Universidade Presbiteriana Mackenzie

CIÊNCIA DE DADOS

Projeto Aplicado 02

Curso Ciência de Dados

Universidade Presbiteriana Mackenzie

2° semestre – ano letivo: 2025

Universidade Presbiteriana Mackenzie

CIÊNCIA DE DADOS

FinData ANALYTICS

Este documento detalha a proposta de projeto da FinData Analytics, uma consultoria de tecnologia e dados focada em fornecer soluções para o setor financeiro. é baseado no uso de técnicas e ferramentas desenvolvidas no curso de Ciência de Dados da Universidade Presbiteriana Mackenzie com a orientação do professor Felipe Albino dos Santos.

ALBERTO NAGEM – 10720713 / ANA JULIA SAMPAIO - 10724346
DIOGO LIMA PEREIRA – 10727456 / GABRIEL MENDES - 10720824

Sumário

Introdução

A concessão de crédito é um pilar da economia, mas está intrinsecamente ligada ao risco de inadimplência. Decisões imprecisas podem resultar em perdas financeiras significativas para credores. Atualmente, muitos dados valiosos sobre o cenário macroeconômico e o comportamento do consumidor são publicados em formatos não estruturados, como os relatórios em PDF do Serasa. O desafio e a oportunidade residem na capacidade de extrair, consolidar e analisar sistematicamente essas informações para aprimorar os modelos de avaliação de risco existentes.

A proposta da FinData Analytics é desenvolver uma solução de ponta a ponta que extrai e processa dados textuais e visuais de relatórios financeiros para construir um modelo preditivo de risco de crédito. Isso é feito por meio de um processo que automatiza a extração e estruturação desses dados, aplica modelos de Machine Learning para gerar previsões de risco (como o VTDD) e entrega os resultados em dashboards visuais, ajudando gestores a tomar decisões mais rápidas e embasadas na concessão de crédito.

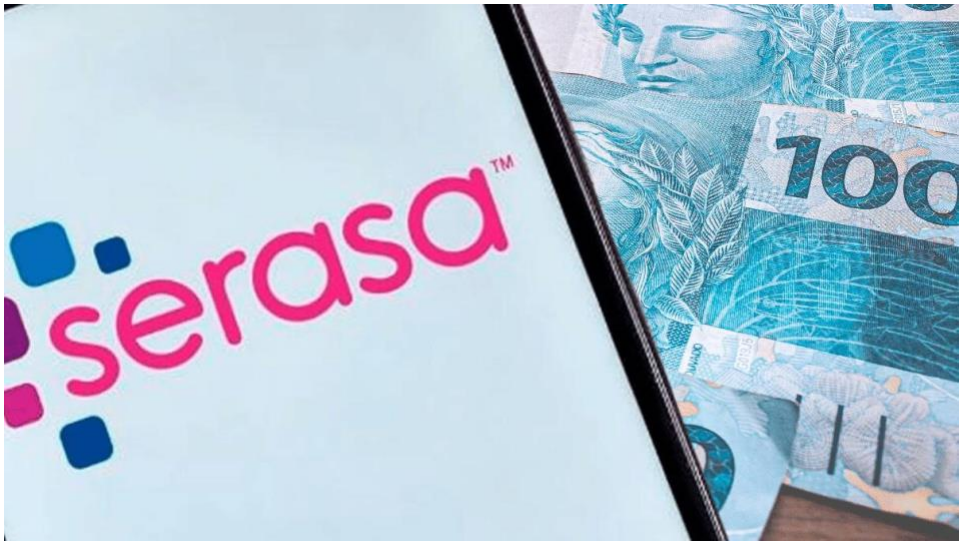

Figura 1 - SERASA

Modelo de Negócios

O modelo de negócios da FinData Analytics fundamenta-se na oferta de soluções analíticas avançadas para instituições financeiras, empresas de crédito e fintechs que buscam aprimorar seus processos de avaliação de risco e tomada de decisão. A proposta de valor central consiste em transformar dados públicos não estruturados — especialmente os relatórios de inadimplência do Serasa Experian — em insights acionáveis, previsões confiáveis e segmentações estratégicas de consumidores.

A solução desenvolvida pela empresa adota um modelo B2B (Business-to-Business), com foco em serviços de análise preditiva, extração automatizada de dados e inteligência de crédito. O produto final combina dois pilares principais:

1. Pipeline de Extração e Estruturação de Dados: automatização completa da coleta, padronização e consolidação de informações provenientes de PDFs e elementos visuais. Esse componente reduz custos operacionais, elimina trabalho manual e garante um fluxo contínuo de dados atualizados.

2. Motor Analítico e Modelos de Machine Learning: aplicação de algoritmos de classificação, agrupamento e previsão para identificar perfis de risco, mapear tendências de inadimplência e projetar indicadores futuros, como o Valor Total das Dívidas (VTDD). Esse módulo fornece previsões mensais de risco e relatórios de performance que auxiliam na concessão de crédito.

Objetivos Específicos e Metodologia

Etapas

Para alcançar o objetivo geral, as seguintes etapas serão executadas:

- Extração de Dados: Utilização da biblioteca pdfplumber (Python) para a extração de texto e aplicação de técnicas de visão computacional (e.g., OpenCV, Tesseract) para extrair informações de tabelas e gráficos contidos nos relatórios do Serasa;
- Estruturação de Dados: Consolidação das informações extraídas em um dataset estruturado, garantindo a integridade e a qualidade dos dados;
- Análise Exploratória de Dados (EDA): Investigação do dataset para identificar tendências, padrões e correlações, além de realizar o tratamento e a engenharia de features necessárias;b
- Desenvolvimento de Modelos: Aplicação de algoritmos de Machine Learning (e.g., Regressão Logística, Random Forest, Gradient Boosting) para treinar um modelo de classificação de perfis de risco;
- Validação e Métricas: Avaliação da performance dos modelos utilizando uma metodologia de validação cruzada e métricas como acurácia, matriz de confusão, precisão e F1-score.

Entregáveis

1. Dataset Consolidado: Base de dados limpa e estruturada, pronta para análise.

2. Código-Fonte: Repositório no GitHub contendo todo o código desenvolvido para extração, tratamento e modelagem.

3. Relatório Técnico: Documento detalhando a metodologia, as análises realizadas, os resultados dos modelos e as conclusões do projeto.

