Um logotipo para computação universitária

Descrição gerada automaticamente

**Data de Submissão: 19/05/2024**

|  |  |  |
| --- | --- | --- |
| *Marluce Taciana Bora* |  |  |
| *Número do aluno: 2024141* |  |  |
| *Título do módulo: Pensamento Estratégico* |  |  |
| *Projeto de Implementação de Modelos de Machine Learning para Risco de Crédito* |  |  |
| *Data de vencimento da avaliação: 19/05/2024* |  |  |
| *Data de Submissão: 19/05/2024* |  |  |

**Declaração**

Ao enviar esta avaliação, confirmo que li a política do CCT sobre má conduta acadêmica e entendo as implicações de enviar trabalhos que não são meus ou não fazem referência adequada a material retirado de terceiros ou de outra fonte.

Declaro que é meu próprio trabalho e que todo o material de terceiros foi devidamente referenciado.

Confirmo ainda que este trabalho não foi submetido anteriormente para avaliação por mim ou por outra pessoa no CCT College Dublin ou em qualquer outra instituição de ensino superior.

Abstrair

Este relatório descreve um projeto abrangente focado na implementação de modelos de aprendizado de máquina para avaliação de risco de crédito. O principal objetivo é desenvolver modelos preditivos capazes de classificar com precisão os clientes com base em seu risco de crédito, apoiando a tomada de decisão no processo de empréstimo. Usando o conjunto de dados "Risco de Crédito Alemão - Com Alvo" disponível na plataforma Kaggle, o projeto seguiu uma metodologia estruturada que incluiu as seguintes etapas:

* Análise Exploratória de Dados (EDA): Exploração e visualização inicial de dados para identificar padrões, tendências e possíveis problemas de qualidade de dados.
* Preparação de dados: Limpar, transformar e preparar os dados para garantir que estejam prontos para modelagem.
* Implementação de Algoritmos de Machine Learning: Treinamento e avaliação de vários modelos de machine learning, incluindo Regressão Logística, Análise Discriminante Linear, K-Nearest Neighbors, Árvore de Decisão, Naive Bayes, Random Forest, Support Vector Machine e XGBoost.
* Avaliação do modelo: Comparação de modelos com base em métricas de desempenho, com ênfase em recall, precisão e estabilidade.
* Melhor Seleção de Modelo: Análise detalhada dos melhores modelos, Regressão Logística (LGR) e Análise Discriminante Linear (LDA), justificando sua escolha com base no desempenho e interpretabilidade.
* Resultados e Conclusões: Interpretação dos resultados obtidos, destacando-se o desempenho superior da Máquina de Vetores de Suporte (SVM) e a robustez dos modelos LGR e LDA.

O relatório conclui com uma visão para análise e melhoria futuras, propondo a implementação de técnicas avançadas e a exploração de novos algoritmos de aprendizado de máquina nos próximos seis meses para aumentar ainda mais a precisão dos modelos preditivos. Um cronograma detalhado do que foi feito e das atividades planejadas é fornecido para garantir o sucesso contínuo do projeto.

Conteúdo

[1. Introdução 4](#_Toc181544203)

[2. Plano de Projeto 4](#_Toc181544204)

[3. Compreensão do negócio 4](#_Toc181544205)

[4. Compreensão de dados 5](#_Toc181544206)

[5. Análise Exploratória de Dados 5](#_Toc181544207)

[5.1 Amostras das análises realizadas. 6](#_Toc181544208)

[6. Preparação de dados 10](#_Toc181544209)

[6.1 Identificação de outliers 10](#_Toc181544210)

[6.2 Tratamento de outliers 11](#_Toc181544211)

[6.3 Codificação de variáveis categóricas 12](#_Toc181544212)

[6.4 Padronização 12](#_Toc181544213)

[6.5 Análise de correlação e mapa de calor 12](#_Toc181544214)

[7. Implementação de aprendizado de máquina 13](#_Toc181544215)

