Universidade Federal do Rio de Janeiro

Escola Politécnica

MBA em Big Data, Business Intelligence e Business Analytics (MB3B)

MACHINE LEARNING APLICADO A RISCO DE CRÉDITO

Autor:

Jonathan das Neves Braz

Orientador:

Manoel Villas Boas Junior, M. Sc.

Coorientador:

Edilberto Strauss, Ph. D.

Examinador:

José Airton Chaves Cavalcante Junior, D. Sc.

Examinador:

Vinicius Drumond Gonzaga, M. Sc.

Examinador:

Flávio Luis Mello, D. Sc.

Rio de Janeiro
Dezembro de 2021

Declaração de Autoria e de Direitos

Eu, **Jonathan das Neves Braz** CPF 112.172.587-28, autor da monografia ***MACHINE LEARNING APLICADO A RISCO DE CRÉDITO***, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na defesa da monografia do curso de Pós-Graduação, Especialização MBA em Big Data, Business Intelligence e Business Analytics da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.

2. Excetuam-se do item 1 eventuais transcrições de texto, figuras, tabelas, conceitos e idéias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.

3. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.

4. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.

5. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.

6. Por ser verdade, firmo a presente declaração.

Rio de Janeiro, _____ de _____ de _____.

_____

Nome Completo

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Av. Athos da Silveira, 149 - Centro de Tecnologia, Bloco H, sala - 212,
Cidade Universitária Rio de Janeiro – RJ - CEP 21949-900.

AGRADECIMENTO

Dedico este trabalho à minha família que contribuiu de forma significativa à minha formação. Este projeto é uma pequena forma de retribuir o investimento e confiança em mim depositados.

RESUMO

Tendo em vista que existem milhões de consumidores inadimplentes no Brasil de acordo com Serasa e as crises financeiras são de longe as situações mais estressantes, a tendência de alta da inadimplência neste ano está no radar dos grandes bancos. Por isso, os bancos podem sofrer perdas financeiras quando o cliente tem uma dívida pendente dificultando assim a instituição financeira de recuperar os pagamentos pelos serviços prestados. O presente trabalho teve como objetivo desenvolver e comparar modelos preditivos para detecção de inadimplência empregando diversos algoritmos de Machine Learning. Para tanto, é necessário explicar o que é Machine Learning, realizar a seleção dos algoritmos e efetuar a predição dos dados. A metodologia adotada baseia-se em pesquisas bibliográficas onde foi realizado a revisão da literatura sobre Aprendizado de Máquina. Diante disso verifica-se que foi obtido um classificador de alto desempenho com AUC de 79% nos testes, o que impõe a constatação de que o melhor algoritmo utilizado nos experimentos é a Floresta Aleatória.

Palavras-Chave: Machine Learning, AUC, Inadimplência.

ABSTRACT

Considering that there are millions of defaulting consumers in Brazil, according to Serasa, and financial crises are by far the most stressful situations, the upward trend in defaults this year is on the radar of large banks. Therefore, banks can suffer financial losses when the customer has an outstanding debt, thus making it difficult for the financial institution to recover payments for the services provided. The present work aimed to develop and compare predictive models for default detection using several Machine Learning algorithms. Therefore, it is necessary to explain what Machine Learning is, perform the selection of algorithms and perform data prediction. The adopted methodology is based on bibliographical research where a literature review on Machine Learning was carried out. Therefore, it is verified that a high performance classifier was obtained with AUC of 79% in the tests, which imposes the observation that the best algorithm used in the experiments is the Random Forest.

Keywords: Machine Learning, AUC, Default.

SIGLAS

AM	Aprendizado de Máquina
AUC	Área sob a curva
IA	Inteligência Artificial
SVM	*Support Vector Machine*
ROC	*Receiver Operating Characteristic*
VP	Verdadeiro Positivo
FP	Falso Positivo
VN	Verdadeiro Negativo
FN	Falso Negativo
CPU	*Central Processing Unit*
RAM	*Random Access Memory*
HD	*Hard Disk*

LISTA DE FIGURAS

LISTA DE QUADROS

Sumário

Capítulo 1

Introdução

1.1 – Tema

Os computadores de hoje têm uma extensa genealogia, pois um dos primeiros dispositivos de computação foi o ábaco. Provavelmente teve suas raízes na China antiga e foi usado nas primeiras civilizações grega e romana. A arquitetura do ábaco é bastante simples, consistindo em elementos de contagem fixados em hastes e à medida que as contas se movem, suas posições representam valores armazenados. Além disso, a busca por máquinas de computação mais aprimorada se tornou mais acentuada, no período entre a Idade Média e a Era Moderna. (BROOKSHEAR, 2013)

Inteligência Artificial (IA) é a ciência de fazer máquinas inteligentes, especialmente programas de computador. Está relacionado à tarefa de usar computadores para entender a inteligência humana, mas a IA não precisa se limitar a métodos que são biologicamente observáveis. A Inteligência é a parte computacional da capacidade de atingir objetivos (MCCARTHY, 2007), como tomada de decisão e resolução de problemas.

Muitas empresas de tecnologia como Google, Microsoft e Amazon fizeram grandes investimentos no setor de Inteligência Artificial (IA). O Google, por exemplo, gastou bilhões adquirindo empresas do setor de IA, bem como contratando milhares de cientista de dados. Sabe-se que, mais e mais empregos exigiram conhecimentos de IA, isso não significa que você precisará aprender linguagens de programação ou saber estatística avançada, mas será fundamental ter uma base sólida dos fundamentos. Além disso, alguns dos principais impulsionadores da IA são: o imenso crescimento dos dados e novas infraestruturas tecnológicas. (TAULLI, 2020)

Com o avanço da tecnologia elas passam a fazer parte de nossas vidas de uma forma tão silenciosa que dificilmente percebemos, como o Aprendizado de Máquina (*Machine Learning*), que é uma subárea de pesquisa da Inteligência Artificial e que tem se destacado em vários setores da indústria.

Muitos dos recursos que temos hoje só são possíveis com o uso do Aprendizado de Máquina como a detecção de fraude, previsão de vendas e sistemas de recomendação de conteúdo.

1.2 – Justificativa

Existem milhões de consumidores inadimplentes no Brasil, de acordo com o indicador de inadimplência do Serasa Experian. O desemprego e as crises econômicas são de longe as situações mais estressantes para os orçamentos familiares. O total de brasileiros com contas vencidas atingiu 63 milhões em abril, 0,7% a mais que no mês anterior, a terceira alta do índice no ano. Em 2021 já são 1,6 milhão de pessoas que não pagaram suas dívidas e acabaram sendo negativadas. (ALMEIDA, 2021)

Em julho, o sistema financeiro concedeu 2,8% a menos em novos empréstimos e financiamentos do que em junho. O número leva em consideração o total de concessões de cada mês. Considerando a média por dia útil, houve queda de 7,2%, anunciou o Banco Central (BC). (RIBEIRO; TAIAR, 2021)

A tendência de alta da inadimplência neste ano está no radar dos quatro maiores bancos de capital aberto do país, que já projetam um possível aumento de suas respectivas taxas de inadimplência nos próximos meses. (BOLZANI; GARCIA, 2021)

1.3 – Objetivos

O objetivo deste trabalho, de um modo geral, é realizar um estudo de caso comparando modelos preditivos através de diversas técnicas de Aprendizagem de Máquina. A principal pergunta a ser respondida por esta pesquisa é: qual o modelo preditivo é mais eficaz e simples para prever o risco de crédito e qual o melhor algoritmo a ser utilizado?

