Ciência de Dados Aplicada: Um Estudo Guiado por Desafios do Kaggle

Henrique Coelho dos Santos[1], **Felipe Barradas Sebastião**[1], **Darlon Vassata**[1], **Thiago Berticelli Ló**[1]

[1]Instituto Federal do Paraná (IFPR)
85814-800 – Cascavel – PR – Brazil

{henriquecoelhodossantos2007,febarradas13}@gmail.com

{darlon.vasata,thiago.lo}@ifpr.edu.br

Abstract. *With the significant increase in data generation across various sectors, new opportunities have emerged for predictions and analyses that support evidence-based decision-making. In this context, the transformation of raw data into relevant information has become a core competency of Data Science. The wide range of techniques available for building predictive models makes it essential to master all stages of the analytical process, from data collection and preparation to modeling and result interpretation. This work is a practical study guided by real-world challenges from the Kaggle platform, aiming to document and apply a structured methodology for developing Data Science projects. For this purpose, the OSEMN cycle (Obtain, Scrub, Explore, Model, Interpret) is adopted. Through case studies involving different types of predictive tasks, such as predicting customer churn in a bank (classification) and predicting house prices in the United States (regression), the procedures adopted in each phase of the cycle are presented, with emphasis on replicable practices and critical interpretations. In addition to highlighting common challenges and pitfalls in the process, this work serves as a support material for projects in educational or competitive environments.*

Resumo. *Com o aumento expressivo da geração de dados em diferentes setores, surgem novas oportunidades para previsões e análises que auxiliam na tomada de decisão baseada em evidências. Nesse contexto, a transformação de dados brutos em informações relevantes tornou-se uma competência central da Ciência de Dados. A diversidade de técnicas disponíveis para construção de modelos preditivos torna essencial o domínio de todas as etapas do processo analítico, desde a coleta e preparação dos dados até a modelagem e interpretação dos resultados. Este trabalho é um estudo prático guiado por desafios reais da plataforma Kaggle, com o objetivo de documentar e aplicar uma metodologia estruturada de desenvolvimento de projetos em Ciência de Dados. Para isso, adota-se o ciclo OSEMN (Obtain, Scrub, Explore, Model, Interpret). Através de estudos de caso envolvendo diferentes tipos de tarefas preditivas, como a predição da evasão de clientes em um banco (classificação) e a predição do preço de casas nos Estados Unidos (regressão), são apresentados os procedimentos adotados em cada fase do ciclo, com ênfase em práticas replicáveis e interpretações críticas. Além de evidenciar os desafios e arma-*

dilhas comuns ao processo, este trabalho oferece um material de apoio para projetos em ambientes educacionais ou competitivos.

1. Introdução

Nas últimas décadas, o mundo foi marcado por uma explosão no volume, variedade e velocidade com que dados são gerados. Essa nova era, impulsionada pela digitalização de serviços, pela internet das coisas (IoT) e pela computação em nuvem, demanda ferramentas e metodologias capazes de transformar dados brutos em *insights* valiosos. Essa capacidade torna-se crucial não apenas para organizações que buscam vantagens competitivas, mas também para governos, instituições de pesquisa e sociedade civil, que se beneficiam da análise de dados para a tomada de decisões baseadas em evidências. Nesse contexto, emerge a Ciência de Dados (*Data Science* - DS) como uma disciplina essencialmente interdisciplinar, que integra estatística, Aprendizado de Máquina (*Machine Learning* - ML), programação, engenharia de dados e conhecimento de domínio para extrair valor de grandes volumes de dados [Provost and Fawcett 2016].

Diante desse cenário, o profissional de DS assume um papel estratégico, sendo responsável por conduzir todas as etapas do ciclo analítico: da coleta e pré-processamento dos dados à modelagem preditiva, interpretação de resultados e comunicação com os tomadores de decisão. No entanto, o domínio desse processo não se dá apenas pelo acúmulo teórico, mas sobretudo pela experiência prática em situações reais. A diversidade de problemas e dados disponíveis demanda um repertório flexível e adaptável de técnicas, aliado a uma compreensão crítica das limitações dos modelos e das armadilhas comuns da inferência estatística [Cleveland 2001].