[8. Balanceamneto de Classes com SMOTE 15](#_Toc181544216)

[9. Construcao do Pipeline e Pré-Processamento 15](#_Toc181544217)

[9.1 Normalização dos Dados (StandardScaler) 15](#_Toc181544218)

[9.2 Seleção de Variáveis e PCA 16](#_Toc181544219)

[9.3 Ajuste de Hiperparâmetros com GridSearchCV 16](#_Toc181544220)

[9.4 Avaliação e Comparação dos Modelos 16](#_Toc181544221)

[9.5 Identificação do Melhor Modelo 17](#_Toc181544222)

[10. Resultados 17](#_Toc181544223)

[11. Desafios 18](#_Toc181544224)

[12. Conclusão 18](#_Toc181544225)

[13. Referência 19](#_Toc181544226)

# Introdução

A gestão eficaz do risco de crédito é essencial para o setor financeiro, particularmente para avaliar a probabilidade de inadimplência de clientes em potencial. Em um ambiente cada vez mais orientado a dados, a aplicação de algoritmos de aprendizado de máquina surge como uma abordagem promissora para prever o comportamento do mutuário.

Este relatório detalha um projeto abrangente focado na implementação de modelos de aprendizado de máquina para avaliação de risco de crédito. O principal objetivo é desenvolver modelos preditivos capazes de classificar com precisão os clientes com base em seu risco de crédito, apoiando a tomada de decisão no processo de empréstimo. O projeto segue uma metodologia de gestão bem definida, com ênfase em etapas como análise exploratória de dados, preparação de dados, implementação de algoritmos de aprendizado de máquina e análise dos resultados.

No âmbito deste projeto, explorarei o conjunto de dados "Risco de Crédito Alemão - Com Alvo" da plataforma Kaggle. Este conjunto de dados fornece informações detalhadas sobre clientes de bancos alemães, incluindo idade, renda, histórico de crédito e status de inadimplência em empréstimos anteriores. A abordagem metodológica visa não apenas desenvolver modelos preditivos robustos, mas também gerar insights que possam aprimorar a análise de risco de crédito, beneficiando tanto as instituições financeiras quanto os potenciais clientes.

# Plano de Projeto

O projeto segue o seguinte plano:

* Definição do problema e objetivos.
* Compreensão dos dados e do contexto de negócios.
* Análise exploratória dos dados.
* Pré-processamento de dados.
* Implementação de algoritmos de aprendizado de máquina.
* Avaliação e otimização dos modelos.
* Apresentação de resultados e conclusões.
* Identificação de desafios e recomendações futuras.

# Compreensão do negócio

O principal objetivo deste projeto é a construção de modelos preditivos a fim de melhorar o processo de tomada de decisão na concessão de empréstimos. Buscando reduzir o risco de inadimplência e aumentar a eficiência operacional das instituições financeiras por meio da aplicação de técnicas avançadas de aprendizado de máquina.

# Compreensão de dados

O conjunto de dados "Risco de Crédito Alemão - Com Alvo" foi usado, fornecendo informações sobre características do cliente, histórico de crédito e status de inadimplência.

Uma captura de tela de um computador

Descrição gerada automaticamente

Fig1: Conjunto de dados de risco de crédito alemão

Um gráfico mostrando uma distribuição de risco de crédito

Descrição gerada automaticamente

Fig 2: Distribuição do risco de crédito

# Análise Exploratória de Dados

Realizei análises estatísticas e visualizações para explorar as relações entre as variáveis e identificar padrões nos dados. Usei uma variedade de gráficos, como histogramas, boxplots e mapas de calor, para visualizar as distribuições e correlações entre as variáveis.

# Amostras das análises realizadas.

Notei alguns insights interessantes ao observar a distribuição etária por gráficos de risco. No primeiro gráfico, mostrando a sobreposição de distribuições etárias para clientes considerados "bons" (verde) e "ruins" (vermelho), notei que a distribuição etária para clientes "bons" parece ser mais ampla e tende a ser mais uniforme em comparação com a distribuição etária para clientes "ruins". No entanto, há uma sobreposição considerável entre as duas distribuições, indicando que a idade por si só pode não ser um fator determinante na previsão do risco de crédito. Outros fatores podem estar influenciando a classificação de risco.