Para atingir este objetivo geral, serão necessários atingir três objetivos específicos, são eles:

- Explicar os principais conceitos de *Machine Learning*;

- Realizar a seleção dos principais algoritmos de *Machine Learning* como regressão logística, árvore de decisão, máquina de vetores de suporte e floresta aleatória
- Efetuar a predição dos dados;

1.4 – Metodologia

A metodologia utilizada no desenvolvimento do presente trabalho baseia-se em pesquisa bibliográfica, aplicação de algoritmos de classificação e uma avaliação comparativa desses algoritmos através de um estudo de caso.

Na pesquisa bibliográfica foi realizado uma revisão da literatura sobre Aprendizado de Máquina e seus diferentes algoritmos de classificação utilizados para predição de inadimplência.

Além disso, na etapa de aplicação de algoritmos de classificação, se buscou avaliar o risco de inadimplência, através da criação de modelos. Na última etapa é feito a comparação dos resultados obtidos.

Foi utilizado um conjunto de dados e dividido aleatoriamente numa fração 30:70, onde a primeira parte foi utilizada para teste e na segunda para treinamento do modelo. Os modelos criados foram analisados através da área sob a curva (AUC). Foram utilizados os seguintes algoritmos de classificação: regressão logística, árvore de decisão, máquina de vetores de suporte (SVM) e floresta aleatória.

1.5 – Descrição

Este trabalho está estruturado da seguinte forma:

No capítulo 2 será apresentado o referencial teórico sobre *Machine Learning*, Aprendizado Supervisionado e métricas de avaliação de modelos preditivos.

O capítulo 3 apresenta a proposta de solução deste trabalho, descrevendo com detalhes as etapas, as técnicas utilizadas para criação dos modelos e as ferramentas utilizadas para a realização do trabalho.

Os resultados obtidos são apresentados no capítulo 4. Nele será explicitado os resultados alcançados em cada etapa do trabalho, utilizando um conjunto de dados. São apresentados 4 experimentos realizados para avaliar o algoritmo proposto.

O capítulo 5 conclui este trabalho apresentando análises sobre os resultados, limitações do trabalho e possíveis trabalhos futuros.

Capítulo 2

Embasamento Teórico

Este capítulo apresenta os principais conceitos de Aprendizado de Máquina e algumas definições importantes. O objetivo é contextualizar o leitor e responder algumas dúvidas, como, por exemplo: O que é? Como se classifica? Quais as suas características? Entre outros questionamentos a cerca do assunto. Neste capítulo, tentaremos responder a essas perguntas.

2.1 – Inteligência Artificial

A inteligência artificial é uma das áreas mais atuais e presentes nas ciências e engenharia, a atividade se iniciou logo após a Segunda Guerra Mundial, e este mesmo nome foi cunhado em 1956. Além do mais, a inteligência artificial engloba uma gigantesca diversidade de subcampos, nos dias de hoje, do geral como aprendizagem e percepção até tarefas específicas, como jogos de xadrez, composição de poesia, direção de um veículo em estrada movimentada e determinação de doenças. A IA é essencial para qualquer tarefa nos dias de hoje e é de fato um campo universal. (RUSSEL; NORVIG, 2013)

Considera-se que a época do surgimento das máquinas inteligentes começou na metade do século XX e Alan Turing se perguntou se as máquinas poderiam raciocinar. A inteligência artificial um subcampo da ciência da computação cresceu velozmente. Do mesmo modo, os humanos já sonharam em criar máquinas com o mesmo nível de inteligência dos humanos. (MOHAMMED; KHAN; BASHIER, 2017)

Além disso, as pesquisas atuais em aprendizado de máquina concentram se em: reconhecimento de padrões, computação cognitiva, processamento de linguagem naturais e outros. Essas linhas de pesquisa têm como objetivo permitir que as máquinas coletem dados através de sentidos semelhantes aos humanos e posteriormente, processe os dados coletados com técnicas de aprendizado de máquina para realizar previsões e tomar decisões no mesmo nível que os humanos. (MOHAMMED; KHAN; BASHIER, 2017)

2.2 – Machine Learning

De acordo com Géron (2017), “aprendizado de máquina é a ciência (e a arte) de programar computadores para que eles possam aprender com os dados.” Para o Samuel (1959 apud GÉRON, 2017), aprendizado de máquina é o campo de estudo que fornece aos computadores a habilidade de aprender sozinhos.

> *Para Mitchell (1997 apud BATISTA, 2003, p. 11), Aprendizado de Máquina - AM - é uma sub-área de pesquisa muito importante em Inteligência Artificial - IA - pois a capacidade de aprender é essencial para um comportamento inteligente. AM estuda métodos computacionais para adquirir novos conhecimentos, novas habilidades e novos meios de organizar o conhecimento já existente*

Desse modo, a teoria do Aprendizado de Máquina é baseada nos princípios do aprendizado indutivo, isto significa, que o modelo é definido por um conjunto de dados ou representações da experiência. O aprendizado indutivo geralmente é feito por algoritmos que processam o conjunto de dados e extraem um modelo que pode explicar ou representar os dados de alguma forma. (BRUNIALTI et al., 2015)

A indução é uma forma de dedução lógica que permite tirar conclusões gerais sobre um determinado conjunto de dados. Caracteriza-se por ser o raciocínio que parte de um conceito específico e o generaliza, ou seja, da parte para o todo. Na indução, você aprende um conceito de fazendo inferências indutivas sobre o conjunto de dados (MONARD; BARANAUSKAS, 2003)

Por exemplo, um programa de aprendizado de máquina poder ser utilizado no filtro de spam, que pode aprender a sinalizar se um e-mail é spam ou não, dado exemplos de e-mails de spam, sinalizados por usuários. Os exemplos que o sistema usa para aprender são chamados de conjunto de treino. Cada exemplo de treinamento é chamado de amostra de treino. (GÉRON, 2017)

Além disso, para Brunialti et al. (2015 p. 203) “Aprendizado de Máquina (AM) é caracterizado pelo desenvolvimento de técnicas que objetivam prover os softwares com a habilidade de melhorar seu desempenho em uma tarefa aprendendo através da experiência (aprendizado indutivo)”.

O Aprendizado de Máquina pode ser ótimo para várias situações como:

1. Para “problemas para os quais as soluções existentes exigem muita configuração manual ou longas listas de regras: um algoritmo de Aprendizado de Máquina geralmente simplifica e melhora o código.” (GÉRON, 2017)

2. “Problemas complexos para os quais não existe uma boa solução quando utilizamos uma abordagem tradicional: as melhores técnicas de Aprendizado de Máquina podem encontrar uma solução”. (GÉRON, 2017)

3. “Ambientes flutuantes: um sistema de Aprendizado de Máquina pode se adaptar a novos dados. Compreensão de problemas complexos e grandes quantidades de dados”. (GÉRON, 2017)

Dessa forma, "o aprendizado indutivo pode ser dividido em supervisionado e não-supervisionado. No aprendizado supervisionado é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido.” (MONARD; BARANAUSKAS, 2003, p.40)

Portanto, existem diferentes tipos de sistemas de aprendizado como o aprendizado supervisionado, que precisa ser treinado com supervisão humana, e não supervisionado, que busca identificar padrões ou grupos através dos dados, como a segmentação de clientes.

A figura 2.1 ilustra a hierarquia de aprendizado. Nessa hierarquia os nós sombreados levam ao aprendizado supervisionado usando classificação, objeto de estudo desta monografia.