A complexidade crescente dos desafios enfrentados pela DS exige uma formação sólida e aplicada, que prepare o estudante não apenas para reproduzir algoritmos, mas para compreender profundamente suas decisões metodológicas e suas consequências. Diante disso, plataformas como [1]*Kaggle*, [2]*DrivenData*, [3]*Zindi* e [4]*AICrowd* têm se consolidado como ambientes relevantes para a experimentação e o aprimoramento de habilidades ao disponibilizarem Bancos de Dados (*Databases* - DB) e tarefas com métricas de avaliação objetivas, elas funcionam como verdadeiros laboratórios de experimentação em DS. Competências como limpeza e transformação de dados, engenharia de atributos, escolha e ajuste de modelos, validação cruzada e interpretação de resultados tornam-se habilidades refinadas através de ciclos iterativos de tentativa e erro orientados a objetivos claros.

Embora compartilhem propostas semelhantes, essas plataformas variam em escopo e foco temático. Por exemplo, enquanto a DrivenData tende a enfatizar aplicações de impacto social, a Zindi concentra-se em problemas pertinentes ao continente africano e a AIcrowd costuma abordar desafios em áreas como robótica e jogos. O Kaggle, por sua vez, destaca-se pela ampla comunidade global, diversidade de problemas e documentação colaborativa, sendo, portanto, a plataforma utilizada para este trabalho. Sua base consolidada de competições públicas permite a reprodução de projetos com dados acessíveis e

[1]https://www.kaggle.com/
[2]https://www.drivendata.org/about/
[3]https://zindi.africa/about
[4]https://www.aicrowd.com

bem documentada, promove a reprodutibilidade e aprofundamento técnico.

O presente trabalho documenta e explora os métodos utilizados no desenvolvimento de projetos em DS, com base no ciclo Obter, Limpar, Explorar, Modelar e Interpretar (*Obtain, Scrub, Explore, Model, and Interpret* - OSEMN) , amplamente difundido na comunidade [Dineva et al. 2018]. Por meio de estudos de caso aplicados a competições do Kaggle, é apresentado de forma replicável as etapas essenciais do processo analítico, construindo um manual com o básico para a iniciação de um projeto de DS. Os desafios escolhidos envolvem a aplicação do ciclo OSEMN em conjuntos de dados com distintas naturezas de tarefas preditivas supervisionadas. O primeiro, uma competição de Regressão (predição do preço de casas nos Estados Unidos), exige que o modelo preveja um valor numérico contínuo. O segundo é uma competição de Classificação (predição da evasão de clientes em um banco), cujo objetivo é atribuir um item a uma categoria predefinida (evadir ou não evadir). Essa diversidade entre a predição de valor contínuo e de categoria discreta permite uma análise crítica aprofundada de aspectos como generalização, viés de amostragem e robustez dos modelos preditivos.

A metodologia adotada segue uma abordagem *hands-on*, com foco na aplicação direta de conceitos em problemas reais. A escolha do ciclo OSEMN se justifica por sua simplicidade conceitual e abrangência prática, facilitando internalizar uma rotina de trabalho coerente e reutilizável em diferentes contextos [Shameti 2024]. Cada etapa é detalhada com exemplos práticos e ferramentas específicas, como as bibliotecas em Python (Pandas, Scikit-learn, Seaborn, entre outras) [VanderPlas 2023], além de métricas de avaliação e técnicas de validação de modelos.

2. Fundamentação Teórica

A DS emerge como uma evolução de campos consolidados da estatística, impulsionada pela necessidade de analisar e extrair valor de um volume crescente de dados [Provost and Fawcett 2016] [Cleveland 2001]. Seu principal objetivo é extrair *insights* e conhecimento para fundamentar a tomada de decisões. Isso é alcançado por meio do tratamento e análises automatizadas, permitindo gerar conclusões robustas mesmo a partir de dados incompletos.

Para atingir essa meta, a DS atua em três frentes principais: exploração, previsão e inferência. A exploração consiste na identificação de padrões e tendências; identificação de anomalias; entendimento das variáveis e formulação das hipóteses iniciais. A previsão utiliza esses padrões para estimar valores futuros. Por fim, a inferência dedica-se a qualificar o grau de certeza das conclusões e hipóteses levantadas [Provost and Fawcett 2016].

Devido a sua natureza interdisciplinar, e por possuir uma forte intersecção com outras áreas, como a Inteligência Artificial (*Artificial Intelligence* - IA), o ML, a Aprendizagem Profunda (*Deep Learning* - DL) e a Mineração de Dados (*Data Mining* - DM) frequentemente causam confusão sobre duas definições e escopo. A Figura 1 apresenta a inter-relação entre essas áreas.

