e-ISSN: 2236-0972
https://doi.org/10.5585/gep.v12i2.18942
Recebido em: 16 dez. 2020
Aprovado em: 12 abr. 2021
Editor Chefe: Luciano Ferreira da Silva
Editor Científico: Flavio Santino Bizarrias
Editor Científico: Renato Penha

MODELOS DE MACHINE LEARNING PARA PREDIÇÃO DO SUCESSO DE STARTUPS

MACHINE LEARNING MODELS FOR PREDICTING SUCCESS OF STARTUPS

Fabiano Rodrigues
Doutor em Administração, FEA-USP
ESPM / PPGA
frodrigues@espm.br

Francisco Aparecido Rodrigues
Doutor em Física, Instituto de Física de São Carlos (USP)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
franscisco@icmc.usp.br

Thelma Valéria Rocha Rodrigues
Doutora em Administração, FEA-USP
ESPM / PPGA
tvrocha@espm.br

Resumo
Este estudo analisa resultados obtidos com modelos de *machine learning* para predição do sucesso de *startups*. Como *proxy* de sucesso considera-se a perspectiva do investidor, na qual a aquisição da *startup* ou realização de *IPO* (*Initial Public Offering)* são formas de recuperação do investimento. A revisão da literatura aborda *startups* e veículos de financiamento, estudos anteriores sobre predição do sucesso de *startups* via modelos de *machine learning*, e *trade-offs* entre técnicas de *machine learning*. Na parte empírica, foi realizada uma pesquisa quantitativa baseada em dados secundários oriundos da plataforma americana Crunchbase, com *startups* de 171 países. O design de pesquisa estabeleceu como filtro *startups* fundadas entre junho/2010 e junho/2015, e uma janela de predição entre junho/2015 e junho/2020 para prever o sucesso das *startups*. A amostra utilizada, após etapa de pré-processamento dos dados, foi de 18.571 *startups*. Foram utilizados seis modelos de classificação binária para a predição: Regressão Logística, *Decision Tree*, *Random Forest*, *Extreme Gradiente Boosting*, *Support Vector Machine* e Rede Neural. Ao final, os modelos *Random Forest* e *Extreme Gradient Boosting* apresentaram os melhores desempenhos na tarefa de classificação. Este artigo, envolvendo *machine learning* e *startups*, contribui para áreas de pesquisa híbridas ao mesclar os campos da Administração e Ciência de Dados. Além disso, contribui para investidores com uma ferramenta de mapeamento inicial de *startups* na busca de *targets* com maior probabilidade de sucesso.

Palavras-chave: Predição do sucesso de startups. Machine learning. Investimento em startups. Plataforma crunchbase.

Abstract
This study analyzes results from machine learning models to predict the success of startups. As a proxy for success, we considered the investor's perspective, according to which startup buyout or IPO (Initial Public Offering) are ways to recover the investment. The literature review addresses startups and funding mechanisms, previous studies on prediction of startup success via machine learning models, and trade-offs between machine learning techniques. The empirical study comprised a quantitative research based on secondary data from the American Crunchbase platform, with startups from 171 countries. The research design used as filter startups founded between June/2010 and June/2015, as well as a prediction window from June/2015 to June/2020 to predict startup success. The final sample, after the data preprocessing stage, comprised 18,571 startups. Six binary classification models were used for success prediction: Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, Support Vector Machine, and Neural Networks. In the end, the Random Forest and Extreme Gradient Boosting models had the best performance in the classification task. This article involving machine learning and startups contributes to research in hybrid fields by combining perspectives from Business and Data Science. Additionally, it contributes to investors with a tool for initial mapping of startups in search of targets with greater probability of success.

Keywords: Startup success prediction. Machine learning. Investment in startups. Crunchbase platform.

Cite como
American Psychological Association (APA)

Rodrigues, F., Rodrigues, F. A., & Rodrigues, T. V. R. (2021, maio/ago.). Modelos de machine learning para predição do sucesso de startups. *Revista de Gestão e Projetos (GeP)*, *12*(2), 28-55. https://doi.org/10.5585/gep.v12i2.18942.

1 Introdução

Nos últimos anos, o número global de unicórnios (*startups* com valor de mercado acima de um bilhão de dólares) vem aumentando. Em dezembro de 2020, 506 *startups* foram enquadradas nesta categoria, com *valuation* acumulado de aproximadamente US$ 1.6 trilhões de dólares (CB Insights, 2020). Por outro lado, apenas metade das *startups* conseguem sobreviver mais de cinco anos (National, 2020).

Startups, por definição, são organizações temporárias usadas para procurar um modelo de negócios repetível e com alta escalabilidade (Blank, 2013). Essa busca por modelos de sucesso é bastante volátil, tanto para os empreendedores como para os investidores. Afinal, *startups* são instituições humanas desenhadas para entregar um novo produto ou serviço sob condições de extrema incerteza (Ries, 2012). As incertezas são múltiplas, de natureza tecnológica, financeira, mercadológica, macroeconômica, de encaixe entre a oferta criada e a necessidade dos consumidores, de condições e forma de gerenciamento, entre outras.

Fundos de Venture Capital (VC) realizam investimentos com o intuito de auferirem retornos advindos do desinvestimento (*exit*), ou seja, quando a *startup* investida é adquirida por outra organização ou abre seu capital por meio de um *IPO – Initial Public Offering*. Empreendedores, além do propósito que movimenta suas organizações, também buscam investimentos para auxiliá-los nas diversas fases do seu empreendimento.

Diante desse cenário de alto risco e potencial alto retorno, prever as *startups* com maiores chances de sucesso pode ajudar o ecossistema empreendedor, reunindo investidores a iniciativas com maior potencial de rentabilização e oferecendo aos empreendedores uma bússola para checar a probabilidade de terem sucesso no desenvolvimento de suas iniciativas.

O sucesso de uma *startup* pode ser definido de muitas formas: pela capacidade de atração de novos investimentos (*funding rounds*), pela velocidade exponencial de crescimento das vendas, pela participação em processos de *M&A* (*Mergers and Acquisitions*), e pela abertura de capital via *IPO*.

Este estudo assume a perspectiva do investidor, usando *exit* como *proxy* para o sucesso. *Exit* é o termo usado para a saída do investidor; isso ocorre quando a *startup* é adquirida por outra empresa (processo de *M&A*) ou realiza uma oferta pública de suas ações (processo de *IPO*). Nos dois casos, os investidores podem vender suas participações para auferir ganhos nas transações. Dessa forma, o objetivo deste estudo consiste em analisar os resultados de modelos de *machine learning* para a predição do sucesso de *startups*. Os modelos preditivos são de classificação binária: sucesso – ou não – das *startups*.

A literatura acadêmica com foco em *machine learning* para predição do sucesso de

startups já conta com frentes de pesquisa nessa direção, que contribuíram para o presente estudo (Liang & Daphne Yuan, 2013; Shan, Cao & Lin, 2014; Bento, 2017; Pan, Gao & Luo, 2018; Gastaud, Carniel & Dalle, 2019; Arroyo et al., 2019). Na última década, vários modelos foram testados para essa tarefa, incluindo *Naive Bayes*, *Decision Trees*, *Random Forest*, *Supported Vector Machine*, entre outros. Os autores citados concentram-se majoritariamente no campo da computação e ciência de dados. Este estudo inspira-se no design de janelas temporais proposto por Arroyo et al. (2019), incluindo apenas *startups* com data de fundação recente, entre junho de 2010 e junho de 2015, para prever seu sucesso entre junho de 2015 e junho de 2020.

Como contribuição acadêmica, realizaram-se algumas mudanças no design da pesquisa desenvolvida por Arroyo et al. (2019) e na estratégia de tratamento dos dados, tais como: tempo das janelas de inclusão e predição; adição de *startups* com período inferior a dois anos para captação na primeira rodada de financiamento; adoção de estratégia conservadora para tratamento de dados faltantes; inclusão de redes neurais do tipo *multi layer perceptron* ao conjunto de modelos avaliados para a tarefa de classificar e inclusão de atributos preditivos aderentes ao contexto empreendedor. Adicionalmente, devido ao *background* dos autores da pesquisa, este estudo contribui com a aproximação dos campos da Ciência de Dados e da Administração.

Para o ecossistema empreendedor, este estudo propõe filtrar *startups* que tenham maior probabilidade de sucesso para os investidores. Em um cenário extremamente competitivo e incerto, onde mais de 90% das *startups* falham (Profitfromtech, 2020), o estudo auxilia na identificação das mais promissoras. Também destaca a relevância de modelos de *machine learning* no processo de varredura inicial de oportunidades de investimento, agregando uma camada para filtrar as *startups* com maior potencial de sucesso.

Como estrutura, este estudo inclui uma revisão teórica sobre o tema, o desenvolvimento do design da pesquisa, tratamento e limpeza dos registros de *startups* oriundos da plataforma Crunchbase (2020), uma análise exploratória, e predição do sucesso de *startups* a partir de modelos de classificação desenvolvidos em linguagem de programação Python.

2 Revisão teórica

Nesta seção são abordados o conceito de *startups*, veículos para financiamento de suas atividades, estudos anteriores sobre predição do sucesso de *startups* e modelos de *machine learning*.

2.1 Startups e veículos de financiamento

Startups, por definição, são organizações temporárias usadas para procurar um modelo de negócios repetível e com alta escalabilidade (Blank, 2013). São ideias que se multiplicaram nas últimas décadas e foram

postas à prova em ambientes de alta incerteza, mas com potencial de atração de demanda.

O surgimento de bases de dados sobre *startups*, como a Crunchbase, facilita o acesso a grande quantidade e variedade de dados estruturados, com potenciais usos para múltiplos stakeholders: fundadores em busca de financiamento, investidores à procura de *startups*, pessoas em busca de ideias para novos empreendimentos, e pesquisadores acadêmicos que se interessam pela dinâmica e pelo ecossistema das *startups* (Crunchbase, 2020).

Crunchbase é uma plataforma mundial criada em 2007, com sede nos Estados Unidos, para profissionais que buscam dados sobre empresas inovadoras. Dalle, den Besten e Menon (2017) identificaram mais de 90 artigos acadêmicos usando o Crunchbase como fonte de dados para estudos em variados campos, como gestão e economia.

