ÁREA TEMÁTICA:
ÁREA 11: AI – ADMINISTRAÇÃO DA INFORMAÇÃO

TÍTULO:

MACHINE LEARNING PARA A PREDIÇÃO DE SUCESSO DE *STARTUPS*: REVISÃO SISTEMÁTICA DA LITERATURA

Resumo
Este estudo visa comparar, por meio de uma revisão sistemática da literatura, estudos que agregaram modelos de *machine learning* para a predição do sucesso de *startups* com o uso da base *Crunchbase* (base de dados mundial sobre startups). Para tanto, foi realizada uma revisão sistemática nas bases Google Scholar e Ebsco em abril de 2020. Após leitura completa de 21 artigos pré selecionados, 8 atenderam aos critérios de elegibilidade (uso da base *Crunchbase* como fonte de dados para rodagem dos modelos de predição e uso de modelos de *machine learning*). Destes, coletou-se os objetivos, tipificação da amostra, variável de interesse do estudo, técnicas de *machine learning* utilizadas e os principais resultados dos modelos de predição. Como resultados, percebe-se o uso majoritário de IPO ou aquisição da startup como *proxy* para sucesso, destaque para uso de algoritmos supervisionados com classificação binária e uso de *deep learning* apenas para estudos nos últimos dois anos. A revisão não apontou estudos nacionais e autores do campo da Administração, apenas pesquisadores relacionados à ciência de dados e computação.

Palavras-chaves: Startup, Aprendizado de Máquina, Predição de Sucesso.

Abstract
This study aims to compare, through a systematic literature review, studies that added machine learning models to predict the success of startups with the use of the Crunchbase. To this end, a systematic review was carried out on the Google Scholar and Ebsco databases in April 2020. After a complete reading of 21 pre-selected articles, 8 met the eligibility criteria (use of the Crunchbase database as a data source for running the prediction models and machine learning models). From these, the objectives, sample typification, variable of interest of the study, machine learning techniques used and the main results of the prediction models were collected. As a result, we perceive the majority use of IPO or acquisition of the startup as a proxy for success, highlighting the use of supervised algorithms with binary classification and the use of deep learning only for studies in the last two years. The review did not mention national studies and authors in the field of Administration, only researchers related to data and computer science.

Key-words: Startup, Machine Learning, Success Prediction.

1 INTRODUÇÃO

Nos últimos anos, o número de unicórnios, ou seja, *startups* com valor de mercado acima de um bilhão de dólares, vem aumentando. Em abril de 2020, 464 *startups* foram enquadradas nesta categoria, com *valuation* acumulado de aproximadamente US$ 1.36 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). Esta 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 necessidade dos consumidores, de condições e forma de gerenciamento, entre outras.

Fundos de Venture Capital (VCs) realizam investimentos com 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 deste cenário, de possível alto retorno e alto risco, a predição de *startups* com maiores chances de sucesso pode colaborar para o ecossistema empreendedor, reunindo investidores a iniciativas com maior potencial de rentabilização e trazendo aos empreendedores uma bussola para checagem de sua capacidade de sucesso durante o desenvolvimento de suas iniciativas.

O sucesso de uma *startup* pode ser definido de muitas formas, como por exemplo, 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*), pela abertura de capital via IPO.

Este estudo assumirá a perspectiva do investidor, com o *exit* como *proxy* para o sucesso. *Exit* é o termo usado para a saída do investidor, isto 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 ações para auferir ganhos nas transações.

A literatura acadêmica, com foco em *machine learning* para predição de sucesso de *startups*, possui frentes de pesquisa nesta direção (LIANG; DAPHNE YUAN, 2013, SHAN; CAO; LIN, 2014, BENTO, 2017; PAN; GAO; LUO, 2018; ARROYO et at.; 2019, GASTAUD; CARNIEL; DALLE, 2019). Na última década uma grande variedade de modelos foi testada para esta tarefa, incluindo *Naive Bayes*, *Decision Trees*, *Random Forest*, *Supported Vector Machine*, entre outros.

O objetivo geral deste estudo consiste em analisar e comparar, por meio de uma revisão sistemática da literatura, estudos que agregaram modelos de *machine learning* para a predição do sucesso de *startups* com o uso da base Crunchbase: design de pesquisa, tipo de modelo usado e operacionalização de sucesso de startups.

2 REVISÃO TEÓRICA

Nesta seção será abordado o conceito de *startups* e veículos para financiamento de suas atividades 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 multiplicam nas últimas décadas, 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 acesso a grande quantidade e variedade de dados estruturados com potenciais usos para múltiplos stakeholders: fundadores em busca de financiamento, investidores a procura de *startups*, pessoas em busca de ideias para novos empreendimentos e pesquisadores acadêmicos que se interessam pela dinâmica e ecossistema das *startups*.

