Abordagem para verificação autoexplicativa de erros em classificadores

Approach to self-explanatory error checking in classifiers

Álvaro Jaime Gomes de Sá Silva[1], **Rhian Costa Branco Cavalcanti**[1]

[1] Análise e Desenvolvimento de Sistemas – Instituto Federal de Pernambuco (IFPE)
Paulista – PE – Brasil

ajgss@discente.ifpe.edu.br, rcbc1@discente.ifpe.edu.br

Resumo. *É perceptível o relevante papel, nas diversas áreas da sociedade, dos sistemas de Inteligência Artificial e Aprendizagem de Máquina, em especial quando utilizado em conjunto com Sistemas de Apoio à Decisão. No entanto, a falta de transparência e explicabilidade desses sistemas tem sido uma preocupação frequente. A Inteligência Artificial Explicável surge nesse cenário como uma solução para esse problema, possibilitando que sistemas inteligentes possam ganhar a confiança de seus usuários. A abordagem deste estudo se mostra relevante para aprimorar tais sistemas em áreas críticas, impulsionando uma tomada de decisão mais confiável e transparente. Portanto, este trabalho propõe uma abordagem para verificar e explicar os erros cometidos por classificadores, contribuindo para a confiança e adoção mais ampla das aplicações de Inteligência Artificial. Nas bases adotadas neste estudo, os resultados obtidos demonstraram a eficiência da proposta ao corrigir todos os erros de uma base, 75% de outra e 66% em mais duas outras, deixando apenas três bases sem correção de um total de nove bases.*

Palavras-chave: Sistemas de Apoio à Decisão, Aprendizagem de Máquina, Inteligência Artificial Explicável

Abstract. *The relevant role, in different areas of society, of Artificial Intelligence and Machine Learning systems is perceptible, especially when used in conjunction with Decision Support Systems. However, the lack of transparency and explainability of these systems has been a frequent concern. Explainable Artificial Intelligence emerges in this scenario as a solution to this problem, enabling intelligent systems to gain the trust of their users. The approach of this study is relevant to improve such systems in critical areas, boosting more reliable and transparent decision-making. Therefore, this work proposes an approach to verify and explain errors made by classifiers, contributing to confidence and wider adoption of artificial intelligence applications. In the bases adopted in this study, the results obtained demonstrated the efficiency of the proposal in correcting all errors in one base, 75% in another and 66% in two others, leaving only three bases uncorrected out of a total of nine bases.*

Keywords: Decision Support System, Machine Learning, Explanable Artificial Intelligence

1. Introdução

É sabido que a Inteligência Artificial (IA) vem desempenhando um papel de extrema importância nas últimas décadas, impulsionando avanços significativos em tecnologias e aplicações. A IA mostra-se promissora ao lidar com desafios complexos, proporcionando soluções eficazes em diversos setores, tais como medicina, química, física e geociência (XU et al., 2021).

Nesse contexto, os classificadores desempenham um papel essencial como componentes centrais da IA, sendo algoritmos que realizam tarefas de classificação, mapeando dados de entrada para categorias ou classes predefinidas (LOH, 2011). Esses modelos têm sido amplamente aplicados em inúmeras situações, tais como identificação de objetos em imagens (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), diagnóstico médico por meio de imagens (LITJENS et al., 2017), e categorização de documentos (JOACHIMS, 1998), demonstrando seu potencial em automatizar tarefas complexas com base em dados.

Entretanto, com a crescente complexidade dos sistemas de IA o processo de tomada de decisão pode ficar comprometido, tornando difícil compreender os motivos pelos quais determinadas previsões são feitas. O entendimento desses motivos é essencial para gerar confiança em tais sistemas e garantir sua aceitação. Tanto é que, mesmo que um sistema interpretável não seja tão preciso quando um sistema de caixa preta (do inglês, *black-box*), ele ainda será preferível, em certas circunstâncias, por causa de sua transparência (RIBEIRO; SINGH; GUESTRIN, 2016). Percebe-se, então, a importância de abordar essa questão. Da mesma forma, essa ausência de explicação torna-se um empecilho significativo para a adoção e confiança nas aplicações de IA em diversos setores, como saúde (BOHR; MEMARZADEH, 2020), finanças (GIUDICI, 2018) e segurança (HOADLEY; LUCAS, 2018).

