Universidade Federal do ABC
Centro de Matemática, Computação e Cognição

Avaliação do impacto da colinearidade em técnicas de inteligência artificial explicável

José Murillo da Silva Lima

Santo André - SP, Agosto/2025

José Murillo da Silva Lima

Avaliação do impacto da colinearidade em técnicas de inteligência artificial explicável

Projeto de Graduação em Computação apresentado ao Centro de Ciências, Computação e Cognição como parte dos requisitos necessários para a obtenção do Título de Bacharel em Ciência da Computação.

Universidade Federal do ABC – UFABC

Centro de Matemática, Computação e Cognição

Orientador: Prof. Dr. Paulo Henrique Pisani

Santo André - SP

Agosto/2025

Resumo

O avanço da tecnologia e o crescimento exponencial da geração de dados têm impulsionado, ano após ano, o uso de algoritmos de aprendizado de máquina em diferentes áreas do conhecimento. Entretanto, muitos desses algoritmos são considerados “caixas-pretas”, devido à complexidade que dificulta a compreensão de como produzem seus resultados. Nesse contexto, surge a área da inteligência artificial explicável (XAI), cujo objetivo é tornar mais transparentes os modelos complexos, fornecendo interpretações sobre seus resultados. Um desafio frequente em análises com múltiplas variáveis é a colinearidade, situação em que uma ou mais variáveis independentes utilizadas no treinamento do modelo apresentam elevada correlação entre si, podendo ser expressas como funções lineares umas das outras. Esse fenômeno pode comprometer a confiabilidade das explicações fornecidas pelas técnicas de XAI. Este trabalho tem o objetivo avaliar o impacto da colinearidade sobre técnicas de inteligência artificial explicável, buscando compreender de que forma essa condição influencia as explicações obtidas.

Palavras-chaves: inteligência artificial explicável, aprendizado de máquina, colinearidade, interpretação de modelos, transparência.

Abstract

The advancement of technology and the exponential growth of data generation have driven, year after year, the use of machine learning algorithms in different fields of knowledge. However, many of these algorithms are considered “black boxes” due to their complexity, which makes it difficult to understand how they produce their results. In this context, the field of explainable artificial intelligence (XAI) has emerged, aiming to make complex models more transparent by providing interpretations of their outcomes. A common challenge in analyses with multiple variables is collinearity, a situation in which one or more independent variables used to train the model are highly correlated with each other and can be expressed as linear functions of one another. This phenomenon can compromise the reliability of the explanations provided by XAI techniques. This work aims to assess the impact of collinearity on explainable artificial intelligence techniques, seeking to understand how this condition influences the obtained explanations.

Keywords: explainable artificial intelligence, machine learning, collinearity, model interpretation, transparency.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Embora o termo Inteligência Artificial tenha sido criado em 1956, a IA permaneceu por mais de meio século como uma área relativamente obscura na ciência, despertando pouco interesse prático (HAENLEIN; KAPLAN, 2019). Atualmente, com o surgimento do Big Data e o avanço do poder computacional, o uso da IA e do Aprendizado de Máquina, um subcampo da IA, tornou-se fundamental em diversas áreas do conhecimento, registrando um crescimento contínuo (MIJWIL et al., 2022).

Os sistemas baseados em IA cresceram a tal ponto que, em muitos casos, quase não há intervenção humana necessária para sua concepção e implementação. Com o desenvolvimento de algoritmos cada vez mais complexos, muitos modelos acabam se tornando "caixas-pretas", pois ocultam informações sobre o processo de aprendizado, as representações internas e o funcionamento final do modelo em formatos que não são, ou são pouco, interpretáveis pelos seres humanos (KOH; LIANG, 2017). No entanto, quando as decisões derivadas desses sistemas afetam diretamente a vida das pessoas, como ocorre, por exemplo, nas áreas da medicina, do direito ou das finanças, surge uma necessidade cada vez maior de compreender como essas decisões são produzidas pelos métodos de IA.

Uma solução que tem evoluído para enfrentar esse desafio é a área de Inteligência Artificial Explicável (XAI), cujo objetivo é desenvolver técnicas de aprendizado de máquina capazes de gerar explicações sobre o funcionamento dos modelos, permitindo que os seres humanos compreendam, confiem e consigam gerenciar de forma eficaz os resultados dos sistemas inteligentes (Barredo Arrieta et al., 2020). Entretanto, aspectos específicos, como a colinearidade entre variáveis, podem afetar a qualidade dos resultados obtidos pelas técnicas de XAI (SALIH et al., 2025). A colinearidade ocorre quando duas ou mais variáveis independentes apresentam alta correlação entre si, o que dificulta identificar quais delas são, de fato, responsáveis pelas variações no resultado do modelo, comprometendo, assim, a interpretação fornecida pelas técnicas explicativas.

