**LEONARDO ADERALDO VARGAS**

**ANÁLISE DE RISCO DE CRÉDITO DIRECIONADA POR MODELAGEM MATEMÁTICA E APRENDIZADO DE MÁQUINA**

Sorocaba - SP

2023

**LEONARDO ADERALDO VARGAS**

**ANÁLISE DE RISCO DE CRÉDITO DIRECIONADA POR MODELAGEM MATEMÁTICA E APRENDIZADO DE MÁQUINA**

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Bacharel em Engenharia de Controle e Automação pela Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Sorocaba.

Orientador: Prof. Dr. Xxxxxxxxxx

Orientador: Prof. Dr. Xxxxxxxxxx

Orientador: Prof. Dr. Galdenoro Botura Junior

Coorientador: Prof. Dr. Leopoldo André Lusquino Filho

Sorocaba - SP

2023

**AGRADECIMENTOS**

Primeiramente agradeço a Deus, pois sempre recorri a Ele em minhas orações pedindo para que meus caminhos permanecessem abertos e iluminados.

Agradeço à minha família, especialmente minha mãe Marcia, meu irmão Gustavo e minhas tias Eulalia e Maria Aparecida, os quais foram responsáveis pela melhor educação, alicerce, amor e incentivos que uma pessoa pode ter.

Agradeço à minha namorada Patrícia, a qual sempre acreditou em mim e me apoiou em todas as circunstâncias.

Agradeço a todos meus amigos da faculdade, os quais foram e ainda são grandes companheiros no quesito acadêmico, profissional e pessoal.

Agradeço a todos meus colegas do Banco Santander, empresa a qual eu atuo, por todos os ensinamentos adquiridos no cotidiano e por me mostrarem que este trabalho possui um grande impacto positivo para a sociedade.

Agradeço aos professores Galdenoro Botura Junior e Leopoldo André Dutra Lusquino Filho pelo interesse no tema, pelas excelentes orientações durante a execução e por acreditarem que eu pudesse fazer um bom trabalho.

Agradeço à faculdade Universidade Estadual Júlio de Mesquita Filho (UNESP), por proporcionar-me uma formação acadêmica excepcional a qual é um grande motivo de orgulho para mim.

**RESUMO**

Processos de concessão de crédito sempre estarão interligados a perdas, portanto, o objetivo é maximizar a assertividade e minimizar os riscos, visto que decisões ruins podem acabar endividando o cliente e acabando com a saúde financeira da empresa. Neste âmbito, este trabalho busca apresentar um processo de concessão de crédito automático auxiliado por métodos quantitativos de modelagem matemática e aprendizado de máquina. Visando combater a inadimplência e erros operacionais, o software desenvolvido nas linguagens Python e PySpark tem como intuito receber uma base de dados de clientes novos, aplicar a modelagem e classificar quais clientes podem ou não receber o crédito. Durante a pesquisa, explicou-se a fundamentação teórica da análise de crédito e do aprendizado de máquina, proporcionando base técnica para o correto entendimento dos processos. A base de dados é fictícia, contudo, a problemática consiste em ajudar pessoas carentes e desbancarizadas a fim de proporcionar uma experiência de empréstimo positiva e segura, portanto, este trabalho explora a fundo o racional analítico por trás de um processo de análise de crédito, desde a análise histórica, construção de novas variáveis e aplicação de modelos de aprendizado de máquina. Com uma ferramenta automática de análise de crédito, a instituição passa a explorar todo o potencial de seus dados e garante que bons pagadores sejam beneficiados com taxas de juros menores e a inadimplência seja minimizada. Com isso, nota-se a grande importância da implantação de metodologias automáticas através de softwares matemáticos por parte das empresas, visto que o processo de concessão de crédito torna-se muito mais seguro, rápido e rentável.

**Palavras-chave:** Concessão de Crédito, Análise de Crédito, Modelagem Matemática, Aprendizado de Máquina.

**ABSTRACT**

Credit granting processes will always be linked to losses, therefore, the objective is always to maximize assertiveness and minimize risks, since bad decisions can end up indebting the customer and ending the company’s financial health. In this context, this work seeks to present an automatic credit granting process aided by quantitative methods of mathematical modeling and machine learning. In order to combat default and operational errors, the software developed in Python and PySpark languages is intended to receive a database of new customers, apply the modeling and classify which customers may or may not receive credit. During the research, the theoretical basis of credit analysis and machine learning was explained, providing a technical basis for the correct understanding of the processes. The database is fictitious, however, the problem is to help needy and unbanked people in order to provide a positive and safe loan experience, therefore, this work deeply explores the analytical rationale behind a credit analysis process, from historical analysis, construction of new variables and application of machine learning models. With an automatic credit analysis tool, the institution starts to explore the full potential of it’s data and ensures that good payers benefit from lower interest rates and defaults are minimized. With this, the great importance of the implantation of automatic methodologies thourgh mathematical software by the companies is noticed, since the process of granting credit becomes much safer, faster and more profitable.

**Keywords:** Credit Granting, Credit Analysis, Mathematical Modeling, Machine Learning.

**Lista de iLUSTRAÇÕES**

Figura 1 – Fluxograma do Ciclo de Crédito 8

Figura 2 - 10

Figura 3 - 19

Figura 4 - 20

Figura 5 – 21

Figura 6 – 21

Figura 7 – 22

Figura 8 – 22

Figura 9 – 23

Figura 10 – 23

**LISTA DE TABELAS**

Tabela 1 – Diagrama Rating x Perda Esperada 10

Tabela 2 – Descrição detalhada das Etapas do CRISP-DM 15

Tabela 3 - 33

Tabela 4 - 33

Tabela 5 - 35

Tabela 6 - 36

Tabela 7 - 36

Tabela 8 - 37

Tabela 9 - 42

Tabela 0 - 43

**Lista de Abreviaturas**

|  |  |  |
| --- | --- | --- |
| SPC Brasil  BACEN | | Serviço de Proteção ao Crédito  Banco Central do Brasil |
| ML | | Machine Learning |
| PD | | Probability of Default |
| EAD | | Exposure at Default |
| LGD | | Loss Given Default |
| EL | | Expected Loss |
| SVM | | Support Vector Machines |
| MC | | Matriz de Confusão |
| VN | | Verdadeiro Negativo |
| VP | | Verdadeiro Positivo |
| FN | | Falso Negativo |
| FP | | Falso Positivo |
| AUC | | Área Sob a Curva |
| KS | | Teste de Hipótese de Kolmogorov-Smirnov |
| CRISP-DM | | Cross Industry Standard Process for Data Mining |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  |  |

**sumário**

**1 INTRODUÇÃO** 1

1.1 Considerações Iniciais 1

1.2 Contexto e justificativa 2

1.3 Objetivos 3

**2 CONCEITUAÇÃO** 4

**2.1 Fundamentos do Risco de Crédito** 4

2.1.1 Crédito e seus princípios 4

2.1.2 O processo de decisão 6

**2.2 Fundamentos de Aprendizado de Máquina** 9

2.2.1 Importância do Aprendizado de Máquina 9

2.2.2 Diferença entre Inferência e Predição 10

2.2.3 Modelos de Aprendizado de Máquina para Classificação 10

2.2.4 Pré-Processamento 23

2.2.5 Otimização de Modelos 23

2.2.6 Métricas de Avaliação 30

**3 MATERIAIS E MÉTODOS** 30

**3.1 Materiais** 30

3.1.1 Python 30

3.1.2 PySpark 31

3.1.3 GIT 32

3.1.4 Plataforma Kaggle 33

**3.2 Métodos** 33

3.2.1 Metodologia CRISP-DM 33

**4 METODOLOGIA** 34

**4.1 Entendimento da Base de Dados** 20

4.1.1 Descrição do Problema 20

4.1.2 Significado das Variáveis 20

4.1.3 Separação em Treino e Teste 20

**4.2 Análise Exploratória** 22

4.2.1 Análise Univariada 22

4.2.1 Análise Bivariada 22

**4.3 Engenharia de Atributos** 22

4.3.1 Features A 22

4.3.2 Features B 22

4.3.3 Features C 22

4.3.4 Features D 22

**4.4 Pré-Processamento de Dados** 22

4.4.1 Valores Ausentes 22

4.4.2 Encoding 22

4.4.3 Escalonamento 22

**4.5 Aplicação de Modelos de Aprendizado de Máquina** 23

4.5.1 Motor de Modelos 23

4.5.2 Análise de Métricas 23

4.5.3 Seleção de Features 23

4.5.4 Otimização de Hiperparâmetros 23

**5 RESULTADOS E DISCUSSÕES** 31

**5.1 Aplicação do Modelo Treinado em Dados Novos** 31

**5.2 Tradução e Quantificação dos Resultados** 31

**5.3 Estudo e Aplicação de Políticas de Crédito** 31

**5.4 Comparação com a Metodologia Tradicional** 31

**6 CONCLUSÃO** 53

**6.1 Resumo dos Resultados Obtidos** 31

**6.2 Resposta da questão de pesquisa** 31

**6.3 Limitações do Estudo** 31

**6.4 Sugestões para trabalhos futuros** 31

**REFERÊNCIAS** 55

**ANEXO** 56

**1) INTRODUÇÃO**

**1.1 Considerações iniciais**

Crédito é uma palavra derivada do latim, significando “confiança”. Ele é uma espécie de empréstimo solicitado por um cliente a alguma instituição financeira visando antecipar algum tipo de gasto, auxiliando o solicitante quando o mesmo não possui o capital.

Implantado durante a Revolução Industrial, o crédito possibilitou que pessoas planejassem investimentos e abrissem seu próprio negócio, empresas adquirissem novas tecnologias e aumentassem sua produção e as instituições financeiras obtivessem lucro e capital de giro para ampliar seu patrimônio. Sebben (2020) afirma que essa situação resultou em inúmeras ascensões sociais e movimentação da economia, portanto, tornou-se uma operação extremamente importante para o desenvolvimento da sociedade em geral.

Pelo fato de ser um empréstimo, a concessão de crédito é realizada sob condições de incerteza, logo, ao antecipar recursos, a instituição naturalmente insere-se em um ambiente extremamente sujeito a perdas, muitas vezes por situações ao acaso que acontecem com os tomadores, mas também por conta de pessoas fraudadoras e má intencionadas que solicitam crédito já sabendo que não o pagarão. Por tratar-se de uma incerteza, esse processo está diretamente ligado a riscos, logo, o credor necessita de algumas garantias as quais serão protocoladas após uma análise criteriosa sobre diversas informações a respeito do cliente.

O risco de crédito nada mais é do que a probabilidade de perda financeira decorrente do não cumprimento de obrigações de pagamento por parte do solicitante, logo, uma boa análise baseia-se em determinadas metodologias a fim de garantir maior confiabilidade e segurança, resultando na criação de uma “nota de cliente”, classificando-o como bom ou mau pagador e definindo quais taxas de juros será submetido.

Tal análise fundamenta-se em inúmeras variáveis compostas de dados internos e externos sobre o cliente, pois é de suma importância definir corretamente o perfil do solicitante a fim de melhorar a gestão, mitigar os riscos, atender às necessidades do cliente e garantir que a instituição financeira receba o pagamento para continuar existindo. Um processo de concessão de crédito mal feito pode endividar um cliente e prejudicar a saúde financeira da empresa, causando severos danos a ambos os lados.

Como envolve diversas variáveis e diversos clientes, as bases de dados costumam ser bastante extensas e analisá-las a olho nu torna-se praticamente impossível. Nesse âmbito, a utilização de softwares baseados em técnicas matemáticas e estatísticas tornam-se fundamentais, pois permitem a classificação em larga escala, de forma automática e segura, criando-se assim o que chamamos de “Modelos de Risco de Crédito”, os quais, segundo Laredo (2010), possuem o objetivo de prever, na data de decisão do crédito, a probabilidade da concessão tornar-se uma perda para o credor.