A fotografia da base de dados Crunchbase, no dia 29 junho de 2020, apresentava 1.064.929 *startups* e informações sobre 906.248 fundadores (extração realizada pelos autores, com a devida autorização da equipe Crunchbase para fins acadêmicos). A base possui um histórico com mais de 325 mil rodadas de financiamento (*funding rounds*), em todas as etapas do ciclo de vida de uma *startup*: *early-*, *growth-* e *mature stage* (nascimento, crescimento e maturidade).

Para exemplificar essa lógica de *funding* no cenário brasileiro, Gereto (2019) analisou o ciclo de investimento e desinvestimento de 436 *startups* brasileiras registradas na base de dados Crunchbase. Nesse estudo, visualizou uma média de 1,7 rodadas de investimento do tipo *seed*; 2,2 rodadas em *early stage;* e 3,6 rodadas em *later stage*, com faixas crescentes captadas ao longo do ciclo de vida da *startup*: US$ 100.000,00 a US$ 1.000.000,00 (*seed*); US$ 1.000.000,00 a US$ 10.000.000,00 (*early stage*); e US$ 10.000.000,00 a US$ 100.000.000,00 (*later stage*). A Tabela 1 apresenta diferentes tipos de financiamento de *startups*

Tabela 1 - Tipos de *funding* no ciclo de vida das *startups*

	Descrição	Fontes ($) mais comuns
Pre-Seed	Estágios embrionários da *startup* (início do seu ciclo de vida)	Investidores anjos, amigos, familiares e capital próprio
Seed-Round	*Startup* começa a ganhar tração, mas ainda está no início de sua operação. Alguns investidores mais qualificados começam a sondar as *startups* para ampliar e diversificar seu portfólio de investimentos.	Investidores anjos, incubadoras, aceleradoras e alguns fundos de *venture capital* (VC) começam sondagens.
Series A	*Startup* já apresenta alguma prova de conceito / modelo. Torna-se possível analisar resultados reais obtidos a partir de rodadas de *funding* anteriores. Começam a entrar investidores que contribuam para que a *startup* atinja novos patamares.	Investidores anjos com maior capacidade, venture capital, *Family offices* e alguns fundos de *private equity*.
Series B	Neste ponto, as expectativas são mais ousadas, como por exemplo: forte expansão territorial, ampliação de canais de vendas e ganho de escala em ritmo mais acelerado.	Investidores parecidos com as rodadas anteriores, com maior apetite pelo potencial de crescimento da *startup*. Nesta fase, investidores anteriores podem auxiliar na captação de novos interessados.
Series C e mais	Este *milestone* indica um aumento na probabilidade de sucesso da *startup*. Neste ponto, provavelmente, o negócio já esteja validado, operando em alta escala e apresentando maior *valuation*. Investidores já começam a pensar em estratégias de saída.	As exigências feitas por novos investidores nesta fase costumam ser maiores, com crescente expectativa sobre controle, dados e *due dilligence*, por exemplo.

Fonte: Adaptado de Losada (2020, p. 124).

No começo do ciclo de vida da *startup*, a competição por recursos é bastante expressiva, tanto pelo volume de ofertas como pela busca de diversificação de ativos por parte dos investidores.

A passagem pelas rodadas de financiamento é uma sinalização de que a *startup* está trilhando um caminho de potencial sucesso para os investidores. A profissionalização do dinheiro investido aumenta, até mesmo com a indicação de executivos de *venture capital* para posições-chave na gestão da *startup* (Cremades, 2016).

Ao longo do ciclo de vida das *startups*, os investidores vão gradativamente aumentando suas preocupações quanto às estratégias de saída (*exit*). As estratégias de saída mais comuns, com potencial geração de ganhos, são um *IPO* ou uma venda competitiva. Em um cenário negativo, caso de fracasso da *startup*, os investimentos realizados são perdidos (Losada, 2020).

2.2 Predição do sucesso de startups: estudos anteriores com machine learning

Nos últimos anos, muitos estudos sobre *startups* têm usado técnicas de *machine learning* como apoio à geração de *insights* e conhecimentos. Pesquisas sobre ecossistemas empreendedores (Nylund & Cohen, 2017; Kemeny, Nathan & Almeer, 2017; Kosterich & Weber, 2018; Basole, Park & Chao, 2019), sucesso de *startups*, mercados específicos como *fintechs* (Hsieh & Li, 2017), seleção de oportunidades para investimento e capital de risco são alguns exemplos. A Tabela 2 sintetiza alguns estudos relacionados à predição do

sucesso de *startups* com técnicas de *machine learning*, a partir de dados da plataforma Crunchbase.

Tabela 2 - Predição do sucesso de *startups*: síntese de estudos relacionados

Referência	Objetivo	Amostra	Variável de Interesse	Técnicas de *Machine Learning*
Liang e Daphne Yuan (2013)	Investigar o papel das relações sociais entre investidores e empresas para a predição do comportamento de investimento.	11.916 *startups*, 12.127 pessoas, 1.122 organizações financeiras	Ocorrência de investimento	***SVM***, **Árvore de Decisão** e *Naive Bayes*
Shan, Cao e Lin (2014)	Prever se um investidor investirá em uma *startup* específica com base em sinais textuais, topológicos e específicos do domínio	214.290 *startups*, 286.659 pessoas, 31.942 investimentos	Investidor efetiva investimento em *startup*	**Regressão Logística**
Bento (2017)	Desenvolver um modelo preditivo para classificar uma *startup* como bem-sucedida, ou não (classificação binária)	86588 *startups* (estados americanos)	*IPO* ou Aquisição da *startup*	**Floresta Aleatória**, *SVM* e Regressão Logística
Pan, Gao e Luo (2018)	Predizer o sucesso de *startups*, definido como um evento que dá uma grande quantia aos fundadores e investidores	+60.000 *startups*	Processo de *M&A* ou *IPO*	***KNN*,** Floresta Aleatória e Regressão Logística
Arroyo et al. (2019)	Desenvolvimento e avaliação de uma abordagem orientada por dados que usa *machine learning* para ajudar os investidores de *VC* a explorar e selecionar as melhores empresas para apoiar.	120.507 *startups*, 34.180 *funding* rounds	Aquisição, rodada de financiamento, *IPO*, fechamento ou nenhum evento	***Gradient Tree Boosting***, Árvore de Decisão, Floresta Aleatória e *SVM*.
Gastaud, Carniel e Dalle (2019)	Predizer o sucesso de *startups* na arrecadação de investimentos em diferentes estágios (*early-*, *growth-* e *late stage*)	65.957 *startups*	Obtenção de *funding* em diferentes estágios (*seed*, *series* A e B)	Floresta Aleatória, *Graph Neural Networks*

Fonte: Elaborado pelos autores (2021).

A variável de interesse é operacionalizada de diversas formas. Dentre os estudos apresentados na Tabela 2, os eventos *IPO* e/ou aquisição de uma *startup* destacam-se como variáveis para o sucesso (Bento, 2017; Pan, Gao & Luo, 2018; Arroyo et al., 2019). Um processo de *M&A*, bem como um *IPO*, representam estratégias de saída (*exit*) para que investidores tentem auferir ganhos.

Outros estudos estabelecem rodadas de investimento como *proxy* para o sucesso. Arroyo et. al (2019) estabeleceram um modelo de predição para múltiplas classes, caracterizadas pelos eventos: aquisição da

startup, fechamento, *IPO*, rodada de financiamento, e nenhum evento.

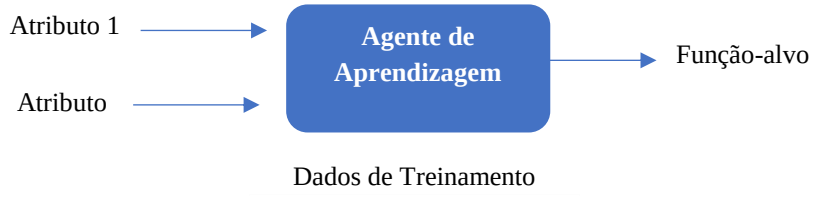
Sobre as técnicas de *machine learning* usadas, destacam-se as de aprendizado supervisionado com classificação binária. As técnicas mais utilizadas foram Regressão Logística, Árvore de Decisão, Floresta Aleatória e *SVM*. Redes *Bayesianas*, *KNN* e *Naive Bayes* foram menos presentes (Liang & Daphne Yuan, 2013; Pan, Gao & Luo, 2018). Recentemente, um estudo usou *Gradient Tree Boosting* no processo de predição do sucesso de *startups* (Arroyo et al., 2019) e outro utilizou *Graph Neural Networks* (Gastaud, Carniel & Dalle, 2019); ambos os estudos obtiveram resultados e modelagens promissoras.

2.3 Modelos de machine learning

Machine learning (*ML*), ou aprendizado de máquina, é um dos principais ramos da Inteligência Artificial. Os impactos dos modelos de *machine learning* são cada vez mais presentes no dia a dia de pessoas e organizações, seja em aplicações para reconhecimento de voz, detecção de fraudes bancárias, sistemas de recomendação de filmes e músicas em plataformas de *streaming*, sistemas de busca de dados, entre outros. Nos bastidores dessas aplicações, os algoritmos de *ML* buscam aprender com grandes volumes de dados, estruturados ou não, identificando padrões para classificação ou regressão.

Do ponto de vista computacional, *machine learning* caracteriza-se como "o estudo de algoritmos de computador que melhoram automaticamente através da experiência" (Ertel, 2017, p. 178). Similarmente, Facelli et al. (2019, p. 3) argumentam que, em *ML*: "... computadores são programados para aprender com a experiência passada. Para tal, empregam um princípio de inferência denominado indução, no qual se obtém conclusões genéricas a partir de um conjunto particular de exemplos". Assim, modelos de *ML* aprendem a induzir hipóteses ou funções que consigam resolver o problema (de regressão ou classificação) a partir do conjunto de dados de treinamento.

Figura 1 - Processo simplificado de aprendizagem



Fonte: Adaptado de Ertel (2017, p. 178).

Um agente de aprendizagem é bem-sucedido se "melhorar seu desempenho (medido por um critério adequado) em dados novos e desconhecidos ao longo do tempo (após muitos exemplos de treinamento)" (Ertel,

2017, p. 178). As tarefas de aprendizagem podem ser divididas em aprendizado supervisionado e não supervisionado.