A IA, ou inteligência computacional, é um objeto de estudo tanto da psicologia e filosofia, ao discorrer sobre o que são processos cognitivos, quanto da computação, ao utilizar agentes computacionais que raciocinam sobre informações do ambiente, criam conhecimento a partir de experiências passadas e agem adequadamente no ambiente

[Poole et al. 1998]. O ML está contido em IA, pois trata dos modelos matemáticos utilizados para ensinar à máquina como lidar com os dados de maneira eficiente, sem que especifique-se como deve o fazer [Mahesh et al. 2020]. A DL é um subconjunto dentro do ML que permite que modelos computacionais aprendam a partir de dados complexos. A DM, por sua vez, tem como objetivo extrair conhecimento útil dos dados. O termo é uma analogia à mineração convencional, na qual se buscam padrões valiosos em meio a grandes volumes de informação, como DB [Kulin et al. 2021].

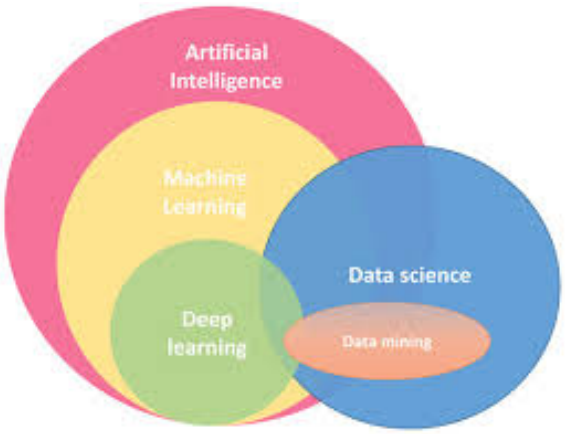

Figura 1. Data science vs. data mining vs. Artificial Intelligence (AI) vs. Machine learning (ML) vs. deep learning. Fonte: [Kulin et al. 2021]

Apesar de suas particularidades, essas áreas compartilham conceitos fundamentais, com destaque para duas abordagens de aprendizado: o aprendizado supervisionado e o aprendizado não supervisionado. O aprendizado supervisionado é treinado com conjuntos de dados rotulados, i.e., cada entrada terá uma saída conhecida. Seu objetivo é compreender as relações entre atributos para prever as saídas de novos dados. Dentro desta abordagem, destacam-se duas tarefas principais: a regressão e a classificação. A regressão visa prever um valor numérico contínuo. Por exemplo, descobrir o preço de casas com base nas suas características. Já a classificação, tem o objetivo de atribuir um item a uma categoria predefinida. Tal como, determinar se um e-mail é spam ou não [Provost and Fawcett 2016]. No aprendizado não supervisionado, o modelo analisa dados não rotulados para descobrir padrões, estruturas ou agrupamentos (clusters) de forma autônoma, sem um resultado esperado previamente definido.

2.1. Ciclo de vida de um projeto de Ciência de Dados

Para atender à necessidade do mercado por soluções com DS que fossem adaptativas, consistentes, objetivas e reiteráveis, foram desenvolvidas abordagens com etapas fundamentais para o processo de trabalho com dados e computação, tais como o Processo Padrão de Indústria Cruzada para Exploração de Dados (*CRoss-Industry Standard Process for Data Mining* - CRISP-DM) e o OSEMN [Shameti 2024]. Nesse artigo adotou-se o termo ciclo de vida de projetos de DS, por este abranger o processo de ponta a ponta, desde o entendimento do problema, até a avaliação da solução final. Essa escolha se justifica porque o termo “projeto de software” é inadequado, visto que o trabalho com dados se assemelha mais a uma pesquisa do que um artefato de software [Provost and Fawcett 2016].

O CRISP-DM é dividido em seis etapas sendo elas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação. A primeira etapa consiste no entendimento do problema e das especificidades do objeto de trabalho, além do estabelecimento de objetivos com uma linguagem natural para a DS. A compreensão dos dados começa na aquisição e familiarização da informação para identificar problemas, possíveis correlações ou subgrupos para criar hipóteses sobre eles. Na etapa de preparação dos dados, constrói-se o conjunto final de informações que será enviado para a modelagem futura, cabe-se dizer que ela ocorre durante todo o projeto e não somente após a segunda etapa. A modelagem do processo é o estágio de escolha dos modelos e refinação de seus parâmetros. Em seguida, é feita a avaliação dos modelos para assegurar que atendem aos objetivos do projeto e medir a precisão alcançada. Finalmente, a implementação do projeto varia em complexidade e, muitas vezes, é feita não pelo cientista de dados, mas sim pelo cliente, podendo ser algo básico, como a construção de relatórios com o modelo, até a aplicação em sistemas complexos de hardware [Shameti 2024].

