Universidade Federal do ABC

Evolução de grafos de genealogia acadêmica

Projeto de Graduação em Computação

Aluno: Arthur Veloso Kamienski
Bacharelado em Ciência da Computação
a.kamienski@aluno.ufabc.edu.br

Orientador: Jesús P. Mena-Chalco
Centro de Matemática, Computação e Cognição
jesus.mena@ufabc.edu.br

Santo André, 16 de Dezembro de 2019

Resumo

A genealogia acadêmica é a área que busca entender o fluxo de conhecimento e o desenvolvimento de uma comunidade acadêmica através das relações de orientação que ocorrem entre professores e seus alunos. Estas relações podem ser estruturadas como grafos que modelam pesquisadores como vértices e suas orientações como arestas direcionadas, demonstrando o sentido de passagem de conhecimento. Estes grafos, apesar de muitas vezes serem representados em sua forma estática para um determinado instante do tempo estão em constante desenvolvimento e evolução através do surgimento de novos pesquisadores e a criação de novas relações de orientação. Este caráter temporal é, no entanto, pouco explorado por outros estudos, mesmo demonstrando um grande potencial de trazer um melhor entendimento do desenvolvimento da ciência e do crescimento de uma comunidade de pesquisadores. Este trabalho utiliza as informações históricas do modo em que estes grafos evoluíram para obter informações sobre suas principais características e sobre o rumo que eles podem tomar no futuro. Utilizando técnicas de manipulação de dados, assim como métodos de aprendizado de máquina, foi definido um processo para a extração de informações temporais de grafos de genealogia acadêmica e o posterior treinamento de um modelo a partir de aprendizado supervisionado utilizando estes dados. Os resultados obtidos pelo método desenvolvido indicaram padrões de crescimento no grafo como a redução da produção de novos pesquisadores ao longo dos anos e as principais características que levam um pesquisador a realizar orientações. Ao fim deste projeto, obteve-se uma forma bem definida para a realização das tarefas indicadas, adicionando uma nova ferramenta para a exploração de grafos evolutivos e de áreas relacionadas.

Palavras-chave: grafos, genealogia acadêmica, evolução, crescimento, predição.

Sumário

1 Introdução

A necessidade de registrar a história e estudar acontecimentos passados é algo natural para a humanidade (Butler, 2011). Desde os tempos antigos, diversos estudiosos tentam desvendar as informações que foram deixadas por seus ascendentes, de modo a entender fenômenos atuais (Basu & Waymire, 2006). Além disso, compreender as suas origens a partir dos registros históricos passa a ser um ponto importante de estudo (Marrou & Marrou, 1982), possibilitando o surgimento de novas áreas e linhas de pesquisa relacionadas.

Com o aumento da volume disponível de dados, iniciativas voltadas ao estudo desses registros tem sido alteradas para aliar julgamentos pessoais com medidas obtidas através da análise de dados (Hicks *et al.* , 2015). É neste contexto que surgem a cientometria 2.0 (Priem & Hemminger, 2010) e a webometria (Thelwall, 2008), abordagens derivadas das tradicionais áreas de infometria que tem como principal objetivo mensurar o conhecimento (Sengupta, 1992).

Diversos estudos já se utilizam dos novos métodos de obtenção de dados para realizar análises sobre o desenvolvimento da ciência (Leydesdorff & Milojević, 2012). Duas abordagens comuns neste âmbito são a análise de citações (Hummon & Dereian, 1989; Moed, 2006) e de co-autoria (Henriksen, 2016; Newman, 2004) em trabalhos acadêmicos. Existem ainda trabalhos que analisam a genealogia acadêmica como meio de avaliar o fluxo de conhecimento cientifico e caracterização dos membros da comunidade acadêmica (Mena-Chalco & Junior, 2013).

