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, permitindo a identificação de grupos e regiões com maior confiabilidade para a concessão de crédito.

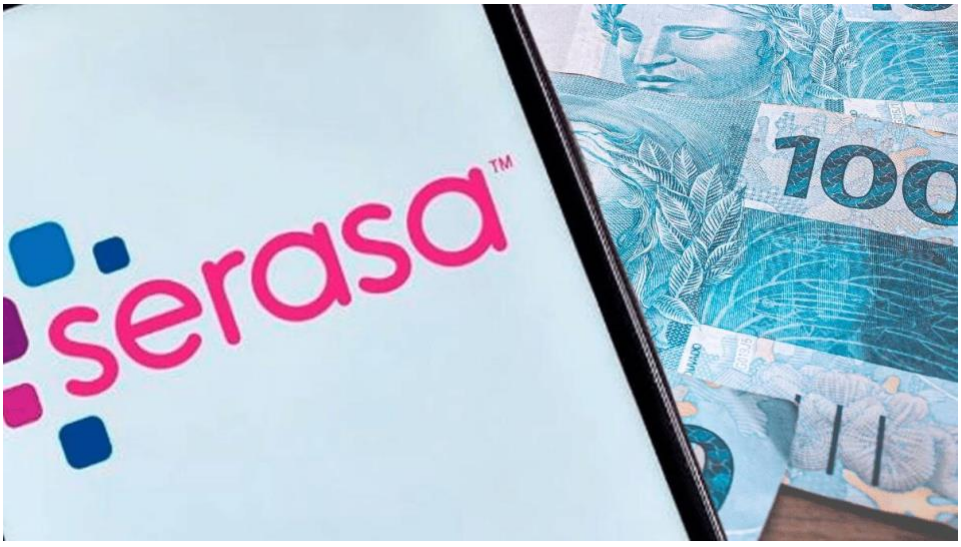

Figura 1- Serasa

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;
- 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.

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:

- VMAF: Valor Médio dos Acordos Fechados (R$);
- DESC_CONC: Descontos Concedidos (R$ bilhões);
- INADIMPLENTES: Número de inadimplentes (em milhões);
- VMPP: Valor Médio por Pessoa (R$);
- DÍVIDAS: Quantidade de dívidas (em milhões);
- VMCD: Valor médio de cada dívida (R$);
- VTDD: Valor total das dívidas (R$ bilhões).

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.