Na aprendizagem supervisionada, foco deste estudo, tenta-se prever uma variável-alvo a partir de um conjunto de atributos, por meio de classificação ou regressão. O termo supervisionado refere-se ao conhecimento inicial do estado da variável dependente do modelo, para cada vetor de atributos dos exemplos contidos no conjunto dos dados de treinamento. Já na aprendizagem não supervisionada, a predição cede espaço para a descrição, seja na forma de agrupamento, associação ou sumarização. Neste caso, não existe um atributo-alvo a priori, buscam-se padrões entre atributos ou objetos contidos no conjunto de dados.

Skiena (2017) argumenta que, dificilmente, um algoritmo de *ML* seja superior a todos, pela diversidade de domínios, contextos do problema de decisão, e da própria característica dos dados.

A Tabela 3 mostra uma avaliação subjetiva, em uma escala de 1 (pior) a 10 (melhor), de algumas técnicas de *ML* em cinco dimensões: poder, facilidade de interpretação, facilidade de uso, velocidade de treinamento e velocidade de predição.

Tabela 3 - Avaliação subjetiva das técnicas de *ML*

Técnica	Poder	Facilidade de Interpretação	Facilidade de Uso	Velocidade de Treinamento	Velocidade de Predição
Regressão linear	5	9	9	9	9
Vizinho mais próximo (*KNN*)	5	9	8	10	2
Naive Bayes	4	8	7	9	8
Árvore de decisão	8	8	7	7	9
Máquinas de vetores de suporte	8	6	6	7	7
Boosting	9	6	6	6	6
Graphical Models	9	8	3	4	4
Deep Learning	10	3	4	3	7

Fonte: Adaptado de Skiena (2017, p. 353).

Há um *trade-off* natural entre as dimensões expostas na Tabela 3. Algumas técnicas propiciam maior facilidade para a interpretação dos resultados (Árvore de Decisão, por exemplo). Já outras, como *Deep Learning*, possuem alto poder de predição e menor facilidade de interpretação.

A Tabela 4 apresenta as técnicas usadas e os principais resultados dos 6 estudos (ver Tabela 2) que usaram como fonte de dados a base Crunchbase. A técnica mais utilizada foi a Floresta Aleatória (*Random Forest*), seguida das técnicas de Regressão Logística e *SVM*. As três técnicas menos utilizadas, empregadas em pesquisas mais recentes, foram *KNN*, *Gradient Tree Boosting* e *Graph Neural Networks*.

Tabela 4 - Técnicas de *machine learning* versus artigos com a base Crunchbase

TÉCNICA VS ARTIGO VS RESULTADO	Liang e Daphne Yuan (2013)	Shan, Cao e Lin (2014)	Bento (2017)	Pan, Gao e Luo (2018)	Arroyo et al. (2019)	Gastaud, Carniel e Dalle (2019)
Regressão logística		Precisão (**0.864**)	Acurácia (0.928)	Acurácia (72.54%)		
Naive Bayes	TPR (54.80%)					
Árvore de decisão	TPR (87.53%)				Precisão (0.09, aquisição) (0.04, *IPO*)	
Floresta aleatória			Acurácia (**0.931**)	Acurácia (**84.53%**)	Precisão (**0.33**, aquisição) (**0.44**, *IPO*)	Precisão (0.63)
KNN				Acurácia (73.33%)		
SVM	TPR (**89.58%**)		Acurácia (0.928)		Precisão (0.00, aquisição) (0.00, *IPO*)	
Gradient tree boosting					Precisão (0.17, aquisição) (0.07, *IPO*)	
Graph neural network						Precisão (**0.65**)

Fonte: Elaborado pelos autores (2021).

Em termos das medidas de avaliação de classificadores, Kubat (2017) aponta que precisão consiste na probabilidade de o classificador estar correto quando classifica um exemplo como positivo, enquanto um alto *recall* significa a porcentagem de exemplos positivos classificados como positivos. A acurácia é o indicador mais simples, consiste na razão entre o número de predições corretas pelo total de predições. A pontuação F1 é a média harmônica entre precisão e *recall*. Já a medida *GMean*, que é bastante usada no caso de *datasets* desbalanceados, basicamente consiste na média geométrica entre a sensibilidade (recall) e a especificidade (porcentagem de exemplos negativos classificados como negativos) do modelo.

Arroyo et. al (2019) obtiveram menor precisão em seus modelos em comparação com Shan, Cao e Lin (2014). As melhores precisões foram obtidas com o modelo de Floresta Aleatória, com 0,33 para *startups* adquiridas e 0,44 para as que realizaram processo de *IPO*. Importante destacar que esse estudo, após reinterpretação e nova operacionalização do conceito de sucesso das *startups*, atingiu precisão superior a 60%. Em resumo, os autores classificaram como bem-sucedidas *startups* que foram adquiridas, que realizaram *IPO*, ou que continuavam recebendo *funding* (o fracasso foi designado apenas para *startups* fechadas ou sem eventos na janela de simulação do estudo).

O primeiro estudo usou janelas temporais para avaliação dos modelos. Nesse caso, estabeleceram-se duas janelas: 1) de aquecimento, com duração de 4 anos, entre ago/2011 e ago/2015 (contendo *startups* não adquiridas, sem *IPO*, operantes e com rodada de

financiamento inferior à série C); e 2) de simulação, entre ago/2015 e ago/2018, para captura do primeiro evento entre as seguintes opções: aquisição, fechamento, *IPO*, rodada de financiamento, e nenhum evento para uma *startup* da janela de aquecimento. O segundo estudo teve outro design, incorporando sinais textuais, topológicos e específicos do domínio, tanto do investidor quanto da *startup*.

Bento (2017), criticado por Arroyo et al. (2019) pela inclusão de *startups* fundadas desde 1985 (data anterior à criação da plataforma Crunchbase), apresentou alta acurácia em seus modelos. O autor trabalhou com o maior número de atributos (158), usando apenas dados de alguns estados americanos de forma agregada. Além disso, o uso da acurácia como métrica de avaliação de desempenho para *datasets* desbalanceados deve ser vista com cautela, pois verifica a proporção de casos corretamente classificados, sejam verdadeiros positivos ou verdadeiros negativos. Como, muitas vezes, o interesse analítico recai na classe minoritária, a acurácia pode ser enviesada pelo desempenho da classe majoritária.

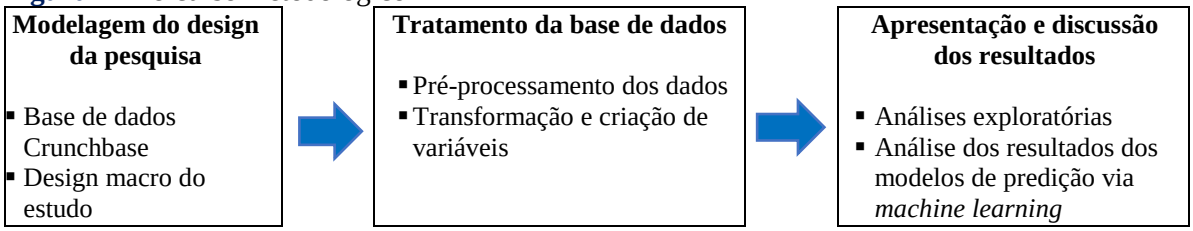
Nota-se que diferentes designs de pesquisa (janelas de simulação, filtros para eleger *startups* válidas para o estudo), quantidade e tipos de atributos usados, as etapas de pré-processamento e transformação dos dados, além das métricas de avaliação de desempenho dos modelos, podem interferir na qualidade e na interpretação dos resultados.

A próxima seção tratará da metodologia do presente estudo, seus materiais, métodos, pré-processamento e tratamento das variáveis, escolha dos atributos, design e operacionalização da pesquisa.

3 Metodologia

O percurso metodológico contempla a modelagem do design da pesquisa, a partir das tabelas da base de dados Crunchbase e do design macro do estudo, bem como o tratamento da base de dados (pré-processamento, limpeza, transformação e criação de novos atributos preditivos), culminando na apresentação e discussão dos resultados advindos de seis modelos de *machine learning* desenvolvidos em linguagem de programação Python. A Figura 2 apresenta os principais passos da pesquisa quantitativa baseada na análise de dados secundários.

Figura 2 - Percurso metodológico



Fonte: Adaptado de Bento (2017, p. 25).

3.1 Base de dados Crunchbase

Após contatos eletrônicos estabelecidos com a equipe técnica da plataforma Crunchbase, sediada nos Estados Unidos, foi cedido aos autores o acesso aos dados para fins acadêmicos por um período de 6 meses (até outubro de 2020). A Tabela 5 apresenta um resumo de todos os arquivos extraídos da base de dados, em formado CSV (*comma-separated values*), no dia 29 de junho de 2020.

Tabela 5 - Arquivos Crunchbase

Nome das Tabelas	Observações na Base	No. de Atributos
Acquisitions	103.783	27
category_groups	744	9
Checksum	17	3
Degrees	362.706	17
event_appearances	410.332	15
Events	21.229	21
funding_rounds	325.766	24
Funds	15.243	15
investment_partners	86.124	14
Investments	494.377	14
Investors	149.143	25
Ipos	32.225	27
Jobs	103.783	27
org_parents	15.538	10
organization_descriptions	633.142	9
Organizations	1.064.929	41
People	1.030.568	22
people_descriptions	554.512	9

Fonte: Adaptado de Crunchbase (2020).

A base de dados possui 18 tabelas, incluindo informações sobre aquisições realizadas (*acquisitions*), tipos de mercados atingidos pelas *startups* (*category_groups*), rodadas de financiamento (*funds*, *funding_rounds*), investidores e investimentos (*investors*, *investments*, *investment_partners*), *IPOS* realizados (*IPOS*), pessoas envolvidas nas *startups* (*people*, *people_descriptions*) e dados sobre as próprias *startups* (*organizations*, *organization_descriptions*). Neste estudo, a tabela *#organizations* é a base central para consolidação e geração do *dataframe* final.

