VII SIMAC
SIMPÓSIO ACADÊMICO
FACULDADE ENG. SALVADOR ARENA

Big Data, Mineração de Dados e IA: Fundamentos e Aplicações na Detecção de Fraudes

Ana Beatriz Canassa Jacinto, anabeatrizcanassajacinto@gmail.com, Faculdade Engenheiro Salvador Arena
Lucas Barbosa Silva, 081230009@faculdade.cefsa.edu.br, Faculdade Engenheiro Salvador Arena

Resumo

Este artigo explora a interseção entre Big Data, Mineração de Dados e Inteligência Artificial (IA), com foco na transformação de grandes volumes de dados em conhecimento útil. Por meio de uma revisão bibliográfica sistematizada e uma aplicação prática com o framework Apache Spark, o estudo investiga como essas tecnologias se complementam em um ecossistema de análise inteligente. A metodologia incluiu experimentação em ambiente Databricks com scripts desenvolvidos em PySpark, utilizando a arquitetura de dados em camadas (Bronze, Silver e Gold) para ilustrar um fluxo típico de mineração de dados. Os resultados evidenciam que a integração entre técnicas de limpeza, tratamento e transformação de dados com ferramentas modernas permite potencializar a geração de insights estratégicos. Conclui-se que a combinação entre conhecimento teórico e prática tecnológica contribui significativamente para a formação técnica e aplicação eficiente dessas ferramentas em diferentes contextos sociais e econômicos.

Palavras-chave: Big Data. Mineração de Dados. Inteligência Artificial. PySpark. Ciência de Dados.

Introdução

O avanço das tecnologias digitais tem levado a um crescimento exponencial na geração de dados provenientes de diversas fontes, como dispositivos móveis, redes sociais, sensores Internet Of Things (IOT, ou, em português, Internet das Coisas) e transações financeiras (CHEN, 2024). A capacidade tecnológica per-capita mundial de guardar informação, duplica, em média, a cada 40 meses desde 1980 (HILBERT & LÓPEZ, 2011). Esse fenômeno impulsionou o desenvolvimento do conceito de Big Data, que se caracteriza pelos três Vs: volume, variedade e velocidade (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Destacando a característica de grande volume de dados, este pode ser definido por conjuntos que alcançam ordens de magnitude de terabytes (10^{12} Bytes) ou superiores, frequentemente chegando a petabytes (10^{15} Bytes). Para fins comparativos, um arquivo de PDF tipicamente varia de alguns Quilobytes (10^{3} bytes) para alguns Megabytes (10^{6} bytes) (Dependendo do seu conteúdo). Para lidar com essa imensa quantidade de informações, O uso da teoria de estrutura Big Data torna-se fundamental já que oferece organização, facilidade de análise e de desenvolvimento de bancos de dados relacionais (Chen, 2024), também são utilizadas tecnologias como Hadoop e Spark, que possibilitam o processamento eficiente e a extração de padrões a partir dos dados que possam ser usados de forma estratégica. (MEMON et al., 2017)
.

No contexto do Big Data, a mineração de dados (Data Mining) desempenha um papel essencial, pois permite a identificação de padrões e tendências ocultas em grandes volumes de informações. Essa abordagem envolve técnicas de aprendizado de máquina, redes neurais e estatísticas avançadas para

transformar dados brutos em conhecimento útil. Aplicações como previsão de falhas em equipamentos por meio de sensores IoT e análise geoespacial para exploração mineral demonstram o impacto da mineração de dados em setores estratégicos (ORACLE, 2022). Dessa forma, a combinação entre Big Data e Data Mining possibilita a tomada de decisões mais precisas e fundamentadas em evidências.

O aprendizado de máquina (Machine Learning) e a Inteligência Artificial (IA) complementam esse ecossistema, automatizando a análise e interpretação dos dados em larga escala. A IA, dividida entre IA fraca e IA forte, busca simular capacidades cognitivas humanas para executar tarefas complexas, como visão computacional e processamento de linguagem natural (XU; WUNSCH, 2005). Dentro desse campo, o aprendizado de máquina emerge como um subcampo fundamental, permitindo que algoritmos identifiquem padrões e tomem decisões com base em dados históricos, sem programação explícita (MITCHELL, 1997). Técnicas como aprendizado supervisionado, não supervisionado e por reforço são amplamente utilizadas em aplicações como diagnóstico médico, análise financeira e sistemas autônomos. Diante desse cenário, este artigo explora a interseção entre Big Data, Data Mining e Inteligência Artificial, destacando seus impactos, desafios e contribuições para a sociedade atual.

