

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Gabriel Tobias Fuhr

**USO DE *MACHINE LEARNING* PARA A CLASSIFICAÇÃO DO
CRÉDITO DE EMPRESAS POR MEIO DE DEMONSTRATIVOS
FINANCEIROS.**

Santa Maria, RS
2022

Gabriel Tobias Fuhr

USO DE *MACHINE LEARNING* PARA A CLASSIFICAÇÃO DO CRÉDITO DE EMPRESAS POR MEIO DE DEMONSTRATIVOS FINANCEIROS.

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharelado em Ciência da Computação**. Defesa realizada por videoconferência.

ORIENTADOR: Prof. Sergio Luis Sardi Mergen

Número do TG: 498
Santa Maria, RS
2022

©2022

Todos os direitos autorais reservados a Gabriel Tobias Fuhr. A reprodução de partes ou do todo deste trabalho só poderá ser feita mediante a citação da fonte.

Endereço: Avenida João Machado Soares, 812

End. Eletr.: gtfuhr@inf.ufsm.br

GABRIEL TOBIAS FUHR

**USO DE MACHINE LEARNING PARA A CLASSIFICAÇÃO DO
CRÉDITO DE EMPRESAS POR MEIO DE DEMONSTRATIVOS
FINANCEIROS.**

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação**.

Aprovado em 10 de Fevereiro de 2022:



Sergio Luis Sardi Mergen, Dr. (UFSM)
(Presidente/Orientador)



Joaquim Vinicius Carvalho Assunção, Dr. (UFSM)



Daniel Lichtnow, Dr. (UFSM)

AGRADECIMENTOS

Agradeço a minha mãe que sempre acreditou na minha vontade de estudar e batalhou para me dar a oportunidade de fazer um curso superior. Agradeço a minha noiva pelo apoio e paciência durante a escrita desse trabalho. E, por último, mas não menos importante, agradeço pelos valiosos aconselhamentos e orientação do Professor Sergio Luis Sardi Mergen.

*O que você sabe não tem valor; o valor
está no que você faz com o que sabe.*

(Bruce Lee)

RESUMO

USO DE *MACHINE LEARNING* PARA A CLASSIFICAÇÃO DO CRÉDITO DE EMPRESAS POR MEIO DE DEMONSTRATIVOS FINANCEIROS.

AUTOR: Gabriel Tobias Fuhr

ORIENTADOR: Sergio Luis Sardi Mergen

Um elemento importante para a economia é a chamada Classificação de Crédito, um conceito indicativo das condições de quitação de créditos concedidos. O processo de classificação é realizado por agências de classificação que usam métodos que não são completamente compreendidos/reproduzíveis, pois dependem de um expertise proprietário. Aumentar a transparência no processo de Classificação de Crédito ajudaria a explicar a razão de certas entidades conseguirem tomar crédito e outras não. Nesse sentido, o presente trabalho visa utilizar dados públicos para predição de Classificação de Crédito de empresas por meio de modelos de aprendizado de máquina. Os modelos utilizados foram: Árvore de Decisão, Floresta Aleatória, *Gradient Boosting* para Classificação de Crédito binária e multiclasse. Como resultados, obteve-se uma maior acessibilidade a respeito da Classificação de Crédito e a obtenção de modelos de livre acesso que sejam autoexplicativos. O modelo com melhor desempenho, tanto para a classificação binária quanto para a multiclasse, foi o *Gradient Boosting*, que apresentou o melhor *F1 score*. Em suma, o presente trabalho mostrou que o uso de aprendizado de máquina para a Classificação de Crédito é uma proposta viável e informativa para esse processo, além de ser reproduzível.

Palavras-chave: Aprendizado de Máquina. Classificação de Crédito. Aprendizado Supervisionado. Software de código aberto. Dados Abertos.

ABSTRACT

USE OF MACHINE LEARNING TO CLASSIFY COMPANIES CREDIT THROUGH FINANCIAL STATEMENTS.

AUTHOR: Gabriel Tobias Fuhr

ADVISOR: Sergio Luis Sardi Mergen

An important element for the economy is the so-called credit rating, a concept indicative of the conditions for repayment of credits granted. The rating process is carried out by rating agencies using methods that are not fully understood/reproducible as they rely on proprietary expertise. Increasing transparency in the credit rating process would help explain why certain entities are able to take out credit and others are not. In this sense, the present work aims to use public data to predict the credit rating of companies through machine learning models. The models used were: Decision Tree, Random Forest, *Gradient Boosting* for binary and multi-class credit rating. As a result, greater accessibility was obtained regarding the credit rating and the obtaining of free access models that are self-explanatory. The model with the best performance, both for binary and multi-class classification, was *Gradient Boosting*, which presented the best *F1 score*. In short, the present work showed that the use of machine learning for credit rating is a viable and informative proposal for this process, in addition to being reproducible.

Keywords: Machine Learning. Credit Rating. Supervised Learning. Open source software. Open Data.

LISTA DE FIGURAS

Figura 2.1 – Representação gráfica da Árvore de Decisão para a classificação de vinhos. .	27
Figura 2.2 – Floresta aleatória genérica.	29
Figura 2.3 – Funcionamento do Gradient Boosting.	30
Figura 5.1 – Representação gráfica simplificada da Árvore de Decisão para o Credit Rating.	50
Figura 7.1 – Representação gráfica completa da Árvore de Decisão para o Credit Rating.	57

LISTA DE GRÁFICOS

Gráfico 3.1 – Publicações de <i>Credit Rating</i>	31
Gráfico 3.2 – Publicações de <i>Credit Rating</i> que incluem o termo <i>Machine Learning</i>	32
Gráfico 3.3 – Publicações de <i>Credit Rating</i> agrupadas por países. Os dez países com mais publicações são exibidos junto com o Brasil.	32
Gráfico 3.4 – Publicações de <i>Credit Rating</i> que incluem o termo <i>Machine Learning</i> agrupadas por países. Os dez países com mais publicações são exibidos junto com o Brasil.	33
Gráfico 5.1 – Matriz de confusão com os dados de teste binários.	44
Gráfico 5.2 – Matriz de confusão com os dados de teste multiclasse.	46

LISTA DE TABELAS

Tabela 2.1 – Notas de risco de crédito.	17
Tabela 2.2 – Matriz de Confusão (Tabela de valor predito vs valor atual.).....	22
Tabela 5.1 – Desempenho da Classificação Binária.	43
Tabela 5.2 – Desempenho da Classificação Multiclasse.	45
Tabela 5.3 – Variáveis mais importantes na Classificação Binária.	47
Tabela 5.4 – Variáveis mais importantes na Classificação Multiclasse.....	47

LISTA DE ABREVIATURAS E SIGLAS

<i>CR</i>	Classificação de Crédito (do inglês <i>Credit Rating</i>)
<i>CV</i>	Validação Cruzada (do inglês <i>Cross-Validation</i>)
<i>DT</i>	Árvore de Decisão (do inglês <i>Decision Tree</i>)
<i>EBA</i>	European Banking Authority
<i>GB</i>	Gradiente Boosting (do inglês <i>Gradient Boosting</i>)
<i>HPO</i>	Otimização de Hiperparâmetros (do inglês <i>Hyperparameter optimization</i>)
<i>KNN</i>	K-vizinhos Mais Próximos (do inglês <i>K-nearest neighbors</i>)
<i>LTR</i>	Classificação a Longo Prazo (do inglês <i>Long-Term Rating</i>)
<i>ML</i>	Aprendizado de Máquina (do inglês, <i>Machine Learning</i>)
<i>NN</i>	Redes Neurais (do inglês <i>Neural Networks</i>)
<i>RF</i>	Floresta Aleatória (do inglês <i>Random Forest</i>)
<i>STR</i>	Classificação a Curto Prazo (do inglês <i>Short-Term Rating</i>)
<i>SVM</i>	Máquinas de Vetores de Suporte (do inglês <i>Support-Vector Machines</i>)
<i>TBT</i>	A Grande Árvore (do inglês <i>The Big Three</i>)

LISTA DE SÍMBOLOS

VP Verdadeiro-positivo

VN Verdadeiro-negativo

FP Falso-positivo

FN Falso-negativo

SUMÁRIO

1	INTRODUÇÃO	14
1.1	OBJETIVOS	15
1.1.1	Objetivo Geral	15
1.1.2	Objetivos Específicos	15
2	REVISÃO DE LITERATURA	16
2.1	CLASSIFICAÇÃO DE CRÉDITO	16
2.1.1	Tipos de Classificação de Crédito	16
2.1.2	Notas e faixas de risco de crédito	17
2.2	APRENDIZADO DE MÁQUINA	19
2.2.1	Tipos de aprendizado	19
2.2.2	Explicabilidade	20
2.2.3	Separação dos dados de entrada	20
2.2.4	<i>Overfitting</i> e <i>UnderFitting</i>	21
2.2.5	Estatística e mensuração de desempenho de ML	21
2.2.6	Otimização de parâmetros em ML	23
2.2.7	Algoritmos Supervisionados de ML	24
2.2.7.1	<i>Neural Networks</i>	25
2.2.7.2	<i>Árvore de Decisão</i>	25
2.2.7.3	<i>Floresta Aleatória</i>	28
2.2.7.4	<i>Gradient Boosting</i>	29
3	TRABALHOS RELACIONADOS	31
3.1	USO DE <i>MACHINE LEARNING</i> NO <i>CREDIT RATING</i>	33
3.1.1	<i>Árvore de Decisão</i>	33
3.1.2	<i>Floresta Aleatória</i>	34
3.1.3	<i>Gradient Boosting</i>	34
3.2	CONSIDERAÇÕES SOBRE OS TRABALHOS RELACIONADOS	35
4	MATERIAIS E MÉTODOS	36
4.1	OBTENÇÃO DE DADOS FINANCEIROS DE EMPRESAS	36
4.1.1	Dados de entrada	36
4.1.2	Dados de saída	36
4.1.3	Problemas de Classificação	37
4.1.4	Associação entre dados de entrada e saída	38
4.2	MODELOS DE APRENDIZADO DE MÁQUINA	39
4.2.1	Árvores de Decisão	40
4.2.2	<i>Floresta Aleatória</i>	40
4.2.3	<i>Gradient Boosting</i>	41
4.2.4	Análise estatística	41
5	RESULTADOS	42
5.1	ANÁLISE DOS ALGORITMOS	42
5.1.1	Classificação Binária	42
5.1.2	Classificação Multiclasse	44
5.2	EXPLICABILIDADE DOS ALGORITMOS	46
5.2.1	Classificação binária	46
5.2.2	Classificação multiclasse	47

5.2.3	Análise das características para as classificações	48
5.2.3.1	<i>Classificação Binária</i>	48
5.2.3.2	<i>Classificação Multiclasse</i>	48
5.2.3.3	<i>Características em comum da classificação Binária e Multiclasse</i>	49
5.3	EXPLICABILIDADE DO MODELO GERADO PARA A ÁRVORE DE DECISÃO	49
6	CONCLUSÕES	51
	REFERÊNCIAS BIBLIOGRÁFICAS	52
7	ÁRVORE DE DECISÃO USADA PARA CR.	57

1 INTRODUÇÃO

Por definição, crédito é a transação de dinheiro entre duas partes, onde o prestador fornece dinheiro ao tomador em troca de um pagamento futuro (ENCYCLOPAEDIA, 2020). Essas transações normalmente envolvem o pagamento de juros ao prestador. É de interesse de quem irá emprestar verificar se a outra parte é apta a honrar com a dívida. O processo de analisar essa competência de pagamento tem por objetivo avaliar o risco de crédito, e é chamada de Classificação de Crédito, do inglês, *credit rating* (CR) (LANGOHR HERWIG; LANGOHR, 2009).

O CR é um elemento importante na economia, uma vez que ajuda a equilibrar a concessão de crédito para quem está apto a quitá-lo. Dessa forma, o CR impede que empresas ou países que estão em uma situação econômica fragilizada contraiam mais dívidas e piores sua situação econômica. Todos os países e corporações que tomam crédito estão sujeitos a terem o seu risco de crédito classificado. Essas informações são interessantes para a manutenção de um mercado que seja capaz de avaliar a credibilidade de empresas e países (SECURITIES; COMMISSION, 2002). Dados a respeito da credibilidade geram impacto sobre a economia, pois eles fornecem uma opinião independente sobre empresas e países que podem receber empréstimos (SEC; COMMISSION, 2017).

Classificar empresas e países fica no encargo de agências de CR. As três maiores agências que contemplam 94% do mercado mundial, conhecidas como *The Big Three* (TBT), são: a Standard & Poor's, Moody's, e Fitch Group (LANGOHR HERWIG; LANGOHR, 2009).

Os critérios básicos de CR utilizados por essas agências podem ser explorados nos seus portais online. Entretanto, os modelos utilizados em CR ainda não possuem código aberto disponível, o que dificulta a total compreensão acerca do seu funcionamento. Além disso, as agências de CR detêm uma base de dados de empresas e países, desde o começo de suas operações, que não está disponível ao público. Dessa forma, por ser uma base de dados reservada às agências de CR, uma empresa ou país pode encontrar dificuldade em reconhecer quais fatores influenciaram no seu CR (RATINGS, 2021; SERVICE, 2021).

Nesse sentido, explorar novos meios de se extrair informações a respeito do CR é uma tarefa relevante, pois ajuda empresas a descobrir os principais fatores que influenciam no processo de classificação. Há inúmeras maneiras de se utilizar e analisar esses dados, seja estatisticamente ou economicamente, ou até mesmo por meio de modelos de aprendizado de máquina.

