Análise Preditiva de potencial de crescimento de Empresas no Continente Americano: Uma Aplicação de Machine Learning Utilizando Dados Financeiros e Macroeconômicos

Sidnei Alves de Almeida
Orientador: Dr. Fernando Elemar Vicente dos Anjos
Co-orientador: MSc. Ivanielly Deyse de Paiva Moura

Resumo

A análise preditiva tem se consolidado como uma ferramenta essencial no setor financeiro, permitindo a antecipação de tendências e a otimização de operações. Este trabalho tem como objetivo investigar o *potencial de crescimento* de 4.394 empresas das Américas, por meio da construção de *modelos* preditivos baseados em dados financeiros e macroeconômicos. Inicialmente, foi realizada uma análise exploratória de dados (*EDA*) para compreender as interações entre variáveis como receita, lucro, índice P/L, capitalização de mercado, PIB, inflação e taxa de juros. A variável-alvo foi definida de forma não supervisionada, a partir de uma segmentação baseada em *PCA (Análise de Componentes Principais)* combinada com a discretização por quantis (*qcut*). Em seguida, foram aplicados e avaliados diversos modelos de classificação supervisionada, incluindo *Logistic Regression*, *Support Vector Machine (SVM)*, *K-Nearest Neighbors (KNN)*, *Naive Bayes*, *Árvores de Decisão*, *Random Forest*, *Gradient Boosting*, *XGBoost* e *Redes Neurais*, utilizando métricas como acurácia, precisão, recall e *F1-score*. Com base nos resultados obtidos, o *Gradient Boosting* destacou-se por sua robustez e elevado desempenho nas métricas avaliadas, apresentando os melhores indicadores entre os modelos testados. Complementarmente, desenvolveu-se uma *aplicação web* em *Streamlit*, com o objetivo de proporcionar uma interface intuitiva que permita aos usuários realizar novas previsões de forma acessível. A complexidade dos dados, marcada por múltiplos *outliers* e distribuições não normais, exigiu o uso de técnicas estatísticas robustas, reforçando a escolha do *Gradient Boosting* por sua capacidade de gerar previsões precisas. Como sugestão para trabalhos futuros, recomenda-se explorar novas abordagens para definição do *potencial de crescimento* e a adoção de técnicas de modelagem ainda mais modernas e sofisticadas.

Palavras Chave

Análise Preditiva, Crescimento Empresarial, Machine Learning, Dados Macroeconômicos, Modelos de Classificação, Previsão de Crescimento, Inteligência Artificial.

Abstract

Predictive analysis has become an essential tool in the financial sector, enabling the anticipation of trends and the optimization of operations. This study aims to investigate the *growth potential* of 4,394 companies across the Americas by constructing predictive *models* based on financial and macroeconomic data. Initially, an exploratory data analysis (*EDA*) was conducted to understand the interactions between variables such as revenue, profit, P/E ratio, market capitalization, GDP, inflation, and interest rates. The target variable was defined in an unsupervised manner through segmentation based on *PCA (Principal Component Analysis)* combined with quantile discretization (*qcut*). Subsequently, several supervised classification models were applied and evaluated, including *Logistic Regression*, *Support Vector Machine (SVM)*, *K-Nearest Neighbors (KNN)*, *Naive Bayes*, *Decision Trees*, *Random Forest*, *Gradient Boosting*, *XGBoost*, and *Neural Networks*, using metrics such as accuracy, precision, recall, and *F1-score*. Based on the results obtained, *Gradient Boosting* stood out due to its robustness and high performance across the evaluated metrics, achieving the best indicators among all tested models. Additionally, a *web application* was developed using *Streamlit* to provide an intuitive interface that allows users to perform new predictions in an accessible way. The complexity of the data, marked by numerous *outliers* and non-normal distributions, required the use of robust statistical techniques, reinforcing the choice of *Gradient Boosting* for its ability to deliver accurate predictions. As a recommendation for future work, it is suggested to explore new approaches for defining *growth potential* and to adopt even more modern and sophisticated modeling techniques.

Keywords

Predictive Analysis, Business Growth, Machine Learning, Macroeconomic Data, Classification Models, Growth Prediction, Artificial Intelligence.

I. Introdução

A análise preditiva tem se consolidado como uma ferramenta crucial para organizações que almejam antecipar cenários, otimizar recursos e orientar a tomada de decisões estratégicas. Essa abordagem, fundamentada na análise de grandes volumes de dados históricos e na aplicação de algoritmos de *machine learning*, permite a identificação de padrões e a projeção de comportamentos futuros com elevada acurácia. No contexto financeiro, a análise preditiva viabiliza a estimativa do desempenho empresarial e o monitoramento de indicadores-chave, promovendo uma gestão mais informada e eficiente *(Mitchell, 1999)*.

O uso de *machine learning* no setor financeiro tem se destacado por sua eficácia na previsão de indicadores financeiros, como liquidez e rentabilidade, com alta precisão. Essa abordagem identifica padrões complexos e não evidentes, oferecendo uma visão mais detalhada sobre o futuro das empresas. Assim, a aplicação de *modelos* preditivos financeiros tem se consolidado como uma ferramenta estratégica no mercado *(Dio, 2022)*.

Este trabalho objetiva analisar o *potencial de crescimento* de 4.394 empresas localizadas nas Américas, utilizando dados financeiros e macroeconômicos especificamente coletados no ano de 2024 para construir *modelos* preditivos. Estudos como o de *Pinto e Carvalho (2020)* demonstram que modelos baseados em aprendizado de máquina, como o XGBoost, podem superar modelos tradicionais na previsão de risco de crédito, utilizando variáveis financeiras e macroeconômicas. Além disso, *Guimarães (2020)* aplicou técnicas de aprendizado de máquina, incluindo *Redes Neurais Multilayer Perceptron* e *Random Forest*, para prever a produtividade da cultura da soja, destacando a eficácia desses métodos em contextos agrícolas. Nossa análise exploratória de dados (*EDA*) será conduzida para investigar variáveis como *receita*, *lucro*, *índice P/L*, *valor de mercado* e *dividendos*, além de fatores macroeconômicos, como *PIB per capita*, *taxa de crescimento do PIB*, *inflação*, *taxa de juros*, *taxa de câmbio* e *taxa de desemprego*. Dados ausentes ou inconsistentes serão tratados para preparar o conjunto de dados para os *modelos* de previsão.

Os algoritmos aplicados incluirão *Árvores de Decisão*, *Random Forest*, *Support Vector Machines (SVM)*, *Redes Neurais* e modelos baseados em *Boosting* avaliados por métricas como acurácia, precisão, *recall*, *F1-score*, *Matriz de Confusão* e *Curva ROC-AUC* com o objetivo de classificar as empresas de acordo com seu *potencial de crescimento*. A comparação entre os *modelos* permitirá identificar a abordagem mais eficaz para as previsões.

A aplicação será integrada a uma plataforma web desenvolvida com o *framework Streamlit*, que fornecerá uma interface intuitiva para análise e previsão em tempo real. Essa integração facilitará a tomada de decisões estratégicas pelos usuários, aprimorando a análise de dados no setor financeiro e oferecendo uma vantagem competitiva *(Dio, 2022)*.

O trabalho está estruturado da seguinte forma: na **Seção I**, apresenta-se a introdução, com os objetivos e a relevância do estudo. A **Seção II** contém a revisão teórica sobre o uso de *machine learning* nas finanças. Na **Seção III**, detalha-se a metodologia, incluindo a coleta de dados, a análise exploratória de dados (*EDA*), pré-processamento e os *modelos* preditivos. A **Seção IV** apresenta a implementação dos *modelos* e os resultados obtidos, acompanhados de sua avaliação. Por fim, a **Seção V** traz as conclusões e sugestões para futuras pesquisas. A **Seção VI** apresenta a bibliografia utilizada na pesquisa.

A. Objetivos

1) Objetivo geral: Desenvolver uma análise preditiva do *potencial de crescimento* de 4.394 empresas localizadas nas Américas, por meio da aplicação de algoritmos de *machine learning*, utilizando dados financeiros e macroeconômicos coletados no ano de 2024, com integração dos resultados a uma aplicação web desenvolvida em *Streamlit*.

2) Objetivos específicos:

- Coletar, tratar e consolidar dados financeiros e macroeconômicos de empresas atuantes no continente americano no ano de 2024.
- Definir e calcular uma métrica representativa do *potencial de crescimento* das empresas com base em variáveis selecionadas.
- Aplicar e comparar algoritmos de *machine learning* supervisionados para previsão da variável definida.
- Avaliar o desempenho dos *modelos* por meio de métricas como acurácia, *precision*, *recall*, *F1-score* e *ROC-AUC*.
- Desenvolver uma aplicação web com *Streamlit* para disponibilizar a análise preditiva de forma acessível e interativa.

II. Revisão Teórica

O *aprendizado de máquina* (*machine learning*) tem se consolidado como uma ferramenta indispensável no setor financeiro, ao oferecer soluções inovadoras para problemas clássicos de análise de dados. Sua capacidade de explorar grandes volumes de informação e de modelar relações não lineares permite uma compreensão mais refinada do comportamento do mercado e dos consumidores. Essa tecnologia é amplamente utilizada na detecção de fraudes, avaliação de risco de crédito e personalização de serviços, contribuindo para a obtenção de vantagem competitiva pelas instituições financeiras que a incorporam *(Breiman, 2001)*.

Uma das aplicações mais relevantes do *machine learning* nas finanças é a automação de processos. Tarefas repetitivas e demoradas, como análise de crédito e detecção de transações fraudulentas, podem ser realizadas com maior eficiência e precisão por meio de algoritmos de *aprendizado de máquina*. Essa automação reduz custos operacionais e melhora a experiência do cliente, ao oferecer serviços mais ágeis e personalizados. *(Yu, 2022)*.

Além disso, o *machine learning* possibilita análises preditivas mais eficazes. *Modelos* avançados permitem prever tendências de mercado, comportamentos de consumidores e riscos financeiros com maior precisão, auxiliando na tomada de decisões estratégicas. Em um ambiente financeiro dinâmico e competitivo, a capacidade de antecipar eventos futuros torna-se uma ferramenta crucial *(Zhang, 2024)*.

A integração dessa tecnologia no setor financeiro representa um avanço significativo, proporcionando ferramentas poderosas para a análise de dados e automatização de processos.

III. METODOLOGIA

A. Coleta e Integração dos Dados

Neste trabalho, serão utilizados dados financeiros e macroeconômicos de **4.394** empresas localizadas nas Américas. A coleta será automatizada por meio de *APIs*, como a *Yahoo Finance* e *Alpha Vantage*, complementada com *datasets* públicos da plataforma *Kaggle*. Para indicadores econômicos, serão utilizadas fontes como o *World Bank*, *Trading Economics* e, se necessário, técnicas de *web scraping* via *API* do *GPT*.