A compreensão dos erros cometidos por esses classificadores pode fornecer informações valiosas para o refinamento de tais modelos, auxiliando na identificação de limitações e falhas, bem como no aprimoramento da qualidade das previsões.

Diante do empecilho exposto, o presente artigo visa ampliar os trabalhos de (OLIVEIRA; NETO, 2022) adicionando dois novos classificadores, cinco novas bases e uma nova execução de todos os experimentos anteriores. Com isso, busca-se demonstrar a viabilidade da proposta em um contexto mais abrangente.

As próximas seções dividem-se em quatro partes: Referencial Teórico, Metodologia, Experimentos e Resultados e Conclusão. O conceito de Sistemas de Apoio à Decisão (SADs) é introduzido no Referencial Teórico, assim como a exposição de trabalhos similares à proposta deste artigo e a explanação do principais conceitos da área de Inteligência Artificial Explicável (XAI, do inglês, eXplainable Artificial Intelligence) e seus tópicos subjacentes. Na seção de Metodologia é exposto o método proposto por este trabalho. As bases de dados escolhidas e a configuração para treinar os modelos, assim como seus resultados, são introduzidas na seção de Experimentos e Resultados. Por fim, na Conclusão tem-se um resumo do que foi realizado neste artigo, concomitantemente com uma discussão sobre os resultados obtidos, assim como a apresentação de possíveis trabalhos futuros.

2. Referencial Teórico

2.1. Sistemas de Apoio à Decisão

No contexto dos Sistemas de Informação, os Sistemas de Apoio à Decisão são ferramentas que auxiliam os gestores a tomar decisões mais informadas e fundamentadas em áreas de planejamento estratégico, com a habilidade de adaptar-se rapidamente a novos problemas (APRIYANSYAH, 2022).

Eles ainda podem ser definidos como sistemas baseados em computador que apoiam atividades de tomada de decisão, incluindo sistemas especialistas e Análise de Decisão Multicritério (MCDA) (MORGE, 2007).

Tais sistemas são construídos visando lidar com problemas de grande porte, nos quais há grande quantidade de informações e incertezas envolvidas. Eles, normalmente, usam IA e aprendizagem de máquina para fazer a análise dos dados a fim de tomar decisões estratégicas.

Os SADs têm sido amplamente aplicados em diversos setores, como saúde (SUTTON et al., 2020), finanças (SHAIKH et al., 2021) e logística (FANTI et al., 2017), fornecendo suporte para problemas que requerem análise de dados em tempo real, avaliação de riscos e simulações. Ao integrar IA com Aprendizagem de Máquina (AM), é possível obter uma maior eficiência e precisão na tomada de decisões, baseada em análises objetivas e dados em grande escala.

2.2. Inteligência Artificial Explicável

A já citada crescente complexidade nos modelos de IA também tem estimulado o aumento do uso de Inteligência Artificial Explicável. Ela apresenta-se como uma área que aborda essa questão da complexidade, visando desenvolver métodos para tornar os modelos de IA mais compreensíveis para os usuários humanos.

A XAI usa técnicas de AM que possam fornecer modelos explicáveis e, ainda assim, manter o alto nível de performance da aprendizagem propriamente dita. Ela também almeja fazer com que humanos possam entender, confiar e gerenciar de forma efetiva a emergente geração de IA parceiras (ARRIETA et al., 2020).

A importância da XAI vai além da simples transparência dos modelos de IA. No campo da medicina, existe uma demanda crescente por soluções de IA que sejam não apenas performáticas, mas transparentes e interpretáveis. Além disso, a explicabilidade pode ajudar a aumentar a confiança em sistemas de IA no futuro (HOLZINGER et al., 2019).

Portanto, nota-se que a aproximação da XAI com a AM é crucial para garantir a transparência e confiabilidade aos modelos. Um dos argumentos a favor do uso da XAI é que ela busca desenvolver abordagens que permitam aos usuários entender como os modelos de AM estão tomando suas decisões, identificar enviesamentos e erros, além de fornecer explicações confiáveis e compreensíveis sobre o processo de tomada de decisão (SAMEK; WIEGAND; MüLLER, 2017).