O Aprendizado de Máquina, do inglês *Machine Learning* (ML), permite o uso de dados para o treinamento de algoritmos. Com o processamento de vários exemplos sobre determinada tarefa, esses algoritmos criam modelos capazes de realizar a mesma tarefa, imitando um comportamento inteligente. Além disso, algumas técnicas de ML geram modelos auto-explicativos, que apresentam de maneira compreensível quais fatores são importantes para tomada de decisão durante a tarefa (ROSCHE et al., 2020).

Dessa forma, o presente trabalho busca realizar a predição das classificações de crédito de empresas usando técnicas de ML e dados financeiros públicos. Ainda, pretende-se obter respostas sobre os principais fatores que influenciam o CR. Para isso, pretende-se utilizar técnicas de ML de alta explicabilidade, que sejam capazes de gerar modelos que sejam fáceis de compreender.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Análise do uso de diferentes modelos baseados em Árvores de Decisão para verificar se dados públicos são suficientes para obter um CR transparente.

1.1.2 Objetivos Específicos

- Mensurar o desempenho das técnicas de ML para realizar a Classificação de Crédito;
- Extrair dados sobre a importância de cada indicador de um demonstrativo financeiro para o CR, de acordo com os modelos treinados;
- Disponibilizar os dados financeiros obtidos ao público, em um formato que seja de fácil uso.
- Disponibilizar o código fonte desenvolvido durante a pesquisa, aumentando a reprodutibilidade dos experimentos.

2 REVISÃO DE LITERATURA

2.1 CLASSIFICAÇÃO DE CRÉDITO

A Classificação de Crédito, do inglês *Credit Rating*, o (CR) é um elemento que busca trazer equilíbrio ao mercado financeiro, pois facilita o crédito para quem consegue pagá-lo. O CR impede que empresas ou países que estão em uma situação econômica fragilizada contraiam mais dívidas e piorem sua situação econômica. Todos os países e corporações que tomam crédito estão sujeitos a terem o seu risco de crédito classificado. Essas informações são interessantes para a manutenção de um mercado que seja capaz de avaliar a credibilidade de empresas e países (SECURITIES; COMMISSION, 2002). Dados a respeito da credibilidade geram impacto sobre a economia, pois fornecem uma opinião independente sobre empresas e países que podem receber empréstimos (SEC; COMMISSION, 2017).

A Classificação de Crédito, propriamente dita, normalmente se resume a uma nota, nota esta que costuma ser um valor discreto. A nota de crédito está normalmente atrelada a um instrumento de investimento, os instrumentos de investimento definem a detenção de ações, notas promissórias ou títulos de renda fixa. No presente trabalho, instrumentos de investimento se referem a títulos de renda fixa, ou títulos de dívida emitidos por empresas de capital aberto. A nota de crédito tem o objetivo de auxiliar um possível investidor a entender o nível de risco relacionado a um instrumento de investimento.

2.1.1 Tipos de Classificação de Crédito

A Classificação de Crédito pode se dividir em duas categorias, classificação de curto prazo, do inglês *Short-Term Rating* (STR) e a classificação de longo prazo, do inglês, *Long-Term Rating* (LTR). O STR fornece a nota de risco para operações de crédito com o prazo considerado curto, de no máximo um ano. Em contrapartida, o LTR se refere a operações em que o prazo é maior que um ano (GLOBAL, 2021).

Geralmente a nota do STR de uma entidade é menor que sua nota LTR. A liquidez, ou seja, sua capacidade de converter ativos em dinheiro, influencia fortemente o STR. (LANGOHR HERWIG; LANGOHR, 2009)

2.1.2 Notas e faixas de risco de crédito

As três grandes, do inglês *The Big Three*, como são chamadas as maiores empresas do setor de CR, definem as faixas mais usadas e conhecidas de risco de crédito. A *European Banking Authority* (EBA) definiu um mapeamento comum de notas de crédito entre as três empresas. As notas definidas pelas BTs e o mapeamento comum de notas é demonstrado pela Tabela 2.1, o mapeamento está disponível na coluna *EU Credit Quality Step*.

Tabela 2.1 – Notas de risco de crédito.

Moody's		Standard&Poor's		Fitch Ratings		EU Credit Quality Step
Long-term	Short-term	Long-term	Short-term	Long-term	Short-term	
Aaa	P-1	AAA	A-1+	AAA	F1+	1
Aa1		AA+		AA+		
Aa2		AA		AA		
Aa3		AA		AA		
A1		A+	A-1	A+	F1	
A2	A	A				
A3	P-2	A	A-2	A	F2	2
Baa1		BBB+		BBB+		
Baa2	P-3	BBB		BBB	F3	
Baa3		BBB	A-3	BBB		
Ba1	Not Prime	BB+	B	BB+	B	
Ba2		BB		BB		
Ba3		BB		BB		
B1		B+		B+		
B2		B		B		
B3		B		B		
Caa1		CCC+	C	CCC+	C	6
Caa2	CCC	CCC				
Caa3	CCC	CCC				
Ca	CC	CC				
	C	C				
C	RD	D	RD	D		
/	SD		D			
/	D		/			

Fonte: Adaptado de (EBA, 2019a; EBA, 2019b; EBA, 2019c).

Conforme mostra a Tabela 2.1, percebe-se que o CR possui diferentes notas para o STR e o LTR. A EBA criou faixas de crédito, originalmente chamadas de *EU Quality Steps*, que explicam e unificam de maneira mais simples o risco de crédito. As faixas de crédito são

criadas conforme o número de instrumentos de determinada nota que acabam inadimplentes. O ato de entrar em inadimplência é comumente chamado de *default*.

De acordo com (The European Commission, 2016), a nota 1 refere-se à entidades que possuem capacidade muito forte para honrar seus compromissos financeiros e está sujeita a risco de crédito muito baixo. Além disso, nota 1 indica que as taxas de inadimplência variam de 0,10% a 0,16%. Entidades com a nota 2 têm forte capacidade de cumprir seus compromissos financeiros e está sujeita a baixo risco de crédito. Entretanto, a nota 2 é um pouco mais suscetível aos efeitos adversos de mudanças nas circunstâncias e condições econômicas, e suas taxas de inadimplência variam de 0,17% a 0,54%. A nota 3 define entidades que possuem capacidade adequada para honrar seus compromissos financeiros e está sujeita a um risco de crédito moderado. Entretanto, condições econômicas adversas ou mudanças nas circunstâncias têm maior probabilidade de levar a um enfraquecimento da capacidade da entidade classificada de honrar seus compromissos financeiros. Para a nota 3, as taxas de inadimplência variam de 0,55% a 2,39% (The European Commission, 2016). As classificações de 1 a 3 costumam se referir a instrumentos de investimentos seguros, do Inglês *investment-grade* (LANGOHR HERWIG; LANGOHR, 2009).

As entidades de nota 4 têm capacidade para honrar seus compromissos financeiros, mas estão sujeitas a um risco de crédito substancial. Essas entidades enfrentam grandes incertezas e exposição a condições adversas de negócios, financeiras ou econômicas, o que pode levar à capacidade inadequada da entidade avaliada em honrar seus compromissos financeiros. Em uma nota 4, as taxas de inadimplência variam de 2,40% a 10,99%. A nota 5 indica capacidade para cumprir compromissos financeiros, mas também o alto risco de crédito. As condições adversas de negócios, financeiras ou econômicas, provavelmente prejudicarão a capacidade ou disposição da entidade avaliada em honrar seus compromissos financeiros. Em uma nota 5, as taxas de inadimplência variam de 11,00% a 26,49%. Por fim, a nota 6 define entidades que têm taxas de inadimplência que variam de 26,50% a 100,00%. As entidades com nota 6 estão atualmente vulneráveis e sujeitas a um risco de crédito muito alto, incluindo inadimplência ou muito próximo a ela. Para cumprir seus compromissos, uma entidade de nota 6 depende de condições comerciais, financeiras e econômicas favoráveis (The European Commission, 2016). As classificações de 4 a 6 costumam se referir a instrumentos especulativos que possuem um risco elevado e altas taxas de juros para compensar o alto risco. Nos Estados Unidos, é comum a denominação de *junk* ou *high-yield* para esse tipo de instrumento, que significam, 'lixo' e 'alta produção', respectivamente, em português. O *junk* é usado para remeter a sua natureza de alto risco e o *high-yield* ao altos valores de juros (LANGOHR HERWIG; LANGOHR, 2009; CHEN; BOYLE, 2020).

Dessa forma, as notas de CR de 1 a 6 podem ser agrupadas em dois grupos: investimentos seguros, ou *investment-grade*, que correspondem as notas 1 a 3, e instrumentos especulativos de alto risco, que correspondem as notas 4 a 6, também conhecidos em Inglês como *high-yield* ou *junk*. Esta divisão das notas de CR, apresenta então, dois grupos de extrema importância

para o setor financeiro.

2.2 APRENDIZADO DE MÁQUINA

O aprendizado de máquina, do inglês *Machine Learning* (ML), é um ramo da ciência da computação e um subconjunto da inteligência artificial que, por meio de experiências anteriores e através de treinamento de máquina, é capaz de aprender a reconhecer padrões (SINGH et al., 2020). O objetivo do ML é aprender com dados fornecidos de experiências antecessoras a fim de se extrair informações sobre o processo analisado (MAHESH, 2019; SINGH et al., 2020).

O treinamento da máquina é feito mediante fornecimento de um conjunto de dados e aplicação de algoritmos específicos, para que seja possível aprender com os dados e fazer previsões ou classificações. Os dados fornecidos aos algoritmos de treinamento podem ser binários, nominais ou até mesmo contínuos (SINGH et al., 2020).

O ML é uma ferramenta poderosa para solucionar diversos tipos de problemas, facilitando tomadas de decisão e ações por meio de suas interpretações. As técnicas de ML estão evoluindo progressivamente e podem ser aplicadas em praticamente todos os domínios (MEKONNEN et al., 2020).

2.2.1 Tipos de aprendizado

Para o ML, há três maneiras de se treinar a máquina para aprender com os dados ou experiências: aprendizado supervisionado, não supervisionado e por reforço (MEKONNEN et al., 2020).

No aprendizado supervisionado, os problemas se dividem em problemas de Classificação ou de Regressão. Quando a variável a ser predita é quantitativa, trata-se de um problema de Regressão. Por outro lado, quando a variável é categórica, trata-se de um problema de Classificação (MEKONNEN et al., 2020).

O aprendizado supervisionado é aquele em que são definidas variáveis de entrada e saída e um algoritmo que cria modelos capazes de relacioná-las (SINGH et al., 2020; MEKONNEN et al., 2020). Para que os modelos gerados sejam capazes de realizar previsões ou classificações, o conjunto de dados é dividido entre dados de treino e teste (MAHESH, 2019). A separação dos dados em teste e treino faz com que o aprendizado supervisionado forneça dados mais precisos e confiáveis (SINGH et al., 2020). Exemplos de algoritmos de aprendizado supervisionado são: Árvores de Decisões, Florestas Aleatórias, *Gradient Boosting*, Redes Neurais.

Por outro lado, o aprendizado não supervisionado inclui apenas as variáveis de entrada e nenhuma saída (SINGH et al., 2020; MEKONNEN et al., 2020). Por não ter nenhuma saída relacionada com as entradas, o algoritmo precisa descobrir por conta própria uma estrutura para

os dados. Esse tipo de técnica pode ser utilizada, por exemplo, para o agrupamento de dados em grupos com base nas suas similaridades, detecção de valores atípicos nos dados e redução de dimensionalidade (MAHESH, 2019; MEKONNEN et al., 2020). Exemplos de algoritmos de aprendizado não supervisionado são: K-means e Apriori.

O aprendizado de reforço é um tipo de aprendizado em que a máquina aprende com os dados e tenta decidir qual a melhor ação a ser tomada de acordo com as condições em que os dados se encontram. Esse tipo de treinamento usa uma sequência de decisões para atingir uma meta em um ambiente incerto e complexo, sendo baseado em uma interação entre um agente e o ambiente. Por meio de tentativa e erro, a máquina encontra a solução para o problema. O objetivo é maximizar o resultado final. Um exemplo de algoritmo de aprendizado de reforço é o *Q-learning*. (DAYAN; NIV, 2008; MNIH et al., 2015)

2.2.2 Explicabilidade

A explicabilidade de um modelo se refere à facilidade em se entender como o modelo de ML chega a uma predição. Áreas como a saúde e finanças costumam exigir explicabilidade em modelos, pois as decisões dos modelos possuem consequências importantes (ROSCHER et al., 2020).

Um modelo pode ser explicável devido a forma que o mesmo é construído, como pode ser visto em Árvores de Decisão, Florestas Aleatórias e *Gradient Boosting*. Um exemplo de algoritmo que não tem a explicabilidade como sua principal característica são as Redes Neurais (BURKART; HUBER, 2021). Uma alternativa mais complexa de se buscar a compreensão de um algoritmo, é a de realizar a engenharia reversa de modelos de ML já treinados. Isso é feito para extrair conhecimento sobre como o algoritmo faz as suas previsões, como pode ser visto em (BENÍTEZ; CASTRO; REQUENA, 1997; OH; SCHIELE; FRITZ, 2019).

2.2.3 Separação dos dados de entrada

Para os algoritmos de aprendizagem supervisionada, é necessário realizar a separação dos dados em treino e teste. Os algoritmos são treinados e otimizados utilizando os dados de treino. Os dados de teste ficam inacessíveis ao algoritmo durante todo o treino.