O Jornalista Steve Lohr apontava da importância que Big Data tinha em sua publicação de 2012 no jornal The New York Times: “The Age of Big Data” (Que pode ser traduzido para “A era da Big Data”). Nesta publicação, o escritor ja diz:

“A história é similar em campos tão variados como os da ciência, esportes, publicidade e saúde pública – Um movimento para descobertas e tomada de decisões lideradas por dados. “É uma revolução,” diz Gary King, diretor do Instituto para Ciência Social Quantitativa de Harvard, “A gente está apenas começando, mas a marcha de quantificação, feito possível pela quantidade enorme de dados, vai impactar o mundo acadêmico, empresarial e governamental. Não existe setor que não será impactado.”

Uma das aplicações mais sensíveis e socialmente relevantes dessa tríade tecnológica — Big Data, mineração de dados e inteligência artificial — está na detecção automatizada de fraudes financeiras. Com a digitalização acelerada da economia, aumentou-se também a vulnerabilidade das transações eletrônicas. Segundo relatório da Serasa Experian (2025), mais da metade da população brasileira (51%) foi vítima de algum tipo de fraude em 2024, sendo as mais frequentes aquelas envolvendo cartões de crédito (47,9%), boletos falsos e golpes via Pix. Em janeiro de 2025, o país registrou 1,24 milhão de tentativas de fraude — uma a cada 2,2 segundos —, com aumento de 41,6% em relação ao mesmo período do ano anterior. No setor bancário, o primeiro trimestre do ano somou quase 2 milhões de tentativas de fraude, com prejuízo estimado em mais de R$ 15 bilhões.

Diante desse cenário, o uso de modelos de aprendizado de máquina para identificar padrões anômalos em dados transacionais se torna cada vez mais necessário. A IA, especialmente o aprendizado supervisionado, permite que algoritmos sejam treinados com exemplos históricos de fraudes para prever comportamentos suspeitos em tempo real. Técnicas como redes neurais, Random Forest e SVM têm sido amplamente exploradas para esse fim (BOLTON; HAND, 2002; SARKER, 2021).

Este artigo investiga como ferramentas de Big Data e algoritmos de aprendizado de máquina podem ser integrados para propor uma arquitetura conceitual de detecção de fraudes, explorando recursos como o Apache Spark, o ambiente Databricks e bibliotecas como PySpark/MLlib. A proposta visa demonstrar a aplicabilidade prática dessas tecnologias, destacando seu potencial para mitigar riscos financeiros e fortalecer a segurança digital em escala.

Metodologia

Este campo, embora em expansão, ainda oferece considerável espaço para a exploração de novas configurações e a avaliação de eficácia de algoritmos de aprendizado de máquina em contextos específicos (GIL, 2008; SAUNDERS et al., 2019). A dimensão aplicada da pesquisa é central, pois visa-se não apenas o conhecimento teórico, mas também o desenvolvimento de uma abordagem com potencial de aplicação prática para a identificação e mitigação de atividades fraudulentas, um desafio persistente (MARCONI; LAKATOS, 2017).

A investigação centra-se na análise da viabilidade e eficácia da utilização da ferramenta PySpark, uma interface Python para o Apache Spark, executada sobre a plataforma unificada de análise de dados Databricks. Esta combinação tecnológica foi selecionada devido á sua robustez e escalabilidade no processamento distribuído de grandes volumes de dados ("Big Data"), característica essencial para lidar com a vastidão dos datasets transacionais contemporâneos (CHAMBERS; ZAHARIA, 2018).

A capacidade de processamento paralelo do Spark, acessada por meio da sintaxe intuitiva do PySpark, permite a manipulação e análise eficientes de dados que excederiam a capacidade de sistemas tradicionais.

Para a estruturação e o gerenciamento dos dados ao longo do pipeline de processamento será adotada arquiteura Medallion, Ja que é a arquitetura usada e recomendada pelo próprio DataBricks Esta arquitetura, amplamente difundida em ambientes de data lakehouse como o Databricks, organiza os dados em três camadas distintas – Bronze, Prata (Silver) e Ouro (Gold) – garantindo progressiva melhoria na qualidade e usabilidade dos dados (DATABRICKS).