4. Apresentação de Resultados: Um storytelling em formato de vídeo, simulando a entrega da solução para um cliente final, focando nos insights de negócio e no valor gerado.

Metodologia de Aquisição e Estruturação de Dados

Esta seção descreve a fonte primária de informações e o processo planejado para a construção do dataset que servirá como alicerce para este projeto. É importante ressaltar que um dos principais entregáveis desta iniciativa é a criação de um dataset estruturado a partir de fontes públicas não estruturadas.

O projeto será desenvolvido em Python, linguagem escolhida pela sua robustez no processamento de dados e pela ampla disponibilidade de bibliotecas específicas para ciência de dados.

As bibliotecas adotadas foram:

- PDFplumber: para a extração de informações textuais e tabulares diretamente dos relatórios em formato PDF;
- Tesseract: para a aplicação de técnicas de visão computacional (em conjunto com ferramentas como OpenCV) para extrair informações de tabelas e gráficos contidos nos relatórios do Serasa, especialmente dados visuais ou de imagens;
- Pandas: para a manipulação e estruturação de dados tabulares;
- NumPy: para o suporte a cálculos numéricos e estatísticos;
- Scikit-learn: para a implementação dos modelos de classificação e agrupamento;
- Matplotlib e Seaborn: para a visualização de dados e análise exploratória.

O conjunto de ferramentas selecionado atende às necessidades do projeto, abrangendo desde a extração dos dados brutos até as etapas de análise e modelagem preditiva.

Origem dos Dados:

A fonte de dados selecionada para este projeto são os relatórios públicos do "Mapa de Inadimplência e Renegociação de Dívidas no Brasil", disponibilizados periodicamente pela Serasa Experian. Esses documentos representam uma fonte de informação rica e de alta credibilidade sobre o cenário do crédito e o comportamento financeiro da população brasileira.

- Os relatórios são publicados em formato PDF e contêm um conjunto diversificado de informações, que incluem:
- Dados Textuais: Análises conjunturais, comentários de especialistas e explicações sobre as tendências observadas.
- Dados Tabulares: Tabelas detalhando a inadimplência por faixas etárias, sexo, faixas de renda, regiões geográficas (estados e capitais) e setores da economia.
- Elementos Visuais: Gráficos (barras, linhas, pizza) e infográficos que ilustram a evolução de indicadores, distribuições percentuais e comparações históricas.

A natureza heterogênea e não estruturada desses documentos constitui o principal desafio técnico a ser superado, justificando a aplicação de técnicas de extração de texto e visão computacional.

Período da Coleta:

O dataset deste projeto será construído progressivamente. A estratégia de coleta foi dividida em duas fases:

1. Fase Inicial (Escopo Mínimo Viável): A análise primária será focada no relatório mais recente disponível no início do projeto, correspondente a Julho de 2025. Esta abordagem permitirá a criação de um snapshot detalhado do cenário de inadimplência atual, servindo como base para o desenvolvimento e a validação inicial dos pipelines de extração de dados e dos modelos preditivos.

OBS: Fase Finalizada juntamente com a primeira entrega.

2. Fase de Expansão (Análise Histórica): Conforme a evolução do projeto e a validação da metodologia, o escopo será expandido para incluir o histórico completo de relatórios mensais do ano de 2025. A incorporação de dados históricos é estratégica e visa enriquecer a análise de múltiplas formas:

- Identificação de Sazonalidade: Permitirá analisar se existem padrões de inadimplência que se repetem em determinados períodos do ano (ex: pós-festas, início de ano).
- Análise de Tendências: Possibilitará a observação da evolução da inadimplência ao longo do tempo, gerando insights sobre o impacto de fatores macroeconômicos.
- Robustez do Modelo: Um dataset com maior variedade temporal tende a gerar modelos de aprendizado de máquina mais robustos e generalizáveis.

A consolidação desses múltiplos relatórios em um único dataset estruturado e coeso será a base fundamental que permitirá a aplicação das análises exploratórias e preditivas propostas nos objetivos deste trabalho.

OBS: Fase Finalizada juntamente com a segunda entrega.

Desenvolvimento

Para a extração e consolidação dos dados dados, foi desenvolvido o script pdf2csv.py, com as seguintes finalidades:

- Extrair as informações de interesse dos relatórios (valores financeiros e quantitativos);
- Consolidar os dados em um formato tabular padronizado (CSV);
- Garantir a consistência e a padronização das variáveis de interesse.

O arquivo serasa.csv não constitui a base de dados primária, mas sim o resultado do processo de extração e tratamento inicial aplicado aos relatórios. Este arquivo representa a versão estruturada dos dados que será utilizada nas etapas subsequentes do projeto.

Principais variáveis da análise são:

- INADIMPLENTES_MI – número de inadimplentes no Brasil (em milhões de pessoas);
- VMPP – valor médio em dívida por pessoa inadimplente (em R$);
- DIVIDAS_MI – quantidade total de dívidas em aberto (em milhões);
- VMCD – valor médio de cada dívida (em R$);
- VTDD_BI – valor total das dívidas, em bilhões de reais;
- VMAF – valor médio dos acordos realizados (em R$);
- DESCONTOS_BI – volume total de descontos concedidos, em bilhões de reais.

Tratamento e Pré-processamento dos Dados

O tratamento da base de dados envolveu as seguintes etapas:

- Limpeza de dados (Data Cleaning): correção de ruídos provenientes da extração, como textos ou símbolos monetários em campos numéricos;
- Conversão de tipos: transformação de variáveis textuais em formatos numéricos adequados para a modelagem;
- Normalização: padronização das escalas monetárias para garantir a comparabilidade entre as variáveis;
- Engenharia de atributos (Feature Engineering): criação de novas variáveis a partir das existentes (ex.: a razão inadimplentes/dívidas) para potencializar o desempenho dos modelos;
- Divisão dos dados (Data Splitting): separação da base em conjuntos de treino (70%) e teste (30%), preservando a ordem cronológica para garantir a consistência temporal.

Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) foi realizada sobre o conjunto de dados consolidado (serasa.csv) com o objetivo de identificar padrões, correlações e tendências temporais.

