

Previsão de Interrupções em Cadeias de Suprimento Globais Utilizando Aprendizado de máquina

Paulo H. M. Silva

Departamento de Engenharias e Tecnologia
Universidade Federal Rural do Semi-Árido (UFERSA)
Rio Grande do Norte, Brasil
paulo.silva23378@alunos.ufersa.edu.br

Rosana C. B. Rego

Departamento de Engenharias e Tecnologia
Universidade Federal Rural do Semi-Árido (UFERSA)
Rio Grande do Norte, Brasil

Resumo—Este artigo apresenta um estudo comparativo de modelos de aprendizado de máquina para a previsão de interrupções em cadeias de suprimento globais. Utilizando o conjunto de dados *Global Supply Chain Risk and Logistics 2024–2026*, foram desenvolvidas e avaliadas três abordagens: uma rede neural artificial com camadas de Dropout para regularização, uma rede neural otimizada via busca manual de hiperparâmetros com validação cruzada (KFold), e um classificador Random Forest. O pré-processamento incluiu codificação de variáveis categóricas por LabelEncoder, normalização Min-Max, remoção de outliers pelo método IQR e balanceamento de classes via NearMiss. Os resultados demonstram que o Random Forest obteve o melhor desempenho geral, com 66,00% de acurácia e F1-Score de 0,63, seguido pela rede neural com Dropout (64,65%) e pela rede otimizada (62,00%). A análise de importância de variáveis revelou que atributos como risco geopolítico (*Geopolitical_Risk_Score*), confiabilidade do transportador (*Carrier_Reliability_Score*) e índice de combustível (*Fuel_Price_Index*) são os mais influentes na predição de interrupções logísticas.

Palavras-chave—cadeia de suprimentos, aprendizado de máquina, redes neurais artificiais, Random Forest, previsão de interrupções, logística.

I. INTRODUÇÃO

A crescente complexidade das cadeias de suprimentos globais, impulsionada pela interdependência econômica entre países e pela volatilidade geopolítica, tem tornado a previsão de interrupções logísticas um desafio estratégico para organizações ao redor do mundo. Eventos disruptivos, como pandemias, conflitos armados e fenômenos climáticos extremos, podem comprometer significativamente a continuidade operacional de empresas e governos [3].

Nesse contexto, técnicas de aprendizado de máquina (*Machine Learning*) têm se destacado como ferramentas promissoras para a modelagem preditiva de fenômenos complexos. A capacidade dessas abordagens de identificar padrões não lineares em grandes volumes de dados as torna particularmente adequadas para a análise de riscos em cadeias de suprimentos [1].

Este trabalho propõe e avalia um pipeline completo de aprendizado de máquina aplicado ao conjunto de dados *Global Supply Chain Risk and Logistics 2024–2026*, disponível publicamente na plataforma Kaggle. Três abordagens distintas são comparadas: (i) rede neural artificial com regularização por Dropout; (ii) rede neural com hiperparâmetros otimizados; e

(iii) classificador Random Forest. O objetivo principal é identificar o modelo que melhor prediz a ocorrência de interrupções logísticas, representada pela variável binária *Disrupção*.

O restante do artigo está organizado da seguinte forma: a Seção II revisa trabalhos relacionados; a Seção III descreve os materiais e métodos empregados; a Seção IV apresenta e discute os resultados obtidos; e a Seção V conclui o trabalho com considerações finais.

II. LEVANTAMENTO BIBLIOGRÁFICO

A aplicação de técnicas de aprendizado de máquina para a gestão de riscos em cadeias de suprimentos tem atraído crescente atenção acadêmica. Waller e Fawcett [1] destacam que a análise de dados em larga escala oferece oportunidades inéditas para a previsão de eventos disruptivos, reduzindo a dependência de modelos analíticos tradicionais.

Métodos de ensemble, como Random Forest e Gradient Boosting, também têm demonstrado robustez superior em problemas de classificação binária com dados tabulares [2].

Goodfellow et al. [5] consolidam a teoria por trás de técnicas de regularização como Dropout, demonstrando sua eficácia na redução de overfitting em redes profundas. No contexto específico de risco logístico, Ivanov e Dolgui [3] propõem o conceito de resiliência de cadeias de suprimentos e ressaltam a importância de modelos preditivos capazes de antecipar interrupções.