Todas as variáveis financeiras e macroeconômicas serão convertidas para dólares americanos (USD) no momento da aquisição, garantindo padronização entre países. Entre as variáveis coletadas estarão: **marketcap**, **revenue_ttm**, **earnings_ttm**, **pe_ratio_ttm**, **price**, **dividend_yield_ttm**, **gdp_per_capita**, **gdp_growth_percent**, **interest_rate_percent**, **inflation_percent**, **unemployment_rate_percent** e **exchange_rate_to_usd**.

B. Ambiente de Desenvolvimento e Ferramentas

Todas as etapas do projeto serão conduzidas em *Python 3.13*, com suporte das bibliotecas *Pandas*, *NumPy*, *Scikit-learn*, *Matplotlib*, *Seaborn*, *Plotly*, *XGBoost*, *Imbalanced-learn*, entre outras. O ambiente *Jupyter Notebook* será utilizado para experimentação e documentação, enquanto o *PyCharm Professional* fornecerá suporte estruturado ao desenvolvimento, depuração e controle de versão.

C. Pré-processamento dos Dados

Será conduzido um processo rigoroso de preparação dos dados, que incluirá:

- **Remoção de duplicatas** e tratamento de valores ausentes e infinitos, com imputação por mediana calculada sobre o conjunto de treino.
- Aplicação de **clipping** nos percentis 0.1 e 99.9 para atenuar outliers extremos.
- Transformação das variáveis numéricas via *Quantile-Transformer* com distribuição *normal*, buscando simetria estatística.
- Inversão dos sinais de **inflation_percent**, **interest_rate_percent** e **unemployment_rate_percent** para assegurar coerência semântica.
- Conversão de tipos e normalização de variáveis com foco na compatibilidade com algoritmos baseados em distância.

Além disso, será realizada *engenharia de **features*** para criar variáveis derivadas mais expressivas, como:

- Razões financeiras: **price_to_arnings**, **market-cap_to_revenue**, **revenue_to_earnings**, etc.;
- Indicadores ajustados ao contexto: **earnings_inflation_impact**, **price_gdp_ratio**, **market-cap_gdp_ratio**, entre outros.

D. Criação da Variável-Alvo

Será aplicada a técnica de *Principal Component Analysis (PCA)* para condensar a variância explicada em poucos componentes. O escore da primeira componente será discretizado por meio da função *qcut*, gerando três grupos balanceados.

A variável resultante, **pe_class**, representará o *potencial de crescimento* das empresas, com as seguintes categorias:

- **0**: Baixo potencial;
- **1**: Médio potencial;
- **2**: Alto potencial.

O número de categorias será validado por meio do *Elbow Method*, aplicado ao algoritmo *K-means*, observando o ponto de inflexão na curva de inércia.

E. Divisão dos Dados e Balanceamento

O conjunto de dados será dividido em 80% para *treinamento* e 20% para *teste*, com estratificação para preservar a proporção das classes em **pe_class**, utilizando a função *train_test_split* do *Scikit-learn*.

Para lidar com o desbalanceamento da variável-alvo, será empregada a técnica *SMOTETomek*, que combina a oversampling da classe minoritária (*SMOTE*) com a remoção de amostras ambíguas nas bordas (*Tomek Links*).

F. Modelagem e Avaliação

Serão utilizados nove algoritmos de *machine learning* supervisionado, selecionados por sua robustez e diversidade:

- **Logistic Regression**: Modelo linear amplamente utilizado para classificação binária, que estima a probabilidade de uma instância pertencer a uma classe com base em uma combinação linear das variáveis preditoras. Apresenta boa interpretabilidade e foi utilizado como modelo base de comparação.
- **Random Forest**: Técnica de *ensemble* baseada em múltiplas árvores de decisão. Cada árvore é construída a partir de uma amostra aleatória dos dados, e o resultado final é obtido por votação. Este modelo mostrou bom desempenho geral, com acurácia e *F1-score* consistentes, além de não apresentar indícios significativos de *overfitting*.
- **Gradient Boosting**: Algoritmo de *boosting* que constrói árvores sequencialmente, corrigindo os erros cometidos pelas anteriores. Apresentou o melhor desempenho entre todos os modelos, com alta acurácia, *F1-score* e excelente capacidade de prever corretamente a classe 0, embora tenha demonstrado leve *overfitting*.
- **XGBoost**: Uma implementação otimizada de *gradient boosting*, que incorpora técnicas de regularização para melhorar a generalização. Teve desempenho competitivo, com ótima pontuação balanceada e resultados consistentes, sem apresentar sinais relevantes de *overfitting*.
- **Support Vector Machine (SVM)**: Modelo que busca encontrar o hiperplano ótimo que separa as classes. Apesar de sua eficácia em alguns contextos, neste projeto apresentou desempenho inferior aos demais, principalmente em *recall* e *F1-score*.
- **K-Nearest Neighbors (KNN)**: Classificador baseado na proximidade dos dados no espaço de características. Embora simples, obteve resultados razoáveis, especialmente em precisão, mas com indícios de *overfitting*.
- **Naive Bayes**: Baseado no teorema de Bayes, assume independência entre as variáveis. Foi avaliado por sua eficiência e baixo custo computacional, porém apresentou desempenho inferior em termos de acurácia da classe 0.

- **Decision Tree**: Modelo baseado em regras de decisão hierárquicas. Fornece boa interpretabilidade e resultados satisfatórios, sendo competitivo em todas as métricas, sem *overfitting* acentuado.
- **Neural Network**: Modelo inspirado no funcionamento do cérebro humano, capaz de capturar padrões complexos nos dados. Apresentou desempenho sólido em todas as métricas, com boa generalização.

Os *modelos* serão estruturados em *pipelines*, com normalização, e treinados com *validação cruzada estratificada* de $k = 5$ *folds*. A escolha final do *modelo* será baseada na métrica de *F1-score* da classe 0, *acurácia balanceada* e valores de *ROC-AUC*.

Para aprimorar a performance do *modelo*, foi realizado um ajuste fino dos *hiperparâmetros* utilizando a técnica de *Grid Search*, disponível na biblioteca *scikit-learn*. O *Grid Search* é uma abordagem de busca exaustiva que testa sistematicamente todas as combinações possíveis de *hiperparâmetros* definidos pelo usuário, permitindo identificar a configuração que maximiza a performance do modelo.

Pesos customizados serão atribuídos às classes de difícil distinção nos *modelos baseados em árvore*, como *Random Forest* e *Gradient Boosting*, para aumentar sua sensibilidade.

G. Justificativa das Ferramentas e Reprodutibilidade

A escolha das bibliotecas será fundamentada em sua maturidade e ampla adoção no campo da *Ciência de Dados*. O uso de *Scikit-learn*, *XGBoost* e *Imbalanced-learn* permitirá a implementação eficiente de técnicas modernas de *machine learning*.

A reprodutibilidade será garantida por meio da definição explícita do parâmetro **random_state** nas etapas de divisão, balanceamento e modelagem. O *modelo* final será exportado com a biblioteca *joblib* para futuras integrações.

H. Fluxo de Execução do Projeto

O fluxo de execução seguirá as etapas descritas na **Figura 1**.

I. Aplicação e Visualização

O *modelo* com melhor desempenho será integrado a uma aplicação web desenvolvida com *Streamlit*, permitindo a realização de previsões de forma acessível e interativa. Gráficos e análises serão gerados com *Plotly*, promovendo visualizações interativas e intuitivas dos resultados.

IV. Resultados

A. Coleta de Dados

Os dados utilizados neste estudo foram previamente coletados por meio de *APIs* e bases públicas, conforme descrito na seção de Metodologia.. Os *datasets* utilizados foram: *Companies_ranked_by_Earnings*, *Companies_ranked_by_Dividend_Yield*, *Companies_ranked_by_Market_Cap*, *Companies_ranked_by_P_E_ratio* e *Companies_ranked_by_Revenue*, todos em formato *.csv*. Após a unificação dos dados utilizando o IDE *PyCharm PRO* e as bibliotecas *Python Pandas* e *Numpy*, os *datasets*

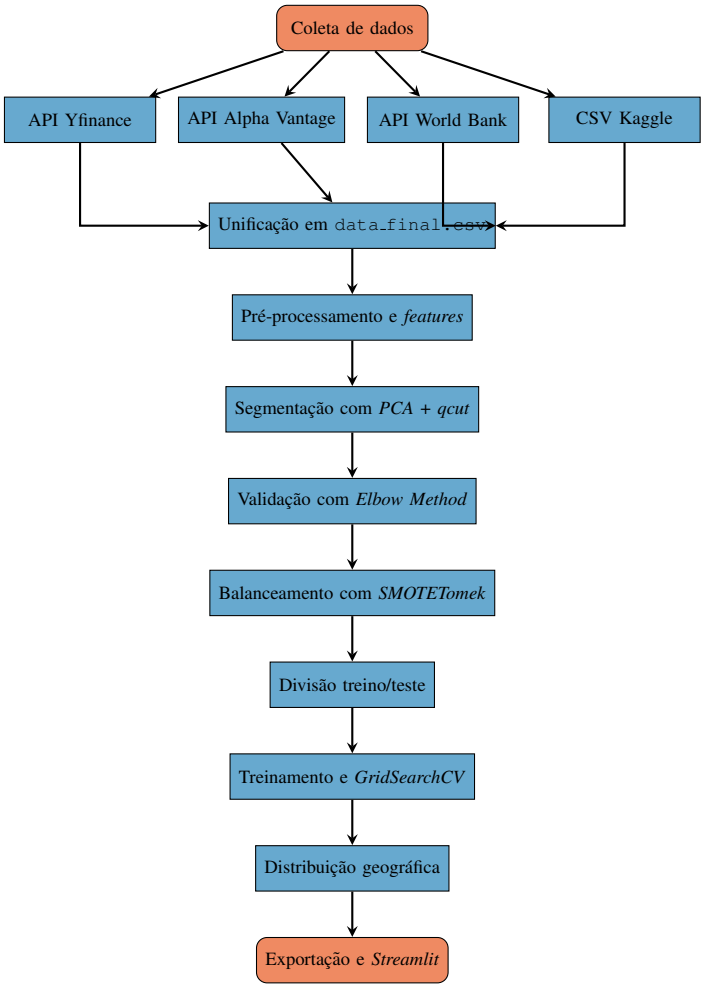

Fig. 1: Fluxograma de Processos

foram filtrados a partir da coluna *'country'* que representa os países de origem das empresas para que apenas restassem empresas localizadas no continente americano; além disso, foram centralizados em um único arquivo, denominado *data_final.csv*. Os dados financeiros utilizados neste estudo foram coletados de diversas empresas localizadas em diferentes países. Todos os valores foram convertidos para dólares norte-americanos (USD) no momento da coleta, levando em consideração a cotação no período dos dados. Essa padronização direta para dólares garante que as comparações financeiras entre as empresas sejam realizadas de forma justa e evita distorções causadas por flutuações cambiais.