**1.2 Contexto e justificativa**

Oliveira Lima (2008) diz que a história mostra que mesmo bons clientes têm potencial de não honrarem suas obrigações financeiras. Segundo levantamentos realizados pelo Serasa e pelo SPC Brasil, a inadimplência brasileira alcançou números alarmantes. Comportando em torno de 65 a 70 milhões de pessoas devedoras, o Brasil conta com uma das maiores inadimplências mundiais.

Ao não pagarem suas dívidas, as instituições financeiras deixam de receber o pagamento e os clientes tendem a endividarem-se cada vez mais, prejudicando todo o ecossistema econômico. Nesse âmbito, bancos e fintechs naturalmente enrijecem seu processo de concessão de crédito e elevam as taxas de juros, prejudicando os bons pagadores de concretizarem muitos sonhos. Empresas deixam de existir, negócios deixam de vender e muitas pessoas não conseguem financiar seu imóvel ou carro após longos anos de trabalho árduo.

Como forma de auxiliar o cenário, as instituições financeiras reinventam-se a todo momento, seja atualizando as políticas ou buscando novas metodologias, resultando em fortes cobranças sobre os analistas de crédito, os quais passam boa parte do tempo analisando informações minuciosamente de forma manual e gastando esforço em partes que podem ser automatizadas.

Visando acabar com este tipo de problema, a proposta é criar um software matemático capaz de classificar corretamente os bons pagadores de forma automática e segura, baseando-se exclusivamente em dados. Além disso, como pauta-se completamente em técnicas matemáticas e estatísticas, esta metodologia capta muitas vezes relações entre as variáveis que humanos só enxergam após cálculos extremamente complexos e que demandam muito tempo.

Dessa forma, ao invés de analistas de crédito gastarem tempo conferindo as informações a olho nu, eles podem direcionar seus esforços para outras questões, como a implementação do modelo e até mesmo a melhoria constante do algoritmo.

**1.3 Objetivos**

O trabalho “Análise de Risco de Crédito direcionada por Modelagem Matemática e Aprendizado de Máquina” apresenta uma introdução ao mundo do crédito auxiliado por técnicas de Matemática Aplicada, Estatística e Machine Learning para a classificação de clientes inadimplentes, a fim de auxiliar a tomada de decisão.

Direcionada pelas técnicas e ferramentas de Ciência de Dados, o software desenvolvido será responsável pela concessão de crédito para pessoas físicas de forma rápida e automática, pautando-se exclusivamente em conceitos matemáticos e estatísticos a fim de maximizar a confiança e minimizar o risco do credor.

A base de dados utilizada será fictícia e advém da plataforma Kaggle, visto que bases de dados de concessão de crédito possuem inúmeras informações sensíveis, contudo, ela simula muito bem um ambiente real de uma instituição financeira.

O trabalho visa aplicar o conhecimento assimilado durante o curso de Engenharia de Controle e Automação em processos de análise de risco e criar um software otimizado para atender as demandas do mercado de crédito.

**2) CONCEITUAÇÃO**

Para correta compreensão do trabalho, é necessário o conhecimento de determinados conteúdos, portanto, apresentou-se alguns fundamentos a respeito de Risco de Crédito e Aprendizado de Máquina a fim de nivelar, ao menos conceitualmente, algumas teorias importantes.

Além disso, o intuito deste trabalho é o desenvolvimento de um software de Aprendizado de Máquina voltado para a correta classificação de pessoas físicas, portanto, embora algumas metodologias sejam análogas para ambos os casos, conceitos exclusivos de classificação de pessoas jurídicas (empresas) não serão abordados.

**2.1 Fundamentos de Risco de Crédito**

* + 1. Crédito e seus princípios

Embora sejam palavras semelhantes, “risco” e “incerteza”, no contexto de crédito, diferem um pouco. Esse risco é uma probabilidade em função de algo eventual e incerto, portanto, não depende de nenhuma das partes. A incerteza ocorre quando a instituição financeira não possui dados sobre o cliente, portanto, torna-se impossível avaliá-lo.

Como dito anteriormente, o risco de crédito nada mais é do que a probabilidade de perda financeira decorrente do não cumprimento de obrigações de pagamento por parte do solicitante. Por tratar-se de uma “operação de confiança”, toda vez que há uma antecipação de recursos há chances da não recuperação do valor e é justamente este risco que o credor aceita passar visto que será recompensado futuramente através dos juros.

Embora o pagamento de juros seja rentável ao banco, deseja-se evitar clientes completamente inadimplentes, pois eles oferecem problemas de rentabilidade e jamais pagarão suas dívidas. Somado ao fato de que o BACEN é um grande balizador sobre a taxa de juros e decide quando a mesma aumenta ou diminui, o cenário de crédito requer mudanças constantes. Dada a situação, o objetivo da análise de risco de crédito é justamente prever os bons e maus pagadores, reduzindo o volume de crédito concedido a pessoas que não poderão honrá-lo ou concedendo volume aos clientes adimplentes.

De acordo com Tchilian (2022), a concessão de crédito aliada a uma boa gestão de riscos representa a principal fonte de renda de instituições financeiras, portanto, para facilitar o ecossistema, criou-se um fluxo chamado Ciclo de Crédito o qual consiste em seis etapas: Captação, Segmentação, Valoração, Manutenção, Cobrança e Recuperação.

A seguir encontra-se um resumo sobre cada uma das etapas:

Figura 1 – Fluxograma do Ciclo de Crédito

****

Fonte: Autoria Própria.

**Captação –** A captação é o ponto de partida do ciclo, onde a instituição financeira busca atrair novos clientes. Isso é feito por meio de estratégias de marketing e vendas, identificando produtos adequados ao público-alvo e direcionando esforços para conquistar novos negócios. A captação pode ser ativa, realizada pelo time comercial da instituição, ou passiva, quando os próprios clientes buscam a instituição.

**Segmentação –** Dado que cada cliente possui diferentes necessidades e perfis financeiros, a segmentação é fundamental. Ela envolve a divisão dos clientes em grupos com características similares, permitindo a oferta de produtos e serviços específicos para atender às necessidades de cada segmento. Isso melhora a eficácia das ofertas e aumenta as chances de satisfação do cliente.

**Valoração –** A etapa de valoração é onde as metodologias de risco de crédito entram em ação. Esses métodos avaliam os clientes com base em diversas variáveis, a fim de classificá-los como bons ou maus pagadores. Além disso, há políticas de crédito, as quais são regras definidas para estabelecer limites e critérios para a concessão de crédito, garantindo consistência nas decisões.

**Manutenção –** A manutenção refere-se à gestão contínua dos relacionamentos com os clientes. Isso envolve o atendimento às necessidades dos clientes ao longo do tempo e o incentivo à fidelidade. Clientes antigos são valorizados, uma vez que sua permanência tende a ser benéfica tanto em termos de rentabilidade quanto de redução do risco de crédito.

**Cobrança e Recuperação –** Esta etapa lida com situações de inadimplência. As políticas de cobrança são implementadas para lidar com clientes que não conseguem cumprir suas obrigações de pagamento. Isso pode envolver medidas como a aplicação de juros, multas ou a renegociação da dívida. O objetivo é minimizar as perdas e manter um equilíbrio entre os lucros da instituição e a recuperação dos valores em débito.

* + 1. O processo de decisão

Na história do sistema financeiro, pelo fato das decisões de empréstimos serem abrangentes e pautarem-se em inúmeras informações, um dos marcos mais significativos foi a introdução dos “5 C’s do Crédito”. De acordo com Sebben (2020), em conjunto, os 5 C’s do Crédito norteiam todo o processo de concessão de crédito criando os principais fatores da análise de risco e auxiliam a expor a probabilidade de um solicitante honrar ou não o pagamento dos empréstimos, sendo fundamentais durante todo o Ciclo de Crédito para que a instituição financeira minimize as perdas e maximize os resultados.

A seguir detalhou-se cada um desses critérios:

**Caráter –** Sendo o elemento básico para decisões de crédito, este critério avalia características pessoais e profissionais do cliente, como sua reputação em termos de integridade e honestidade. Informações como histórico de pagamentos em dia, dívidas em atraso, antecedentes criminais, informações de bureus, entre outras questões fazem parte da avaliação do comportamento e imagem do indivíduo.

**Capacidade –** Refere-se à validação sobre as condições do tomador pagar suas dívidas, avaliando questões como renda e suas fontes, conhecimento técnico, área de atuação profissional, despesas e outras dívidas existentes, além de adequar os valores das prestações e prazos respeitando as limitações do cliente a fim de não o endividar.

**Colateral –** É a garantia do pagamento do empréstimo a qual o credor pode recorrer em casos de inadimplência do solicitante, portanto, são bens de valor como casas, carros, entre outros, avaliando-se o valor desses bens e sua liquidez a fim de decidir se é suficiente para cobrir o valor do empréstimo.

**Condições –** Indica as condições referentes ao contexto econômico no qual o empréstimo será realizado, avaliando as características socioeconômicas do tomador e do mercado nacional, como estabilidade econômica da região, as taxas de juros e as condições do mercado a fim de definir se o momento para concessões é propício.

**Capital –** Apresenta uma análise interna sobre as finanças da instituição a fim de garantir que ela possui o dinheiro solicitado pelo cliente, avaliando questões como patrimônio líquido, balanço financeiro, projeções de rentabilidade, bens e fluxo financeiro interno.

Essa metodologia serviu por muito tempo como método de avaliação da capacidade de um cliente obter crédito, sendo um conjunto de critérios utilizados pelos credores para balizar a concessão e entender se o mesmo é elegível ou não ao empréstimo solicitado. Mesmo sendo comprovadamente eficaz, à medida que a complexidade das transações financeiras e a quantidade de dados disponíveis aumentaram ao longo dos anos, tornou-se improvável a manutenção de técnicas manuais. Nesse contexto, a introdução de modelos estatísticos como metodologia para a concessão de crédito foi amplamente aceita pelas empresas, pois eles fornecem objetividade e precisão na avaliação do risco de crédito de um cliente.

Há dois tipos de modelos capazes de discriminar bons e maus pagadores, sendo eles o Credit Scoring e Behavior Scoring. O primeiro pauta-se nas variáveis cadastrais e informações de mercado externas a fim de definir o score do cliente, ou seja, é como se fosse uma “foto” das informações e aplicado em situações que a instituição não conhece; o segundo é composto por variáveis comportamentais para descobrir a interação dos clientes em relação aos pagamentos, produtos e características históricas como se fosse um “filme” das informações, portanto, trata-se de um modelo bem mais complexo. Embora sejam modelos diferentes, seu intuito é análogo e a diferença são as informações utilizadas, todavia, a saída de ambos métodos são similares, sendo ela a probabilidade do cliente tornar-se inadimplente.

A probabilidade retornada pelo modelo é muitas vezes expressa como uma pontuação de risco e transformada em uma espécie de ranking, sendo uma medida chave que orienta decisões importantes. Na avaliação de risco de crédito, os principais indicadores que desempenham papéis críticos na quantificação do risco envolvido em empréstimos e operações de crédito são a PD, LGD e EAD.

Esses indicadores fornecem insights fundamentais para as instituições financeiras tomarem decisões, alocarem capital adequadamente e implementarem estratégias eficazes de gerenciamento de risco, garantindo que os empréstimos concedidos sejam sustentáveis e lucrativos. A seguir há a definição de cada um deles:

***Probability of Default – A*** Probabilidade de Inadimplência (PD) é um dos pilares centrais da análise de risco de crédito. Calculada através de modelos estatísticos avançados, essa métrica estima a chance de um cliente não honrar suas obrigações de pagamento.

***Loss Given Default –*** A Perda em Caso de Inadimplência (LGD) refere-se ao percentual da exposição que se espera ser perdida em caso de inadimplência. Essa métrica considera o percentual de recuperação esperado em caso de inadimplência.

***Exposure at Default –*** A Exposição em Caso de Inadimplência (EAD) quantifica o valor de exposição sujeito a ser perdido durante uma concessão de crédito em caso de inadimplência do cliente. Essa métrica calcula o valor real esperado caso o cliente não pague o empréstimo.