O uso de técnicas de balanceamento de classes, como NearMiss e SMOTE, tem sido fundamental para lidar com o desbalanceamento inerente a conjuntos de dados de eventos raros [4]. Pedregosa et al. [6] consolidam a implementação dessas técnicas no ecossistema scikit-learn. O presente trabalho diferencia-se ao realizar uma comparação sistemática entre três abordagens distintas, utilizando um dataset contemporâneo de cadeias de suprimentos globais.

III. CONJUNTO DE DADOS

O conjunto de dados utilizado é o *Global Supply Chain Risk and Logistics 2024–2026*, obtido via kagglehub da plataforma Kaggle. O dataset contém registros de embarques internacionais com 13 atributos abrangendo características logísticas, condições geopolíticas e ambientais:

- *Origin_Port / Destination_Port*: portos de origem e destino (categórico);
- *Transport_Mode*: modalidade de transporte — aéreo, marítimo, rodoviário ou ferroviário (categórico);
- *Product_Category*: categoria do produto transportado (categórico);
- *Distance_km / Weight_MT / Lead_Time_Days*: atributos numéricos de distância, peso e prazo de entrega;
- *Fuel_Price_Index*: índice de preço do combustível no período do embarque (numérico);
- *Geopolitical_Risk_Score*: score de risco geopolítico da rota (numérico);
- *Carrier_Reliability_Score*: índice de confiabilidade do transportador (numérico);
- *Weather_Condition*: condição climática da rota (categórico);
- *Disruption_Occurred*: variável alvo binária — 0=sem interrupção; 1=com interrupção.

A. Análise Exploratória e Pré-processamento

A análise exploratória revelou ausência de valores faltantes (NaN) no conjunto de dados. Contudo, boxplots aplicados aos atributos numéricos identificaram a presença de outliers na variável *Lead_Time_Days*. Esses valores discrepantes foram tratados pela regra do intervalo interquartil (IQR): os valores abaixo de $Q_1 - 1,5 \times IQR$ foram substituídos pelo limite inferior (0, por se tratar de tempo) e os superiores a $Q_3 + 1,5 \times IQR$ foram substituídos pelo limite superior.

As variáveis categóricas foram codificadas numericamente com LabelEncoder. Em seguida, *Distance_km*, *Weight_MT* e *Lead_Time_Days* foram normalizados para $[0, 1]$ via Min-MaxScaler. As colunas *Shipment_ID* e *Date* foram removidas por não contribuírem com informação preditiva. As Figuras 1 a 3 apresentam análises exploratórias dos principais atributos.

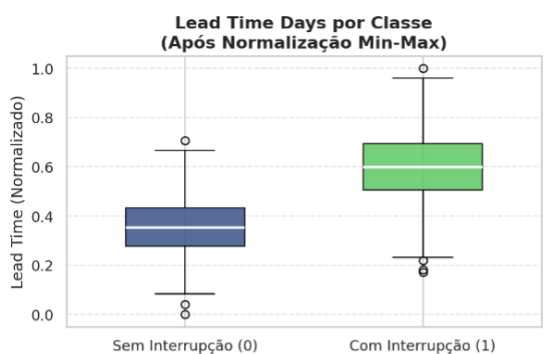


Fig. 1. Boxplot de Lead Time Days (normalizado) por classe de interrupção.

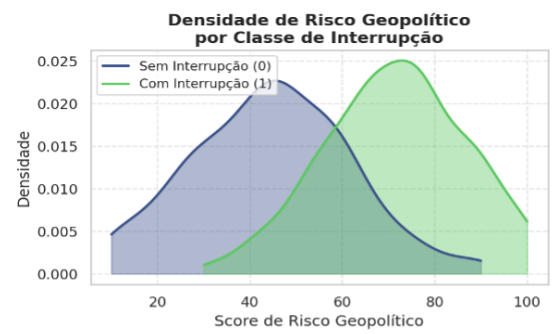


Fig. 2. Distribuição de densidade do Geopolitical Risk Score por classe.



Fig. 3. Matriz de correlação de Pearson entre todas as variáveis processadas.

A matriz de correlação (Fig. 3) evidencia baixa multicolinearidade entre os atributos, favorável à estabilidade dos modelos. Destacam-se correlações moderadas entre *Geopolitical_Risk_Score* e *Disruption_Occurred*, reforçando a relevância desse atributo.