Os parâmetros financeiros presentes nesses *datasets* empresariais incluem:

- **earnings_ttm** — Lucros (últimos 12 meses)
- **dividend_yield_ttm** — Rendimento de dividendos (últimos 12 meses)
- **marketcap** — Capitalização de mercado
- **pe_ratio_ttm** — Índice Preço/Lucro (P/L)
- **revenue_ttm** — Receita total (últimos 12 meses)
- **price** — Preço do ativo

A coleta de dados macroeconômicos foi realizada por meio de ferramentas de *API*, utilizando fontes como o *Trading Economics*, a *API* do *World Bank*, e outras plataformas especializadas. Essas *APIs* forneceram informações essenciais

sobre os dados econômicos de diversos países, como PIB per capita, taxa de crescimento do PIB, inflação, taxa de juros, taxa de câmbio e taxa de desemprego. No entanto, ao tentar obter os dados sobre a economia de Bermudas, não foi possível encontrá-los diretamente nas APIs. Como alternativa, foi necessário recorrer ao processo de *scraping* utilizando a *API* do *GPT* para buscar as informações faltantes.

Os parâmetros macroeconômicos coletados para cada país incluem:

- **gdp_per_capita_usd** — PIB per capita (em dólares)
- **gdp_growth_percent** — Taxa de crescimento do PIB para 2024
- **inflation_percent** — Taxa de inflação
- **interest_rate_percent** — Taxa de juros
- **exchange_rate_to_usd** — Taxa de câmbio (em relação ao dólar)
- **unemployment_rate_percent** — Taxa de desemprego

Esses parâmetros foram organizados de maneira a permitir uma análise detalhada sobre o desempenho financeiro das empresas, sendo essenciais para as etapas subsequentes do trabalho.

B. Análise Exploratória

A Análise Exploratória de Dados (*AED*) é uma etapa essencial no processo de análise, pois permite compreender as principais características dos dados, identificar padrões e detectar possíveis inconsistências. De acordo com *Rao et al. (2021)*, essa abordagem inicial é fundamental para garantir a qualidade dos dados antes da aplicação de *modelos* estatísticos ou de *machine learning*, prevenindo interpretações errôneas e resultados distorcidos. Além disso, a *AED* é crucial para a detecção de valores ausentes ou *outliers* que possam afetar a análise subsequente, como destacado por *Rao et al. (2021)*.

Durante a Análise Exploratória de Dados (*EDA*), realizamos as seguintes etapas para avaliar a consistência e integridade do conjunto de dados:

1) **Verificação do Tamanho do *Dataframe*:** A função *shape* da biblioteca *Pandas* foi utilizada com o objetivo de determinar o número de linhas e colunas no conjunto de dados, garantindo um entendimento inicial sobre sua estrutura. O resultado dessa verificação indicou um *shape* de (4.394, 13), ou seja, o conjunto de dados contém **4.394 observações** e **13 variáveis**.
2) **Verificação de Valores Faltantes ou Nulos:** Uma análise para identificar a presença de valores nulos ou ausentes em cada coluna foi realizada, etapa essencial para evitar problemas na modelagem e análises subsequentes.
3) **Verificação de Valores Duplicados:** A análise foi concentrada na coluna *name*, verificando a existência de registros duplicados.
4) **Verificação do Nível de Agrupamento de Valores Numéricos:** Uma análise de dispersão foi realizada com o objetivo de identificar a necessidade de normalização ou transformação de variáveis.

C. Pré-processamento de Dados

O pré-processamento de dados é uma etapa crucial na análise de dados, envolvendo a preparação e transformação dos dados brutos em um formato adequado para análise. Essa fase é essencial para garantir que os dados sejam precisos, completos e relevantes, aumentando a confiabilidade das conclusões obtidas a partir deles, conforme destacado por *Han, Kamber e Pei (2012)*, que enfatizam técnicas como limpeza de dados, integração, transformação e redução para melhorar a qualidade e utilidade dos dados em mineração e aprendizado de máquina.

1) Limpeza e Manipulação: A limpeza e manipulação de dados são etapas essenciais para garantir a qualidade das informações utilizadas em análises. A limpeza envolve identificar e corrigir erros, remover duplicatas e tratar valores ausentes, preparando os dados para análises confiáveis *(AWS, 2023)*. A manipulação de dados transforma dados brutos em informações estruturadas, permitindo a extração de *insights* significativos. Esse processo é fundamental para garantir que os dados estejam em um formato adequado para análise, conforme destacado por *Han, Kamber e Pei (2012)*. Conforme o processo de pré-processamento apresentado na Metodologia, a limpeza foi feita na seguinte ordem:

1) **Remoção de Valores Duplicados:** Foi aplicada a função *drop_duplicates(keep='first')* da biblioteca *Pandas*, eliminando entradas redundantes na variável *name*, o que contribuiu para evitar viés na modelagem.
2) **Tratamento de Valores Faltantes e Infinitos:** Valores infinitos foram inicialmente convertidos para *NaN*, seguidos por preenchimento utilizando a mediana da respectiva *feature*, calculada exclusivamente no conjunto de treino para evitar vazamento de dados. Ao final do processo, aplicou-se *np.nan_to_num()* para garantir a ausência de valores inválidos.
3) **Clipping de Valores Extremos:** Valores foram limitados aos percentis 0.1 e 99.9 para cada variável, preservando a estrutura dos dados e mitigando o impacto de *outliers* severos.
4) **Conversão de Tipos de Dados:** As colunas numéricas foram convertidas para o tipo *float*, assegurando compatibilidade com os algoritmos de aprendizado de máquina.
5) **Transformação Quantílica:** Aplicou-se a transformação com o *QuantileTransformer* da biblioteca *scikit-learn* com a distribuição de saída configurada como *normal*. Esta abordagem demonstrou ser mais eficaz que normalizadores convencionais para dados financeiros com distribuições assimétricas.
6) **Padronização da Direção Semântica:** As variáveis macroeconômicas **inflation**, **interest_rate_percent** e **unemployment** tiveram seus sinais invertidos para garantir coerência semântica (valores maiores representam melhor cenário).

2) Seleção e Preparação dos Dados:: A primeira etapa na construção do *modelo* preditivo foi a seleção das variáveis relevantes para a análise do potencial de crescimento (*pc_class*). As variáveis escolhidas abrangem tanto indicadores financeiros das empresas quanto fatores macroeconômicos que influenciam o ambiente em que essas empresas operam. As variáveis que compõem o cálculo do *pc_class* são:

- **marketcap**: Capitalização de mercado das empresas, expressa em bilhões. Este indicador fornece uma medida do tamanho relativo da empresa no mercado.

- **revenue_ttm**: Receita total dos últimos 12 meses (*TTM*), expressa em bilhões. Reflete a capacidade da empresa em gerar receita.
- **pe_ratio_ttm**: Razão Preço/Lucro (*P/E ratio*) dos últimos 12 meses, um indicador do quanto os investidores estão dispostos a pagar por cada unidade de lucro gerado pela empresa.
- **price**: Preço dos ativos da empresa, utilizado para calcular a atratividade da empresa em relação aos seus lucros.
- **earnings_ttm**: Lucro total dos últimos 12 meses (*TTM*), expressa em bilhões. Este parâmetro ajuda a avaliar a rentabilidade da empresa.
- **inflation_rate**: Taxa de inflação anual, um fator macroeconômico que afeta diretamente o poder de compra e as condições de mercado.
- **gdp_growth_2024**: Taxa de crescimento do Produto Interno Bruto (*PIB*) do país para o ano de 2024. Reflete as perspectivas econômicas gerais do país e influencia o ambiente de negócios.
- **gdp_per_capita**: *PIB* per capita, que indica a riqueza média gerada por habitante no país onde a empresa está localizada.

Essas variáveis foram escolhidas por sua capacidade de fornecer uma visão abrangente sobre a saúde financeira das empresas e as condições econômicas do país no qual elas estão inseridas. A literatura sugere que a combinação desses indicadores pode proporcionar uma avaliação precisa das perspectivas de crescimento das empresas (*GrahamDodd*, 2008), (*Damodaran*, 2012).

3) Engenharia de parâmetros: A *engenharia de parâmetros* consistiu na transformação e criação de novos atributos derivados de indicadores financeiros e macroeconômicos, com o objetivo de capturar relações mais complexas e contextuais entre os dados. A transformação das variáveis foi uma etapa crítica para garantir a consistência e a comparabilidade entre os dados.

- **Transformações Logarítmicas:** Variáveis como *earnings_ttm*, *revenue_ttm*, *marketcap* e *price* passaram por transformações logarítmicas do tipo *np.log1p(x + offset)* com *offset* dinâmico, a fim de lidar com valores negativos ou nulos e reduzir a assimetria positiva.
- **Criação de Razões Financeiras Robustas:** Foram introduzidas variáveis como:
  - *price_to_revenue*, *price_to_earnings*
  - *marketcap_to_revenue*, *marketcap_to_earnings*
  - *revenue_to_earnings*, *price_to_marketcap*

  As razões foram estabilizadas com constantes nos denominadores e uso de *np.clip* para evitar divisões por zero e valores excessivos.
- **Variáveis de Interação Econômica:** Novas variáveis foram criadas com base na interação entre os dados da empresa e o contexto macroeconômico:
  - *price_gdp_ratio*: preço relativo ao PIB per capita
  - *marketcap_gdp_ratio*: capitalização de mercado relativa ao PIB per capita
  - *earnings_inflation_impact*: lucros ajustados pela inflação
  - *revenue_interest_impact*: receita ajustada pela taxa de juros

  Tais variáveis foram especialmente relevantes para a separação da classe 0, que se mostrou mais sensível ao contexto econômico.

4) Balanceamento de Classes: Devido ao desbalanceamento da variável alvo, foi utilizada a técnica *SMOTETomek*, que combina o *SMOTE* (aumento sintético das classes minoritárias) com *Tomek Links* (remoção de amostras ambíguas nas fronteiras de decisão), resultando em uma distribuição mais equilibrada e menos propensa a *overfitting*.