As etapas desta análise incluíram:

- Análise de estatísticas descritivas: para examinar a distribuição de inadimplentes, as médias de dívidas e outras métricas centrais;
- Criação de visualizações gráficas: elaboração de histogramas, boxplots e gráficos de séries temporais para evidenciar padrões como sazonalidade (ex.: aumento da inadimplência em determinados períodos);
- Investigação de correlações: análise da relação entre variáveis, como a influência da quantidade de dívidas no valor médio por pessoa ou na taxa de inadimplência.

Análise de estatísticas descritivas

Após a etapa de aquisição e estruturação, foi construída uma base consolidada a partir dos relatórios “Mapa da Inadimplência” do Serasa referentes ao período de outubro de 2024 a julho de 2025.

O dataset final contempla, para cada mês, os indicadores macroeconômicos citados no início da seção: (INADIMPLENTES_MI, VMPP, DIVIDAS_MI, VMCD, VTDD_BI, VMAF e DESCONTOS_BI)

A EDA teve como objetivo compreender a dinâmica desses indicadores ao longo do tempo, identificar tendências e relações entre variáveis e preparar a base para a fase de modelagem preditiva.

A análise a seguir concentra-se em três variáveis principais extraídas dos relatórios do Serasa Experian: o Valor Total das Dívidas (VTDD_BI) e o Número de Inadimplentes (INADIMPLENTES_MI).

	PERIODO	INADIMPLENTES_MI	VMPP	DIVIDAS_MI	VMCD	VTDD_BI	VMAF	DESCONTOS_BI	t	PERIODO_FULL
0	out/24	73.10	5504.33	276.08	1457.48	402.03	734.83	10.51	0	Outubro 2024
1	nov/24	73.79	5558.30	274.54	1493.92	410.14	590.74	15.60	1	Novembro 2024
2	dez/24	73.51	5496.69	275.68	1465.73	404.07	560.00	13.40	2	Dezembro 2024
3	jan/25	74.60	5617.00	281.25	1489.90	419.00	676.00	10.20	3	Janeiro 2025
4	fev/25	75.00	5837.49	286.11	1530.28	437.00	698.00	12.10	4	Fevereiro 2025
5	mar/25	75.70	5793.66	287.60	1526.41	438.00	714.00	17.70	5	Março 2025
6	abr/25	76.60	5968.71	294.10	1555.33	457.00	790.00	10.50	6	Abril 2025
7	mai/25	77.00	6036.94	298.50	1558.68	465.00	839.00	11.70	7	Maio 2025
8	jun/25	77.80	6128.26	304.50	1567.05	477.00	772.00	9.90	8	Junho 2025
9	jul/25	78.20	6177.74	307.50	1570.17	482.00	736.00	11.11	9	Julho 2025

Figura 2 - Output EDA Dados SERASA

Criação de visualizações gráficas (Indo mais a fundo)

Nesta seção, serão apresentados os principais insights obtidos através da visualização gráfica, focando na evolução das variáveis-chave para o projeto de risco de crédito: o Valor Total das Dívidas (VTDD) e o Número de Inadimplentes (INADIMPLENTES) no Brasil.

Evolução do número de inadimplentes

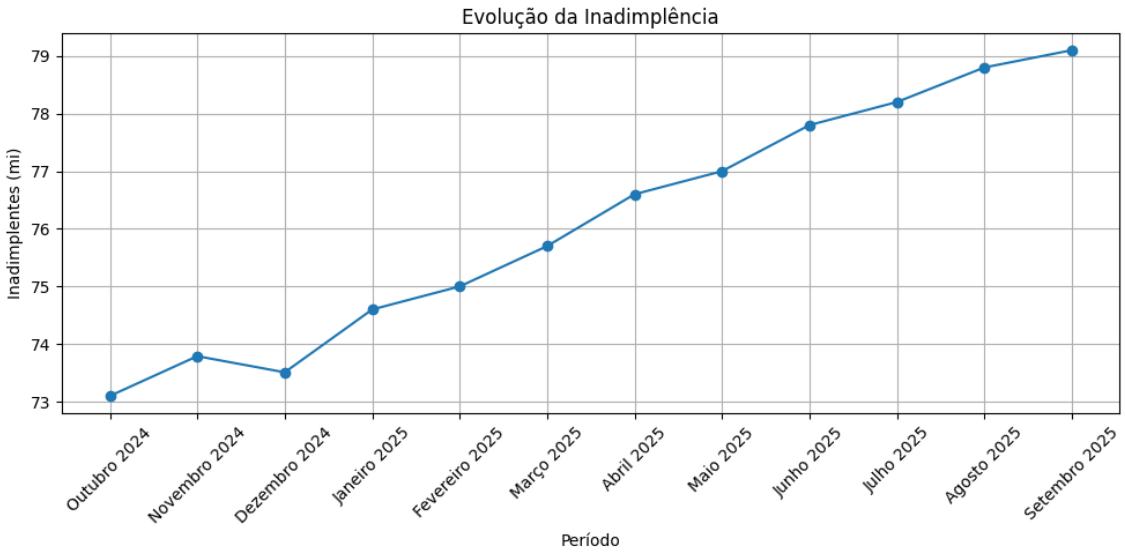

Figura 3 - Evolução da Inadimplência no Brasil

A Figura 3 – Evolução da Inadimplência no Brasil apresenta a série temporal do número de inadimplentes (INADIMPLENTES_MI) entre out/24 e set/25, enquanto a Tabela 1 – Número de Inadimplentes no Brasil consolida os mesmos valores, com a participação relativa de cada mês no total do período.

Número de Inadimplentes no Brasil Out/24 - Set/25			
Mês	**Ano**	**% Total**	**Qtd. Total**
Outubro	2024	39.90%	73.1M
Novembro	2024	40.53%	73.79M
Dezembro	2024	40.49%	73.51M
Janeiro	2025	41.02%	74.6M
Fevereiro	2025	41.30%	75M
Marco	2025	41.66%	75.7M
Abril	2025	41.97%	76.6M
Maio	2025	42.10%	77M
Junho	2025	41.57%	77.8M
Julho	2025	41.94%	78.2M
Agosto	2025	48.31%	78.8M
Setembro	2025	48.47%	79.1M
Total:			**913.2M**

Tabela 1 - Número de Inadimplentes no Brasil Out/24 - Set/25

A tabela detalha a contagem de inadimplentes por mês no período de Outubro de 2024 a Setemrbo de 2025, em milhões de indivíduos.