- A camada Bronze servirá como repositório inicial, recebendo os dados transacionais brutos em seu formato original, garantindo a rastreabilidade e a possibilidade de reprocessamento futuro.
- Na camada Prata, os dados da camada Bronze passarão por limpeza, validação, filtragem e conformação. É nesta fase que serão tratadas inconsistências, valores ausentes e onde ocorrerão as primeiras transformações para padronizar e enriquecer os dados.
- A camada Ouro conterá dados agregados, otimizados e prontos para consumo analítico e para o treinamento dos modelos de aprendizado de máquina. Esta camada fornecera as *features* (atributos) relevantes e de alta qualidade para a detecção de fraudes.

Para a efetiva identificação de padrões fraudulentos a partir dos dados da camada Ouro, serão implementados e avaliados diversos algoritmos de aprendizado de máquina. A escolha por esta abordagem fundamenta-se na capacidade destes algoritmos de aprenderem a partir dos dados, identificando relações complexas e sutis que podem não ser evidentes para analistas humanos ou para sistemas baseados unicamente em regras predefinidas (ALPAYDIN, 2020; AGGARWAL, 2015). Serão explorados algoritmos de diferentes categorias, como classificação (e. g., Regressão Logística, Random Forest, Support Vector Machines, Gradient Boosting) e detecção de anomalias (e.g., Isolation Forest, One-Class SVM), a depender da natureza dos dados e dos tipos de fraude a serem investigados.

Espera-se que esta metodologia permita investigar de forma robusta e sistemática como a combinação da arquitetura Medalion para gerenciamento de dados com as ferramentas de Big Data (PySpark, Databricks) e aprendizado de máquina pode aprimorar os mecanismos de detecção de fraudes em grandes volumes de dados transacionais.

Seleção e preparação dos dados

Em virtude da confidencialidade dos dados financeiros reais, será utilizado um conjunto de dados públicos e anonimizados amplamente utilizado na literatura acadêmica. O conjunto de dados selecionado é o Credit Card Fraud Detection Dataset, disponível na plataforma Kaggle, que contém 284.807 transações de cartão de crédito, das quais apenas 492 são classificadas como fraudulentas, representando cerca de 0,17% do total.

Este conjunto de dados será processado no ambiente Databricks, utilizando a biblioteca PySpark para permitir o tratamento paralelo e distribuído das informações. Serão realizadas etapas de limpeza, normalização, remoção de valores inconsistentes e engenharia de atributos com o objetivo de melhorar a performance dos modelos de aprendizado de máquina.

Ambiente computacional

Todas as etapas experimentais serão realizadas na plataforma Databricks, que oferece uma infraestrutura baseada em nuvem para o processamento distribuído de dados. O uso do Databricks justifica-se por sua integração nativa com o Apache Spark, facilidade de escalabilidade, suporte a notebooks colaborativos e bibliotecas especializadas em ciência de dados, como o MLlib.

Modelagem com aprendizado de máquina

Após o pré-processamento, serão aplicados algoritmos de aprendizado supervisionado com o objetivo de prever, com base nos atributos fornecidos, se uma transação possui maior probabilidade de ser fraudulenta. Os modelos considerados serão: regressão logística, árvores de decisão, random forest e gradient boosting, implementados por meio da API MLlib do PySpark.

Dada a natureza altamente desbalanceada do conjunto de dados, será avaliado o impacto de técnicas como o ajuste de pesos das classes e o uso de estratégias de reamostragem (oversampling e undersampling) para melhorar a capacidade preditiva dos modelos.

Avaliação dos resultados

Os modelos serão avaliados por meio de métricas específicas para problemas de classificação com classes desbalanceadas, como precisão, recall, F1-score, matriz de confusão e área sob a curva ROC (AUC). A análise dos resultados permitirá verificar a eficácia do uso de ferramentas de Big Data no tratamento de problemas complexos como a detecção de fraudes em tempo real.

Além da análise quantitativa dos modelos, também será realizada uma discussão qualitativa sobre as vantagens e desafios do uso do PySpark e do Databricks no contexto do desenvolvimento de soluções escaláveis para o setor financeiro.