2.3. Aprendizagem de Máquina

No mesmo raciocínio, a AM exerce um papel considerável no desenvolvimento de modelos de IA, incluindo classificadores. A AM trata-se da capacidade dos sistemas de IA de aprender a partir de dados, identificar padrões, tomar decisões e fazer previsões com base no que aprendeu. A preocupação principal do campo de aprendizagem de máquina é com a questão de como construir aplicações de computador que melhoram automaticamente através da experiência (MITCHELL, 1997).

Tais aplicações abarcam soluções de mineração de dados que aprendem a detectar transações fraudulentas de cartões de crédito, sistemas de filtragem de informações que aprendem as preferências de leitura dos usuários e veículos autônomos que aprendem a dirigir em vias públicas.

Em contrapartida, o recorrente problema da falta de explicabilidade mantém-se nos modelos de AM. E mesmo que os sistemas de AM venham alcançando maiores desempenhos preditivos, o aumento de sua presença na sociedade demonstrou a importância e a necessidade de interpretabilidade nesses sistemas (CARVALHO; PEREIRA; CARDOSO, 2019).

Ao explorar a abordagem proposta neste trabalho, que visa a verificação autoexplicativa de erros em classificadores, ficará indubitável perceber que a conexão harmoniosa entre AM, XAI e SADs oferece uma oportunidade de avanço no campo da IA.

Integrando a capacidade de aprendizagem dos modelos de AM com a transparência e explicabilidade fornecidas pela XAI, busca-se construir sistemas de IA confiáveis e compreensíveis, capazes de apontar os erros cometidos pelos classificadores e aprimorar a tomada de decisão em diversas áreas da sociedade.

Logo, neste artigo serão utilizados classificadores amplamente conhecidos e utilizados, como Random Forest (BREIMAN, 2001) e XGBoost (CHEN; GUESTRIN, 2016). Eles são reconhecidos por sua eficácia em tarefas de classificação e, por meio da abordagem proposta, serão explorados para identificar e verificar os erros cometidos, fornecendo inferências autoexplicativas que auxiliem na compreensão dos motivos pelos quais esses erros ocorrem.

3. Metodologia

Nesta seção será apresentada explicação geral da estratégia de tomada de decisão sugerida, além de conceitos importantes sobre o uso da orientação como sinal de problemas com as inferências dos modelos inteligentes. Após isso, também serão abordadas informações sobre a identificação e explicação de erros.

3.1. Visão geral da abordagem de tomada de decisão com verificação de erros autoexplicativa

Uma visão geral do processo de tomada de decisão proposto pode ser observada na Figura 1. A estratégia de decisão proposta neste trabalho considera que um novo problema de decisão (P) será avaliado por um Tomador de Decisão (TD) e esse respectivo TD deve escolher uma solução (S) para resolver dado problema com a ajuda de um SAD. No âmbito deste artigo, um modelo de decisão inicial (*m0*), que atua como classificador, deve ser selecionado a partir de diversas alternativas disponíveis em uma Base de Dados de Modelos (BDM) previamente criada. O modelo escolhido (*m*) será usado para fazer uma inferência (*inf*) sobre o P. Uma explicação (*exp1..n*) sobre a *inf*, podendo ser uma ou mais, será extraída com o auxílio de um ou mais Geradores de Explicação (GE1..n).

