RENATA LARIOS ALCANTARA DE FREITAS

MODELO DE SAÚDE FINANCEIRA PESSOAL PARA PREVISÃO DE RISCO DE CRÉDITO UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA ATRAVÉS DE DADOS PROVENIENTES DO OPEN FINANCE

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.

SÃO PAULO

2024

RENATA LARIOS ALCANTARA DE FREITAS

MODELO DE SAÚDE FINANCEIRA PESSOAL PARA PREVISÃO DE RISCO DE CRÉDITO UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA ATRAVÉS DE DADOS PROVENIENTES DO OPEN FINANCE

Monografia apresentada ao Programa de Educação Continuada da Escola Politécnica da Universidade de São Paulo, para obtenção do título de Especialista, pelo Programa de Pós-Graduação em Engenharia de Dados e Big Data.

Área de concentração: Tecnologia da Informação – Engenharia/ Tecnologia/ Gestão

Orientador: MSc. Ana Claudia Rossi

SÃO PAULO

2024

FICHA CATALOGRÁFICA

Freitas, Renata Larios Alcantara de
MODELO DE SAÚDE FINANCEIRA PESSOAL PARA PREVISÃO DE RISCO DE CRÉDITO UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA ATRAVÉS DE DADOS PROVENIENTES DO OPEN FINANCE / R. L. A. Freitas -- São Paulo, 2014. 48 p.

Monografia (Especialização em Engenharia de Dados & Big Data) - Escola Politécnica da Universidade de São Paulo. PECE – Programa de Educação Continuada em Engenharia.

1.Aprendizado de máquina 2.Análise de desempenho 3.Risco de crédito 4.Open Finance 5.Previsão de inadimplência I.Universidade de São Paulo. Escola Politécnica. PECE – Programa de Educação Continuada em Engenharia II.t.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por sua bondade infinita e misericórdia, que têm sido meu refúgio e alicerce em todas as circunstâncias.

À minha família, cujo amor incondicional, paciência e compreensão foram essenciais em cada passo desta jornada. O apoio e o incentivo de vocês formaram os pilares que me sustentaram, permitindo crescer, tanto pessoal quanto profissionalmente.

À professora orientadora Ana Claudia Rossi, por acreditar em mim mesmo nos momentos em que duvidei de minha própria capacidade. Sua sabedoria, paciência e orientações precisas foram determinantes para superar os desafios e alcançar este marco tão significativo em minha vida acadêmica.

Estendo ainda meu agradecimento aos amigos, colegas de sala, professores do PECE e a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

CURSO ENGENHARIA DE BIG DATA

Coord.: Prof. Solange N. Alves de Souza
Vice-Coord.: Prof. Pedro Luiz Pizzigatti Corrêa

Perspectivas profissionais alcançadas com o curso:

A especialização foi fundamental para aprofundar meus conhecimentos em Big Data e engenharia de dados, explorando aspectos de arquitetura, tecnologias e metodologias que contribuíram para desenvolver uma visão mais crítica e precisa. Além de aprimorar minhas habilidades técnicas, essa formação também ampliou minhas possibilidades de atuação e me deu a chance de causar um impacto positivo na área.

RESUMO

A análise do risco de crédito é fundamental para as instituições financeiras, pois pode acarretar perdas diretas e imediatas. Este estudo examina a aplicabilidade de modelos de aprendizado de máquina na previsão de risco de crédito utilizando dados provenientes do Open Finance. Foram empregados os modelos Support Vector Machine, Decision Trees, Bagging, AdaBoost e Random Forest, e comparados quanto à sua precisão preditiva através de métricas de padrão de desempenho em classificação. Os resultados mostram que o AdaBoost se destacou como o mais eficaz, demonstrando um bom equilíbrio entre as métricas e maior capacidade de identificar casos positivos. A conclusão sugere que existem oportunidades explorando modelos como redes neurais profundas (Deep Learning) ou técnicas de aprendizado por reforços, que pode trazer benefícios significativos em cenários com grandes volumes de dados e relações não lineares complexas.

Palavras-chave: Aprendizado de máquina, Risco de crédito, Open Finance, Previsão de inadimplência, Análise de desempenho

ABSTRACT

The analysis of credit risk is essential for financial institutions, as it can lead to direct and immediate losses. This study examines the applicability of machine learning models in predicting credit risk using data from Open Finance. The models Support Vector Machine, Decision Trees, Bagging, AdaBoost, and Random Forest were employed and compared regarding their predictive accuracy through standard classification performance metrics. The results show that AdaBoost stood out as the most effective, demonstrating a good balance among metrics and a greater ability to identify positive cases. The conclusion suggests opportunities for exploring models such as deep neural networks (Deep Learning) or reinforcement learning techniques, which can bring significant benefits in scenarios with large volumes of data and complex nonlinear relationships.

Keywords: Machine learning, Credit risk, Open Finance, Default prediction, Performance analysis

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC - Area Under the Curve

BAGGING - Bootstrap Aggregating

DT - Decision Trees

NBDRA - NIST Big Data Reference Architecture

RF - Random Forest

ROC - Receiver Operating Characteristic Curve

SVM - Support Vector Machines

SUMÁRIO

1 INTRODUÇÃO

Os modelos de aprendizado de máquina têm se destacado como ferramentas centrais na previsão de risco de crédito, impulsionando avanços significativos no setor financeiro. Técnicas como Support Vector Machines (SVM), Árvores de Decisão e modelos de ensemble (como Bagging, AdaBoost e Random Forest) têm demonstrado grande eficácia em lidar com volumes massivos de dados e identificar padrões complexos, proporcionando previsões mais precisas e personalizadas. Esses modelos possibilitam estimativas de risco mais confiáveis, atendendo às demandas de um mercado financeiro cada vez mais dinâmico e baseado em dados (BREIMAN, 2001; OMARINI, 2020).

Nesse cenário, a previsão de risco de crédito continua sendo um pilar essencial para a estabilidade e eficiência do setor financeiro. A integração de técnicas avançadas de aprendizado de máquina com novas fontes de dados tem potencializado o desenvolvimento de soluções mais robustas e adaptadas às necessidades dos consumidores e instituições.

Ao mesmo tempo, o advento do Open Finance tem transformado o ecossistema financeiro ao permitir o compartilhamento automatizado de dados cadastrais e transacionais entre instituições. Esse modelo operacional, além de modernizar o Sistema Financeiro Nacional (SFN), reduz custos e aumenta a eficiência operacional, criando um ambiente favorável para a aplicação de tecnologias emergentes, como inteligência artificial, big data e blockchain (MELNYCHENKO et al., 2020).

A combinação de dados integrados pelo Open Finance com modelos de aprendizado de máquina redefine o panorama da análise de crédito. A utilização dessas informações mais ricas e contextuais permite aprimorar a precisão dos modelos preditivos, resultando em ferramentas mais eficazes e personalizadas na concessão de crédito. Essa convergência entre aprendizado de máquina, previsão de crédito e o cenário do Open Finance representa uma oportunidade única para inovação e transformação do setor financeiro.

1.1 Problema

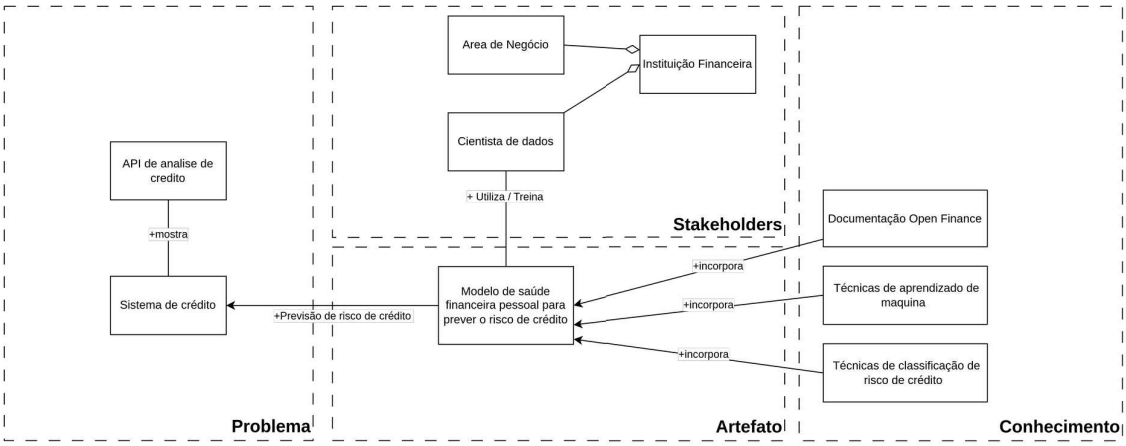
O problema tratado nesta monografia é como adaptar e aplicar modelos de aprendizado de máquina de forma eficaz para prever o risco de crédito, utilizando os dados disponibilizados pelo Open Finance. Isso envolve desafios relacionados à qualidade, integridade e complexidade dos dados, além de questões técnicas para garantir que os modelos sejam precisos, escaláveis e interpretáveis no contexto financeiro.