Dos 41 atributos da tabela *#organizations*, alguns foram descartados por não agregarem valor ao processo de predição do sucesso das *startups*, tais como: e-mail, número de telefone, urls dos perfis de *linkedin*, *facebook* e *twitter*, endereço, CEP e descrições. O atributo "UUID" (sigla para código de 128 bits denominado *Universally Unique Identifier*) consiste na chave primária dessa tabela, com valores únicos para identificação de cada *startup*.

3.2 Design macro do estudo

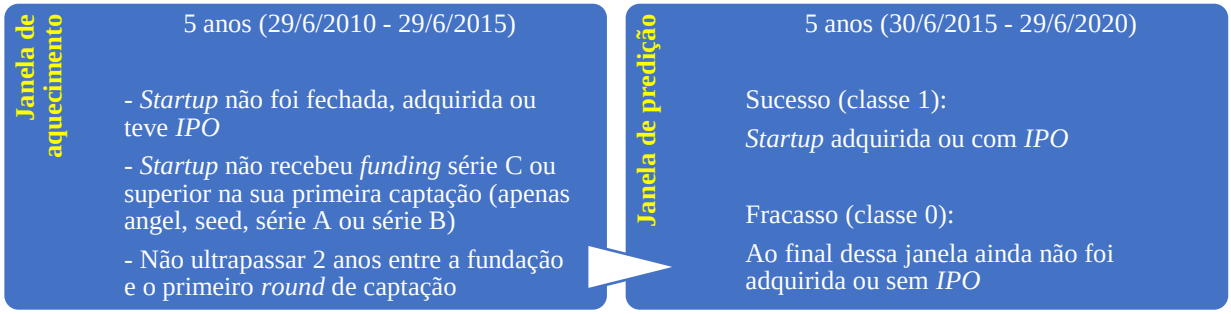
O design deste estudo foi similar ao desenvolvido por Arroyo et al. (2019), com a

utilização de uma janela temporal para a escolha das *startups* que fizeram parte do *dataset* final. A partir da base completa, com 1.064.926 *startups*, extraída do Crunchbase no dia 29 de junho de 2020, foram filtradas *startups* com data de fundação entre os dias 29/6/2010 e 29/6/2015, tendo-se assim uma janela de 5 anos para inclusão de *startups*. Esse período foi denominado janela de aquecimento. Foram incluídas apenas *startups* que receberam investimentos até a série C, mais especificamente: tipo *angel* (capital anjo), *seed* (capital semente), série A e série B.

Adicionalmente aos filtros citados, o fator 'tempo até a primeira rodada de captação' foi inserido, com uma janela máxima de 2 anos após a fundação para a *startup* captar seu primeiro financiamento. Esse critério buscou aproximar a visão do investidor, que deseja crescimento rápido dos negócios, sendo o tempo para a primeira rodada de investimentos uma variável *proxy* para a velocidade de atração de recursos.

A Figura 3 mostra o design macro do estudo.

Figura 3 – Design macro do estudo



Fonte: Adaptado de Arroyo et al. (2019, p. 124235).

A predição do sucesso das *startups* mapeadas na janela de aquecimento foi feita entre os dias 30/6/2015 e 29/6/2020. Nesse intervalo, foram identificadas as *startups* da janela de aquecimento que tiveram sucesso na janela de predição, com a variável *target*

apresentando 1 (um) se a *startup* foi adquirida ou realizou *IPO* (*proxy* de sucesso neste trabalho), ou 0 (zero), caso contrário. Na próxima seção é apresentado o processo de tratamento dos dados, em sintonia com o design de pesquisa proposto.

4 Tratamento da base de dados

Nesta etapa foi realizado todo o pré-processamento dos dados, incluindo a limpeza da base de dados, transformação e criação de novos atributos, bem como a análise de potenciais *outliers*. Todas as etapas de tratamento da base de dados foram executadas por meio da linguagem Python de programação.

4.1 Limpeza inicial da base de dados

Para a predição do sucesso das *startups*, alguns atributos se destacaram como potenciais

variáveis preditoras, a saber: o número de rodadas de financiamento (*num_funding_rounds*), a categoria/indústria atendida pela *startup* (*category_groups_list*), o volume total arrecadado ao longo de todas as rodadas de financiamento recebidas (*total_funding_usd*), a data de fundação da *startup*, cuja ausência pode significar um registro equivocado (*founded_on*), o país de origem (*country_code*) e a data de sua última rodada de financiamento (*last_funding_on*)

A Tabela 6 sumariza o processo de limpeza inicial da base de dados.

Tabela 6 – Processo inicial de limpeza da base de dados

Etapa da limpeza	Número de registros excluídos	Tamanho da amostra	%
BASE DE DADOS ORIGINAL (tabela #organizations)		1.064.929	100%
Remoção de registros sem o número de rodadas de financiamento (*num_funding_rounds*)	890.917	174.012	16,34%
Remoção de registros sem o volume total de financiamento obtido pela *startup* (*total_funding_usd*)	47.294	126.718	11,90%
Remoção de *startups* sem identificação de nome (*name*)	4	126.714	11,90%
Remoção de registros sem identificação da data da última rodada de financiamento (*last_funding_on*)	0	126.714	11,90%
Remoção de *startups* que não tenham a sua data de fundação (*founded_on*)	3.671	123.043	11,55%
Remoção de *startups* sem identificação das indústrias que atendem (*category_groups_list*)	2.231	120.812	11,34%
Remoção de *startups* sem identificação do país de origem (*country_code*)	454	120.358	11,30%
Janela de aquecimento: apenas *startups* fundadas entre os dias 29/6/2010 e 29/6/2015, que não tenham sido fechadas nesse período nem recebido rodada de financiamento série C ou superior	76.483	43.875	4,12%
Manutenção apenas de *startups* com status de empresa, removendo-se assim as que apresentam status de investidor ou empresa/investidor	978	42.897	4,03%
Retirada de *startups* com os seguintes investimentos em sua primeira captação: *post ipo equity, post ipo debt, post ipo secondary e secondary market*	221	42.676	4,01%
Manutenção apenas de *startups* com primeiras rodadas de *funding* abaixo de série C (*angel, seed*, *series* A e *series* B)	12.696	29.980	2,82%
Retirada de *startups* com data de fundação "posterior" ao primeiro *round* de captação	2.083	27.897	2,62%
Manutenção apenas de *startups* com tempo até a primeira rodada de captação inferior a 720 dias (2 anos)	9.336	18.561	1,74%
TAMANHO DA AMOSTRA APÓS LIMPEZA INICIAL		18.561	1,74%

Fonte: Elaborado pelos autores (2021).

Decidiu-se também excluir 978 *startups* que apresentam status de investidora ou status híbrido, ou seja, empresa e investidora. Dessa forma, foram mantidas no estudo apenas *startups* que não vinculem atividades de investimento ao seu *core business*. Além disso, após análise das formas de captação no primeiro

funding, decidiu-se excluir *startups* com financiamentos típicos de empresas já maduras, como rodadas após um *IPO* e ligadas a captação via mercado mobiliário.

Após a limpeza inicial dos dados, chegou-se a 18.561 *startups* com dados em todos os atributos e que atendiam aos critérios estabelecidos na janela de aquecimento. Dando seguimento à etapa de pré-processamento dos dados, decidiu-se excluir da tabela *#organizations* colunas (atributos) sem aderência ao estudo.

4.2 Transformação e criação de novos atributos

Transformações e novas *features* foram criadas a partir das tabelas *#organizations*, *#funding_rounds* e *#people* para melhor caracterizar as rodadas de financiamento, as indústrias atendidas pelas *startups*, os locais de origem e o atributo-alvo (*target*).

O critério para a escolha e criação de atributos considerou atributos preditivos aderentes ao universo e cenário competitivo das *startups*. Foram considerados atributos importantes para os investidores, dentro do horizonte temporal proposto no estudo: valor do primeiro *funding*, tempo entre *funding rounds*, número de rodadas de investimento, número de colaboradores, número de mercados atendidos, país de origem e principal indústria.

A Tabela 7 mostra o esquema dos novos atributos.

Tabela 7 – Visão geral dos novos atributos

Rodadas de Financiamento	Indústrias atendidas	Local de origem	Atributo-alvo (*target*)
• Tempo até a primeira rodada de financiamento • Tempo médio entre rodadas de financiamento • Valor da primeira rodada de captação • Tipo da primeira rodada de financiamento	• Categoria principal • Número de indústrias atendidas	• Região de origem	• Sucesso (ou não) da *startup*

Fonte: Elaborado pelos autores (2021).

Sobre as rodadas de financiamento, decidiu-se explorar:

- O tempo até a primeira rodada de financiamento como *proxy* da velocidade de atração de recursos para a *startup*. Essa variável foi operacionalizada pela diferença entre a data da primeira rodada de financiamento e a data de fundação da *startup*. A data do primeiro *funding* foi obtida pela manipulação prévia da tabela *#funding_rounds*.
- O tempo médio entre rodadas também permite visualizar o espaçamento temporal da ação dos investidores junto às *startups*. Essa variável foi operacionalizada com dados da tabela *#organizations*, sendo o intervalo entre a última rodada de financiamento (‘*last_funding_on*) e a data de fundação da *startup* (‘*founded_on*) dividido pela

quantidade de *funding rounds* obtidos pela *startup* (‘*num_funding_rounds*).

- O valor da primeira rodada de captação se relaciona à capacidade de atração inicial de capital. Valores mais altos podem indicar que os investidores possuem uma visão otimista sobre o futuro da *startup*. Esse valor foi obtido pelo valor do primeiro *funding* de cada *startup* na tabela *#funding_rounds*, coletando-se a variável ‘*raised_amount_usd’*.
- As *startups* são financiadas de diversas formas, sendo capital semente (*seed*), investimento anjo (*angel*) e série A prevalentes na primeira rodada de financiamento das *startups* encontradas na plataforma Crunchbase, com destaque para o capital semente (75,52%). A criação dessa variável teve o intuito de averiguar se distintas formas de captação inicial desempenham papéis diferenciados para a predição do sucesso de *startups*.