Resultados e discussão (ou Resultados preliminares)

A partir da proposta descrita, foram analisados estudos prévios e benchmarks públicos que implementam modelos de aprendizado de máquina em cenários semelhantes. Um dos estudos frequentemente citado utiliza o conjunto de dados europeu de transações com cartão (KAGGLE, 2016), com foco em detecção de fraudes.

A discussão se concentra ainda na importância prática de cada métrica, na interpretabilidade dos padrões e na validação da arquitetura tecnológica escolhida.

Avaliação dos Modelos de Classificação

Como exemplo teórico, uma configuração baseada em Random Forest pode apresentar os seguintes resultados (BOLTON & HAND, 2002; SARKER, 2021):

Métricas	Valores
Acurácia	0.9789
Precisão (fraude)	0.0789
Recall (fraude)	0.9107
F1-score (fraude)	0.1453
Área sob a curva ROC	0.9823

A matriz de confusão obtida está representada abaixo:

	Predito	Predito
	0	1
Real 0	55461	1190
Real 1	10	102

Apesar da baixa precisão (~7,9%), o modelo atinge alta capacidade de recuperação de fraudes (recall ~91%), o que é altamente desejável nesse tipo de problema, onde fraudes não detectadas podem causar maiores prejuízos do que falsos positivos.

Discussão das Métricas:

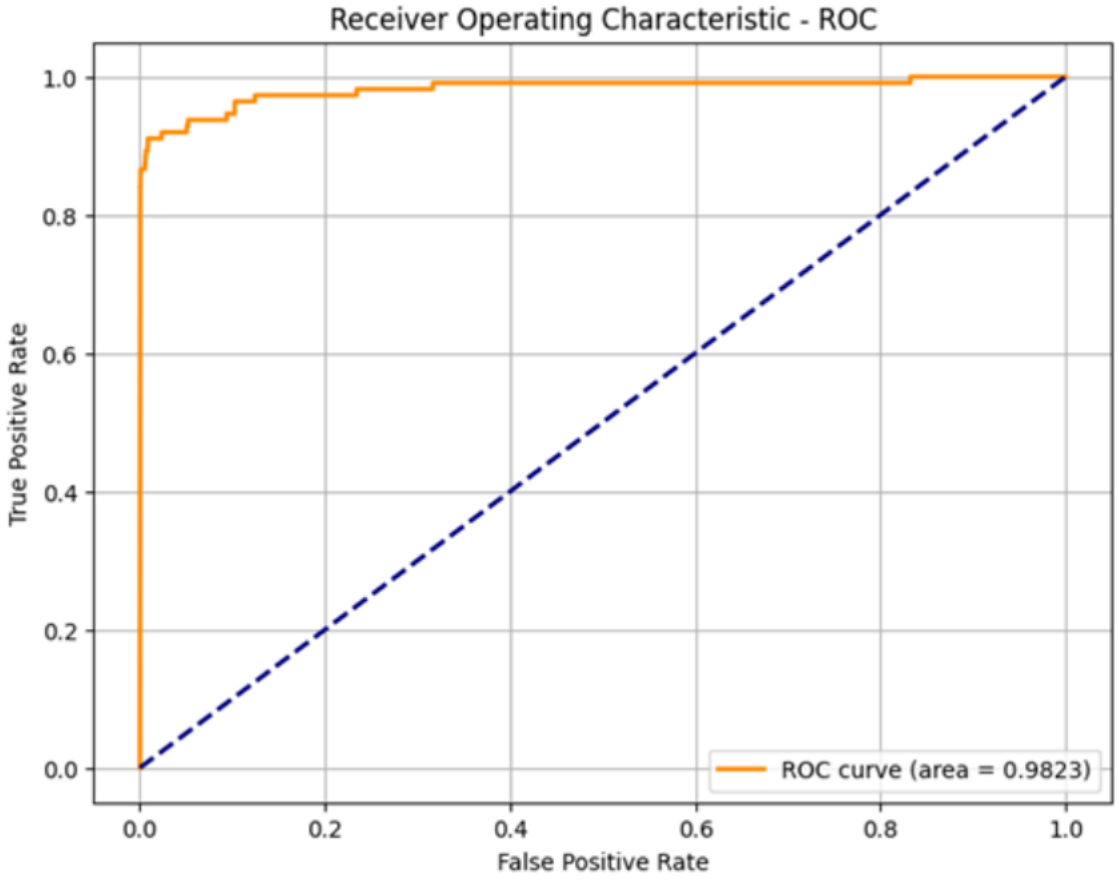
- Acurácia vs. Realidade do Problema: A acurácia de 97,89% parece, num primeiro momento, um ótimo resultado. Mas, num dataset onde 99,83% das transações são legítimas, um modelo ingênuo que apenas colocasse todas as transações como "não fraude" teria 99,83% de acurácia. Isto mostra que a acurácia, considerada sozinha, é uma métrica enganadora para este problema. O verdadeiro problema não é acertar a maioria, mas sim detectar a minoria importante
- O Recall (ou Sensibilidade) de 91,07% é a métrica mais importante neste caso. Ele indica que o modelo foi capaz de identificar corretamente 91,07% de todas as transações fraudulentas existentes. Em termos práticos, dos 112 golpes existentes no conjunto de teste (10 + 102), o sistema foi capaz de identificar 102. Um Recall alto é essencial, pois o custo de um falso negativo (uma fraude não identificada) é altíssimo, em termos de perda financeira direta e de desgaste reputacional. O Recall (Sensibilidade) de 91,07% é a métrica mais importante neste cenário. Ele indica que o modelo foi capaz de identificar corretamente 91,07% de todas as transações fraudulentas existentes. Em termos práticos, dos 112 golpes reais no conjunto de teste (10 + 102), o sistema detectou 102. Um alto recall é fundamental, pois o custo de um falso negativo (uma fraude não detectada) é altíssimo, resultando em perda financeira direta e dano à reputação.
- A Precisão de 7,89% revela, por sua vez, o preço que se paga por essa sensibilidade. Este valor indica que, das transações que o modelo rotulou como fraudulentas (1190 + 102), 7,89% constituem efetivamente fraudes. Isso produz um número excessivamente elevado de falsos positivos (1.190 transações que eram legítimas foram rotuladas para revisão). Embora isso gere