A avaliação do risco de crédito tem um papel relevante para as instituições financeiras, uma vez que os empréstimos podem resultar em perdas reais e imediatas. (ANICETO et al., 2020)

1.2 Objetivo

O objetivo principal desta monografia é uma analise e adequação de modelos de classificação de saúde financeira pessoal para prever o risco de crédito, através de dados provenientes do Open Finance, explorando técnicas de aprendizado de máquina.

Figura 1 – Modelo Conceitual



Fonte: Adaptado de Design Science Methodology for Information Systems and Software Engineering (WIERINGA, 2014)

A figura 1 descreve o ciclo de design e validação de um sistema para análise de risco de crédito (problema), onde o objetivo principal é a construção de um artefato, o sistema de análise de crédito que incorpora técnicas de aprendizado de máquina e classificação de risco, para processar dados financeiros provenientes de fontes do Open Finance. Para isso é realizado revisão da literatura de técnicas de classificação de risco de crédito, técnicas de aprendizado de Máquina e Documentação Open Finance, para atender os stakeholders que são os Cientistas de Dados que atuam diretamente na criação e treinamento do modelo e a área de Negócio que representa os interesses estratégicos e ambos pertencem a uma Instituição Financeiras.

O ciclo de design e validação do sistema, abordando a identificação de requisitos das partes interessadas, a construção e aplicação de técnicas computacionais e a incorporação de regulamentações externas para produzir resultados robustos e aplicáveis no contexto financeiro. (WIERINGA, 2014)

Os objetivos específicos para alcançar a meta proposta são:

Objetivos Específicos

- Construção e Análise do Conjunto de Dados: Construir um conjunto de dados baseado em API do Open Finance disponibilizada pelo BACEN (Banco Central do Brasil), para classificação risco de crédito;

- Preparação dos Dados: Preparar os dados, incluindo a limpeza, normalização e transformação dos dados para garantir que sejam adequados para o treinamento dos modelos de aprendizagem de máquina;

- Treinamento e Validação dos Modelos: Support Vector Machine (SVM), Decision Trees (DT), Bootstrap Aggregating (Bagging), AdaBoost e Random Forest (RF) para determinar quais são mais eficazes na previsão do risco de crédito.

- Análise Comparativa: Comparar os modelos de aprendizagem de máquina com base em métricas como como Acurácia Geral (Overall Accuracy – ACC), Erro Tipo I (T1E – Sensitivity, Erro Tipo II (T2E – Specificity, Curva ROC (Receiver Operating Characteristic Curve) e AUC (Area Under the Curve) da curva ROC, destacando os pontos positivos e negativos de cada modelo no contexto de previsão de risco de crédito.

- Seleção do Melhor Modelo: Identificar o modelo mais eficaz para a classificação de risco de crédito com base nos dados do Open Finance.

1.3 Justificativa

A análise de crédito ao consumidor é um pilar essencial para garantir a saúde financeira tanto dos indivíduos quanto das instituições financeiras. (XOLANI et al 2020)

Existem vários estudos que investigam o uso de modelos de aprendizado de máquina na análise de risco de crédito, concentrando-se em sua eficiência em várias bases de dados (ASSEF et al. 2019, PŁAWIAK et al. 2020, ZHONG et al.2014). Embora exista uma quantidade significativa de estudos nesta área, os resultados ainda não são conclusivos, o que destaca a complexidade da tarefa e a importância de levar em conta as particularidades de cada situação específica (DASTILE et al. 2020).

O processo inicial de construção do conjunto de dados é essencial para garantir a qualidade e a integridade dos dados de entrada, que são fundamentais para o desempenho dos modelos. A literatura sobre classificação em finanças, como o artigo sobre análise de crédito bancário, destaca a importância de dados históricos e detalhados para a construção de modelos robustos, evidenciando que uma base de dados bem estruturada contribui para previsões mais precisas (ASSEF et al. 2019).

A etapa de preparação dos dados é crucial para garantir a eficácia dos modelos. Segundo Breiman (1996), a estabilidade dos modelos de Bagging e Random Forest depende fortemente da qualidade dos dados de entrada. Essas técnicas aplicadas a dados de Open Finance ajudam a garantir que as variáveis estejam devidamente estruturadas para maximizar a acurácia dos modelos.

A escolha de algoritmos como Support Vector Machines (SVM), Decision Trees (DT), Bagging, AdaBoost e Random Forest para a previsão de risco de crédito está fundamentada nas suas capacidades de lidar com grandes volumes de dados e em sua eficácia para classificação complexa. Segundo Breiman (2001) aponta que esse método é altamente eficaz em cenários onde a diversidade entre as árvores melhora a generalização do modelo, enquanto o artigo sobre Bagging explora como a combinação de modelos pode reduzir a variância e melhorar a robustez.

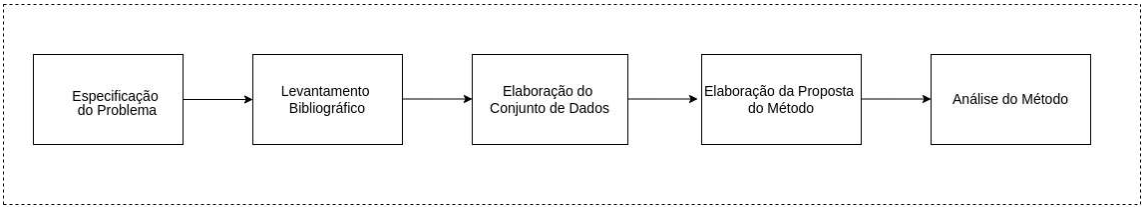
A análise comparativa dos modelos será realizada com base em métricas de desempenho, como Acurácia, Erro Tipo I e II, Curva ROC e AUC, que são padrões na avaliação de modelos de classificação. As técnicas comparativas destacadas por Ben-David (1995) sobre manutenção de monotonicidade e métricas de avaliação são úteis para avaliar as nuances de cada modelo em cenários específicos. Esse passo permite identificar pontos fortes e limitações de cada abordagem, fornecendo uma visão detalhada sobre quais modelos se adaptam melhor ao contexto de risco de crédito.

Dessa forma, este trabalho justifica-se pela necessidade crescente de modelos preditivos eficazes que possam alavancar os dados de Open Finance para prever o risco de crédito de maneira precisa e confiável. A utilização de técnicas de aprendizado de máquina possibilita a criação de modelos que podem se adaptar rapidamente a novas informações, reduzindo riscos e promovendo a estabilidade financeira. A base teórica e metodológica, fundamentada nos artigos selecionados, oferece um embasamento sólido para o desenvolvimento, avaliação e seleção dos melhores modelos para a previsão de risco de crédito.

1.4 Metodologia

A pesquisa foi realizada seguindo os ciclos de Design Science, incorporando elementos como problem-solving, construção de artefatos e avaliação empírica. (WIERINGA, 2014).

Figura 2 – Fluxo Metodológico



Fonte: Adaptado de Design Science Methodology for Information Systems and Software Engineering (WIERINGA, 2014)

A figura 2 apresenta o fluxo da metodologia utilizada baseada no ciclo de Design Science, seguindo as principais etapas:

1. Especificação do problema: Nesta etapa inicial, define-se claramente o escopo do estudo, abrangendo:

- Identificação e detalhamento do problema relacionado a como adaptar e aplicar modelos de aprendizado de máquina para prever o risco de crédito baseado em dados do Open Finance.
- Estabelecimento do objetivo principal da pesquisa, analisar e adequar os modelos de aprendizado de máquina prever o risco de crédito, através de dados provenientes do Open Finance.
- Justificativa do estudo, destacando sua relevância no contexto acadêmico e prático.

2. Levantamento bibliográfico: Realiza-se uma revisão sistemática de literatura, abordando:

- Princípios e frameworks do Open Finance.
- Técnicas e algoritmos de aprendizagem de máquina.
- Estudos e métricas associadas à previsão de risco de crédito.