Para as indústrias atendidas:

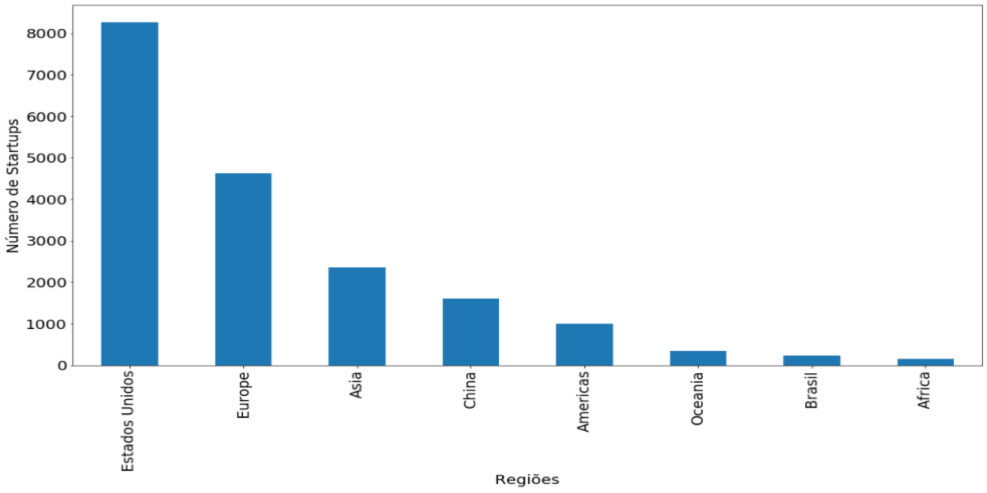
- Inicialmente, vale destacar que o atributo ‘*category_groups_list’*, na tabela *#organizations*, tinha 9.955 valores únicos, os quais foram identificados na fase de pré-processamento com o uso da linguagem de programação Python. Esses 9.950 valores são derivados de diversas combinações entre as 43 indústrias representadas na plataforma Crunchbase. As cinco indústrias com maior número de *startups* na amostra são: *commerce and shopping*, *apps*, *financial services*, *data and analytics*, e *advertising*; somadas representam 40% das *startups*. Nota-se grande concentração entre um e cinco mercados atendidos, ou ao menos declarados como potenciais mercados consumidores pelas próprias *startups* na plataforma Crunchbase.
- Similar ao trabalho de Bento (2017), a categoria principal se refere ao foco do posicionamento do negócio da *startup*, ou seja, primeira indústria declarada como *proxy* de seu *core business*.
- O número de indústrias atendidas refere-se à quantidade de indústrias que cada *startup* declara atender com seu modelo de negócios. Essa variável pode ser considerada como *proxy* da diversificação mercadológica de cada *startup*.

Sobre os locais de origem, decidiu-se pela seguinte transformação:

- A base de dados limpa possui *startups* de 141 países. Estados Unidos e Europa, com 8.267 e 4.619, respectivamente, representam cerca de 70% de todas as *startups* deste estudo. Dessa forma, decidiu-se agrupar os

países pelo seu continente de origem, com exceção dos Estados Unidos, China (que sozinha representa quase 9% da amostra), e Brasil (pela condução das etapas da pesquisa em território brasileiro). A Figura 4 mostra a distribuição das *startups* após a transformação do atributo '*country_code'* em continentes de origem.

Figura 4 – Locais de origem das *startups*



Fonte: Elaborado pelos autores (2021).

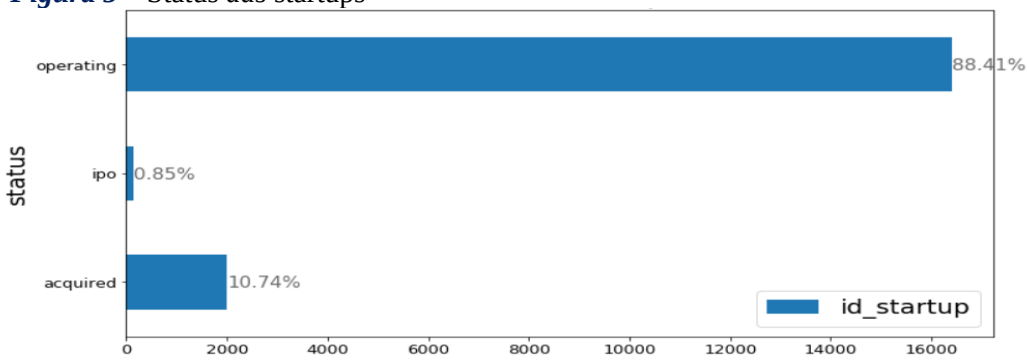
Quanto à variável que se deseja predizer (*target*), decidiu-se pela seguinte operacionalização, de acordo com o objetivo do estudo:

- *Startups* que conseguiram fazer um processo de *IPO* ou foram adquiridas serão relacionadas à "Classe 1", ou seja, denominadas com bem-sucedidas de acordo com o design desta pesquisa.

Startups ainda em operação, mas sem *IPO* ou aquisição serão alocadas para a "Classe 0" (sem sucesso, ao menos ainda).

Conforme se observa na Figura 5, a amostra comporta menos de 1% de *startups* que fizeram *IPO*. Tem-se 11,59% de *startups* bem-sucedidas pelo critério de sucesso estabelecido neste estudo, caracterizando um problema de classificação com dados desbalanceados.

Figura 5 – Status *das* startups



Fonte: Elaborado pelos autores (2021).

A Tabela 8 mostra os atributos do *dataset* da pesquisa após o processo de transformação e criação de novas *features*.

Tabela 8 – Atributos após transformações e criação de novas *features*

Atributo	Descrição	Tipo
Uuid	Código único para identificação de cada *startup* (chave primária da tabela *#organizations*)	Nominal
Name	Nome da *startup*	Nominal
categoria_ principal	Identificação do posicionamento mercadológico da *startup*, ou seja, a indústria principal que declara atender	Categórica
new_num_ mercados	Diversidade declarada de indústrias atendidas pelo modelo de negócio da *startup*	Categórica
num_ colaboradores	Número de colaboradores informados na plataforma. Essa variável é apresentada em faixas, por exemplo: 1-10 funcionários. Apenas mudança do nome do atributo '*employee count'*.	Categórica
num_funding_ rounds	Número total de rodadas de financiamento recebidas ao longo da trajetória da *startup*	Numérica
tipo_investimento_ primeiro_funding	Tipo de captação recebida na primeira rodada de financiamento da *startup* (por exemplo: *seed, angel, series* A, *series* B, etc)	Categórica
valor_primeiro_ funding	Valor recebido pela *startup* em sua primeira captação (em US$)	Numérica
tempo_ate_ primeiro_round	Tempo transcorrido (em dias) entre a fundação da *startup* e seu primeiro *round* de captação	Numérica
tempo_medio_ entre_fundings	Tempo médio (em dias) entre as rodadas de captação recebidas pela *startup*	Numérica
regiao_startup	Continentes de origem da *startup*, com exceção dos Estados Unidos e China, pela representatividade destes na base de dados, e do Brasil, pela pesquisa ser conduzida em território brasileiro	Categórica
status_sucesso	Classe 1 para *startups* com *IPO* ou adquiridas, significando sucesso no presente estudo. Classe 0 se seguem em operação, mas ainda sem aquisição ou *IPO* dentro da janela de simulação.	Binária (0, 1)

Fonte: Elaborado pelos autores (2021).

5 Apresentação e discussão dos resultados

Este item apresenta análises exploratórias sobre o sucesso na amostra de 18.561 *startups*, bem como os modelos de predição via *machine learning* e seus resultados.

5.1 Análise exploratória

Algumas análises exploratórias já foram realizadas na seção sobre Tratamento dos Dados. Nesta seção, a exploração dos dados avançará nas seguintes dimensões: a) Análise descritiva das variáveis numéricas; b) Análise de correlação entre as variáveis numéricas do *dataset* final; c) Análise de dispersão com visualização de *startups* bem-sucedidas (Classe 1) e sem sucesso (Classe 0); e d) Análise percentual da distribuição de *startups* em operação e adquiridas/*IPOs*, por indústria principal atendida.

Com isso, pretende-se visualizar a existência de padrões prévios antes da rodagem dos modelos de *machine learning* para classificação binária.

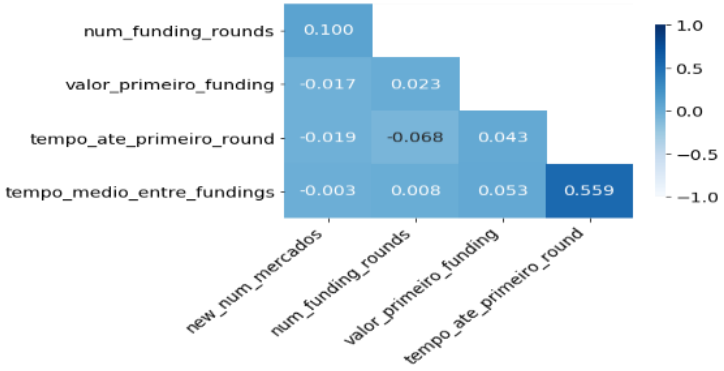
Tabela 9 – Análise descritiva inicial (variáveis numéricas do conjunto de dados final)

Atributo Numérico	Média	Desvio padrão	Coeficiente de Variação	Mínimo	Máximo
New_num_mercados	3,51	1,73	49,29%	1	14
Num_funding_rounds	2,57	1,81	70,43%	1	22
Valor_primeiro_funding (US$)	1.792.994	13.956.656	778,40%	468	1.080.000.000
Tempo_ate_primeiro_round (dias)	309	196	63,43%	1	719
Tempo_medio_entre_funding (dias)	380	205	53,95%	1	1579

Fonte: Elaborado pelos autores (2021).

A Tabela 9 apresenta uma análise descritiva das variáveis numéricas do *dataset*. Notam-se dados enviesados à direita nas variáveis numéricas, principalmente no valor da primeira rodada de financiamento. Não foram realizadas discretização dos dados e retirada de potenciais *outliers*, como fez Bento (2017), pois assume-se que os valores são possíveis dentro do cenário de investimentos em *startups* e que, por isso, podem ser mantidos nos modelos de predição.