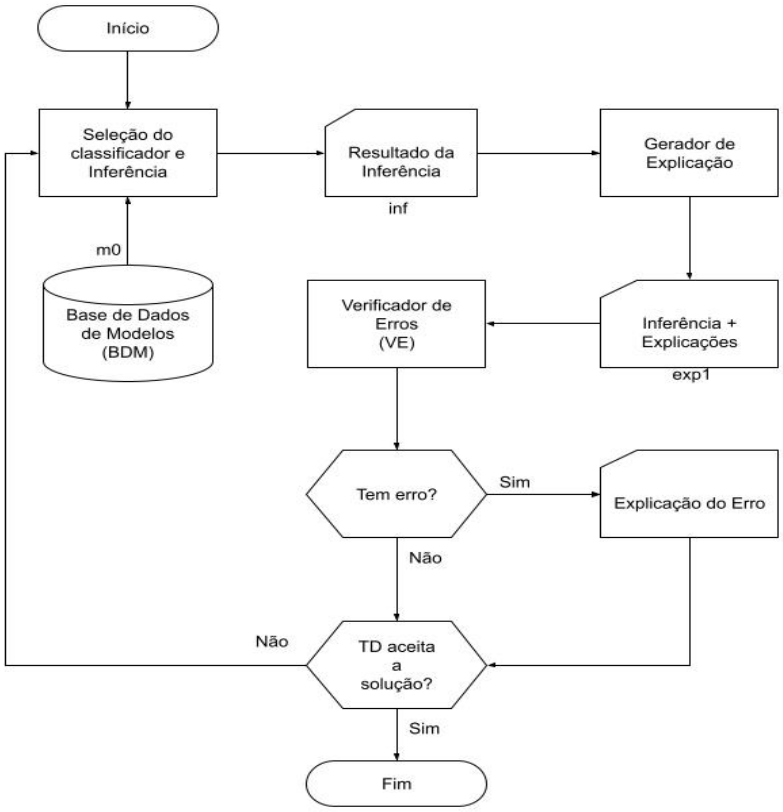
Em seguida, o Verificador de Erros (VE) utiliza a *inf* e sua respectiva explicação *exp1..n* para validá-la, avaliando se houve ou não um erro. Caso um possível erro seja detectado, a *inf* anterior e a *exp1..n* serão melhoradas com a explicação do VE e, após isso, são submetidas à avaliação do TD. Espera-se que, ao receber uma explicação de erro, o TD possa investigar mais a fundo a solução, tendo a possibilidade de mudar de opinião antes de tomar uma decisão final. Se o TD estiver em uma situação de dúvida, ele pode decidir usar um modelo de decisão diferente (*m1*), com o objetivo de reduzir a probabilidade de erros de classificação.

As seguintes subseções irão expor mais detalhes sobre como projetar um verificador de erros classificador e uma abordagem para implementar a detecção de erros autoexplicativa.

3.2. Explicações de inferência como possíveis evidências de erro

Conforme discutido na Subseção 2.2, as explicações são essenciais para que os usuários possam ter confiança no sistema que usam. No entanto, as explicações que podem ser extraídas dos modelos preditivos são um tanto diversificadas, seja na forma ou no significado. Portanto, ao usar um algoritmo, as explicações devem ser quantificadas e devem ser, pelo menos, correlacionadas com inferências corretas ou incorretas.

Figura 1. Visão geral da abordagem de decisão auxiliada pela detecção de erros autoexplicativa



Nessa lógica, abaixo apresenta-se uma lista não exaustiva de perguntas cujas explicações poderiam possivelmente contribuir para um verificador de erros:

- Qual é a acurácia do modelo em relação às N instâncias de treinamento mais similares?
- Quais atributos que foram avaliados como discriminadores importantes de classe são utilizados por este classificador?
- Quão próximo esta instância prevista como classe X está de uma instância de treinamento da classe X?
- Quão distante esta instância prevista como classe X está da instância de treinamento mais próxima de uma classe diferente de X?
- Qual é a acurácia de teste do classificador?
- Qual é a acurácia da classe prevista pelo classificador?

Notadamente, algumas das perguntas antecedentes foram empregadas na seção de Experimentos e Resultados com o objetivo de evidenciar a exequibilidade da proposta. Também vale salientar que, por essa ser apenas uma lista exemplificativa e não absoluta, qualquer outra pergunta cujas explicações possam ser quantificadas para uso em classificadores são válidas para avaliação e utilização em abordagens de verificação de erros. Para facilitar ainda mais a compreensão dessa proposta, abaixo está exposto um pseudocódigo utilizado para gerar as Bases de Dados Verificadoras de Erros de treinamento e teste (BD-VE), que foram utilizadas para treinar e testar os classificadores de VE.