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 que usaram 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*, com informações sobre 906.248 fundadores (extração realizada pelo autor, com a devida autorização da equipe *Crunchbase* para fins acadêmicos). 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*.

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 (VCs) 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, ventures capital, Family offices e alguns fundos de *private equity.*
Series B	A rodada de captação neste momento alcança maiores volumes. 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.

	Descrição	Fontes ($) mais comuns
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 um 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)

Nota-se que ao longo do ciclo de vida da *startup*, caracterizado como *early*, *growth* e *mature stage*, o volume das rodadas de financiamento tendem a aumentar, pela expectativa de crescimento das *startups* e pelos resultados mercadológicos e financeiros que apresentam aos potenciais investidores.

Para exemplificar esta lógica de *funding*, no cenário nacional, Gereto (2019) analisou o ciclo de investimento e desinvestimento de 436 startups brasileiras registradas na base de dados Crunchbase. Neste estudo visualizou uma média de 1.7 rodadas de investimento do tipo seed, 2.2 rodadas em early e 3.6 rodadas em later stage, com faixas crescentes captadas ao longo do ciclo de vida da startup: US$ 100.000,00 e US$ 1.000.000,00 (seed), US$ 1.000.000,00 e US$ 10.000.000,00 (early stage) e US$ 10.000.000,00 e US$ 100.000.000,00 (later stage).

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. Em rodadas mais avançadas é interessante perceber um potencial efeito rede dos investidores na busca por dinheiro novo que contribua para o sucesso de sua investida. A *startup* começa a ficar muito grande, na visão dos investidores, para que o investimento realizado não gere retorno.

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 ventures capital para posições-chave na gestão da *startup* (CREMADES, 2016).

Ao longo do ciclo de vida das *startups*, 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, em caso de fracasso da *startup*, os investimentos realizados são perdidos, gerando um *write-off* do investimento dos fundos (LOSADA, 2020, p.129).

2.2 Modelos de *machine learning*

Machine Learning (ML) é “o estudo de algoritmos de computador que melhoram automaticamente através da experiência” (ERTEL, 2017, p. 178).

Similarmente, Facelli et al. (2019, p. 3) argumenta 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.

A Figura 1 apresenta a lógica simplificada do processo de aprendizagem:

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, há a tentativa de predição de uma variável alvo a partir de um conjunto de atributos, por meio da 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 de 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 à priori, busca-se padrões entre atributos ou objetos contidos no conjunto de dados.

Como foco deste estudo consiste na predição do sucesso de *startups*, há muitos modelos de *machine learning* com características preditivas vistos na seção 2.1.2. Skiena (2017) argumenta que dificilmente um algoritmo de ML é superior a todos, pela diversidade de domínios, contexto do problema de decisão e da própria característica dos dados.

A Tabela 2 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 2 - 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 (SVM)	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 2. Algumas técnicas propiciam maior facilidade para a interpretação dos resultados (por exemplo, Árvore de Decisão). Já outras, como *Deep Learning*, possuem alto poder de predição e baixa facilidade de interpretação.

3. Método

Para condução desta revisão sistemática, recorreu-se às seguintes etapas de pesquisa: critérios de elegibilidade, critérios de busca e fontes de informação, seleção de estudos e coleta de dados, bem como medidas de comparação.

Quanto aos critérios de elegibilidade, decidiu-se pela inclusão de estudos recentes (últimos 10 anos). Além disso, incluiu-se apenas artigos que usaram o repositório de dados sobre *startups Crunchbase*, como fonte primária de dados para predição do sucesso de startups e que usaram modelos de machine learning, independentemente do algoritmo usado.

Utilizou-se como critério de busca as seguintes palavras-chaves: *startup sucess prediction* AND *machine learning*. Os estudos foram coletados usando-se o *Google Scholar* e a base de dados *Ebsco*, durante o mês de abril de 2020.

Após leitura do título e resumo, 21 estudos foram previamente selecionados por terem foco na predição de sucesso de startups no mundo. Após leitura completa dos artigos, 8 atenderam aos critérios de elegibilidade (uso da base *Crunchbase* como fonte de dados para rodagem dos modelos de predição e uso de modelos de *machine learning*). Destes, coletou-se os objetivos, tipificação da amostra, variável de interesse do estudo, técnicas de machine learning utilizadas e os principais resultados dos modelos de predição.