Para este tipo de estudo, grafos são objetos comumente utilizadas para a realização de análises (David & Hayden, 2012). Através da utilização de métricas tradicionais da Teoria dos Grafos, as relações capturadas pelos dados podem ser estruturadas de modo a facilitar o entendimento e a captura de fenômenos (Rossi *et al.* , 2018). Contudo, estes grafos são normalmente considerados em seu formato estático, apesar de capturarem também informações temporais que os caracterizam como grafos evolutivos.

Neste trabalho, procura-se explorar a capacidade evolutiva de grafos de genealogia acadêmica de modo a extrair um novo conhecimento a respeito da maneira como é dado o crescimento da comunidade científica. Através de técnicas de manipulação de grafos, aliada com métodos

de Aprendizado de Máquina para identificar padrões históricos nos dados, busca-se entender não só o passado, mas também o futuro da ciência.

Com isso, foram definidas três perguntas focais para guiar o desenvolvimento do trabalho:

1. Quais as características que ditam o surgimento de novos pesquisadores?
2. Existe diferença entre o crescimento passado e o atual?
3. Qual será o cenário da comunidade científica em médio prazo?

Estas perguntas, apesar de abordarem aspectos que se sobrepõe, correspondem a três ideias distintas que serão discutidas ao longo do trabalho. Primeiro, procura-se entender quais são os principais fatores responsáveis por expandir a comunidade científica à nível de um único pesquisador. Segundo, deseja-se provar que, através da análise do passado, é possível obter uma previsão confiável do futuro, indicando também que a ciência se desenvolve de maneira constante e homogênea. Por fim, a última pergunta foca em desenhar uma noção de como o estado atual da ciência se desenvolverá para o futuro, tentando entender não só quais pesquisadores adquirirão grande importância, mas também quais serão as áreas de maior impacto e maior produção em termos de recursos humanos.

1.1 Objetivo geral

O objetivo geral deste Projeto de Graduação em Computação é o desenvolvimento de um método computacional de predição do crescimento de grafos de genealogia acadêmica, de forma a gerar conhecimento sobre a maneira como essa evolução é dada e estimar um possível cenário futuro de uma comunidade de pesquisadores.

1.2 Objetivos específicos

- Identificar características de evolução para comunidades de pesquisadores;
- Explorar diferentes técnicas de análise de evolução de grafos, buscando identificar aquela mais adequado à tarefa proposta;

- Verificar relações entre diferentes áreas do conhecimento e a forma como evoluem ao longo do tempo; e
- Estimar o cenário científico brasileiro para diferentes momentos no futuro.

1.3 Organização do trabalho

Este trabalho está dividido em sete seções, incluindo esta seção introdutória. Estas seções buscam explicar os resultados obtidos pela aplicação das técnicas desenvolvidas, explorando também o caminho percorrido para a obtenção de tais resultados.

A Seção 2 descreve conceitos relevantes para o entendimento da metodologia aplicada e dos resultados obtidos neste trabalho. São discutidos nesta seção os conceitos de grafos de genealogia e grafos evolutivos, assim como técnicas de aprendizado de máquina.

A Seção 3 explora o conjunto de dados utilizado para a obtenção dos resultados. Além de indicar a origem dos dados, a seção também explora características gerais do conjunto e do grafo de genealogia descrito por ele. As informações contidas na seção são importantes não só para o embasamento dos resultados, mas também para a compreensão do contexto em que estão inseridos.

A metodologia criada para a obtenção dos resultados está descrita na Seção 4. Esta seção explica os passos seguidos para a obtenção de dados temporais do grafo, posteriormente utilizados para a análise da evolução do grafo, e para o treinamento de modelos de predição que podem ser utilizados para analisar uma evolução futura.

Os resultados são explorados na Seção 5. A discussão destes resultados é focada tanto na análise do crescimento do grafo em períodos passados, assim como a análise das previsões obtidas pelos modelos desenvolvidos.

Todos os passos realizados durante o transcorrer do projeto estão indicados na Seção 6. Esta seção apresenta um diagrama de Gantt indicando quais foram as tarefas realizadas e quando cada uma foi desenvolvida.