3. Elaboração do conjunto de dados do Open Finace: Nesta etapa, executa-se:

- Criação de dados simulados do ambiente Open Finance.
- Limpeza, organização e pré-processamento dos dados para adequação aos modelos de aprendizado de máquina.

4. Elaboração da proposta do método: Consiste na seleção de modelos de aprendizagem de máquina, incluindo:

- Definição e implementação de algoritmos para previsão de risco de crédito.
- Configuração do processo de treinamento dos modelos, utilizando validação cruzada e otimização de hiperparâmetros.

5. Análise do Método: Após a aplicação dos modelos:

- Avaliação do desempenho com base em métricas como precisão, recall, F1-score e AUC-ROC.
- Comparação entre diferentes modelos para identificar vantagens, limitações.
- Discussão dos resultados, incluindo impacto, desafios e sugestões de melhorias.

1.5 Estrutura do Trabalho

O presente trabalho está organizado em 4 capítulos.

No capítulo 1 é apresentado o escopo e motivação da monografia, com a especificação do contexto, do problema, do objetivo, da justificativa e da metodologia.

No capítulo 2 são apresentados os fundamentos teóricos das técnicas de aprendizado de máquina utilizadas no contexto de classificações para análise de risco de crédito. Support Vector Machine, Decision Trees, Bagging (Bootstrap Aggregating), AdaBoost e Random Forest.

No capítulo 3 é apresentado a elaboração do método proposto neste trabalho. É apresentado as características dos dados do Open Finance, realizado o treinamento, aplicação de modelos de máquina, é apresentada a analise dos resultados, são comparados os modelos de aprendizagem de máquina com base em métricas como como Acurácia Geral (Overall Accuracy – ACC), Erro Tipo I (T1E – Sensitivity, Erro Tipo II (T2E – Specificity, Curva ROC (Receiver Operating Characteristic Curve) e AUC (Area Under the Curve) da curva ROC, apontando os pontos positivos e negativos de cada modelo no contexto de previsão de risco de crédito.

No capítulo 4 são apresentadas as conclusões da monografia, as principais considerações da pesquisa e algumas limitações do estudo.

2 CONCEITOS DE PREVISÃO DE RISCO DE CRÉDITO, MODELOS DE APRENDIZADO DE MÁQUINA E AVALIAÇÃO DE DESEMPENHO DOS MODELOS

Neste capítulo são apresentados os conceitos relacionados à previsão de crédito, com foco em técnicas de aprendizado de máquina amplamente utilizadas para classificação de risco. Serão discutidas as características de métodos específicos, como Support Vector Machines (SVM), Decision Trees (DT), Bagging, AdaBoost e Random Forest (RF), além de métricas essenciais para avaliar a eficácia desses modelos na previsão de crédito.

2.1 Big Data e Previsão de Crédito

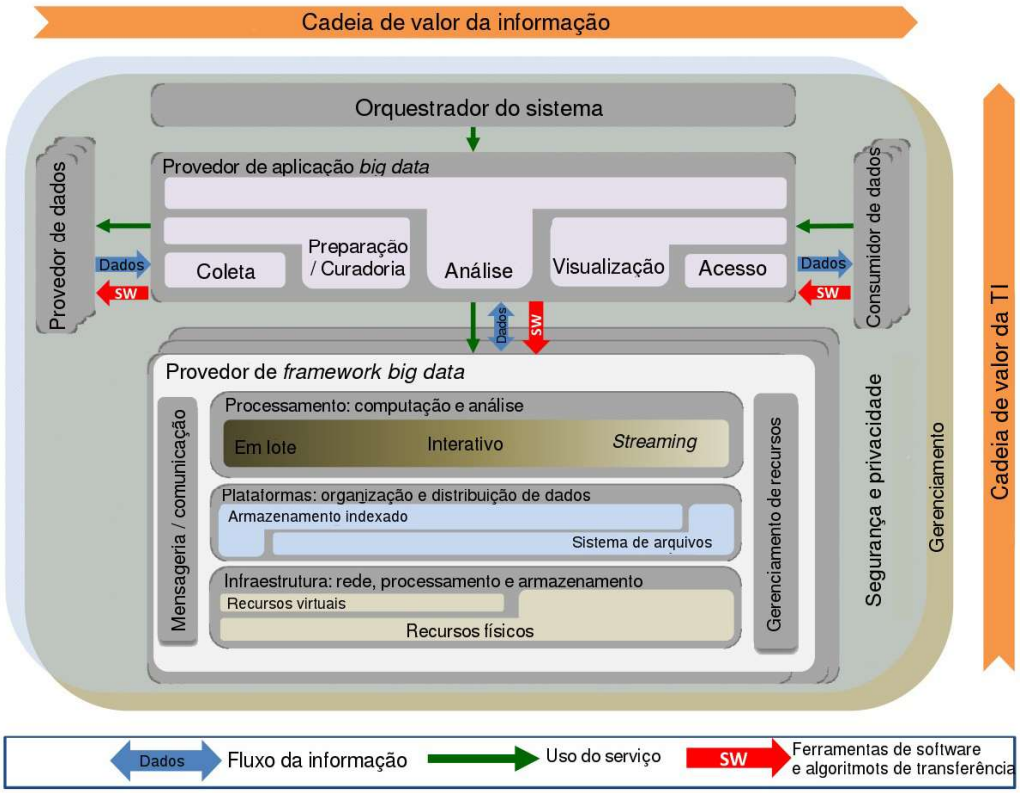
A capacidade de coletar, processar e analisar grandes volumes de dados tem permitido que bancos e outras instituições financeiras melhorem a predição de crédito. Big Data oferece ferramentas para lidar com a complexidade e o volume de informações relacionadas ao comportamento de crédito dos consumidores. (ANICETO et al., 2020)

O uso de Big Data viabiliza não apenas a aplicação de diferentes técnicas de machine learning, mas também o treinamento eficiente em conjuntos de dados amplos e diversificados. A figura 4 apresenta a arquitetura de referência NIST para Bigdata, neste estudo o contexto do Provedor de Dados são os dados financeiros utilizados originados de APIs disponibilizadas pelo Open Finance, regulamentadas pelo Banco Central do Brasil (BACEN).

Esses dados estruturados representam informações financeiras pessoais que servem como insumos brutos para o sistema desenvolvido, atendendo à etapa inicial de coleta de informações dentro do modelo. No contexto do Provedor de Framework Big Data, o trabalho se concentra nas camadas de processamento, plataforma e infraestrutura. As técnicas de aprendizado de máquina, configuram-se como

ferramentas para o processamento e análise dos dados. O foco do trabalho está associado ao Provedor de Aplicação Big Data, onde ocorre a construção do sistema de análise de risco de crédito, incluindo as etapas de geração dos dados do Open Finance, preparação dos dados e sua análise com o uso de algoritmos de aprendizado de máquina. A aplicação dessas técnicas gera resultados que são posteriormente disponibilizados aos stakeholders cientistas de dados.

Figura 4 - NIST Big Data Reference Architecture (NBDRA)



Fonte: NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. (CHANG, 2015)

2.2 Conceitos de Previsão de Crédito

A previsão de crédito visa classificar a confiabilidade financeira de consumidores a partir de dados comportamentais e transacionais, como aqueles provenientes de operações com cartões de crédito. Esses dados, ao revelarem padrões de gastos e tendências financeiras, ajudam as instituições a avaliar o risco associado a potenciais clientes. A análise de crédito busca identificar perfis de clientes com maior probabilidade de inadimplência, oferecendo insights valiosos para a gestão de risco financeiro. Com a evolução de técnicas de aprendizado de máquina, a previsão de crédito tornou-se mais precisa, ao se beneficiar da capacidade de processamento de grandes volumes de dados interdependentes e da identificação de padrões complexos de comportamento financeiro (ASSEF et al. 2019; BEN-DAVID, 1995).

2.3 Conceitos de Técnicas/Modelos de Aprendizado de Máquina

Nesta seção, são detalhadas as principais técnicas de aprendizado de máquina utilizadas na previsão de crédito, cada uma com características específicas para lidar com dados complexos e volumes amplos.

2.3.1 Support Vector Machines (SVM)

O SVM é um método de aprendizado supervisionado que busca encontrar um hiperplano que separe os dados em grupos homogêneos, trabalha com vetores de entrada (x1,y1),(x2,y2),...,(xm,ym), onde xi são os vetores de características em um espaço dimensional d, e $y_i \in \{-1,1\}$ representa as classes das observações. A técnica utiliza funções Kernel para mapear dados para um espaço de características superior e construir um hiperplano que maximize a margem entre as classes. (ANICETO et al., 2020)

Os Kernels mais comuns incluem:

- **Linear**: Usado para dados linearmente separáveis.