Para medida de comparação, adotou-se métricas de desempenho dos algoritmos de *machine learning*, como acurácia e precisão. Além disso, buscou-se comparar artigos com técnicas mais clássicas (regressão logística, *naive bayes*, *support vector machine* e baseados em árvores) e uso de *deep learning*.

4. Resultados

Nos últimos anos, muitos estudos sobre *startups* têm usado técnicas de *machine learning* como apoio à geração de insights e conhecimento sobre *startups*. 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.

O uso de técnicas de *machine learning* para amostras pequenas só foi encontrado em dois estudos. Santos da Silva (2016) buscaram sua amostra de 55 respondentes por meio de incubadoras e parques tecnológicos portugueses. Martinez (2019), conseguiu 91 *startups* após endereçar sua pesquisa para 969 *startups* holandesas presentes na base de dados techleap.nl.

A Tabela 3 sintetiza os oito estudos relacionados à predição do sucesso de *startups* com técnicas de *machine learning* (com o uso da base Crunchbase).

Tabela 3 - Predição de sucesso de *startups*: síntese de estudos relacionados

Referência	Objetivo	Base de Dados	Amostra	Variável de Interesse	Técnicas de *Machine Learning*
Xiang et al. (2012)	Explorar a modelagem de itens de textos em artigos de notícias para aprimorar os atributos tradicionais da previsão de fusões e aquisições.	*Techcrunch* e *Crunchbase* (um dos primeiros artigos a utilizar esta base de dados)	59631 *startups*, 105795 pessoas, 5915 aquisições, 38617 artigos Techcrunh	Aquisição da *startup*	**Bayesian Networks**, SVM, Logistic Regression
Liang e Daphne Yuan (2013)	Investigar o papel das relações sociais entre investidores e empresas para a predição do comportamento de investimento.	*Crunchbase*	11916 *startups*, 12127 pessoas, 1122 organizações financeiras (com 4 graus de separação do Facebook)	Ocorrência de investimento	**SVM** (rbf como kernel), **Árvore de Decisão** (algoritmo CART) e Naive Bayes (modelo Bernoulli)
Shan, Cao e Lin (2014)	Prever se um investidor investirá em uma start-up específica com base em sinais textuais, topológicos e específicos do domínio, tanto do investidor quanto da start-up	*Crunchbase*	214290 *startups*, 286659 pessoas, 31942 investimento s	Investidor efetiva investimento em *startup*	**Regressão Logística**, CRF
Bento (2017)	Desenvolver um modelo preditivo para classificar uma *startup* como bem-sucedida, ou não (classificação binária)	*Crunchbase*	86588 *startups* (estados americanos)	IPO ou Aquisição da startup	**Floresta Aleatória**, SVM e Regressão Logística

Referência	Objetivo	Base de Dados	Amostra	Variável de Interesse	Técnicas de *Machine Learning*
Pan, Gao e Luo (2018)	Predizer o sucesso de *startups*, definido como um evento que dá uma grande quantia aos fundadores, investidores e primeiros funcionários da empresa, especificamente através de um processo de fusões e aquisições ou um IPO	*Crunchbase*	+60.000 *startups*	Processo de M&A ou IPO	**KNN,** Floresta Aleatória e Regressão Logística
Sharchilev (2018)	Prever se uma empresa, que já garantiu financiamento inicial (semente ou anjo), atrairá uma nova rodada de investimentos (série A ou superior) em um horizonte de um ano.	*Crunchbase* (menções no *LinkedIn* e *Business Insider*, por exemplo)	21.947 *startups*	Rodada de financiamento em um ano	WBSSP, incluindo regressão logística, GBDT - Catboost
Arroyo et al. (2019)	Desenvolvimento e avaliação de uma abordagem orientada a dados que usa aprendizado de máquina para ajudar os investidores de VC a explorar e selecionar as melhores empresas para apoiar.	*Crunchbase*	120.507 *startups*, 34.180 *funding* rounds	Aquisição, rodada de financiamento, IPO, fechamento ou nenhum evento dentro da janela de simulação.	**Gradient Tree Boosting**, Árvore de Decisão, Floresta Aleatória e SVM.

Referência	Objetivo	Base de Dados	Amostra	Variável de Interesse	Técnicas de *Machine Learning*
Gastaud, Carniel e Dalle (2019)	Predizer o sucesso de startups na arrecadação de investimentos em diferentes estágios (*early*, *growth* e *late-stage*)	*Crunchbase*	65.957 *startups*	Obtenção de *funding* em diferentes estágios (seed, series A e series B)	Floresta Aleatória, Graph Neural Networks

Fonte: Elaborado pelo autor (2020)

O atributo target sucesso é operacionalizado de diversas formas. Dentre os estudos apresentados na Tabela 3, os eventos IPO e/ou aquisição de uma *startup* destacam-se como variáveis explicativas para o sucesso (XIANG et al., 2012; 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 uma proxy para o sucesso. Sharchilev (2018) procura prever se uma empresa, que já garantiu financiamento inicial (semente ou anjo), atrairá uma nova rodada de investimentos (série A ou superior) em uma janela de simulação dinâmica de um ano. Arroyo et. al (2018) estabeleceram um modelo para cada classe tipo de evento: aquisição, fechamento, IPO, rodada de financiamento e nenhum evento para uma *startup* da amostra.

