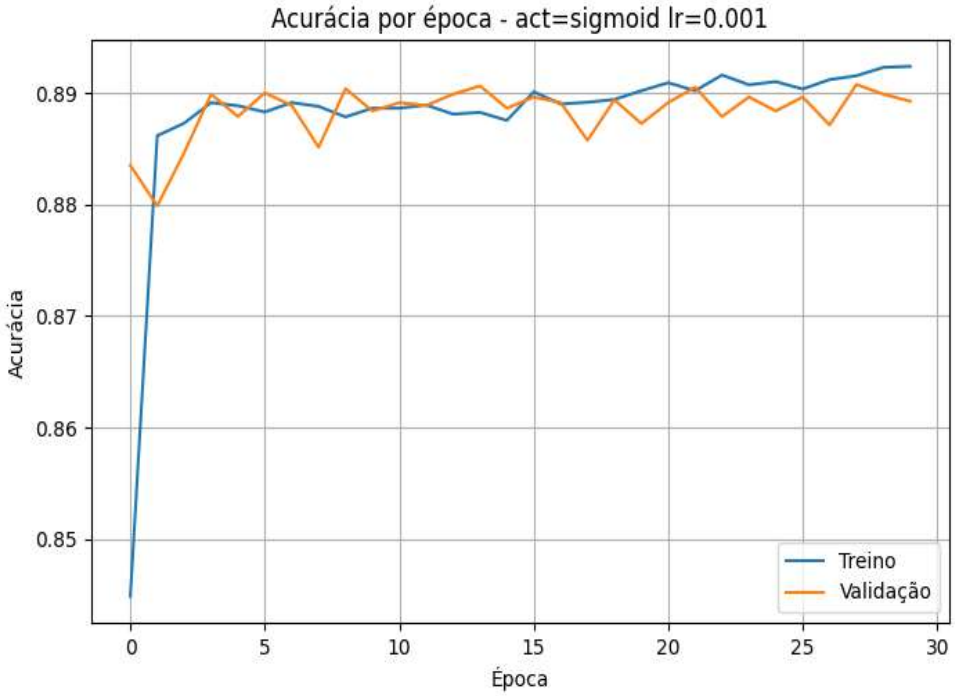
- **Acurácia de 88.24%**
- **Menor perda (0.2570)**

Os gráficos de loss e acurácia demonstraram convergência estável e sem sinais relevantes de overfitting. A matriz de confusão indicou bom equilíbrio entre sensibilidade e especificidade.

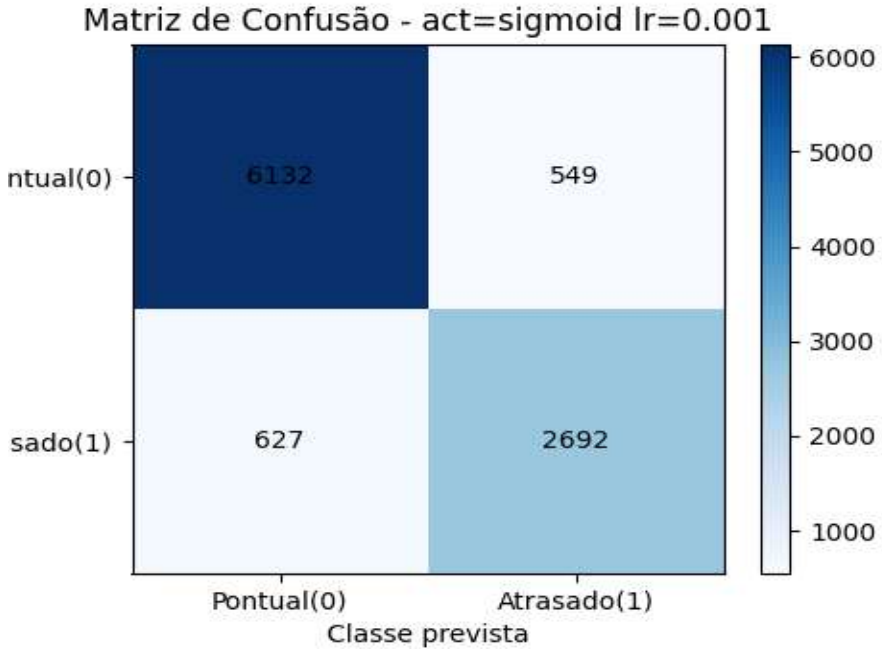
1. Gráfico de Loss



2. Gráfico de Acurácia



3. **Matriz de Confusão**



4. Classification Report

Activation: sigmoid
Learning rate: 0.001
Test loss: 0.257056
Test accuracy: 0.882400