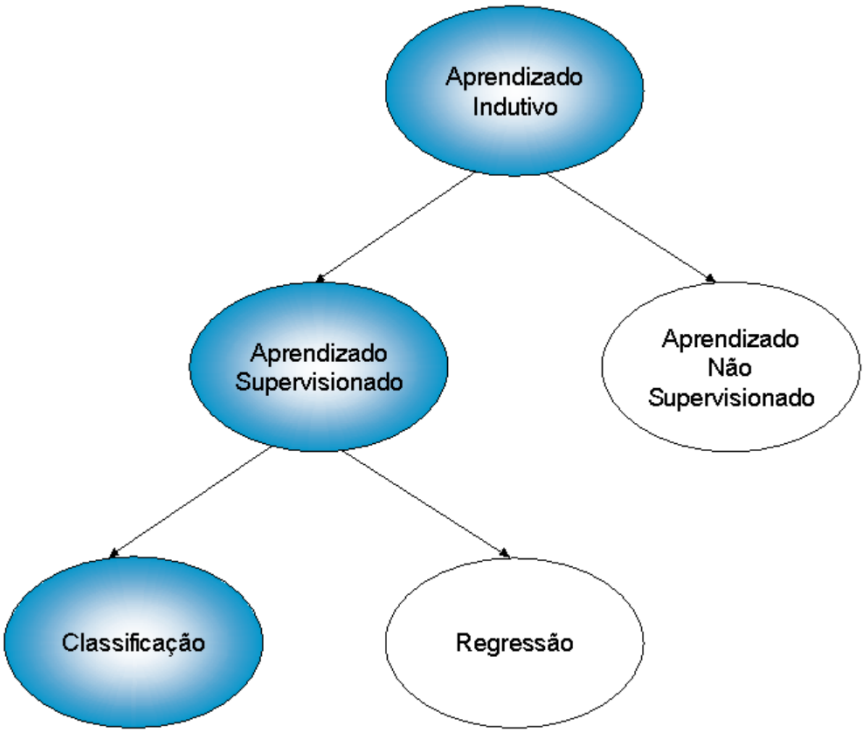

Figura 2.1 – Hierarquia do Aprendizado.

Fonte: Monard; Baranauskas

Além disso, nos problemas de regressão estamos tentando prever “respostas contínuas. Por exemplo, mudanças na temperatura, flutuações na demanda de energia e movimento de ações na bolsa. Aplicações típicas incluem previsão de carga de eletricidade e negociação algorítmica.” (HAYUM, 2019)

2.3 – Aprendizado Supervisionado

Os modelos de aprendizado supervisionado são divididos em problemas de classificação e regressão. Nos problemas de classificação estamos tentando prever “respostas categóricas. Modelos de classificação classificam os dados de entrada em classes. Aplicações típicas incluem classificação de imagens de satélites, reconhecimento de padrões de fala.” (HAYUM, 2019)

Além disso, para Amidi (2018, p.2) “Dado um conjunto de dados $\{x^{(1)}, \ldots, x^{(m)}\}$ associados a um conjunto de resultados $\{y^{(1)}, \ldots, y^{(m)}\}$, nós queremos construir um classificador que aprende como predizer y baseado em x”, conforme mostrado na figura 2.2.

	X_1	X_2	$\cdots$	X_m	Y
T_1	x_{11}	x_{12}	$\cdots$	x_{1m}	y_1
T_2	x_{21}	x_{22}	$\cdots$	x_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
T_n	x_{n1}	x_{n2}	$\ldots$	x_{nm}	y_n

Figura 2.2 – Conjunto de exemplos

Fonte: Monard; Baranauskas

De acordo com Monard e Baranauskas (2003), o objetivo do indutor (ou programa de aprendizado) é extrair o classificador apropriado de um conjunto de dados rotulados. Você pode então usar o classificador, que é a saída do indutor, para classificar novos exemplos (sem rótulo) e prever corretamente o rótulo de cada dado. O classificador pode então ser avaliado levando em consideração sua precisão ou outras propriedades desejadas para determinar sua eficácia para a tarefa atual.

A figura 2.3 abaixo mostra como funciona o modelo de aprendizado supervisionado:

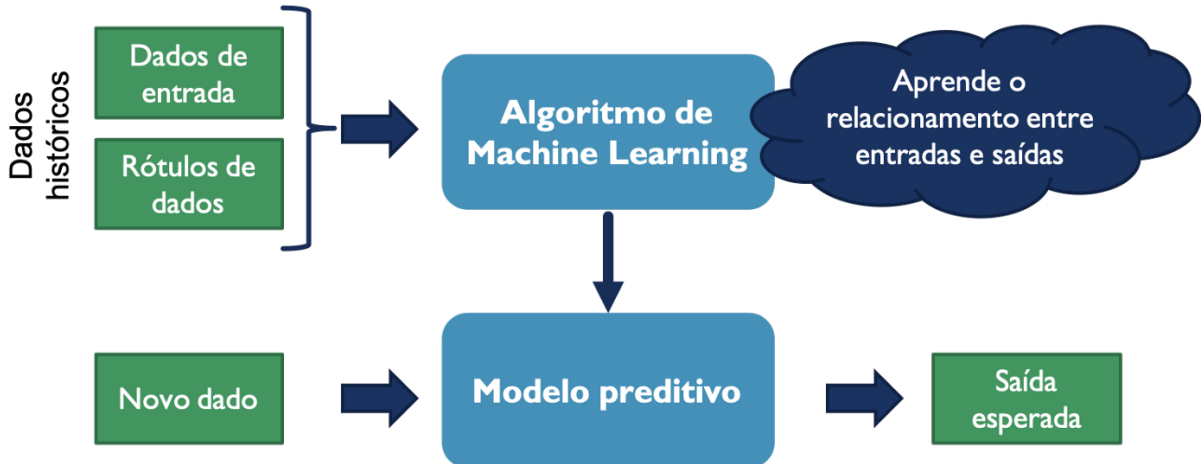

Figura 2.3 – Representação do aprendizado supervisionado.

Fonte: Escovedo, 2020

Sousa (2019, p. 43) define o Aprendizado Supervisionado da seguinte forma:

> *"No Aprendizado Supervisionado temos problemas onde há uma entrada e uma saída bem definida, tendo ideia que existe uma relação entre eles. Ou seja, temos os dados com quais queremos analisar e temos o que procurar, a tarefa e aprender o mapeamento desse percurso. Para tal, há uma fase denominada treinamento, em que alguns dados são inseridos no sistema, e o objetivo dele e encontrar parâmetros que se ajuste a dados desconhecidos do conjunto de teste, ou seja, esse tipo de aprendizado tem como finalidade prever o resultado dado um conjunto de amostras de treinamento, junto com seus rótulos de treinamento."*

Além disso, Mohammed, Khan e Bashier (2017, p. 7) definiram o objetivo do Aprendizado Supervisionado:

> *"Na aprendizagem supervisionada, o objetivo é inferir uma função ou mapeamento dos dados de treinamento que são rotulados. Os dados de treinamento consistem no vetor de entrada X e vetor de saída Y de rótulos ou tags. Um rótulo ou tag do vetor Y é a explicação de seu respectivo exemplo de entrada do vetor de entrada X. Juntos, eles formam um exemplo de treinamento. Em outras palavras, os dados de treinamento incluem exemplos de treinamento. Se a rotulagem não existir para o vetor de entrada X, então X é um dado não rotulado. Por que esse aprendizado é chamado de aprendizado supervisionado? O vetor de saída Y consiste em rótulos para cada exemplo de treinamento presente nos dados de treinamento."*

O conjunto de dados possui um atributo, chamado rótulo "que descreve o fenômeno de interesse, isto é, a meta que se deseja aprender e poder fazer previsões a respeito." (MONARD; BARANAUSKAS, 2003, p.43) Conforme mencionado anteriormente, os rótulos geralmente pertencem a um conjunto de classes discretas para classificação e valores contínuos para regressão.

Por exemplo, analisando um conjunto de dados sobre o número de quartos, tamanho da casa e localização no setor imobiliário com objetivo de prever o preço é um problema de regressão, pois o preço é uma variável contínua. Outro exemplo, analisando um conjunto de dados sobre tumor com objetivo de prever se é maligno ou benigno é um problema de classificação, pois a variável é discreta binária.