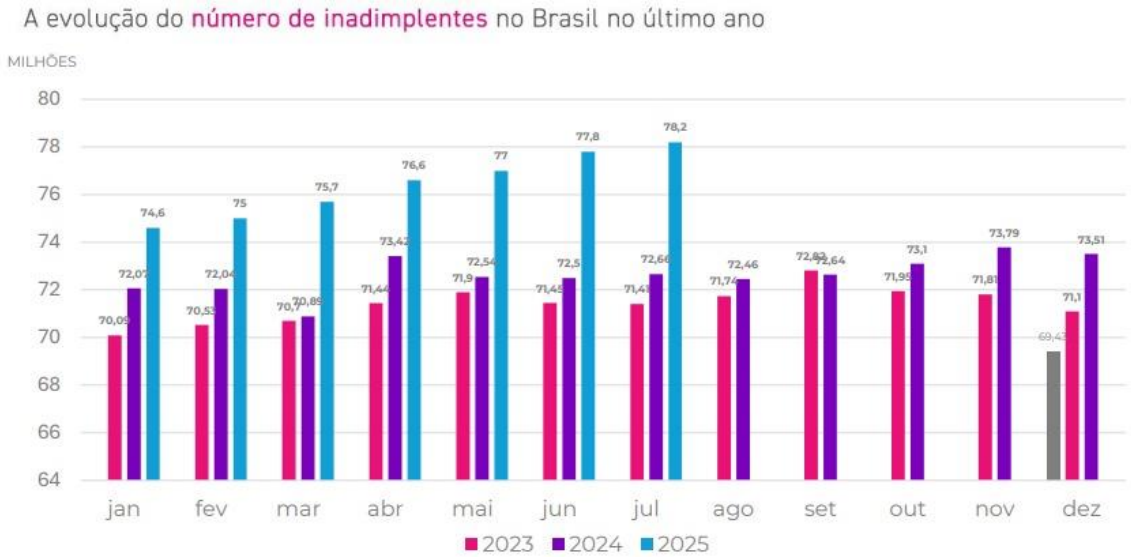

Figura 2 - Evolução de Inadimplência no Brasil

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.

	VMAF	DESC_CONC	INADIMPLENTES	VMPP	DIVIDAS	VMCD	VTDD
Abril 2025	790,00	10,5	76,6	5.968,71	294,1	1.555,33	457
Dezembro 2024	560,00	13,4	73,51	5.496,69	275,68	1.465,73	404,07
Fevereiro 2025	698,00	12,1	75,0	5,837,49	286,11	1.530,28	437
Janeiro 2025	676,00	10,2	74,60	5.617,00	281,25	1.489,90	419
Julho 2025	736,00	11,11	78,2	6.177,74	307,5	1.570,17	482
Junho 2025	772,00	9,90	77,8	6.128,26	304,5	1.567,05	477
Maio 2025	839,00	11,7	77	6.036,94	298,5	1.558,68	465
Marco 2025	714,00	17,7	75,7	5.793,66	287,6	1.526,41	438
Novembro 2024	590,74	15,6	73,79	5.558,3	274,54	1.493,92	410,14
Outubro 2024	734,83	10.51	73,10	5.504,33	276,08	1.457,48	402,03

Figura 3 - Output EDA Dados Serasa

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.

Definição das Métricas de Avaliação

O desempenho dos modelos foi avaliado por meio das seguintes métricas:

- Acurácia: proporção de classificações corretas;
- Matriz de Confusão: análise detalhada dos acertos e erros (Verdadeiros Positivos, Falsos Positivos, Verdadeiros Negativos e Falsos Negativos);
- Precision e Recall: métricas essenciais para avaliar o impacto de falsos positivos (risco de conceder crédito indevidamente) e falsos negativos (risco de negar crédito a bons pagadores).

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

Considerações

A conclusão da segunda etapa deste projeto marca um avanço significativo, transitando da fase de planejamento para a execução e entrega de resultados concretos. O trabalho realizado até o momento concentrou-se na aquisição, estruturação e análise exploratória dos dados, culminando na criação do dataset serasa.csv, um ativo fundamental para as próximas fases. A extração de informações de relatórios em PDF, uma fonte de dados notadamente não estruturada, foi concluída com sucesso, validando a metodologia e as ferramentas escolhidas.

A Análise Exploratória de Dados (EDA) permitiu uma primeira imersão nos padrões de inadimplência, revelando correlações e tendências iniciais que confirmam o potencial dos dados para a construção de modelos preditivos robustos. Os desafios técnicos previstos, como a variabilidade no layout dos documentos, foram superados por meio do desenvolvimento de um pipeline de tratamento e limpeza de dados eficaz.

Com um conjunto de dados consolidado, limpo e bem compreendido, o projeto está agora preparado para iniciar a fase de desenvolvimento analítico. As próximas etapas serão focadas na aplicação dos algoritmos de aprendizado de máquina supervisionado (Classificação) e não supervisionado (Agrupamento) definidos na metodologia. O objetivo será traduzir os dados estruturados em insights acionáveis, desenvolvendo modelos capazes de segmentar perfis de consumidores e prever o risco de crédito, conforme o escopo original proposto pela FinData Analytics.

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

Figura 3 - Output EDA Dados Serasa, NAGEM, Alberto; SAMPAIO, Ana Julia de Almeida; PEREIRA, Diogo Lima; MENDES, Gabriel Torres.

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.

Link 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