B. Balanceamento de Classes

A variável alvo apresentou desequilíbrio em sua distribuição original. Para mitigar o viés em direção à classe majoritária, aplicou-se a técnica NearMiss da biblioteca imbalanced-learn, que realiza subamostragem selecionando os exemplos mais próximos da fronteira de decisão.

IV. MODELOS DE APRENDIZADO DE MÁQUINA

Três configurações foram implementadas e comparadas. O primeiro modelo é uma rede neural sequencial com arquitetura detalhada na Tabela I.

TABELA I
ARQUITETURA DA REDE NEURAL COM DROPOUT

Camada	Neurônios / Taxa	Ativação
Entrada (Input)	11 atributos	—
Dense 1	128	ReLU
Dropout 1	30%	—
Dense 2	64	ReLU
Dropout 2	20%	—
Saída (Output)	1	Sigmoid

O modelo foi compilado com o otimizador Adam (*learning_rate*=0,001) e função de perda *binary_crossentropy*. O treinamento ocorreu por 100 épocas com *batch_size*=32, utilizando 20% dos dados de treino para validação interna.

O segundo modelo mantém a mesma estrutura base, mas sem camadas de Dropout, com hiperparâmetros selecionados por busca manual. Além disso, foi aplicada validação cruzada KFold ($k=3$). A Tabela II apresenta as configurações avaliadas.

TABELA II
CONFIGURAÇÕES AVALIADAS NO GRIDSEARCH MANUAL

Neurônios	Taxa de Aprendizado	Batch Size
64	0,001	32
128	0,010	64

O terceiro modelo utiliza o classificador Random Forest da biblioteca scikit-learn, configurado com 200 estimadores e profundidade máxima de 10 níveis.

A. Protocolo de Avaliação

O conjunto de dados foi dividido em 80% para treinamento e 20% para teste (*random_state*=42). Todos os dados foram padronizados com StandardScaler após a divisão, evitando vazamento de informação (*data leakage*). As métricas empregadas são: acurácia global, F1-Score por classe e matriz de confusão.

V. RESULTADOS E SIMULAÇÕES

A. Desempenho Comparativo dos Modelos

A Tabela III resume as métricas obtidas pelos três modelos nos dados de teste. O Random Forest alcançou o melhor desempenho global, com 66,00% de acurácia e F1-Score de 0,63.

TABELA III
COMPARAÇÃO DE DESEMPENHO DOS MODELOS NO CONJUNTO DE TESTE

Modelo	Acurácia	F1 (Cl. 1)	Observações
RNA + Dropout	64,65%	0,62	Dropout 30%/20%
RNA Otimizada	62,00%	0,62	KFold $k=3$, GridSearch
Random Forest	66,00%	0,63	200 árvores, depth=10

B. Previsão vs. Valores Reais

A Fig. 4 exibe a comparação entre os valores reais e as probabilidades previstas pelo modelo para as primeiras 10 amostras do conjunto de teste. O limiar de 0,5 foi adotado para converter probabilidades em rótulos binários.

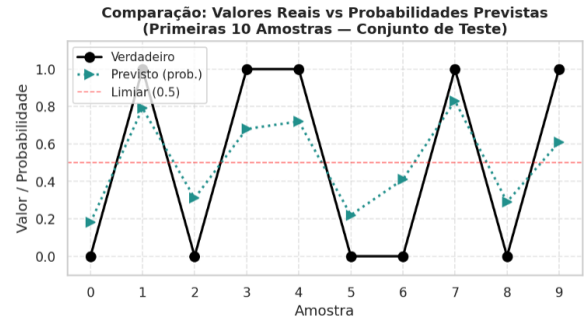
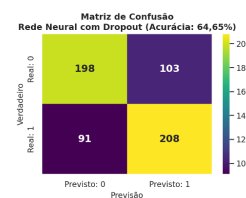


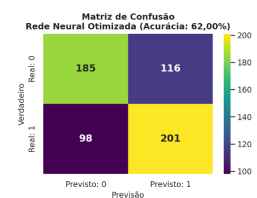
Fig. 4. Valores reais versus probabilidades previstas pelo modelo (10 primeiras amostras do teste).