Um gráfico mostrando a distribuição etária

Descrição gerada automaticamente

Figura 3: Distribuição etária

Além disso, o gráfico Contagem de Idade por Risco mostra a contagem de clientes em diferentes faixas etárias dividida por risco de crédito. Nota-se que as faixas etárias mais representadas tanto para os grupos "bons" quanto para os "ruins" estão entre aproximadamente 20 e 40 anos. No entanto, a proporção de clientes "ruins" parece ser maior nas faixas etárias mais jovens, enquanto a proporção de clientes "bons" aumenta nas faixas etárias mais velhas. Isso sugere que os clientes mais jovens podem ter maior probabilidade de serem classificados como "ruins", enquanto os clientes mais velhos têm maior probabilidade de serem classificados como "bons". Essas análises nos ajudaram a entender melhor o comportamento das variáveis.

Um gráfico de contagem de idade por risco

Descrição gerada automaticamente

Figura 4: Contagem de idade por risco

Os gráficos indicam como os valores de crédito são distribuídos entre os grupos de risco "bom" e "ruim". Ao analisar a distribuição dos valores de crédito, observei que a tabela de violino mostra uma densidade maior em valores de crédito menores para o grupo de risco "ruim". Isso sugere que os clientes com crédito mais baixo têm maior probabilidade de serem classificados como "ruins". Por outro lado, o boxplot revela que o valor médio de crédito para o grupo de risco "bom" é maior, o que sugere que os clientes com crédito mais alto têm maior probabilidade de serem classificados como "bons".

Além disso, examinei a variabilidade e a presença de valores discrepantes nos valores de crédito entre os grupos de risco. Observamos que a presença de muitos outliers no grupo de risco "ruim" no boxplot pode indicar maior variabilidade nos valores de crédito concedidos a clientes de maior risco. Essa observação é complementada pelo gráfico de violino, que mostra uma distribuição mais ampla e menos concentrada para o grupo de risco "ruim". Essas análises nos forneceram informações sobre como os valores de crédito se relacionam com os grupos de risco e nos ajudaram a entender melhor os padrões nos dados.

Um gráfico de vários pontos vermelhos e azuis

Descrição gerada automaticamente com confiança média

Figura 5: Distribuição do valor do crédito por nível de risco

Outra análise revela que a distribuição dos empréstimos classificados como "ruins" é mais ampla e tem maior densidade em prazos mais longos. Isso sugere que os empréstimos com durações mais longas têm maior probabilidade de serem classificados como de alto risco ("ruins").

Um diagrama de um diagrama mostrando uma variedade de formas

Descrição gerada automaticamente com confiança média

Figura 6: Distribuição da duração do empréstimo por risco de crédito

Olhando para o gráfico de Distribution of Housing bu Risk, pode-se ver que há um número significativamente maior de indivíduos com um risco bom ("bom") em comparação com um risco ruim ("ruim"). Isso sugere que possuir uma casa pode estar associado a um melhor risco de crédito, ou seja, indivíduos que possuem suas casas geralmente têm maior estabilidade financeira, o que pode ser um indicador de melhor risco de crédito.

Aqueles que alugam ("aluguel") podem ter diferentes níveis de estabilidade financeira. O aluguel não indica necessariamente alto risco de crédito, mas pode refletir uma faixa de renda variável ou preferências pessoais. A distribuição mais equilibrada pode sugerir que o aluguel não é um forte indicador de risco de crédito por si só.

Indivíduos que vivem de graça ("de graça") podem ter situações diferentes, como morar com a família. Esta categoria pode ter menos pessoas, mas analisar a distribuição de riscos dentro desta categoria pode revelar se a moradia gratuita está associada a alguma forma de suporte financeiro ou dependência, influenciando o risco de crédito.