Ao final do treinamento e otimização, o modelo é avaliado com os dados de teste, simulando como seria o desempenho do modelo de ML ao classificar dados que ele nunca processou. Com essa técnica, pode-se prever como o modelo se sairia no mundo real, onde as entradas que são dadas ao modelo são desconhecidas no momento do treino.

Um modelo que obtém ótimo desempenho com os dados que treinou pode ter apenas decorado os valores vistos durante o treino e ter um desempenho fraco com dados que nunca

viu (GÉRON, 2019). Então, o desempenho de um modelo deve sempre ser avaliado com dados que o mesmo não encontrou durante o treino, pois isso evita o descobrimento de falsas relações nos dados.

Um processo bastante utilizado na estruturação dos dados de treino é chamado de Validação Cruzada, do inglês, *Cross-Validation* (CV). Este processo visa aumentar a confiabilidade sobre as métricas de desempenho de um modelo. A CV divide os dados de treino em N partes, então N-1 partes são utilizadas para otimizar os parâmetros e a última é utilizada para testar o desempenho do modelo com estes parâmetros. Esse processo é repetido até todas as diferentes porções de treino serem utilizadas como teste. A média da métrica de desempenho em todas as partes testadas é usada para avaliar os parâmetros.

2.2.4 *Overfitting e UnderFitting*

Os dados utilizados em um problema de ML não irão sempre compor um problema trivial a ser resolvido pois, de maneira resumida, os dados podem ser compostos por sinal e ruído. Os sinal se referem à características dos dados que auxiliam a predizer a variável de saída. Já o ruído é o dado que não traz nenhum benefício durante a predição da variável de saída.

Há dois problemas comuns que podem ocorrer durante o treinamento. O primeiro deles é o *Overfitting*, onde o modelo de ML treina usando os dados de treino, obtém um desempenho significativo classificando os dados de treino, mas não consegue repetir o mesmo desempenho nos dados de teste. Quando isso ocorre, significa que o modelo acabou aprendendo não só os sinal nos dados de treino, mas também o seu ruído. Ou seja, o modelo não consegue utilizar o seu conhecimento para entradas que nunca processou. Para combater o *Overfitting*, pode-se utilizar o ajuste de parâmetros do modelo, ou regularização, com o intuito de limitar a sua capacidade de aprendizado, evitando que ele aprenda o ruído dos dados.

O *Underfitting*, por sua vez, se refere a quando o modelo tem um desempenho fraco ao classificar instâncias de treino e de teste. Isso demonstra que o modelo que está sendo utilizado não tem a capacidade necessária para aprender o sinal que estão presentes nos dados. A utilização de modelos mais complexos ou o ajuste de parâmetros que tornam o modelo mais flexível podem ajudar a minimizar esse problema (GÉRON, 2019; DIETTERICH, 1995)

2.2.5 Estatística e mensuração de desempenho de ML

Um aspecto muito importante para o ML é a avaliação do desempenho. Por isso, a melhor abordagem para avaliação do desempenho depende do problema a ser abordado (JAPKOWICZ; SHAH, 2015). Há diversos tipos de métricas para avaliar o desempenho de um

algoritmo ML (HANDELMAN et al., 2019). No presente trabalho, serão abordadas métricas para a mensuração do desempenho de métodos supervisionados de classificação utilizando ML.

Matriz de Confusão (tabela de valor predito vs valor atual)

Uma classificação binária ou predição pode resultar em quatro resultados diferentes, conforme a combinação entre o valor predito pelo modelo e o valor atual. São os seguintes: Verdadeiro-positivo, Verdadeiro-negativo, Falso-positivo e Falso-negativo. A combinação entre os valores é demonstrada pela Tabela 2.2. Essa tabela também pode ser chamada de Matriz de Confusão de um modelo, pois evidencia exatamente onde o modelo está acertando ou cometendo erros.

Tabela 2.2 – Matriz de Confusão (Tabela de valor predito vs valor atual.)

		Predito	
		Positivo	Negativo
Atual	Positivo	Verdadeiro-positivo (VP)	Falso-negativo (FN)
	Negativo	Falso-positivo (FP)	Verdadeiro-negativo (VN)

Fonte: Extraído e adaptado de (GÉRON, 2019).

O Verdadeiro-positivo é quando o valor predito é positivo e a predição é correta. O Verdadeiro-negativo é quando o valor predito é negativo e a predição é correta. O Falso-positivo é quando o valor predito é positivo e a predição é incorreta. O Falso-negativo é quando o valor predito é negativo e a predição é incorreta.

Esses conceitos ajudam a identificar problemas quando uma classificação possui um desbalanceamento de classes. Um problema de classificação desbalanceado pode ser observado, por exemplo, quando existem poucas observações positivas a serem classificadas. Nesse caso, o alto valor de verdadeiros-negativos e baixo valor de falso-positivos do modelo pode apresentar resultados expressivos mesmo sem real desempenho do modelo. (POZZOLO et al., 2015)

Precisão

Algoritmos de classificação podem ser avaliados de acordo com sua precisão. A precisão de um modelo é dada pela razão entre o número de previsões positivas identificadas corretamente e o número total da amostra da entrada, conforme mostra a Equação 2.1 (HANDELMAN et al., 2019). Aumentar a precisão tem como foco reduzir o número de falso-positivos.

$$Precisao = \frac{VP}{VP + FP} \quad (2.1)$$

Recall

O *Recall* mede o número de previsões positivas corretas feitas a partir de todas as previsões positivas que poderiam ter sido feitas.

A precisão mostra o contexto geral sobre as previsões positivas corretas de todas as previsões positivas. O *recall*, por sua vez, pode mostrar as previsões positivas perdidas. Otimizar o *recall* tem como objetivo reduzir os falso-negativos. A sua fórmula pode ser observada pela Equação 2.2.

$$Recall = \frac{VP}{VP + FN} \quad (2.2)$$

Medida F1

Uma métrica que busca trazer equilíbrio entre a precisão e o *recall* é a pontuação *F1*, que traz a média harmônica entre as duas medidas. A fórmula pode ser observada na Equação 2.3. Os possíveis valores para a pontuação F1 vão de zero até um.

$$F1 = \frac{VP}{VP + \frac{FN+FP}{2}} \quad (2.3)$$

Abordagem Ingênu

Com base nos dados do problema sendo analisado, pode ser feita a avaliação de como uma abordagem simplística e ingênu

2.2.6 Otimização de parâmetros em ML

Cada modelo de ML dispõe de diversos parâmetros que influenciam no aprendizado. A escolha certa dos parâmetros pode influenciar no desempenho final. Pensando nisso, técnicas de otimização de parâmetros (HPO), do inglês, *hyperparameter-optimization*, são utilizadas.

Essas técnicas são aplicadas nos dados de treino e costumam otimizar o desempenho da predição. Elas exploram um espaço de possíveis parâmetros, buscando otimizar alguma uma métrica de desempenho escolhida para o modelo. Algumas técnicas simples podem ser utilizadas, como a experimentação exaustiva com todas as combinações de parâmetros e também como a busca Aleatória por melhores combinações de parâmetros. (FEURER; HUTTER, 2019; YANG; SHAMI, 2020).

2.2.7 Algoritmos Supervisionados de ML

Para a ciência de dados, não há uma definição de um algoritmo único que seja capaz de resolver todos os problemas. Dessa forma, há a possibilidade de se testar algoritmos que se adéquem melhor a cada caso analisado (MAHESH, 2019).

O ML é uma área ampla, e oferece diversas ferramentas e algoritmos a serem aplicados, tais como Árvore de Decisão, Floresta Aleatória e Gradiente Boosting, entre outros. O ML também conta com outros algoritmos disponíveis, tais como Redes Neurais, Máquinas de Vetores de Suporte e K-ésimo vizinho mais próximo (AYODELE, 2010).

Os algoritmos usados em CR são normalmente supervisionados, pois os mesmos costumam ter variáveis de entrada descrevendo uma empresa ou pessoa e então uma variável de saída como a sua Classificação de Crédito ou probabilidade de inadimplência. Devido a isso, modelos supervisionados serão abordados durante o trabalho.

2.2.7.1 *Neural Networks*

As redes neurais, do inglês *Neural Networks*, representam a subseção de aprendizado profundo, do inglês *Deep Learning*, do aprendizado de máquina. A técnica consiste em imitar o funcionamento de uma rede neural biológica para modelar dados.

Basicamente, uma rede neural é composta por uma rede de nós, divididas por várias camadas, podendo elas serem ocultas, de entrada e de saída. Além disso, as camadas podem ser simplesmente ou densamente conectadas. Cada conexão possui um peso e um viés, sendo que os pesos definem a força de uma conexão entre os nós. Ou seja, os pesos definem o quanto a saída de um nó irá influenciar na entrada de outro nó. Já o viés é uma constante adicionada à saída de um nó. As saídas da camada de saída são utilizadas como a predição da rede. Durante o treinamento, os pesos são ajustados para otimizar o desempenho de predição da rede (LECUN et al., 1989).

Esta técnica foi citada pela primeira vez na metade do século XX (MCCULLOCH; PITTS, 1943). Entretanto, o processo de treinar as redes neurais era computacionalmente caro. Esta percepção em relação ao custo operacional mudou com a aplicação do conceito de *back-propagation*, que foi criado por (KELLEY, 1960) e aprimorado por (LINNAINMAA, 1976).

A aplicação do algoritmo de *backpropagation* permitiu o treinamento das redes neurais e a obtenção de resultados expressivos em diversas áreas (WIDROW; RUMELHART; LEHR, 1994; LECUN et al., 1989).

No entanto, uma das desvantagens das redes neurais é a sua baixa capacidade explicativa, pois o funcionamento das camadas ocultas de uma rede neural treinada é pouco compreensível a seres humanos em relação a como a rede toma as suas decisões. Existem técnicas que tentam aumentar a capacidade de explicação das redes neurais, entretanto, elas envolvem a engenharia reversa dos algoritmos (BENÍTEZ; CASTRO; REQUENA, 1997; OH; SCHIELE; FRITZ, 2019).

Durante a Classificação de Crédito, é interessante que modelos utilizados sejam compreensíveis sobre como são feitas as predições. Assim, os utilizadores desses modelos podem analisar os motivos de cada predição ou ao menos as principais influências dos dados no processo de decisão.

Devido a isso, a partir deste ponto, apenas técnicas que contenham alta explicabilidade serão apresentadas e exploradas mais a fundo. Dentre as técnicas existentes, destacam-se as Árvores de Decisão, Florestas Aleatórias e o *Gradient Boosting*, (ROSCHE et al., 2020).

2.2.7.2 *Árvore de Decisão*

A Árvore de Decisão, do inglês *Decision Tree* (DT), é um algoritmo de aprendizado supervisionado (SINGH et al., 2020; MAHESH, 2019). As DTs podem ser utilizadas para regressão ou classificação de dados (MEKONNEN et al., 2020). Os resultados são obtidos

graficamente sob a forma de uma árvore. Os nós representam um evento com seus atributos e as folhas representam as regras e as condições de decisão e o valor que o nó pode assumir (LOH, 2014; SOMVANSHI et al., 2016; MAHESH, 2019).

A DT divide os dados em conjuntos homogêneos, sejam eles dois ou mais, baseando-se no divisor mais significativo entre as variáveis independentes (MEKONNEN et al., 2020). Em outras palavras, a árvore de decisão aprende quais são os pontos que melhor separam os dados de acordo com as suas categorias.

De forma genérica, uma vez construída a árvore, pode-se usá-la para a tarefa de classificação usando o seguinte algoritmo:

1. Iniciar pelo nó raiz.
2. Observar a regra condicional estipulada para este nó. Caso a regra seja satisfeita, segue-se o caminho esquerdo, caso contrário, segue-se o caminho direito. Repita esse passo até chegar a um nó folha.
3. Utilizar as informações de classe do nó folha para fazer a predição.

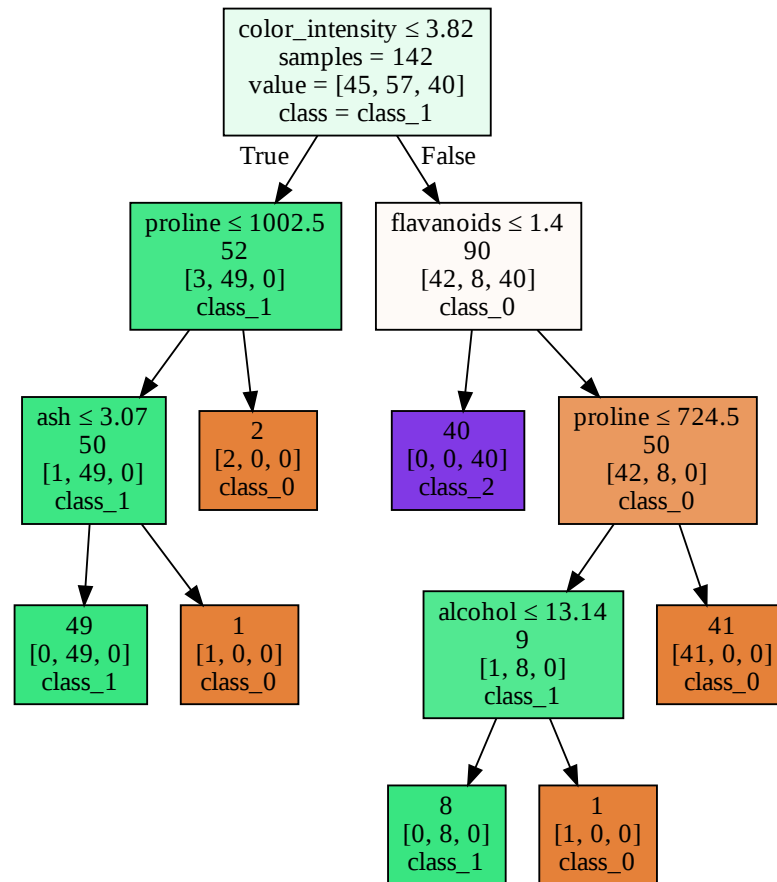
Para ilustrar, uma árvore de decisão que classifica vinhos em três classes diferentes foi criada. Os dados utilizados descrevem exemplares de vinhos através de características como o teor alcoólico, cor, intensidade, alcalinidade e outras características. Fica ao encargo da árvore de decisão descobrir, durante o treino, quais características e de qual forma elas separam os vinhos entre as três características. O algoritmo utilizado para a criação desta árvore está disponível na biblioteca *scikit-learn* (PEDREGOSA et al., 2011; BUITINCK et al., 2013). Já o código-fonte que utiliza este algoritmo está disponível no GitHub¹ (FUHR, 2021). Os dados estão disponíveis gratuitamente em (DUA; GRAFF, 2017).