2.4 – Regressão logística

"Modelos lineares são uma classe de modelos amplamente utilizados na prática e extensivamente estudados nas últimas décadas, com raízes que remontam a mais de cem anos." (MÜLLER; GUIDO, 2017) A saída de uma função linear por meio da função de limiar cria um classificador. Além disso, a natureza rígida do limiar também causa um pouco de problemas. Esses problemas podem ser amplamente resolvidos suavizando a função de limiar, aproximando o limite rígido com uma função contínua. (RUSSEL; NORVIG, 2013)

"O processo de ajuste dos pesos desse modelo para minimizar a perda em um conjunto de dados é chamado de **regressão logística**. Não há solução fácil de forma fechada para encontrar o valor ótimo de w com esse modelo, mas o cálculo da descida pelo gradiente é simples." (RUSSEL; NORVIG, 2013) A regressão logística é frequentemente usada para estimar a probabilidade de uma instância pertencer a uma classe específica.

2.5 – Máquinas de vetores de suporte

"Máquinas de vetores de suporte, em inglês *Support Vector Machines*, são algoritmos de classificação que maximizam as margens entre as instâncias mais próximas, dessa forma, é criado um vetor otimizado que é então utilizado para classificar novas instâncias." (AMARAL, 2016, p. 45) Nas palavras de Russel e Norvig (2013), a máquina de vetor de suporte é a abordagem mais popular no momento sendo um excelente método para realizar os primeiros experimentos. Existem três propriedades que tornam o SVM interessante:

1. Os SVMs constroem um **separador de margem máxima**.
2. Os SVMs criam uma separação linear em hiperplano.
3. Os SVMs são um método não paramétrico.

2.6 – Árvore de decisão

O algoritmo de árvore de decisão segue uma estratégia de divisão para atingir os objetivos: ele sempre testa primeiro o atributo mais importante. Essa estratégia segue dividindo o problema em subproblemas menores, que podem então ser resolvidos recursivamente. O

atributo mais importante é aquele que mais faz diferença na classificação de uma amostra. Portanto, almejamos assim obter a classificação correta, com um número reduzido de testes. (RUSSEL; NORVIG, 2013)

2.7 – Floresta aleatória.

Floresta aleatória é um método de aprendizagem para tarefas de classificação e regressão funciona através de uma coleção de árvores de decisão, onde cada árvore difere ligeiramente das outras. Para tarefas envolvendo classificação a saída do método é a classe selecionada pela maioria das árvores. A ideia por trás das florestas aleatórias é que qualquer árvore pode fazer previsões razoavelmente boas, mas provavelmente se ajustará demais em parte dos dados. Se muitas árvores forem construídas, todas as quais funcionando bem e com sobre ajuste diferentes, podemos reduzir a quantidade de sobre ajuste calculando a média de seus resultados. É possível demonstrar através da matemática a redução no *overfitting* e manter o poder preditivo das árvores. (MÜLLER; GUIDO, 2017)

2.8 – Métricas de avaliação de desempenho

Em um contexto de classificação binária as principais métricas utilizadas para avaliar o desempenho dos classificadores estão listados abaixo, pois elas são importantes para avaliar a qualidade e a performance dos modelos de Aprendizado de Máquina em dados desconhecidos. Assim, utilizar diferentes indicadores conseguimos analisar a eficiência dos modelos preditivos em prever uma determinada variável. Os principais indicadores usados para avaliar a qualidade dos classificadores são:

A matriz de confusão, ilustrada na figura 2.4, é uma forma de expressar o desempenho de um método de classificação que é o número de Verdadeiros Positivos (VP), Falsos Positivos (FP), Verdadeiros Negativos (VN) e Falsos Negativos (FN). Com base nessas informações, outras métricas podem ser derivadas, facilitando a comparação de diferentes modelos. (LOPEZ et al., 2018)

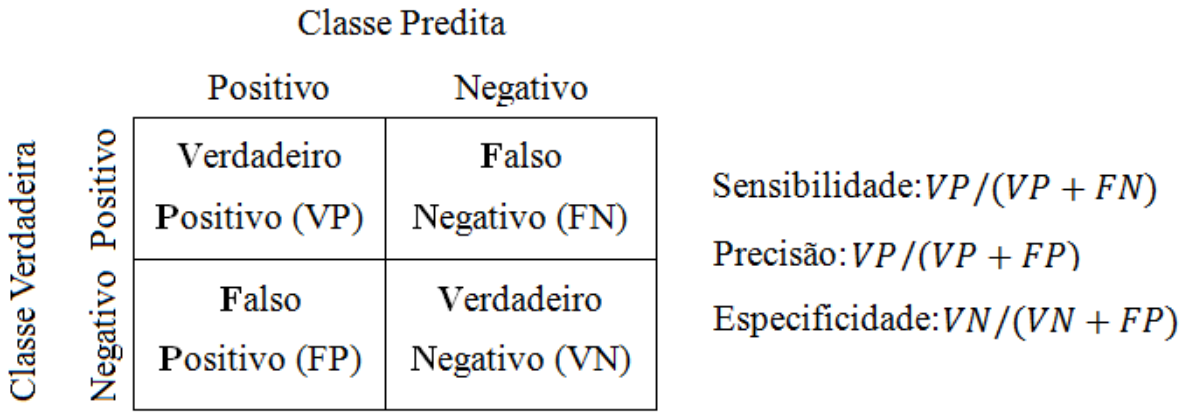

Figura 2.4 – Matriz de confusão

Fonte: Filho et al. 2015

A Acurácia que "é a razão do total de amostras classificadas corretamente (VP + VN) dividido pelo número total de amostras (P+N)." (LOPEZ et al., 2018, p. 28) Mas para conjunto de dados desbalanceados, em que a proporção entre as classes não são as mesmas, a acurácia não é um bom indicador.

A Precisão é uma métrica que mede a número de exemplos classificados como verdadeiros positivos dividido pela soma entre os classificados como verdadeiros positivos e falsos positivos.

Recall (Revocação), também conhecido como Sensibilidade é uma métrica que mede a quantidade de amostras positivas que foram classificadas corretamente pelo modelo, ou seja, é a quantidade de amostras positivos verdadeiras divididos pelo total de positivos.

A Área sob a Curva ROC (AUC) é uma métrica importante para conjunto de dados que possuem classes desbalanceadas, pois nela mede-se a área sob uma curva formada pelo gráfico entre a taxa de amostras positivas e a taxa de amostras falso positivos.

Capítulo 3

Proposta de Solução

O Capítulo 3 apresenta a proposta de solução do problema e mostra a sequência dos passos seguidos para realizar o experimento. Além disso, são apresentadas as tecnologias utilizadas e a contextualização do estudo de caso, as etapas para a realização dos experimentos.

3.1 – Tecnologias utilizadas

Todas as implementações foram feitas utilizando a linguagem de programação Python. Na primeira etapa foi usado a biblioteca Pandas, que é uma ferramenta de análise e manipulação de dados de código aberto, e a biblioteca NumPy, que é uma ferramenta de computação numérica oferecendo diversas funções matemáticas de código aberto e fácil de usar.

Foi utilizado o colab que é um serviço na nuvem que permite escrever códigos em Python e possui a seguinte configuração: dois processadores Intel Xeon CPU @ 2.20GHz, 12.48 GB de memória RAM disponível e 108GB de HD

Além disso, será feita a obtenção dos dados dos clientes seguida da limpeza dos dados, caso seja indispensável, e em seguida será feita a visualização dos dados através da análise exploratória com a biblioteca Seaborn que fornece uma área de interação de alto nível para a criação de gráficos estatísticos.