- **Radial Basis Function (RBF)**: Adequado para padrões mais complexos, permitindo a deformação do hiperplano para capturar não linearidades.

Na previsão de crédito, o SVM é eficaz para distinguir consumidores com diferentes perfis de risco ao utilizar uma função de kernel que transforma os dados em um espaço dimensional elevado, tornando-se útil para problemas com múltiplas variáveis interdependentes. Além disso, o SVM possui uma forte capacidade de generalização, embora possa demandar um alto custo computacional em grandes conjuntos de dados(BEN-DAVID, 1995).

2.3.2 Decision Trees (DT)

As Árvores de Decisão são modelos baseados em uma estrutura de árvore, onde os dados são organizados em uma estrutura hierárquica para classificação ou regressão. (ANICETO et al., 2020)

Seguem a estrutura de um "árvore invertida", onde:

- **Nó raiz:** Representa a primeira decisão baseada em um atributo com maior ganho de informação.
- **Nós internos:** Representam decisões subsequentes baseadas em atributos escolhidos interativamente.
- **Folhas:** Representam as classes ou valores preditivos finais, indicando o caminho de decisões que levaram a elas.

Esse método é simples de interpretar e permite visualizar como as decisões são tomadas com base em variáveis específicas, o que é vantajoso na análise de risco de crédito. No entanto, as Árvores de Decisão podem se tornar instáveis, sendo suscetíveis ao sobreajuste, especialmente com dados com variabilidade indesejada(ASSEF et al. 2019)

2.3.3 Bagging (Bootstrap Aggregating)

O Bagging é uma técnica de aprendizado de máquina baseada em ensemble que cria múltiplas versões de um modelo ao aplicar amostras bootstrap (amostras com reposição) no conjunto de dados original, combinando os resultados por votação ou média. (ANICETO et al., 2020)

Funciona criando múltiplos subconjuntos de dados por meio de amostragem com reposição (bootstrap) de um conjunto de treinamento original. Cada subconjunto é usado para treinar um modelo base (geralmente Decision Trees), e os resultados são combinados por meio de votação (classificação) ou média (regressão).

Na previsão de crédito, o Bagging ajuda a reduzir a variância do modelo, criando classificadores robustos a partir de árvores de decisão. Essa técnica é especialmente eficaz em métodos instáveis, melhorando a acurácia sem aumentar significativamente o viés (BREIMAN, 1996).

2.3.4 AdaBoost

O AdaBoost, um método de ensemble adaptativo, ajusta iterativamente o peso das observações, focando em amostras onde os erros são mais frequentes. Cada novo classificador é ajustado para corrigir os erros das iterações anteriores, o que resulta em um modelo mais preciso. (ANICETO et al., 2020)

Segue as seguintes as etapas:

- **Inicialização dos pesos:** No início, cada observação recebe o mesmo peso.
- **Treinamento iterativo:** Em cada iteração, é treinado um modelo base (frequentemente Decision Trees simples), e os erros de classificação são calculados.
- **Ajuste de pesos:** Observações mal classificadas recebem maior peso na próxima iteração, direcionando o próximo modelo a corrigir esses erros.

- **Combinação dos modelos:** Os modelos treinados são combinados por meio de uma votação ponderada, onde os pesos refletem a precisão de cada modelo.

Na previsão de crédito, o AdaBoost pode melhorar a capacidade de identificar consumidores de alto risco, mas é sensível a dados ruidosos, o que pode limitar sua aplicação em certos cenários (BREIMAN,2021).

2.3.5 Random Forest (RF)

O RF é um método de aprendizado de máquina baseado em ensemble, que utiliza múltiplas árvores de decisão para melhorar a robustez e a precisão preditiva. (ANICETO et al., 2020)

Segue as seguintes as etapas:

- **Construção de árvores:** O RF cria diversas árvores de decisão, onde cada uma é treinada em um subconjunto dos dados, gerado por amostragem bootstrap. Em cada nó, uma seleção aleatória de características é considerada para a divisão, introduzindo diversidade entre as árvores.
- **Combinação de resultados:** Para classificação, as predições são combinadas por votação majoritária. Para regressão, a média das predições das árvores é usada.

Na previsão de crédito, o Random Forest é robusto a dados ruidosos e variáveis altamente correlacionadas, o que o torna ideal para cenários com múltiplas variáveis interrelacionadas (BREIMAN,2021).

2.4 Métricas de Desempenho

Para avaliar a eficácia dos modelos de aprendizado de máquina na previsão de crédito, são utilizadas várias métricas de desempenho. Essas métricas permitem entender a precisão e a capacidade de generalização de cada modelo, destacando suas vantagens e limitações no contexto de análise de risco.

A Acurácia Geral (Overall Accuracy – ACC) é a proporção de previsões corretas em relação ao total de previsões realizadas. No contexto de previsão de crédito, a acurácia indica a porcentagem de consumidores classificados corretamente como de alto ou baixo risco. No entanto, essa métrica pode ser insuficiente para avaliar modelos em cenários de dados desbalanceados, onde a proporção entre classes é desigual (BEN-DAVID, 1995).

O Erro Tipo I ou Sensibilidade (T1E – Sensibilidade), mede a capacidade do modelo de identificar corretamente os consumidores de alto risco (inadimplentes). Alta sensibilidade é desejável na previsão de crédito, pois minimiza o risco de conceder crédito a clientes que apresentam maior probabilidade de inadimplência(ASSEF et al. 2019).

O Erro Tipo II ou Especificidade (T2E – Especificidade), representa a capacidade do modelo de identificar corretamente os consumidores de baixo risco (adimplentes). Uma alta especificidade reduz o número de classificações incorretas de consumidores confiáveis como de alto risco, evitando decisões injustas e conservadoras demais(BEN-DAVID, 1995).

A Curva ROC (Receiver Operating Characteristic Curve) é uma representação gráfica que ilustra o desempenho de um modelo em termos de Sensibilidade (Taxa de Verdadeiros Positivos) e 1-Especificidade (Taxa de Falsos Positivos). A área sob a curva ROC, ou AUC (Area Under the Curve), quantifica a capacidade de discriminação do modelo, indicando sua eficiência em classificar corretamente os consumidores como de alto ou baixo risco. Quanto maior a AUC, melhor o modelo é

na distinção entre classes, sendo um indicador robusto na avaliação de modelos preditivos na análise de risco de crédito(ASSEF et al. 2019).

Este capítulo abordou os conceitos de previsão de crédito e aprendizado de máquina, detalhando técnicas específicas e métricas de desempenho que orientam a escolha e a avaliação dos modelos para uma análise precisa e confiável de risco de crédito. A fundamentação teórica aqui apresentada fornecerá a base para o desenvolvimento e a análise dos modelos nos capítulos seguintes.

3 VALIDAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Neste capítulo é elaborada a ponte entre os conceitos fundamentais discutidos anteriormente e a sua aplicação prática. Durante a condução do estudo, é criado um conjunto de dados preenchido com os dados do Open Finance através de API disponibilizada pelo BACEN, feito um pré-processamento no conjunto de dados, para assegurar a qualidade e a uniformidade dos dados. Aplicado diferentes modelos de aprendizagem de máquina - como Support Vector Machines (SVM), Decision Trees (DT), Bagging (Bootstrap Aggregating), AdaBoost e Random Forest (RF).

3.1 Contexto de Aplicação

O cenário estudado baseia-se no ecossistema de Open Finance, um modelo regulatório que promove a integração de serviços financeiros através do compartilhamento seguro de dados entre instituições. No Brasil, o Banco Central regulamenta e supervisiona essa iniciativa, oferecendo APIs padronizadas para acesso às informações financeiras dos clientes. (OMARINI, 2020).

Para conduzir a análise de crédito, foi criado um conjunto de dados baseado nas APIs do Open Finance disponibilizadas pelo BACEN. Estas APIs incluem informações sobre contas de crédito, faturas, limites de crédito e transações, conforme detalhado a seguir:

Os dados foram gerados, visando a criação de um ambiente de teste representativo para as APIs de contas de crédito do Open Finance, conforme especificações estabelecidas pelo Banco Central do Brasil (BACEN, 2024).