O conjunto de dados de vinhos consiste em um problema de classificação multiclasse, tendo três classes diferentes. A distribuição de dados entre as três classes é de 59, 71 e 48, tendo no total 178 dados. Sendo treze características que descrevem cada exemplar de vinho. Para o exemplo de classificação de vinhos, a DT obtida por meio do algoritmo é mostrada pela Figura 2.1. A raiz e cada nó de decisão possui quatro informações, conforme mostrado na Figura 2.1. A primeira é a regra de decisão sendo avaliada, como por exemplo em '*color_intensity < 3.82*'. O nó também exibe o número de observações que atendem todas as regras de decisão até aquele nó, como em '*samples = 142*'. As duas últimas informações, mostram respectivamente, a distribuição do número de observações entre as três classes e a classe mais representativa das observações.

Os nós folhas contém, respectivamente, o número de observações, distribuição das classes e a classe predita. Os dados apresentados representam um problema de classificação, em que os dados de entrada descrevem os vinhos através de várias características. Os vinhos por sua vez, são separados em três classes e a classe é o dado de saída.

¹<https://github.com/gtfuhr/tcc>

Figura 2.1 – Representação gráfica da Árvore de Decisão para a classificação de vinhos.



Fonte: Próprio autor.

Conforme mostrado pela Figura 2.1, observa-se um caminho lógico a ser seguido na árvore, caso um novo vinho venha a ser classificado. Cada nó da árvore contém uma pergunta que pode ser verdadeira ou falsa. Conforme o resultado da pergunta, o algoritmo continua para a esquerda ou direita. Caso o caminho leve até outro nó intermediário, o processo de decisão se repete. Quando o algoritmo atinge uma folha da árvore, o nó alcançado indicará qual a classe do vinho está sendo predita.

Para o exemplo citado, os nós folhas contêm valores de classe ou valores discretos. Por outro lado, caso a DT fosse de Regressão, os nós folhas conteriam valores quantitativos preditos para a instância.

A DT é um algoritmo simples, mas que serve como base para outros algoritmos de ML, como será visto mais adiante. Entre as vantagens, destaca-se a possibilidade de se utilizarem dados nominais e até mesmo tolerar a ausência de valores (SOMVANSHI et al., 2016). A facilidade de interpretação e o rápido ajuste são outras vantagens da DT, sendo indicada para processos que requerem o uso de mínima memória (MEKONNEN et al., 2020).

Além disso, a árvore de decisão se destaca por sua explicabilidade, ou seja, sua elevada capacidade de ser auto explicativa, característica esta que é herdada por algoritmos que a utilizam como base.

2.2.7.3 Floresta Aleatória

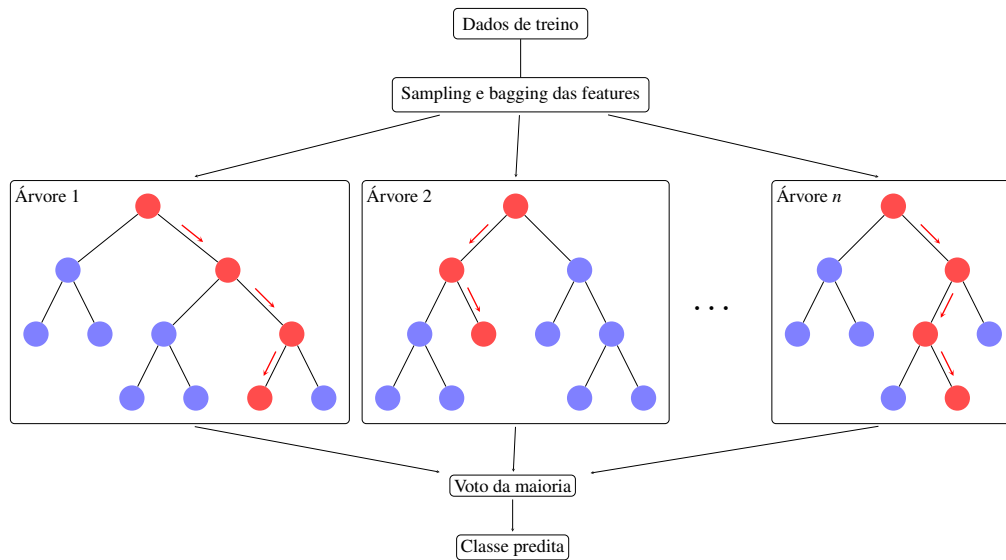
As Florestas Aleatórias (do inglês *Random Forests*) são combinações de DTs (CHEN et al., 2011; BREIMAN, 2001). Assim como as DTs, as RFs são um algoritmo de aprendizado supervisionado (MORE; RANA, 2017). Uma RF é um algoritmo de *bagging* que combina a saída de diversas DTs para fornecer uma previsão. Cada árvore da floresta é criada selecionando aleatoriamente um subconjunto de atributos a considerar durante cada nó construído. Dessa forma, cada árvore se torna única, e a totalidade das árvores explora uma maior quantidade de relações entre os dados de entrada e saída. Para a obtenção de uma previsão final, é feita uma eleição da classe por maioria dos votos no caso de classificação (AYYADEVARA, 2018).

Uma floresta aleatória genérica é mostrada na Figura 2.2, em que diversas DTs inicializadas randomicamente são utilizadas em conjunto. Cada árvore toma uma decisão dada uma nova entrada. O caminho de cada decisão de cada árvore pode ser observado pelos nós e setas vermelhas indicados na Figura 2.2. Todas as árvores da RF participam no processo de decisão através de um processo de votação. A RF seleciona uma classe através do voto da maioria das árvores.

O processo de eleição é feito sempre que um Floresta Aleatória tem de fazer uma previsão, nesse processo, todas as Árvores de Decisão escolhem uma classe, e a classe mais votada é eleita como a previsão da Floresta Randômica. O desempenho da RF converge para um limite, conforme o número de DTs utilizadas aumenta e a floresta usa a seleção aleatória de características para criar nós de decisão. (BREIMAN, 2001)

A vantagem de se utilizar uma RF é a possibilidade de se trabalhar com valores ausentes e valores atípicos (MORE; RANA, 2017). Além disso, a RF melhora o poder preditivo em relação a DT, pois introduz elementos aleatórios no modelo. A capacidade preditiva de uma RF está relacionada com a correlação entre as árvores da floresta e com o poder preditivo de cada árvore. Uma correlação elevada entre as DTs eleva a taxa de erro da RF, e um maior poder preditivo por parte das DTs reduz o erro (CHEN et al., 2011).

Figura 2.2 – Floresta aleatória genérica.



Fonte: Adaptado de (ILLUSTRATING..., 2019).

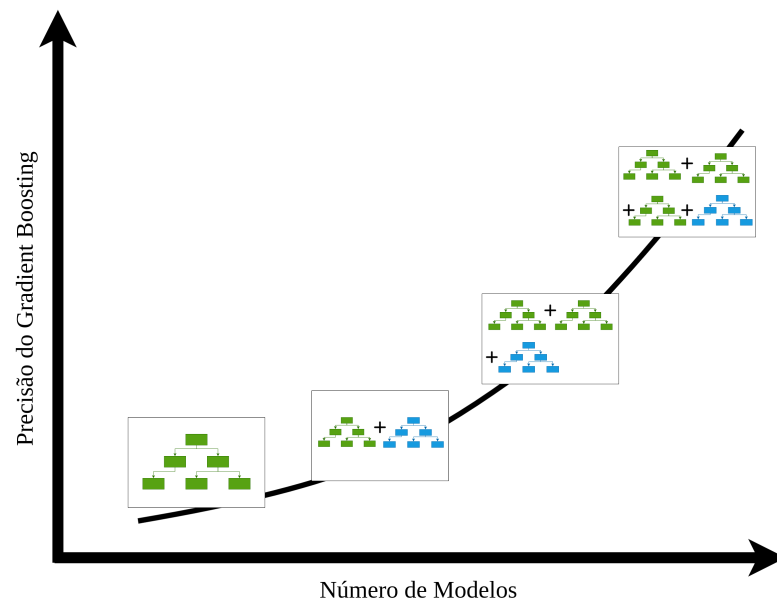
2.2.7.4 Gradient Boosting

O *Gradient Boosting* (GB), diferentemente das RF, faz previsões adotando uma construção sequencial de DTs para obter as previsões. Dessa forma, cada DT tenta corrigir o erro da árvore anterior, o que melhora o poder preditivo do modelo (AYYADEVARA, 2018). Assim como as DTs e RFs, trata-se de um algoritmo de aprendizado supervisionado (LU et al., 2019).

Para o algoritmo de Gradient Boosting realizar uma predição, os dados de entrada são passados para todos os modelos treinados. O primeiro modelo terá a sua saída, que dirá qual a probabilidade inicial de que o dado de entrada pertence a classe X, Y ou Z. Os próximos modelos irão adicionar as suas saídas, comumente chamadas de resíduos, na probabilidade inicial, ao final do processo a probabilidade de pertencimento em cada classe pode ser obtida.

O conceito de modelo aditivo pode ser observado pela Figura 2.3, em que o desempenho do modelo converge a um limite, de acordo com o aumento do número de DTs utilizadas. O eixo X da Figura 2.3 representa o número de modelos, e o eixo Y representa o seu desempenho.

Figura 2.3 – Funcionamento do Gradient Boosting.



Fonte: Próprio autor.

O GB é um algoritmo poderoso capaz de prever conjuntos de dados heterogêneos, altamente correlacionados, perdidos e até mesmo categóricos, levando a modelo interpretáveis por meio da construção de um modelo aditivo (LU et al., 2019).

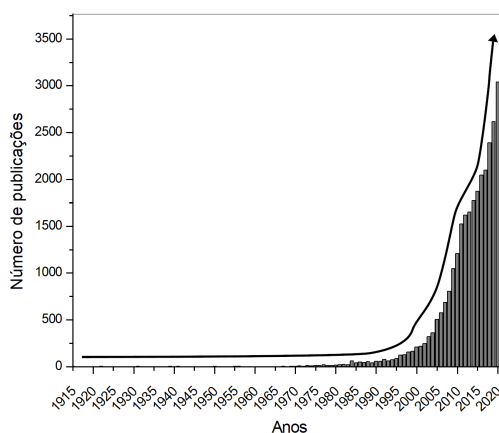
3 TRABALHOS RELACIONADOS

A presente pesquisa se valeu da base de dados SCUPUS da editora Elsevier para buscar informações sobre trabalhos relacionados. Estas informações foram buscadas através da pesquisa por termos no mecanismo de busca da SCOPUS. A ferramenta SCOPUS retornou dados como o número de publicações sobre o termo pesquisado ao longo dos anos. Vale ressaltar que os números obtidos representam apenas uma amostra de todas as publicações disponíveis e têm como objetivo auxiliar o entendimento sobre os trabalhos relacionados. Outras bases além da SCOPUS podem ser utilizadas para recolher amostras sobre o número de trabalhos científicos.

Os dados foram apresentados seguindo a metodologia de Petersen (PETERSEN et al., 2008). Que orienta os pesquisadores em definir uma questão para a pesquisa e conduzir uma pesquisa sobre os artigos relevantes para a questão levantada. Com isso, é realizada uma extração de dados sobre a relevância da questão utilizando a extração de palavras-chaves dos artigos relacionados, no presente trabalho, foi utilizada como ferramenta a base de dados SCOPUS para o levantamento de dados sobre o número de publicações.

Nas últimas décadas, houve um aumento do número de publicações a respeito da Classificação de Crédito, de acordo com a base de dados da SCOPUS (Elsevier), conforme mostra o Gráfico 3.1. Observa-se que trabalhos científicos já exploram o tema *Credit Rating* desde 1915. As publicações acerca do assunto vêm crescendo intensamente desde meados de 1985 e até o momento, seu ápice de publicações está em 2021. O aumento no número de publicações demonstra que este é um assunto relevante.

Gráfico 3.1 – Publicações de *Credit Rating*.

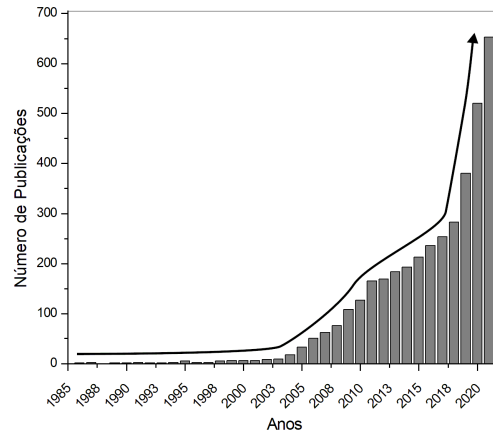


Fonte: Próprio autor. Dados obtidos pesquisando por (credit AND rating) no banco de dados SCOPUS (Elsevier) (DATABASE, 2021).