Na próxima etapa, que é de análise e modelagem, será criado um modelo preditivo utilizando o algoritmo de aprendizado supervisionado, que será implementado por meio dos dados obtidos dos clientes para que possamos prever contingente que ficará inadimplente ou não. Para isso, será utilizada a biblioteca de Aprendizado de Máquina chamada scikit-learn.

scikit-learn é uma biblioteca de aprendizado de máquina de código aberto que compreende diversos algoritmos de classificação, agrupamento e regressão incluindo árvore de decisão, floresta aleatória, máquinas de vetores de suporte e k-means.

A sequência a seguir apresenta as etapas que serão realizadas no projeto de Aprendizado de Máquina aplicado a risco de crédito.

1. Compreender o problema e definir objetivos
2. Coletar e analisar os dados
3. Preparar os dados
4. Construir o modelo
5. Avaliar o modelo
6. Apresentar os resultados

A figura 3.1 apresenta o fluxo geral do projeto de AM

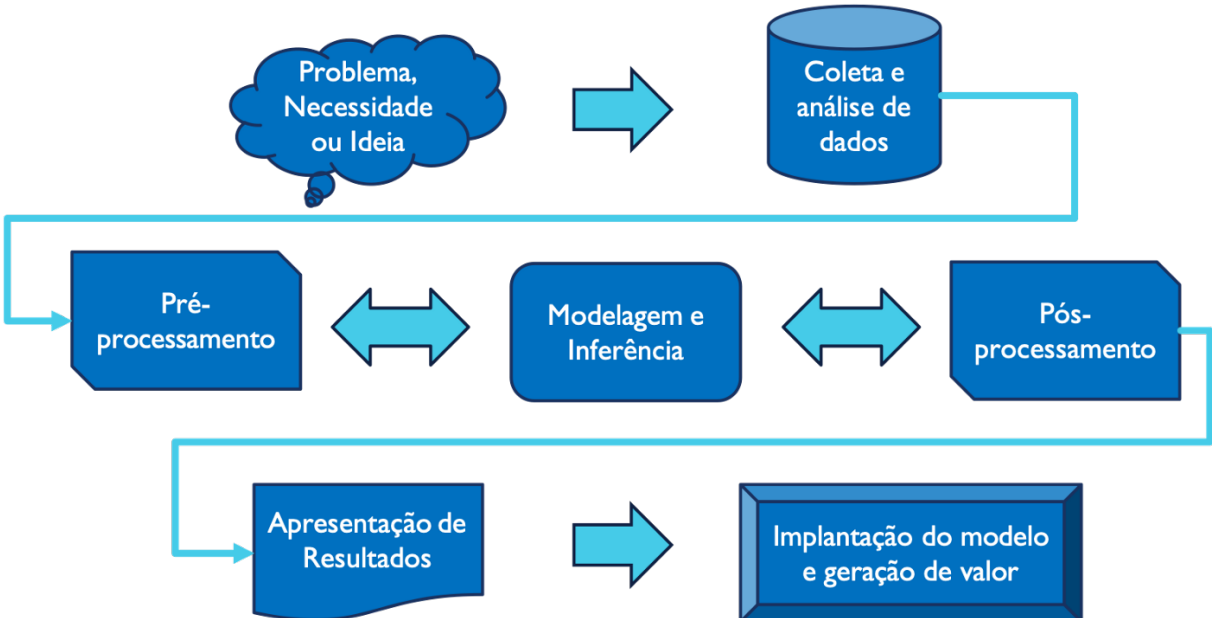

Figura 3.1 – Fluxo geral do projeto de Machine Learning.

Fonte: Escovedo; Koshiyama, 2020

3.2 –Descrição do problema

Nos setores financeiros, por exemplo, existe a necessidade de prever quais clientes deixaram de pagar suas dívidas, ou seja, quais clientes ficaram inadimplentes. Os bancos podem sofrer diversas perdas em produtos de concessão de crédito e uma possível causa dessas perdas é a dívida pendente do cliente dificultando o banco de recuperar os pagamentos pelos serviços prestados.

Por isso, foi estudado técnicas de Aprendizado de Máquina Supervisionado, a qual possui diversos algoritmos que vêm sendo utilizados no mercado. Será analisado um banco de dados de clientes para verificar quantos clientes possuem a probabilidade de inadimplência. Há muitos modelos de Aprendizado de Máquina para prever quais clientes possuem a possibilidade de ficarem inadimplentes, de forma que os bancos possam cancelar ou reduzir as linhas de crédito de clientes de risco para minimizar as perdas.

Deste modo, há a necessidade de se encontrar o melhor modelo que atenda satisfatoriamente esse caso.

3.3 – Obtenção dos dados

Esta pesquisa, utilizou uma base de dados pública que contém informações de uma instituição financeira atuante no mercado de crédito que estão disponibilizados no Kaggle, UCI *Machine Learning Repository* e outros.

Sabe-se que Kaggle é uma plataforma online que reúne especialistas em aprendizado de máquina e cientistas de dados, visto que os usuários podem explorar, pesquisar e publicar conjuntos de dados, construir modelos em uma plataforma de ciência de dados, e colaborar com outros especialistas em aprendizado de máquina. Do mesmo modo, você também pode resolver diversos desafios de ciência de dados, além de participar de competições.

UCI *Machine Learning Repository* é uma plataforma que mantém 588 conjuntos de dados para a comunidade de Aprendizado de Máquina onde é possível explorar, pesquisar e publicar bancos de dados. (DUA; GRAFF, 2019)

O conjunto de dados utilizado contém 1000 entradas com 20 atributos. Nesse conjunto de dados, cada entrada representa uma pessoa que recebe crédito de um banco. Cada pessoa será classificada como alto ou baixo risco de crédito.

O quadro 3.1 apresenta a relação de variáveis presente no conjunto de dados e utilizada no estudo como o saldo na poupança, o histórico de crédito, o status da conta corrente e assim por diante para avaliar o risco de crédito usando o Aprendizado de Máquina.

Quadro 3.1 – Lista de variáveis

Variável	**Descrição do item**	**Valores possíveis**
checking_status	Status da conta corrente existente	< 0 DM; 0 <= x < 200 DM; >= 200 DM; no checking account.
duration	Duração em meses	Numeric.
credit_history	Histórico de crédito	critical/other existing credit; existing paid; delayed previously; no credits/all paid; all paid.
purpose	Motivo do crédito	car (new); car (used); furniture/equipment; radio/television; domestic appliances; repairs; education; retraining; business; others.
credit_amount	Quantidade de crédito	Numeric.
savings_status	Poupança / títulos	< 100 DM; 100 <= x <500 DM; 500 <= x < 1000 DM; x >= 1000 DM; no savings account.
employment	Tempo no trabalho atual	unemployed; < 1 year; 1 <= x < 4 years; 4 <= x < 7 years; x >= 7 years.
installment_commitment	Taxa de parcelamento em porcentagem da renda disponível	Numeric.
personal_status	Estado civil e sexo	male : divorced/separated;female : divorced/separated/married; male : single; male : married/widowed; female : single.
other_parties	Outros devedores/ fiadores	none; co-applicant; guarantor.
residence_since	Tempo residência atual	Numeric.
property_magnitude	Propriedade	real estate; life insurance; no known property; car.
age	Idade	Numeric.

other_payment_plans	Outros planos de parcelamento	bank; stores; none.
housing	Habitação	rent; own; for free.
existing_credits	Número de créditos existentes neste banco	Numeric.
job	Emprego	unemployed/ unskilled - non-resident; unskilled – resident; skilled employee / official; management/ self-employed/highly qualified employee/ officer;
num_dependents	Número de pessoas responsáveis por fornecer manutenção	Numeric.
own_telephone	Telefone	none; yes, registered under the customers name.
foreign_worker	Trabalhador estrangeiro	yes; no.