1.1 Justificativa

Conforme apresentado na introdução deste trabalho, a Inteligência Artificial Explicável (XAI) é fundamental para possibilitar que seres humanos compreendam, confiem e consigam gerenciar as decisões produzidas por modelos de Machine Learning. Entretanto, estudos como o de Salih et al. (2025) destacam que essas técnicas podem apresentar limitações quando há alta correlação entre as variáveis independentes utilizadas para treinar os modelos. Essa situação, conhecida como colinearidade, afeta a capacidade das técnicas de XAI em identificar, de forma precisa, quais variáveis realmente influenciam

as previsões do modelo, comprometendo a qualidade e a confiabilidade das explicações geradas.

Apesar de alguns trabalhos abordarem esse problema, a questão da multicolinearidade ainda não é explorada e investigada de forma adequada na literatura (SALIH, 2024).

1.2 Objetivos

Este trabalho tem como objetivo avaliar o impacto da colinearidade sobre técnicas de inteligência artificial explicável em modelos treinados com diferentes algoritmos de aprendizado de máquina. Busca-se compreender como a colinearidade afeta a estabilidade das explicações geradas, identificar quais modelos e quais técnicas geram rankings de importância mais estáveis, e investigar se tratamentos nos dados, como seleção de variáveis ou redução de dimensionalidade, podem melhorar a performance das explicações. Dessa forma, o estudo visa contribuir para o desenvolvimento de modelos mais transparentes e confiáveis em contextos onde variáveis altamente correlacionadas estão presentes.

2 Fundamentação Teórica

2.1 Colinearidade

Colinearidade é um fenômeno que ocorre quando duas ou mais variáveis preditoras de um modelo estatístico apresentam uma relação linear significativa (DORMANN et al., 2013), de modo que uma variável pode ser aproximadamente expressa como combinação linear das demais. No caso de colinearidade exata, essa relação é perfeita, ou seja, uma variável é completamente determinada pelas outras.

Formalmente, considera-se que existe colinearidade entre duas variáveis X_i e X_j quando existem coeficientes α e β tais que:

$$X_j \approx \alpha X_i + \beta, \tag{2.1}$$

sendo que, na colinearidade perfeita, a relação é exata:

$$X_j = \alpha X_i + \beta. \tag{2.2}$$

Esse fenômeno reduz a independência entre as variáveis preditoras e pode afetar tanto modelos estatísticos clássicos, como Regressão Linear e Regressão Logística, quanto técnicas modernas de Inteligência Artificial Explicável (XAI) (SALIH et al., 2025).

Segundo Dormann et al. (2013), colinearidade também é chamada de multicolinearidade. No entanto, de acordo com Kim (2019), colinearidade refere-se ao caso em que uma variável apresenta forte relação linear com outra, enquanto multicolinearidade ocorre quando uma variável apresenta forte relação linear com duas ou mais variáveis.

Um indicador simples e frequentemente utilizado para detectar colinearidade é o coeficiente de correlação de Pearson entre pares de variáveis preditoras. O coeficiente de correlação de Pearson (r) mede o grau de associação linear entre duas variáveis X e Y e é definido como:

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y}, \tag{2.3}$$

onde $\mathrm{Cov}(X, Y)$ é a covariância entre X e Y, definida por:

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X})(Y_k - \bar{Y}), \tag{2.4}$$

e σ_X e σ_Y representam os respectivos desvios-padrão das variáveis X e Y.

O valor de r varia no intervalo $[-1, 1]$: valores próximos de 1 indicam forte correlação linear positiva; valores próximos de -1 indicam forte correlação linear negativa; e valores

próximos de 0 indicam ausência de dependência linear (KIRCH, 2008). Quanto maior o valor absoluto de r, maior é a dependência linear entre as variáveis, sugerindo a presença de colinearidade. De forma particular, quando $|r| = 1$, temos uma situação de colinearidade perfeita, em que uma variável é completamente determinada pela outra, e quando $|r| = 0$ temos a ausência de qualquer nível de colinearidade (KIM, 2019). Dormann et al. (2013) recomenda o limiar $|r| \geq 0{,}7$ como indicativo de quando a colinearidade começa a distorcer significativamente as estimativas dos modelos.

Em dados do mundo real, sempre existe algum grau de colinearidade entre variáveis preditoras, que pode surgir por diferentes motivos (DORMANN et al., 2013):

- **Colinearidade intrínseca:** ocorre quando variáveis colineares são diferentes manifestações de um mesmo processo subjacente, muitas vezes não mensurável (variável latente). Por exemplo, o peso e a altura de uma pessoa estão fortemente relacionados, pois ambos são representações do tamanho corporal. Variáveis assim tendem a apresentar alta correlação linear entre si.