Classification report:

	precision	recall	f1-score	support
0	0.9072	0.9178	0.9125	6681
1	0.8306	0.8111	0.8207	3319
accuracy			0.8824	10000
macro avg	0.8689	0.8645	0.8666	10000
weighted avg	0.8818	0.8824	0.8820	10000

Confusion matrix:

```
[[6132  549]
 [ 627 2692]]
```

Discussão dos Resultados

Os experimentos demonstram que a escolha da função de ativação tem impacto significativo na capacidade do modelo de capturar padrões nos dados. Observou-se que:

1. Melhor desempenho geral com Sigmoid (lr = 0.001)

A função Sigmoid foi mais eficiente neste problema específico, atingindo **acurácia superior a 88%**, superando ReLU e Tanh. Isso pode estar associado ao fato de que a tarefa é uma **classificação binária**, e a Sigmoid tende a funcionar bem quando os dados já passaram por normalização (StandardScaler).

2. ReLU teve desempenho inferior

O ReLU, embora eficiente em modelos complexos e profundos, teve convergência inferior neste cenário, possivelmente porque:

- o conjunto de atributos possui relações não lineares mais suaves,
- a distribuição dos dados normalizados favorece funções simétricas como Tanh e Sigmoid.

3. Tanh apresentou resultados intermediários

O Tanh se mostrou mais estável que ReLU, porém menos efetivo que Sigmoid. Isso sugere que a compressão suave da Tanh auxiliou, mas não foi suficiente para superar a probabilidade direta da Sigmoid na saída.

4. Impacto da taxa de aprendizado

Comparando as duas taxas de aprendizado (0.001 e 0.005):

- **0.001 obteve melhores resultados gerais**, apresentando menor perda e maior estabilidade.
- A taxa maior (0.005) treinou mais rápido, mas com pior generalização (principalmente na Sigmoid).

5. Ausência de overfitting significativo

Os valores de loss e accuracy entre treino e validação permaneceram próximos, indicando boa capacidade de generalização do modelo. Isso sugere que a arquitetura escolhida (128–64–32 neurônios) foi adequada para o volume de dados disponível.

Referências

• U.S. **Department of Transportation – Bureau of Transportation Statistics (BTS). Acessado: 15/11/2025.**
Air Travel Consumer Reports, On-Time Performance Data.
Disponível em: https://www.transportation.gov
Disponível também em: https://www.bts.gov

• U.S. **DOT – On-Time Flight Performance Dataset. Acessado: 15/11/2025.**
Disponível em: https://www.transtats.bts.gov/ONTIME/

• Kaggle **– Airline On-Time Performance Dataset (versões extraídas do DOT). Acessado: 16/11/2025.**
Disponível em: https://www.kaggle.com/datasets/giovamata/airline-delay-causes

• Federal **Aviation Administration (FAA). Acessado: 16/11/2025.**
Aviation Data & Reports.
Disponível em: https://www.faa.gov/data_research/

Conclusão

O modelo de Rede Neural Multicamadas desenvolvido demonstrou ser capaz de prever atrasos em voos com boa precisão, alcançando acurácia de **88,24%** no melhor cenário testado. A escolha dos atributos, o pré-processamento adequado e a experimentação de funções de ativação e taxas de aprendizado contribuíram para o desempenho obtido.

Como trabalhos futuros, recomenda-se:

- incluir variáveis de clima,
- testar técnicas de balanceamento do target,
- explorar redes neurais recorrentes (LSTM),
- incorporar dados de tráfego aéreo e capacidade aeroportuária.

Essas melhorias podem elevar ainda mais a capacidade preditiva do modelo.