D. Análise Bivariada:

A **Figura 2** apresenta o mapa de calor com os coeficientes de correlação de *Spearman* entre variáveis financeiras e macroeconômicas. Esta métrica não paramétrica é adequada para identificar relações monotônicas mesmo em distribuições não normais, sendo ideal para dados econômicos heterogêneos.

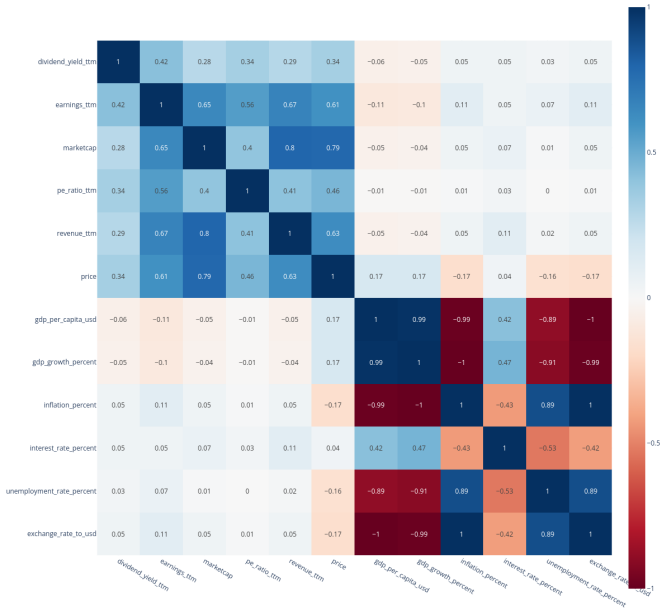

Fig. 2: Correlação Spearman

Relações Intrassetoriais: Métricas Corporativas: As métricas financeiras como *earnings_ttm*, *revenue_ttm*, *marketcap* e *price* exibem correlações fortemente positivas entre si, com destaque para:

- **marketcap vs revenue_ttm**: $\rho = 0.8$
- **marketcap vs price**: $\rho = 0.79$
- **earnings_ttm vs revenue_ttm**: $\rho = 0.67$

Essas correlações confirmam que empresas com maiores receitas tendem a possuir maior capitalização de mercado, o que é consistente com a lógica de *valuation* baseada em fluxo de caixa.

Outro ponto importante é a correlação entre *pe_ratio_ttm* e *price* ($\rho = 0.46$), sugerindo que múltiplos de preço-lucro ainda são relevantes na precificação de ações, embora moderadamente, possivelmente refletindo a influência de fatores externos, como expectativas de crescimento e risco percebido.

Dinâmica Macroeconômica: Crescimento, Juros e Inflação: As variáveis macroeconômicas demonstram padrões sistemáticos que refletem relações estruturais da economia:

- *gdp_per_capita_usd* e *gdp_growth_percent* possuem correlação perfeita ($\rho = 0.99$), sugerindo que países com maior PIB per capita tendem a ter taxas de crescimento econômico mais estáveis.
- Há correlação negativa quase perfeita entre *gdp_growth_percent* e *inflation_percent* ($\rho = -1$), e entre *gdp_growth_percent* e *interest_rate_percent* ($\rho = -0.43$), o que está alinhado com o modelo IS-LM, no qual o crescimento desacelera em resposta ao aumento das taxas de juros e à inflação.
- *unemployment_rate_percent* também está fortemente correlacionada negativamente com o crescimento e o PIB per capita ($\rho \approx -0.9$), evidenciando a clássica relação de Okun.

Política Monetária e Dinâmica Cambial: A variável *exchange_rate_to_usd* apresenta correlações negativas acentuadas com:

- *gdp_per_capita_usd*: $\rho = -1$
- *gdp_growth_percent*: $\rho = -0.99$
- *price*: $\rho = -0.17$

Tais correlações sugerem que uma moeda desvalorizada frente ao dólar está fortemente associada a economias menos desenvolvidas ou em desaceleração, além de possível impacto negativo sobre o valor de mercado de ativos denominados em moeda local.

Além disso, o câmbio também mostra correlação inversa com o *interest_rate_percent* ($\rho = -0.42$), o que pode refletir a atuação de bancos centrais em países emergentes que elevam os juros como forma de conter a depreciação cambial, prática comum em contextos de fuga de capital.

4. Relações Fracas, mas Estratégicas: Algumas variáveis demonstram correlações fracas, mas que merecem atenção estratégica:

- *dividend_yield_ttm* apresenta correlação positiva fraca com variáveis como *earnings_ttm* ($\rho = 0.42$) e *pe_ratio_ttm* ($\rho = 0.34$), o que pode indicar que empresas lucrativas ainda mantêm políticas de distribuição de dividendos, embora em um mercado com foco crescente em reinvestimento de lucros.
- A ausência de correlações significativas entre *dividend_yield_ttm* e variáveis macroeconômicas sugere que as decisões de *payout* são mais *firm-specific* do que dependentes do ambiente econômico externo.

5. Sinais de Assimetrias de Mercado: Notam-se assimetrias notáveis:

- Enquanto variáveis internas das empresas (lucros, receita, preço) mantêm coesão interna, suas correlações com variáveis macroeconômicas são majoritariamente fracas.
- Isso pode sinalizar que, no curto prazo, o desempenho de ações reflete mais características idiossincráticas da empresa do que o ciclo econômico, o que reforça a hipótese de eficiência semi-forte dos mercados.

E. Análise Univariada

A *análise univariada* é uma técnica estatística que examina cada variável de forma isolada, buscando descrever sua distribuição e identificar padrões, desvios ou anomalias. Por meio de métricas como média, mediana e desvio padrão, além de representações gráficas, ela permite compreender o comportamento dos dados e detectar possíveis problemas, como valores discrepantes. Esse tipo de análise é essencial para orientar a construção de *modelos* preditivos e decisões sobre ajustes nas variáveis *(Montgomery et al., 2021; Bowerman e O'Connell, 2016)*.

1) Análise de Distribuição: Para a *análise de distribuição*, foi gerada uma grade de histogramas utilizando as bibliotecas *Plotly* e *Plotly Express* para verificar a distribuição das variáveis. Assim, foi possível visualizar a frequência das observações em diferentes intervalos e a presença de alguns *outliers*. A **Figura 3** apresenta a distribuição dos dados por meio de uma grade de histogramas.

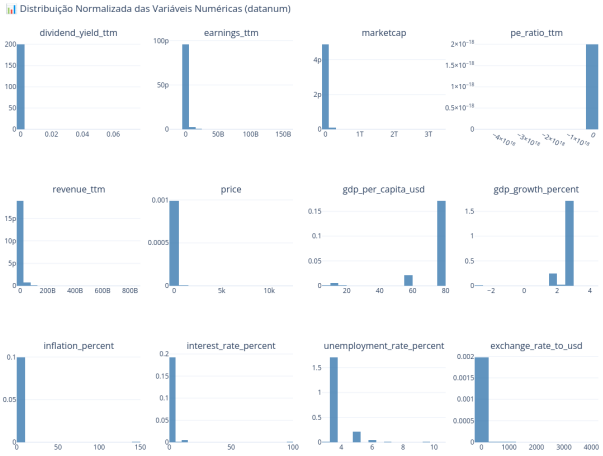

Fig. 3: Grade de Histogramas

2) Assimetria e Dispersão: A distribuição dos dados indica que:

- Muitas variáveis possuem valores concentrados próximos a zero, com alguns valores muito altos *(outliers)*;
- Variáveis como **revenue_ttm**, **marketcap** e **gdp_per_capita_usd** seguem distribuições altamente assimétricas;
- Algumas variáveis macroeconômicas, como **inflation_rate** e **interest_rate**, apresentam grande concentração em valores baixos.

3) Impacto na Modelagem: Dado esse comportamento dos dados, algumas considerações devem ser feitas na escolha dos modelos:

1) Modelos baseados em distância, como **KNN** e **SVM**, podem ter dificuldades devido à presença de *outliers* e escalas distintas;
2) Árvores de decisão e seus derivados, como **Random Forest** e **XGBoost**, são mais robustos a *outliers* e distribuições assimétricas;
3) A regressão logística pode ser aplicada, mas pode exigir transformações logarítmicas para lidar com a assimetria;
4) Técnicas de normalização, como **Min-Max Scaling** ou **log-transform**, podem melhorar a performance dos modelos.
5) A presença de *outliers* ou dados atípicos pode distorcer significativamente os resultados do método *Elbow*, pois

esses valores extremos aumentam a variância *intra-cluster (WCSS)*, dificultando a identificação clara do ponto de inflexão que indica o número ótimo de agrupamentos.

4) Identificação de Outliers:: A dispersão dos dados de treino foi analisada por meio de gráficos como *boxplots*. Os valores que se apresentaram como *outliers*. As figuras **Figura 4** e **Figura 5** compõem uma representação gráfica em forma de *Boxplot* na variável *price* antes e depois do tratamento dos *outliers*.

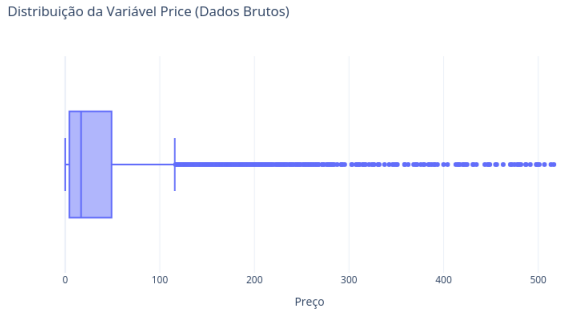

Fig. 4: Boxplot Antes do Tratamento

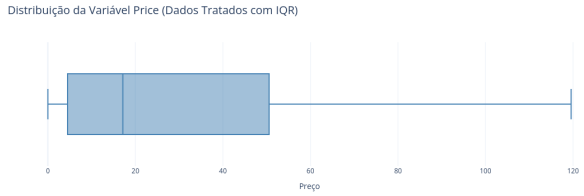

Fig. 5: Boxplot Depois do Tratamento

A remoção ou ajuste de *outliers* é essencial para aprimorar a precisão e a estabilidade de modelos de clusterização, como *KMeans*, métodos de avaliação de variação *intra-cluster* e modelos preditivos, como regressão logística, *KNN* e *SVM*, que podem ser sensíveis a valores extremos, resultando em classificações imprecisas *(Freitas, 2019)*. Reduzir a influência dos *outliers* assegura que o modelo reflita o comportamento real dos dados, evitando viés e *overfitting*, além de melhorar a capacidade de generalização do modelo *(Hendrycks et al., 2018)*. Contudo, o tratamento de *outliers* deve ser realizado com cautela, considerando o contexto específico de cada tarefa de aprendizado de máquina, já que em algumas situações, manter *outliers* pode ser benéfico *(Zamoner, 2014)*.