3.4 – Preparação dos dados

Os dados do mundo real geralmente contêm muitos elementos ausentes e inconsistências, ou seja, erros de digitação. Por isso, para que o aprendizado de máquina seja bem-sucedido, depois que os dados são obtidos, eles devem ser limpos, organizados e preparados.

Este processo é uma etapa importante e muitas vezes as pessoas gastam 80% do tempo nesta fase. Ter um conjunto de dados limpo melhora a precisão dos modelos. Foi utilizada a biblioteca *Label Enconding* para converter todas as variáveis categóricas em numéricas.

3.5 – Criação e avaliação do modelo

Nesta etapa após a verificação e seleção das variáveis, o conjunto de dados foi dividido em duas amostras aleatórias, conforme mostrado na figura 3.2, uma para o treinamento do modelo com 70% do conjunto de dados e outra para teste do modelo com 30% do conjunto de dados. As amostras de treinamento e de teste serão avaliadas pelos modelos de Aprendizado de Máquina.

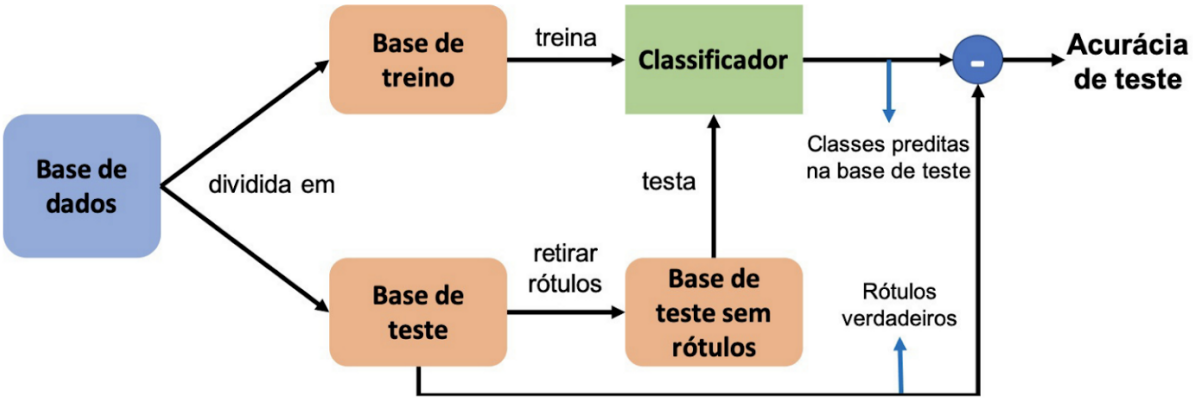

Figura 3.2 – Arquitetura da divisão do conjunto de dados.

Fonte: Escovedo; Koshiyama, 2020

Capítulo 4

Resultados Obtidos ou Esperados

4.1 – Resultado 1

A partir das propostas tecnológicas descritas no capítulo 03, apresenta-se a sequência de resultados obtidos para a solução do problema estudado.

Neste capítulo iremos apresentar os resultados obtidos pela regressão logística, máquinas de vetores de suporte, árvore de decisão e floresta aleatória. Para a análise dos resultados foram utilizadas as seguintes métricas: acurácia, precisão, *recall*, f1-*score*, AUC e matriz de confusão.

A implementação da regressão logística exibiu os seguintes resultados com uma acurácia de 0.7000 e um valor da área sob a curva ROC de 0.5324, a precisão para adimplência foi de 0.7314 e para a inadimplência foi de 0.5690, nesse mesmo contexto, no *recall* o valor foi de 0.8762 para a adimplência e 0.3367 para a inadimplência, o f1-*score* ficou em 0.7973 para adimplência e 0.4231 para inadimplência.

Na matriz de confusão ocorreu uma discriminação de 33 amostras como verdadeiro positivo, 25 amostras como falso positivo, 65 amostras como falso negativo e 177 amostras como verdadeiro negativo, conforme mostrado na figura 4.1.

```
Relatório de Classificação:
              precision    recall  f1-score   support

         bad     0.5690    0.3367    0.4231        98
        good     0.7314    0.8762    0.7973       202

    accuracy                         0.7000       300
   macro avg     0.6502    0.6065    0.6102       300
weighted avg     0.6783    0.7000    0.6751       300

Acurácia: 0.7000

AUC: 0.5324
```

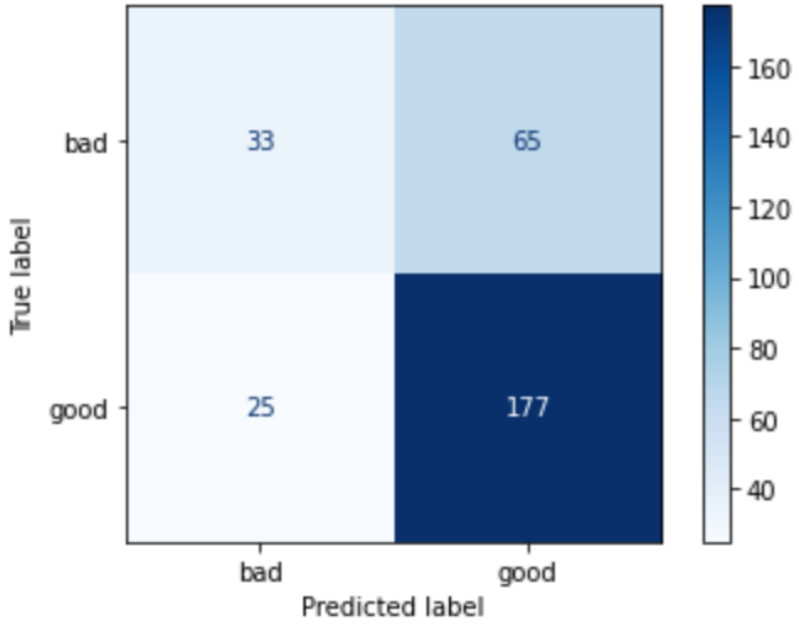

Figura 4.1 – Relatório de classificação.

A implementação da Máquina de Vetores de Suporte exibiu os seguintes resultados com uma acurácia de 0.7267 e um valor da área sob a curva ROC de 0.7102, a precisão para adimplência foi de 0.7273 e para a inadimplência foi de 0.7222, nesse mesmo contexto, no *recall* o valor foi de 0.9505 para a adimplência e 0.2653 para a inadimplência, o f1-*score* ficou em 0.8240 para adimplência e 0.3881 para inadimplência.

Na matriz de confusão ocorreu uma discriminação de 26 amostras como verdadeiro positivo, 10 amostras como falso positivo, 72 amostras como falso negativo e 192 amostras como verdadeiro negativo, conforme mostrado na figura 4.2.