O termo *Machine Learning* (ML) pode ser encontrado em trabalhos científicos sobre Classificação de Crédito (CR) desde 1985, conforme mostra o Gráfico 3.2. A partir do ano

2000, observa-se um rápido aumento na aparição do termo ML em publicações de CR. Isso demonstra que é uma abordagem que vem sendo amplamente adotada.

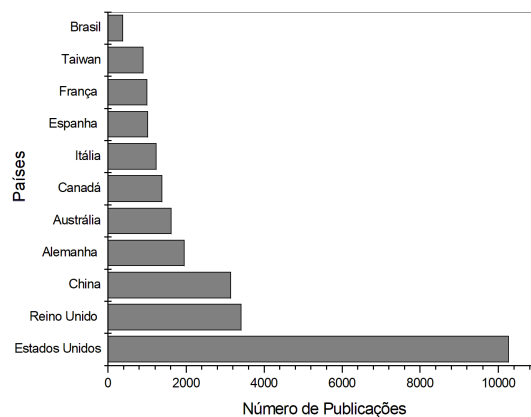
Gráfico 3.2 – Publicações de *Credit Rating* que incluem o termo *Machine Learning*



Fonte: Próprio autor. Dados obtidos pesquisando (credit AND rating) AND (machine AND learning) no banco de dados SCOPUS (Elsevier) (DATABASE, 2021).

A distribuição de publicações agrupadas de acordo com países de origem, conforme o Gráfico 3.3, apresenta os Estados Unidos com o maior número de publicações. Em relação as publicações nacionais, o Brasil se encontra em 21^o. Nota-se que o Brasil ainda tem um longo caminho a percorrer para se estabelecer entre os países que mais produzem conhecimento científico sobre a área.

Gráfico 3.3 – Publicações de *Credit Rating* agrupadas por países. Os dez países com mais publicações são exibidos junto com o Brasil.

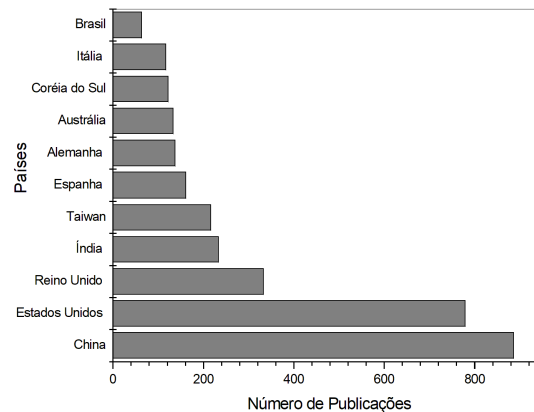


Fonte: Próprio autor. Dados obtidos pesquisando por (credit AND rating) no banco de dados SCOPUS (Elsevier) (DATABASE, 2021).

Em relação ao uso de ML em CR, pelo Gráfico 3.4, podemos ver que os Estados Unidos

ainda permanecem como os maiores pesquisadores da área. Já o Brasil aparece ainda como décimo nono colocado, o que mostra que, o uso de ML tem oportunidade para crescimento nas publicações nacionais de CR.

Gráfico 3.4 – Publicações de *Credit Rating* que incluem o termo *Machine Learning* agrupadas por países. Os dez países com mais publicações são exibidos junto com o Brasil.



Fonte: Próprio autor. Dados obtidos pesquisando (credit AND rating) AND (machine AND learning) no banco de dados SCOPUS (Elsevier) (DATABASE, 2021).

3.1 USO DE *MACHINE LEARNING* NO *CREDIT RATING*

Como visto no início do capítulo 3, houve um aumento do uso de ML em CR. Dentre essas publicações, pode-se observar que algoritmos com maior poder explicativo são explorados para a obtenção do CR, haja vista que a Classificação de Crédito possui um impacto real na economia.

A partir da seção 3.1.1, publicações que exploram o uso de ML em CR serão apresentadas e discutidas, mostrando exemplos de como essas técnicas compreensíveis podem ser utilizadas no setor de análise de crédito.

3.1.1 **Árvore de Decisão**

Devido as características do algoritmo, as DTs podem ser utilizadas em diversos tipos de problemas e áreas, como engenharia, saúde e até mesmo no reconhecimento de números manuscritos como visto em (CHARBUTY; ABDULAZEEZ, 2021). Em Shahbazi (2020) (SHAHBAZI, 2020), dados bancários de pessoas foram explorados por meio de DT. Para o CR do setor bancário, os autores utilizaram uma abordagem em que foram selecionadas variáveis por

meio de uma DT e, os dados foram separados em treino e teste. Os resultados mostraram que a abordagem de DT para o setor bancário obteve elevada precisão para o CR (SHAHBAZI, 2020).

Yunyan, Weiwei e Quijiu (2021) (YUNYAN; WEIWEI; QIUJU, 2021) utilizaram uma DT para o CR no setor de finanças e comércio eletrônico, em que os dados de crédito utilizados eram de uma plataforma de empréstimos online. Os autores construíram uma árvore que minimizou os riscos de crédito e maximizou os retornos da operação da plataforma estudada (YUNYAN; WEIWEI; QIUJU, 2021).

Outro estudo, realizado por Chern et al. (2021) (CHERN et al., 2021), propôs o uso de DTs para o CR, explorando a explicabilidade do modelo, devido as regras serem de fácil compreensão. O estudo utilizou dados de crédito de uma das maiores empresas de empréstimos Taiwanesa, obtendo um modelo simples e com boa interpretabilidade para CR (CHERN et al., 2021). Dessa maneira o uso de DTs para CR é uma escolha segura e viável.

3.1.2 Floresta Aleatória

As RFs são amplamente utilizadas no setor de finanças, desde a predição do movimento de instrumentos no setor financeiro. Em Kumar (2006) (KUMAR; THENMOZHI, 2006), as RFs são utilizadas para a predição de movimentos em ações. Já em Zou (2015), RFs são usadas para a previsão de movimentos em contratos futuros no setor financeiro (ZOU; PENG; LUO, 2015). Yeh (2012) (YEH; LIN; HSU, 2012) utilizou RF para o CR, em que as informações contábeis de corporações foram utilizadas como dados de entrada para o CR dessas corporações. O estudo explorou o uso de informações disponíveis ao mercado, e mostrou que as informações financeiras ajudam na determinação do CR, quando utilizadas por RFs (YEH; LIN; HSU, 2012).

Wang (2021), por sua vez, também utilizou RF em CR de corporações. O trabalho de Wang (2021) ressalta que técnicas de inteligência artificial (IA) têm obtido resultados positivos para CR. O estudo comparou redes neurais artificiais e RFs. Como resultado, os autores obtiveram que as RFs atingiram os melhores desempenhos no CR das corporações de crédito usando demonstrativos financeiros como variáveis de entrada. Dessa forma, as RFs também obtiveram resultados competitivos quando os demonstrativos financeiros passados ao algoritmo representavam uma linha de tempo de três anos (WANG; KU, 2021). Sendo assim, a aplicação de RF para CR se mostra como um grande potencial para o setor financeiro.

3.1.3 Gradient Boosting

O uso de GB também vem sendo explorado no setor financeiro. Tian (2020), apresentou o GB como um método eficiente de CR, quando foi comparado com outras técnicas de ML como DT, *Support-Vector Machines* e *Multilayer Perceptron*. Os resultados obtidos mostraram

o GB como sendo a melhor técnica (TIAN et al., 2020), quando utilizados dados de crédito originados de uma empresa que faz o CR de empresas. Já Chang (2018) (CHANG; CHANG; WU, 2018) utilizou GB para o CR de instituições financeiras. A aplicação de GB foi comparada com regressão logística. O GB obteve resultados superiores e demonstrou ser uma técnica válida para a o desenvolvimento de modelos de CR (CHANG; CHANG; WU, 2018).

Em Youlve et al. (2021)(YOULVE; KAIYUN; JIANGTIAN, 2021), o estudo utilizou a técnica de pré-processamento de dados chamada PCA para diminuir a dimensionalidade dos dados de entrada e fez uso de GB para criar um modelo de CR. O estudo concluiu que a aplicação de ML no CR tem potencial, e mostrou que os melhores resultados foram obtidos utilizando GB, quando comparados com *Support-Vector Machines* e DTs. Os dados de crédito sobre empréstimos utilizados no estudo foram fornecido pelo *National Bureau of Statistics*. (YOULVE; KAIYUN; JIANGTIAN, 2021).

A interpretabilidade do GB também é explorada pelo estudo (ZHOU et al., 2021), onde dados de crédito de uma empresa de empréstimos entre pessoas são usados para avaliar o uso de GB no CR. No estudo de (ZHOU et al., 2021), o GB mostra desempenho positivo.

Pode-se observar o uso do GB para a modelagem de risco de falência em Carmona et al. (2019). Os dados de falência usados para a previsão de falência são do setor bancário americano. O estudo mostrou melhores resultados do GB, quando comparado a DTs e RFs. A interpretabilidade dos modelos também foi explorada através de uma discussão acerca das variáveis eleitas pelos modelos como mais importantes no processo de previsão de falência. Pode-se concluir que o lucro antes dos impostos e os lucros acumulados são ótimos indicadores da saúde financeira de uma empresa (CARMONA; CLIMENT; MOMPARDLER, 2019). Dessa maneira, assim como as DTs e RFs, o GB mostra grande potencial para o setor financeiro.

3.2 CONSIDERAÇÕES SOBRE OS TRABALHOS RELACIONADOS

Após o estudo de trabalhos relacionados, pode-se observar a baixa reprodutibilidade dos estudos encontrados, pelo uso de dados pagos pelos autores e inacessíveis aos leitores. A baixa reprodutibilidade pode impedir que as técnicas de modelagem sejam comparadas utilizando o mesmo conjunto de dados, dificultando o avanço na área.

Além disso, poucos trabalhos exploram a importância das características encontradas pelos modelos de ML citados a partir da seção 2.2.7.2. Nota-se que a extração da importância das variáveis de entrada, que pode ser obtida a partir das abordagens com maior explicabilidade, como DTs, RFs e o GB, poderia ser melhor explorada.

4 MATERIAIS E MÉTODOS

4.1 OBTENÇÃO DE DADOS FINANCEIROS DE EMPRESAS

Os algoritmos de ML supervisionados necessitam de variáveis de entrada e saída para o seu treinamento. Dados prontos para o uso em algoritmos de ML não costumam estar disponíveis gratuitamente, como visto na Seção 3.2.

4.1.1 Dados de entrada

Uma pesquisa sobre possíveis dados de entrada contendo informações sobre empresas de capital aberto dos EUA foi realizada, visto que o país possui regulações consolidadas sobre a transparência e o fácil acesso a dados.

A prospecção por dados públicos de demonstrativos financeiros de empresas com capital aberto foi realizada por meio da base de dados online *Financial Statements Data Sets*, publicada pela Comissão de Valores Mobiliários dos Estados Unidos. Os dados estão disponíveis no formato XBRL, do inglês *Extensible Business Reporting Language*, formato padrão para todas as informações de empresas que são reguladas pela Comissão de Valores Mobiliários dos Estados Unidos.

A linguagem XBRL não teve uma forte adesão em instituições financeiras brasileiras, e pouco material sobre a linguagem está disponível em Português. Para o processamento dos dados em XBRL, um *script* em *Python* foi utilizado e está disponível na plataforma Github (FUHR, 2021).

4.1.2 Dados de saída

Para os dados de saída, dados de CR sobre as empresas foram obtidos utilizando os dados históricos de Classificação de Crédito publicado pelas três maiores agências de crédito, Standard & Poor's, Moody's, e Fitch Group. Os dados são publicados seguindo a norma 17g-7 (Council of European Union, 2014).

As classificações de crédito feitas pelas agências de crédito são disponibilizadas pelo *Center of Municipal Finance*, uma organização sem fins lucrativos que busca aumentar a qualidade do mercado de títulos de crédito. Para a busca automatizada desses dados foi utilizado um *script* que se encontra disponível em (FINANCE, 2021).

Para serem usadas como dados de saída, as classificações de crédito foram padronizadas em notas de um a seis usando a Tabela 2.1. Para realizar a padronização, foi utilizada a coluna *EU Credit Quality Step* que mapeia diferentes notas de crédito para uma nota específica.

4.1.3 Problemas de Classificação

Neste trabalho, dois problemas distintos de classificação são considerados: Classificação binária e classificação multiclasse.

Para a classificação binária, as notas de crédito de um a três foram mapeadas como Investimento, enquanto que as notas de quatro a seis foram mapeadas como Especulativa. O mapeamento das notas de crédito das agências de crédito em notas de um a seis pode ser vista na tabela 2.1.

De acordo com essa classificação, os demonstrativos financeiros obtidos são compostos por:

- 24250 instâncias de Instrumentos de Investimento.
- 19771 instâncias de Instrumentos Especulativos.

Para a classificação multiclasse, cada uma das seis notas será categorizada como uma classe distinta.

De acordo com essa classificação, os demonstrativos financeiros obtidos são compostos por:

- 1819 instâncias com a nota um (1).
- 6363 instâncias com a nota dois (2).
- 16068 instâncias com a nota três (3).
- 10636 instâncias com a nota quatro (4).
- 6141 instâncias com a nota cinco (5).
- 2994 instâncias com a nota seis (6).

No total, foram 44021 instrumentos disponíveis no conjunto de dados.