***Perda Esperada –*** A Perda Esperada (EL) é o valor esperado de perdas que uma instituição financeira espera incorrer em suas exposições de crédito. Essa é a principal métrica de risco de crédito, pois representa uma estimativa das perdas prováveis com base na probabilidade de inadimplência, na recuperação esperada e no valor exposto.

Segundo Oliveira Lima (2008), os termos probabilidade de inadimplência (PD), perda dada a inadimplência (LGD) e exposição a inadimplência (EAD) começaram a dar um contorno mais técnico aos cálculos necessários para a fixação do capital regulatório. Os valores eram estipulados, na grande maioria dos casos, em arbitrários e conservadores. Isso requeria um capital regulatório maior do que o necessário para fazer frente à perda. A abordagem avançada reconhecia esse conservadorismo e passou a permitir que cada instituição pudesse desenvolver modelos internos de fixação dos valores de PD, LGD e EAD. O grande benefício seria medir de forma mais adequada o risco de crédito de sua carteira e assim manter um capital regulatório adequado a essa exigência. Durante este processo, o maior desafio é atender as margens de risco que a instituição está disposta a correr de forma rentável e lucrativa, ao mesmo tempo que proporciona bons produtos e serviços ao cliente.

Para exemplificar, pode-se pensar no financiamento de um veículo. Supondo que o veículo esteja avaliado em R$100.000,00; o financiamento esteja atrelado à garantia de que em caso de inadimplência a instituição recuperará o carro e, consequentemente, 60% do valor inicial; sabe-se que a EAD é de R$60.000,00 e a LGD de 40%. Pensando em um cliente com PD de 10%, a EL pode ser definida como

portanto, para um veículo avaliado em R$100.000,00, a perda esperada seria de R$2.400,00 para clientes com essa PD.

Dessa forma, percebe-se a importância da existência de um modelo matemático bem calibrado para cálculo de PD. Dada a situação, este trabalho comporta a criação de um Modelo de Aprendizado de Máquina capaz de estimar a PD de um cliente a fim de auxiliar os fluxos de concessão de crédito de forma automática e confiável.

* 1. **Fundamentos de Aprendizado de Máquina**
     1. *Importância do Aprendizado de Máquina*

Aprendizado de Máquina é uma subárea da Ciência da Computação responsável pela confecção de algoritmos com a finalidade de reconhecer padrões em dados de forma automática e sem programação prévia. Pautando-se completamente em técnicas matemáticas e estatísticas, este ramo é a base para sistemas de Inteligência Artificial.

Segundo o matemático e cientista da computação Dr. Andrew Ng, um dos grandes especialistas da área, os algoritmos de Aprendizado de Máquina tem a capacidade de revolucionar a sociedade assim como as Revoluções Industriais, o uso de energia elétrica e o descobrimento da Internet. Grandes empresas como Amazon, Google, Apple, Microsoft, grandes bancos entre outras instituições investem muito dinheiro na área, justamente porque esses modelos mudam a forma que nós entendemos grandes volumes de dados e permitem a criação de cenários que há anos atrás pareciam impossíveis, como a confecção de veículos autônomos, detecção de doenças, reconhecimento de linguagem humana, sistemas de recomendação, otimização de rotas físicas, simulações financeiras, entre outras possibilidades.

Devido ao grande poder de decisão dos algoritmos, a demanda por profissionais capacitados intensificou-se nos últimos anos, visto que cientistas de dados definem o rumo de uma instituição, proporcionando destaque competitivo em relação a concorrência. Dessa forma, nota-se que técnicas de Aprendizado de Máquina tornaram-se um dos principais segmentos das Ciências Exatas.

* + 1. *Diferença entre Inferência e Predição*

Kumar (2022) afirma que “Inferência” e “Predição” são conceitos muito relevantes em questões de modelagem, todavia, embora tenham intersecções, são propostas diferentes. Do ponto de vista matemático, a inferência concentra-se em extrair conclusões sobre uma população através de uma amostra. Com um viés muito mais próximo da Estatística Clássica, ela é o processo de avaliar a relação entre a variável dependente e as variáveis independentes, ligando-se a questões como “Quais preditores estão associados com a variável alvo?”, “Como a mudança dos preditores influência na magnitude da variável alvo?”. Durante questões inferências, costuma-se utilizar testes de hipótese e intervalos de confiança para rejeitar ou não determinadas hipóteses.

Por sua vez, a predição simplesmente pauta-se em determinados dados históricos a fim de realizar previsões futuras. Em suma, um modelo matemático é treinado para entender o comportamento médio de um sistema com informações passadas e prever o futuro com a maior assertividade possível. Embora não tenha um caráter estritamente estatístico, ainda sim modelos de predição são construídos com base em técnicas matemáticas e estatísticas avançadas.

* + 1. *Modelos de Aprendizado de Máquina para Classificação*

Modelos de Classificação estão contidos no Aprendizado de Máquina Supervisionado. Segundo Géron (2019), no Aprendizado Supervisionado, os dados são apresentados ao algoritmo com os dados de entrada acompanhados dos resultados, chamados de rótulos. A partir dos rótulos, o modelo é treinado e estima uma função matemática capaz de classificar novas amostras.

Sendo assim, o principal objetivo desses modelos é utilizar dados históricos e previamente rotulados para construir uma representação matemática dos padrões presentes na amostra de treinamento. Essa representação é capturada na forma de um modelo capaz de generalizar e prever, para novas instâncias nunca vistas anteriormente, a qual classe elas pertencem.

Essa classe nada mais é do que a representação de uma probabilidade. Isso significa que, após passar pela equação, a nova instância terá uma determinada probabilidade de pertencer a classe negativa e outra probabilidade de pertencer a classe positiva. No contexto de um modelo de crédito, esse novo elemento terá uma probabilidade estimada de ser qualificado para o empréstimo e uma probabilidade estimada de não ser qualificado.

* + - 1. ***Modelos Lineares de Regressão***

Modelos de Regressão são uma ferramenta muito importante na modelagem estatística. Segundo Morettin e Singer (2021), uma regressão é uma técnica para modelar a relação entre variáveis independentes e uma variável dependente a fim de estimar o valor esperado de uma variável resposta. Os dois modelos lineares mais famosos desse tipo são a regressão linear (voltada para inferir uma variável dependente contínua) e a regressão logística (voltada para inferir uma variável dependente qualitativa).

Em geral, a equação para um modelo de regressão linear pode ser definida como:

em que é o intercepto, o coeficiente angular, *xi* são as variáveis utilizadas, i corresponde ao erro aleatório os quais representam desvios entre os valores observados e preditos.

Bruce e Bruce (2017) afirmam que o intuito desse modelo é estimar os coeficientes que ajustem-se à amostra de treino, entendam os padrões e tornem-se uma equação representativa do fenômeno desejado. Em outras palavras, “estimar o quanto Y mudará quando X mudar em determinada quantidade”. Na figura abaixo, os pontos azuis são cada uma das amostras de uma base de dados e a reta em vermelho é representa a equação do modelo de regressão linear.

Figura 8 – Modelo de Regressão Linear



Fonte: GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems**.**p.92.

Como a regressão linear possui variável dependente contínua, ela não pode ser utilizada para prever a classe de um indivíduo. Nesse âmbito, os estatísticos criaram a regressão logística, a qual é uma extensão da regressão linear, adaptada para lidar com problemas em que a variável de dependente é categórica e sua distribuição é uma distribuição de Bernoulli, como é o caso de eventos binários.

No contexto de modelagem de risco de crédito, a variável dependente é uma variável binária que indica se um indivíduo é considerado um bom ou mau pagador. Dessa forma, ao definir-se a classe positiva como 1 e a negativa como 0, a variável dependente será a probabilidade da instância de pertencer a classe 1.

Para que isso ocorre, necessita-se de uma forma capaz de transformar uma variável dependente contínua contida no intervalo [-∞, ∞] para uma variável categórica binária contida no intervalo [0,1]. A função responsável por transformar valores contínuos em termos de probabilidade é a Função Sigmóide, a qual é uma transformação não-linear. Dessa forma, após a aplicação da Sigmóide na equação a Regressão Linear, tem-se a equação da regressão logística:

sendo assim possível estabelecer uma probabilidade de corte para a classe de interesse acima da qual considera-se um registro como pertence àquela classe, ou, em outras palavras, para definir os limites de separação entre um bom e mau pagador.

Nesse contexto, há também os ajustes em relação aos parâmetros. Conforme Morettin e Singer (2021), os parâmetros são ajustados pelo método da máxima verossimilhança, o qual possui o objetivo de encontrar os coeficientes do modelo que possuem a maior probabilidade de obter uma distribuição análoga a dos dados reais.

Os coeficientes estimados na regressão logística representam a razão entre a probabilidade de sucesso e a probabilidade de falha da variável dependente para cada unidade de mudança nas variáveis independentes. Em outras palavras, eles representam a contribuição relativa de cada variável na previsão da probabilidade da classe e indicam como elas afetam as chances de um indivíduo ser um mau pagador.

Figura 9 – Modelo de Regressão Logística



Fonte: GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems.p.114.

* + - 1. ***Modelos Bayesianos***

Bussab e Morettin (2017) expressam que a probabilidade condicional é um conceito importante na estatística o qual permite lidar com incertezas em situações de eventos dependentes. Nesse âmbito, o Teorema de Bayes é um mecanismo o qual descreve a forma de atualizar a probabilidade de uma hipótese com base em novas evidências. Sua equação é definida como:

sendo P(A) a probabilidade do evento A, P(B) a probabilidade do evento B, P(B|A) a probabilidade a priori e P(A|B) a probabilidade a posteriori.

A partir do Teorema de Bayes, pode-se modelar um fenômeno dada as variáveis, criando-se o Naive Bayes. Bruce e Bruce (2017) descrevem o Naive Bayes como um algoritmo o qual pauta-se na probabilidade de observação de valores preditores, dado um resultado, para estimar a probabilidade de observar o resultado dado um conjunto de preditores. Em outras palavras, o algoritmo cria tabelas de contingência e calcula as probabilidades de cada classe para cada uma das variáveis e, a partir dos resultados, compreende se uma nova instância pertencerá a classe 1 ou 0.

Embora seja uma abordagem muito interessante, este modelo assume que as variáveis são independentes entre si, todavia, é muito raro encontrar um fenômeno de independência como este na prática. Sua equação é expressa como:

sendo a probabilidade dos atributos ocorrerem dada a classe , a probabilidade marginal da classe e a probabilidade condicional da nova instância ser um bom ou mau pagador dado um conjunto de atributos .

Figura 10 – Modelo de Naive Bayes exemplificado para uma Target (PD) e 2 Variáveis (Idade e Se Já Deu Atraso)



Fonte: Autoria Própria.

* + - 1. ***Modelos de Suporte Vetorial***

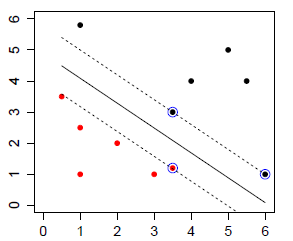
Géron (2019) estabelece que o SVM (Support Vector Machines) é um dos modelos mais versáteis na literatura científica. Inspirado puramente em Álgebra Linear, este algoritmo é capaz de realizar classificações lineares e não lineares de conjuntos de dados de alta complexidade. Embora poderoso, por possuir matemática de alta complexidade, este algoritmo demanda intenso poder computacional, sendo recomendado para conjuntos de dados pequenos ou médios.

A ideia fundamental de uma SVM é identificar o melhor hiperplano separador entre as classes, de tal forma que a margem entre os pontos mais próximos de ambas as classes seja a maior possível. Esses pontos mais próximos são chamados de vetores de suporte e, por serem os pontos mais difíceis de serem classificados dado que estão super próximos, são responsáveis por determinarem a localização e a orientação do hiperplano de separação, asseguram Morettin e Singer (2021).

A equação para um Modelo de SVM de Classificação Linear é definida como:

sendo o viés, os coeficientes de Lagrange associados a cada vetor de suporte Xi, X a nova instância a qual deseja-se classificar e f(x) a probabilidade do novo indivíduo ser um bom ou mau pagador.