- Crescimento Constante: Assim como o valor das dívidas, o número de inadimplentes demonstra um crescimento quase constante no período;
- Pico em Setembro/2025: O ápice é alcançado em Julho de 2025 sendo o último mês analisado, registrando 79.1 milhões de inadimplentes;
- Total de Inadimplentes no Período: A soma da quantidade de inadimplentes no período de Outubro/2024 a Setembro/2025 totaliza 913.2 milhões.

Os resultados indicam que:

- O número de inadimplentes parte de aprox. 73,1 milhões no ponto de partida e atinge 79,1 milhões em seu último registro, o que representa um crescimento acumulado em torno de 6 milhões de pessoas em onze meses;
- A contribuição mensal é relativamente equilibrada: cada mês responde por cerca de 8,0% a 8,6% do total de inadimplentes do período analisado, o que reforça a ideia de nível persistentemente elevado de inadimplência, sem quedas bruscas;
- A trajetória é claramente ascendente, com pequenos ajustes pontuais, evidenciando deterioração gradual da capacidade de pagamento da população.

A Tabela 1 gerada pelo script EDA foi atualizada com base no CSV consolidado e substitui a versão inicial do relatório, que havia sido construída com valores preliminares.

Valor médio por pessoa (VMPP)

A Figura 4 – Valor Médio por Pessoa (VMPP) mostra a evolução do valor médio em dívida por inadimplente.

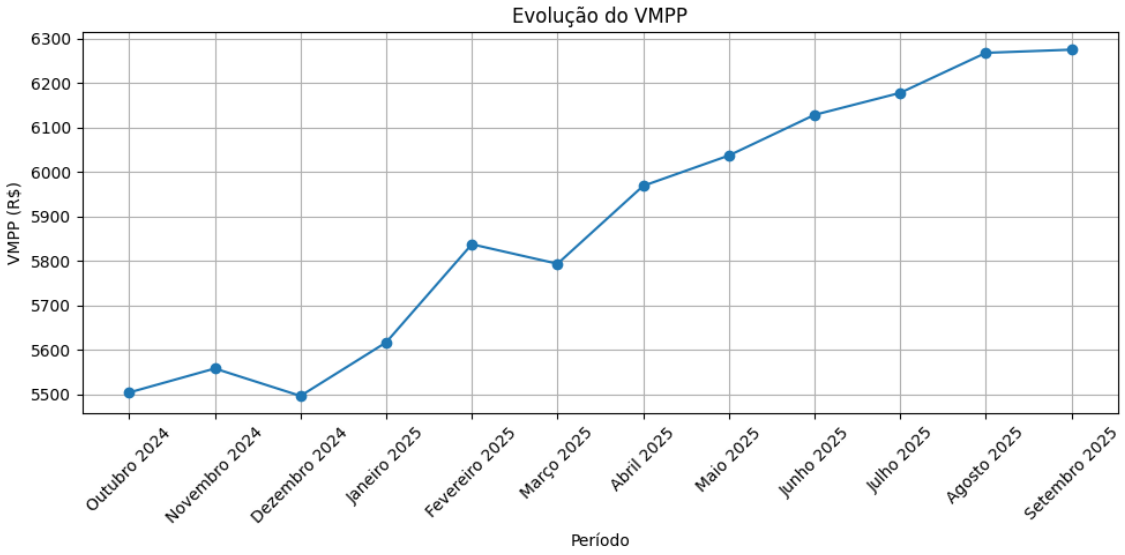

Figura 4 - Valor Médio por Pessoa (VMPP)

Observa-se que:

- O VMPP gira em torno de R$ 5,5 mil em out/24, evoluindo para cerca de R$ 6,27 mil em set/25;
- A tendência é de crescimento contínuo, indicando que não apenas há mais pessoas inadimplentes, como cada pessoa, em média, concentra dívidas maiores ao longo do tempo;
- Esse comportamento reforça o aumento do risco de crédito: a inadimplência deixa de ser apenas um fenômeno de volume de pessoas e passa também a ser um fenômeno de maior intensidade por indivíduo.

Quantidade de dívidas e valor médio por dívida (DIVIDAS_MI e VMCD)

A dinâmica da quantidade de dívidas (DIVIDAS_MI) e do valor médio de cada dívida (VMCD) é apresentada nas Figuras 5 e 6:

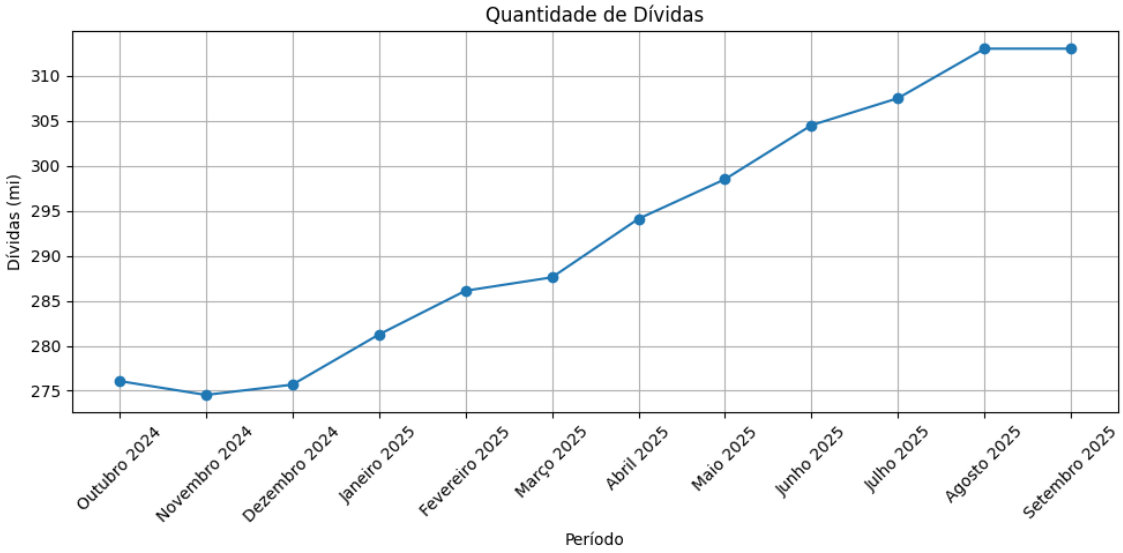

Figura 5 - Quantidade de dívidas no Brasil

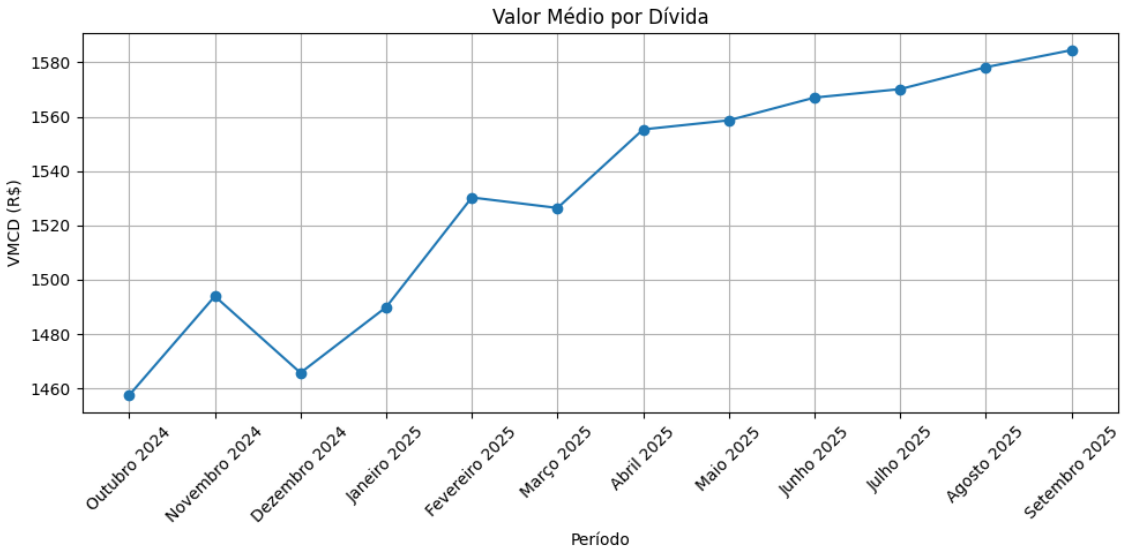

Figura 6 - Valor médio de cada dívida no Brasil

Principais insights:

- O número total de dívidas cresce de aproximadamente 276 milhões em out/24 para algo em torno de 307 milhões em jul/25;
- O valor médio de cada dívida (VMCD) permanece em patamar elevado e também apresenta tendência de alta, saindo de cerca de R$ 1,46 mil e se aproximando de R$ 1,57 mil no final da série;
- A combinação de mais dívidas e dívidas mais caras contribui diretamente para o comportamento crescente do VTDD_BI (valor total de dívidas), discutido na subseção seguinte.

Valor total das dívidas (VTDD_BI)

A **Figura 7 – Valor Total das Dívidas (VTDD)** apresenta a série temporal do montante total devido pelos inadimplentes, em bilhões de reais.

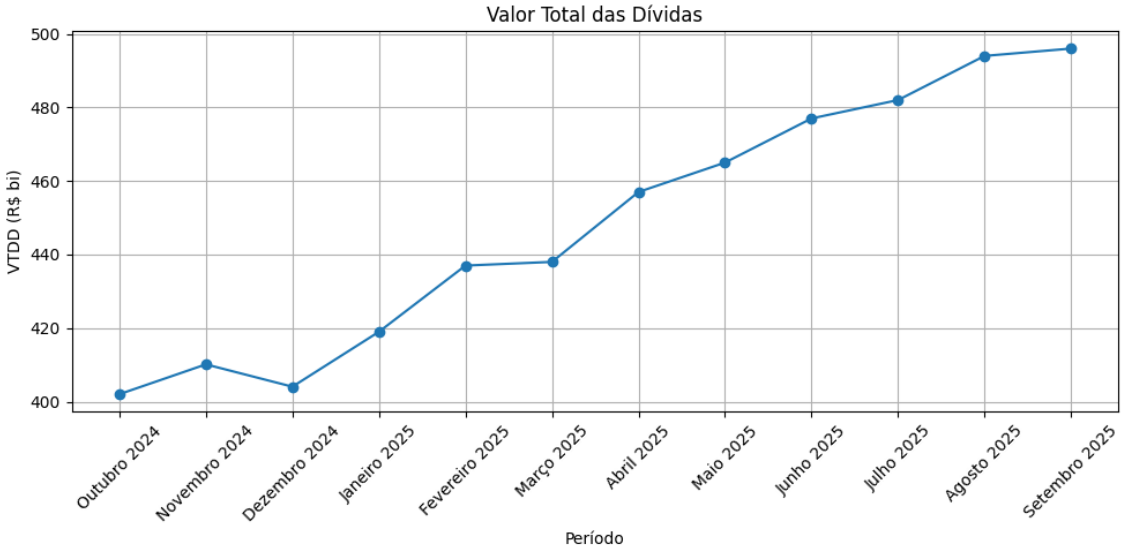

Figura 7 - Valor Total das Dívidas (VTDD)

A análise mostra que:

- Em **out/24**, o VTDD é de aproximadamente **R$ 402 bilhões**;
- Após uma leve oscilação nos meses seguintes, observa-se um movimento de alta mais consistente a partir de **jan/25**, quando o valor chega a **R$ 419 bilhões**;
- O indicador continua crescendo até **set/25**, atingindo cerca de **R$ 496 bilhões**, o que representa um aumento acumulado expressivo no período;
- A tendência de crescimento do VTDD é coerente com o aumento simultâneo do número de inadimplentes, da quantidade de dívidas e dos valores médios (VMPP e VMCD).

Essa visão consolidada reforça a hipótese de **agravamento do risco de crédito** ao longo da janela analisada.