Finalmente, a Seção 7 descreve as conclusões e as principais contribuições deste trabalho, apresentando também as limitações e possíveis abordagens para pesquisas futuras baseadas.

2 Fundamentação teórica

O desenvolvimento deste trabalho se baseia no estudo de grafos de genealogia acadêmica e a possível extração de conhecimento através da análise de seu caráter evolutivo com o uso de técnicas de Aprendizado de Máquina. A fim de clarificar estes conceitos, assim como trazer definições úteis para o entendimento futuro do trabalho, esta seção trás um maior detalhamento sobre os grafos e técnicas utilizadas.

2.1 Grafos de genealogia acadêmica

Interações entre pesquisadores no meio acadêmico podem ocorrer de inúmeras maneiras distintas (Kogan, 2000). Ao mesmo tempo em que se relacionam em seus ambientes de trabalho e nos diversos eventos dos quais participam, relações são criadas também através da colaboração em pesquisa (Vinkler, 1993) e da formação de novos pesquisadores por meio de orientações acadêmicas (Damaceno *et al.* , 2017).

Pesquisadores que trabalharam em conjunto para a escrita de um artigo científico, por exemplo, podem ser caracterizados como co-autores deste trabalho, e identifica-se esta interação como uma relação de co-autoria (Mena-Chalco *et al.* , 2012). De forma similar, ao basear-se em um artigo de outro pesquisador e utilizá-lo para a escrita de um novo artigo relacionado, o autor gera uma relação de citação entre ambos os trabalhos, que pode ser também expandida metonimicamente para seus respectivos criadores (Grácio & Oliveira, 2014).

Por outro lado, uma parcela da designação de pesquisadores (em especial aqueles que se dedicam também a lecionar) é a transmissão de seu conhecimento e a formação de novas gerações que possam dar continuidade à expansão de sua área. Este processo de passagem de conhecimento científico através da herança intelectual e geração de novos pesquisadores é caracterizado pelas relações de orientação e co-orientação acadêmica entre professores e alunos, e ao seu estudo é dado o nome de genealogia acadêmica (Sugimoto, 2014).

Ao se tratar de relações entre um grande número de entidades, a representação através de grafos aparece como algo natural para facilitar o seu entendimento (Gargiulo *et al.* , 2016). Para os casos expostos, de interações entre pesquisadores, pode-se definir cada pesquisador