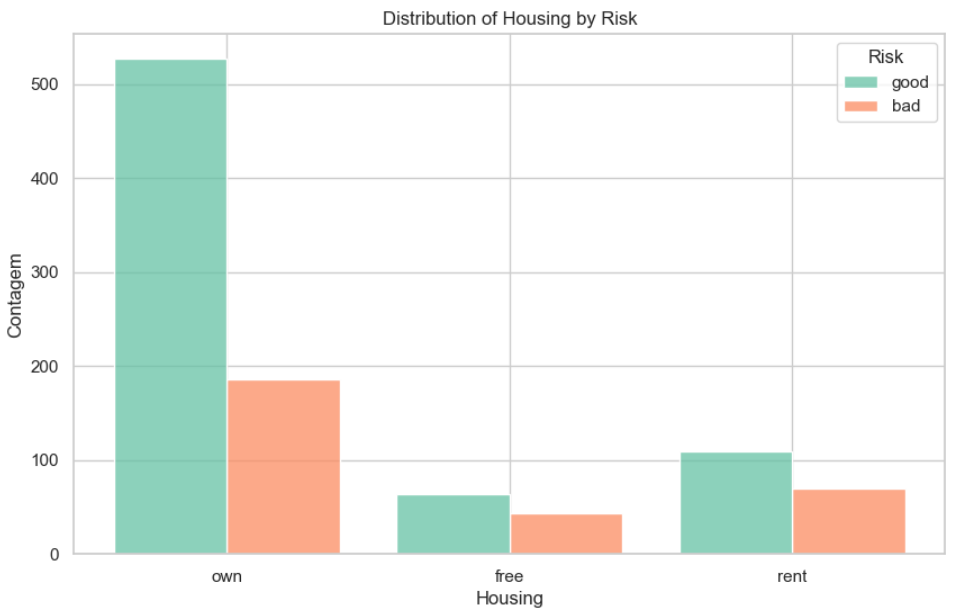


Figura 7: Distribution of Housing by Risk

Tambem podemos observar que os homens tendem a solicitar empréstimos com mais frequência em comparação às mulheres. No entanto, quando analisamos a média dos grupos de homens e mulheres, verificamos que há uma distribuição equilibrada entre ambos os sexos no que diz respeito às finalidades dos empréstimos. Isso indica que, embora os homens peçam mais empréstimos em termos de quantidade, os motivos pelos quais homens e mulheres solicitam crédito são semelhantes em proporção. Em outras palavras, os motivos para solicitar empréstimos não diferem significativamente entre os gêneros.

A diagram of a person's body

Description automatically generated

Figura 8: Comparison of Sex and Purpose

Várias outras análises foram realizadas para explorar padrões complexos e relações entre as variáveis. Devido à extensão e complexidade das análises, seus detalhes podem ser visualizados no projeto Python que será submetido junto com este relatório. O código-fonte e as visualizações complementares oferecem uma visão mais profunda e técnica das abordagens e insights desenvolvidos ao longo do projeto.

# Preparação de dados

As etapas de pré-processamento incluíram tratamento de valores ausentes, codificação de variáveis categóricas, padronização de variáveis numéricas e seleção de características. Isso garantiu que os dados estivessem em um estado ideal para a criação de modelos preditivos precisos e eficazes.

# Identificação de outliers

Usei uma função para detectar outliers em cada coluna do DataFrame calculando os limites inferior e superior com base no intervalo interquartil (IQR). Valores fora desses limites foram considerados outliers.

A blue squares with white text

Description automatically generated

Figura 9: Outlier Count by Column.

Os resultados indicaram a presença de outliers em várias colunas, incluindo "Idade", "Trabalho", "Valor do crédito" e "Duração".

# Tratamento de outliers

Substitui valores discrepantes por limites calculados para minimizar seu impacto negativo.