Sobre 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 (XIANG et al., 2012; LIANG; DAPHNE YUAN, 2013; PAN; GAO; LUO, 2018).

Recentemente, dois estudos usaram Gradient Tree Boosting no processo de predição de sucesso de *startups* (SHARCHILEV, 2018; ARROYO et al, 2019) e um estudo utilizou Graph Neural Networks (GASTAUD; CARNIEL; DALLE, 2019), estudos com resultados e modelagens promissoras pelo uso de Deep Learning.

Pela Tabela 3, nota-se que alguns estudos trazem fontes adicionais aos dados oriundos da plataforma *Crunchbase*, como por exemplo: *LinkedIn*, *Business Insider* e *Techcrunch* (XIANG et al., 2012; SHARCHILEV, 2018). Tais estudos buscam enriquecer os dados *Crunchbase* e extrair relações entre sucesso de *startups* e menções realizadas em redes sociais e textos sobre notícias.

A Tabela 4 apresenta as técnicas usadas e os principais resultados (TPR: True Positive Rate dos dados agregados, acurácia e precisão) dos 6 estudos (Tabela 2) que usaram – exclusivamente - 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, são KNN, Gradient Tree Boosting e Graph Neural Networks.

Tabela 4 - Técnicas de *machine learning* versus artigos com 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 (0.45)	
Floresta aleatória			Acurácia (**0.931**)	Acurácia (**84.53%**)	Precisão (0.64)	Precisão (0.63)
KNN				Acurácia (73.33%)		
SVM	TPR (**89.58%**)		Acurácia (0.928)		Precisão (**0.68**)	
Gradient tree boosting					Precisão (**0.68**)	
Graph neural networks						Precisão (**0.65**)

Fonte: Elaborado pelo autor (2020)

Arroyo et. al (2019), obtiveram uma menor precisão de seus modelos comparando-se com Shan, Cao e Lin (2014).

O primeiro estudo usou janelas temporais para avaliação dos modelos. Neste caso, estabeleceram duas janelas: de aquecimento 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); 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 start-up.

Bento (2017), que apresentou alta acurácia em seus modelos, trabalhou com o maior número de atributos (158), sem uso de janelas temporais como Arroyo et al. (2018), usando dados de alguns estados americanos de forma agregada.

5. Considerações finais

Apesar da base de dados Crunchbase ser majoritariamente formada pelo auto preenchimento dos empreendedores / startups, o que pode causar deficiência na acuracidade dos seus dados, seu volume de dados se mostra promissor para pesquisas variadas sobre a temática do empreendedorismo.

Sobre predição de sucesso de startups, esta revisão mostrou oportunidades e desafios quanto à operacionalização da proxy “sucesso de startups”, design da pesquisa, seleção de atributos e escolha dos modelos de machine learning.

Quanto à operacionalização da proxy sobre o sucesso de startups, a estratégia mais comum designa como bem sucedida a startup que foi adquirida ou teve seu IPO.

Alguns estudos também levaram em consideração rodadas de financiamento intermediárias, principalmente as mais avançadas, série A, B ou C, como indicativo de sucesso. Afinal, nestes casos, a startup se mostra atrativa para captação de funding com valores mais expressivos.

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

Os modelos de classificação binária, de aprendizado supervisionado, foram os mais utilizados, com destaque para modelos clássicos como floresta aleatória, regressão logística e *support vector machine.* Em 2019, alguns estudos têm usado técnicas de deep learning na tentativa de predição de sucesso de startups.

Por fim, destaca-se que nesta revisão de literatura, majoritariamente, os estudos sobre predição de sucesso de startups com uso de machine learning são conduzidos por pesquisadores de ciência de dados e computação. Neste ponto reside oportunidade para pesquisadores do campo da Administração conduzirem estudos paralelos sobre o tema, sozinhos ou em conjunto com as áreas citadas.

Esta aproximação pode agregar conhecimento do contexto, útil para o pré processamento e seleção de atributos, importantes fatores para a qualidade de modelos de machine learning. Importante também para trazer novas interpretações e significados aos resultados oriundos dos modelos de machine learning.