Relatório de Classificação:

	precision	recall	f1-score	support
bad	0.7222	0.2653	0.3881	98
good	0.7273	0.9505	0.8240	202
accuracy			0.7267	300
macro avg	0.7247	0.6079	0.6060	300
weighted avg	0.7256	0.7267	0.6816	300

Acurácia: 0.7267

AUC: 0.7102

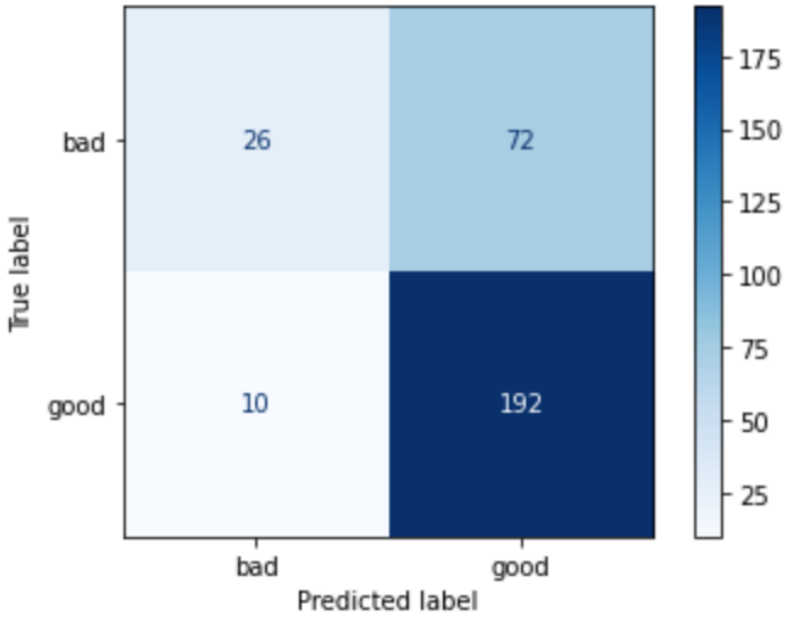

Figura 4.2 – Relatório de classificação

A implementação da Árvore de decisão exibiu os seguintes resultados com uma acurácia de 0.6900 e um valor da área sob a curva ROC de 0.6463, a precisão para adimplência foi de 0.7685 e para a inadimplência foi de 0.5258, nesse mesmo contexto, no *recall* o valor foi de 0.7723 para a adimplência e 0.5204 para a inadimplência, o f1-*score* ficou em 0.7704 para adimplência e 0.5231 para inadimplência.

Na matriz de confusão ocorreu uma discriminação de 51 amostras como verdadeiro positivo, 46 amostras como falso positivo, 47 amostras como falso negativo e 156 amostras como verdadeiro negativo, conforme mostrado na figura 4.3.

```
Relatório de Classificação:
              precision    recall  f1-score   support

         bad     0.5258    0.5204    0.5231        98
        good     0.7685    0.7723    0.7704       202

    accuracy                         0.6900       300
   macro avg     0.6471    0.6463    0.6467       300
weighted avg     0.6892    0.6900    0.6896       300

Acurácia: 0.6900

AUC: 0.6463
```

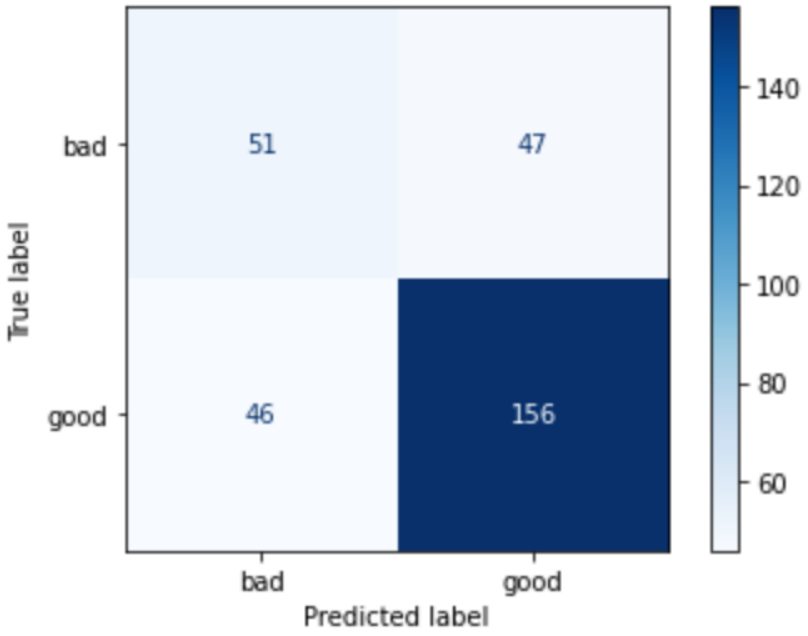

Figura 4.3 – Relatório de classificação

A implementação da Floresta aleatória exibiu os seguintes resultados com uma acurácia de 0.7467 e um valor da área sob a curva ROC de 0.7972, a precisão para adimplência foi de 0.7561 e para a inadimplência foi de 0.7037, nesse mesmo contexto, no *recall* o valor foi de 0.9208 para a adimplência e 0.3878 para a inadimplência, o f1-*score* ficou em 0.8304 para adimplência e 0.5000 para inadimplência.

Na matriz de confusão ocorreu uma discriminação de 38 amostras como verdadeiro positivo, 16 amostras como falso positivo, 60 amostras como falso negativo e 186 amostras como verdadeiro negativo, conforme mostrado na figura 4.4.

```
Relatório de Classificação:
              precision    recall  f1-score   support

         bad     0.7037    0.3878    0.5000        98
        good     0.7561    0.9208    0.8304       202

    accuracy                         0.7467       300
   macro avg     0.7299    0.6543    0.6652       300
weighted avg     0.7390    0.7467    0.7224       300

Acurácia: 0.7467

AUC: 0.7972
```

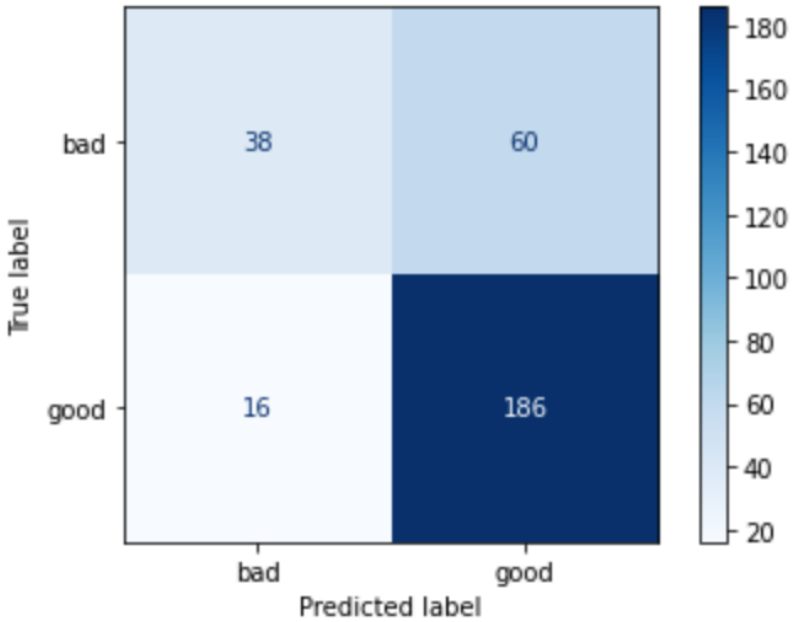

Figura 4.4 – Relatório de classificação

A figura 4.5 ilustra os resultados dos experimentos realizados, mostrando as métricas da área sob a curva (AUC) e acurácia. O modelo que apresentou os melhores resultados nos experimentos tanto na área sob a curva quanto na acurácia foi a Floresta aleatória exibindo AUC de 79,7% e acurácia de 74,6% e o segundo modelo que apresentou os melhores resultados foi o SVM com AUC de 71% e acurácia de 72,6%. A floresta aleatória teve um desempenho 12% superior que o SVM em termos de área sob a curva.