Figura 2. Pseudocódigo usado para gerar as Bases de Dados Verificadoras de Erro

```
criacao_verificador_de_erros_bd(bd_original, modelo_treinado):
  verificador_de_erros_bd = []
  for amostra in bd:
      previsao = modelo_treinado.previsao(amostra.entradas)
      if previsao == amostra.real_classe:
          VE_classe = "CORRETO"
      else:
          VE_classe = "ERRO"
      explicacoes = extracao_explicacoes(modelo_treinado, amostra, previsao)
      verificador_de_erros_bd.append(explicacoes, VE_classe)
  return verificador_de_erros_bd
```

3.3. Classificador do Verificador de Erros e explicações

Para fazer a detecção dos erros, após a montagem da BD-VE, serão empregadas técnicas de Aprendizagem de Máquina supervisionada que serão aplicadas por classificadores. Estes serão introduzidos na seção de Experimentos e Resultados.

É importante ressaltar que pode haver diferenças entre os exemplos que contêm amostras da VE_classe “CORRETO” em relação aos que contêm “ERRO”. Por esse motivo, pode ser necessário tomar alguma medida para equilibrar a base de dados, garantindo uma compreensão adequada das métricas explicativas que representam inferências corretas ou incorretas. Essa medida poderá envolver o aumento da classe minoritária “ERRO”ou, de maneira mais prudente, a redução da classe majoritária “CORRETO”.

Um motivo para se empregar um classificador VE transparente, que é um VE que pode ser interpretado como um algoritmo de Árvore de Decisão (DT), é que ele pode ser inspecionado diretamente para extrair as explicações do erro. Caso contrário, é necessário utilizar classificadores explicáveis personalizados ou classificadores regulares, juntamente com técnicas de explicação post-hoc.

Duas são as vantagens da abordagem proposta: (i) quando há a aceitação de uma maior automação na inferência de decisões, o VE pode ser empregado de maneira automatizada para tentar, ininterruptamente, outro modelo, até que qualquer erro de inferência seja eliminado e (ii) ao invés de se dedicar à inspeção de múltiplas explicações provenientes de um modelo de decisão, o TD pode se dedicar à inspeção da explicação do VE somente quando um erro é constatado, preservando, assim, seus recursos cognitivos.

4. Experimentos e Resultados

Nesta seção, são expostas as bases de dados escolhidas para este trabalho. Também são apresentadas as configurações para treinar os modelos e os verificadores de erro, assim como seus resultados e quais erros foram detectados e quantos foram corrigidos. O código utilizado para gerar os resultados foi desenvolvido na linguagem Python em conjunto com a biblioteca de aprendizagem de máquina Scikit-Learn.

4.1. Bases de dados analisadas e geração do modelo de decisão

As seguintes bases de dados para *benchmark* foram escolhidas por suas características distintas, tais como número de instâncias, classes e atributos. Vale ressaltar que, por serem diversificadas, elas

foram selecionadas para avaliar a generalidade dessa proposta. Pode-se conferir as peculiaridades de cada base na Tabela 1.

Tabela 1. Características das bases utilizadas

Base de Dados	*Número de Amostras*	*Número de Atributos*	*Número de Classes*
Auto	193	19	3
Breast Cancer	683	9	2
South German Credit	1000	20	2
Heart	297	13	2
Ai4i-2020	10000	14	3
Zoo	101	16	7
Glass	214	9	6
Thyroid	3772	21	3
Wine	178	12	15

Para realizar o tratamento adequado de cada base, o pré-processamento delas ocorreu da seguinte forma: (i) remoção das duplicatas e instâncias com valores nulos, (ii) normalização dos atributos numéricos e (iii) concessão de rótulos aos atributos categóricos. Na base Auto, que normalmente não é usada para classificação e, sim, regressão, foram feitas modificações para que a coluna ”preço” fosse substituída pelos rótulos de classe *low*, *medium* e *high*.

As bases de dados foram separadas em três partes: a primeira parte com aproximadamente 60% usada para o treinamento do modelo, a segunda com 20% usada para testar os modelos e terceira com também 20% usada para casos de decisão simulados, cujo valor foi chamado de Acurácia da Simulação. As partes de teste e decisão simulada não foram usadas para treinamento do modelo.