4.1.4 Associação entre dados de entrada e saída

Os dados de entrada (demonstrativos financeiros) e de saída (notas de crédito) foram associados através do nome das empresas. Ou seja, os demonstrativos financeiros referentes a uma empresa são aproveitados apenas caso haja notas de crédito correspondentes a essa empresa. Além disso, o ano da coleta também é usado para atrelar os dados de entrada com os de saída. Ou seja, os demonstrativos financeiros de uma empresa coletados em um ano específico são atrelados à última Classificação de Crédito disponível para o ano considerado.

A associação entre entrada e saída foi realizada utilizando a linguagem *Python* e está disponível na plataforma Github (FUHR, 2021).

Após a associação e a remoção de variáveis com dados faltantes, os dados financeiros de entrada se reduziram às seguintes 37 variáveis:

- Ativos;
- Passivo;
- Ganho/Lucro;
- Núm. de Ações;
- Gastos com Juros;
- Ativos correntes;
- Patrimônio Líquido;
- Passivo Circulante;
- Valor da ação comum;
- Patrimônio de marca;
- Remuneração em Ações;
- Ativos a longo prazo;
- Caixa ou Equivalentes;
- Custos com transações;
- Núm. de Ações criadas;
- Patrimônio permanente;
- Crédito a curto prazo;
- Perda de renda líquida;
- Passivos a longo prazo;
- Despesa/Benefício do IR;
- Lucro/Prejuízo por Ação;
- Depreciação/Amortização;
- Aumento/Redução de caixa;
- Aumento/Redução do Lucro;
- Ganhos com investimentos;
- Ganho/Perda com operações;
- Valor declarado de sócios;
- Propriedade e Equipamentos;
- Perda de lucro por operação;
- Núm. Máx de ações de sócios;
- Aumento/Redução de impostos;
- Ganho/Perda com financiamento;
- Valor declarado de ação comum;
- Investimentos em propriedades;
- Lucro/Prejuízo por Ação Diluída;
- Aumento/Redução Patrimônio Líquido;
- Soma do Passivo e Patrimônio Líquido.

4.2 MODELOS DE APRENDIZADO DE MÁQUINA

Nessa etapa, foram analisados três modelos de ML que possuem alta explicabilidade em suas tomadas de decisão, sendo eles DT, RF e GB. A linguagem de programação utilizada foi *Python*. Os dados de entrada e saída estão descritos na Seção 4.1. Os dados foram separados entre treino e teste, sendo 70% para treino e 30% para teste. A função utilizada para a separação randômica dos dados entre treino e teste é oriunda da biblioteca *scikit-learn* (PEDREGOSA et al., 2011).

Com os dados de treino é feita a otimização de parâmetros de forma randômica. Para isso, foi utilizado o método *RandomizedSearchCV* da biblioteca *scikit-learn* (PEDREGOSA et al., 2011). O *RandomizedSearchCV* faz uma busca aleatória das melhores combinações de parâmetros utilizando a Validação Cruzada (CV) para mensurar o desempenho da combinação sendo testada. A CV ajuda a escolher parâmetros para o modelo, pois escolhe a combinação de parâmetros do modelo que melhor desempenham utilizando múltiplas combinações de conjuntos de treino e teste diferentes.

Para cada modelo, os dados de treino e teste são normalizados utilizando diferentes técnicas de normalização de dados numéricos da biblioteca *scikit-learn*. Três abordagens são utilizadas:

- **Originais:** Os dados numéricos originais foram utilizados como entrada na primeira abordagem.
- **Normalização Z-Score** implementada com o nome de *StdScaler* da biblioteca *scikit-learn* (PEDREGOSA et al., 2011). Na normalização dos dados utilizando o Z-Score ou *StdScaler*, os dados de treino são estudados, encontrando o desvio padrão e a média de cada variável. A média é subtraída dos valores. O resto da subtração é então dividido pelo desvio padrão, normalizando os dados. O *StdScaler* com as médias e desvios padrões calculados é então aplicado aos dados de teste. Os dados resultantes são processados pelo algoritmo na segunda abordagem.
- ***Robust Scaler*** da biblioteca *scikit-learn* (PEDREGOSA et al., 2011). Na normalização de dados utilizando o *Robust Scaler*, os dados de treino entre o primeiro e terceiro quartil são estudados, encontrando o desvio padrão e a média de cada variável. A técnica é mais resistente a valores atípicos, do inglês, *outliers*, pois apenas os quartis centrais dos dados são considerados. A média é subtraída dos valores. O resto da subtração é então dividido pelo desvio padrão, normalizando os dados. O *StdScaler* com as médias e desvios padrões calculados é então aplicado aos dados de teste. Os dados resultantes são processados pelo algoritmo na terceira abordagem.

4.2.1 Árvores de Decisão

O algoritmo de DT utilizado está presente na biblioteca *scikit-learn* (PEDREGOSA et al., 2011). O campo de parâmetros explorado para encontrar as melhores combinações é composto por cada variável de ajuste do algoritmo e os diferentes valores que podem ser utilizados durante o treino. O espaço a ser otimizado pode ser visto a seguir:

- 'criterion': ['gini', 'entropy'];
- 'max_features': ['auto', 'sqrt', 'log2', None];
- 'max_depth': [3, 4, 5, 8, 10, 12, 15, 20, None];
- 'min_samples_split': [5, 10, 15, 20, 30, 0.01, 0.02, 0.03, 0.04, 0.06, 0.08, 0.1];
- 'min_samples_leaf': [3, 4, 6, 8, 10, 15, 0.01, 0.02, 0.03, 0.04, 0.05];
- 'splitter': ['best', 'random'];
- 'max_leaf_nodes': [5, 10, 15, 20, 25, None];

4.2.2 Floresta Aleatória

A implementação de RF utilizada está presente na biblioteca *scikit-learn* (PEDREGOSA et al., 2011). O campo de parâmetros explorado para encontrar os melhores parâmetros é composto por cada parâmetro do algoritmo e os diferentes valores que o mesmo pode assumir durante o treino. O espaço a ser otimizado pode ser visto a seguir:

- 'n_estimators': [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]
- 'bootstrap': [True, False]
- 'warm_start': [True, False]
- 'criterion': ['gini', 'entropy']
- 'max_depth': [3, 5, 8, 10, 15, 20, 25, 30, None]
- 'min_samples_leaf': [3, 4, 6, 8, 10, 15, 20, 25, 30, 0.01, 0.02, 0.03, 0.04, 0.05]
- 'max_leaf_nodes': [5, 10, 15, 20, 25, 30, 40, None]
- 'min_samples_split': [5, 10, 15, 20, 30, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1]
- 'max_features': ['auto', 'sqrt', 'log2', None]

4.2.3 Gradient Boosting

O algoritmo de GB utilizado está presente na biblioteca *xgboost* (CHEN; GUESTRIN, 2016). O campo de parâmetros explorado para encontrar os melhores parâmetros é composto por cada parâmetro do algoritmo e os diferentes valores que o mesmo pode assumir durante o treino. O espaço a ser otimizado pode ser visto a seguir:

- 'min_child_weight': [1, 2, 3, 4, 5, 6, 8, 10]
- 'gamma': [0.5, 1, 1.5, 2, 5]
- 'subsample': [0.6, 0.8, 1.0]
- 'colsample_bytree': [0.6, 0.8, 1.0]
- 'max_depth': [3, 4, 5]
- 'n_estimators': [100, 200, 300, 400, 500, 600]

4.2.4 Análise estatística

A avaliação do desempenho e capacidade preditiva dos modelos foi realizada de acordo com a precisão, *Recall* e a pontuação F1. A métrica otimizada durante a otimização de parâmetros foi o F1-Score, ao invés da Precisão, pois viu-se que a métrica busca um equilíbrio entre a Precisão e o *Recall*. O uso de *Recall* é recomendado para problemas onde a distribuição dos valores da variável de saída é desbalanceada, segundo (BROWNLEE, 2016).

5 RESULTADOS

A seguir são mostrados os resultados obtidos através de modelos de Árvore de Decisão, Floresta Aleatória e *Gradient Boosting* para o problema de classificação binária e multiclasse. A análise de desempenho pode ser vista na seção 5.1, já a exploração da explicabilidade da DT, RF e GB é feita na seção 5.2.

5.1 ANÁLISE DOS ALGORITMOS

Os algoritmos testados tiveram seu desempenho avaliado em dois diferentes cenários, sendo o primeiro a classificação binária e o segundo a classificação multiclasse. O desempenho foi analisado utilizando as métricas F1, *Recall* e Precisão, tanto para os dados de treino quanto os de teste. Cada processamento da otimização de parâmetros durou cerca de dez horas para a Classificação Binária e onze horas para a classificação Multiclasse.

5.1.1 Classificação Binária

Recapitulando, a classificação binária divide os degraus de classificação entre dois conhecidos grupos: *investment-grade*, ou instrumentos de investimento e *high-yield bonds*, ou instrumentos especulativos, como visto na tabela 2.1 e explicado na seção 2.1.2. Os algoritmos foram treinados e testados com o objetivo de classificar corretamente um demonstrativo financeiro entre essas duas classes.

Na Tabela 5.1 são mostrados os resultados das métricas de desempenho dos algoritmos no cenário de classificação binária. As linhas estão ordenadas pelo valor da métrica F1.

Tabela 5.1 – Desempenho da Classificação Binária.

Algoritmo	Normalização	Treino			Teste		
		F1	Precisão	Recall	F1	Precisão	Recall
Abordagem Ingênua	-	70,00%	54,90%	100,00%	71,00%	55,40%	100,00%
Árvore de decisão	Originais	82,10%	79,20%	86,90%	81,50%	78,40%	86,20%
Árvore de decisão	Standard Scaler	86,50%	85,40%	85,00%	85,40%	84,10%	84,00%
Floresta Aleatória	Robust Scaler	86,10%	84,30%	88,20%	85,40%	83,50%	87,40%
Floresta Aleatória	Originais	86,40%	84,70%	88,40%	85,80%	84,00%	87,80%
Floresta Aleatória	Standard Scaler	87,70%	86,20%	89,30%	86,90%	85,20%	88,40%
Árvore de decisão	Robust Scaler	94,40%	93,80%	93,80%	89,80%	88,70%	89,80%
Gradient Boosting	Standard Scaler	95,00%	94,40%	95,50%	91,70%	90,80%	92,60%
Gradient Boosting	Originais	94,70%	94,10%	95,30%	92,00%	91,10%	93,00%
Gradient Boosting	Robust Scaler	94,30%	93,70%	94,90%	92,00%	91,10%	92,90%

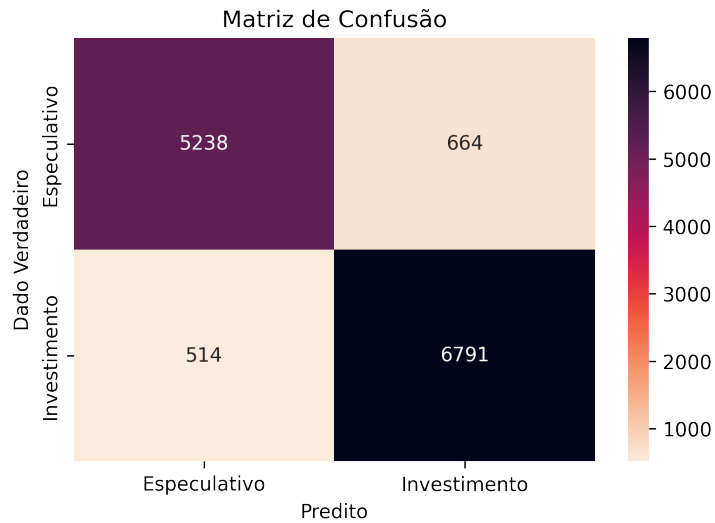
Todos os algoritmos testados obtiveram resultados superiores à abordagem ingênua. As Árvores de Decisão estão no início da escala de desempenho, sendo superadas pelas Florestas Aleatórias. Já o *Gradient Boosting* obteve os melhores resultados, o que se repetiu com as três diferentes abordagens de normalização. Isso mostra a flexibilidade do algoritmo em trabalhar com dados normalizados e não normalizados.

O modelo com o melhor desempenho na classificação binária, obteve os seguintes parâmetros eleitos como os melhores durante a otimização:

- 'min_child_weight': 5
- 'gamma': 1.5
- 'subsample': 1.0
- 'colsample_bytree': 0.6
- 'max_depth': 5
- 'n_estimators': 400

A matriz de confusão do Gráfico 5.1 exibe os erros e acertos da versão do algoritmo *Gradient Boosting* que obteve a melhor pontuação F1.

Gráfico 5.1 – Matriz de confusão com os dados de teste binários.



Os valores elevados da diagonal principal da matriz de confusão 5.1 demonstram um bom desempenho durante a classificação. Por exemplo, os valores 6791 e 5238 indicam instrumentos que foram corretamente classificados instrumentos Especulativos que foram classificados como instrumentos de investimento e especulativos, respectivamente.

A diagonal secundária da matriz 5.1, por sua vez, representa os erros do modelo. O canto inferior esquerdo indica que 514 instrumentos de Investimento foram classificados erroneamente como Especulativos. Já o canto superior direito indica que 664 instrumentos Especulativos foram classificados erroneamente como Investimento.

5.1.2 Classificação Multiclasse

Recapitulando, a classificação multiclasse utiliza os degraus de crédito, que vão de um a seis, como valores da variável de saída no treinamento dos algoritmos de Aprendizado de Máquina. Os algoritmos foram treinados e testados com o objetivo de classificar corretamente um demonstrativo financeiro entre essas seis classes. As classes podem ser vistas na tabela 2.1, na coluna *EU Credit Quality Step* e são exploradas na seção 2.1.2.