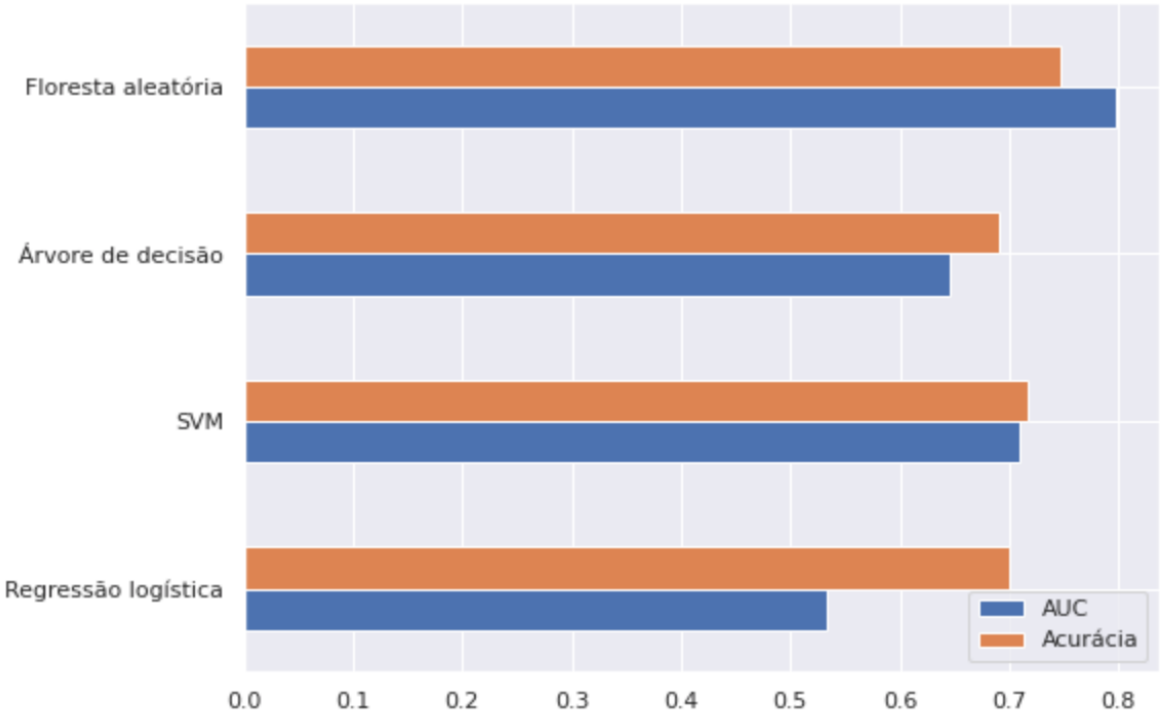

Figura 4.5 – Resultado da avaliação de desempenho

Capítulo 5

Conclusão e Trabalhos Futuros

5.1 – Conclusão

Neste trabalho, vimos conceitos e técnicas de Aprendizado de Máquina para classificação de risco de crédito utilizando uma base de dados pública disponível na internet. Foi utilizado os principais algoritmos de *Machine Learning* para prever a inadimplência como regressão logística, máquinas de vetores de suporte, árvore de decisão e floresta aleatória.

O modelo preditivo que apresentou os melhores resultados foi floresta aleatória, pois foi obtido um classificador com excelente desempenho com valor da área sob a curva ROC de 79% nos experimentos. Portanto, a proposta de solução apresentado ajudou a atingir a cada um dos objetivos apresentados no trabalho de conclusão.

O problema descrito na seção 3.2 foi resolvido como demostrado na seção 4.1 em que foi construído diversos classificadores utilizando algoritmos de Aprendizado de Máquina Supervisionado para tratar as situações mencionadas.

5.2 – Trabalhos Futuros

Como sugestão para trabalhos futuros, indica-se realizar o balanceamento da base de dados e realizar a otimização de hiperparâmetros, ou seja, escolher os melhores hiperparâmetros para os modelos preditivos, assim resultará em uma melhor performance do classificador.

Além disso, outra sugestão seria implementar outros modelos de Aprendizado de Máquina como Redes Neurais e comparar com os resultados obtidos nos experimentos.

Referências Bibliográficas

ALMEIDA, Marília. Mais de 1,5 milhão de pessoas ficaram inadimplentes em 2021. **Invest**, 27 mai. 2021. Disponível em: https://invest.exame.com/mf/mais-de-um-milhao-e-meio-de-pessoas-se-tornaram-inadimplentes-em-2021. Acesso em: 6 out. 2021.

AMARAL, Fernando. **Aprenda mineração de dados:** teoria e prática. Alta Books Editora, 2016.

AMIDI, Afshine. AMIDI, Shervine. **SUPER VIP Cheatsheet: Aprendizado de Máquina**. Stanford University, 2018.

BOLZANI, Isabela; GARCIA, Larissa. Bancos projetam aumento da inadimplência nos próximos meses. **Folha de S. Paulo**, 7 mai. 2021. Disponível em: https://www1.folha.uol.com.br/mercado/2021/05/bancos-projetam-aumento-da-inadimplencia-nos-proximos-meses.shtml. Acesso em: 6 out. 2021.

BROOKSHEAR, J. Glenn. **Ciência da Computação - 11ed: Uma Visão Abrangente**. Bookman Editora, 2013.

BRUNIALTI, Lucas. et al. **Aprendizado de Máquina em Sistemas de Recomendação Baseados em Conteúdo Textual:** Uma Revisão Sistemática. In: Simpósio Brasileiro de Sistemas de Informação (SBSI), 11. 2015.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. Irvine, CA: University of California, School of Information and Computer Science, 2019. Disponível em: http://archive.ics.uci.edu/ml. Acesso em: 8 dez. 2021.

ESCOVEDO, Tatiana. KOSHIYAMA, Adriano. **Introdução a Data Science:** Algoritmos de Machine Learning e métodos de análise. Casa do Código, 2020.

ESCOVEDO, Tatiana. Machine Learning: Conceitos e Modelos — Parte I: Aprendizado Supervisionado. **Medium**, 2020. Disponível em: https://tatianaesc.medium.com/machine-learning-conceitos-e-modelos-f0373bf4f445. Acesso em: 12 set. 2021.

FILHO, Geraldo. et al. **ResiDI Um Sistema de Decisão Inteligente para Infraestruturas Residenciais via Sensores e Atuadores Sem Fio**. XXXIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos, SBRC'15, 2015.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. Alta Books, 2019.

HAYUM, Luiz. **Um estudo comparativo de técnicas de Machine Learning na classificação de fácies:** aplicações nos campos de Hugoton e Panoma e campo de

Namorado. Tese (Graduação em Engenharia de Petróleo). Rio de Janeiro: UFRJ/ Escola Politécnica, 2019.

LOPEZ, Martin. et al. **Aprendizado de Máquina em Plataformas de Processamento Distribuído de Fluxo:** Análise e Detecção de Ameaças em Tempo Real. 2018. Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC).

MCCARTHY, John. **What is Artificial Intelligence?** Computer Science Department – Stanford University, Stanford, CA, 2007.

MOHAMMED, M.; KHAN, M.; BASHIER, E. **Machine Learning:** Algorithms and Applications. Taylor & Francis Group, 2017.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre Aprendizado de Máquina. In **Sistemas Inteligentes - Fundamentos e Aplicações**, Rezende, S.O. (ed.), Editora Manole, pp. 89-114, 2003.

MÜLLER, Andreas; GUIDO, Sarah. **Introduction to Machine Learning with Python**. O'Reilly Media, 2017.

RIBEIRO; TAIAR, Alex; Estevão. Concessões de crédito caem 2,8% e inadimplência fica estável em julho. **Valor**, 27 ago. 2021. Disponível em: https://valor.globo.com/financas/noticia/2021/08/27/concessoes-de-credito-caem-28percent-e-inadimplencia-fica-estavel-em-julho.ghtml. Acesso em: 6 out. 2021.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência artificial**. Rio de Janeiro: Elsevier, 2013.

SOUSA, Maria**. Uma análise do algoritmo K-Means como introdução ao Aprendizado de Máquinas**. Tese (Graduação). Universidade Federal do Tocantins – Curso de Matemática, 2019.

TAULLI, Tom. **Introdução à Inteligência Artificial: Uma abordagem não técnica**. Novatec Editora, 2020.