Não foi realizado nenhum ajuste de hiperparâmetro, pois a proposta principal deste trabalho é avaliar a viabilidade e a generalidade da mesma. Vale ressaltar que o uso de acurácia nos ensaios desse trabalho foi permitida devido ao fato das bases possuírem níveis baixo de desbalanceamento de classe. Sete classificadores foram escolhidos: Árvore de Decisão, Floresta Aleatória (RF), K-vizinhos mais próximos (KNN), Máquina de Vetor de Suporte (SVM), Regressão Logística (LR), Redes Neurais Artificiais (ANN) e eXtreme Gradient Boosting (XGBoost).

Os hiperparâmetros padrões foram utilizados: “gini” para o parâmetro criterion no DT, 100 para o parâmetro n_estimators no RF, 5 para o parâmetro n_neighbors no KNN, 1 para o parâmetro C no SVM, “l2” para o parâmetro penalty no LR, 100 para o parâmetro hidden_layer_sizes no ANN e “log_loss” para o parâmetro loss no XGBoost.

A base de modelos de cada base de dados foi criada com 10 modelos. Cada modelo foi selecionado de forma aleatória dentro os classificadores supracitados. Com a adoção desses classificadores, espera-se diversificar os métodos de inferência já que tais classificadores possuem diferentes estratégias de aprendizado, em especial os de RF e XGBoost por suas características mais robustas.

4.2. Avaliação e métricas do verificador de erros

Os resultados base para os melhores e segundo melhores modelos podem ser observados na Tabela 2. A partir dela, nota-se que a fase de aprendizagem foi apropriadamente conduzida por causa da aproximação entre os valores de acurácia de Teste e de Treino dos classificadores das bases Heart, South German, Breast Cancer, Ai4i-2020, Thyroid e Wine.

Tabela 2. Resultados obtidos pelos classificadores aplicados às bases analisadas

Base de Dados	Primeiro e Segundo melhores modelos	Acurácia (Treino)	Acurácia (Teste)	Acurácia (Simulação)
Auto	SVC	0,9130	0,8205	0,9230
	DT	1,0000	0,7692	0,8205
Breast Cancer	LR	0,9755	0,9635	0,9562
	SVC	0,9755	0,9635	0,9635
South German Credit	MLP	0,7666	0,7450	0,7050
	SVC	0,7966	0,7400	0,7450
Heart	SVC	0,8483	0,8666	0,7796
	LR	0,8539	0,8666	0,7796
Ai4i-2020	MLP	0,9986	0,9990	0,9880
	KNN	1,0000	0,9980	0,9975
Zoo	RF	1,0000	0,9047	0,9500
	XGB	1,0000	0,9047	0,9500
Glass	XGB	1,0000	0,8604	0,8139
	RF	1,0000	0,7906	0,7674
Thyroid	XGB	1,0000	0,9947	0,9960
	RF	1,0000	0,9933	0,9946
Wine	LR	1,0000	0,9722	0,9722
	SVC	1,0000	0,9444	0,9722

A maior diferença de acurácia entre os conjuntos de treinamento e teste foi na base Glass, aproximadamente 14% de diferença. Logo depois, vem a base Zoo com uma diferença de 10% entre esses conjuntos. Em seguida, a terceira maior diferença fica na base Auto com 9% aproximadamente. As outras bases de dados não foram maiores que 3%, cujo valor refere-se a base Wine. Já o valor da acurácia de simulação em relação ao valor da acurácia de treinamento apresentou uma diferença mais alarmante na base Glass, com um valor aproximado de 19%. Enquanto isso, os outros valores de diferença entre esses conjuntos não foram maiores que 7%. que é o valor da base Heart.

Cada classificador do VE foi treinado com uma BD-VE conforme exposto na Subseção 3.3. Para realizar as explicações de cada inferência foram usadas as seguintes métricas:

- Acurácia semelhante (AS): a acurácia do modelo ao considerar as k amostras de treinamento mais semelhantes;
- Distância da mesma classe (DMC): a distância do padrão avaliado ao padrão de treinamento mais próximo da mesma classe previsto pelo modelo;
- Distância de outra classe (DOC): distância do padrão avaliado ao padrão de treino mais próximo de uma classe diferente daquela prevista pelo modelo.