Tabela 1 - Dados, Critérios de análise e as condições utilizadas

API/Condição	Descrição	Campos Utilizados	Critérios de Análise
API /accounts (accounts.csv)	Informações sobre contas de crédito, incluindo tipo de cartão, limite disponível e rede associada. Dados ajustados pela classe social do titular.	Tipo de cartão, limite disponível, rede associada.	-
API /accounts/{creditCardAccountId}/bills (bills.csv)	Simula faturas de contas, detalhando valores totais, mínimos, e status de pagamento. Probabilidades de inadimplência variam conforme classe social.	isPaid, payments, billTotalAmount.	Detectar múltiplas faturas não pagas consecutivas pelo campo **isPaid** e **payments.**
API /accounts/{creditCardAccountId}/limits (limits.csv)	Detalha o limite de crédito, valores usados e disponíveis. Percentuais ajustados por classe social.	usedAmount, limitAmount.	Calcular percentual de uso: **usedAmount / limitAmount**. Classificar como risco elevado se acima de 80%.
API /accounts/{creditCardAccountId}/ transactions (transactions.csv)	Histórico de transações, com valores, categorias de gastos e parcelamentos. Inclui penalidades para faturas não pagas.	transactionDate, valor das transações.	Contar transações realizadas nos últimos 30 dias. Identificar risco para clientes com mais de 20 transações neste período.

API /personal/qualifications (qualifications.csv)	Informações financeiras dos titulares, incluindo renda mensal, ocupação e classe social. Dados refletem padrões reais.	incomeAmount.	Relacionar total das faturas (**billTotalAmount**) com renda mensal (**incomeAmount**). Classificar como risco elevado se comprometimento excessivo da renda.
API /personal/financial-relations (financial_relations.csv)	Relações financeiras dos titulares, como patrimônio e renda anual. Valores ajustados para refletir padrões econômicos reais.	Patrimônio declarado, renda anual informada.	-
Condição: Histórico de faturas não pagas	Identificação de clientes com múltiplas faturas consecutivas em aberto.	isPaid, payments.	Combinar dados das APIs de faturas e transações para identificar faturas não pagas.
Condição: Uso excessivo do limite de crédito	Análise do percentual de limite utilizado por cliente.	usedAmount, limitAmount.	Avaliar uso do limite pelo cálculo **usedAmount / limitAmount**. Classificar como risco elevado se uso superior a 80%.
Condição: Capacidade de pagamento comprometida	Avaliação da relação entre dívida total e renda mensal.	billTotalAmount, incomeAmount.	Relacionar total das faturas ao valor da renda mensal declarada.
Condição: Histórico de transações frequentes	Identificação de clientes com alta frequência de transações nos últimos 30 dias.	transactionDate.	Contar transações no período e associar a outros fatores de risco (atrasos, uso excessivo do limite).

Fonte: Próprio autor

A tabela 1 sintetiza as principais especificações utilizadas para a geração dos dados do Open Finance. A seguir, é detalhado como cada uma:

- **API /accounts (accounts.csv):**

  Inclui informações sobre contas de crédito, como tipo de cartão, limite disponível, e rede associada. A distribuição dos tipos de cartões e limites foi baseada na classe social do titular.

- **API /accounts/{creditCardAccountId}/bills (bills.csv):**

  Simula as faturas associadas às contas, detalhando valores totais, mínimos, e status de pagamento. As probabilidades de inadimplência variaram por classe social, refletindo padrões reais.

- **API /accounts/{creditCardAccountId}/limits (limits.csv):**

  Fornece detalhes sobre o limite de crédito, valores usados e disponíveis, com percentuais ajustados por classe social.

- **API /accounts/{creditCardAccountId}/transactions (transactions. csv):**

  Contém o histórico de transações, incluindo valores, categorias de gastos, e parcelamentos. Foram adicionados dados como penalidades para faturas não pagas.

- **API /personal/qualifications (qualifications.csv):**

  Fornece informações financeiras detalhadas sobre os titulares das contas, incluindo renda mensal, ocupação e classe social. Os dados refletem a distribuição de classes sociais e padrões de renda baseados na realidade brasileira.

- **API /personal/financial-relations (financial_relations.csv):**

  Detalha as relações financeiras dos titulares, incluindo patrimônio declarado e renda anual informada. Os valores foram ajustados para refletir a proporção entre patrimônio e renda, considerando padrões econômicos reais.

Para a aplicabilidade dos modelos, vamos utilizar algumas condições para verificar se existe alto risco de inadimplência:

- Histórico de mais de uma fatura não paga.
- Uso excessivo do limite de crédito.
- Capacidade de pagamento comprometida pela renda mensal.
- Histórico de transações com alta frequência.

A seguir, é detalhado como cada uma dessas condições foi analisada e vinculada aos dados fornecidos pelas APIs.

- **Histórico de mais de uma fatura não paga:**

  Para identificar se uma fatura foi quitada, a análise combinou dados da API de faturas (/accounts/{creditCardAccountId}/bills) e transações (/accounts/{creditCardAccountId}/transactions). A API de faturas contém informações como o campo isPaid, que indica diretamente se uma fatura foi paga, além do campo payments, que registra os valores quitados. Caso este campo estivesse vazio ou indicasse um pagamento parcial, a fatura foi classificada como não paga. Esse cruzamento de dados permitiu detectar

clientes com múltiplas faturas consecutivas em aberto, alinhado ao cenário de risco descrito no estudo (BACEN, 2024).

- **Uso excessivo do limite de crédito:** Este fluxo reflete uma abordagem sistemática e iterativa para garantir resultados robustos e confiáveis. A Etapa 1 aborda os métodos e ferramentas utilizados para criar e organizar os dados. A etapa 2 detalha o processo de treinamento e teste do modelo, incluindo a configuração de parâmetros e as métricas aplicadas. Por fim, a Etapa 3 apresenta a avaliação comparativa entre diferentes abordagens e os Resultados Obtidos, oferecendo uma análise do desempenho e das implicações das soluções propostas.

O percentual de uso do limite foi calculado utilizando os campos usedAmount (limite já utilizado) e limitAmount (limite total) fornecidos pela API /accounts/{creditCardAccountId}/limits. A fórmula usedAmount / limitAmount foi empregada para avaliar o percentual de limite utilizado por cada cliente. Caso esse valor ultrapasse 80%, o cliente é classificado como risco elevado, pois tal comportamento pode indicar uma alta dependência de crédito e dificuldade em gerir os recursos disponíveis.

- **Capacidade de pagamento comprometida pela renda mensal**

A capacidade de pagamento foi avaliada considerando a relação entre o valor total das faturas e a renda mensal declarada do cliente. Os dados para essa análise foram extraídos das APIs de faturas (billTotalAmount no /accounts/{creditCardAccountId}/bills) e de qualificações financeiras

(incomeAmount no /accounts/{creditCardAccountId}/qualifications). Foi calculada a razão billTotalAmount / incomeAmount para identificar clientes que comprometem uma grande parte de sua renda mensal no pagamento de dívidas, refletindo maior risco de inadimplência.

- **Histórico de transações com alta frequência**

O número total de transações realizadas nos últimos 30 dias foi calculado utilizando a API /accounts/{creditCardAccountId}/transactions, com base no campo transactionDate. Para clientes que realizaram mais de 20 transações nesse período, foi identificado um padrão de comportamento financeiro que pode refletir tanto uma gestão eficiente quanto um risco elevado, especialmente quando analisado em conjunto com atrasos no pagamento e uso excessivo do limite (BACEN, 2024).

3.2 Processo de Avaliação de Modelos

O processo de avaliação dos modelos foi estruturado para garantir uma análise comparativa com as práticas estabelecidas de aprendizado de máquina em cenários de desequilíbrio de classes.

Diversos algoritmos de aprendizagem de máquina foram aplicados para prever o risco de crédito, incluindo:

- **Support Vector Machines (SVM)**

  Classifica clientes com base em características para identificar aqueles com maior ou menor risco de crédito.

- **Decision Trees (DT)**

  Estrutura decisões sobre aprovação ou recusa de crédito com base em regras derivadas de dados históricos.

- **Bootstrap Aggregating (Bagging)**

  Combina múltiplos modelos para prever o risco de crédito com maior estabilidade e menor variância.

- **AdaBoost**

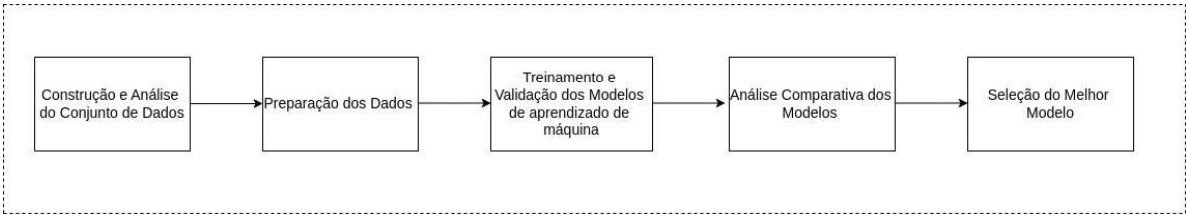
  Ajusta iterativamente erros de previsão para melhorar a identificação de clientes com risco de crédito.