Distribuição da Inadimplência por Faixa Etária

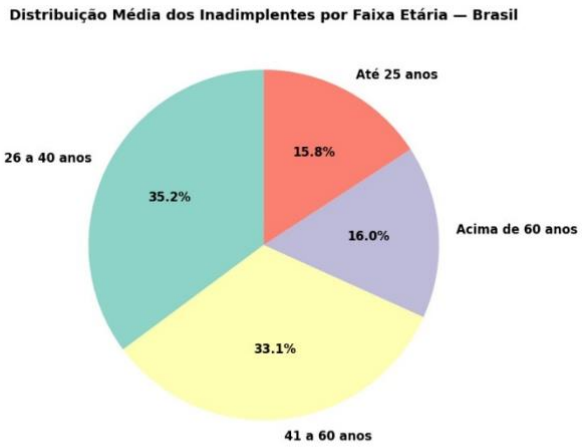

Figura 8 - Distribuição dos inadimplentes por faixa etária

O gráfico acima apresenta a distribuição média do número de inadimplentes no Brasil, segmentada por quatro faixas etárias principais. Este insight é crucial para a fase de Agrupamento (Clustering), que visa segmentar consumidores em perfis com características similares.

- Faixa de Maior Risco: A faixa etária com a maior proporção de inadimplentes é a de 26 a 40 anos, representando 35.2% do total médio;
- Faixa de Risco Secundário: A segunda maior proporção é a de 41 a 60 anos, que corresponde a 33.1% dos inadimplentes;
- Concentração de Risco: As duas faixas intermediárias de idade (26 a 60 anos) somam juntas 68.3% da média total de inadimplentes, indicando que o foco principal do risco de crédito se concentra neste grupo demográfico;
- Faixas com Menor Proporção:
    - grupo Até 25 anos representa 15.8%;
    - grupo Acima de 60 anos representa 16.0%.

Investigação de correlações

Em resumo, a EDA mostrou que, entre um período de quase 1 ano de análise:

- O número de inadimplentes aumentou em 6 milhões de brasileiros passando de aprox. 73,1 para aprox. 79,1 milhões de pessoas;
- O valor médio em dívida por pessoa e o valor médio por dívida cresceram ao longo do período;
- O valor total das dívidas (VTDD) apresentou tendência claramente ascendente, aproximando-se de R$ 496 bilhões em set/25;
- As correlações indicam forte dependência entre VTDD, volume de dívidas e valores médios, justificando o uso dessas variáveis em modelos preditivos;
- A distribuição percentual mostra que a inadimplência é persistente e relativamente homogênea ao longo dos meses, com leve agravamento ao final da série.

Esses achados fundamentam a etapa de modelagem preditiva, cujo objetivo é projetar o VTDD para os cinco meses subsequentes e antecipar a evolução do risco de crédito.

Análise Preditiva e Projeção do Risco de Dívidas:

Após a fase de Análise Exploratória de Dados (EDA) e a modelagem, foi desenvolvido um Modelo Preditivo com o objetivo de projetar o Valor Total das Dívidas (VTDD) no Brasil para os próximos cinco meses (Outubro de 2025 a Fevereiro de 2026).

A escolha de VTDD como variável-alvo é coerente com a proposta de negócio da FinData Analytics: oferecer ao mercado uma visão antecipada do estoque total de dívidas em inadimplência, permitindo a elaboração de estratégias de crédito, cobrança e renegociação.

Métricas de Performance do Modelo:

Embora o SERASA disponibilize relatórios de períodos anteriores, nem todos seguem o mesmo padrão estrutural dos PDFs utilizados neste projeto. Em particular:

- Relatórios mais antigos apresentam layouts distintos, o que exige novos ajustes de código e regras de parsing/OCR para que seus dados possam ser incorporados com segurança ao mesmo CSV;
- Relatórios mais recentes ainda não estavam disponíveis no momento da consolidação da base;
- Por uma questão de consistência metodológica, foram utilizados apenas os 11 relatórios que compartilham o mesmo modelo de PDF, garantindo que o pipeline de extração gere um dataset íntegro e comparável.

Dessa forma, a amostra de 11 meses representa todos os dados viáveis e padronizados até a fase atual do projeto. A expansão histórica (anos anteriores) e futura (novos meses de 2025 e 2026) está planejada como evolução natural do pipeline, à medida que os códigos forem adaptados para outros modelos de PDF.

Apesar desta pequena amostragem de dados, o modelo de série temporal foi validado com as métricas de acurácia a seguir e os resultados são sumarizados nas Tabelas 2 - Métricas de acurácia dados registrados e 3 - Métricas de acurácia futuro; e ilustrados no gráfico “Previsão de Dívidas para os Próximos 5 Meses – Brasil (em Bilhões de R$)”:

- Coeficiente de Determinação: 0.7728
    - O modelo explica 77,28% da variabilidade dos dados históricos do VTDD. Visto que a adequação é confirmada por valores mais próximos de 1, este resultado sinaliza um resultado satisfatório da acurácia do modelo para a previsão;
- Erro Quadrático Médio: R$ 3.8 bilhões
    - Este valor representa a média da magnitude dos erros de previsão do modelo, indicando que, em média, as previsões passadas do VTDD desviaram em cerca de R$ 3.8 bilhões dos valores reais. Dado o volume total das dívidas e o número de amostragens, esta é uma margem de erro elevada.

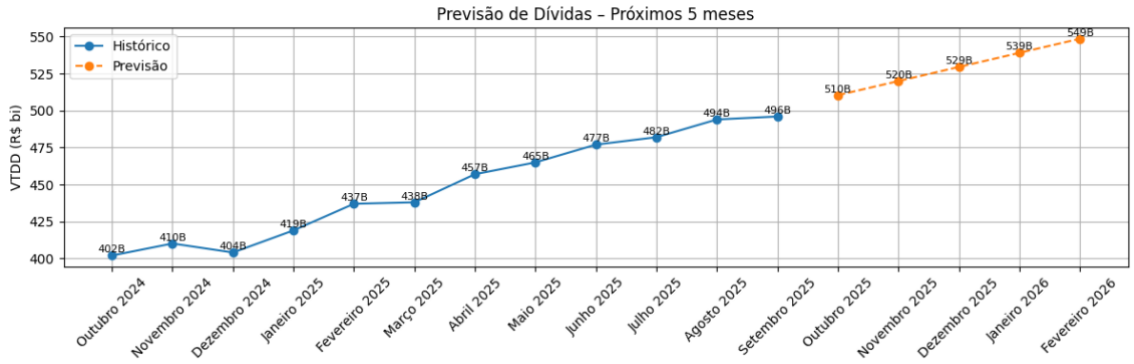

Figura 9 - Previsão de Dívidas para os Próximos 5 Meses – Brasil (em Bilhões de R$)