Os classificadores aplicados nas bases de dados desse artigo produziram BD-VE diferentes cada um. Além disso, a classe majoritária foi reduzida por meio do método de *random undersampling*, cuja técnica visa evitar o desequilíbrio das classes por meio da seleção aleatória de exemplos da classe majoritária e a exclusão deles na base de treinamento.

Em seguida, cada BD-VE foi usada para treinar os classificadores abordados neste artigo: DT, RF, KNN, SVM, LR, ANN e XGBoost. O número total de elementos para cada base, o número de erros de classificação usando o classificador com maior acurácia, assim como quantos desses erros foram detectados pelo VE e corrigidos pelo mesmo, podem ser vistos na Tabela 3.

Com base nessa mesma tabela nota-se que o número total de erros é relativamente baixo quando comparado ao número total de elementos. Da mesma forma, percebe-se também que o VE conseguiu identificar uma quantidade bastante razoável de erros no total.

O VE saiu-se excepcionalmente bem ao corrigir todos os erros na base Zoo. Na base Glass, foram corridos aproximadamente 66% dos erros identificados, deixando apenas cinco erros de dez não corrigidos. Já nas bases Breast Cancer e South German Credit foram corrigidos aproximadamente 33% e 32% do erros identificados respectivamente. Em seguida, na base Thyroid foram corrigidos dois erros de três, enquanto na base Wine três erros de quatro foram corrigidos. Por último, percebe-se que o VE não corrigiu os erros das bases Auto, Heart e Ai4i-2020, situação que pode ser explorada por trabalhos futuros a fim de melhorar a eficiência do VE.

Tabela 3. Estatísticas das bases utilizadas exibindo o total de elementos, total de erros de classificação, total de erros reais e total de erros corrigidos

Baases de Dados	Total de Elementos	Erros	Erros Reais Identificados	Erros Reais Corrigidos Pela Mudança de Modelo
Breast Cancer	137	4	6	2
Auto	39	3	0	0
Soutch German	200	44	47	15
Heart	59	13	9	0
Ai4i - 2020	2000	4	4	0
Zoo	20	1	12	12
Glass	43	7	15	10
Thyroid	754	1	3	2
Wine	36	1	4	3

5. Conclusão

Através da integração da XAI com a AM e os SADs, este trabalho busca promover a transparência, confiabilidade e compreensibilidade dos modelos de IA. A verificação autoexplicativa dos erros fornece informações valiosas sobre as razões pelas quais os classificadores falham em determinados casos, permitindo a identificação de limitações, aprimoramento dos modelos e refinamento das estratégias de tomada de decisão.

Percebe-se que, com os resultados obtidos, a proposta deste estudo é válida para detectar boa parte dos erros feitos pelos modelos de decisão e, em alguns casos, corrigindo todos os erros, como foi o caso da base Zoo. Esse resultado mostra a relevância do VE como uma ferramenta promissora para aprimorar a tomada de decisão em tarefas de classificação.

A inclusão dos classificadores XGBoost e RF na abordagem fortaleceu a inspeção automatizada de explicações, visto que esses classificadores são reconhecidos por sua eficiência e capacidade de lidar com problemas complexos. Com a adição desses modelos mais potentes e das novas bases, amplia-se o escopo do artigo antecedente e reforça-se a viabilidade da proposta de verificação autoexplicativa de erros em uma variedade de bases de dados.

Como trabalhos futuros, é válido explorar a aplicação da abordagem proposta em diferentes domínios e problemas específicos, para avaliar sua generalização e adaptabilidade. Além disso, é importante considerar a utilização de métricas específicas para avaliar a qualidade das explicações geradas pelos VE e desenvolver técnicas adicionais para melhorar a interpretabilidade das informações fornecidas.

Outra sugestão de pesquisa é a investigação de métodos para aprimorar a eficiência e a escalabilidade da verificação autoexplicativa, permitindo a análise de grandes volumes de dados e dando suporte a modelos de IA mais complexos. Esses avanços tecnológicos poderão contribuir para uma adoção mais ampla e confiável dos sistemas de IA em diversas áreas de aplicação.