Figura 10 – Modelo de SVM para Classificação Linear Binária



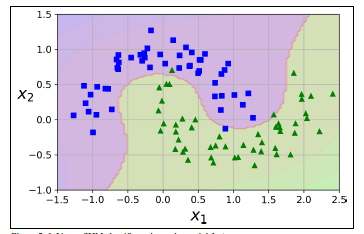
Fonte: MORETTIN, Pedro A.; SINGER, Julio M. Estatística e Ciência de Dados.p.297.

Determinadas situações são complexas de modelar, portanto, aplica-se uma metodologia denominada Kernel Trick, a qual ajusta uma função kernel responsável por calcular o produto escalar entre os pontos nesse espaço a fim de avaliar a similaridade entre dois pontos do espaço de características. Dessa forma, pode-se mapear os dados para um espaço de características de maior dimensionalidade e então traçar um hiperplano linear.

A equação para um Modelo de SVM de Classificação Não-Linear é definida como:

sendo a função de kernel responsável por calcular a similaridade entre o vetor de suporte e a nova instância no espaço de alta dimensão.

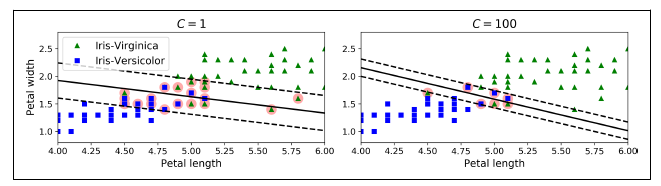
Figura 11 – Modelo de SVM para Classificação Não-Linear Binária



Fonte: GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems.p.124.

A violação das margens ocorre quando pontos de dados ficam no interior das margens ou do lado incorreto do hiperplano de separação, portanto, eles foram classificados incorretamente ou estão muito próximos aos limiares de decisão. Para lidar com esses problemas, há dois tipos de cenários denominados como Margem Rígida e Margem Suave. A primeira refere-se a um modelo de hiperplano com margens mais estreitas nas quais há a diminuição de erros de classificação; ao passo que a segunda retrata um modelo com margens mais largas e permissivas. O controle da margem deriva do parâmetro C, o qual controla o equilíbrio entre viés e variância e define o quanto o modelo aceita erros de classificação (Géron, 2019).

Figura 11 – Detalhes sobre as Margens de um SVM para Classificação Linear Binária



Fonte: GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems.p.122.

* + - 1. ***Modelos Baseados em Distância***

Bruce e Bruce (2017) afirmam que modelos baseados em Distância destacam-se por sua simplicidade conceitual e abrangência. No ramo de classificação, o mais famoso é o KNN (K-Nearest Neighbors).

O KNN compara a nova instância com os *K* elementos mais próximos baseados nas variáveis utilizadas. Essa comparação é realizada via cálculos geométricos de distância e, a depender da distância escolhida, os resultados podem ser distintos. Dentre as distâncias mais famosas estão a Euclidiana, Manhattan e Similaridade de Cossenos.

A distância Euclidiana é a fórmula clássica para cálculos de distância em um espaço tridimensional. Através das coordenadas, ela mede a distância linear direta entre dois pontos no espaço. Além disso, ela é recomendada para variável contínuas.

A distância de Manhattan é similar a Euclidiana, todavia, neste caso ela torna-se interessante em situações as quais a distância linear direta entre dois pontos não é permitida ou não atende a problema em questão. Dessa forma, a distância é calculada apenas em relação a ângulos retos. Assim como a Euclidiana, ela também é recomendada para variável contínuas.

Finalmente, a similaridade de cossenos mede a similaridade direcional entre dois vetores em um espaço tridimensional. Os cálculos são realizados através do produto escalar e da norma dos vetores e o resultado final varia de -1 a 1, ao passo que quanto maior o valor, maior a similaridade entre os vetores. Recomenda-se este tipo de distância tanto para variáveis contínuas quanto qualitativas.

Figura 11 – Modelo de KNN para Classificação



Fonte: BRUCE, Peter; BRUCE Andrew. Practical statistics for data scientists**.** p.219, fig. 6-2.

* + - 1. ***Modelos de Árvore***

Árvores de Decisão são modelos amplamente utilizados, sendo talvez os mais famosos. Conhecidas pelo seu alto poder preditivo e fácil entendimento, esses algoritmos tornaram-se extremamente relevantes tanto na literatura quanto nas mais variadas aplicações. Inspiradas na estrutura de uma árvore, são modelos baseados em regras sequenciais de fácil interpretabilidade.

Morettin e Singer (2021) relatam que uma Árvore de Decisão é uma espécie de fluxograma no qual as observações percorrem uma série de condições determinadas pelas variáveis do modelo a fim de resultar em uma decisão final. Parte-se de um nó raiz, passando gradualmente pelos nós filhos de tal forma que escolhe-se o atributo mais informativo para realizar a divisão em cada etapa. Ao final do processo, encontra-se uma folha representativa da classe à qual a instância pertence.

Figura 12 – Modelo de Árvore de Decisão para Classificação



Fonte: BRUCE, Peter; BRUCE Andrew. Practical statistics for data scientists**.** p.228, fig. 6-3.

A escolha da variável de maior relevância é feita recursivamente através de medidas de impureza as quais objetivam produzir subconjuntos mais puros. Quanto mais puro um nó, maior seu poder decisório. Nesse âmbito, as medidas mais famosas para analisar impureza são o Índice Gini e a Entropia.

Segundo Géron (2019), o Índice Gini mede a impureza de um nó com base na probabilidade de classificações incorretas ao selecionar dois elementos da amostra e atribuí-los a diferentes classes, portanto, quanto maior seu valor, mais impuro está o nó e, consequentemente, menor seu poder preditivo. Dessa forma, na construção de uma árvore, o algoritmo testará todas as variáveis e escolherá aquela à qual resulta no menor valor de Gini como nó raiz. Sua equação pode ser definida como:

sendo a proporção de observações da classe .

Em relação a Entropia, Géron (2019) expressa que ela mede a pureza dos dados em determinado nó da árvore. Pautando-se na teoria da informação, ela reflete a incerteza em uma distribuição de probabilidade. Dessa forma, na construção de uma árvore, o algoritmo também testará todas as variáveis e escolherá aquela à qual resulta no menor valor de entropia como nó raiz. À medida que a árvore se aprofunda, espera-se que os nós filhos possuam valores de entropia maiores que o nó pai. Sua equação pode ser definida como:

sendo a proporção de observações da classe i.

A principal diferença entre essas medidas está na maneira de avaliação da impureza. Géron (2019) declara que enquanto o Índice Gini concentra-se nas classificações incorretas, a Entropia baseia-se na incerteza da distribuição de probabilidade. Em termos práticos, ambas geralmente levam à mesma decisão, todavia, em certos casos há leves divergências dependendo da problemática.

* + - 1. ***Modelos de Ensemble***

O avanço do poder computacional proporcionou o desenvolvimento de técnicas mais robustas as quais foram responsáveis por grandes resultados, sendo a mais famosa denominada como Ensemble. Métodos Ensemble são projetos para combinarem diversos preditores fracos de forma a criar um preditor forte e robusto, assegura Géron (2019). Há três tipos de métodos de Ensemble: Bagging, Boosting e Stacking. Para este trabalho, abordou-se os dois primeiros.

Figura 13 – Representação de um Ensemble para Classificação



Fonte: GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems.p.148.

Morettin e Singer (2021) alegam que a técnica de Bagging é um método para gerar múltiplas versões de um preditor a partir de vários conjuntos de treinamento a fim de diminuir a variância desse modelo e proporcionar predições mais fidedignas. O principal algoritmo para esta metodologia é a Random Forest, a qual é criada a partir da combinação de diversas Árvores de Decisão. Cada árvore da Random Forest é criada através de amostragens aleatórias e com reposição de linhas e colunas da base de dados de tal forma que todos os elementos possuem a mesma probabilidade de serem selecionadas. Em casos de classificação, a classe final será àquela com maior ocorrência entre os diversos preditores.

Figura 14 – Modelo de Random Forest para Classificação



Fonte: MORETTIN, Pedro A.; SINGER, Julio M. Estatística e Ciência de Dados.p.329.

O Boosting é uma técnica amplamente reconhecida e altamente poderosa no campo de aprendizado de máquina. Ao contrário do Bagging, o Boosting, como explicado por Morettin e Singer (2021), envolve a geração sequencial de árvores de decisão com base na atualização de pesos para cada elemento no conjunto de treinamento. O processo inicia-se com a criação de uma árvore de decisão inicial, seguida pelo cálculo dos resíduos iniciais que representam a diferença entre as probabilidades a priori e a probabilidade a posteriori. Os exemplos classificados incorretamente recebem pesos mais altos, tornando-os mais influentes nas iterações subsequentes. Essas árvores, construídas sequencialmente, visam corrigir as classificações errôneas e aprimorar a precisão do modelo. O Gradient Boosting é o exemplo mais conhecido dessa técnica e é notável por seu uso de métodos de otimização iterativa, conferindo-lhe uma capacidade preditiva fantástica. O objetivo principal do Boosting é reduzir o viés do modelo por meio de correções progressivas dos erros de previsão, como descrito por Géron (2019).

Figura 15 – Modelo de Gradient Boosting para Classificação



Fonte: GÉRON, Aurélien. Hands-on machine learning with scikit-learn, keras and tensorflor: concepts, tools and techniques to build intelligent systems.p.155.

* + 1. *Pré-Processamento*

Sabe-se que, embora os modelos de Aprendizado de Máquina sejam criados com base em técnicas matemáticas e estatísticas, sua implementação final se dá na forma de software. Isso implica que todos os processos subjacentes ao aprendizado, desde a aquisição e o pré-processamento dos dados até a construção e o treinamento do modelo, devem ser traduzidos para linguagem de máquina para que o computador possa executar essas tarefas de forma precisa e eficaz.

No entanto, a realidade dos dados do mundo real é frequentemente complexa e desafiadora. Erros humanos, falhas operacionais e outras fontes de ruído podem introduzir uma alta frequência de dados que não se conformam às expectativas ou são divergentes do comportamento ideal. Nesse contexto, a qualidade dos dados é frequentemente discutível, o que pode prejudicar a eficácia do modelo de Aprendizado de Máquina. Sendo assim, entra em cena o pré-processamento de dados, uma etapa crítica na preparação dos dados para a modelagem.

O principal objetivo do pré-processamento é melhorar a qualidade e a utilidade dos dados, tornando-os mais adequados para o treinamento de modelos de Aprendizado de Máquina. Isso envolve uma série de técnicas e transformações aplicadas aos dados brutos, como limpeza para corrigir erros, tratamento de valores ausentes, normalização e seleção de características relevantes.

O pré-processamento não é apenas uma etapa necessária, mas também estratégica. Uma base de dados de alta qualidade e bem tratada é fundamental para garantir que o modelo aprenda com precisão os padrões presentes nos dados e, por sua vez, faça previsões precisas em novos exemplos. Portanto, a qualidade dos dados é um dos fatores críticos que determinam o sucesso de um projeto de Aprendizado de Máquina. Dessa forma, abordou-se três técnicas de pré-processamento: Binary Encoder, Min-Max Scaler e KNN Imputer.

* + - 1. ***Binary Encoder***

Encoding é uma técnica muito útil quando lida-se com variáveis categóricas. Sua função é a aplicação de um processo de discretização a fim de transformar dados categóricos em dados discretos ou, em outras palavras, transformar classes em números. Em modelos de crédito, ao invés de termos uma variável binária expressa como “Pagou” ou “Não Pagou”, pode-se alterá-la para 1 ou 0, possibilitando assim o entendimento do software por parte do computador.