Como forma de simplificar e fornecer uma solução mais objetiva e moderna, foi proposto o ciclo OSEMN, que organiza o fluxo de trabalho com ações subsequentes claras. A primeira etapa, Obter (*Obtain*), é essencial, pois sem dados não é possível realizar produções no campo da DS. Essa fase consiste na reunião e coleta de dados provenientes de diferentes fontes, que serão posteriormente utilizados nas análises. Em seguida, na etapa de Limpar (*Scrub*), busca-se tornar os dados utilizáveis, filtrando linhas e colunas e aplicando transformações necessárias, como a remoção de valores nulos, duplicatas ou inconsistências. A terceira etapa, Explorar (*Explore*), baseia-se na compreensão da estrutura dos dados obtidos, na identificação de padrões, tendências e possíveis relações entre variáveis que possam influenciar os resultados desejados. Na etapa de Modelar (*Model*), os dados previamente tratados são utilizados para a construção de modelos estatísticos ou de aprendizado de máquina, com o objetivo de gerar previsões ou classificações. Nessa fase, o pesquisador busca configurar os parâmetros dos modelos de forma que o desempenho seja maximizado em relação aos objetivos definidos. Por fim, na etapa de Interpretar (*Interpret*), os resultados gerados são analisados, discutidos e apresentados, transformando as descobertas obtidas ao longo do processo em informações úteis, muitas vezes com valor aplicado em contextos comerciais, operacionais ou sociais [Shameti 2024].

2.2. Ciência de Dados competitiva

A DS competitiva faz referência à programação competitiva, mas com foco nos desafios da área de DS. Ela vem ganhando popularidade ao decorrer dos últimos anos e aumentando o nicho dentre os cientistas de dados, inspirando empresas e universidades à hospedar e fornecer grandes bancos de dados para solucionar problemas de seus interesses. Por exemplo, a Netflix lançou, em 2006, uma competição para prever a nota que um usuário daria para filmes de acordo com seu histórico de avaliações, incentivando o desenvolvimento de abordagens inteligentes e criativas para com os dados disponíveis. Nesse contexto, diversas plataformas que promovem competições a fim de incentivar o estudo e a inovação na área da DS foram criadas e ganharam espaço no mercado. Dentre essas plataformas pode-se destacar a Kaggle, a DrivenData, a Zindi e a AIcrowd [Banachewicz et al. 2022].

Diversos estudos da DS se baseiam em competições hospedadas por plataformas

como a Kaggle, documentando as ferramentas e técnicas utilizadas para cada estudo de caso. Para fomentar a pesquisa e participação o livro [Banachewicz et al. 2022] foi criado e é relevante, pois introduz os estudantes aos conceitos básicos desses campeonatos. Além disso, a documentação de projeto desenvolvidos em desafios no Kaggle é uma prática comum, como pode ser exemplificado em [Jang et al. 2017] no qual os autores utilizam dois casos de estudo e discorrem das metodologias que foram úteis ao decorrer de todo o processo da competição, desde a exploração até os resultados. Outro exemplo é o trabalho de Lu et al. (2021) [Lu et al. 2021], que teve como objetivo construir um modelo para prever dados. Neste trabalho o autor utilizou da metodologia CRISP-DM para documentar a modelagem de um caso real visando definir melhores estratégias comerciais. Portanto, as competições para criação de modelos de predições de dados são uma tendência do mercado, tanto para incentivar cientistas de dados,quanto para criar soluções inovadoras úteis no contexto global.

2.3. Métricas de Avaliação

A avaliação adequada de modelos preditivos é fundamental para garantir que as soluções desenvolvidas atendam aos objetivos propostos. Diferentes tipos de problemas demandam métricas específicas:

2.3.1. Para Classificação

- **Acurácia (*Accuracy*):** Proporção de predições corretas sobre o total de amostras avaliadas. É simples de interpretar, mas pode ser inadequada quando há desbalanceamento entre classes, pois ignora a distribuição dos rótulos.
- **Precisão (*Precision*):** Proporção de verdadeiros positivos entre todas as instâncias classificadas como positivas pelo modelo. É especialmente relevante em cenários onde o custo de um falso positivo é alto, como em detecção de fraudes.
- **Revocação (*Recall/Sensitivity*):** Proporção de verdadeiros positivos identificados entre todos os casos realmente positivos. É essencial quando o custo de um falso negativo é alto, como em diagnósticos médicos ou detecção de falhas.
- **F1-Score:** Média harmônica entre precisão e revocação. Fornece uma medida equilibrada do desempenho quando é necessário considerar simultaneamente ambos os erros (falsos positivos e falsos negativos), sendo útil em datasets desbalanceados.
- **Área Sob a Curva ROC (*Area Under the ROC Curve – AUC-ROC*):** Mede a capacidade do modelo de distinguir entre classes ao longo de todos os limiares de decisão possíveis. Valores próximos de 1 indicam excelente separação entre classes; valores próximos de 0,5 indicam desempenho equivalente ao acaso.