- **Random Forest (RF)**

  Utiliza várias árvores de decisão para melhorar a precisão na previsão do risco de crédito ao considerar múltiplas variáveis.

Esses modelos foram selecionados com base em sua relevância, características distintas e eficácia comprovada em análises de classificação, especialmente em previsões de risco de crédito. (ANICETO et al., 2020)

Figura 4 – Etapas do Processo de Avaliação



Fonte: Próprio Autor

A figura 4 apresenta um fluxo das etapas do processo de avaliação, adaptado do artigo , dividido em cinco etapas principais descritas abaixo:

- **Construção e Análise do Conjunto de Dados:**

  Construir um conjunto de dados baseado em API do Open Finance disponibilizada pelo BACEN (Banco Central do Brasil), para classificação risco de crédito;

- **Preparação dos Dados:**

  Descrever as etapas de preparação dos dados, incluindo a limpeza, normalização e transformação dos dados para garantir que sejam adequados para o treinamento dos modelos de aprendizagem de máquina;

- **Treinamento e Validação dos Modelos de aprendizado de máquina:**

  Support Vector Machine (SVM), Decision Trees (DT), Bagging (Bootstrap Aggregating), AdaBoost e Random Forest (RF) para determinar quais são mais eficazes na previsão do risco de crédito.

- **Análise Comparativa dos Modelos:**

  Comparar os modelos de aprendizagem de máquina com base em métricas como como Acurácia Geral (Overall Accuracy – ACC), Erro Tipo I (T1E – Sensitivity, Erro Tipo II (T2E – Specificity, Curva ROC (Receiver Operating Characteristic Curve) e AUC (Area Under the Curve) da curva ROC, destacando os pontos positivos e negativos de cada modelo no contexto de previsão de risco de crédito.

- **Seleção do Melhor Modelo:**

  Identificar o modelo mais eficaz para a classificação de risco de crédito com base nos dados do Open Finance.

3.3 Aplicação do Processo

Descreve as etapas da aplicação do processo proposto, abrangendo desde a geração e pré-processamento dos dados até a avaliação comparativa dos resultados. A etapa 1 e 2 aborda os métodos e ferramentas utilizados para criar e organizar os dados. A etapa 3 detalha o processo de treinamento e teste do modelo, incluindo a configuração de parâmetros e as métricas aplicadas. A etapa 4 apresenta a avaliação comparativa entre diferentes abordagens e os Resultados Obtidos e a etapa 5 descreve o modelo mais eficaz.

Etapa 1: Construção e Análise do Conjunto de Dados e Etapa 2: Preparação dos Dados:

Para a implementação dos scripts, foram empregadas bibliotecas do Python, tais como Faker para criação de dados aleatórios e csv para exportação dos resultados gerados. A biblioteca Faker permite simular dados fictícios, mas realistas, como nomes de empresas, datas, valores financeiros e outros detalhes transacionais, essenciais para a diversidade e autenticidade dos dados. Outras funções nativas de Python, como random.choice e random.randint, foram aplicadas para sorteios e definições de valores com controle e variação, assegurando a representatividade dos cenários.

Todos os métodos desenvolvidos respeitam os formatos de dados e as exigências das APIs do BACEN, como identificação única de contas (creditCardAccountId), numeração

de CNPJ e MCCs (Merchant Category Codes) válidos, datas no padrão ISO e valores monetários com precisão decimal.

Todos os dados gerados foram exportados para arquivos CSV utilizando a biblioteca csv, garantindo conformidade com os campos obrigatórios e opcionais definidos pelo BACEN.

Etapa 3: Treinamento e Validação dos Modelos de aprendizado de máquina

Os dados foram divididos em conjuntos de treino (70%) e teste (30%). Para atender às particularidades de cada modelo a abordagem de treinamento e avaliação foi ajustada, segue detalhes:

No modelo SVM para o processo de pré-processamento, os dados foram padronizados utilizando o StandardScaler, garantindo que todas as variáveis fossem normalizadas para uma escala padrão. Para a configuração do modelo, foi empregada uma máquina de vetores de suporte com kernel linear, juntamente com o balanceamento das classes para mitigar possíveis problemas de desbalanceamento nos dados. Além disso, foi realizado o cálculo de probabilidades para a classificação, permitindo uma análise probabilística dos resultados. A validação do modelo foi conduzida por meio de validação cruzada com 5 folds, um método que assegura maior robustez e confiabilidade na avaliação do desempenho do modelo, reduzindo o risco de overfitting.

Para a configuração do modelo de árvore de decisão, a profundidade máxima foi limitada a 5, com o objetivo de evitar problemas de overfitting e manter a simplicidade do modelo. Além disso, foi realizado o balanceamento de peso entre as classes para lidar com possíveis desbalanceamentos no conjunto de dados. A validação do modelo foi efetuada utilizando validação cruzada com 5 folds, permitindo avaliar a capacidade de generalização do modelo de forma robusta e confiável.

Para a configuração do modelo Bagging, foi utilizado o BaggingClassifier com 50

árvores de decisão, garantindo a diversidade dos modelos de base. O método empregou subconjuntos aleatórios gerados por meio de bootstrap a partir do conjunto de treino, permitindo maior robustez e variabilidade no treinamento. As árvores de decisão tiveram a profundidade máxima limitada para evitar overfitting, e os pesos das classes foram balanceados para tratar possíveis desbalanceamentos nos dados. A avaliação da robustez e da capacidade de generalização do modelo foi realizada por meio de validação cruzada com 5 folds.

Na configuração do modelo AdaBoost, foi utilizado o AdaBoostClassifier com 50 estimadores baseados em árvores de decisão de profundidade limitada, garantindo simplicidade e controle do overfitting. O algoritmo SAMME foi empregado para combinar previsões ponderadas, ajustando dinamicamente os pesos das amostras durante o treinamento, o que permite ao modelo focar progressivamente nos exemplos mais difíceis. A avaliação da capacidade de generalização e robustez foi realizada por meio de validação cruzada com 5 folds, assegurando uma análise confiável do desempenho do modelo.

Para a configuração do modelo Random Forest, foram utilizadas 100 árvores de decisão com amostragem por bootstrap, garantindo diversidade entre os modelos de base. A profundidade máxima das árvores foi limitada a 5, com o objetivo de evitar overfitting e manter a simplicidade do modelo. Além disso, foi aplicado o balanceamento de pesos entre as classes para lidar com possíveis desbalanceamentos nos dados. A avaliação da robustez e capacidade de generalização do modelo foi realizada por meio de validação cruzada com 5 folds, proporcionando uma análise confiável do desempenho.

Etapa 4: Análise Comparativa dos Modelos

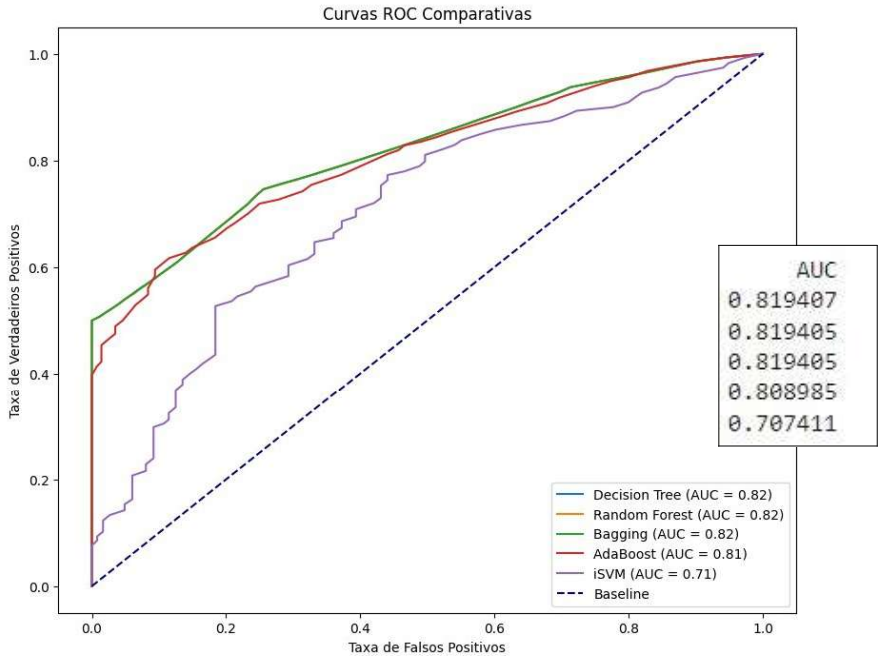
A avaliação do desempenho dos modelos de aprendizado de máquina foi realizada com base em métricas de acurácia, sensibilidade, especificidade, precisão e área sob a curva ROC (AUC). A seguir, apresenta-se uma análise detalhada de cada modelo, considerando as métricas observadas:

Figura 5 - Acurácia Geral, Sensibilidade, Especificidade e Precisão

	Modelo	Acurácia	Sensibilidade	Especificidade	Precisão
0	Decision Tree	0.550020	0.513358	0.980983	0.996859
1	Random Forest	0.550020	0.513358	0.980983	0.996859
2	Bagging	0.550020	0.513358	0.980983	0.996859
3	AdaBoost	0.690667	0.682463	0.787110	0.974149
4	iSVM	0.476590	0.447725	0.815901	0.966203

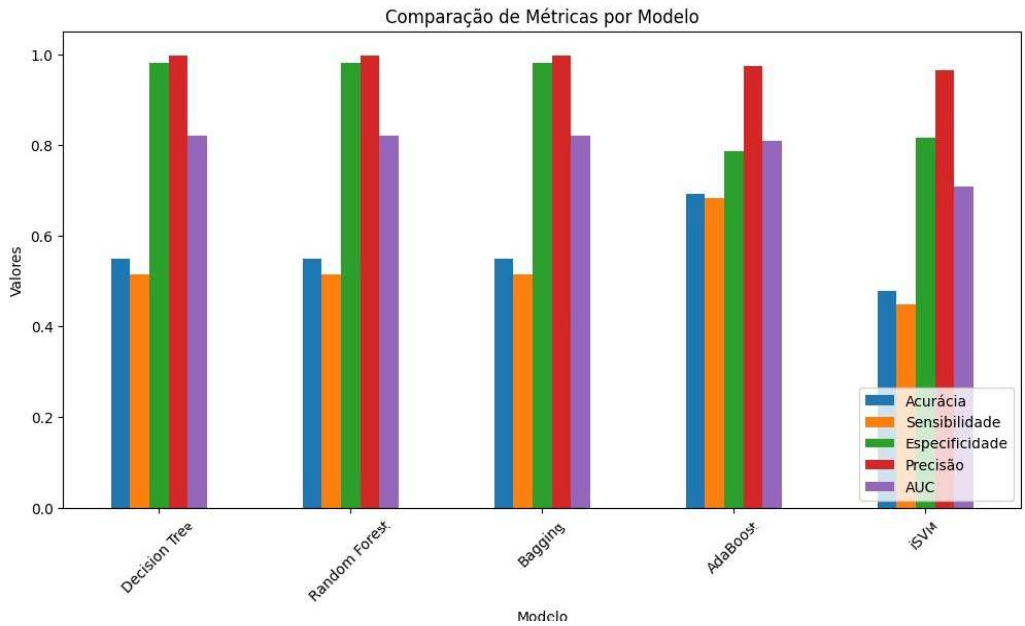
Fonte: Próprio Autor

Figura 6 - Curva ROC e AUC da curva ROC



Fonte: Próprio Autor

Figura 7 - Comparação Geral



Fonte: Próprio Autor

Com base nos resultados apresentados nas figuras 5, 6 e 7, é possível avaliar o desempenho de cada modelo de classificação com base nas métricas de Acurácia, Sensibilidade, Especificidade, Precisão e AUC (Área sob a Curva ROC).

O modelo **Decision Tree** apresentou um desempenho limitado em termos de acurácia (55,0%) e sensibilidade (51,3%), indicando dificuldade em identificar corretamente as amostras positivas. Apesar disso, a especificidade foi elevada (98,0%), refletindo boa capacidade de identificar as amostras negativas. A precisão também foi alta (99,7%), sugerindo que, entre as predições positivas, a maioria estava correta. O AUC foi de 0,82, mostrando que o modelo possui uma capacidade razoável de separar as classes, mas não é o mais robusto no conjunto avaliado.

O **Random Forest** teve resultados idênticos ao modelo Decision Tree, com acurácia, sensibilidade, especificidade e precisão nos mesmos valores (55,0%, 51,3%, 98,0% e 99,7%, respectivamente). Isso indica que, embora seja uma técnica

baseada em ensemble, as configurações aplicadas limitaram sua capacidade de superar o desempenho da árvore de decisão isolada. O AUC de 0,82 sugere uma separação das classes equivalente à do Decision Tree, sem ganho significativo em termos de performance.

O modelo **Bagging** também apresentou desempenho semelhante aos modelos anteriores, com valores idênticos para todas as métricas (acurácia de 55,0%, sensibilidade de 51,3%, especificidade de 98,0% e precisão de 99,7%). O AUC de 0,82 reflete novamente um desempenho equivalente aos outros modelos baseados em árvores de decisão, indicando que o ensemble com Bagging não trouxe uma melhoria expressiva no cenário avaliado.

O modelo **AdaBoost** foi o que apresentou o melhor desempenho geral, destacando-se pela acurácia de 69,0% e sensibilidade de 68,2%, indicando maior eficiência na identificação de amostras positivas. A especificidade, embora mais baixa em comparação aos modelos anteriores (78,7%), foi suficiente para um bom equilíbrio entre as classes. A precisão foi alta (97,4%), destacando a capacidade do modelo de realizar predições positivas corretas. O AUC de 0,81, embora ligeiramente menor que o dos modelos baseados em árvores, reflete um desempenho competitivo em cenários com maior complexidade.

Por fim, o modelo **iSVM** apresentou o pior desempenho geral, com acurácia de 47,6% e sensibilidade de apenas 44,7%, evidenciando dificuldades em identificar corretamente as amostras positivas. A especificidade foi moderada (81,5%), enquanto a precisão, embora elevada (96,6%), não foi suficiente para compensar as limitações nas demais métricas. O AUC de 0,71 confirma o desempenho inferior em comparação aos outros modelos, indicando uma menor capacidade de separação entre as classes.

Em resumo, o AdaBoost foi o modelo mais eficaz, enquanto o iSVM apresentou o pior desempenho. Decision Tree, Random Forest e Bagging tiveram resultados semelhantes, destacando a necessidade de ajustes nas configurações para explorar melhor o potencial dos ensembles.

Etapa 5: Seleção do Melhor Modelo

Entre os modelos avaliados, o AdaBoost destacou-se como a abordagem mais eficiente, apresentando a maior acurácia (69,0%) e sensibilidade (68,2%), demonstrando excelente capacidade de identificar amostras positivas e de lidar com cenários mais complexos. Além disso, sua precisão elevada (97,4%) assegurou que a maioria das predições positivas estivesse correta, tornando-o particularmente robusto em aplicações onde a identificação de casos positivos é crucial. Embora outros modelos, como Decision Tree, Random Forest e Bagging, tenham mostrado boa especificidade e precisão, seus desempenhos gerais foram inferiores ao AdaBoost devido à menor sensibilidade e acurácia. Por outro lado, o SVM apresentou o pior desempenho, com métricas significativamente mais baixas, indicando que não foi adequado para o conjunto de dados avaliado. Assim, o AdaBoost foi selecionado como o melhor modelo, equilibrando eficácia, robustez e capacidade de generalização.

3.4 Resultados Obtidos

Os resultados obtidos evidenciaram diferenças importantes no desempenho dos modelos de classificação analisados. O modelo Decision Tree apresentou limitações na identificação de amostras positivas, apesar de demonstrar boa capacidade em classificar corretamente as amostras negativas e uma alta precisão nas predições positivas. O Random Forest, embora seja um modelo de ensemble mais avançado, teve desempenho idêntico ao da árvore de decisão, indicando que as configurações aplicadas não permitiram explorar todo o seu potencial.

De maneira similar, o Bagging também apresentou resultados equivalentes aos do Decision Tree e Random Forest, sugerindo que as limitações configuracionais restringiram os benefícios esperados do ensemble. Por outro lado, o AdaBoost destacou-se como o modelo mais eficaz, mostrando maior equilíbrio entre as métricas avaliadas, com uma excelente capacidade de identificar amostras positivas e realizar predições robustas em cenários complexos.