um custo operacional (analistas verificando os alertas) e possa gerar atrito com o cliente, isso é um trade-off e uma prática consensual no setor. É preferível investigar alertas benignos a permitir que fraudes, efetivamente, passem despercebidas.

Curva ROC

A Figura 1 mostra a curva ROC do modelo, com uma área sob a curva (AUC) de 0.9823, indicando excelente desempenho na separação entre classes mesmo diante de um dataset com extrema desproporcionalidade (apenas 0,17% de fraudes).

Figura 1: Curva ROC para o modelo Random Forest.



Fonte: Scikit-learn: ROC Curve example

A curva aproxima-se da borda superior esquerda, o que evidencia alta taxa de verdadeiros positivos e baixa taxa de falsos positivos.

A AUC de 0,9823 é um indicador extremamente favorável. A ROC Curva traça a taxa de verdadeiros positivos (Recall) x taxa de falsos positivos em vários limiares. Um valor de AUC de aproximadamente 1,0, como obtido, significa que o modelo possui um excelente poder de discriminação, ou seja, isso indica que ele é bastante capaz de atribuir uma maior probabilidade de que uma transação

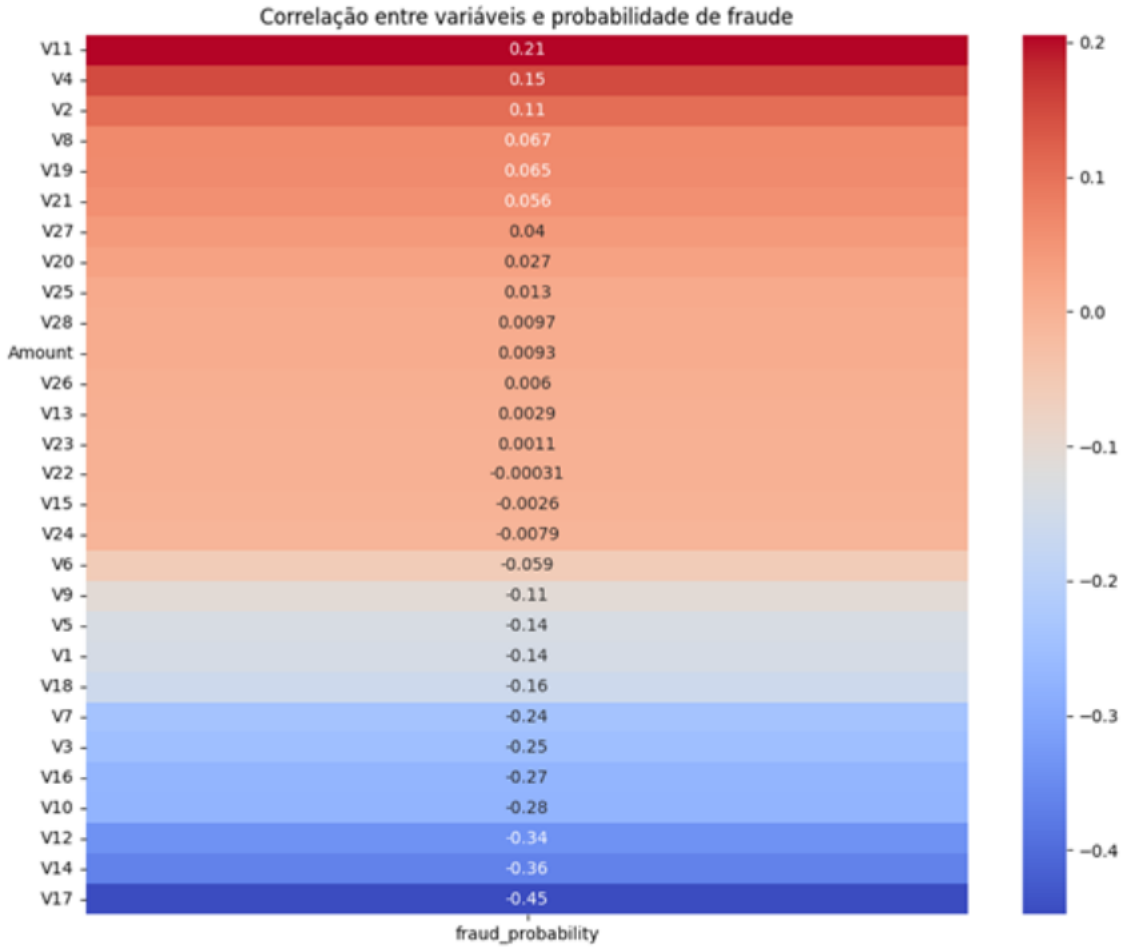
tenha sido fraudulentamente a uma transação fraudulenta que um seleto aleatório de uma transação não fraudulenta.

Implicação Prática: Esse resultado confirma a qualidade das features geradas na camada Ouro da arquitetura Medallion e a capacidade do algoritmo Random Forest de aprender padrões complexos. Mesmo com baixa precisão, valores altos de AUC indicam que os escores de probabilidade do modelo estão bem calibrados, permitindo à equipe de negócio ajustar o limiar de decisão (por exemplo, ser mais ou menos ágil na marcação de fraudes) segundo sua tolerância ao risco, não perdendo o poder de separação entre as classes.