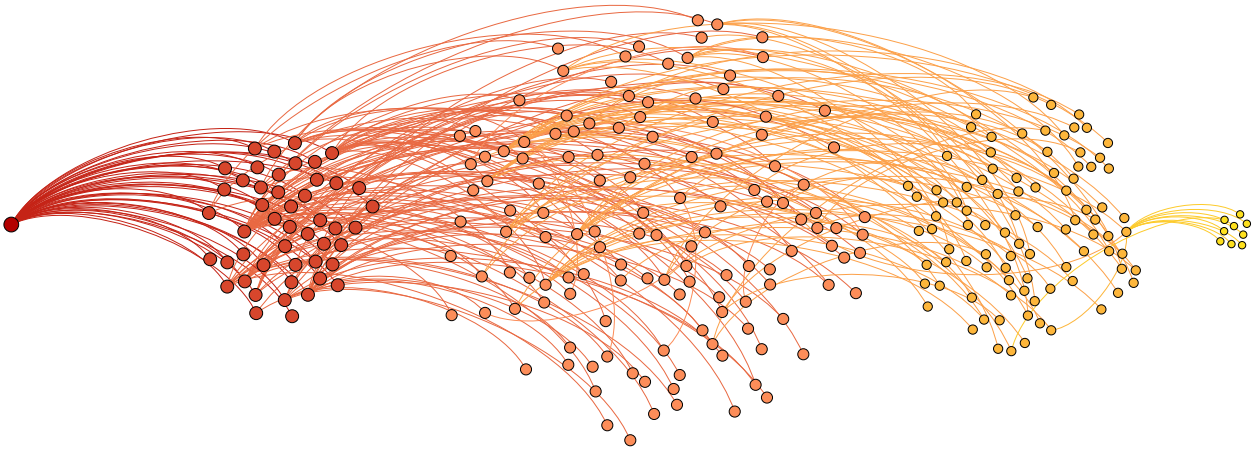

Figura 1: *Grafo de genealogia acadêmica apresentando todos os descendentes do pesquisador Jacob Palis, representado pelo vértice vermelho mais à esquerda. Cada geração a partir de Palis está representada por uma cor mais clara e tamanho menor dos vértices. A cor de cada aresta é a mesma do vértice de origem, representando sua unidirecionalidade.*

como um vértice e cada relação (seja ela qualquer uma das apresentadas) como uma aresta pertencentes ao grafo que os representa. Grafos que modelam as relações de co-autoria são chamados de "grafos de co-autoria" (Yan & Ding, 2009), enquanto àqueles que modelam relações de citação é dado o nome "grafos de co-citação" (Batagelj, 2003).

Finalmente, os grafos que representam relações de orientação acadêmicas são designados por "grafos de genealogia acadêmica", nome proveniente da noção tradicional de genealogia, a qual indica tanto a origem quanto o desenvolvimento de um ramo de uma atividade humana (Sugimoto *et al.* , 2011). Esses grafos são comumente direcionados, dado que as relações são em sua maioria unidirecionais, de modo que um autor é citado por outro ou que um professor orienta um aluno. Vale notar também que, como um pesquisador pode receber mais de uma orientação de um mesmo professor, estes grafos podem também ser caracterizados como multigrafos. A Figura 1 apresenta um exemplo de grafo de genealogia contendo todos os descendentes do pesquisador Jacob Palis. As informações utilizadas para compor o grafo foram obtidas a partir da plataforma Lattes[1], utilizando o método desenvolvido por Damaceno *et al.* (2017).

Apesar de suas semelhanças com árvores genealógicas tradicionais, as quais expõe relações de parentesco entre indivíduos através das ramificações de suas famílias, vale apontar que gra-

[1] http://lattes.cnpq.br, último acesso em 24 de agosto de 2019.

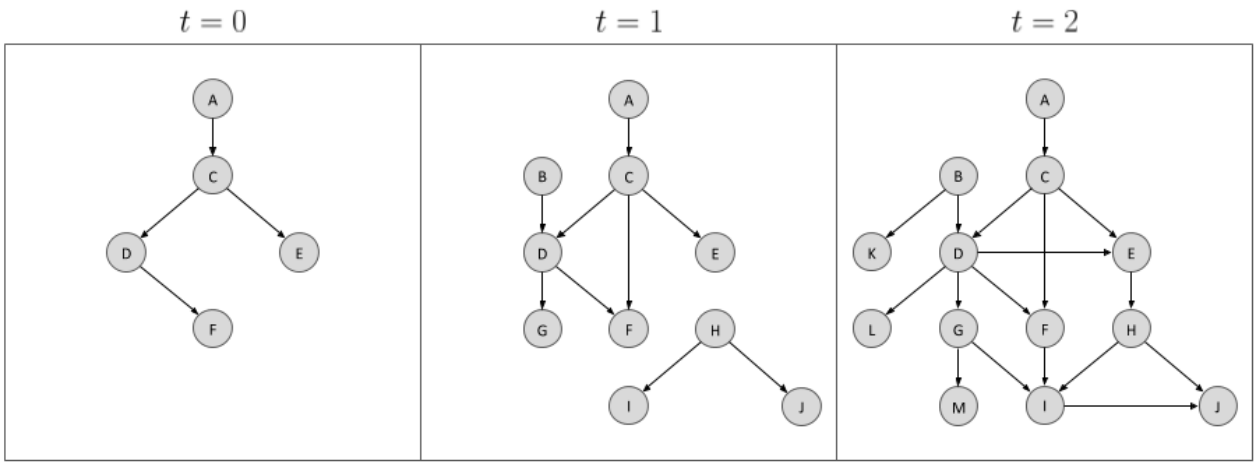

Figura 2: *Exemplo de grafo evolutivo mostrando a sua evolução com o passar do tempo, indicado pela variável t acima de cada grafo da figura.*