- **Colinearidade composicional:** ocorre em dados em que as variáveis representam partes de um todo e, portanto, não são independentes entre si. Por exemplo, se medirmos a proporção de três tipos de frutas em uma cesta (maçãs, laranjas e bananas) que sempre somam 100%, um aumento na proporção de maçãs implica necessariamente uma diminuição na proporção de laranjas ou bananas. Esse tipo de relação gera colinearidade entre as variáveis.

- **Colinearidade incidental:** ocorre quando variáveis preditoras parecem colineares por acaso, por exemplo, em amostras pequenas, quando nem todas as combinações de condições possíveis estão presentes.

2.2 Trabalhos relacionados

Alguns trabalhos foram desenvolvidos com o objetivo de compreender e mitigar os efeitos da colinearidade nos algoritmos de Inteligência Artificial Explicável (XAI). Basu e Maji (2022) apresentaram um framework matemático para corrigir a multicolinearidade no cálculo de valores de Shapley. Eles argumentam que a versão tradicional do SHAP assume independência entre atributos, o que não é realista e pode distorcer a atribuição de importância. Para resolver isso, propuseram um ajuste matricial que corrige os valores de Shapley considerando as correlações entre as variáveis, de modo que a importância individual se torne independente da multicolinearidade. Além disso, estenderam essa correção para calcular efeitos combinados de pares (ou grupos) de variáveis correlacionadas, somando os valores ajustados. O método foi validado em problemas reais de classificação,

demonstrando boa eficiência computacional e explicações mais consistentes em presença de variáveis correlacionadas.

Dando continuidade à discussão sobre os impactos da multicolinearidade na explicabilidade, Salih et al. (2022) propuseram um novo critério de estabilidade para avaliar técnicas de XAI aplicadas à classificação de demência com base em imagens de ressonância magnética. Os autores treinaram modelos de classificação para diferenciar pacientes com e sem demência e analisaram a robustez dos rankings de importância gerados por métodos explicativos. A principal contribuição foi o desenvolvimento do NMR (*Normalized Movement Rate*), um critério que quantifica a estabilidade das explicações fornecidas por técnicas de XAI, especialmente em cenários com multicolinearidade entre atributos. Os resultados demonstraram que o NMR melhora a confiabilidade na identificação de variáveis informativas, contribuindo para uma personalização mais segura do monitoramento clínico.

A partir da mesma motivação, Salih et al. (2024) apresentaram o método *Modified Index Position* (MIP) como uma solução simples e agnóstica ao modelo para ajustar os rankings de importância de variáveis gerados por métodos de XAI, como o SHAP, especialmente em contextos com colinearidade entre atributos. A abordagem consiste em remover iterativamente a variável mais importante apontada pela técnica de XAI, retreinar o modelo e reaplicar a explicabilidade, observando como as demais variáveis mudam de posição no ranking. Isso permite reordenar a importância original de forma a refletir dependências entre as variáveis. Aplicado a uma tarefa de classificação de gênero (homem ou mulher) com base em nove fenótipos cardíacos, o método demonstrou rankings mais robustos e menos sensíveis à colinearidade em comparação ao SHAP tradicional, sendo validado por análise de componentes principais e plausibilidade biológica.

Complementando esse estudo, Salih et al. (2025) investigaram as limitações dos métodos SHAP e LIME ao aplicá-los separadamente em dois conjuntos de dados distintos. Em cada experimento, treinaram quatro modelos de classificação (LightGBM, Regressão Logística, Árvore de Decisão e SVC) e analisaram como cada variável era explicada por cada modelo. Os resultados mostraram que ambos os métodos apresentaram forte dependência do modelo, com variações na ordem e na direção da contribuição das variáveis em cada classificador. Além disso, a presença de colinearidade comprometeu a interpretação, já que variáveis correlacionadas recebiam baixa importância por serem explicadas por outras. Para mitigar esses efeitos, os autores propuseram o uso da métrica NMR para avaliar a estabilidade das explicações entre os modelos e o método MIP para ajustar a importância das variáveis considerando a multicolinearidade.