PERIODO	VTDD_REAL	VTDD_PRED
out/24	402.03	395910000
nov/24	410.14	405372857
dez/24	404.07	414835714
jan/25	419.00	424298571
fev/25	437.00	433761429
45717	438.00	443224286
abr/25	457.00	452687143
mai/25	465.00	462150000
jun/25	477.00	471612857
jul/25	482.00	481075714
ago/25	494.00	490538571
set/25	496.00	500001429

Tabela 2 - Métricas de acurácia dados registrados

PERIODO	VTDD_PREVISTA
out/25	510412121
nov/25	519946807
dez/25	529481492
jan/26	539016177
fev/26	548550862
46082	558085548
abr/26	567620233
mai/26	577154918
jun/26	586689604
jul/26	596224289
ago/26	605758974
set/26	615293660

Tabela 3 – Métricas de acurácia futuro

Bases Teóricas dos Métodos

A metodologia do projeto emprega duas abordagens complementares:

- Classificação (Aprendizagem Supervisionada): utilizada para prever a probabilidade de inadimplência. A base teórica consiste na estimação de P(y|X), onde y representa a variável alvo (inadimplência). Os algoritmos aplicados incluem Regressão Logística, Random Forest e Gradient Boosting;
- Agrupamento (Aprendizagem Não Supervisionada): empregado para segmentar consumidores em perfis com características similares, sem o uso de rótulos pré-definidos. A base teórica fundamenta-se na minimização da distância intra-grupo e na maximização da distância entre grupos. O algoritmo aplicado é o K-Means.

Essa abordagem combinada permite, primeiramente, a descoberta de padrões latentes nos dados e, subsequentemente, a avaliação do risco de crédito dentro dos segmentos identificados.

Conclusões da modelagem

Em conjunto com a EDA, a modelagem preditiva permite concluir que:

- O Brasil vive, no período analisado, um contexto de inadimplência elevada e crescente, tanto em número de pessoas quanto em valor total devido;
- O modelo de regressão linear, ainda que simples e com R^2 moderado, captura a tendência de alta do VTDD e projeta que o indicador deve se aproximar de R$ 532 bilhões em dezembro de 2025;
- As limitações de amostra decorrem principalmente de questões técnicas de padronização dos PDFs e disponibilidade dos relatórios, e não de falhas metodológicas na construção do pipeline;
- A partir da estrutura já implementada (extração + EDA + modelagem), torna-se viável expandir o projeto para janelas temporais maiores e modelos mais sofisticados (por exemplo, regressão múltipla com mais preditores, modelos ARIMA ou algoritmos baseados em aprendizado de máquina).

Dessa forma, o objetivo do Projeto Aplicado II é alcançado: a FinData Analytics demonstra ser capaz de transformar relatórios PDF não estruturados em um dataset analítico, extrair insights por meio da EDA e produzir projeções quantitativas sobre o risco de crédito a partir de modelos preditivos.

Cronograma de Atividades

Etapa	Atividades Principais	Prazo de Entrega
1. Kick-off	Definir grupo	11/agosto
	Definir empresa e área de atuação	04/setembro
	Definir dados, objetivos e cronograma	05/setembro
	Entrega da Etapa 1	05/setembro
2. Exploração e preparação de dados	Definir bibliotecas Python e repositório GitHub	10/setembro
	Análise exploratória da base de dados	17/setembro
	Tratamento e preparação dos dados	24/setembro
	Definir bases teóricas dos métodos analíticos e cálculo de acurácia	01/outubro
	Entrega da Etapa 2	03/outubro
3. Desenvolvimento analítico	Aplicar métodos analíticos definidos à base de dados	08/outubro
	Calcular acurácia e comparar métodos	12/outubro
	Descrever resultados preliminares e gerar protótipo	17/outubro
	Definir modelo de negócio e elaborar storytelling inicial	22/outubro
	Entrega da Etapa 3	24/outubro
4. Conclusão e entrega final	Redação do relatório técnico final	04/novembro

Finalização da apresentação storytelling em PPT	10/novembro
Organização final do repositório GitHub	14/novembro
Gravação e edição do vídeo de apresentação (YouTube)	18/novembro
Entrega da Etapa 4 (Final)	21/novembro

Conclusão

A conclusão deste projeto marca a transição bem-sucedida da fase de planejamento para a entrega da solução completa, conforme proposto pela FinData Analytics. O processo concentrou-se inicialmente na aquisição, estruturação e Análise Exploratória de Dados (EDA), culminando na criação do dataset serasa.csv, um ativo fundamental que validou o pipeline de extração de informações de relatórios em PDF, uma fonte de dados notadamente não estruturada. Os desafios técnicos previstos, como a variabilidade no layout dos documentos, foram superados, permitindo o desenvolvimento de um pipeline de tratamento e limpeza de dados eficaz.

A EDA forneceu insights críticos: as visualizações gráficas demonstraram um crescimento constante no Valor Total das Dívidas (VTDD) e no Número de Inadimplentes, e revelaram que a maior concentração de risco se encontra nas faixas etárias de 26 a 60 anos. A modelagem preditiva, por sua vez, garantiu a robustez das projeções, indicando que o VTDD deve alcançar aproximadamente R$ 548.55 bilhões até Fevereiro de 2026.

Com o desenvolvimento analítico concluído, a FinData Analytics cumpre seu objetivo de traduzir os dados estruturados em insights acionáveis. As soluções desenvolvidas, baseadas na aplicação de algoritmos de aprendizado (Classificação e Agrupamento), permitem:

- Segmentar perfis de consumidores para entender o risco;
- Prever o risco de crédito, aprimorando os modelos de avaliação das instituições financeiras.