Referências

APRIYANSYAH, H. Literature review of: Decision support system: Organization, human resources and knowledge management. *Dinasti International Journal of Economics, Finance & Accounting*, v. 3, n. 2, p. 136–148, 2022. Disponível em: https://dinastipub.org/DIJEFA/article/view/1246. Acesso em: 28 jul. 2023.

ARRIETA, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, v. 58, p. 82–115, 2020. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103. Acesso em: 28 jul. 2023.

BOHR, A.; MEMARZADEH, K. The rise of artificial intelligence in healthcare applications. *Artificial Intelligence in healthcare. Academic Press*, p. 25–60, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128184387000022. Acesso em: 28 jul. 2023.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. Disponível em: https://link.springer.com/article/10.1023/a:1010933404324. Acesso em: 28 jul. 2023.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Eletronics*, v. 8, n. 8, p. 832, 2019. Disponível em: https://www.mdpi.com/2079-9292/8/8/832. Acesso em: 28 jul. 2023.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 785–794, 2016. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939785. Acesso em: 28 jul. 2023.

FANTI, M. P. et al. A decision support system for cooperative logistics. *IEEE Transactions on Automation Science and Engineering*, v. 14, n. 2, p. 732–744, 2017. Disponível em: https://ieeexplore.ieee.org/abstract/document/7855636. Acesso em: 28 jul. 2023.

GIUDICI, P. Fintech risk management: A research challenge for artificial intelligence in finance. *Frontiers in Artificial Intelligence*, v. 1, p. 1, 2018. Disponível em:

https://www.frontiersin.org/articles/10.3389/frai.2018.00001/full. Acesso em: 28 jul. 2023.

HOADLEY, D. S.; LUCAS, N. J. Artificial intelligence and national security. *Congressional Research Service Washington, DC*, 2018. Disponível em: https://a51.nl/sites/default/files/pdf/R45178.pdf. Acesso em: 28 jul. 2023.

HOLZINGER, A. et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 9, n. 4, p. e1312, 2019. Disponível em: https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1312. Acesso em: 28 jul. 2023.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg*, p. 137–142, 1998. Disponível em: https://link.springer.com/chapter/10.1007/BFb0026683. Acesso em: 28 jul. 2023.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012. Disponível em: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html. Acesso em: 28 jul. 2023.

LITJENS, G. et al. A survey on deep learning in medical image analysis. *Medical image analysis*, v. 25, p. 60–88, 2017. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1361841517301135. Acesso em: 28 jul. 2023.

LOH, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, v. 1, n. 1, p. 14–23, 2011. Disponível em: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.8. Acesso em: 28 jul. 2023.

MITCHELL, T. M. *Machine Learning*. 1st. ed. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997.

MORGE, M. The hedgehog and the fox an argumentation-based decision support system. *International Workshop on Argumentation in Multi-Agent Systems. Berlin, Heidelberg: Springer Berlin Heidelberg*, p. 114–131, 2007. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-78915-4_8. Acesso em: 28 jul. 2023.

OLIVEIRA, F. R. D. S.; NETO, F. B. L. Self-explanatory error checking capability for classifier-based decision support systems. *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Montevideo, Uruguay*, p. 1–6, 2022. Disponível em: https://ieeexplore.ieee.org/document/9981853. Acesso em: 28 jul. 2023.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016. Disponível em: https://arxiv.org/abs/1606.05386. Acesso em: 28 jul. 2023.

SAMEK, W.; WIEGAND, T.; MüLLER, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017. Disponível em: https://arxiv.org/abs/1708.08296. Acesso em: 28 jul. 2023.

SHAIKH, A. A. et al. Decision support system for customers during availability of trade credit financing with different pricing situations. *RAIRO-Operations Research*, v. 55, n. 2, p. 1043–1061,

2021. Disponível em: https://www.rairo-ro.org/articles/ro/abs/2021/03/ro200331/ro200331.html. Acesso em: 28 jul. 2023.

SUTTON, R. T. et al. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, v. 3, n. 1, p. 17, 2020. Disponível em: https://www.nature.com/articles/s41746-020-0221-y. Acesso em: 28 jul. 2023.

XU, Y. et al. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, v. 2, n. 4, 2021. Disponível em: https://www.cell.com/article/S2666-6758(21)00104-1/fulltext. Acesso em: 28 jul. 2023.