Na Tabela 5.2 são mostrados os resultados das métricas de desempenho dos algoritmos no cenário de classificação multiclasse. As linhas estão ordenadas pelo valor da métrica F1.

Tabela 5.2 – Desempenho da Classificação Multiclasse.

Algoritmo	Normalização	Treino			Teste		
		F1	Precisão	Recall	F1	Precisão	Recall
Abordagem Ingênua	-	19,40%	36,40%	36,40%	19,60%	36,60%	36,60%
Árvore de decisão	Robust Scaler	45,50%	49,00%	49,00%	44,80%	48,40%	48,40%
Árvore de decisão	Originais	46,80%	48,80%	48,80%	46,10%	48,10%	48,10%
Árvore de decisão	Standard Scaler	51,10%	52,50%	52,50%	49,30%	50,80%	50,80%
Floresta Aleatória	Robust Scaler	54,00%	56,40%	56,40%	53,00%	55,30%	55,30%
Floresta Aleatória	Standard Scaler	60,30%	61,80%	61,80%	58,20%	59,80%	59,80%
Floresta Aleatória	Originais	61,70%	63,20%	63,20%	58,80%	60,30%	60,30%
Gradient Boosting	Standard Scaler	78,90%	79,00%	79,00%	70,00%	70,30%	70,30%
Gradient Boosting	Originais	78,70%	78,80%	78,80%	70,10%	70,30%	70,30%
Gradient Boosting	Robust Scaler	78,30%	78,50%	78,50%	70,10%	70,40%	70,40%

As abordagens algorítmicas, novamente, superaram a abordagem ingênua. Tanto as Árvores de decisão quanto as Florestas Aleatórias obtiveram resultados modestos, sendo que o F1 não chegou à marca dos 60%. Já o *Gradient Boosting* novamente obteve os melhores resultado. Curiosamente, o desempenho não sofreu um impacto considerável ao mudar a abordagem de normalização.

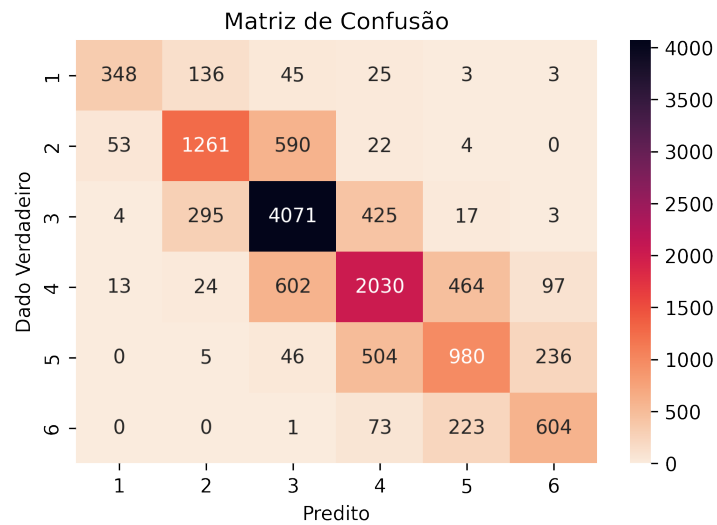
Além disso, este foi o algoritmo que mais sofreu perda de desempenho ao passar a usar a base de teste em vez da de treino. Essa observação levanta a hipótese de que alguns dos falso-positivos/falso-negativos encontrados sejam devido ao *overtraining* do modelo.

O modelo com o melhor desempenho na classificação multiclasse, obteve os seguintes parâmetros eleitos como os melhores durante a otimização:

- 'min_child_weight': 1
- 'gamma': 2
- 'subsample': 1.0
- 'colsample_bytree': 0.6
- 'max_depth': 5
- 'n_estimators': 300

Para melhor entendimento do desempenho do *Gradient Boosting*, a matriz de confusão do Gráfico 5.2 descreve detalhadamente os erros e acertos do modelo.

Gráfico 5.2 – Matriz de confusão com os dados de teste multiclasse.



A matriz de confusão mostra que a diagonal principal contém os valores mais elevados, o que indica a acertabilidade do modelo em relação a classificação em todas as classes. A concentração dos erros ao redor da diagonal principal mostra que o modelo cometeu mais erros leves do que erros graves. Isso porque, classificar uma empresa de *rating* 1 como 2 é um erro menos grave do que classificar uma empresa de *rating* 1 como 6.

5.2 EXPLICABILIDADE DOS ALGORITMOS

Com o uso de Aprendizado de Máquina, pode-se inferir, para cada algoritmo, quais características são mais importantes no processo de Classificação de Crédito.

Para obter as características mais importantes para cada algoritmo, é necessário utilizar a API da biblioteca *scikit-learn*. Esta biblioteca disponibiliza o atributo *feature_importances* que contém a lista de variáveis de entrada e sua determinada importância para o modelo gerado. A biblioteca calcula estes valores com base no processo de treinamento dos modelos baseados em DT, que é o caso de RF e GB. Quanto mais uma característica é utilizada para dividir os dados em conjuntos homogêneos, maior é a sua importância.

As próximas seções descrevem as características mais importantes para o *Gradient Boosting*, uma vez que este foi o algoritmo que obteve os melhores resultados.

5.2.1 Classificação binária

A Tabela 5.3 mostra as características mais importantes descobertas durante o treinamento para a classificação binária.

Tabela 5.3 – Variáveis mais importantes na Classificação Binária.

Variável	Descrição	Importância
Aumento/Redução do Lucro	Aumento ou redução de lucro.	15,69%
Lucro/Prejuízo por Ação Diluída	Lucro/Prejuízo por ação diluída.	9,28%
Despesa/Benefício do IR	Despesa referente a impostos de operação.	6,62%
Soma do Passivo e Patrimônio Líquido	Soma do Passivo e Patrimônio Líquido	6,23%
Aumento/Redução Patrimônio Líquido	Aumento ou redução do patrimônio líquido.	5,97%
Valor da ação comum	Valor calculado de uma ação comum.	3,22%
Patrimônio Líquido	Diferença entre o valor de ativo e passivo.	3,20%
Valor declarado de ação comum	Valor declarado de uma ação comum.	2,87%
Núm. de Ações	Número de ações da entidade avaliada.	2,84%
Patrimônio permanente	Patrimônio Líquido menos o temporário.	2,66%

A tabela 5.3 indica que as dez características mais importantes detém 58,58% de importância no processo de decisão. Em outras palavras, mais da metade do processo de decisão é guiada por apenas dez características, o que ajuda a identificar a importâncias dessas informações.

5.2.2 Classificação multiclasse

A Tabela 5.4 mostra as características mais importantes descobertas durante a o treinamento para a classificação multiclasse.

Tabela 5.4 – Variáveis mais importantes na Classificação Multiclasse.

Variável	Descrição	Importância
Aumento/Redução do Lucro	Posses ou bens da empresa.	9,31%
Despesa/Benefício do IR	Despesa referente a impostos de operação.	7,21%
Patrimônio Líquido	Diferença entre o valor de ativo e passivo.	5,00%
Lucro/Prejuízo por Ação	Lucro/Prejuízo por ação.	4,71%
Lucro/Prejuízo por Ação Diluída	Lucro/Prejuízo por ação diluída.	4,54%
Ativos	Posses ou bens da empresa.	4,22%
Aumento/Redução Patrimônio Líquido	Aumento ou redução de lucro.	3,93%
Patrimônio permanente	Patrimônio Líquido menos o temporário.	2,83%
Núm. de Ações	Número de ações da entidade avaliada.	2,77%
Soma do Passivo e Patrimônio Líquido	Soma do Passivo e Patrimônio Líquido	2,70%

Pode-se notar, na tabela 5.4, que as dez características mais importantes fazem parte de 47,22% do processo de decisão. Ou seja, aproximadamente metade do processo de decisão é afetada por apenas dez características, o que ressalta a importâncias dessas informações.

5.2.3 Análise das características para as classificações

5.2.3.1 Classificação Binária

Conforme mostra a Tabela 5.3, a variável mais importante para o CR binária foi o aumento/redução do lucro. O lucro/prejuízo por ação diluída ficou em segundo lugar, e é um indicador amplamente utilizado para a saúde financeira de uma corporação, conforme visto em (FERNANDO; KINDNESS, 2021). Já a característica referente a passivo e patrimônio líquido é importante, pois remete aos endividamentos e as posses que podem ser facilmente liquidadas, e isto reflete a capacidade da empresa de quitar dívidas. A variação do Patrimônio Líquido da empresa também aparece como importante para a tomada de decisão do modelo, e esta informação está associada ao estado de retração ou expansão do crescimento.

Valores relacionados às ações das empresas aparecem em seguida, mostrando que o modelo utiliza o conhecimento e opinião do mercado financeiro sobre o valor e saúde financeira das corporações.

5.2.3.2 Classificação Multiclasse

Conforme mostra a Tabela 5.4, a variável mais importante para o CR multiclasse foi o aumento/redução do lucro. O mesmo ocorreu na classificação binária, o que mostra que a capacidade de uma empresa ter lucro pode ser um forte indicador sobre a sua capacidade de crédito, o que fica alinhado com os conceitos vistos em (SEC; COMMISSION, 2017). Outras variáveis que foram relevantes para a classificação binária também surgiram como importantes na classificação multiclasse, o que reforça a sua relevância. O gasto com impostos também aparece como indicativo de poder de crédito. O patrimônio líquido e o lucro/prejuízo por ações aparecem em seguida, que são indicadores de liquidez e lucratividade, respectivamente, de uma companhia. Dentre as variáveis apresentadas, encontra-se também o valor de Ativos, que representa as posses e direitos econômicos da empresa.

5.2.3.3 Características em comum da classificação Binária e Multiclasse

Pode-se observar nas Tabelas 5.3 e 5.4, que oito características são comuns entre as dez mais importantes na classificação binária e na multiclasse. As variáveis importantes em comum em ambas as classificações, demonstram que a classificação binária e multiclasse possuem características comuns. As características em comum em ambas as classificações já eram esperadas, visto que, as classes binárias são derivadas das classes usadas pelo problema que envolve múltiplas classes.

Em primeiro lugar, nas duas classificações, há um foco na manutenção do lucro da empresa, que indica se uma empresa está se mantendo rentável ou não. A Despesa ou Benefício com impostos é um indicador referente à quanto as operações da empresa estão sendo taxadas. Quanto maior esse valor, maior a receita da empresa sendo avaliada pelos modelos.

O Patrimônio Líquido aparece como valor fixo e também com a sua variação, o que demonstra que os modelos estão analisando o tamanho atual e também a expansão de operações ou a retração da empresa. Nota-se também a seleção de características relacionadas às ações das empresas, mostrando que os modelos consideram o quanto a empresa está sendo valorizada pelo mercado de ações, e como isso, relaciona-se com o seu lucro ou prejuízo.

5.3 EXPLICABILIDADE DO MODELO GERADO PARA A ÁRVORE DE DECISÃO

A explicabilidade de uma árvore de decisão está relacionada à facilidade em compreender o modelo gerado, a partir da visualização de sua árvore.

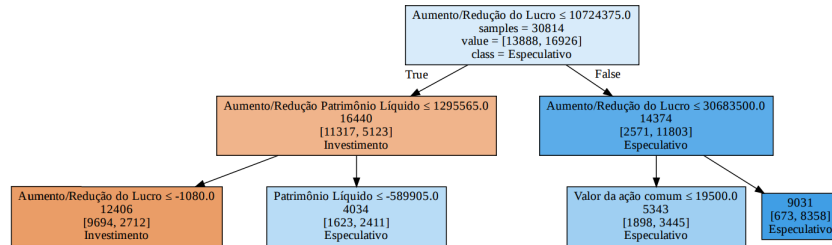
Para exemplificar, foi gerada uma árvore de decisão enxuta que considera apenas as dez características selecionadas previamente por meio do melhor modelo de CR (*Gradient Boosting*). A redução no número de características de entrada para a árvore foi feita para limitar a sua complexidade, ou seja, foi feita para criar um modelo visualmente mais simples e didático. Além disso, foram usados parâmetros que limitaram o número de características usadas durante o treinamento. Para facilitar o entendimento da DT gerada, não foi aplicada nenhuma normalização nos dados, ou seja, os valores estão em suas escalas originais.

A árvore gerada aparece na Figura 5.1. Em tons de laranja, estão representados os nós folhas que representam a classificação das empresas como Empresas Especulativas. Em tons de azul, estão representados as folhas que representam a classificação das empresas como Investimento.

Para fins de ilustração, a figura exhibe apenas os três primeiros níveis da árvore. Neste níveis aparecem as características com o maior poder de classificação. Entre elas, estão o 'Aumento/Redução do Lucro', 'Aumento/Redução do Patrimônio Líquido' e 'Patrimônio Líquido', contra apenas uma aparição de características relacionadas às ações da empresa. A versão completa da árvore pode ser observada no Apêndice 7.

A árvore de decisão gerada obteve 79,58% na pontuação F1, 79,65% de *Recall* e 79,55% de precisão. O resultado é pior do que o obtido quando se usa um modelo baseado no conjunto total de características, o que já era esperado. Mesmo assim, o desempenho pode ser considerado satisfatório.

Figura 5.1 – Representação gráfica simplificada da Árvore de Decisão para o Credit Rating.



Fonte: Próprio autor.

Como a Árvore de Decisão para a classificação multiclasse obteve resultados estatisticamente piores que os melhores modelos na classificação multiclasse, seu modelo não será explorado visualmente.