A próxima etapa de desenvolvimento, reside na comunicação e entrega da solução. O objetivo é apresentar os resultados e o valor gerado por meio de um storytelling em formato de vídeo, simulando a entrega da solução ao cliente final. Assim, este projeto não só validou o processo de dados e a análise preditiva, mas também resultou na solução completa, pronta para ser comunicada e implementada no setor financeiro

Glossário

Este glossário define os termos técnicos e conceituais essenciais utilizados ao longo deste trabalho, com foco em ciência de dados, aprendizado de máquina e o domínio financeiro do projeto;

Acurácia (Accuracy):

Métrica de avaliação de modelos de classificação que mede a proporção de previsões corretas (positivas e negativas) em relação ao total de previsões realizadas. Embora seja uma métrica comum, deve ser usada com cautela em datasets desbalanceados;

Agrupamento (Clustering):

Técnica de aprendizado de máquina não supervisionado que visa agrupar um conjunto de dados em subconjuntos (clusters), de modo que os dados em um mesmo cluster sejam mais similares entre si do que com os de outros clusters. Utilizado neste projeto para segmentar perfis de consumidores;

Análise Exploratória de Dados (EDA):

Processo de investigação inicial de um conjunto de dados para descobrir padrões, anomalias, testar hipóteses e verificar premissas com o auxílio de estatísticas descritivas e visualizações gráficas;

Aprendizado de Máquina (Machine Learning):

Subcampo da inteligência artificial que se concentra no desenvolvimento de algoritmos que permitem aos computadores aprenderem padrões e tomar decisões a partir de dados, sem serem explicitamente programados para cada tarefa;

Classificação (Classification):

Tarefa de aprendizado de máquina supervisionado que consiste em atribuir uma categoria ou rótulo a uma observação com base em suas características. Neste projeto, é usada para classificar perfis de risco (ex: alto risco, baixo risco);

Dados Não Estruturados:

Dados que não possuem um modelo de dados pré-definido ou não são organizados de maneira padronizada. Exemplos incluem textos em linguagem natural, imagens, vídeos e arquivos de áudio. Os PDFs do Serasa são uma fonte primária de dados não estruturados neste projeto;

Dataset:

Termo em inglês para conjunto de dados. Refere-se a uma coleção organizada de dados, geralmente em formato tabular (linhas e colunas), que serve como base para análise, treinamento e teste de modelos;

Engenharia de Atributos (Feature Engineering):

Processo de criação de novas variáveis (features) a partir de variáveis existentes em um dataset. O objetivo é melhorar o desempenho dos modelos de aprendizado de máquina, fornecendo-lhes informações mais relevantes e preditivas;

GitHub:

Plataforma de hospedagem de código-fonte baseada no sistema de controle de versão Git. É amplamente utilizada para armazenar projetos, facilitar a colaboração entre desenvolvedores e garantir a transparência e reprodutibilidade de trabalhos acadêmicos e de software;

Granularidade:

Nível de detalhe ou profundidade da informação contida nos dados. Dados com alta granularidade são mais específicos (ex: inadimplência por cidade), enquanto dados com baixa granularidade são mais agregados (ex: inadimplência por país);

Inadimplência:

O não cumprimento de uma obrigação financeira, como o pagamento de uma dívida ou parcela de um empréstimo, dentro do prazo estipulado no contrato;

Insights:

Compreensões profundas e revelações importantes obtidas a partir da análise de dados, que podem ser utilizadas para embasar a tomada de decisões estratégicas

Matriz de Confusão (Confusion Matrix):

Tabela utilizada para visualizar o desempenho de um modelo de classificação. Ela detalha os acertos e erros, dividindo as previsões em Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e Falsos Negativos (FN);

Modelo Preditivo:

Um modelo estatístico ou de aprendizado de máquina treinado com dados históricos para fazer previsões sobre eventos futuros. No contexto deste projeto, o objetivo é prever o risco de crédito de um determinado perfil ou região;

Pipeline de Dados:

Uma série de etapas automatizadas para o processamento de dados, que vai desde a extração inicial (coleta) até a transformação (limpeza, estruturação) e o carregamento em um destino para análise ou modelagem;

Precision e Recall:

Duas métricas cruciais para modelos de classificação. A Precisão mede, de todas as previsões positivas feitas, quantas estavam corretas. A Revocação (ou Sensibilidade) mede, de todos os casos positivos reais, quantos o modelo conseguiu identificar;

Processamento de Linguagem Natural (PLN):

Área da inteligência artificial que capacita os computadores a compreenderem, interpretar e gerar a linguagem humana. Neste trabalho, é aplicado para extrair informações relevantes dos textos contidos nos relatórios em PDF;

Visão Computacional (Computer Vision):

Campo da inteligência artificial que treina computadores para interpretar e compreender o mundo visual. No projeto, refere-se às técnicas usadas para

extrair dados de elementos visuais, como tabelas e gráficos, dentro dos arquivos PDF.

Referências Bibliográficas

Figuras

Figura 1 - Serasa: SPC e Serasa - Indenização por dano moral. Jusbrasil, 2016. Disponível em: https://www.jusbrasil.com.br/artigos/spc-e-serasa-indenizacao-por-dano-moral/111821975. Acesso em: 4 set. 2025.

Figura 2 - Output EDA Dados SERASA, NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 3 - Evolução da Inadimplência no Brasil, NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 4 - Valor Médio por Pessoa (VMPP), NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 5 - Quantidade de dívidas no Brasil, NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 6 - Valor médio de cada dívida no Brasil, NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 7 - Valor Total das Dívidas (VTDD), NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 8 - Distribuição dos inadimplentes por faixa etária, NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Figura 9 - Previsão de Dívidas para os Próximos 5 Meses – Brasil (em Bilhões de R$), NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

Tabelas

Fontes

SERASA. Mapa da inadimplência e renegociação de dívidas no Brasil. [S. l.: s. n.], [s.d.]. Disponível em: https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/. Acesso em: 4 set. 2025.

Links de acesso

Repositório ONLINE

NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres. Projeto Aplicado 02 - Data Science Mackenzie 2025. [S. l.], 2025. Disponível em: https://github.com/GrupoMackenzie/ProjetoAplicado02-DataScience-Mackenzie-2025

Vídeo de Apresentação

https://youtu.be/YaVFC-WLAnw