Explorando outra abordagem para lidar com esse desafio, Salih (2025) propôs o método *Additive Effects of Collinearity (AEC)* para superar as limitações de métodos de XAI em contextos com colinearidade entre variáveis. O autor argumenta que técnicas como o SHAP e LIME assumem independência entre os atributos, o que pode distorcer os rankings

de importância. O AEC contorna esse problema ao decompor modelos multivariáveis em modelos univariáveis, estimando o efeito isolado de cada variável e, depois, somando esses efeitos considerando suas interdependências. O método foi validado em tarefas de regressão e classificação com dados simulados e reais, utilizando regressão logística e regressão linear. A partir das explicações geradas, o autor utilizou a métrica NMR para avaliar o impacto da colinearidade, concluindo que o AEC é mais robusto e estável do que o SHAP tradicional.

Por fim, Salih (2024) realizou uma revisão sistemática da literatura para investigar como as técnicas de Inteligência Artificial Explicável (XAI) lidam com a multicolinearidade entre variáveis. Após filtrar artigos das bases Web of Science, Scopus e IEEE Xplore, foram identificadas apenas sete abordagens que tratam explicitamente esse problema. Os autores destacam que não existe, até o momento, uma técnica de XAI que por natureza mitigue o impacto da dependência entre atributos. Além disso, observam que as soluções existentes são limitadas: ou adaptadas para métodos específicos (como o SHAP), ou restritas a explicações locais, concluindo que são necessários avanços metodológicos que considerem interações complexas entre atributos correlacionados, tanto na geração quanto na visualização das explicações.

3 Metodologia Experimental

Neste trabalho, o desempenho de técnicas de inteligência artificial explicável será avaliado em dados que possuem variáveis com colinearidade. Para isso, serão utilizados conjuntos de dados que apresentam diferentes graus de correlação entre atributos. A análise será conduzida por meio da aplicação de algoritmos de classificação combinados com técnicas de explicabilidade, com o objetivo de investigar o impacto da colinearidade na estabilidade e coerência das explicações geradas.

A seguir, são descritos os conjuntos de dados utilizados, os algoritmos de classificação selecionados e os critérios adotados para a avaliação dos modelos e das técnicas de XAI.

3.1 Conjuntos de dados

Para atingir os objetivos deste trabalho, serão utilizados três conjuntos de dados: dois compostos por informações do mercado financeiro e um conjunto sintético. Todos os conjuntos representam problemas de classificação, nos quais a variável *target* corresponde a um rótulo.

1. **Default of Credit Card Clients**: Conjunto de dados com informações de clientes de cartões de crédito em Taiwan, utilizado para análise de risco e previsão de inadimplência. A variável alvo indica se o cliente entrou em inadimplência.

2. **Corporate Credit Rating**: Conjunto de dados com informações financeiras de grandes empresas americanas, coletadas entre 2010 e 2016. É utilizado para prever os ratings de crédito, que classificam o risco de inadimplência das empresas.

3. **Conjunto de dados sintéticos**: Conjunto de dados gerado com a biblioteca Scikit-learn.

A Tabela 1 apresenta um resumo dos conjuntos de dados:

Tabela 1 – Resumo dos datasets utilizados no experimento

Dataset	Qtd. de instâncias	Qtd. de colunas	Qtd. de classes	Desbalanceado?
Default of Credit Card Clients	30.000	25	2	Sim (77,9% classe 0)
Corporate Credit Rating	2.029	31	10	Sim (76,8% em BBB, BB e A)
Conjunto de Dados Sintéticos	100.000	21	2	Não (balanceado)

A seguir serão apresentados mais detalhes sobre os conjuntos de dados.

3.1.1 Default of Credit Card Clients

Este conjunto de dados, obtido no Kaggle [1], contém 30 mil registros e 25 variáveis (24 explicativas e uma variável alvo). Ele reúne, além de dados cadastrais como sexo, idade, nível de escolaridade e estado civil, informações financeiras como limite de crédito, histórico de pagamentos, valores de faturas e pagamentos mensais de clientes de cartão de crédito em Taiwan, coletados entre abril e setembro de 2005.

Todas as variáveis explicativas são numéricas, embora algumas representem categorias. A variável alvo indica se o cliente entrou em *default* (inadimplência) no pagamento do mês seguinte (valor 1) ou não (valor 0). O conjunto é desbalanceado, com 77,9% das amostras pertencentes à classe 0 e 22,1% à classe 1.