Figura 6 – Correlação das variáveis numéricas



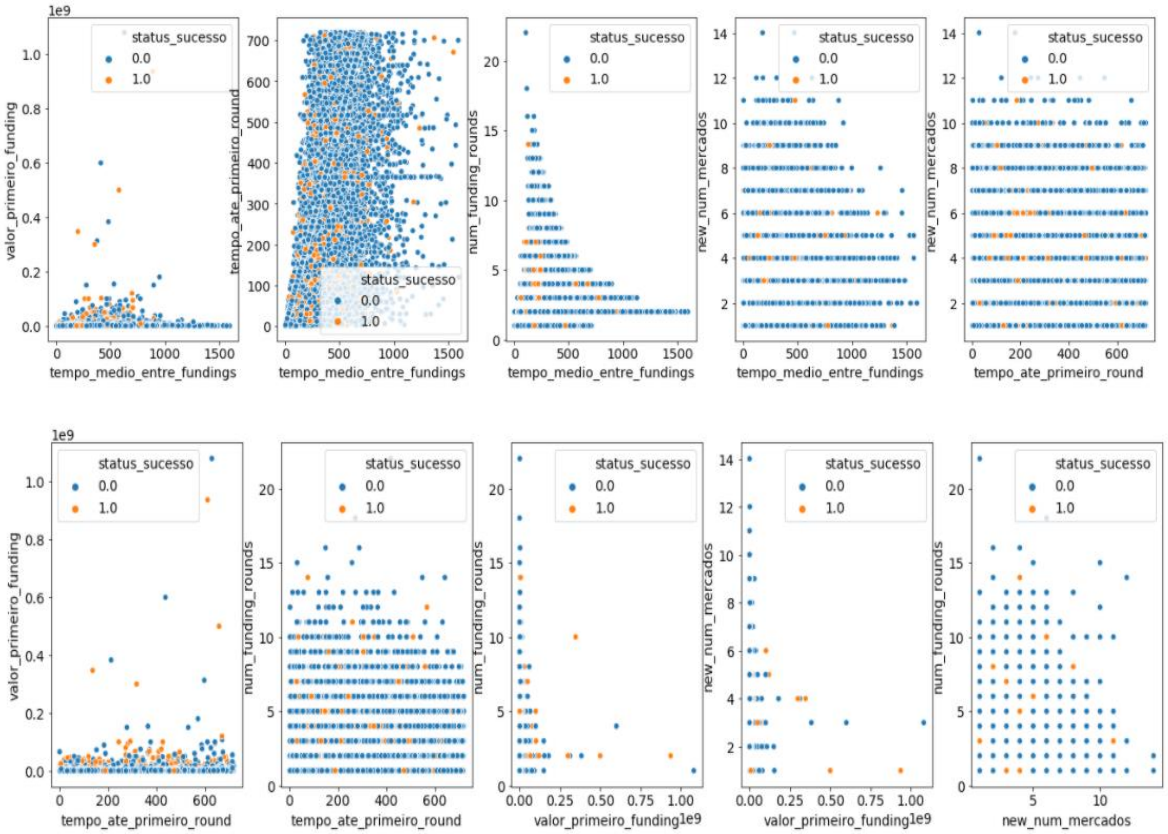
Fonte: Elaborado pelos autores (2021).

A Figura 6 apresenta uma análise de correlação entre os atributos preditivos numéricos. A correlação mais intensa foi entre o tempo até a primeira rodada de captação e o

tempo médio entre *funding rounds* (0,559). As demais relações não apresentam correlação aparente, demonstrando baixa multicolinearidade entre as variáveis preditivas.

Os gráficos de dispersão com pares de atributos preditivos (Figura 7) não apresentam clara distinção entre *startups* Classe 0 e Classe 1 (sucesso).

Figura 7 – Análise de dispersão



Fonte: Elaborado pelos autores (2021).

Essa dificuldade de distinguir entre as variáveis preditoras e a variável *target* é mais um incentivo para se tentar modelos de *machine learning* que busquem auxiliar a tarefa de classificação binária.

Ainda na fase de exploração de potenciais padrões, a Tabela 10 apresenta a distribuição do percentual de *startups* bem-sucedidas (*IPO* + Aquisição), *startups* com *IPO* e *startups* adquiridas, nas 10 indústrias mais representativas da amostra - cerca de 63% das 18.561 *startups*.

Tabela 10 – Distribuição das *startups* por status (top 10 indústrias na amostra)

Indústria	Número de *Startups*	% sucesso	% sucesso via IPO	% sucesso via Aquisição
Commerce and Shopping	2435	11,01	0,49	10,51
Apps	2069	11,79	0,68	11,12
Financial Services	1072	11,10	0,75	10,35
Data and Analytics	978	18,71	0,10	18,61
Advertising	911	13,94	0,55	13,39
Information Technology	881	13,28	0,23	13,05
Artificial Intelligence	878	13,67	0,23	13,44
Health Care	870	9,08	0,69	8,39
Consumer Electronics	827	9,07	0,60	8,46
Internet Services	804	11,94	0,75	11,19

Fonte: Elaborado pelos autores (2021).

Da lista das 10 principais indústrias, nenhuma ultrapassou 20% de sucesso na janela de simulação (média de 11,59%). Destacam-se *startups* voltadas para *Data and Analytics*, com 18,71% de sucesso. Verifica-se também que o sucesso, independentemente da indústria, provém mais dos processos de aquisição do que de *IPO*. Dentre as 47 indústrias presentes no *dataset*, nenhuma ultrapassou a taxa de sucesso de 20% dentro da janela de simulação deste estudo, exceto *Platforms* com apenas 2 *startups*.

5.2 Modelos de classificação binária via machine learning

Os algoritmos selecionados para a tarefa de predição do sucesso de *startups* seguiram a maioria dos modelos descritos na Tabela 2 da revisão da literatura, sendo eles: Regressão Logística (*baseline*), *Decision tree*, *Random Forest*, *XGBoost*: *Extreme Gradient Boosting*, *SVM*: *Support Vector Machine* e Rede neural (*MLP: Multi-Layer Perceptron*).

O *dataset* é altamente desbalanceado, com 16.409 registros alocados na Classe 0 e 2.152 na Classe 1, indicando uma proporção de desbalanceamento de 7,62:1. Como o interesse é a predição da classe minoritária, optou-se por tratar o desbalanceamento usando o objeto *RandomOverSample* da biblioteca *Imblearn* do Python, apenas nos dados de treinamento. *Random oversampling* consiste na seleção de exemplos da classe minoritária, com substituição e adição destes ao conjunto de dados de treinamento. Os dados foram balanceados após a separação do *dataset* em dados para treinamento (80%) e dados para teste

(20%), com estratégia de partição balanceada dos registros.

Para a rodagem dos modelos, criou-se um pré-processador para a preparação dos dados, incluindo *one-hot encoder* para transformação das variáveis categóricas, e normalização das numéricas com *MinMaxScaler*. Esse pré-processador foi utilizado nos *pipelines* do Python para rodagem dos modelos. Todos os modelos tiveram validação cruzada com 5 *folds*. Além disso, buscou-se realizar *tunning* dos seus hiperparâmetros por meio da função *GridSearch* do Python. A Tabela 11 apresenta os melhores parâmetros de cada modelo após o processo de *tunning*.

Tabela 11 – Melhores parâmetros dos modelos

Modelos	Parâmetros testados	Melhores parâmetros (via *GridSearch*)
Regressão logística (RL**)**	C: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]	C=3
***Decision Tree* (*DT*)**	*max_depth*: [3, 4, 5, 6, 7, 9, 11, 13, 15, 17]	*Tree max. depth* = 17
***Random Forest* (*RF*)**	*max_depth*: [3, 4, 5, 6, 7, 9, 11, 13, 15]	*Random Forest max. depth* = 15
***XGBoost* (*XG*)**	*learning_rate*: [0.01, 0.015, 0.02, 0.04, 0.06]	*learning rate* = 0,06
***Support Vector Machine* (*SVM*)**	C: [0.1, 1, 10, 100] *Gamma*: [1, 0.1, 0.01, 0.001], *kernel*: ['rbf']	C = 10 *Gamma* = 0,1 *kernel* = Rbf
Redes Neurais (RN)	*alpha*: [0.1, 0.01, 0.02], *hidden_layer_sizes*: [3, 5, 15, 25, 50, 100, 1000]	*alpha* = 0,01 *hidden layer sizes* = 1000

Fonte: Elaborado pelos autores (2021).

5.3 Resultados dos modelos de classificação

A Tabela 12 apresenta as principais medidas de avaliação do modelo para cada classe: precisão, *recall*, *f1-score* e *AUC* (*Area Under the Curve*).

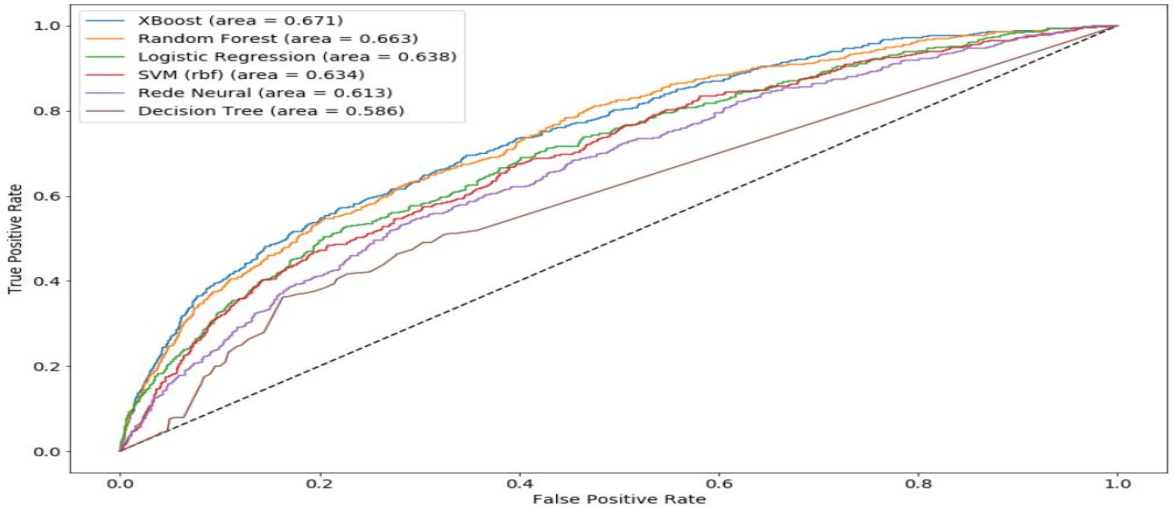
Tabela 12 – Índices para avaliação dos modelos de classificação

Modelo	*AUC*	Acurácia	Acurácia balanceada	*GMean*	Classe 0			Classe 1 (sucesso)		
					Precisão	***Recall***	**F1**	**Precisão**	***Recall***	**F1**
RL	0,638	0,628	0,638	0,638	0,93	0,63	0,75	0,19	0,65	0,29
DT	0,586	0,711	0,586	0,563	0,91	0,75	0,82	0,18	0,42	0,25
RF	0,663	0,779	0,669	0,645	0,93	0,81	0,87	0,26	0,51	0,35
XG	0,671	0,715	0,671	0,669	0,94	0,73	0,82	0,23	0,61	0,33
SVM	0,634	0.592	0,634	0,631	0,93	0,58	0,71	0,18	0,69	0,28
RN	0,613	0,733	0,613	0,592	0,92	0,65	0,84	0,21	0,46	0,28

Fonte: Elaborado pelos autores (2021).