fos de genealogia acadêmica não podem ser formalmente/computacionalmente caracterizados como árvores (grafos conexos acíclicos Bondy *et al.* (1976)). Ainda que em algumas regiões a estrutura um grafo deste tipo seja de fato uma árvore, existem exceções que impedem que o seu todo seja classificado como uma. Um exemplo disso são ciclos que ocorrem quando um pesquisador é orientado em seu doutorado por outro, enquanto ambos foram orientados em seus mestrados por um mesmo orientador em um momento anterior. Além disso, os grafos costumam não ser totalmente conexos, havendo uma grande componente que conecta a maior parte dos pesquisadores e diversas outras componentes isoladas com uma quantidade pequena de vértices, se assemelhando, portanto, à uma floresta.

Ainda assim, a nomenclatura utilizada em árvores genealógicas familiares é útil para descrever as relações entre dois pesquisadores, dado que facilitam o entendimento pela similaridade com exemplos do cotidiano. Deste modo, designam-se aqui os orientadores como "pais acadêmicos" de seus alunos os quais, de forma oposta, designam-se "filhos acadêmicos". Essa convenção também é expandida para outras relações, tais como irmãos, primos, avôs, bisavôs, entre outros.

2.1.1 Grafos evolutivos

Os grafos descritos anteriormente, apesar de serem representados em sua forma estática para um dado instante no tempo, levam consigo também a informação temporal do momento e da duração na qual uma relação ocorreu, assim como o momento de surgimento

de cada vértice. Deste modo, um grafo pode ser observado para diferentes instantes, tendo sua forma alterada pela criação de vértices e arestas com o passar do tempo a partir. Esses grafos podem, portanto, ser caracterizados como grafos que evoluem conforme o tempo (Aggarwal & Subbian, 2014). Um exemplo deste tipo de grafo pode ser observado na Figura 2.

De fato, orientações acadêmicas ocorrem em períodos distintos para pesquisadores distintos, assim como pesquisadores surgem conforme recebem as suas primeiras orientações. Analisando um período passado, muitas das orientações e dos pesquisadores presentes em um grafo mais recente não existirão, e um crescimento poderá ser notado conforme avançam-se os anos.

A representação estática de grafos evolutivos, apesar de útil para o estudo de um dado período de tempo e de seu passado, não leva em conta seu estado de constante evolução. Esta informação pode ser explorada, entretanto, ao se analisar o estado do grafo em diferentes momentos do passado, entendendo a maneira como se deu seu desenvolvimento até o presente (Kostakos, 2009).

Grafos evolutivos têm sido utilizados em diversas aplicações, como para a detecção de anomalias na *web* (Papadimitriou *et al.* , 2010), análise e evolução de redes sociais (Doreian & Stokman, 2013) e para a modelagem de interações entre proteínas (Vázquez *et al.* , 2003). Existem ainda iniciativas para a extração de regras de evolução de grafos, como o trabalho de Berlingerio *et al.* (2009), que busca a obtenção de regras simples baseadas em mudanças locais ocorridas ao longo do tempo.

Ainda que diversos estudos tenham sido feitos para a exploração e o entendimento dos grafos de genealogia, são poucos os esforços que se propõe a analisá-los como grafos evolutivos e explorar o potencial de sua informação temporal (Boaventura *et al.* , 2014).

2.2 Técnicas de aprendizado de máquina

A área de aprendizado de máquina é um ramo da ciência da computação que incorpora diversos campos e sub-áreas distintas, as quais todas se utilizam de um mesmo princípio comum: o uso de algoritmos computacionais que aprendem a partir da experiência previa-

mente obtida (El Naqa & Murphy, 2015). Isto é, dado um algoritmo computacional (também chamado usualmente de "modelo"), sua performance em uma determinada tarefa melhora conforme ele adquire experiência obtida pela análise de um ambiente.