5) Determinação do Número de Clusters com o Método Elbow: Com base nas técnicas descritas na Metodologia, a construção da variável-alvo **pc_class**, representando o *potencial de crescimento* das empresas, foi necessário definir previamente a quantidade ideal de classes que melhor segmentasse as observações segundo seus padrões de variabilidade.

Para isso, aplicou-se o *Elbow Method*, técnica que consiste em avaliar a variação da inércia (ou soma das distâncias intra-grupo) para diferentes quantidades de agrupamentos, observando o ponto de inflexão em que o ganho de performance começa a se estabilizar. Os testes foram conduzidos com os dados previamente tratados, após remoção de *outliers* com base na regra do *intervalo interquartil (IQR)*. O gráfico gerado revelou uma queda acentuada da inércia até o ponto correspondente a três agrupamentos, onde se observa um "cotovelo" na curva, indicando que este número de divisões é o mais eficiente em termos de balanceamento entre complexidade e explicabilidade. Dessa forma, definiu-se a segmentação dos dados em três grupos principais, que posteriormente foram rotulados com os valores inteiros de 0 a 2 na variável **pc_class**, representando diferentes níveis de *potencial de crescimento*.

A **Figura 6** apresenta o gráfico gerado pelo *Elbow Method*, aplicado ao conjunto de dados numéricos tratados.

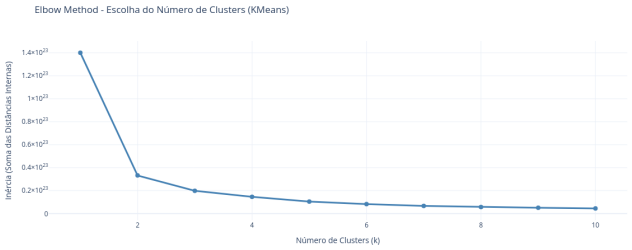

Fig. 6: Gráfico da Curva de inércia

6) Segmentação das Empresas com PCA e Discretização: Com base nas variáveis previamente selecionadas e tratadas, adotou-se uma abordagem combinada de análise de componentes principais (*Principal Component Analysis - PCA)* e discretização por quantis (*qcut*) para a criação da variável **pc_class**, representando o *potencial de crescimento* das empresas.

O uso do *PCA* teve como objetivo reduzir a dimensionalidade do conjunto de dados e concentrar a variância explicada em poucas componentes, permitindo assim uma visão mais robusta das principais direções de variabilidade dos dados financeiros e macroeconômicos. Essa redução torna o processo de agrupamento mais estável e menos sensível a ruídos ou redundâncias nas variáveis.

A discretização dos escores do *PCA* com a função *qcut*, que divide os dados em faixas com a mesma proporção de observações, foi utilizada para criar três categorias de crescimento. Essa abordagem oferece maior interpretabilidade, ao mesmo tempo que garante o balanceamento entre as classes, sendo especialmente vantajosa em contextos onde a classificação supervisionada depende de uma variável-alvo bem definida e balanceada.

Segundo *Fraunhofer Institute (2024)*, a combinação de técnicas de redução de dimensionalidade com métodos de discretização adaptativa permite transformar dados contínuos em categorias informativas, facilitando tarefas preditivas e análises interpretáveis em problemas de *machine learning*. O estudo destaca que, ao refinar o espaço de discretização nas direções mais relevantes do *embedding*, é possível obter representações mais eficientes e informativas dos dados, beneficiando aplicações reais de *aprendizado de máquina*.

Dessa forma, a criação da variável **pc_class** foi guiada tanto por critérios estatísticos quanto pela necessidade de interpretabilidade e equilíbrio entre as classes para os modelos supervisionados subsequentes.

Na **Figura 7**, apresenta-se um gráfico de dispersão ilustrando as classes geradas pelo *método híbrido* utilizado.

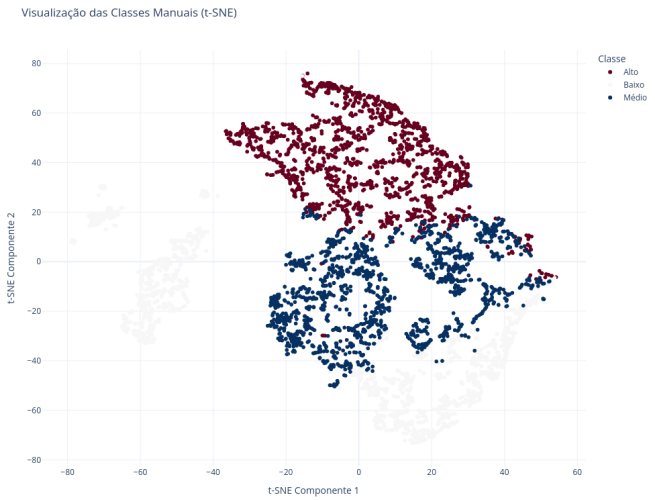

Fig. 7: Representação das Classes

7) Interpretação de Classes: Após a aplicação do *PCA + Qcut*, as empresas foram organizadas em três *classes* distintas, cada uma representando um grupo com perfis de comportamento financeiro e macroeconômico semelhantes. Para compreender as características que definem cada grupo, foi realizada uma análise das médias das variáveis numéricas dentro de cada *classe*.

Essa abordagem permite identificar padrões centrais sem depender de uma fórmula de índice pré-definida, facilitando a interpretação dos grupos formados. As médias das principais variáveis financeiras e macroeconômicas por *classe* estão apresentadas na **Tabela 1**.

A partir da análise desses valores, foi possível classificar as *classes* segundo o *potencial de crescimento* das empresas, sendo elas caracterizadas como baixo, médio e alto potencial, conforme evidenciado pelas diferenças nos indicadores-chave, tais como receita, lucro, valor de mercado e preço das ações.

8) Divisão do Dataset em Treino e Teste: Com o objetivo de realizar uma avaliação confiável e imparcial do desempenho do modelo preditivo, o conjunto de dados foi segmentado em dois subconjuntos distintos: **treino** e **teste**. Essa etapa é fundamental para evitar o *data leakage*, situação em que informações do conjunto de teste influenciam o treinamento do modelo, comprometendo a validade da avaliação.

A divisão dos dados foi executada por meio da função *train_test_split*, disponibilizada pela biblioteca *Scikit-Learn*, amplamente utilizada na área de Ciência de Dados devido à sua robustez e facilidade de uso. Nessa divisão, foi adotada a proporção de **80%** dos dados para o conjunto de treino e **20%** para o conjunto de teste.

O **conjunto de treino** foi utilizado para ajustar os parâmetros internos do modelo, permitindo que ele aprenda os padrões existentes nos dados. Já o **conjunto de teste** foi reservado exclusivamente para a avaliação final do desempenho do modelo, fornecendo uma estimativa realista de sua capacidade de generalização para novos dados não observados durante o processo de aprendizado.

Substituição por Valores Numéricos: Para facilitar a interpretação e utilização dos resultados no modelo de análise, os rótulos das *classes* foram convertidos em valores numéricos representando o *potencial de crescimento* que encontramos na coluna *pc_class* de cada empresa:

- **Potencial Baixo**: 0;
- **Potencial Médio**: 1;
- **Potencial Alto:** 2;

Essa transformação permite que as *classes* geradas sejam interpretadas de forma mais objetiva em análises subsequentes, como em modelos de previsão ou classificação, facilitando sua aplicação em contextos de *aprendizado supervisionado*.

A coluna categórica, denominada *pc_class*, foi utilizada como variável dependente nos *modelos* de classificação. Esse processo garantiu que cada classe tivesse aproximadamente o mesmo número de registros, reduzindo problemas de desbalanceamento, o que é essencial para o desempenho de *modelos* supervisionados *(Brownlee, 2020)*.

9) Modelagem: A modelagem em *machine learning* (ML) tem como objetivo identificar padrões e prever resultados com base em dados. A aplicação de *modelos* de classificação é essencial para categorizar dados em diferentes grupos, sendo amplamente utilizada em problemas como previsão de categorias ou diagnósticos. O processo de modelagem envolve o treinamento de algoritmos que ajustam seus parâmetros para minimizar o erro de previsão. Os modelos testados apresentaram desempenhos variados. O *Gradient Boosting* destacou-se com os melhores resultados em todas as métricas, incluindo *F1-score* de 0.94 e acurácia balanceada de 0.84. O *Random Forest* e o *XGBoost* também obtiveram desempenhos consistentes, enquanto modelos mais simples, como *Naive Bayes* e *SVM*, demonstraram limitações na classificação da classe 0.

Configuração de Pesos e Modelos: Para lidar com a dificuldade de classificar corretamente a classe 0, foram atribuídos pesos diferenciados:

- **Random Forest:** *class_weight* = {*0: 10, 1: 1, 2: 1*}, penalizando fortemente erros na classe 0.
- **Gradient Boosting:** Implementação de pesos personalizados via *sample_weight*, aplicando a mesma lógica de 10x para a classe 0.

F. Escolha do Modelo

A escolha do *modelo* de classificação ideal é uma etapa crucial no processo de análise de dados, uma vez que impacta diretamente na qualidade das previsões. Para determinar o *modelo* mais eficaz, são utilizados diversos critérios de avaliação, como *Acurácia*, *Precisão*, *Recall* e *F1 Score*, que fornecem uma visão abrangente do desempenho do *modelo* em diferentes aspectos. Além disso, a *Curva ROC* (*Receiver Operating Characteristic Curve*), que analisa a relação entre a *Taxa de Verdadeiros Positivos* (*True Positive Rate*) e a *Taxa de Falsos Positivos* (*False Positive Rate*).

As métricas relacionadas aos resultados obtidos das previsões dos *modelos* de classificação estão presentes na **Tabela 2**.

A *Acurácia* reflete a proporção de previsões corretas, enquanto a *precisão* mede a proporção de verdadeiros positivos entre os positivos previstos, o *recall* avalia a capacidade de identificar os positivos reais e o *F1* combina *precisão* e *recall* em uma única métrica.