Nesse âmbito, uma técnica robusta para encoding é o Binary Encoder. Ele é um encoder o qual transforma variáveis categóricas em variáveis discretas binárias de forma inteligente. Ao invés de criar uma coluna para cada categoria, ele cria uma quantidade menor de colunas as quais representam a categoria desejada através de uma tabela-verdade. Esta abordagem é interessante pois ela reduz a dimensionalidade de forma eficiente, possibilitando a criação de uma base de dados menos esparsa, facilitando os cálculos matemáticos do modelo e simplificando o poder computacional demandado.

Exemplificou-se o processo simulando a discretização de uma variável categórica denominada como “Faixa de Idade”. Através do Binary Encoder, pôde-se transformar nove categorias em apenas três colunas, criando uma representação compacta e eficaz das informações categóricas para uso em modelos de Aprendizado de Máquina.

Tabela 1 – Aplicação do Binary Encoder



Fonte: Autoria Própria.

* + - 1. ***Min-Max Scaler***

Durante a etapa de treinamento, determinados modelos entendem a importância das variáveis de forma diferente devido a escalas de magnitude distintas. Variáveis em unidades maiores tendem a influenciarem majoritariamente o modelo, criando assim um algoritmo incompatível com a verdadeira situação.

Técnicas de escalonamento são utilizadas em Aprendizado de Máquina para ajustar as escalas das variáveis e, assim, garantir que todas as variáveis tenham o peso ideal no processo de treinamento do modelo. Dessa forma, pode-se comparar variáveis de forma justa e significativa, inserindo confiabilidade no resultado do modelo.

Uma renomada metodologia de escalonamento é o Min-Max Scaler. Essa técnica é popularmente conhecida por normalização e redimensiona as variáveis de um conjunto de dados para um intervalo específico contido em [0,1]. Sua notoriedade deriva da manutenção da estrutura original dos dados mesmo após o redimensionamento e da preservação de outliers, garantindo o caráter informativo. Sua equação pode ser definida por:

sendo x a variável.

Exemplificou-se o processo de escalonamento através do Min-Max Scaler para as variáveis “idade” e “salário”. Após a aplicação, percebe-se que ambas variáveis possuem a mesma escala.

Tabela 2 – Aplicação do Min-Max Scaler



Fonte: Autoria Própria.

* + - 1. ***KNN Imputer***

Em bases de dados reais é comum encontrar variáveis com valores ausentes. Ignorar essas informações e excluí-las da análise não é uma boa prática visto que resultam na perda de dados importantes, piorando a performance do modelo.

A imputação de valores faltantes tornou-se essencial para modelos de Aprendizado de Máquina justamente para permitir que essas informações sejam levadas em consideração pelo algoritmo. Embora fundamental, a imputação deve ser realizada da forma correta. Preencher os valores de uma variável categórica com sua média não faria sentido, assim como simplesmente preenchê-las com um valor sem pensar nos ganhos e consequências.

Nesse âmbito, uma das técnicas mais consolidadas na literatura denomina-se KNN Imputer. O KNN Imputer é um método de preenchimento de dados faltantes o qual utiliza o modelo KNN por trás. O elemento faltante encontra as *k* observações mais próximas baseando-se nas outras variáveis e então substitui o valor nulo pela média (em caso de variáveis quantitativas) ou moda (em caso de variáveis qualitativas) dessas *k* observações. Essa abordagem é interessante pois traz estimativas mais precisas e interpretáveis.

Geralmente, o primeiro modelo treinado não possui a melhor performance possível. I

* + 1. *Otimização de Modelos*

Geralmente, o primeiro modelo treinado não possui a melhor performance possível. Isso ocorre pois costuma-se decidir o melhor modelo dentre vários testados e, então, otimizá-lo para alcançar resultados ainda melhores. Para alcança-los, há diversos métodos, sendo os principais conhecidos como Engenharia de atributos e otimização de hiperparâmetros.

Engenharia de atributos é o processo de criação de novas variáveis a partir das variáveis iniciais de uma base de dados, bem como a seleção das melhores. Seu objetivo é melhorar a qualidade dos dados e fornecer informações as quais sejam ainda mais relevantes que os dados brutos. Esta abordagem permite que o modelo entenda melhor os padrões e ajuste-se melhor ao desafio proposto, ao mesmo tempo que reduz a dimensionalidade dos dados e a complexidade do modelo, diminuindo a demanda por poder computacional.

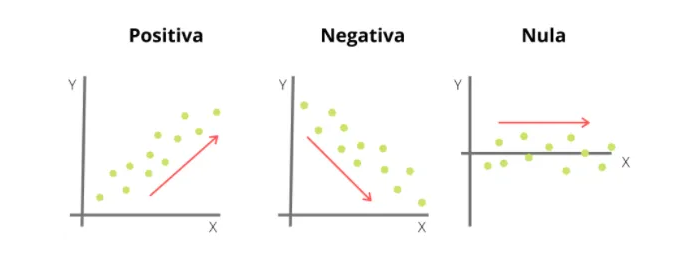
Hiperparâmetros são configurações definidas antes do treinamento de um modelo. Eles representam características construtivas, como o número de vizinhos mais próximos para um KNN, o número de profundidade de uma Árvore de Decisão ou o número de neurônios de uma Rede Neural, por exemplo. A otimização de hiperparâmetros visa achar a melhor configuração possível a fim de aperfeiçoar o modelo, conquistando os melhores resultados possíveis.

Dessa forma, para a engenharia de atributos e otimização de hiperparâmetros utilizou-se três técnicas consolidadas, sendo elas a Análise de Correlação, Variance Threshold, Mutual Information, Feature Importance e Bayes Search.

* + - 1. ***Análise de Correlação***

Para análise de variáveis quantitativas, uma das abordagens mais famosas é a análise de correlações, sendo a mais utilizada a correlação de Pearson. Morettin e Singer (2021) expressam que a correlação de Pearson mede a associação linear entre duas variáveis quantitativas. Contida no intervalo de [-1,1], uma correlação próxima de 0 indica correlação fraca, próxima de 1 correlação forte positiva e próxima de -1 correlação forte negativa.

Figura 16 – Análise de Correlação de Pearson



Fonte: Testes de Correlação. Disponível em: < https://ivanildo-batista13.medium.com/testes-de-correla%C3%A7%C3%A3o-3cb0a37e0f2>

Sua equação pode ser definida como

sendo e os respectivos valores de ambas as variáveis, e , e o desvio padrão das variáveis, e N o número de amostras na base de dados.

Em modelos de Aprendizado de Máquina, a análise de correlação fornece suporte para a engenharia de atributos, pois caso muitas variáveis estejam fortemente correlacionadas entre si, pode-se manter apenas uma delas e assim reduzir a dimensionalidade da base de dados sem perder informações relevantes.

* + - 1. ***Variance Threshold***

Bruce e Bruce (2017) afirmam que a variância de uma amostra é uma medida de variabilidade a qual indica a distância entre os valores da média aritmética ou, em outras palavras, mostra o quão dispersos estão os dados. Sua equação pode ser definida como

sendo o valor da amostra, a média da amostra e N o número de amostras.

Quanto maior a variância, mais variados estão os dados. Modelos de Aprendizado de Máquina beneficiam-se de variáveis com variâncias significativas, pois variáveis constantes simplesmente não agregam informações relevantes para o modelo.

Para exemplificar, pode-se pensar na variável “quantidade de empréstimos tomados”. Se todos os clientes da base de dados não possuírem um empréstimo, pode-se excluí-la, visto que ela não agregará nenhum tipo de informação importante.

Um método muito famoso para esta tarefa é o Variance Threshold, o qual elimina variáveis abaixo de um limiar pré-definido de variância. A ideia é que recursos com baixa variação. Dessa forma, além de reduzir a dimensionalidade da base de dados, o modelo será alimentado apenas com variáveis com possibilidades reais de agregarem positivamente.

* + - 1. ***Mutual Information***

Para a análise variáveis categóricas, um dos métodos mais robustos é o Mutual Information. Esta técnica é uma medida estatística que quantifica a dependência entre duas variáveis aleatórias, sendo um excelente método para entender se a variável resposta possui dependência com a variável de entrada. Basicamente, calcula-se a probabilidade de duas classes ocorrem juntas e a probabilidade de ocorrem separadas, as quais posteriormente são utilizadas para calcular a entropia e finalmente entender o quanto a variável de entrada fornece de informação para predizer a variável resposta. Valores mais altos significam maior dependência entre a variável de entrada e a variável resposta e sua equação pode ser definida como

sendo a probabilidade de ocorrem juntas, p(x) a probabilidade de ocorrer o evento x e p(y) a probabilidade de ocorrer o evento y.

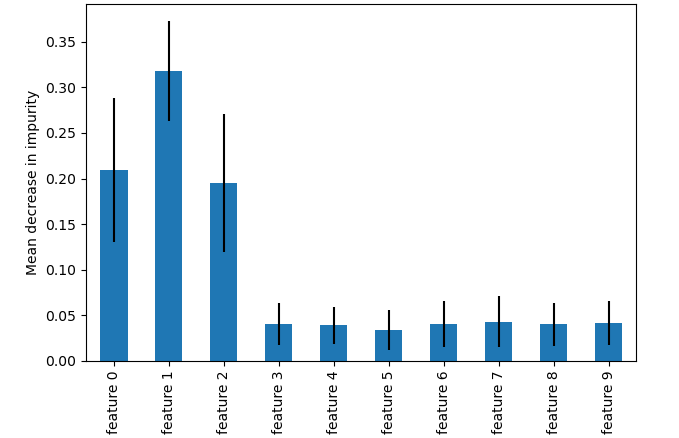
* + - 1. ***Feature Importance***

Modelos de Árvore possuem um processo embutido muito interessante denominado Feature Importance. Essa técnica é uma medida a qual avalia o grau de contribuição de cada variável de entrada no desempenho de um modelo de Aprendizado de Máquina. Ela fornece um ranking o qual indica a relevância de cada característica em relação à variável de resposta do modelo.

A pontuação do ranking geralmente é definida pela função de custo associada ao modelo, sendo as mais famosas o Índice Gini e a Entropia, abordados anteriormente. Resumidamente, o modelo calcula um desses indicadores e entende a proporção de importância daquela variável frente as demais. Esse método é interessante pois pode-se escolher as *N* melhores variáveis do modelo ou, em necessidades mais abrangentes, escolher àquelas as quais possuam importância maior que um limiar pré-definido.

Além de robusta, o Feature Importance possui fácil interpretação, portanto, é extremamente disseminado em casos em que a explicação da decisão do modelo é questionada.

Figura 17 – Feature Importance



Fonte: Feature importances with a forest of trees. Disponível em: < https://scikit-learn.org/stable/auto\_examples/ensemble/plot\_forest\_importances.html>

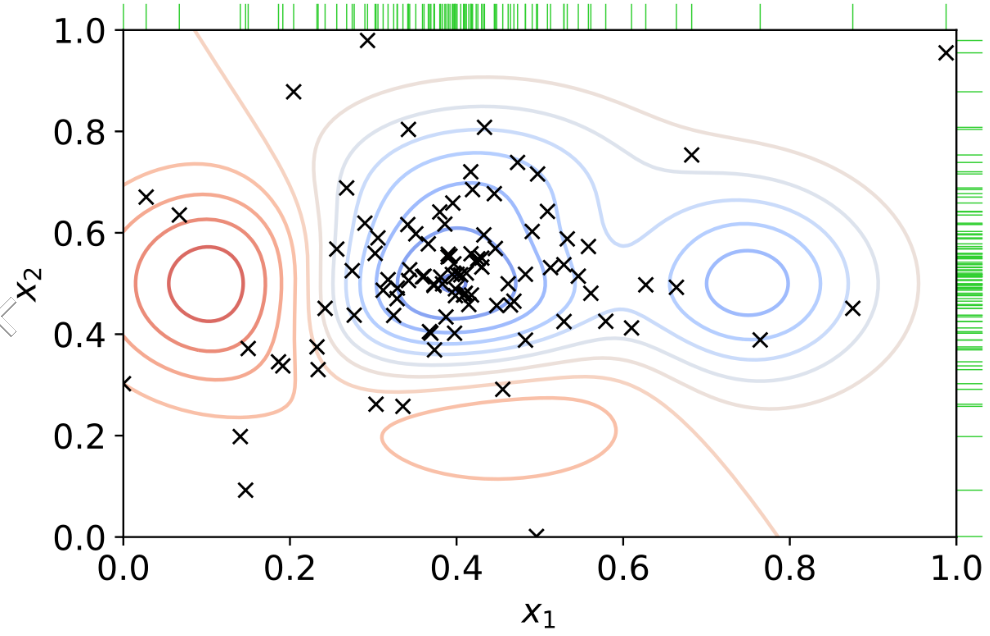
* + - 1. ***Bayes Search***

O Bayes Search é uma técnica de otimização de hiperparâmetros cuja utiliza abordagem bayesiana para encontrar a melhor combinação possível. Resumidamente, a ideia do Bayes Search é construir um modelo probabilístico que relaciona os hiperparâmetros do modelo com a métrica de avaliação escolhida.