Os classificadores que se destacaram, em termos gerais, foram o *Extreme Gradient Boosting (XG)* e o *Random Forest (RF)*, considerados como classificadores conjuntos (*ensemble classifiers*). Tais resultados seguiram a mesma direção do estudo conduzido por Arroyo et al. (2019).

Figura 8 – Curvas ROC dos modelos analisados



Fonte: Elaborado pelos autores (2021).

O índice *AUC* (*Area Under the Curve*), que considera a área sob a curva em um gráfico que relaciona a taxa de falsos positivos (especificidade) e a taxa de verdadeiros positivos (sensibilidade), ficou em quase 70% (0,671) no caso do *Extreme Gradient Boosting*, e 0,663 para o *Random Forest*, que se destacaram entre os classificadores (vide Figura 8).

Os classificadores *SVM* e *DT* tiveram menor índice *AUC*. Na acurácia balanceada, os mesmos classificadores tiveram destaque (*XG* = 0,671 e *RF* = 0,663).

Para o objetivo deste estudo, a precisão por classe, principalmente da classe minoritária (sucesso das *startups*) é o índice mais importante, pois indica o percentual de verdadeiros positivos no total de registros classificados como positivos. Os classificadores tiveram precisão entre 0,18 e 0,26 para classificação na Classe 1. Destacaram-se *XG* e *RF*, com 23% e 26%, respectivamente.

A escolha de um melhor classificador é complexa, pois um único índice não engloba todas as possibilidades de análise. Também depende do objetivo de cada estudo. A Tabela 13 apresenta um conjunto de índices para a análise conjunta dos dois melhores classificadores neste estudo.

Tabela 13 – Análise comparativa XG e RF

Classificador	*AUC*	*TPR (Recall)*	*FNR*	Precisão
RF	0,663	51,2%	18,6%	26,4%
XG	0,671	61,4%	27,1%	22,9%

Fonte: Elaborado pelos autores (2021).

Como o *Recall* (razão entre o número de *startups* classificadas na Classe 1 pelo número total de startups que, de fato, foram bem-sucedidas) é menos relevante para o presente estudo, pois números acima de 50% demonstram uma boa classificação de *startups* elegíveis para investimentos bem-sucedidos, o foco recairá na Precisão. Nesse quesito sobressaiu o classificador *Random Forest* com uma precisão de 26,4% (com menor *FNR – False Negative Rate*, indicando menor probabilidade de indicação aos investidores de *startups* com menor potencial de sucesso).

Em rápida análise, os números parecem baixos, mas vale lembrar que o índice médio de sucesso entre as 18.571 *startups* da amostra é de 11,59%. Tomando-se como exemplo a precisão do modelo *Random Forest* (26,4%), alcança-se uma melhoria de 128% frente aos 11,59% (média de sucesso das *startups*). Do ponto de vista do investidor, o filtro gerado pelo classificador coloca a proporção de escolha de *startups* para *funding* em uma proporção de 3,8 para 1, ou seja, chance de sucesso de 1 *startup* para cada 3,8 *startups* adicionadas ao portfólio do investidor.

Esse resultado (26,4%), mesmo com diferenças no design de pesquisa e tratamento dos dados, é próximo ao obtido por Arroyo et al. (2019), que alcançou uma precisão de 33% para aquisição de *startups* (modelo *Random Forest*) e 44% para *IPO* (modelo *Random Forest*).

Vale também destacar dois pontos do design de pesquisa que podem colaborar para atingir taxas de precisão dessa magnitude:

a) O processo de tratamento de dados desta pesquisa foi bastante rígido quando comparado aos estudos apresentados na Tabela 2. Decidiu-se eliminar todos os registros (*startups*) com ausência de valores nos atributos preditores (ver Tabela 8 para mais detalhes). Essa decisão foi motivada pela análise exploratória dos dados e pela própria natureza dos dados presentes na plataforma Crunchbase. Como existe autodeclaração dos dados na plataforma, decidiu-se apenas pela manutenção de *startups* com dados em todos os atributos, minimizando-se assim a inclusão de registros falhos.

Essa decisão segue a premissa de que *startups* com dados completos, em todas as *features* analisadas, possuem informações mais confiáveis.

b) Foram incluídos atributos preditivos que tivessem consistência com o universo e o cenário competitivo das *startups*, conforme mostra a Tabela 8. Alguns estudos anteriores incluíram informações indicadas na plataforma Crunchbase de baixa relevância para os investidores, como existência ou não de perfil *Linkedin*/*Facebook*, contato

telefônico ou mesmo e-mail, (Arroyo et al. 2019). Tais fatores não agregam diferenciais competitivos às *startups*, mesmo que – eventualmente - possam trazer algum tipo de melhoria nos modelos de classificação.

Os modelos baseados em árvores, também chamados de *White-Box Models*, possibilitam a análise da importância dos atributos (*features*) no modelo de classificação.

Tabela 14 – Importância dos atributos (*feature importance*)

Random Forest (RF)		***Extreme Gradient Boosting (XG)***	
Peso	**Atributo**	**Peso**	**Atributo**
19,38%	Valor_primeiro_funding	**10,83%**	Região_startup_Estados Unidos
14,11%	Tempo_medio_entre_fundings	**3,62%**	Valor_primeiro_funding
10,73%	Tempo_ate_primeiro_funding	**2,82%**	Num_colaboradores_1-10
7,57%	Num_funding_rounds	**2,78%**	Num_funding_rounds
6,11%	Região_startup_Estados Unidos	**2,77%**	Região_startup_China
5,42%	New_num_mercados	**2,69%**	Num_colaboradores_11-20
3,04%	Num_colaboradores_1-10	**2,65%**	Tipo_investimento_primeiro_funding_series_a
2,31%	Tipo_investimento_primeiro_funding_series_a	**2,36%**	Tipo_investimento_primeiro_funding_angel
2,03%	Tipo_investimento_primeiro_funding_angel	**1,89%**	Categoria_principal_Data and Analytics
1,91%	Região_startup_China	**1,85%**	Categoria_principal_Apps

Fonte: Elaborado pelos autores (2021).

De forma geral, os atributos mais valorizados pelos dois modelos baseados em árvores são semelhantes. Os dez atributos mais importantes no modelo *Random Forest* explicam 72,61% da classificação, enquanto no modelo *Extreme Gradient Boosting*, apenas 34,26%, indicando maior dispersão explicativa dos atributos.

O valor da primeira captação ocupa a primeira ou segunda posição nos classificadores. Como país de origem da *startup*, os Estados Unidos se destacam no modelo *XG*, com a China ocupando a quinta posição. Aliás, Estados Unidos e China são atributos preditivos relevantes nos dois modelos (*XG* e *RF*). Destaca-se também a natureza da primeira rodada de captação ser do tipo série A, um indicativo da importância de valores iniciais maiores para uma caminhada bem-sucedida da *startup*.

A variável "new_num_mercados", relacionada à diversidade de mercados atendidos (autodeclaração das *startups*), não se mostrou relevante para o classificador *XG*, apenas para o *RF*. Já a quantidade de colaboradores ("num_colaboradores") teve presença nos dois, em faixas baixas no *RF* e desconhecida no *XG*, indicando que *startups* bem-sucedidas contam com um número menor de colaboradores.

A questão do tempo de *funding*, tanto o período médio entre rodadas de captação quanto o tempo decorrido entre a criação da *startup* e sua primeira captação, não se mostrou relevante no modelo *XG*. Por outro lado, a questão temporal teve grande relevância no modelo de

classificação *RF*, ocupando a segunda e terceira colocação.

Interessante notar a baixa incidência do atributo caracterizador da principal indústria atendida pela *startup* no classificador *RF*. Esse achado é contraintuitivo, pois a literatura de negócios indica forte associação entre o tipo de indústria e sua capacidade de ser atrativa na geração de resultados financeiros positivos (Porter, 2005). No caso do *XG*, a indústria de *Data and Analytics* está entre os dez atributos mais relevantes. Vale lembrar que essa indústria detém o maior percentual de *startups* bem-sucedidas na amostra, 18.71%.

No estudo de Arroyo et al. (2019), outras variáveis se mostraram mais importantes nos modelos *RF* e *XG*, como idade da *startup*, ter um perfil ativo no *LinkedIn*, a quantidade de fundadores, número de telefone e *e-mail* informado na plataforma Crunchbase, o valor total arrecadado nas rodadas de financiamento; e posicionamento da *startup* nas indústrias de saúde, ciência e engenharia, no caso do modelo *XG*. No presente estudo, diferentes indústrias, como *Data Analytics* e *Apps*, explicam os resultados do modelo *XG*, potencialmente demonstrando a velocidade das mudanças e reorientação de investimentos no ecossistema empreendedor. Além disso, conforme mencionado no capítulo introdutório, variáveis pouco relevantes para investidores que foram introduzidas como preditoras no modelo de Arroyo et al. (2019) mostraram-se relevantes para a explicação dos modelos de predição, como a inclusão de perfil no *LinkedIn*, *e-mail* e telefone na plataforma Crunchbase.