Após a aplicação dessa função, verifiquei as mudanças na estatística descritiva, observando uma redução na média e no desvio padrão para algumas variáveis, indicando uma menor dispersão dos dados.

Uma tabela de números e números

Descrição gerada automaticamente

Figura 10: Tabela de Estatística descritiva

\*A idade média (Idade) diminuiu ligeiramente.

\*O desvio padrão da idade (Idade) também diminuiu, indicando uma menor dispersão dos dados em torno da média.

\*O valor máximo para a idade foi ajustado para 64,5, o que mostra que os valores discrepantes foram tratados.

\*A média e o desvio padrão para o valor do crédito foram reduzidos, sugerindo uma diminuição na dispersão dos valores do crédito.

\*O valor máximo de Duração foi ajustado para 42, indicando que os utliers também foram tratados nesta coluna.

# Codificação de variáveis categóricas

Para permitir que os modelos de aprendizado de máquina funcionem com variáveis categóricas, transformei essas variáveis em variáveis de codificação one-hot.

Uma captura de tela de um computador

Descrição gerada automaticamente

Figura 11: Codificação de variáveis categóricas

# Padronização

A padronização dos dados foi importante para que todas as variáveis numéricas tivessem a mesma escala, facilitando a realização de diversos algoritmos de aprendizado de máquina. Usei o método RobustScaler para esta tarefa, que é robusto para outliers.

A análise da estatística descritiva após a padronização confirma que as colunas foram corretamente padronizadas, com médias próximas de zero e desvios-padrão próximos de um.

Análises mais detalhadas podem ser visualizadas no projeto Python que será enviado junto com este relatório.

# Análise de correlação e mapa de calor

A análise de correlação ajuda a entender as relações entre as variáveis independentes e a variável-alvo. Usei um mapa de calor para visualizar as correlações.

Um gráfico com números e uma linha vermelha

Descrição gerada automaticamente com confiança média

Figura 11: Mapa de calor

Essa visualização permitiu a identificação de relações fortes e fracas entre as variáveis, auxiliando na seleção de características relevantes para modelagem

# Implementação de aprendizado de máquina

Vários algoritmos foram treinados e avaliados, incluindo regressão logística, árvores de decisão, floresta aleatória, SVM, XGBoost, entre outros. A avaliação dos modelos foi realizada por meio de validação cruzada e métricas de desempenho.

A graph of a comparison of machine learning

Description automatically generated

Figura 12: Comparação dos Modelos de Aprendizagem de Maquinas

* Regressão Logística (LGR): 0,883371 (0,067871)

Reconvocação média: 88,33% Desvio padrão: 6,78% Modelo estável com alta reconvocação, indicando que identifica bem as amostras positivas.

* Análise discriminante linear (LDA): 0,870533 (0,070737)

Recordação média: 87,05% Desvio padrão: 7,07% Bom desempenho com uma ligeira variação entre as dobras.

* K-Vizinhos mais próximos (KNN): 0,838311 (0,075017)

Recall médio: 83,83% Desvio padrão: 7,50% Recall menor em comparação com os modelos LGR e LDA, mas ainda bom.

* Classificador de árvore de decisão (CART): 0,766250 (0,084334)

Recall médio: 76,62% Desvio padrão: 8,43% Desempenho inferior e relativamente estável.

* Gaussiano Naive Bayes (NB): 0,745714 (0,080858)

Reconvocação média: 74,57% Desvio padrão: 8,08% Desempenho semelhante ao CART, mas com maior variação entre dobras.

* Floresta aleatória (RF): 0,870740 (0,061147)

Recall médio: 87,07% Desvio padrão: 6,11% Bom desempenho e estabilidade razoável.

* Máquina de vetores de suporte (SVM): 0,955443 (0,039990)

Recall médio: 95,54% Desvio padrão: 3,99% Excelente desempenho com muito pouca variação, indicando alta precisão e robustez.

