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.

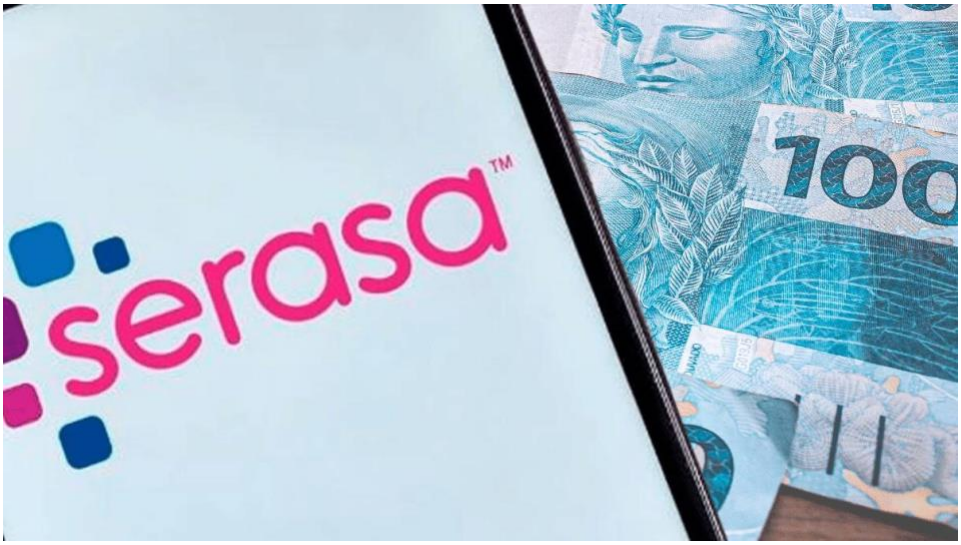

Figura 1-Serasa

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.

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.

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.

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

O presente documento representa a conclusão da fase de concepção e estruturação do projeto. Até o momento, o trabalho foi concentrado no planejamento estratégico e na fundamentação teórica da solução proposta pela FinData Analytics. As etapas concluídas incluem a definição clara do problema de negócio — a mitigação de riscos de inadimplência —, a identificação de uma fonte de dados pública e relevante (relatórios do Serasa) e o desenho da metodologia técnica que será empregada, envolvendo a extração de dados textuais e visuais.

A viabilidade do projeto se ampara na acessibilidade dos dados e na maturidade das ferramentas de código aberto, como a linguagem Python e suas bibliotecas, que serão utilizadas para construir o pipeline de processamento e análise. O escopo inicial, focado no relatório mais recente de Julho de 2025, foi deliberadamente definido para garantir uma entrega de valor controlada e permitir a validação da abordagem antes de uma possível expansão para dados históricos.

É importante considerar que a fase de execução apresentará desafios técnicos inerentes à natureza não estruturada dos dados. A variabilidade no layout dos relatórios em PDF e a complexidade na extração de informações de gráficos são pontos de atenção que exigirão testes e ajustes contínuos nos algoritmos a serem desenvolvidos.

Com o planejamento detalhado finalizado e os objetivos bem definidos, o projeto encontra-se agora pronto para avançar para a próxima fase: o desenvolvimento técnico e a implementação da solução proposta.

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;

Análise Exploratória de Dados (AED):

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:

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;

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;

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;

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;

Repositório:

No contexto do Git e GitHub, um repositório é um local de armazenamento centralizado para todos os arquivos e o histórico de versões de um projeto. Ele permite o rastreamento de todas as alterações feitas no código;

Risco de Crédito:

A probabilidade de perda financeira decorrente do não pagamento (inadimplência) de um empréstimo ou outra obrigação de crédito por parte de um tomador;

Software:

Conjunto de instruções, programas, dados e documentação que comanda o funcionamento de um computador ou dispositivo. No escopo deste projeto, refere-se principalmente à linguagem de programação Python e suas bibliotecas especializadas (ex: pdfplumber, Pandas, Scikit-learn);

Visão Computacional:

Campo da inteligência artificial que treina computadores para "ver" e interpretar o conteúdo de imagens e vídeos. Neste projeto, será utilizada para extrair dados de elementos visuais, como gráficos e tabelas, contidos nos 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.

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. Acesso em: 5 set. 2025.