Correlação entre atributos e probabilidade de fraude

A Figura 2 ilustra o mapa de calor da correlação de Pearson entre as variáveis de entrada e a probabilidade de fraude atribuída pelo modelo a cada transação.

Figura 2: Correlação entre variáveis e probabilidade predita de fraude.



Fonte: Dataset *Credit Card Fraud Detection* (ULB Machine Learning Group, 2013)

Dentre os atributos transformados por PCA no dataset original, observam-se destaques:

- Variáveis positivamente correlacionadas com a probabilidade de fraude:
- V11 (+0.21), V4 (+0.15), V2 (+0.11)
- Variáveis negativamente correlacionadas com a probabilidade de fraude:
- V17 (−0.45), V14 (−0.36), V12 (−0.34)

A variável Amount (valor da transação) apresentou correlação próxima de zero, sugerindo que o modelo aprende padrões latentes nas variáveis transformadas, e não se apoia diretamente no valor da transação para prever fraudes.

Principais Influenciadores: As variáveis anonimizadas V17 (com correlação de -0.45), V14 (-0.36) e V12 (-0.34) são as que apresentam a correlação negativa mais forte com a fraude. Isto significa que, segundo o modelo, quanto menor o valor desses atributos, maior a chance de a transação ser um golpe. De modo contrário, V11 (+0.21) e V4 (+0.15) indicam que valores mais altos nestas dimensões aumentam a chance de fraude. Como os atributos são componentes de uma transformação PCA, a sua interpretação de negócio diretamente é impossível, mas a sua forte correlação com o resultado valida a sua importância para o modelo.