TABELA I: Médias das variáveis financeiras e macroeconômicas por classe

Variável	Classe 0	Classe 1	Classe 2
Dividend Yield (TTM)	3,29E-08	1,01E-08	3,90E-08
Earnings (TTM)	5,27E+05	8,12E+08	3,68E+08
Market Cap	8,02E+08	1,27E+10	5,10E+09
PE Ratio (TTM)	3,51	22,97	13,22
Revenue (TTM)	5,93E+08	7,44E+09	4,34E+09
Price	17,01	83,81	53,48
GDP per capita (USD)	76,0	76,0	76,0
GDP growth (%)	2,60	2,60	2,60
Inflation (%)	2,50	2,50	2,50
Interest rate (%)	4,88	4,88	4,88
Unemployment rate (%)	3,70	3,70	3,70
Exchange rate to USD	1,00	1,00	1,00

TABELA II: Métricas de Avaliação dos Modelos

Modelo	Acurácia	Precisão	Recall	F1-Score	Overfitting	Acurácia Classe 0	Pontuação Balanceada
Random Forest	0.900	0.900	0.900	0.899	0.192	0.823	0.767
Gradient Boosting	0.940	0.941	0.940	0.940	0.116	0.969	0.843
XGBoost	0.925	0.928	0.925	0.925	0.144	0.857	0.797
Logistic Regression	0.778	0.806	0.778	0.772	0.051	0.983	0.755
SVM	0.600	0.758	0.600	0.568	0.074	0.959	0.630
KNN	0.840	0.845	0.840	0.838	0.323	0.717	0.686
Naive Bayes	0.771	0.783	0.771	0.769	0.032	0.614	0.643
Decision Tree	0.883	0.883	0.883	0.883	0.170	0.870	0.774
Neural Network	0.901	0.908	0.901	0.900	0.099	0.782	0.765

Com base nos resultados obtidos, o *Gradient Boosting* se destacou como o modelo de melhor desempenho, apresentando os maiores valores de *Acurácia*, *Precisão*, *Recall* e *F1-Score*. Esse modelo demonstrou ser altamente eficaz na classificação das instâncias, justificando sua escolha para a análise final. Sua robustez é atribuída à capacidade de combinar múltiplas árvores de decisão de forma sequencial, onde cada árvore busca corrigir os erros cometidos pelas anteriores, permitindo uma generalização mais confiável dos dados *Friedman, 2001*. A **Figura 8** representa visualmente a comparação entre as métricas de avaliação de cada modelo, evidenciando o desempenho superior do *Gradient Boosting* em relação aos demais algoritmos analisados.

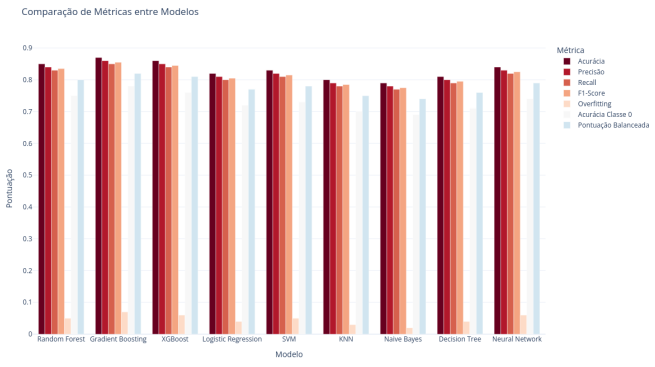

Fig. 8: Comparação de Métricas de Avaliação

A **Figura 9** apresenta uma representação gráfica (*Matriz de Confusão*) em forma de *heatmap* dos resultados das previsões geradas pelo modelo *Gradient Boosting* utilizando a biblioteca *Plotly Express*:

A *Matriz de Confusão* quantifica os acertos e erros de classificação do modelo. As previsões corretas estão dispostas na diagonal principal, enquanto os erros aparecem fora dessa

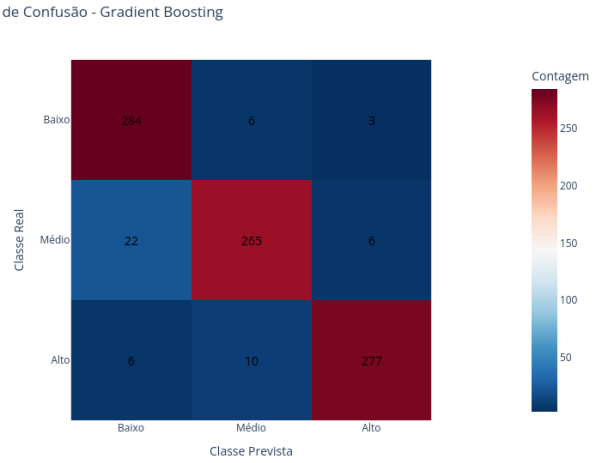

Fig. 9: Matriz de Confusão

diagonal. Os principais valores observados foram:

- **Classe Baixo**: 284 acertos, 6 classificados como *Médio*, e 3 como *Alto*.
- **Classe Médio**: 265 acertos, 22 classificados como *Baixo*, e 6 como *Alto*.
- **Classe Alto**: 277 acertos, 6 classificados como *Baixo*, e 10 como *Médio*.

O modelo obteve excelente desempenho geral, com alta taxa de acertos em todas as classes. O pequeno número de erros está concentrado em confusões entre classes adjacentes (por exemplo, *Médio* sendo confundido com *Baixo* ou *Alto*), o que é compreensível considerando a natureza contínua do *potencial de crescimento* e a possível sobreposição entre limites de classificação.

Para cada classe, foi calculada a *curva ROC* e a respectiva *área sob a curva* (*AUC - Area Under the Curve*), que indica

a capacidade do modelo escolhido em separar as classes corretamente, vide a **Figura 10**.

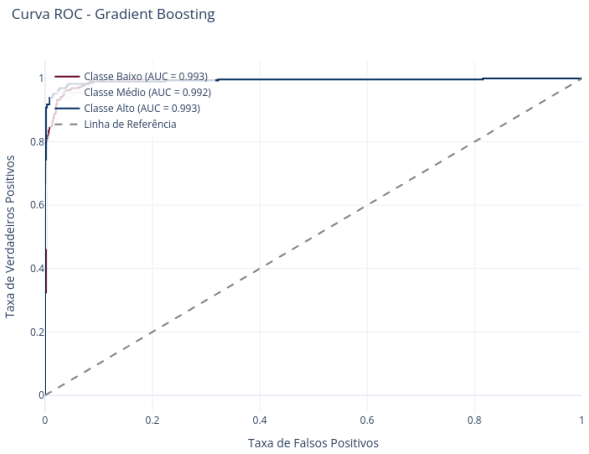

Fig. 10: Curva ROC

A *Curva ROC* ilustra o desempenho do modelo na distinção entre as três classes: *Baixo*, *Médio* e *Alto potencial de crescimento*. Os valores da área sob a curva (AUC) para cada classe foram os seguintes:

- *Classe Baixo*: AUC = 0,993
- *Classe Médio*: AUC = 0,992
- *Classe Alto*: AUC = 0,993

Os elevados valores de *AUC* (próximos de 1) indicam uma excelente capacidade discriminativa do modelo para todas as classes. A proximidade das curvas ao canto superior esquerdo do gráfico reforça o alto desempenho preditivo. Além disso, a similaridade entre os valores de *AUC* demonstra que o modelo é equilibrado, sem favorecer excessivamente nenhuma das categorias.

G. Análise de Features Mais Importante

A análise das *features* mais importantes busca identificar as variáveis com maior influência nas previsões do modelo. Segundo *Breiman (2001)*, essa análise é fundamental para entender a tomada de decisões do modelo e otimizar seu desempenho, enquanto *Molnar (2020)* ressalta que a seleção de *features* relevantes pode melhorar a acurácia e a interpretabilidade. A **Figura 11** apresenta um gráfico de barras com as 6 *features* mais importantes, destacando as variáveis que mais impactam o desempenho do modelo.

H. Ajustes Finos no Modelo

Conforme metodologia definida, *Grid Search* foi aplicado ao classificador, com o objetivo de otimizar métricas importantes como *Acurácia*, *Precisão*, *Recall* e *F1 Score*. O processo consistiu em definir uma grade de valores para cada hiperparâmetro relevante, e o *Grid Search* avaliou o desempenho do modelo para cada combinação, utilizando validação cruzada para garantir uma estimativa robusta da performance.

Os *hiperparâmetros* selecionados como os mais eficazes para a previsão dos dados foram:

- **ccp_alpha:** 0.0

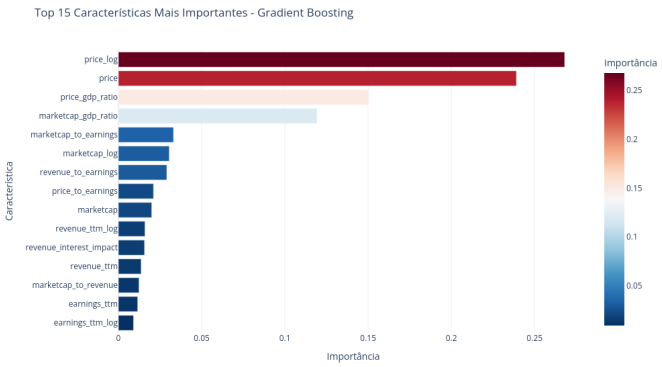

Fig. 11: Features Mais Importantes do Modelo

- **criterion:** friedman_mse
- **init:** None
- **learning_rate:** 0.05
- **loss:** log_loss
- **max_depth:** 5
- **max_features:** None
- **max_leaf_nodes:** None
- **min_impurity_decrease:** 0.0
- **min_samples_leaf:** 1
- **min_samples_split:** 2
- **min_weight_fraction_leaf:** 0.0
- **n_estimators:** 200
- **n_iter_no_change:** None
- **random_state:** 42
- **subsample:** 0.8
- **tol:** 0.0001
- **validation_fraction:** 0.1
- **verbose:** 0
- **warm_start:** False

Esses parâmetros foram ajustados com o objetivo de otimizar o desempenho do modelo, garantindo um bom equilíbrio entre viés e variância. Por exemplo, o parâmetro *learning_rate* com valor 0,05 determina a taxa com que o modelo ajusta os pesos durante o treinamento, influenciando diretamente a velocidade de aprendizagem e a capacidade de generalização. O *max_depth* limitado a 5 restringe a complexidade das árvores, ajudando a evitar o sobreajuste (*overfitting*). O critério *friedman_mse* é utilizado para medir a qualidade das divisões, otimizando a redução de erro em árvores de regressão. O valor de *subsample* igual a 0,8 indica que cada árvore é treinada com 80% dos dados disponíveis, o que contribui para a redução da variância do modelo. O número de estimadores (*n_estimators*) foi definido como 200, fornecendo uma quantidade suficiente de árvores para melhorar a performance preditiva sem comprometer excessivamente o custo computacional. Parâmetros como *min_samples_leaf*, *min_samples_split* e *min_impurity_decrease* controlam a forma como as árvores são construídas, evitando divisões muito específicas que poderiam levar ao *overfitting*.