Dado um conjunto inicial de hiperparâmetros, testa-se de maneira iterativa a combinação dos valores desse conjunto e avalia-se a métrica de avaliação para cada combinação. O algoritmo possui caráter adaptativo, portanto, a atualização ocorre apenas para hiperparâmetros inseridos nos espaços de busca mais promissoras. A combinação de hiperparâmetros escolhida será aquela a qual apresente a maior probabilidade de retornar as melhores métricas de avaliação ou, em outras palavras, a que proporciona ao modelo a melhor performance dentre todas as combinações testadas.

A principal vantagem desta metodologia é o uso da informação adquirida durante a busca para selecionar novos valores de hiperparâmetros de forma iterativa e adaptiva. Além de eficiente, essa técnica pauta-se em teorias consolidadas e de fácil interpretação, como o Teorema de Bayes citado anteriormente.

Figura 18 – Bayes Search



Fonte: Hyperparameter optimization. Disponível em: < https://en.wikipedia.org/wiki/Hyperparameter\_optimization>

* + 1. *Métricas de Avaliação e Validação*

Como todo modelo de Aprendizado de Máquina está relacionado a probabilidades, inevitavelmente ele comentará erros, portanto, é de suma importância compreender sua performance. Para entender seu desempenho, abordou-se as principais metodologias e métricas de problemas de Classificação para concessão de crédito, sendo elas: acurácia, precisão, sensibilidade, AUC, KS, Hold-Out e Validação Cruzada.

* + - 1. ***Matriz de Confusão***

Conforme Bussab e Morettin (2017), *Inferência Estatística* é o processo de fazer afirmações sobre as características de uma população com base em informações dadas por amostras. Ao submetermos uma amostra a um modelo probabilístico, qualquer que seja a decisão tomada, existem dois tipos de erro: Erro Tipo I e Erro Tipo II. O primeiro ocorre ao rejeitar-se a hipótese nula quando ela é verdadeira e o segundo ocorre ao não rejeitar-se a hipótese nula quando ela é falsa.

Laredo (2010) amplia essa perspectiva ao contextualizar esses conceitos em problemas de concessão de crédito. No cenário financeiro, o Erro Tipo I ocorre quando recusa-se uma operação que seria lucrativa para o credor caso acontecesse e o Erro Tipo II quando aprova-se uma operação a qual dará prejuízo à instituição. Para representar essas situações e quantificar a performance de um modelo de classificação, os estatísticos desenvolveram a Matriz de Confusão.

De acordo com Peter Bruce (2017) e Andrew Bruce (2017), uma Matriz de Confusão é uma matriz quadrada utilizada para comparar os valores preditos do modelo com os valores reais. Sua diagonal é composta pelos acertos do modelo e os demais valores são erros cometidos. Durante a classificação de um elemento, há quatro situações possíveis, sendo elas Verdadeiro Negativo, Verdadeiro Positivo, Falso Negativo e Falso Positivos.

Figura 2 – Matriz de Confusão para Classes Binárias



Fonte: BRUCE, Peter; BRUCE Andrew. **Practical statistics for data scientists.** p.201, fig. 5-5

Um Verdadeiro Negativo (VN) refere-se aos casos em que o modelo previu corretamente uma instância como pertencente à classe negativa e de fato ela pertence à classe negativa, ou seja, o cliente é adimplente e o modelo afirma que ele pagará o empréstimo.

Um Verdadeiro Positivo (VP) refere-se aos casos em que o modelo previu corretamente uma instância como pertencente à classe positiva e de fato ela pertence à classe positiva, ou seja, o cliente é inadimplente e o modelo afirma que ele não pagará o empréstimo.

Um Falso Negativo (FN) refere-se aos casos em que o modelo previu incorretamente uma instância como pertencente à classe negativa mas ela pertence à classe positiva, ou seja, o cliente é inadimplente e o modelo afirma que ele pagará o empréstimo.

Um Falso Positivo (FP) refere-se aos casos em que o modelo previu incorretamente uma instância como pertencente à classe positiva mas ela pertence à classe negativa, ou seja, o cliente é adimplente e o modelo afirma que ele pagará o empréstimo.

* + - 1. ***Acurácia***

Conforme Peter Bruce(2017) e Andrew Bruce (2017), a acurácia é uma métrica simples a qual quantifica a proporção de previsões corretas feita pelo modelo em relação ao total de previsões. Embora indique a performance geral do modelo, em problemas de classes desbalanceadas ela não performa bem, pois pode ser enganosa visto que nesses casos uma classe é muito mais comum que a outra.

* + - 1. ***Precisão***

Conforme Peter Bruce (2017) e Andrew Bruce (2017), a precisão quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas pelo modelo. Em suma, ela representa a capacidade de um modelo em prever corretamente a classe positiva, portanto, bons valores de precisão significam a diminuição do Erro Tipo I.

* + - 1. ***Sensibilidade***

Conforme Peter Bruce (2017) e Andrew Bruce (2017), a sensibilidade quantifica a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias que realmente são positivas na amostra. Em suma, ela representa a capacidade de um modelo em capturar a classe positiva, portanto, bons valores de sensibilidade significam a diminuição do Erro Tipo II.

* + - 1. ***F1-Score***

Conforme Peter Bruce (2017) e Andrew Bruce (2017), o f1-score representa a média harmônica entre a precisão e a sensibilidade. Ele é uma métrica valiosa para problemáticas de dados desbalanceados, ou seja, quando há muito mais amostras de determinada classe em relação a outra. Em suma, ela representa o equilíbrio entre a capacidade do modelo de prever a classe positiva e de capturar a classe positiva, portanto, bons valores de f1-score significam um bom desempenho do modelo de classificação.

* + - 1. ***Curva ROC e AUC***

Conforme Peter Bruce (2017) e Andrew Bruce (2017), nota-se que há uma troca entre precisão e sensibilidade, ou seja, inferir a classe positiva impõe que mais amostras da classe negativa serão classificadas como positivas e vice-versa. Dada a situação, o classificador ideal atenderia a classificação de ambas as classes corretamente. A métrica responsável por avaliar essa troca denomina-se “Curva de Característica Operatória Receptora” (Curva ROC).

A curva ROC é uma representação gráfica da taxa de VP em função da taxa de FP para diferentes pontos de corte em um modelo de classificação.

A partir dessa curva criada, pode-se calcular a área sob a curva (AUC), a qual é uma métrica contida no intervalo [0,1], onde um valor maior indica melhor desempenho do modelo. Esta abordagem é muito útil em problemas de classificação binária com classes desbalanceadas.

Figura 3 – Curva ROC e AUC



Fonte: BRUCE, Peter; BRUCE Andrew. Practical statistics for data scientists**.** p.204, fig. 5-6; p.206, fig. 5-7

* + - 1. ***KS***

Segundo Laredo (2010), outra métrica muito famosa para modelos de crédito é o teste de hipótese de Kolmogorov-Smirnov (KS). Este teste estatístico é empregado para avaliar a capacidade discriminativa de um modelo, medindo a diferença acumulada entre as distribuições de probabilidade das classes de bons e maus clientes.

O valor do KS é calculado como a maior distância entre as curvas de distribuição de probabilidade acumulada, podendo variar no intervalo [0, 1]. Quanto mais próximo de 1, mais evidente a separação entre as duas classes, indicando melhor poder de discriminação do modelo.

Figura 4 – Kolgomorov-Smirnov (KS)



Fonte: SAS: Calculating KS Statistics, Listen Data. https://www.listendata.com/2016/01/sas-calculating-ks-test.html

* + - 1. ***Holdout e Validação Cruzada***

Ao desenvolver-se modelos de Aprendizado de Máquina, deve-se garantir que eles possuam boa capacidade de generalização para dados não vistos. Nesse âmbito, para simular seu desempenho em dados novos, há duas técnicas fundamentais: Holdout e Cross Validation.

Holdout é um método que consiste em separar, de forma aleatória, uma parcela dos dados para treinamento e o restante para teste. Os dados de treinamento servem para que o modelo monte uma equação matemática capaz de ajustar-se aos dados e entender os padrões contidos na amostra, ao passo que os dados de teste simulam dados novos aos quais serão submetidos ao modelo.

Como baseia-se na aleatoriedade, esta técnica está sujeita ao viés de seleção de amostra, logo, o modelo pode capturar padrões existentes apenas na amostra treino e performar de maneira não satisfatória nos dados de teste, resultando em uma alta variância.

Figura 5 – Holdout



Fonte: Métodos de Reamostragem, Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná. http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html

No Cross Validação, a amostra é dividida aleatoriamente e separada em K grupos. De forma repetitiva, o mesmo modelo é treinado *k* vezes, sendo *K-1* grupos utilizados para treino e um grupo para teste.

Iterativamente, a amostra de teste muda e o resultado final é uma média aritmética das métricas de todos os treinamentos, portanto, ao final do processo, o modelo consegue reduzir bastante a variância e fornecer um resultado mais robusto e fidedigno da performance do modelo.

Figura 6 – Cross Validation



Fonte: Métodos de Reamostragem, Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná. http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html

* + - 1. ***Overfitting e Underfitting***

Após a obtenção das métricas de treino e teste, pode-se compará-las a fim de analisar se o modelo possui generalização. Conforme Géron (2019), um modelo possui boa generalização quando as métricas de treino e teste são satisfatórias e possuem valores relativamente próximos. Embora o conceito seja simples, nem sempre os modelos alcançam bons resultados e há disparidade entre os cenários de treino e teste.

Denomina-se Overfitting o cenário o qual o modelo possui bons resultados em dados de treino, mas performa abaixo do esperado em dados de teste. Em suma, isso significa que o modelo foi capaz de criar uma equação matemática tão complexa ao ponto de assimilar até mesmo os ruídos da amostra de treino.

O cenário de Underfitting é o inverso, logo, o modelo possui maus resultados tanto em dados de treino quanto de teste. Em suma, o modelo não conseguiu de criar uma equação calibrada a qual fosse representativa e capaz de captar os padrões na amostra de treino.

Em casos de Overfitting, Géron (2019) afirma que a variância é alta, visto que esse algoritmo está extremamente sujeito aos dados de entrada; em casos de Underfitting, o viés é alto, pois o modelo não consegue detectar o padrão de nossos dados. Existem diversas técnicas para tratar ambos os casos, entretanto, espera-se que o modelo tenha uma troca justa entre ambos e consiga performar bem de forma generalizada.

Figura 7 – Trade-Off Viés x Variância



Fonte: Métodos de Reamostragem, Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná. http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html

**3) MATERIAIS E MÉTODOS**

**3.1 Materiais**

*3.1.1* *Python*

Conforme Alexandre Levada (2021), a necessidade de lidar com grandes volumes de dados pressupõe, de maneira mandatória, a utilização de linguagens de programação. Como há muitas opções, deve-se escolher uma que atenda melhor as necessidades do problema. Pela versatilidade, facilidade e extensa documentação, optou-se pelo Python, visto que ela é a linguagem de programação mais utilizada mundialmente.