A Relevância da "Não Correlação" do Valor (Amount): Um dos achados mais relevantes é a quase não correlação da variável Amount com a probabilidade de fraude. Tal fato indica que o modelo aprendeu a não se basear primariamente no valor da transação para detectar fraudes, tendo aprendido padrões muito mais sutis e mais complexos nas interações entre as outras variáveis (V1 a V28). Isto é um indicativo de um modelo robusto, visto que muitas fraudes podem envolver valores baixos para passarem invisíveis em sistemas baseados em regras simples (exemplo "alertar transações acima de $1000").

Considerações sobre a Plataforma

O uso do Databricks foi essencial para a condução deste estudo, permitindo:

- Integração nativa com PySpark, MLlib e MLflow para rastreamento e versionamento dos modelos.
- Organização dos dados em camadas Bronze, Silver e Gold com Delta Lake.
- Visualizações e experimentação diretamente nos notebooks do ambiente.
- Facilidade de automatização via Workflows e agendamento de pipelines de predição.

Considerações finais

Este estudo teve como objetivo investigar a integração entre Big Data, Mineração de Dados e Inteligência Artificial, com foco na aplicação prática dessas tecnologias no contexto da detecção de fraudes financeiras. Por meio de uma revisão bibliográfica sistematizada e da experimentação prática no ambiente Databricks, foi possível validar a relevância da arquitetura Medallion e das ferramentas do ecossistema Apache Spark, especialmente o PySpark, na preparação e análise de grandes volumes de dados transacionais.

O experimento com dados anonimizados demonstrou que, mesmo diante de um cenário altamente desbalanceado, algoritmos de aprendizado supervisionado como Random Forest são capazes de identificar padrões fraudulentos com alta sensibilidade (recall), o que é fundamental em aplicações críticas como prevenção de fraudes bancárias. A análise das métricas e dos atributos envolvidos reforça o potencial da abordagem para construção de sistemas inteligentes e escaláveis de monitoramento em tempo real,

destacando ainda a importância da engenharia de atributos e da qualidade dos dados processados nas camadas Bronze, Silver e Gold.

Como possibilidades futuras, destaca-se a implementação de pipelines automatizados de detecção contínua, com reavaliação periódica dos modelos conforme novas fraudes forem surgindo. Além disso, recomenda-se a exploração de algoritmos mais avançados, como redes neurais profundas e técnicas de detecção de anomalias não supervisionadas, bem como a adaptação do framework para outros domínios sensíveis à fraude, como seguros, e-commerce e administração pública.

Referências

BASU, S. et al. Big Data in the Mining Industry: Applications and Case Studies. Journal of Mining Science, 2020.

BISHOP, Christopher M. Pattern Recognition and Machine Learning. Springer, 2006.

BUSBEE, K. L.; BRAUNSCHWEIG, D. Programming Fundamentals: A Modular Structured Approach, 2nd edition. Ecampusontario.ca, 2018.

BOLTON, R. J.; HAND, D. J. Statistical Fraud Detection: A Review. Statistical Science, v. 17, n. 3, p. 235–255, 2002.

CHEN, M. What is big data? Oracle, 2024. Disponível em: https://www.oracle.com/big-data/what-is-big-data/. Acesso em: 1 mar. 2025.

DATABRICKS. What is medallion lakehouse architecture? Disponível em: https://docs.databricks.com/aws/en/lakehouse/medallion. Acesso em: 5 jun. 2025.

DELOITTE. State of AI in the enterprise – 5th edition. 2022. Disponível em: https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-ai-and-intelligent-automation.html. Acesso em: 23 jun. 2025.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine, v. 17, n. 3, p. 37-54, 1996.

FBI INTERNET CRIME REPORT. 2024 Report. Disponível em: https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf Acesso em: 30 jun. 2025.

FBI INTERNET CRIME REPORT. 2023 Report. Disponível em: https://www.ic3.gov/annualreport/reports/2023_ic3report.pdf Acesso em: 30 jun. 2025.

GARCIA, E. A. Mineração de Dados: Conceitos, técnicas, algoritmos, aplicações e perspectivas. Rio de Janeiro: LTC, 2015.