Este estudo incluiu apenas *startups* fundadas entre junho de 2010 e junho de 2015, utilizando um conceito de janela temporal similar ao design do estudo de Arroyo et al. (2019). Com isso, buscou-se minimizar o viés de sobrevivência das *startups* inseridas no *dataset* final, bem como dados inconsistentes pela natureza retroativa da inserção de dados. Bento (2017), por exemplo, utilizou *startups* americanas fundadas entre 1985 e 2014, data bastante anterior à criação da plataforma Crunchbase.

Os modelos que, no geral, tiveram melhores índices foram os baseados em *ensembles* de árvores: o *Extreme Gradient Boosting* (*XG*) com *AUC* de 0,671 e o *Random Forest*, com 0,663. Algoritmos do tipo

6 Considerações finais

Este trabalho, inspirado pelo design de pesquisa com janelas temporais proposto por Arroyo et al. (2019), analisou seis modelos de *machine learning* para a predição do sucesso de *startups*, a saber: Regressão Logística, *Decision Tree*, *Random Forest*, *Extreme Gradient Boosting*, *Support Vector Machine*, e Redes Neurais.

Modelo de redes neurais do tipo *multi layer perceptron*, pouco usado na tarefa de prever o sucesso de *startups* (vide Tabela 2), mostrou-se promissor para estudos futuros. No presente estudo, gerou 21% de precisão, ocupando a terceira posição na medida de avaliação entre os classificadores testados.

ensemble classifiers, ou de aprendizado conjunto, também se destacaram nos estudos de Bento (2017), Pan, Gao e Luo (2018) e Arroyo et al. (2019).

O classificador *Random Forest* atingiu uma Taxa de Verdadeiros Positivos de 51,2% e Taxa de Falsos Negativos de 18,6%. A precisão do *RF* para classificar corretamente a Classe 1, ou seja, *startups* bem-sucedidas, foi de 26,4% (superior ao *XG*, que obteve 22,9% de precisão).

Em vista da quantidade de dados faltantes na plataforma Crunchbase, a estratégia conservadora usada neste estudo para tratá-los considerou apenas *startups* com valores presentes em todos os atributos preditivos. Além disso, conforme a Tabela 8, incluíram-se apenas variáveis preditivas ligadas ao contexto de investimentos em *startups* envolvendo as seguintes categorias: tipo de indústria e número de mercados atendidos, número de colaboradores, rodadas de financiamento (tempo e valores) e local de origem.

A modelagem e interpretação de modelos de *machine learning* requer técnica e domínio do contexto de negócios envolvido no problema de classificação ou regressão. Pelo *background* dos autores desta pesquisa, este estudo contribui ao aproximar o campo da Ciência de Dados ao da Administração. Esperamos, com isso, fomentar pesquisas conjuntas para avançar o conhecimento acadêmico.

Do ponto de vista dos investidores, que geralmente investem em mais de uma *startup* e montam um portfólio delas, uma precisão de 26,4% equivale a uma taxa de acerto de 1 para 3,8 *startups* na carteira. Pelo design desta pesquisa, a Classe 0, considerada como fracasso, é constituída por *startups* que continuam em operação e recebendo *funding*, muito longe de serem caracterizadas como uma opção de investimento a ser descartada.

Assim, o uso de modelos de *machine learning* para classificação binária tem sua aplicabilidade para os investidores, principalmente na fase de mapeamento de *startups* promissoras, com taxa de acerto de 1 para cada 3,8 *startups* identificadas pelo classificador. Tal precisão é valiosa em um cenário extremamente competitivo, onde mais de 90% das *startups* falham e menos de 1% se tornam unicórnios, ou seja, *startups* com valor de mercado (*valuation*) acima de um bilhão de dólares (Profitfromtech, 2020).

O mecanismo de classificação proposto neste estudo pode agregar um filtro inicial (uma espécie de *screening* – varredura) de opções, para que os investidores possam alocar mais tempo em análises posteriores envolvendo a parte financeira, operacional e mercadológica, buscando um aprofundamento analítico na busca de *targets* com maior probabilidade de sucesso.

As mudanças realizadas no design de pesquisa proposto por Arroyo et al. (2019) abrem frentes para pesquisadores avançarem, tanto no processo de tratamento dos dados faltantes usando a estratégia conservadora deste estudo, quanto na criação de novos atributos pertinentes ao ecossistema empreendedor.

Outras contribuições são a estratégia para lidar com a natureza desbalanceada das classes, os testes com outros modelos, e novas opções para escolha dos melhores hiperparâmetros.

Referências

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *IEEE Access*, *7*, 124233–124243. https://doi.org/10.1109/ACCESS.2019.2938659

Basole, R. C., Park, H., & Chao, R. O. (2019). Visual Analysis of Venture Similarity in Entrepreneurial Ecosystems. *IEEE Transactions on Engineering Management*, 66(4), 568–582. https://doi.org/10.1109/TEM.2018.2855435

Bento, F. R. S. R. (2017). *Predicting Start-up Success with Machine Learning*. Master Program in Information Management. Instituto Superior de Estatística e Gestão da Informação. Universidade Nova de Lisboa. Recuperado de: https://run.unl.pt/bitstream/10362/33785/1/TGI0132.pdf. Acesso em 14/mai/2020.

Blank, S. (2013). *The Four Steps to the Epiphany: successful strategies for products that win*. Pescadero: K&S Ranch Press.

CB Insights. (2020). *The Complete List of Unicorn Companies*. Recuperado de: https://www.cbinsights.com/research-unicorn-companies. Acesso em 10/dez/2020.

Cremades, A. (2016). *The Art of Startup Fundraising: pitching investors, negotiating the deal, and everything else entrepreneurs need to know*. Hoboken: John Wiley & Sons.

Crunchbase. Crunchbase *Platform*. Disponível em: https://www.Crunchbase.com/. Acesso em: 12 mar. 2020.

Dalle, J.-M., Den Besten, M. & Menon, C. (2017). Using Crunchbase for economic and managerial research. *OECD Science, Technology and Industry Working Papers*. Recuperado de: https://pdfs.semanticscholar.org/aa83/4b1ddd1d6c96bde1c6e526be6bb2a99ad011.pdf. Acesso em 07/jun/2020.

Ertel, W. (2017). *Introduction to Artificial Intelligence*. 2nd ed. London: Springer.

Facelli, K., Lorena, A. C., Gama, J., & de Carvalho, A. C, P. L. F. (2019). *Inteligência Artificial: uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC.

Gastaud, C., Carniel, T., & Dalle, J.-M. (2019). The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage. *arXiv preprint arXiv:1906.03210*. Recuperado de: https://arxiv.org/abs/1906.03210. Acesso em 03/jun/2020.

Gereto, M. A. S. (2019). *Caracterização dos ciclos de investimentos de venture capital em startups brasileiras em termos de rodadas de investimento e estratégias de desinvestimento a partir de dados da* Crunchbase**.** Dissertação de mestrado em administração. FGV - Faculdade Getúlio Vargas, 2019. Recuperado de; http://bibliotecadigital.fgv.br/dspace;/handle/10438/27468. Acesso em 01/jun/2020.

Hsieh, K.-H., & Li, E. Y. (2017). Progress of Fintech industry from venture capital point of view. *In: Proceedings of The 17th International Conference on Electronic Business*. ICEB, Dubai, p. 297-301. Recuperado de: http://iceb.johogo.com/proceedings/2017/ICEB_2017_paper_36-WIP.pdf. Acesso em 3/jun/2020.

Kemeny, T., Nathan, M., & Almeer, B. (2017). Using Crunchbase to explore innovative ecosystems in the US and UK. *Birmingham Business School Discussion Paper Series*. Recuperado de: http://epapers.bham.ac.uk/3051/1/bbs-dp-

2017-01-nathan.pdf. Acesso em 01/abr/2020.

Kosterich, A., & Weber, M. S. (2018). Starting up the News: The Impact of Venture Capital on the Digital News Media Ecosystem. *International Journal on Media Management, 20*(4), 239–262. https://doi.org/10.1080/14241277.2018.1563547

Kubat, M. (2017). *An Introduction to Machine Learning*. 2nd ed. Suiça: Springler.

Liang, E., & Daphne Yuan, S.-T. (2013). Investors Are Social Animals: Predicting Investor Behavior using Social Network Features via Supervised Learning Approach. *In: Proceedings of the Workshop on Mining and Learning with Graphs (MLG-2013), Chicago*. Recuperado de: http://chbrown.github.io/kdd-2013-usb/workshops/MLG/doc/liang-mlg13.pdf. Acesso em 03/jun/2020.

Losada, B. (2020). *Finanças para Startups: o essencial para empreender, liderar e investir em startups*. São Paulo: Editora Saint Paul.

National. *Small Business Failure Rate*. Recuperado de: https://www.national.biz/2019-small-business-failure-rate-startup-statistics-industry/. Acesso em 06/abr/2020.

Nylund, P. A., & Cohen, B. (2017). Collision density: driving growth in urban entrepreneurial ecosystems. *International Entrepreneurship and Management Journal, 13(*3), 757–776. https://doi.org/10.1007/s11365-016-0424-5

Pan, C., Gao, Y., & Luo, Y. (2018). Machine Learning Prediction of Companies 'Business Success. *CS229: Machine Learning*, Stanford University. Recuperado de: http://cs229.stanford.edu/proj2018/report/88.pdf. Acesso em 25/mar/2020.

Porter, M. E. (2005). *Estratégia Competitiva*. Rio de Janeiro: Campus.

Profitfromtech (2020). *The Ultimate List of Startup Statistics for 2020.* Recuperado de: https://www.profitfromtech.com/startup-statistics/. Acesso em 01/out/2020.

Ries, E. (2012). *A Startup Enxuta*. 1ª ed. São Paulo: Leya.

Shan, Z., Cao, H., & Lin, Q. (2014). Capital Crunch: Predicting Investments in Tech Companies. *CS221 Project:* Crunchbase *Investment Prediction*. Training, 5831(32462), 32462. Recuperado de: http://www.zifeishan.org/files/capital-crunch.pdf. Acesso em 12/jun/2020.

Skiena, S. S. (2017). *The Data Science Design Manual*. Suiça: Springer.