Um resumo dos preditores utilizados é apresentado a seguir:

1. **ID**: Identificador único de cada cliente.
2. **LIMIT_BAL**: Valor total do crédito concedido (em dólares taiwaneses), incluindo crédito pessoal e familiar.
3. **SEX**: Gênero do cliente (1 = masculino, 2 = feminino).
4. **EDUCATION**: Nível educacional (1 = pós-graduação, 2 = universidade, 3 = ensino médio, 4 = outros, 5 e 6 = desconhecido).
5. **MARRIAGE**: Estado civil (1 = casado, 2 = solteiro, 3 = outros).
6. **AGE**: Idade do cliente, em anos.
7. **PAY_0** até **PAY_6**: Status do pagamento mensal entre abril e setembro de 2005, respectivamente. Os valores indicam:
   - -1 = pagamento em dia;
   - 1 a 8 = atraso de 1 a 8 meses.
   - 9 = atraso de 9 meses ou mais.
8. **BILL_AMT1** até **BILL_AMT6**: Valor da fatura do cartão nos meses de setembro a abril de 2005, respectivamente (em dólares taiwaneses).
9. **PAY_AMT1** até **PAY_AMT6**: Valor do pagamento realizado nos meses de setembro a abril de 2005, respectivamente (em dólares taiwaneses).

Na Figura 1, temos a matriz de correlação das variáveis:

[1] <https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset>

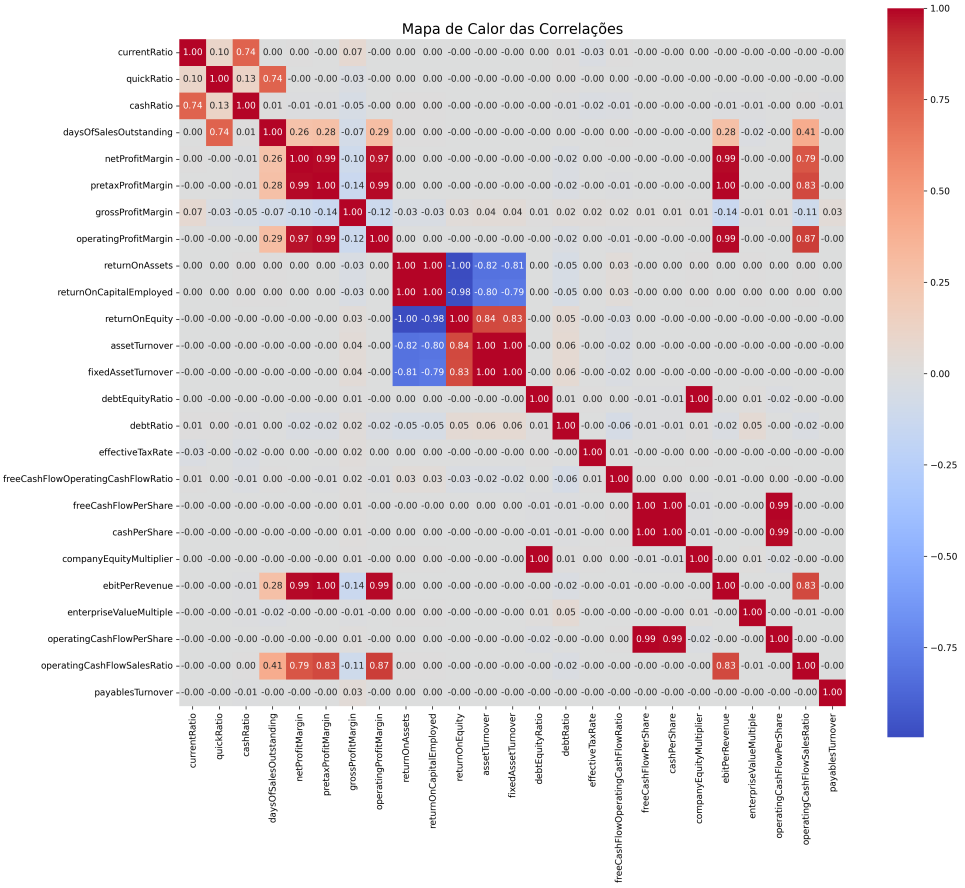

Figura 1 – Matriz de correlação das variáveis do dataset *Default of Credit Card Clients*.

A partir da matriz de correlação apresentada na Figura 1, é possível observar que as variáveis `BILL_AMT1`, `BILL_AMT2`, `BILL_AMT3`, `BILL_AMT4`, `BILL_AMT5` e `BILL_AMT6` possuem forte correlação entre si, caracterizando-se como colineares. Além disso, percebe-se que a variável `PAY_5` apresenta forte correlação com `PAY_4` e `PAY_6`.

3.1.2 Corporate Credit Rating

O *Corporate Credit Rating Dataset*, disponível no Kaggle [2], contém 2029 registros e 31 variáveis, com dados financeiros de grandes empresas americanas listadas na NYSE e na Nasdaq entre 2010 e 2016. Destas, 30 são variáveis preditoras e uma é a variável alvo, correspondente à classificação de crédito atribuída às empresas. O conjunto inclui indicadores de liquidez, lucratividade, endividamento, fluxo de caixa e dados cadastrais, sendo utilizado para prever ratings de crédito — classificações emitidas por agências que indicam o risco de inadimplência das companhias.