2.3.2. Para Regressão

- **Erro Absoluto Médio *(Mean Absolute Error – MAE)*:** Média dos valores absolutos das diferenças entre predições e valores reais. É robusto a outliers e mantém a interpretação direta na mesma unidade da variável alvo.

- **Raiz do Erro Quadrático Médio *(Root Mean Squared Error – RMSE)*:** Raiz quadrada da média dos erros ao quadrado. Penaliza desvios maiores com mais intensidade, sendo útil quando grandes erros são particularmente indesejáveis.
- **Coeficiente de Determinação (R^2):** Mede a proporção da variância da variável dependente explicada pelo modelo. Varia entre 0 e 1 para modelos usuais, mas pode assumir valores negativos quando o modelo é pior que uma predição baseada na média.
- **Erro Percentual Absoluto Médio *(Mean Absolute Percentage Error – MAPE)*:** Média dos erros absolutos expressos em porcentagem do valor real. Útil para comparar desempenho entre diferentes escalas, porém sensível a valores reais próximos de zero.

3. Materiais e Métodos

O presente artigo é desenvolvido utilizando a linguagem de programação Python na plataforma *Google Colaboratory* (*Colab*) [5]. As principais bibliotecas utilizadas incluem a Pandas, para a manipulação das tabelas dos dados, a Numpy, para operações numéricas, a Matplotlib para visualização de dados e a Scikit-learn para implementações padronizadas de algoritmos de ML, pré-processamento, validação cruzada e métricas de avaliação. Para a estruturação dos processos dos estudos de caso, foi adotado o ciclo de vida OSEMN, haja vista a simplicidade e objetividade do ciclo. Para garantir a replicabilidade do projeto, cada etapa é documentada nas seções dedicadas aos estudos de caso, com a anexação dos códigos e materiais utilizados.

Foi utilizada a plataforma Kaggle, visando a obtenção dos DB. A escolha dos desafios foi fundamentada em: um estudo de caso com a análise de regressão e outro com a análise de classificação. Dessa forma, o presente estudo diversifica e avalia o que foi mais positivo em dois projetos de predição diferentes dentro dos supervisionados. Portanto, utilizou-se da competição "*Playground Series - Season 4, Episode 1*", a qual o objetivo é classificar se um cliente deixará o estabelecimento bancário (variável binária) e apresenta desafios típicos de classificação, incluindo desbalanceamento de classes e múltiplas variáveis categóricas e numéricas. Além disso foi utilizado o DB "*House Prices - Advanced Regression Techniques*", focado em estimativa de valores de imóveis, em que o objetivo é prever o preço de venda (variável contínua) exemplificando desafios comuns em regressão, como tratamento de outliers, transformações de variáveis e engenharia de atributos complexa.

A limpeza dos dados teve como objetivo eliminar os problemas das DB escolhidas, sendo eles os dados faltantes e os dados inconsistentes. Eles devem ser tratados a fim de evitar erros na modelagem final e na análise estatística. Dados faltantes são aqueles que em linhas específicas das colunas não possuem valor algum, *i.e.*, são nulos. Os dados inconsistentes podem ser classificados como aqueles que representam anomalias dentro dos padrões da DB específica, *e.g.*, erros ortográficos, valores que de ao analisar graficamente estão fora do padrão (*outliers*), entre outras atipicidades [Oliveira et al. 2004]. Para a normalização, foi utilizado estratégias de remoção e preenchimento sendo cada parte da limpeza específica para a DB e a anormalidade encontrada. Logo, as decisões de cada estudo de caso vão ser descritas em suas seções.