6 CONCLUSÕES

O trabalho mostrou o uso de ML para o problema de CR com classificação binária e multiclasse. Além disso, o trabalho apresenta uma análise referente às variáveis mais importantes no CR de empresas como instrumentos de Investimento ou Especulativos.

O trabalho utilizou como base dados públicos de demonstrativos financeiros e não utilizou fontes pagas para obtê-los. Isso aumenta a reprodutibilidade dos resultados obtidos, e promove a acessibilidade ao assunto discutido, especialmente se considerarmos que os trabalhos relacionados utilizam fonte de dados privadas ou inacessíveis aos leitores.

Com o uso de *Gradient Boosting*, foi obtida uma precisão de 90% na Classificação de Crédito Binária em dados de teste e 70% de precisão na classificação multiclasse. Os resultados indicam que *Gradient Boosting* é uma boa alternativa para a Classificação de Crédito utilizando dados financeiros. De modo geral, os resultados alcançados indicam que é possível obter resultados positivos utilizando apenas dados públicos e algoritmos de código aberto para modelar problemas de Classificação de Crédito.

O uso de algoritmos de Aprendizado de Máquina com alta explicabilidade possibilitou um aprendizado sobre o processo de Classificação de Crédito. Isso se deve ao fato de que, com ML, é possível realizar a extração das características mais importantes para os problemas de classificação trabalhados. O processo de extração realizado permitiu entender quais características mais auxiliariam durante a Classificação de Crédito. Dentre essas características analisadas, notou-se que a variável mais importante para a Classificação de Crédito binária e multiclasse nos modelos com melhor desempenho foi o aumento e redução do lucro.

Para finalizar, foi feita a disponibilização dos *scripts Python* criados para a prospecção e preparação dos dados públicos para algoritmos de ML. Os *scripts Python* utilizados para o treinamento, teste e análise das abordagens de ML também foram liberados ao público com o fim de aumentar a reprodutibilidade do experimento realizado.

REFERÊNCIAS BIBLIOGRÁFICAS

- AYODELE, T. O. Types of machine learning algorithms. **New advances in machine learning**, InTech, v. 3, p. 19–48, 2010.
- AYYADEVARA, V. K. Gradient boosting machine. In: . Apress, 2018. p. 117–134. Disponível em: <https://doi.org/10.1007/978-1-4842-3564-5_6>.
- BENÍTEZ, J. M.; CASTRO, J. L.; REQUENA, I. Are artificial neural networks black boxes? **IEEE Transactions on neural networks**, IEEE, v. 8, n. 5, p. 1156–1164, 1997.
- BREIMAN, L. Random forests. Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, 2001. Disponível em: <<https://doi.org/10.1023/a:1010933404324>>.
- BROWNLIE, J. Machine learning mastery with python. **Machine Learning Mastery Pty Ltd**, v. 527, p. 100–120, 2016.
- BUITINCK, L. et al. API design for machine learning software: experiences from the scikit-learn project. In: **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**. [S.l.: s.n.], 2013. p. 108–122.
- BURKART, N.; HUBER, M. F. A survey on the explainability of supervised machine learning. AI Access Foundation, v. 70, p. 245–317, jan. 2021. Disponível em: <<https://doi.org/10.1613/jair.1.12228>>.
- CARMONA, P.; CLIMENT, F.; MOMPARTLER, A. Predicting failure in the us banking sector: An extreme gradient boosting approach. **International Review of Economics & Finance**, Elsevier, v. 61, p. 304–323, 2019.
- CHANG, Y.-C.; CHANG, K.-H.; WU, G.-J. Application of extreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. **Applied Soft Computing**, Elsevier, v. 73, p. 914–920, 2018.
- CHARBUTY, B.; ABDULAZEEZ, A. Classification based on decision tree algorithm for machine learning. **Journal of Applied Science and Technology Trends**, v. 2, n. 01, p. 20–28, 2021.
- CHEN, C. C. M. et al. Methods for identifying SNP interactions: A review on variations of logic regression, random forest and bayesian logistic regression. Institute of Electrical and Electronics Engineers (IEEE), v. 8, n. 6, p. 1580–1591, nov. 2011. Disponível em: <<https://doi.org/10.1109/tcbb.2011.46>>.
- CHEN, J.; BOYLE, M. J. **High-Yield Bond Definition**. 2020. Disponível em: <https://www.investopedia.com/terms/h/high_yield_bond.asp>.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **CoRR**, abs/1603.02754, 2016. Disponível em: <<http://arxiv.org/abs/1603.02754>>.
- CHERN, C.-C. et al. A decision tree classifier for credit assessment problems in big data environments. **Information Systems and e-Business Management**, Springer, v. 19, n. 1, p. 363–386, 2021.
- Council of European Union. **Electronic Code of Federal Regulations (e-CFR) Disclosure requirements. (SEC) 240.17g-7**. 2014. <<https://www.law.cornell.edu/cfr/text/17/240.17g-7>>.

DATABASE, S. R. **Research Citation Data**. 2021. Dados adquiridos do Scopus Research Database, <<https://www.scopus.com/>>.

DAYAN, P.; NIV, Y. Reinforcement learning: The good, the bad and the ugly. **Current Opinion in Neurobiology**, v. 18, n. 2, p. 185–196, 2008. ISSN 0959-4388. Cognitive neuroscience. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0959438808000767>>.

DIETTERICH, T. Overfitting and undercomputing in machine learning. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 27, n. 3, p. 326327, set. 1995. ISSN 0360-0300. Disponível em: <<https://doi.org/10.1145/212094.212114>>.

DUA, D.; GRAFF, C. **UCI Machine Learning Repository**. 2017. Disponível em: <<http://archive.ics.uci.edu/ml>>.

EBA, E. B. A. E. **Amended Mapping of Fitch Ratings Investors Service credit assessments under the Standardised Approach**. 2019.

_____. **Amended Mapping of Moodys Investors Service credit assessments under the Standardised Approach**. 2019.

_____. **Amended Mapping of S&P Global Ratings Investors Service credit assessments under the Standardised Approach**. 2019.

ENCYCLOPAEDIA, T. E. of. **Credit**. 2020. Disponível em: <<https://www.britannica.com/topic/credit>>.

FERNANDO, J.; KINDNESS, D. **Earnings Per Share (EPS)**. 2021. Disponível em: <<https://www.investopedia.com/terms/e/eps.asp>>.

FEURER, M.; HUTTER, F. Hyperparameter optimization. In: **Automated machine learning**. [S.l.]: Springer, Cham, 2019. p. 3–33.

FINANCE, C. F. M. **Rating History**. [S.l.]: GitHub, 2021. <https://github.com/govwiki/rating_history>.

FUHR, G. G. T. **Use of Machine Learning to classify companies credit through financial statements**. [S.l.]: GitHub, 2021. <<https://github.com/gtfuhr/tcc>>.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. [S.l.]: O'Reilly Media, 2019.

GLOBAL, S. **S&P Global Ratings definitions**. 2021. Disponível em: <<https://www.spglobal.com/ratings/en/research/articles/190705-s-p-global-ratings-definitions-504352>>.

HANDELMAN, G. S. et al. Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods. *American Roentgen Ray Society*, v. 212, n. 1, p. 38–43, jan. 2019. Disponível em: <<https://doi.org/10.2214/ajr.18.20224>>.

ILLUSTRATING the random forest algorithm in TikZ. 2019. *Tex Stack Exchange*. [Online:]. Disponível em: <<https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz>>.

JAPKOWICZ, N.; SHAH, M. Performance evaluation in machine learning. In: . Springer International Publishing, 2015. p. 41–56. Disponível em: <https://doi.org/10.1007/978-3-319-18305-3_4>.

KELLEY, H. J. Gradient theory of optimal flight paths. **Ars Journal**, v. 30, n. 10, p. 947–954, 1960.

KUMAR, M.; THENMOZHI, M. Forecasting stock index movement: A comparison of support vector machines and random forest. In: **Indian institute of capital markets 9th capital markets conference paper**. [S.l.: s.n.], 2006.

LANGOHR HERWIG; LANGOHR, P. **The Rating Agencies and Their Credit Ratings - What They Are, How They Work, and Why They are Relevant**. [S.l.]: Wiley Finance, 2009.

LECUN, Y. et al. Backpropagation applied to handwritten zip code recognition. **Neural computation**, MIT Press, v. 1, n. 4, p. 541–551, 1989.

LINNAINMAA, S. Taylor expansion of the accumulated rounding error. **BIT Numerical Mathematics**, Springer, v. 16, n. 2, p. 146–160, 1976.

LOH, W.-Y. Fifty years of classification and regression trees. Wiley, v. 82, n. 3, p. 329–348, jun. 2014. Disponível em: <<https://doi.org/10.1111/insr.12016>>.

LU, H. et al. **Accelerating Gradient Boosting Machine**. 2019.

MAHESH, B. Machine learning algorithms -a review. *International Journal of Science and Research (IJSR)*, 01 2019.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, n. 4, p. 115–133, 1943.

MEKONNEN, Y. et al. Review—machine learning techniques in wireless sensor network based precision agriculture. *The Electrochemical Society*, v. 167, n. 3, p. 037522, jan. 2020. Disponível em: <<https://doi.org/10.1149/2.0222003jes>>.

MNIH, V. et al. Human-level control through deep reinforcement learning. **nature**, Nature Publishing Group, v. 518, n. 7540, p. 529–533, 2015.

MORE, A. S.; RANA, D. P. Review of random forest classification techniques to resolve data imbalance. In: **2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)**. [S.l.: s.n.], 2017. p. 72–78.

OH, S. J.; SCHIELE, B.; FRITZ, M. Towards reverse-engineering black-box neural networks. In: **Explainable AI: Interpreting, Explaining and Visualizing Deep Learning**. [S.l.]: Springer, 2019. p. 121–144.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PETERSEN, K. et al. Systematic mapping studies in software engineering. In: **12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12**. [S.l.: s.n.], 2008. p. 1–10.

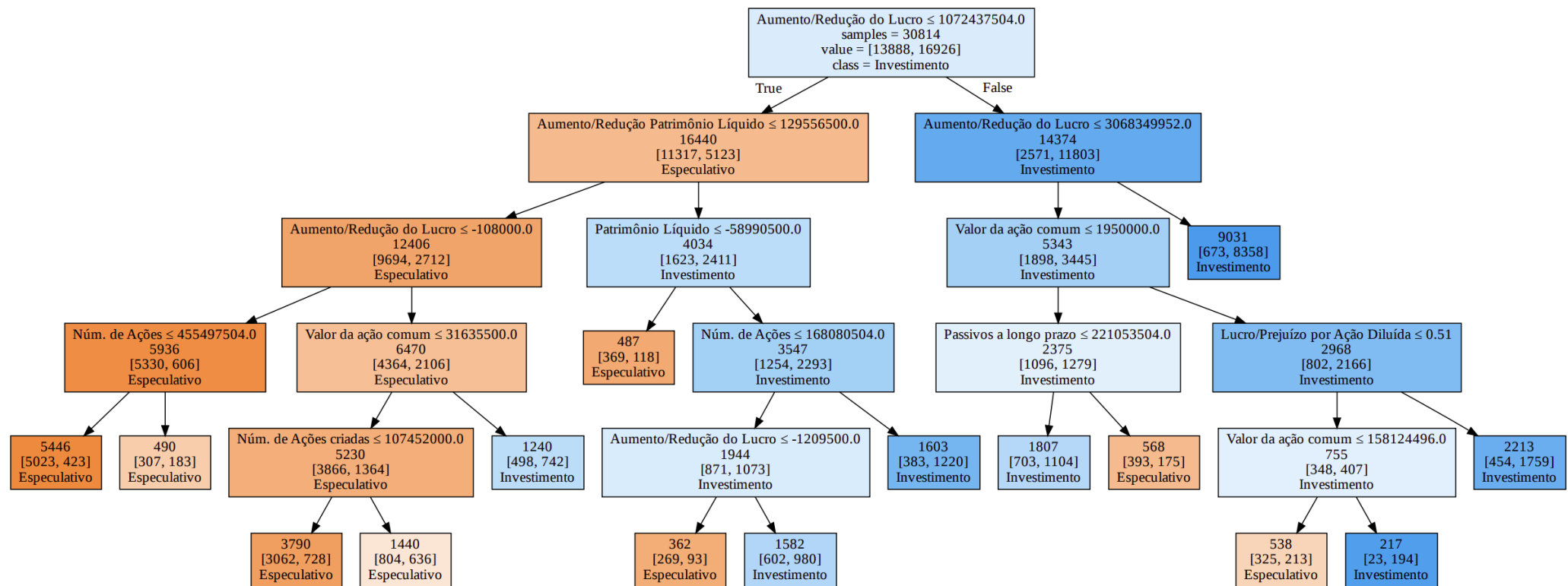
POZZOLO, A. D. et al. Calibrating probability with undersampling for unbalanced classification. In: **IEEE. 2015 IEEE Symposium Series on Computational Intelligence**. [S.l.], 2015. p. 159–166.

RATINGS, S. G. **Ratings Criteria and Models**. 2021. Disponível em: <<https://disclosure.spglobal.com/ratings/en/regulatory/ratings-criteria>>.

ZOU, Z. B.; PENG, H.; LUO, L. K. The application of random forest in finance. In: **Mechanical, Information and Industrial Engineering**. [S.l.]: Trans Tech Publications Ltd, 2015. (Applied Mechanics and Materials, v. 740), p. 947–951.

7 ÁRVORE DE DECISÃO USADA PARA CR.

Figura 7.1 – Representação gráfica completa da Árvore de Decisão para o Credit Rating.



Fonte: Próprio autor.