* XGBoost (XGBM): 0.844139 (0.056622)

Reconvocação média: 84,41% Desvio padrão: 5,66% Bom desempenho, mas inferior a LGR, LDA, KNN e RF.

A análise de recall indica que o SVM foi o modelo com maior eficiência em capturar corretamente as classes positivas, tornando-o ideal para aplicações que priorizam essa métrica, especialmente em dados desbalanceados. O Random Forest também apresentou bom desempenho com recall alto e baixa variabilidade, sendo uma escolha confiável e robusta. Em suma, SVM e Random Forest destacam-se como opções sólidas para tarefas onde maximizar o recall é essencial.

# Balanceamneto de Classes com SMOTE

Em problemas de classificação com classes desbalanceadas, como o deste projeto (70% da classe "True" e 30% da classe "False"), o modelo pode ser tendencioso e priorizar a classe majoritária, prejudicando a identificação correta da classe minoritária. O SMOTE (Synthetic Minority Over-sampling Technique) foi utilizado para gerar amostras sintéticas da classe minoritária, equilibrando as classes. Esse balanceamento melhora a capacidade do modelo de generalizar e prever ambas as classes com maior precisão, especialmente em cenários em que a classe minoritária tem alta relevância (como a detecção de clientes inadimplentes). Ao balancear as classes, assegura-se que o modelo recebe informações equilibradas, essencial para a construção de um modelo mais robusto e menos tendencioso.

# Construcao do Pipeline e Pré-Processamento

Para otimizar o treinamento e a eficiência do modelo, foi desenvolvido um pipeline de pré-processamento que realiza as etapas essenciais de normalização, seleção de variáveis e redução de dimensionalidade. Isso incluiu:

# 9.1 Normalização dos Dados (StandardScaler)

Aplicado StandardScaler para melhorar o desempenho e a convergência dos modelos.

# 9.2 Seleção de Variáveis e PCA

A aplicação de SelectKBest e PCA ajudou a selecionar as variáveis mais significativas e reduzir a dimensionalidade, aumentando a interpretabilidade e reduzindo o risco de sobreajuste. Esses passos facilitam a análise e asseguram que o modelo se concentre nas informações mais relevantes para a classificação.

# 9.3 Ajuste de Hiperparâmetros com GridSearchCV

Para maximizar o desempenho de cada modelo, realizei uma busca em grid de hiperparâmetros para modelos distintos. Utilizando o GridSearchCV, os parâmetros ideais para cada classificador foram determinados, focando em métricas de performance essenciais (como o F1 Score). Esta abordagem permite comparar modelos com configurações otimizadas, facilitando a escolha de modelos que melhor atendam aos objetivos do projeto.

# 9.4 Avaliação e Comparação dos Modelos

Após a otimização dos hiperparâmetros, cada modelo foi avaliado usando as métricas de accuracy, precision, recall e f1-score. Essas métricas fornecem uma visão completa sobre a eficácia do modelo em prever corretamente ambas as classes. Além disso, o uso da matriz de confusão permite uma avaliação visual e interpretável do desempenho, especialmente no que tange a erros de classificação entre as classes positivas e negativas.

A graph of a bar chart

Description automatically generated with medium confidence

Figura 13: Model Performance Comparison

# 9.5 Identificação do Melhor Modelo

Com os resultados das métricas de avaliação, foi possível identificar o SVM como o modelo de melhor desempenho, com um F1 Score de 0.798, seguido pelo Random Forest e o XGBoost. A escolha do SVM indica que ele se adequou bem aos dados e que pode oferecer uma boa generalização em previsões futuras.

Em resumo, cada etapa do pipeline — desde o balanceamento inicial até o ajuste final de hiperparâmetros — foi fundamentada na busca por maximizar a precisão do modelo na identificação de clientes inadimplentes, um objetivo central do projeto.

# Resultados

Support Vector Machine (SVM) : Este modelo se destacou ao alcançar um F1 Score de 0.7986, o que indica uma boa combinação de precisão e recall. A SVM é conhecida por sua eficácia em conjuntos de dados complexos e em alta dimensão, o que a torna uma escolha robusta para este tipo de problema.