[5]https://colab.google/

A etapa de exploração tem o objetivo de entender e transformar os dados em informações, obtendo, por fim, conhecimento. Para isso será feita a investigação do comportamento e relações das variáveis com representações numéricas e gráficas. As medidas de posições e tendências da estatística oferecem características fundamentais na interpretação dos dados. Também se faz útil a análise visual das DB, visto que mesmo com média, variância e correlação idênticos os padrões apresentados em gráficos podem ser diferentes, haja vista os conjuntos dentro do Quartetos de Anscombe. Logo atributos como Média ($\bar{x}$), Moda (Mo), Mediana (Md) e Quartis (Q_k), junto com gráficos de dispersão, diagramas de caixa e histogramas vão oferecer uma compreensão melhor do conjunto total. Além disso, medidas de distribuição vão servir para entender as variações entre os dados, tais como a Amplitude Total (AT), a Variância Populacional (σ), a Variância Amostral (S^2), o Desvio Padrão Populacional (σ), o Desvio Padrão Amostral (S) e o Coeficiente de Variação (CV) [Anscombe 1973].

A etapa de modelagem objetivou a construção de soluções preditivas a partir dos dados previamente tratados e explorados. Foram considerados diferentes métodos estatísticos e computacionais, selecionados conforme a natureza de cada conjunto de dados e o tipo de tarefa envolvida. O processo foi conduzido de forma sistemática, com a definição de estratégias para ajuste dos modelos e avaliação do desempenho obtido em diferentes divisões dos dados, reduzindo o risco de sobreajuste e a possibilidade de conclusões imprecisas. Durante esse processo, serão realizados testes com variações nos parâmetros internos dos modelos, observando-se seus efeitos sobre os resultados, de modo a identificar configurações compatíveis com os dados utilizados. A modelagem será tratada como uma etapa iterativa, na qual as decisões serão orientadas por critérios de avaliação definidos previamente e pelos resultados empíricos obtidos ao longo dos experimentos. As decisões específicas estão registradas nos estudos de caso, respeitando as particularidades de cada base analisada [Caruana and Niculescu-Mizil 2006, Koch et al. 2017].

A etapa de interpretação se dedicou à análise dos resultados obtidos na modelagem. O objetivo central é compreender o comportamento dos modelos frente aos dados avaliando a confiabilidade das previsões geradas. Isso inclui examinar a coerência das saídas com o conhecimento do domínio, investigar possíveis distorções nos padrões aprendidos e verificar o impacto de diferentes variáveis na construção das respostas. Ao longo dessa etapa, buscou-se extrair conclusões que não apenas justifiquem o desempenho alcançado, mas também ofereçam insights relevantes sobre os dados e sua estrutura [Murdoch et al. 2019]. As métricas adotadas na avaliação serão definidas com base no tipo de tarefa e nos critérios estabelecidos para cada problema, respeitando as diretrizes das competições selecionadas. Assim como nas demais fases do ciclo, a etapa de interpretação é documentada detalhadamente, permitindo a replicação e o aperfeiçoamento dos métodos apresentados [Caruana and Niculescu-Mizil 2006].

As etapas descritas ao longo deste trabalho serão retomadas e aprofundadas nas seções específicas dos estudos de caso, onde estão documentados os algoritmos e bibliotecas utilizadas em cada abordagem.

4. Estudo de Caso 1: Predição da Rotatividade de Clientes Bancários

O primeiro estudo de caso investiga um problema clássico de classificação binária no setor bancário: a predição de *churn* (evasão) de clientes. O conjunto de dados utilizado foi disponibilizado na competição *Playground Series - Season 4, Episode 1"*[6] e foi gerado sinteticamente por meio de um modelo de aprendizado profundo treinado sobre o dataset real *Bank Customer Churn Prediction"*. Embora sintético, o conjunto preserva distribuições estatísticas semelhantes às observadas em contextos reais, mantendo a relevância analítica.

O objetivo consiste em prever se um cliente irá encerrar seu relacionamento com o banco com base em 13 atributos descritivos. O dataset de treinamento contém 165.034 registros e o conjunto de teste possui 110.023 instâncias. A métrica de avaliação adotada foi a AUC-ROC, apropriada para contextos com desbalanceamento moderado entre classes.

Os atributos disponíveis incluem:

- **id**: Identificador único do registro
- **CustomerId**: Identificador único do cliente
- **Surname**: Sobrenome do cliente
- **CreditScore**: Pontuação de crédito
- **Geography**: País de residência (França, Alemanha ou Espanha)
- **Gender**: Gênero do cliente
- **Age**: Idade
- **Tenure**: Tempo de relacionamento com o banco (0–10 anos)
- **Balance**: Saldo bancário
- **NumOfProducts**: Número de produtos bancários
- **HasCrCard**: Indicação de posse de cartão de crédito
- **IsActiveMember**: Indicação de cliente ativo
- **EstimatedSalary**: Salário estimado
- **Exited**: Variável alvo (evasão)

Pré-processamento e Análise Exploratória

A primeira etapa da fase *Scrub* incluiu a inspeção da estrutura dos dados via `info()` e `describe()` do Pandas. O conjunto de dados não apresentava valores ausentes, como esperado em datasets sintéticos. Contudo, algumas variáveis numéricas estavam armazenadas como ponto flutuante quando deveriam ser inteiras, incluindo Age, HasCrCard e IsActiveMember. Essas colunas foram convertidas para o tipo *int* por meio do método `astype(int)` para garantir consistência semântica.