A utilização do *Grid Search* da *scikit-learn* possibilitou uma otimização sistemática e transparente dos *hiperparâmetros*. Apesar de não ter ocorrido um aumento considerável nas métricas de desempenho, esses parâmetros representam a melhor configuração obtida para o modelo. Esse processo

de ajuste fino é uma prática fundamental para melhorar a generalização e a capacidade preditiva do *modelo* em dados novos, o que é essencial em problemas de *machine learning* em contextos reais (*Akiba et al. 2019*).

I. Análise de Distribuição das Previsões

1) Potencial de Crescimento por País: O gráfico na **Figura 12** apresenta a distribuição do *potencial de crescimento* por país, evidenciando as diferenças nas previsões geradas pelo modelo para os países analisados. A análise visual permite observar quais países se destacam positivamente em termos de *potencial de crescimento* e quais apresentam um desempenho mais modesto.

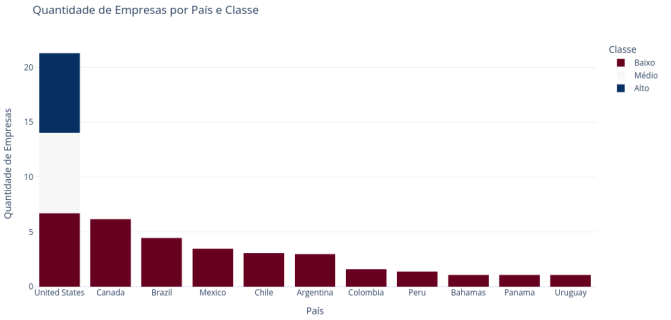

Fig. 12: Potencial de Crescimento por País

O gráfico mostra nitidamente que os **Estados Unidos** concentram a maior parte das empresas com alto e muito alto *potencial de crescimento*. Este fato é coerente com a literatura, visto que os EUA têm um dos mercados de capitais mais desenvolvidos do mundo, com forte presença de empresas listadas na bolsa e grande volume de capitalização de mercado *(La Porta et al., 1997)*.

Além disso, o **Canadá** aparece como segundo país com mais empresas com bom desempenho previsto, embora em escala bem menor. Países latino-americanos, como **Brasil**, **México** e **Argentina**, possuem um número muito reduzido de empresas com alto potencial, o que pode estar associado a fatores macroeconômicos como instabilidade cambial, políticas fiscais restritivas e baixa participação no mercado global *(World Bank, 2023)*.

2) Potencial de Crescimento por Continente: Além da análise por país, também investigamos o *potencial de crescimento* a nível continental representado na **Figura 13**. Essa abordagem permite identificar tendências globais, como a superioridade de certos continentes no que diz respeito ao *potencial de crescimento*, o que pode indicar diferentes condições econômicas ou oportunidades de crescimento.

O gráfico por continente reforça a análise anterior: a **América do Norte** domina amplamente todas as classes de *potencial de crescimento*, em especial as classificações mais elevadas. Isso está alinhado com estudos que indicam que empresas localizadas em regiões com infraestrutura de mercado consolidada, acesso facilitado a financiamento e ambiente regulatório estável tendem a crescer mais *(Demirgüç-Kunt & Levine, 2001)*.

Já a **América do Sul** e **América Central** aparecem com participação residual, com predominância de empresas na categoria de potencial **baixo a mediano**. Isso pode ser interpretado

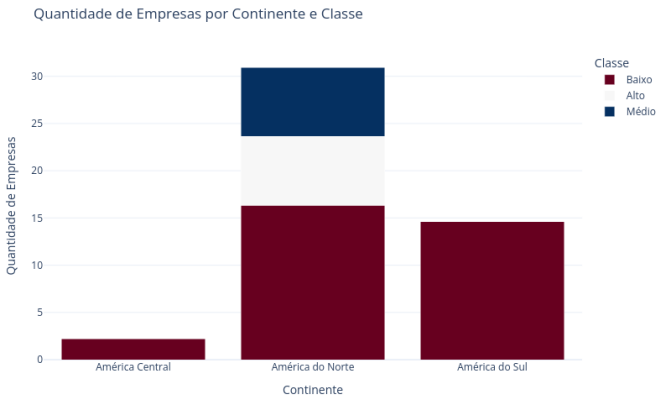

Fig. 13: Potencial de Crescimento por Continente

como reflexo de economias emergentes que enfrentam desafios estruturais para a escalabilidade dos negócios *(OECD, 2022)*.

J. Erros

Compreender as fontes de erro é fundamental para aprimorar a precisão das previsões. Segundo *Silva (2023)*, erros em modelos preditivos podem surgir devido à qualidade dos dados, à escolha inadequada do modelo ou à utilização incorreta de variáveis. A identificação dessas fontes permite que os modelos sejam ajustados de forma mais precisa, resultando em previsões mais confiáveis.

1) Variáveis que Mais Influenciam nos Erros: A **Figura 14** apresenta as variáveis que mais impactam os erros de previsão, com base na análise da diferença entre as médias dos valores dos parâmetros, agrupados conforme a classificação do modelo (previsão correta ou incorreta). Esse tipo de visualização facilita a identificação dos fatores críticos que devem ser ajustados para melhorar a performance e A identificação das fontes de erro é essencial para melhorar a acurácia das previsões. Conforme destacado por Silva (2023), essas falhas podem decorrer da qualidade dos dados, da escolha inadequada do modelo ou da incorreta seleção de variáveis. Tais *insights* orientam aprimoramentos contínuos no processo preditivo. Segundo Silva (2023), erros em modelos preditivos podem surgir devido à qualidade dos dados, à escolha inadequada do modelo ou à utilização incorreta de variáveis. A identificação dessas fontes permite que os modelos sejam ajustados de forma mais precisa, resultando em previsões mais confiáveis. modelo, especialmente quando se trata de erros previsíveis. A utilização de visualizações desse tipo é uma prática recomendada para melhorar a transparência dos modelos e a interpretação de resultados *(Inglis, Parnell & Hurley, 2021)*.

A visualização revela que a variável com maior impacto nas diferenças entre previsões corretas e erradas é o **pe_ratio_ttm**.

O *price-to-earnings ratio* (**pe_ratio_ttm**) é amplamente utilizado em finanças como indicador de avaliação de empresas, refletindo a relação entre o preço da ação e o lucro por ação. A presença dessa variável como principal fonte de erro nas previsões sugere que empresas com relações preço/lucro extremas, seja por expectativas de crescimento exageradas ou por lucros instáveis, representam maior desafio para os modelos de classificação.

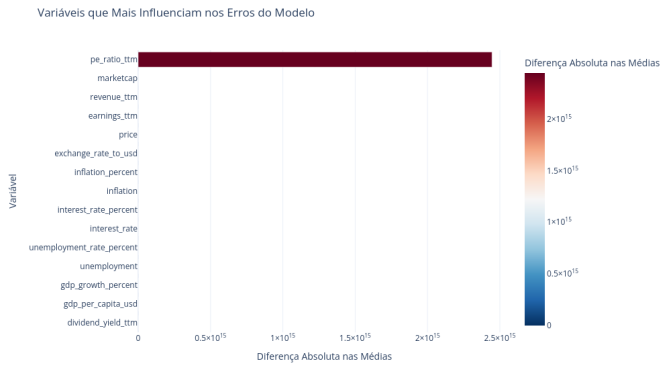

Fig. 14: Variáveis que Mais Influenciam nos Erros de Previsão

K. Implementação

A implementação de *modelos* de *machine learning* consiste em disponibilizar um *modelo* treinado para fazer previsões ou tomar decisões com base em novos dados, transformando *modelos* analíticos em ferramentas práticas para as organizações *(Databricks, 2022)*.

O **Streamlit** é um *framework Python* de código aberto que permite criar aplicativos web interativos para ciência de dados e *machine learning* de maneira simples, sem a necessidade de experiência em desenvolvimento web *streamlit*. Ao integrar *modelos* com *Streamlit*, é possível construir interfaces intuitivas para interação, visualização de resultados e análise de dados de forma acessível e eficiente *(streamlit Doc., 2019)*.

1) Entrada de Dados: A entrada de dados foi implementada de maneira a possibilitar que o usuário realize o *upload* de arquivos no formato *.csv*, representada na **Figura 15**. A utilização do formato *.csv* facilita a transferência de dados tabulares, sendo amplamente compatível com diversas ferramentas e sistemas de análise, o que contribui para a eficiência e a acessibilidade do processo de análise de dados.

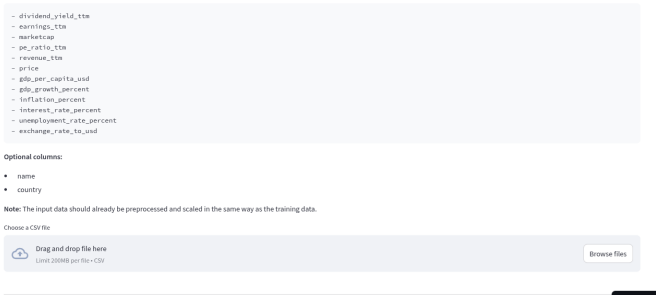

Fig. 15: Entrada de Dados Via *upload*

2) Previsão: Após o *upload* do arquivo *.csv* contendo os dados das empresas, o *modelo* de *aprendizado de máquina* realiza a previsão do *potencial de crescimento* assim que o usuário aciona o botão *"Prever"*, localizado abaixo da visualização geral do *dataset*, vide **Figura 16**.

3) Saída de Dados: Após o usuário clicar no botão *"Prever"*, o *modelo* gera as previsões do *potencial de crescimento* para cada empresa presente no *dataset*. Como resultado, o *dataset* com as previsões é exibido diretamente na interface do aplicativo juntamente com uma análise estatística dos valores previstos. Esta visualização permite que o usuário observe os

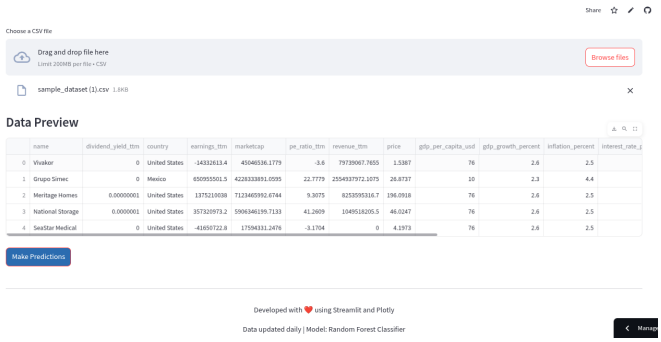

Fig. 16: Visualização dos Dados e Botão para Prever

dados originais juntamente com a nova coluna, que contém os valores previstos de *potencial de crescimento*.