Random Forest: Um modelo de ensemble que combina múltiplas árvores de decisão, mostrando-se eficaz na redução de overfitting e melhorando a precisão nas previsões.

XGBoost: Outro modelo de boosting que se destacou por sua performance, embora tenha ficado um pouco atrás do SVM em termos de F1 Score.

Precisão: A proporção de verdadeiros positivos em relação ao total de previsões positivas. Essa métrica é vital em cenários onde a penalidade por falso positivo é alta.

Recall: A proporção de verdadeiros positivos em relação ao total de positivos reais. É crucial para entender a capacidade do modelo de capturar todos os inadimplentes.

F1 Score: Uma média harmônica entre precisão e recall, oferecendo um único valor que representa o desempenho do modelo em equilibrar ambas as métricas. Este foi o critério principal que utilizei para comparar os modelos.

No final, embora tanto o Random Forest quanto o XGBoost apresentassem desempenhos competitivos, o modelo SVM foi escolhido como o melhor devido à sua superioridade em equilibrar precisão e recall, refletindo a importância de capturar corretamente a inadimplência para a mitigação de riscos nas instituições financeiras. A escolha deste modelo enfatiza a relevância de métodos analíticos no fortalecimento da capacidade preditiva e na segurança das decisões de crédito.

A green and white rectangle with numbers

Description automatically generated

Tabela 2: Tres Melhores Modelos

A graph of different colored squares

Description automatically generated with medium confidence

Figura 14: Machine Learning Model Comparison

# Desafios

Alguns desafios enfrentados durante o projeto incluíram o tratamento de valores ausentes e o ajuste de hiperparâmetros do modelo.

# Conclusão

A conclusão deste relatório evidencia o sucesso do projeto de implementação de modelos de aprendizado de máquina na avaliação do risco de crédito, destacando a importância da análise preditiva na tomada de decisões financeiras. Através de uma abordagem estruturada que abrange desde a análise exploratória de dados até a escolha do modelo mais eficaz, foi possível identificar padrões relevantes e características que impactam a classificação do risco de crédito.

Os resultados mostraram que a Máquina de Vetores de Suporte (SVM) se destacou como o modelo de melhor desempenho, não apenas por sua precisão, mas também por sua robustez em capturar as nuances dos dados. Este modelo, juntamente com o Random Forest e o XGBoost, ofereceu uma base sólida para a identificação de clientes em risco de inadimplência, contribuindo significativamente para a mitigação de riscos nas instituições financeiras.

Além disso, a realização de um pipeline de pré-processamento eficiente e a utilização de técnicas como o balanceamento de classes através do SMOTE reforçaram a eficácia dos modelos, permitindo uma generalização adequada nas previsões e melhorando a capacidade do sistema em lidar com dados desbalanceados.

Por fim, o relatório também enfatiza a necessidade de continuidade no aprimoramento dos modelos, propondo a exploração de novas técnicas e algoritmos nos próximos meses. Essa visão para o futuro é fundamental, pois o ambiente financeiro está em constante evolução, e a adaptação a novas informações e tendências será crucial para manter a relevância e a precisão dos modelos desenvolvidos.

# Referência

Bache, K. & Lichman, M. (2013). Repositório de aprendizado de máquina UCI. Irvine, CA: Universidade da Califórnia, Escola de Informação e Ciência da Computação. Disponível em: <http://archive.ics.uci.edu/ml>.

Kaggle. (s/d). Desempenho dos alunos em exames. Disponível em: <https://www.kaggle.com/spscientist/students-performance-in-exams>.

OCDE. (2010). Resultados do PISA 2009: Aprendendo a Aprender – Engajamento do Aluno, Estratégias e Práticas (Volume III). Paris: Publicação da OCDE. Disponível em: <http://dx.doi.org/10.1787/9789264083943-en>