As variáveis preditoras podem ser agrupadas da seguinte forma:

[2] <https://www.kaggle.com/datasets/agewerc/corporate-credit-rating>

- **Índices de Liquidez:** `currentRatio`, `quickRatio`, `cashRatio`, `daysOfSalesOutstanding`;

- **Indicadores de Lucratividade:** `grossProfitMargin`, `operatingProfitMargin`, `pretaxProfitMargin`, `netProfitMargin`, `effectiveTaxRate`, `returnOnAssets`, `returnOnEquity`, `returnOnCapitalEmployed`;

- **Índices de Endividamento:** `debtRatio`, `debtEquityRatio`;

- **Indicador de Desempenho Operacional:** `assetTurnover`;

- **Indicadores de Fluxo de Caixa:** `operatingCashFlowPerShare`, `freeCashFlowPerShare`, `cashPerShare`, `operatingCashFlowSalesRatio`, `freeCashFlowOperatingCashFlowRatio`

A variável alvo, `Rating`, representa a classificação de crédito das empresas. Observa-se que esta variável está desbalanceada, com concentração significativa em poucas classes, especialmente nas categorias **BBB**, **BB** e **A**, que juntas correspondem a mais de 75% dos registros. A Tabela 2 apresenta a distribuição detalhada da variável alvo.

Tabela 2 – Distribuição da variável alvo (`Rating`) no dataset *Corporate Credit Rating Dataset*

Rating	**Quantidade**	**Frequência (%)**
BBB	671	33.07
BB	490	24.15
A	398	19.62
B	302	14.88
AA	89	4.39
CCC	64	3.15
AAA	7	0.34
CC	5	0.25
C	2	0.10
D	1	0.05
Total	2029	100.00

A seguir na Figura 2, apresenta-se a matriz de correlação entre as variáveis do dataset:

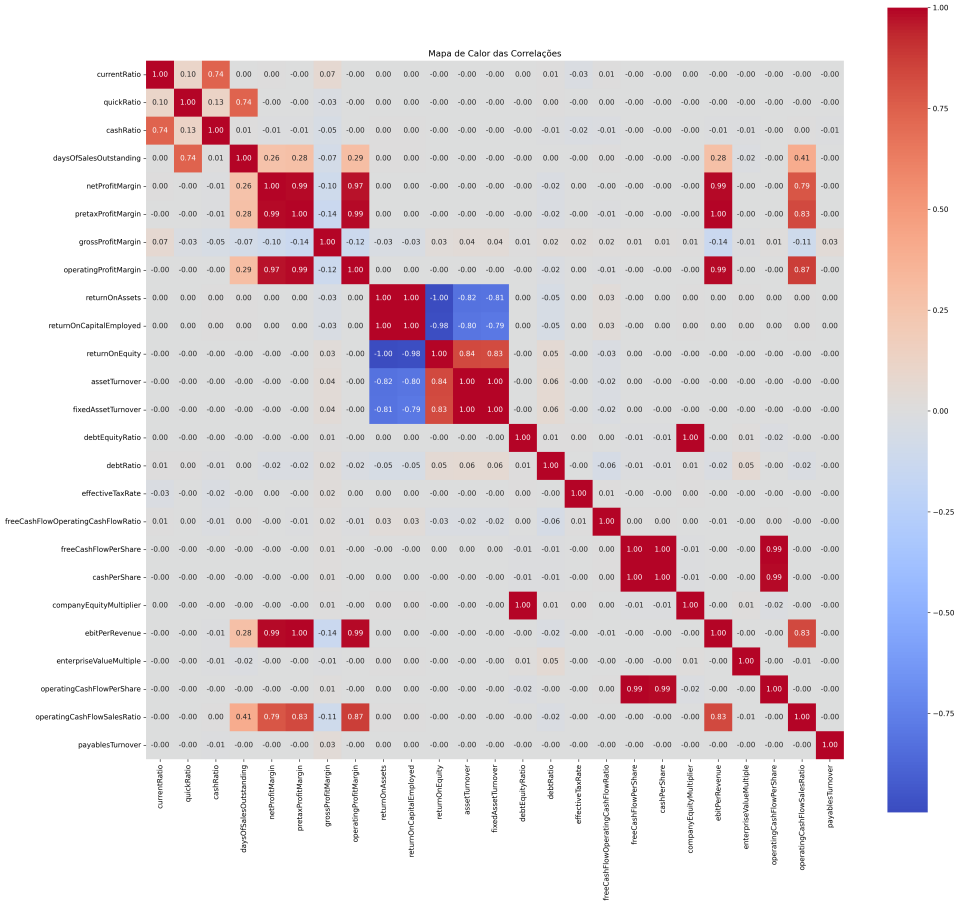

Figura 2 – Matriz de correlação das variáveis do dataset *Corporate Credit Rating Dataset*.