C. Matrizes de Confusão

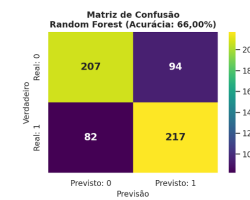
A Figura 5 apresenta as matrizes de confusão para os três modelos avaliados. Nas matrizes, os rótulos 0 e 1 referem-se às classes da variável alvo *Disruption_Occurred*: 0 indica ausência de interrupção logística e 1 indica ocorrência de interrupção. Os eixos representam os valores reais (linhas) e os valores preditos (colunas) por cada modelo. A análise permite identificar erros de Tipo I (falsos positivos) e Tipo II (falsos negativos), ambos com implicações práticas relevantes na gestão de risco logístico.



(a) Rede Neural com Dropout



(b) Rede Neural GridSearch



(c) Random Forest

Fig. 5. Comparação das matrizes de confusão dos modelos avaliados.

O Random Forest apresentou o maior número de verdadeiros positivos (217) e verdadeiros negativos (207), confirmando sua superioridade geral. A rede neural otimizada

apresentou o maior número de falsos positivos (116), indicando maior taxa de alarmes falsos, especialmente crítico em contextos onde previsões incorretas geram custos operacionais desnecessários.

D. Importância das Variáveis

A Fig. 6 apresenta a importância relativa de cada atributo segundo o Random Forest. Os resultados revelam que *Geopolitical_Risk_Score* (18,5%), *Carrier_Reliability_Score* (16,2%) e *Fuel_Price_Index* (13,8%) são os três atributos mais determinantes.

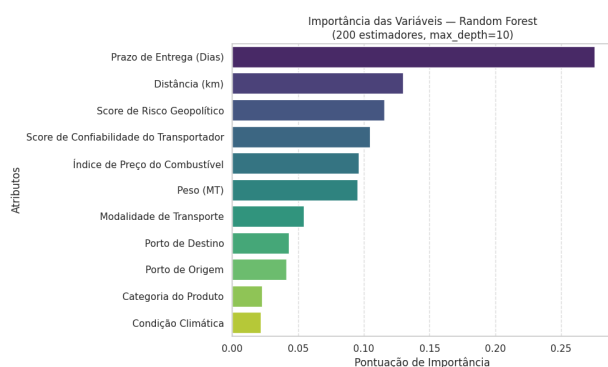


Fig. 6. Importância relativa das variáveis segundo o classificador Random Forest.

Esses resultados são coerentes pois instabilidades geopolíticas aumentam a probabilidade de atrasos nas rotas; baixa confiabilidade dos transportadores eleva o risco de falhas operacionais; e oscilações no preço do combustível impactam a pontualidade dos embarques. Variáveis como porto de origem e porto de destino apresentaram menor importância relativa, sugerindo que a identificação geográfica dos terminais tem impacto secundário frente aos índices de desempenho e risco.

VI. CONCLUSÕES

Este trabalho comparou modelos de aprendizado de máquina para predição de interrupções em cadeias de suprimentos, com pipeline completo de pré-processamento e avaliação. O Random Forest apresentou o melhor desempenho (66,00% de acurácia, $F1 = 0,63$), superando as redes neurais. A rede com Dropout obteve resultado intermediário (64,65%), superior à versão sem regularização (62,00%), indicando a eficácia da técnica. As curvas de perda sugerem boa convergência, sem overfitting relevante.

Como trabalhos futuros, recomenda-se explorar modelos com dependência temporal e técnicas de balanceamento, como SMOTE.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Os autores declaram que utilizaram ferramentas de inteligência artificial, incluindo ChatGPT, exclusivamente para auxiliar na melhoria da linguagem e legibilidade deste

manuscrito. O uso dessas ferramentas foi realizado sob supervisão humana, e os autores revisaram e editaram cuidadosamente o conteúdo, assumindo total responsabilidade pelo texto final publicado.

REFERENCES

- [1] M.S. Waller e S.E. Fawcett, "Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management," *J. Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013.
- [2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [3] D. Ivanov e A. Dolgui, "A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0," *Production Planning & Control*, vol. 32, no. 9, pp. 775–788, 2021.
- [4] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [5] I. Goodfellow, Y. Bengio e A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016.
- [6] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.