Criada em 1991 como uma linguagem de alto nível e orientada a objetos, ela prova diariamente sua robustez e importância. Combinando uma sintaxe bem definida e legível, ela possui inúmeros Frameworks de matemática, estatística e computação científica, portanto, fornece base completa para trabalhar com aprendizado de máquina.

Além disso, ela é utilizada em muitas outras áreas da computação, facilitando a integração de softwares e profissionalizando todo o ecossistema. Por fim, ela conta com a maior comunidade do mundo, logo, correção de bugs e inconsistências tornam-se naturalmente mais rápidas. A seguir encontram-se as bibliotecas utilizadas neste trabalho:

***Pandas –*** Pandas é a biblioteca mais famosa em Ciência de Dados e a principal para manipulação de dados estruturados. Através desse Framework, pode-se realizar consultas, limpezas, tratamentos, análises e as mais variadas transformações possíveis em tabelas. Por pautar-se nas estruturas de dados Series e Dataframes, ela possui performance muito superior aos processos tradicionais de computação como laços de repetição e condicionais.

***Numpy e Scipy –*** Numpy e Scipy destinam-se para a parte de modelagem matemática e computação científica, comportando desde temas simples como Trigonometria e Funções Logarítmicas, até temas avançados como Álgebra Linear, Cálculo Diferencial e Integral, Cálculo Multivariado, Cálculo Numérico, entre tantos outros.

***Matplotlib e Seaborn –*** Matplotlib e a Seaborn voltam-se para visualização de dados. A partir delas, pode-se criar inúmeros tipos de gráficos e personalizá-los da maneira que o usuário deseja, contribuindo positivamente com a apresentação de resultados e Storytelling.

***Stats e Statsmodels –*** Stats e Statsmodels abordam a parte de modelagem estatística, comportando tópicos de Estatística Descritiva, Teoria das Probabilidades, Estatística Inferencial, Regressões e Séries Temporais.

***Scikit Learn –*** Ao lado do Pandas e Numpy, o Scikit Learn completa a “Tríade de Principais Frameworks para a Ciência de Dados”. Ela é responsável por disponibilizar praticamente todos os tópicos de Aprendizado de Máquina existentes em Python. Modelos de Classificação, Regressão, Clusterização, Redução de Dimensionalidade, Métricas de Avaliação, Pré-Processamento de Dados e Transformações Matemáticas são apenas alguns exemplos desta ferramenta incrível.

*3.1.2* *PySpark*

Apache Spark é um Framework construído em 2009 para trabalhar com grandes volumes de dados. Ele é um mecanismo para aplicações de processamento distribuído o qual ocorre na memória e, por esse motivo, trabalha de forma eficiente e veloz (em torno de 100 vezes mais rápido que outros sistemas).

A execução do Spark consiste em dividir os dados em diversas partes iguais, com cada parte sendo executada em um Cluster diferente de forma paralela. Após todo processamento, os Clusters juntam os dados novamente e o resultado é exposto.

PySpark é uma interface para Apache Spark em Python, possibilitando o uso de APIs Python com as qualidades que o Spark oferece. Muito robusto e eficiente, ele divide-se em duas estruturas de dados principais:

***RDD’s –*** RDD’s consistem em coleções de dados distribuídos que são inalteráveis após a criação. Cada conjunto de dados em um RDD é dividido em partições lógicas e comporta dados estruturados e não estruturados.

***Dataframes –*** Dataframes consistem em conjuntos de dados organizado em colunas, sendo equivalentes a tabelas de bancos de dados. Comportam dados estruturados e semiestruturados.

Por ser extremamente potente, o uso do PySpark neste trabalho destinou-se para a transformações mais exigentes, como aplicação de agrupamentos ou funções de janela. Combinando com o Pandas, os resultados costumam ser muito positivos. A seguir apresentam as principais bibliotecas de PySpark utilizadas neste trabalho:

***Spark Session*** – A biblioteca Spark Session é responsável por iniciar a Sessão Spark, permitindo o uso do Framework sem limitações.

***SQL Functions –*** Esta é a principal biblioteca do PySpark. Através dela, pode-se programar com sintaxe e funções muito parecidas com as do Python, mas com toda eficiência que o Spark proporciona. Além disso, é possível também utilizar comandos em SQL (linguagem de consulta a bancos de dados), os quais são extremamente úteis em inúmeras situações.

*3.1.3* *GIT*

Trabalhos que envolvem software sempre estão sujeitos a mudanças, visto que os produtos e serviços são muito dinâmicos e a sociedade está em constante evolução, portanto, é muito comum que o código-fonte sofra alterações com o passar do tempo.

Nesse âmbito, uma técnica muito importante são os sistemas de controle de versão, os quais são responsáveis pelo registro de todas as modificações de código. A partir deles, pode-se criar versões de um software, possibilitando flutuar em estágios antigos e novos sem comprometer o projeto.

O GIT é o sistema de controle de versão mais famoso mundialmente. Baseando-se em repositórios, ele permite que as versões de um software fiquem salvas. Além disso, ele fornece diversos ganhos como proteção criptográfica, alto desempenho e a possibilidade dos desenvolvedores programarem ao mesmo tempo, tornando-se assim fundamental para o cotidiano de um profissional de tecnologia.

Para este trabalho, utilizou-se o GIT para salvar as versões do código-fonte no servidor do GitHub (serviço gratuito para gerenciar repositórios) através de um repositório de código aberto a fim de incentivar a colaboração no meio acadêmico e permitindo que o público interessado possa consultar o projeto.

*3.1.4* *Kaggle*

A plataforma Kaggle é a maior comunidade de cientistas, engenheiros e analistas de dados do mundo. Nela, encontram-se inúmeras bases de dados fictícias e reais destinadas a profissionais que desejam treinar suas habilidades técnicas e analíticas. Ela também é responsável por realizar competições, as quais são desafios impostos por empresas do mundo todo com uma premiação em dinheiro para quem solucionar o problema.

Pelo fato de risco de crédito envolver diversas informações sensíveis, utilizar uma base de dados verdadeira é praticamente impossível sem passar pelos processos burocráticos de confiabilidade, dessa forma, optou-se por uma base de dados de competição do Kaggle, simulando perfeitamente um ambiente real de uma instituição financeira.

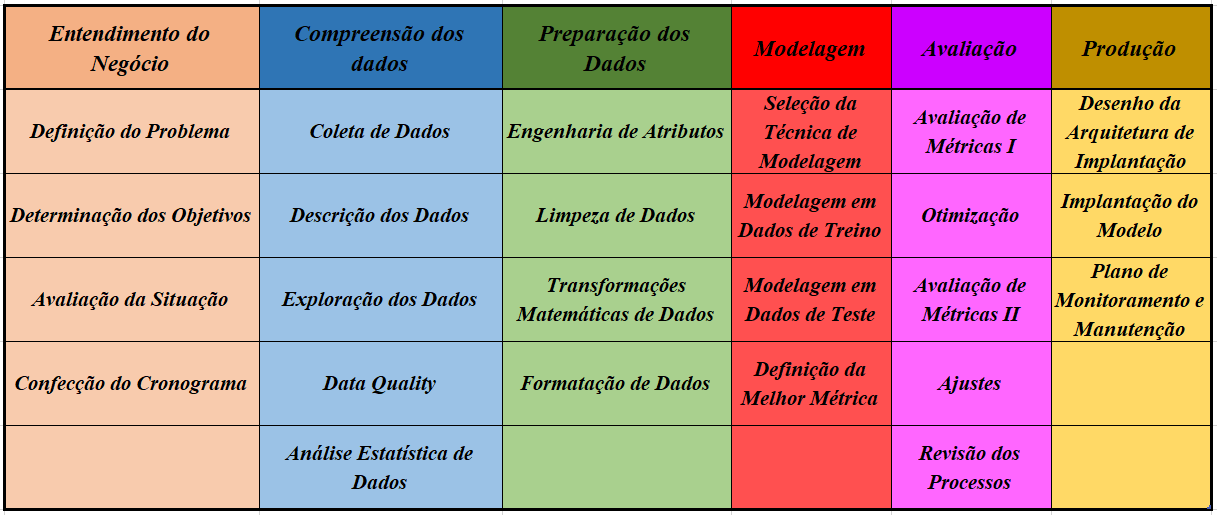
**3.2 Métodos**

3.2.1 Metodologia CRISP-DM

De acordo com Rafaela Lima (2021), ao longo dos anos, a capacidade computacional aumentou exponencialmente, portanto, uma quantidade astronômica de dados passou a ser gerada diariamente. Para utilizar todo o potencial de seus dados, demandou-se a criação de uma metodologia para projetos de Ciência de Dados, assim dando início ao CRISP-DM.

CRISP-DM é um processo de Mineração de Dados criado em 1996. Sua principal função é dividir projetos complexos de Ciência de Dados em partes menores, facilitando a execução das tarefas e o entendimento dos pontos por parte de pessoas não-técnicas. Pautando-se na constante evolução, essa metodologia permite que todas as etapas sejam revisitadas ao longo do projeto, corrigindo falhas e agilizando as entregas. A seguir há um esboço das etapas:

Tabela 2 – Descrição detalhada das Etapas do CRISP-DM

****

Fonte: Autoria Própria.

**Entendimento do Negócio** - Inicialmente, leva-se em consideração todo o contexto do negócio e da empresa e define-se o objetivo do projeto, identificando as necessidades e alinhando as expectativas. Esta provavelmente é a parte mais importante do CRISP-DM, pois um problema mal definido condena as demais etapas a caminhos distintos do esperado.

**Compreensão dos Dados** – Posteriormente, realiza-se uma análise exploratória dos dados a fim de compreender as informações presentes em cada uma das variáveis, bem como suas características estatísticas a fim de inferir seu comportamento. Além disso, nesta etapa também são aplicados testes de hipótese para levantar possíveis insights visando enriquecer a modelagem.

**Preparação dos Dados** – Esta etapa é responsável pelo tratamento e limpeza de dados, combinação de variáveis a fim de criar novos atributos e transformação dos dados no formato correto através de algumas transformações matemáticas e estatísticas para auxiliar a modelagem. Caso os dados estejam em formatos diferentes, alguns modelos de aprendizado de máquina funcionarão da forma errada, prejudicando o reconhecimento de padrões esperado.

**Modelagem** – A parte da modelagem consiste na aplicação dos métodos quantitativos definidos para solucionar o problema em questão. Durante o processo, os melhores modelos são testados e otimizados para que consigam realizar a tarefa corretamente. A má escolha de um modelo pode comprometer o projeto, portanto, esta etapa é fundamental.

**Avaliação** – Nesta fase avalia-se a performance do modelo escolhido. Inicialmente define-se as métricas mais importantes para o problema em questão e, caso apresente bons resultados durante a avaliação, constata-se que o modelo está apto para ser levado adiante.

**Produção** – Finalmente, após todos os testes necessários, é hora de implantar oficialmente o modelo de aprendizado de máquina. Nesta etapa, necessita-se do suporte da área de Engenharia de Machine Learning, Engenharia de Dados, Engenharia de Software e Arquitetos de Cloud, pois além de criar toda uma infraestrutura robusta para que o modelo rode, demanda-se metodologias de monitoramento e novos treinamentos a fim de garantir que o modelo esteja sempre calibrado e apto para realizar suas tarefas com a melhor performance possível.

Como problemas de análise de crédito tendem a ser complexos, extensos e muito importantes para uma instituição, aplicou-se a metodologia CRISP-DM como direcionamento para o problema de concessão de crédito proposto a fim de facilitar o andamento do projeto como um todo.