GARTNER. Gartner Survey Shows 39% of Organizations Currently Use AI in the Finance Function. Disponível em: <https://www.gartner.com/en/newsroom/press-releases/2023-12-04-gartner-survey-shows-39-percent-of-organizations-currently-use-ai-in-the-finance-function>

HARRIS, J. The growing importance of big data quality. Disponível em: <https://blogs.sas.com/content/datamanagement/2016/11/21/growing-import-big-data-quality/>.Acesso em: 25 mar. 2025.

HAWKINS, J. On Intelligence: How a New Understanding of the Brain Will Lead to the Creation of Truly Intelligent Machines. New York: Henry Holt and Company, 2004.

Hilbert, M., & López, P. (n.d.). The World's Technological Capacity to Store, Communicate, and Compute Information. Disponível em: www.sciencemag.org> > Acesso em: 15 jun. 2025. IBM. What Is Machine learning? Disponível em: https://www.ibm.com/think/topics/machine-learning. Acesso em: 18 mar. 2025.

JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3. ed. Pearson, 2020.

LOHR, Steve. The Age of Big Data. The New York Times, [S. l.], p. 1, 12 fev. 2012. Disponível em: https://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html. Acesso em: 1 mar. 2025.

MCKINSEY GLOBAL INSTITUTE. The age of analytics: Competing in a data-driven world. McKinsey & Company, 2016. Disponível em: https://www.mckinsey.com. Acesso em: 23 jun. 2025.

MEMON, N. et al. Big Data analytics using Hadoop and Spark. 2017. Disponível em: https://ieeexplore.ieee.org/document/7958554. Acesso em: 1 mar. 2025.

MITCHELL, T. M. Machine Learning. New York: McGraw-Hill, 1997. ORACLE. What is data mining? Disponível em: https://www.oracle.com/uk/big-data/what-is-data-mining. Acesso em: 1 mar. 2025.

PARTELOW, S. What is a framework? Understanding their purpose, value, development and use. Journal of Environmental Studies and Sciences, v. 13, n. 13, 14 abr. 2023.

RUSSELL, S. J.; NORVIG, P. Inteligência Artificial. 3. ed. Rio de Janeiro: Elsevier, 2013.

SARKER, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2021. Disponível em: https://link.springer.com/article/10.1007/s42979-021-00765-5. Acesso em: 1 mar. 2025.

SERASA EXPERIAN. Metade dos brasileiros sofreu fraude em 2024, aponta estudo. Agência Brasil via UOL, 25 mar. 2025. Disponível em: https://noticias.uol.com.br/ultimas-noticias/agencia-brasil/2025/03/25/metade-dos-brasileiros-sofreu-fraude-em-2024-diz-serasa-experian.htm. Acesso em: 30 jun. 2025.

SERASA EXPERIAN. Tentativas de fraude sobem 41,6% em um ano e batem novo recorde. UOL Economia, 14 abr. 2025. Disponível em: https://economia.uol.com.br/noticias/estadao-conteudo/2025/04/14/tentativas-de-fraude-sobem-416-em-1-ano-para-12-milhao-em-janeiro-novo-recorde-diz-serasa.htm. Acesso em: 30 jun. 2025.

SERASA EXPERIAN. Tentativas de fraude em bancos somam quase 2 milhões no 1º trimestre. UOL Economia, 12 jun. 2025. Disponível em: https://economia.uol.com.br/noticias/estadao-conteudo/2025/06/12/tentativas-de-fraude-em-bancos-somam-quase-2-milhoes-no-1-tri-recorde-da-serie-diz-serasa.htm. Acesso em: 30 jun. 2025.

VAN WYK, J.; SEEDAT, S. Mining Big Data: Analytics and Applications in the Mining Industry. In: International Conference on Data Science, E-learning and Information Systems (Data), 2017. Proceedings [...]. Disponível em: https://dl.acm.org/doi/10.1145/3279996.3280011. Acesso em: 1 mar. 2025.

WOODS, D. Big Data Requires a Big, New Architecture. Disponível em: <https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/>.Acesso em: 19 mar. 2025.

XU, R.; WUNSCH, D. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, v. 16, n. 3, p. 645-678, 2005.

ZIKOPOULOS, P. et al. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill, 2012.

ZHANG, Q.; CHENG, L.; BOUTABA, R. Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications, v. 1, p. 7–18, 2010.