A partir da matriz de correlação apresentada na Figura 2, observa-se colinearidade entre boa parte das variáveis. Destacam-se as variáveis `netProfitMargin`, `pretaxProfitMargin`, `grossProfitMargin`, `operatingProfitMargin`, `returnOnAssets`, `returnOnCapitalEmployed`, `returnOnEquity`, `assetTurnover`, `fixedAssetTurnover`, `debtEquityRatio`, `freeCashFlowPerShare`, `cashPerShare`, `companyEquityMultiplier`, `ebitPerRevenue`, `operatingCashFlowPerShare` e `operatingCashFlowSalesRatio`, que possuem correlação maior ou igual a 0,8 com uma ou mais variáveis preditoras.

3.1.3 Conjunto de dados sintéticos

Assim como em Salih (2025), o conjunto de dados sintético foi gerado utilizando o método `make_classification` da biblioteca Scikit-learn, com os mesmos parâmetros utilizados no trabalho citado. Foram geradas 100 mil amostras com 16 variáveis preditoras (todas numéricas), nomeadas de `f1` a `f16`, das quais nove são informativas, cinco redundantes e duas representam ruído. Para introduzir diferentes níveis de colinearidade, foram adicionadas quatro novas variáveis (`f17`, `f18`, `f19` e `f20`) como combinações lineares de `f1`, `f2`, `f3` e `f4`, respectivamente.A variável alvo é balanceada e apresenta apenas duas classes, 0 ou 1.

A Figura 3 apresenta matriz de correlação dos dados:

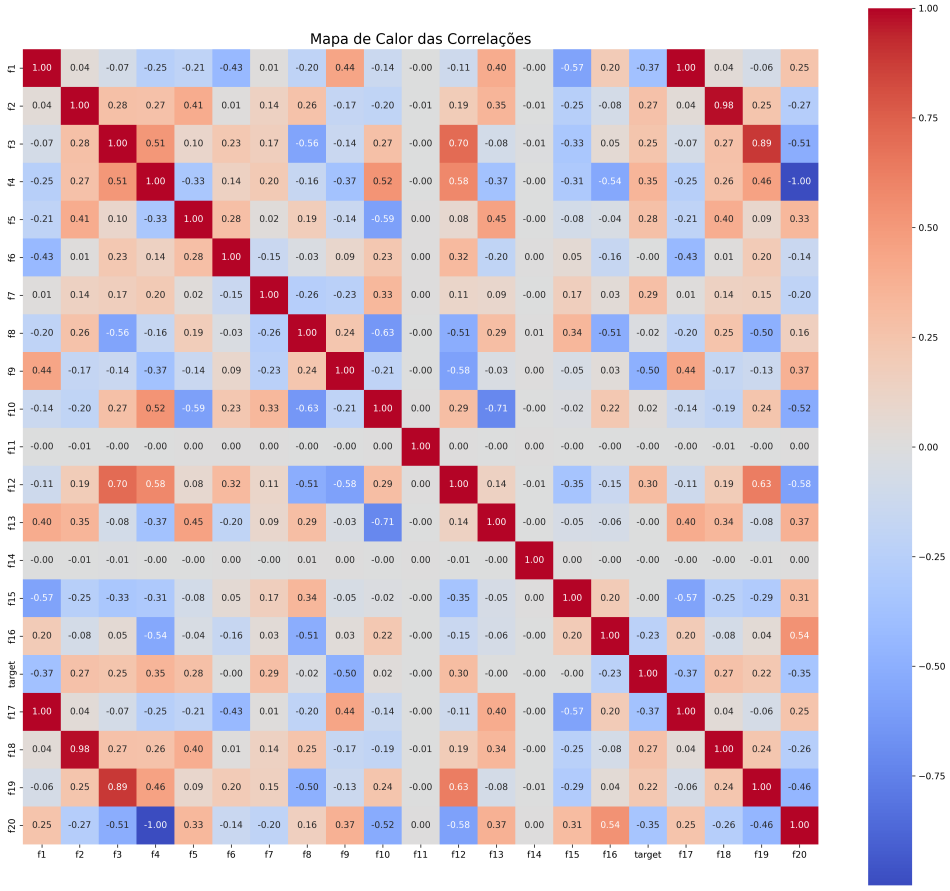

Figura 3 – Matriz de correlação das variáveis do conjunto de dados sintéticos.