REFERÊNCIAS

ARROYO, J. et al. Assessment of machine learning performance for decision support in venture capital investments. **IEEE Access**, v. 7, p. 124233–124243, 2019.

BASOLE, R. C.; PARK, H.; CHAO, R. O. Visual Analysis of Venture Similarity in Entrepreneurial Ecosystems. **IEEE Transactions on Engineering Management**, v. 66, n. 4, p. 568–582, 1 nov. 2019.

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

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

DALLE, J.-M.; MENON, C. **Using Crunchbase for economic and managerial research**. OECD Science, Technology and Industry Working Papers, n. 2017/08, 2017. Disponível em: https://pdfs.semanticscholar.org/aa83/4b1ddd1d6c96bde1c6e526be6bb2a99ad011.pdf. Acesso em: 07 jun. 2020.

ERTEL, W. **Introduction to Artificial Intelligence**. 2nd ed. London: Springer, 2017.

FACELLI, K. et al. **Inteligência Artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2019.

GASTAUD, C. CARNIEL, T.; DALLE, J.-M. 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**, 2019. Disponível em: https://arxiv.org/abs/1906.03210. Acesso em: 03 jun. 2020.

GERETO, M. A. S. **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. Disponível em; http://bibliotecadigital.fgv.br/dspace;/handle/10438/27468. Acesso em: 01 jun. 2020.

HSIEH, K.-H.; LI, E. Y. **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, 2017. Disponível em: http://iceb.johogo.com/proceedings/2017/ICEB_2017_paper_36-WIP.pdf. Acesso em: 3 jun. 2020.

KEMENY, T.; NATHAN, M.; ALMEER, B. **Using Crunchbase to explore innovative ecosystems in the US and UK**. Birmingham Business School Discussion Paper Series, 2017. Disponível em: http://epapers.bham.ac.uk/3051/1/bbs-dp-2017-01-nathan.pdf. Acesso em: 01 abr 2020.

KOSTERICH, A.; WEBER, M. S. Starting up the News: The Impact of Venture Capital on the Digital News Media Ecosystem. **International Journal on Media Management**, v. 20, n. 4, p. 239–262, 2 out. 2018.

LIANG, E.; DAPHNE YUAN, S.-T. **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, 2013. Disponível em: http://chbrown.github.io/kdd-2013-usb/workshops/MLG/doc/liang-mlg13.pdf. Acesso em: 03 jun. 2020.

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

MARTINEZ, D. C. **Startup Success Prediction in the Dutch Startup Ecosystem**. Master of Science in Management of Technology. Delft University of Technology, 2019. Disponível em: https://repository.tudelft.nl/islandora/object/uuid%3A1adc2972-db09-4583-b2da-05fd4e462941. Acesso em: 15 jun 2020.

NYLUND, P. A.; COHEN, B. Collision density: driving growth in urban entrepreneurial ecosystems. **International Entrepreneurship and Management Journal**, v. 13, n. 3, p. 757–776, 1 set. 2017.

PAN, C.; GAO, Y.; LUO, Y. **Machine Learning Prediction of Companies' Business Success**. 2018. Disponível em: http://cs229.stanford.edu/proj2018/report/88.pdf. Acesso em: 25 mar 2020.

SANTOS DA SILVA, D. **Portuguese Startups: a success prediction model**. Master's Dissertation in Finance and Tax. FEP - Faculdade de Economia da Cidade do Porto, 2016. Disponível em: https://repositorio-aberto.up.pt/bitstream/10216/86067/2/157022.pdf. Acesso em: 22 jun 2020.

SHAN, Z.; CAO, H.; LIN, Q. **Capital Crunch: Predicting Investments in Tech Companies**. 2014. CS221 Project: Crunchbase Investment Prediction. Disponível em: http://www.zifeishan.org/files/capital-crunch.pdf. Acesso em: 12 jun 2020.

SHARCHILEV, B. et al. **Web-based startup success prediction**. In: International Conference on Information and Knowledge Management, Proceedings. Anais... Association for Computing Machinery, 17 out. 2018.

SKIENA, S. S. **The Data Science Design Manual**. Suiça: Springer, 2017.

XIANG, G. et al. **A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch**. In: Sixth International AAAI Conference on Weblogs and Social Media. 2012. Disponível em: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4621/5071. Acesso em 11 jun. 2020.

ZHAO, X.; ZHANG, W.; WANG, J. **Risk-hedged venture capital investment recommendation**. In: RecSys 2015 - Proceedings of the 9th ACM Conference on Recommender Systems. Anais...Association for Computing Machinery, 16 set. 2015.