A análise exploratória revelou que certos atributos não contribuem para o poder preditivo do modelo. As variáveis id e CustomerId são apenas identificadores e não apresentam relação estrutural com churn. O atributo Surname apresentou elevada cardinalidade (14.516 valores únicos), resultando em baixa utilidade preditiva. Observou-se também que HasCrCard possui impacto mínimo na variável alvo, dado que as taxas de churn entre clientes com e sem cartão (20,5% e 20,4%, respectivamente) são praticamente idênticas. Assim, optou-se por remover id, CustomerId, Surname e HasCrCard, resultando em um conjunto final com 11 colunas (10 características e a variável alvo).

[6] https://www.kaggle.com/competitions/playground-series-s4e1/overview

As variáveis categóricas Geography e Gender foram transformadas por *Target Encoding*, substituindo categorias pela média da variável alvo condicional a cada grupo. As variáveis numéricas foram padronizadas com *StandardScaler*, assegurando médias nulas e desvios padrão unitários, procedimento crucial para modelos sensíveis à escala.

A distribuição da variável Exited evidenciou desbalanceamento moderado: 79,0% dos clientes permaneceram (130.113) e 21,0% evadiram (34.921). Esse cenário reforça a necessidade de métricas robustas como AUC-ROC, que são menos afetadas por distribuições assimétricas.

Padrões Identificados na Exploração de Dados

Histogramas agrupados e normalizados revelaram padrões relevantes:

- **Geografia**: A Alemanha apresentou a maior taxa relativa de churn (32%), superando França (16%) e Espanha (17%). Embora a França concentre maior volume absoluto de clientes, a propensão à evasão é mais pronunciada no público alemão, possivelmente influenciado por fatores regulatórios, competitivos ou culturais.
- **Gênero**: A taxa de churn foi de aproximadamente 25% entre clientes do gênero feminino, contra 16,5% para clientes do gênero masculino, sugerindo que o gênero é um preditor relevante.
- **Número de Produtos**: Clientes com 2 produtos apresentaram a menor taxa de evasão (7,6%). Por outro lado, clientes com 3 ou 4 produtos exibiram taxas superiores a 80%, um comportamento contraintuitivo que pode refletir tentativas tardias de retenção ou insatisfação com produtos adicionais.
- **Status de Atividade**: Clientes inativos apresentaram churn de 26,9%, enquanto clientes ativos apresentaram 14,3%, reforçando o papel do engajamento como fator de retenção.
- **Tenure**: Clientes com tenure igual a 0 exibiram a maior taxa proporcional (27%), indicando um período crítico de adaptação inicial. Entre 1 e 10 anos, as taxas variam de 19% a 22%, sem tendência claramente monotônica.

A análise por diagramas de caixa (*boxplots*) revelou:

- **CreditScore**: Distribuições muito similares entre evasores e não evasores; medianas próximas a 650 e forte sobreposição, indicando baixo poder preditivo isolado.
- **Age**: A idade mediana dos evasores (cerca de 45 anos) é superior à dos não evasores (37 anos), indicando maior propensão de churn entre clientes mais velhos.
- **Balance**: Notou-se forte assimetria: clientes que permaneceram possuem mediana de saldo próxima de zero, enquanto evasores apresentam mediana em torno de 100.000. Clientes com altos saldos parecem migrar para instituições com condições mais vantajosas.
- **EstimatedSalary**: Distribuições praticamente idênticas entre classes, indicando baixo poder discriminativo.

Os dados foram divididos em treinamento e teste na proporção 70/30, com estratificação para preservar a proporção original das classes. O conjunto de treinamento resultante possui 115.523 instâncias e o conjunto de teste, 49.511.

Modelagem e Otimização

Foram considerados inicialmente Random Forest, XGBoost e CatBoost. Optou-se por aprofundar a modelagem com o CatBoost devido ao seu tratamento nativo de variáveis categóricas, resistência a overfitting por meio do *ordered boosting*, eficiência computacional e bom desempenho em cenários com desbalanceamento.