**4) Metodologia**

**4.1 Entendimento da Base de Dados**

*4.1.1* *Descrição do Problema*

*4.1.2* *Significado das Variáveis*

*4.1.3* *Separação em Treino e Teste*

**4.2 Análise Exploratória**

*4.2.1* *Análise Univariada*

*4.2.2* *Análise Bivariada*

**4.3 Engenharia de Atributos**

*4.3.1* *Features A*

*4.3.2* *Features B*

*4.3.3* *Features C*

*4.3.4* *Features D*

**4.4 Pré-Processamento de Dados**

*4.4.1* *Valores Ausentes*

*4.4.2* *Encoding*

*4.4.3* *Escalonamento*

**4.5 Aplicação de Modelos de Aprendizado de Máquina**

*4.5.1* *Motor de Modelos*

*4.5.2* *Avaliação de Métricas*

*4.5.3* *Seleção de Features*

*4.5.4* *Otimização de Hiperparâmetros*

*4.5.5* *Retenção do Modelo Treinado*

**5) RESULTADOS E DISCUSSÕES**

**5.1 Aplicação do Modelo Treinado em Dados Novos**

**5.2 Tradução e Quantificação dos Resultados**

**5.3 Estudo e Aplicação de Políticas de Crédito**

**5.4 Comparação com a Metodologia Tradicional**

**6) CONCLUSÃO**

**6.1 Resumo dos Resultados Obtidos**

**6.2 Resposta da questão de pesquisa**

**6.3 Limitações do Estudo**

**6.4 Sugestões para trabalhos futuros e próximos passos**

**REFERÊNCIAs**

[1] LAREDO SICSÚ, Abraham. **CREDIT SCORING: DESENVOLVIMENTO, IMPLANTAÇÃO E ACOMPANHAMENTO.** São Paulo: Blucher, 2010. Disponível em: <https://www.blucher.com.br/credit-scoring\_9788521205333>. Acesso em: 22 janeiro 2023.

[2] SEBBEN, Renivaldo José. **ANÁLISE DE RISCO DE CRÉDITO E COBRANÇA: COMO CONCEDER CRÉDITO COM SEGURANÇA E RECUPERAR CRÉDITOS INADIMPLENTES.** Novatec Editora Ltda, 2020. Disponível em: <https://www.amazon.com.br/gp/product/8575228269/ref=as\_li\_tl?ie=UTF8&camp=1789&creative=9325&creativeASIN=8575228269&linkCode=as2&tag=novatec03-20 >. Acesso em 12 de março 2023.

[3] GUIMARÃES XAVIER, Caroline. **RISCO NA ANÁLISE DE CRÉDITO**. Tese (Bacharel em Ciências Contábeis) – Departamento de Ciências Contábeis, Universidade Federal de Santa Catarina. Florianópolis, p.70, 2011. Acesso em: 15 fevereiro 2023.

[4] JORGE CHAIA, Alexandre. **MODELOS DE GESTÃO DO RISCO DE CRÉDITO E SUA APLICABILIDADE AO MERCADO BRASILEIRO.** Tese (Mestrado em Administração) – Departamento de Administração, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo. São Paulo, p.126, 2003. Acesso em: 15 fevereiro 2023.

[5] ARAÚJO, Elaine Aparecida; MONTREUIL CARMONA, Charles Ulises de. **DESENVOLVIMENTO DE MODELOS CREDIT SCORING COM ABORDAGEM DE REGRESSÃO LOGÍSTICA PARA A GESTÃO DA INADIMPLÊNCIA DE UMA INSTITUIÇÃO DE MICROCRÉDITO.** Contabilidade Vista & Revista, Minas Gerais, vol. 18, n. 3, p. 107 – 131, set.2007. Acesso em: 20 fevereiro 2023.

[6] SHELCI SILVA, Juelline. **GERENCIAMENTO INTEGRADO DE RISCOS: MODELOS DE PREDIÇÃO DE RISCO DE CRÉDITO EM MACHINE LEARNING PARA A IDENTIFICAÇÃO DE ATIVOS PROBLEMÁTICOS EM UMA INSTITUIÇÃO FINANCEIRA.** Tese (Mestrado Profissional em Economia) – Departamento de Economia, Faculdade de Administração Contabilidade e Economia, Universidade de Brasília. Brasília, p.74, 2022. Acesso em: 20 fevereiro 2023.

[7] FORTI, Melissa. **TÉCNICAS DE MACHINE LEARNING APLICADAS NA RECUPERAÇÃO DE CRÉDITO DO MERCADO BRASILEIRO**. Tese (Mestrado em Economia) – Escola de Economia de São Paulo, Fundação Getulio Vargas. São Paulo, 2018. Acesso em: 21 fevereiro 2023.

[8] SANTOS, Patrick Ferreira dos. **USO DE TÉCNICAS DE MACHINE LEARNING PARA ANÁLISE DE RISCO DE CRÉDITO.** Tese (Mestrado Profissional em Economia) – Departamento de Economia, Faculdade de Administração Contabilidade e Economia, Universidade de Brasília. Brasília, p.57, 2022. Acesso em: 19 fevereiro 2023.

[9] ARAÚJO, João Paulo Bezerra de. **INTERPRETABILIDADE DE MODELOS DE MACHINE LEARNING: APLICAÇÃO NO MERCADO DE CRÉDITO.** Tese (Bacharel em Engenharia Elétrica) – Universidade Federal do Ceará. Fortaleza, p.73, 2020. Acesso em: 19 de fevereiro 2023.

[10] MONTOYA, Anna; ODINTSOV, Kirill; KOTEK, Martin. **HOME CREDIT DEFAULT RISK.** Kaggle Competition. Disponível em: < https://www.kaggle.com/competitions/home-credit-default-risk/overview>. Acesso em: 10 janeiro 2023.

[11] MORETTIN, Pedro A.; SINGER, Julio M. **ESTATÍSTICA E CIÊNCIA DE DADOS**. LTC, 2022. Disponível em: < https://www.grupogen.com.br/livro-estatistica-e-ciencia-de-dados-pedro-alberto-morettin-e-julio-da-motta-singer-editora-ltc-9788521638162>. Acesso em: 23 fevereiro 2023.

[12] GÉRON, Aurélien. **HANDS-ON MACHINE LEARNING WITH SCIKIT-LEARN, KERAS AND TENSORFLOR: CONCEPTS, TOOLS AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS**. O’Reilly, 2019. Disponível em: < https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ >. Acesso em: 25 Janeiro 2023.

[13] HARRISON, Matt. **MACHINE LEARNING POCKET REFERENCE.** O’Really, 2019. Disponível em: < https://www.oreilly.com/library/view/machine-learning-pocket/9781492047537/ >. Acesso em: 27 Janeiro 2023.

[14] MORETTIN, Pedro A.; BUSSAB, Wilton De O. **ESTATÍSTICA BÁSICA**. São Paulo: Saraiva, 2017. Disponível em: < https://www.saraiva.com.br/estatistica-basica---morettin---saraiva-21397/p >. Acesso em: 10 janeiro 2023.

[15] CAMARGO, Bruna Emy. **NÚMERO DE INADIMPLENTES VOLTA A CRESCER E CHEGA A 65 MILHÕES DE PESSOAS EM JANEIRO**. Estadão.com.br, São Paulo, 16 de fevereiro de 2023. Disponível em: < https://www.estadao.com.br/economia/numero-inadimplentes-cresce-65-milhoes-pessoas-janeiro/#:~:text=Quatro%20em%20cada%20dez%20brasileiros,m%C3%AAs%20do%20ano%2C%20segundo%20pesquisa&text=O%20n%C3%BAmero%20de%20inadimplentes%20no,rela%C3%A7%C3%A3o%20a%20dezembro%20de%202022. >. Acesso em: 19 de março de 2023.

[16] **MAPA DA INADEIMPLÊNCIA E NEGOCIAÇÃO DE DÍVIDAS NO BRASIL**. Serasa, São Paulo, janeiro de 2023. Disponível em: <https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/>. Acesso em: 19 de março de 2023.

[17] LIMA, Rafaela Somavila. **CRIAÇÃO DE PROJETO DE CIÊNCIA DE DADOS UTILIZANDO A METODOLOGIA CRISP-DM EM CONFORMIDADE COM A LGPD**. Tese (Especialização em Ciência de Dados e Suas Aplicações) – Departamento Acadêmico de Informática, Universidade Tecnológica Federal do Paraná. p.35, 2021. Acesso em: 21 de março de 2023.

[18] LEVADA, Alexandre Luis Magalhães. **PROGRAMAÇÃO CIENTÍFICA COM PYTHON.** Departamento de Computação, Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos. p.107, 2021. Acesso em: 22 de março de 2023.

[19] TCHILIAN, Felipe. **CICLO DE CRÉDITO: ENTENDA E OTIMIZE A JORNADA DO CLIENTE.** ClearSale, 2022. Disponível em: <https://blogbr.clear.sale/ciclo-de-credito>. Acesso em 22 de março de 2023.

[20] BRUCE, Peter; BRUCE Andrew. **PRACTICAL STATISTICS FOR DATA SCIENTISTS.** O’Really, 2017. Disponível em: < https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/>. Acesso em: 27 de abril de 2023.

[21] **MÉTODOS DE REAMOSTRAGEM.** Laboratório de Estatística e Geoinformação, Universidade Federal do Paraná. Disponível em: <http://cursos.leg.ufpr.br/ML4all/apoio/reamostragem.html>. Acesso em: 22 de agosto de 2023.

[22] PARK, Sung. **UNDERSTAND AND USE A BUSINESS CREDIT RISK SCORE.** Experian, 2020. Disponível em: <https://blogbr.clear.sale/ciclo-de-credito>. Acesso em 15 de julho de 2023.

[23] Bhalla, Deepanshu. **SAS: CALCULATING KS STATISTICS.** Listen DATA, 2016. Disponível em: < https://www.listendata.com/2016/01/sas-calculating-ks-test.html>. Acesso em 22 de julho de 2023.

[24] Kumar, Ajitesh. **MACHINE LEARNING: INFERENCE & PREDICTION DIFFERENCE.** Data Analytics, 2022. Disponível em: < https://vitalflux.com/machine-learning-inference-prediction-difference/#:~:text=Prediction%20is%20the%20process%20of,the%20predictor%20and%20response%20variables.>. Acesso em 23 de agosto de 2023.

[25] PEREIRA, Pedro Miguel Pinhal. **ANÁLISE DE RISCO DE CRÉDITO USANDO ALGORITMOS DE MACHINE LEARNING.** Tese (Mestrado em Matemática Financeira) – Departamento de Matemática, Universidade de Lisboa. 2020. Acesso em: 23 de agosto 2023.

[26] ABREU, Mariana da Conceição Ferreira. **MODELOS DE AVALIAÇÃO DE RISCO DE CRÉDITO.** Tese (Mestrado em Economia na especialização de Economia Financeira) –Universidade de Coimbra. 2020. Acesso em: 23 de setembro 2023.

[27] SILVA, Daniel de Oliveira Silva. **OTIMIZAÇÃO DE HIPERPARÂMETROS DE ALGORITMOS DE MACHINE LEARNING APLICADO NO CONTEXTO DE ANÁLISE DE RISCO DE CRÉDITO.** Trabalho (Especialização em Ciência de Dados) –Universidade Tecnológica Federal do Paraná. 2022. Acesso em: 14 de setembro 2023.

[28] OLIVEIRA LIMA, Jorge Cláudio Cavalcante de. **A IMPORTÂNCIA DE CONHECER A PERDA ESPERADA PARA FINS DE GERENCIAMENTO DO RISCO DE CRÉDITO.** Revista do BNDES, Rio de Janeiro, V.15, N.30, P.271-302, 2008. Acesso em: 27 de agosto 2023.

[29] SELAU, Lisiane Priscila Roldão. **MODELAGEM PARA CONCESSÃO DE CRÉDITO A PESSOAS FÍSICAS EM EMPRESAS COMERCIAIS: DA DECISÃO BINÁRIA PARA A DECISÃO MONETÁRIA.** Tese (Doutorado em Administração) – Universidade Federal do Rio Grande do Sul. 2012. Acesso em: 30 de agosto 2023.

[30] BORIN, Edson. **CAPACITAÇÃO PROFISSIONAL EM TECNOLOGIAS DE INTELIGÊNCIA ARTIFICIAL.** Instituto de Computação, Universidade Estadual de Campinas. 2023. Acesso em: 30 de agosto 2023.

**ANEXO**

**Código Fonte do Software Desenvolvido**