Além disso, uma opção adicional é disponibilizada ao usuário: um botão para *Download* do *dataset*. Ao clicar neste botão, o usuário tem a possibilidade de baixar o arquivo *.csv* contendo os dados originais, agora com a coluna de previsões inclusa. O arquivo é automaticamente salvo na pasta de *Downloads* do computador do usuário, garantindo uma fácil acessibilidade para futuros acessos e análises. Este processo é evidenciado nas **Figuras 17 e 18**.

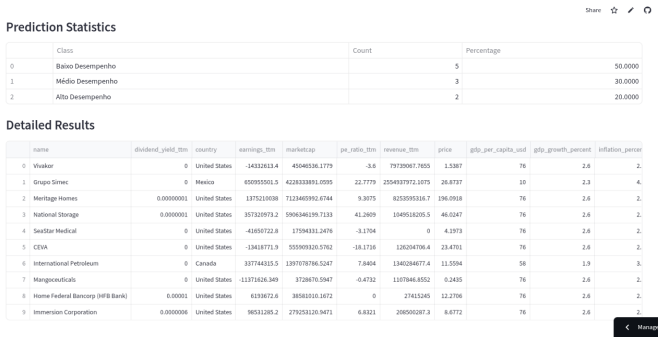

Fig. 17: Dados Previstos e Análise Estatística

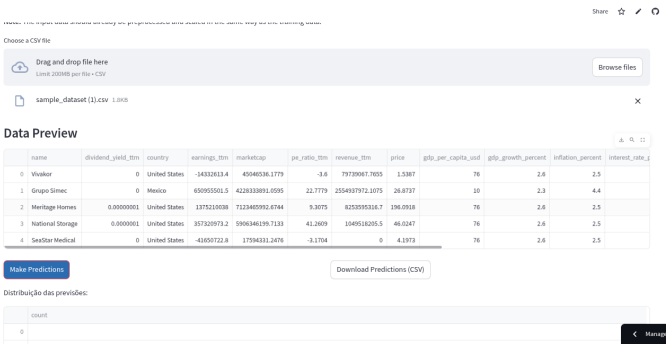

Fig. 18: Botão de Download

V. Conclusão e Sugestões para Futuras Pesquisas

A aplicação de *modelos* de *aprendizado de máquina* para a previsão do *potencial de crescimento* de empresas, com base em dados financeiros e macroeconômicos, demonstrou-se eficaz. O tratamento de um conjunto de dados caracterizado por distribuições assimétricas, presença de *outliers* e variáveis com

escalas heterogêneas exigiu a utilização de técnicas robustas de pré-processamento e modelagem, capazes de gerar previsões confiáveis mesmo diante de um cenário de alta complexidade.

Dentre os algoritmos avaliados, o *Gradient Boosting* destacou-se por apresentar o melhor desempenho preditivo. A análise das métricas obtidas evidenciou sua superioridade em termos de acurácia, equilíbrio entre classes e área sob a *Curva ROC*, o que indica elevada capacidade discriminativa. A *Matriz de Confusão* também revelou uma classificação consistente das empresas em diferentes níveis de *potencial de crescimento*, com baixos índices de falsos positivos e falsos negativos, tornando o modelo particularmente adequado para apoiar decisões estratégicas.

Adicionalmente, a construção de uma interface interativa utilizando a biblioteca *Streamlit* possibilitou a implementação de um sistema acessível, capaz de oferecer visualizações intuitivas e interpretáveis, mesmo por usuários sem conhecimento técnico avançado. Esta abordagem favorece a integração entre ciência de dados e gestão empresarial, ao disponibilizar uma ferramenta prática para análise preditiva em contextos corporativos e institucionais.

Entre as limitações observadas, destaca-se a dependência do modelo à qualidade e abrangência dos dados de entrada. A ausência de variáveis qualitativas e contextuais, como fatores políticos, indicadores de governança corporativa e aspectos setoriais, pode limitar a capacidade de generalização do modelo em diferentes cenários econômicos ou regiões geográficas. Além disso, a classificação do *potencial de crescimento* baseou-se em uma métrica composta, cuja estrutura, embora fundamentada teoricamente, ainda carece de validação empírica mais ampla.

Para pesquisas futuras, sugere-se a incorporação de variáveis adicionais que representem aspectos subjetivos ou qualitativos das empresas, bem como o uso de abordagens mais avançadas, como *modelos híbridos* (combinando algoritmos de aprendizado supervisionado e não supervisionado) e técnicas de *deep learning*. A realização de testes em contextos temporais distintos e a aplicação em setores específicos também podem contribuir para a avaliação da robustez e da adaptabilidade do modelo proposto.

Conclui-se que a presente pesquisa representa uma contribuição relevante para o campo da análise preditiva no ambiente empresarial, ao demonstrar o potencial do *Gradient Boosting* na classificação de empresas segundo seu desempenho futuro estimado. Os resultados obtidos reforçam a importância da integração entre métodos quantitativos e ferramentas computacionais no apoio à tomada de decisão estratégica baseada em dados.

Para visualizar o *aplicativo web* desenvolvido e testar as funcionalidades do *modelo* preditivo, acesse o seguinte link: `https://tcc-sidnei93.streamlit.app/`*aplicativo web*.

VI. Bibliografia

[1] Mitchell, T. M. *Machine Learning and Data Mining*. Communications of the ACM, vol. 42, no. 11, pp. 30-36, 1999. Disponível em: https://dl.acm.org/doi/pdf/10.1145/319382.319388. Acesso em: 23 abr. 2025.

[2] Dio, J. *Aplicações de Machine Learning no Setor Financeiro*. Journal of Finance, 2022, vol. 45, pp. 112-125.

[3] Breiman, L. *Random Forests*. Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324. Acesso em: 23 abr. 2025.

[4] Yu, H. *Application of Machine Learning Algorithm in Financial Industry*. In: E. Azar and A.N. Haddad, editors, *Artificial Intelligence in the Gulf*, pp. 95–115. Springer, Cham, 2022. DOI: https://doi.org/10.1007/978-3-030-97874-7_126.

[5] Zhang, L. *The Application of Machine Learning in Finance: Situation and Challenges*. SSRN Electronic Journal, 2024. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5120025. Acesso em: 23 abr. 2025.

[6] Rao, A. S.; Vardhan, B. V.; Shaik, H. *Role of Exploratory Data Analysis in Data Science*. 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1457-1461, IEEE, 2021. Disponível em: https://ieeexplore.ieee.org/document/9546555https://ieeexplore.ieee.org/document/9546555. Acesso em: 23 abr. 2025.

[7] Han, J., Kamber, M., & Pei, J. *Data Mining: Concepts and Techniques*. 3rd ed., Morgan Kaufmann, 2012.

[8] AWS. *Limpeza e Manipulação de Dados: Técnicas e Ferramentas*. Amazon Web Services, 2023.

[9] Astera. *Transformação de Dados Brutos em Informações Estruturadas*. 2023.

[10] Graham, B., Dodd, D. *Security Analysis: Sixth Edition*. McGraw-Hill, 2008.

[11] Damodaran, A. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*. Wiley, 2012.

[12] Montgomery, D., Runger, G. *Applied Statistics and Probability for Engineers*. Wiley, 2021.

[13] Bowerman, B., O'Connell, R. *Business Statistics in Practice*. McGraw-Hill, 2016.

[14] Miller, R. *Quantitative Analysis for Business*. McGraw-Hill, 1991.

[15] Hoaglin, D., Mosteller, F., Tukey, J. *Understanding Robust and Exploratory Data Analysis*. Wiley, 1983.

[16] Brownlee, J. *Imbalanced Classification: Best Practices and Algorithms*. Machine Learning Mastery, 2020. Disponível em: https://machinelearningmastery.com/imbalanced-classification-with-python/. Acesso em: 23 abr. 2025.

[17] Akiba, T., Sano, S., Ohta, T., Koyama, M., and Aizawa, A. *Optuna: A Next-generation Hyperparameter Optimization Framework*. 2019. Disponível em: https://arxiv.org/abs/1907.10900. Acesso em: 23 abr. 2025.

[18] Databricks. *Deploying Machine Learning Models in Production: Best Practices*. 2022.

[19] Streamlit. *Streamlit Documentation*. 2019. Disponível em: https://docs.streamlit.iohttps://docs.streamlit.io. Acesso em: 23 abr. 2025.

[20] Pinto, P. and Carvalho, M. *Previsão de risco de crédito utilizando aprendizado de máquina: um estudo de caso com XGBoost*. 2020.

[21] Guimarães, R. *Aplicação de técnicas de aprendizado de máquina para previsão da produtividade da soja utilizando Redes Neurais e Random Forest*. 2020.

[22] Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Springer, 2020. Disponível em: https://christophm.github.io/interpretable-ml-book/. Acesso em: 23 abr. 2025.

[23] Chen, T., Guestrin, C. *XGBoost: A scalable tree boosting system*. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, 2016. Disponível em: https://dl.acm.org/doi/10.1145/2939672.2939785. Acesso em: 23 abr. 2025.

[24] La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1997). Legal determinants of external finance. *The Journal of Finance*, 52(3), 1131–1150. Disponível em: https://scholar.harvard.edu/shleifer/files/legaldeterminants.pdf. Acesso em: 23 abr. 2025.

[25] Demirgüç-Kunt, A., & Levine, R. (2001). Financial structure and economic growth: A cross-country comparison of banks, markets, and development. *MIT press*.

[26] OECD. (2022). Latin American Economic Outlook 2022: Towards a Green and Just Transition. *OECD Publishing*. Disponível em: https://www.oecd-ilibrary.org/development/latin-american-economic-outlook/22185671. Acesso em: 23 abr. 2025.

[27] Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, 47(2), 427–465. Disponível em: https://onlinelibrary.wiley.com/doi/abs/

10.1111/j.1540-6261.1992.tb04398.x. Acesso em: 23 abr. 2025.

[28] World Bank. (2023). World Development Indicators. *https://data.worldbank.org*. Acesso em: 23 abr. 2025.

[29] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. https://projecteuclid.org/euclid.aos/1013203451. Acesso em: 23 abr. 2025.

[30] Fraunhofer Institute. (2024). A sparse grid based method for generative dimensionality reduction of high-dimensional data. *Fraunhofer Institute for Algorithms and Scientific Computing*. Disponível em: https://www.ais.fraunhofer.de/en/business-areas/bioinformatics/publications.html. Acesso em: 27 maio 2025.