Uma Busca em Grade (*Grid Search*) foi conduzida para identificar hiperparâmetros ideais, explorando combinações de profundidade máxima das árvores e número total de árvores (25 combinações, 5×5), avaliadas via validação cruzada com 5 folds e métrica AUC-ROC. Os melhores resultados foram obtidos com profundidade máxima igual a 10 e 150 árvores.

O modelo final foi treinado com os 115.523 registros do conjunto de treinamento e avaliado no conjunto de teste. As métricas obtidas foram:

- **AUC (Treino)**: 0,8925 (aprox. 89,3
- **AUC (Teste)**: 0,8883 (aprox. 88,8
- **Kaggle (Submissão)**: 0,88750 (aprox. 88,8

A curva ROC demonstra excelente capacidade discriminativa, mantendo-se substancialmente acima da linha de referência de um classificador aleatório. A diferença de cerca de 1,5 ponto percentual entre treino e teste indica leve sobreajuste, mas dentro de limites aceitáveis.

Análise das Matrizes de Confusão

A partir desses valores, foram obtidos:

- **Acurácia**: 92,8
- **Precisão**: 88,0
- **Revocação**: 76,0
- **Especificidade**: 97,3
- **F1-Score**: 81,5

Dentre os padrões observados:

- **Alta Especificidade**: Com 97,3% de verdadeiros negativos, o modelo reduz falsos alarmes e evita mobilização desnecessária de recursos em ações de retenção.
- **Revocação Moderada**: O modelo identifica 76,0% dos casos reais de evasão; embora haja 24% de falsos negativos, o desempenho é coerente com o desbalanceamento da base.
- **Excelente Precisão**: A precisão de 88,0% indica alta confiabilidade quando o modelo sinaliza um cliente como evasor.
- **Coerência Entre Treino e Teste**: A proximidade nos padrões de erro sugere boa capacidade de generalização sem sobreajuste substancial.

Referências

Anscombe, F. J. (1973). Graphs in statistical analysis. *The american statistician*, 27(1):17–21.

Banachewicz, K., Massaron, L., and Goldbloom, A. (2022). *The Kaggle Book: Data analysis and machine learning for competitive data science*. Packt Publishing.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1):21–26.

Dineva, K., Atanasova, T., et al. (2018). Osemn process for working over data acquired by iot devices mounted in beehives. *Curr. Trends Nat. Sci*, 7(13):47–53.

Jang, H., Kim, S., and Lam, T. (2017). Kaggle competitions: Author identification & statoil/c-core iceberg classifier challenge. *Dept. School Inform., Comput., Eng. Indiana Univ., Bloomington, IN, USA, Tech. Rep*.

Koch, P., Wujek, B., Golovidov, O., and Gardner, S. (2017). Automated hyperparameter tuning for effective machine learning. In *proceedings of the SAS Global Forum 2017 Conference*, pages 1–23. SAS Institute Inc. Cary, NC.

Kulin, M., Kazaz, T., De Poorter, E., and Moerman, I. (2021). A survey on machine learning-based performance improvement of wireless networks: Phy, mac and network layer. *Electronics*, 10(3):318.

Lu, J., Cairns, L., and Smith, L. (2021). Data science in the business environment: customer analytics case studies in smes. *Journal of Modelling in Management*, 16(2):689–713.

Mahesh, B. et al. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Oliveira, P., Rodrigues, F., and Henriques, P. (2004). Limpeza de dados-uma visão geral. *Data Gadgets*, pages 39–51.

Poole, D., Mackworth, A., and Goebel, R. (1998). Computational intelligence: a logical approach. 1998. *Google scholar google scholar digital library digital library*.

Provost, F. and Fawcett, T. (2016). *Data Science para Negócios*. Alta Books, 1st edition.

Shameti, K. (2024). *COMPARISON OF METHODOLOGICAL APPROACHES: CRISP-DM VERSUS OSEMN METHODOLOGY USING LINEAR REGRESSION AND STATISTICAL ANALYSIS*. PhD thesis, EPOKA University.

VanderPlas, J. (2023). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly.

	Predito: 0	**Predito: 1**	**Real: 0 (Não Evadiu)**
89.147	2.091	**Real: 1 (Evadiu)**	5.234
19.051			

Tabela 1. Matriz de confusão para o conjunto de treinamento.

	Predito: 0	**Predito: 1**	**Real: 0 (Não Evadiu)**
38.062	1.072	**Real: 1 (Evadiu)**	2.489
7.888			

Tabela 2. Matriz de confusão para o conjunto de teste.