Por fim, o modelo SVM apresentou o pior desempenho geral, evidenciando dificuldades na classificação de amostras positivas e uma capacidade inferior de separação entre as classes em comparação aos demais modelos. Em síntese, o AdaBoost foi identificado como o modelo mais eficiente, enquanto os outros modelos, especialmente o iSVM, apresentaram limitações significativas no contexto avaliado.

4 CONCLUSÃO

Este estudo apresentou uma análise detalhada de diferentes modelos de aprendizado de máquina aplicados à previsão de crédito utilizando dados gerados no contexto do Open Finance. O principal objetivo foi identificar a abordagem mais eficiente para essa tarefa, considerando métricas como acurácia, sensibilidade, especificidade, precisão e AUC. Entre os modelos avaliados, o AdaBoost destacou-se como o mais eficaz, demonstrando um bom equilíbrio entre as métricas e maior capacidade de identificar casos positivos. Contudo, modelos como Decision Tree, Random Forest e Bagging apresentaram desempenhos similares, indicando que as configurações restritivas limitaram seu potencial. O modelo SVM, por outro lado, mostrou-se menos adequado para o cenário analisado.

Uma limitação importante deste trabalho está relacionada à natureza da base de dados utilizada, que foi gerada pelo autor e não representa dados reais de instituições financeiras. Essa característica pode ter influenciado o desempenho dos modelos e, consequentemente, os resultados obtidos. A quantidade de dados e variáveis também foi restrita, o que pode ter limitado a capacidade dos modelos de generalizar para cenários mais complexos. Essas limitações sugerem que os resultados apresentados não são totalmente representativos de um contexto real de mercado, como aquele encontrado em instituições financeiras.

4.1 Contribuições do Trabalho

As principais contribuições deste estudo são:

Análise comparativa de cinco modelos (Decision Tree, Random Forest, Bagging, AdaBoost e SVM) com base em métricas relevantes, como acurácia, sensibilidade, especificidade, precisão e AUC. Essa análise forneceu pontos positivos e limitações de cada abordagem.

A pesquisa contribuiu para a compreensão de como os dados disponibilizados pelo Open Finance podem ser utilizados para prever o risco de crédito, demonstrando a importância de modelos de ensemble em cenários complexos.

4.2 Trabalhos Futuros

Para trabalhos futuros, é recomendada a aplicação dos modelos a bases de dados reais de instituições financeiras, o que permitiria uma análise mais robusta e representativa. Além disso, a expansão do conjunto de dados, tanto em quantidade quanto em diversidade de variáveis, pode melhorar a capacidade de generalização dos modelos.

Outras abordagens, como uso de redes neurais profundas (Deep Learning) ou técnicas de aprendizado por reforço, pode trazer benefícios significativos, especialmente em cenários com grandes volumes de dados e relações não lineares complexas.

REFERÊNCIAS BIBLIOGRÁFICA

ANICETO, M.C., BARBOZA, F.; KIMURA, H. **Machine learning predictivity applied to consumer creditworthiness.** *Futur Bus J*6, 37, 2020. https://doi.org/10.1186/s43093-020-00041-w

ASSEF, F.; STEINER, M.T.; NETO, P.J.S.; DE BARROS FRANCO, D.G. **Classification algorithms in financial application: credit risk analysis on legal entities.** IEEE Lat Am Trans, 17(10), 1733-1740, 2019. https://doi.org/10.1109/TLA.2019.8985866

BANCO CENTRAL DO BRASIL, BACEN. **Resolução conjunta nº 1, de 4 de maio de 2020.** Diário Oficial da União, Brasília, 5 mai. 2020, seção 1, p. 34-38. Disponível em: https://normativos.bcb.gov.br/Lists/Normativos/Attachments/51028/Res_Conj_0001_v4_P.pdf. Acesso em: 28 de Agosto de 2024.

BEN-DAVID, A. M**onotonicity maintenance in information-theoretic machine learning algorithms.** Machine Learning, v. 19, n. 1, p. 29-43, 1995.

BREIMAN, L. **Bagging predictors.** Machine Learning, v. 24, n. 2, p. 123-140, 1996.

BREIMAN, L. **Random forests.** Machine Learning, v. 45, n. 1, p. 5-32, 2001.

Chang, Wo L. **NIST Big Data Interoperability Framework: Volume 6, Reference Architecture.** No. Special Publication (NIST SP)-1500-6. 2015.

CSV File Format Documentation. **CSV Module in Python**. 2024. Disponível em: https://docs.python.org/3/library/csv.html. Acesso em: 2 de Outubro de 2024.

DASTILE, X.; CELIK, T.; POTSANE, M. , **Statistical and machine learning models in credit scoring: A systematic literature survey**. Applied Soft Computing, Volume 91, 2020, 106263, ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2020.106263.

FAKER Documentation. **Fake Daker: Fata Generator**. 2024. Disponível em: https://faker.readthedocs.io/en/stable/. Acesso em: 2 de Outubro de 2024.

FENG S., XINGCHAO Z., GANG ., FAWAZ E. A. **A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique.** Applied Soft Computing, Volume 98, 2021, 106852, ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2020.106852.

LUO, C. **A comprehensive decision support approach for credit scoring.** *Industrial Management & Data Systems*, Vol. 120 No. 2, pp. 280-290, 2020. https://doi.org/10.1108/IMDS-03-2019-0182

MELNYCHENKO, O.; RUBIN, D.; OMARINI, A.; KARAKAS, S. **Open Finance: The Next Frontier of Financial Services.** Journal of Digital Banking, 5(2), 183-194, 2020. https://doi.org/10.2139/ssrn.3520628

MELO, E. A. **Implantação do open finance no Brasil: desafios e efeitos potenciais.** 106 f. 2023. Dissertação (Mestrado Profissional em Administração e Controladoria) – Faculdade de Economia, Administração, Atuária e Contabilidade, Universidade Federal do Ceará, Fortaleza, 2023.

OMARINI, A. **Banks and Fintechs: How to Develop a Digital Open Banking Approach for the Bank's Future.** International Business Research, 13(3), 1-13, 2020. https://doi.org/10.5539/ibr.v13n3p1

OPEN BANKING BRASIL. **Open Banking Brasil API Documentation: Credit Cards.** 2024. Disponível em: https://openbanking-brasil.github.io/openapi/swagger-apis/credit-cards/?urls.primaryName=2.3.0#/. Acesso em: 2 de Outubro de 2024.

PŁAWIAK, P.; ABDAR, M.; PŁAWIAK, J.; MAKARENKOV, V.; ACHARYA, U.R. **DGHNL: a new deep genetic hierarchical network of learners for prediction of credit scoring.** Inf Sci, 516, 401-418, 2020. https://doi.org/10.1016/j.ins.2019.11.070

PYTHON. **Python Random Module: Randomness in Python.** 2024. Disponível em: https://docs.python.org/3/library/random.html. Acesso em: 2 de Outubro de 2024.

VIEIRA, J.R.C.; BARBOZA, F.; SOBREIRO, V.A.; KIMURA, H. **Machine learning models for credit analysis improvements: predicting low-income families' default.** Applied Soft Computing, Volume 83, (105):640. https://doi.org/10.1016/j.asoc.2019.105640

WIERINGA, Roel J. **Design Science Methodology for Information Systems and Software Engineering.** Springer, Heidelberg, New York, Dordrecht, London, 2014. ISBN 978-3-662-43838-1. Disponível também em formato eBook: ISBN 978-3-662-43839-8. DOI: 10.1007/978-3-662-43839-8.

ZHONG, H.; MIAO, C.; SHEN, Z.; FENG, Y. C**omparing the learning effectiveness of BP, ELM, I-ELM, and SVM for corporate credit ratings.** Neurocomputing, 128, 285-295, 2014. https://doi.org/10.1016/j.neucom.2013.09.035

APÊNDICE A – CÓDIGOS, BASE DE DADOS UTILIZADAS NO DESENVOLVIMENTO DO ESTUDO

Os códigos e a base de dados utilizados no desenvolvimento deste trabalho estão disponibilizados em um repositório público no GitHub.

O repositório contém:

- Códigos-fonte: Scripts utilizados para a análise, processamento de dados e geração de resultados descritos no corpo da monografia.
- Base de dados: Conjunto de dados brutos e processados utilizados no estudo, acompanhados de descrições detalhadas.

Para acessar os arquivos, utilize o seguinte link: https://github.com/RenataLarios/Monografia---Modelo-de-Risco-de-Credito-com-Machine-Learning-e-dados-do-Open-Finance.git