3.2 Configuração do experimento

Os conjuntos de dados serão divididos em dados de treino (70%) e teste (30%). Para cada conjunto, serão treinados e avaliados quatro algoritmos de classificação: SVM, Random Forest, LightGBM e XGBoost. Esses algoritmos foram escolhidos por serem amplamente utilizados, não serem naturalmente explicáveis e atenderem às limitações de recursos computacionais disponíveis para este trabalho. A implementação será realizada com a biblioteca Scikit-learn (PEDREGOSA et al., 2011), utilizando inicialmente os parâmetros padrão de cada modelo.

O desempenho preditivo dos algoritmos será avaliado com as métricas F1-score e Acurácia Balanceada, também disponíveis na biblioteca Scikit-learn. Essas métricas permitem uma avaliação equilibrada do desempenho geral do modelo, sem penalizar excessivamente falsos negativos ou falsos positivos. Após o treinamento e avaliação, serão aplicadas as técnicas SHAP e LIME, por meio das bibliotecas shap [3] e lime [4], respectiva-

[3] <https://shap.readthedocs.io/en/latest/>
[4] <https://lime-ml.readthedocs.io/en/latest/>

mente, para gerar explicações sobre as decisões dos modelos. A estabilidade das explicações será avaliada por meio da métrica NMR (*Normalized Movement Rate*).

Posteriormente, o experimento será repetido utilizando dados transformados por PCA e dados com variáveis redundantes removidas, permitindo comparar o impacto da colinearidade na estabilidade das explicações geradas.

Referências

Barredo Arrieta, A. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, v. 58, p. 82–115, 2020. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253519308103>. Citado na página 1.

BASU, I.; MAJI, S. Multicollinearity correction and combined feature effect in shapley values. In: LONG, G.; YU, X.; WANG, S. (Ed.). *AI 2021: Advances in Artificial Intelligence*. Cham: Springer International Publishing, 2022. p. 79–90. ISBN 978-3-030-97546-3. Citado na página 4.

DORMANN, C. F. et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, v. 36, n. 1, p. 27–46, 2013. Disponível em: <https://nsojournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0587.2012.07348.x>. Citado 2 vezes nas páginas 3 e 4.

HAENLEIN, M.; KAPLAN, A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, v. 61, n. 4, p. 5–14, 2019. Disponível em: <https://doi.org/10.1177/0008125619864925>. Citado na página 1.

KIM, J. H. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, Daegu Catholic University, v. 72, n. 6, p. 558–569, dez. 2019. Epub 2019 Jul 15. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/>. Citado 2 vezes nas páginas 3 e 4.

Pearson's correlation coefficient. In: KIRCH, W. (Ed.). *Encyclopedia of Public Health*. Dordrecht: Springer Netherlands, 2008. p. 1090–1091. ISBN 978-1-4020-5614-7. Disponível em: <https://doi.org/10.1007/978-1-4020-5614-7_2569>. Citado na página 4.

KOH, P. W.; LIANG, P. Understanding black-box predictions via influence functions. In: PRECUP, D.; TEH, Y. W. (Ed.). *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 1885–1894. Disponível em: <https://proceedings.mlr.press/v70/koh17a.html>. Citado na página 1.

MIJWIL, M. M. et al. Has the future started? the current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, v. 3, n. 1, 2022. Disponível em: <https://ijcsm.researchcommons.org/ijcsm/vol3/iss1/13>. Citado na página 1.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado na página 12.

SALIH, A. et al. Investigating explainable artificial intelligence for mri-based classification of dementia: a new stability criterion for explainable methods. In: *2022 IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2022. p. 4003–4007. Citado na página 5.

SALIH, A. M. *Explainable Artificial Intelligence and Multicollinearity : A Mini Review of Current Approaches*. 2024. Disponível em: <https://arxiv.org/abs/2406.11524>. Citado 2 vezes nas páginas 2 e 6.

SALIH, A. M. Explainable artificial intelligence for dependent features: Additive effects of collinearity. In: *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*. New York, NY, USA: Association for Computing Machinery, 2025. (ICAAI '24), p. 94–99. ISBN 9798400718014. Disponível em: <https://doi.org/10.1145/3704137.3704152>. Citado 2 vezes nas páginas 5 e 11.

SALIH, A. M. et al. Characterizing the contribution of dependent features in xai methods. *IEEE Journal of Biomedical and Health Informatics*, v. 28, n. 11, p. 6466–6473, 2024. Citado na página 5.

SALIH, A. M. et al. A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, v. 7, n. 1, p. 2400304, 2025. Disponível em: <https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/aisy.202400304>. Citado 3 vezes nas páginas 1, 3 e 5.