Devido a esta definição ampla, é natural que existam diferentes técnicas para as mais variadas aplicações. Atualmente, algoritmos de aprendizado de máquina são utilizadas para encontrar informações em grandes bases de dados (Witten *et al.* , 2016), reconhecer e entender linguagens naturais (Collobert & Weston, 2008), diagnosticar doenças em exames de pacientes (Kononenko, 2001), dirigir carros de forma autônoma (Chen *et al.* , 2015), dentre diversas outras aplicações.

Para desenvolver e aplicar algoritmos de aprendizado de máquina geralmente são necessários três elementos: uma tarefa bem definida a ser realizada, um meio de determinar o desempenho do algoritmo e uma fonte de experiência, também chamado de conjunto de treinamento (Mitchell *et al.* , 1997). Para a tarefa de identificar palavras faladas, por exemplo, tem-se como medida de desempenho a taxa de acertos do algoritmo e um conjunto de palavras faladas como uma possível fonte de experiência.

A partir desses elementos o algoritmo pode aprender com a análise (utilizando a medida de desempenho definida) dos resultados obtidos pela sua aplicação no conjunto de treinamento. Ao final do processo de aprendizagem o algoritmo deve estar pronto para ser generalizado para um conjunto qualquer, estando pronto para atuar na aplicação desejada.

Os resultados obtidos pela aplicação de um algoritmo após seu treinamento em um conjunto de dados novo, distinto daquele utilizado para o treinamento, podem ser entendidos como previsões sobre os resultados reais que não estão disponíveis ou que ainda não foram observados. Em outras palavras, baseando-se em informações sobre um estado atual, como também sobre estados passados, um algoritmo pode aprender padrões que já foram observados em dados históricos para gerar uma previsão sobre um estado seguinte (Franklin, 2005). Essas previsões podem ser utilizadas, então, para a análise de informações e cenários desconhecidos (pelo seu não acontecimento, a impossibilidade de seu acontecimento, ou o dessaber de seu acontecimento), permitindo que sejam estudados e utilizados para a geração de novos conhecimentos ou para a tomada de decisões.

Existem diversas técnicas e algoritmos de aprendizado de máquina distintos que aplicam estes conceitos para realizar as tarefas descritas anteriormente. Tradicionalmente, técnicas de aprendizado de máquina são divididas em três tipos principais, sendo esses o aprendizado supervisionado, o não-supervisionado e o por reforço (Bishop *et al.* , 1995).

As técnicas de aprendizado supervisionado se baseiam em proporcionar um *feedback* direto ao algoritmo, oferecendo os resultados esperados para as entradas presentes no conjunto de treinamento, de modo a incentivar o algoritmo a fazer ajustes para se adequar. Esse tipo de aprendizado é ideal para aplicações onde se deseja categorizar elementos, dividindo-os em grupos distintos (classificação) ou variáveis contínuas (regressão). São exemplos de algoritmos que se enquadram nessa categoria: (i) *Naive Bayes* (Rish *et al.* , 2001), (ii) máquinas de vetores de suporte (Suykens & Vandewalle, 1999), (iii) regressão logística (Hosmer Jr *et al.* , 2013), entre outros.

O aprendizado não-supervisionado, por outro lado, traz a abordagem contrária, não disponibilizando os resultados para o algoritmo e forçando-o a fazer inferências sobre os dados obtidos. Aplicações dessa forma de aprendizado são, por exemplo, descobrir grupos semelhantes (*clusters*), com algoritmos como o *k-means* (Alsabti *et al.* , 1997) e detectar anomalias em conjuntos de dados, com algoritmos como o *Local Outlier Factor* (Breunig *et al.* , 2000).

Por fim, técnicas de aprendizado por reforço se baseiam em encontrar ações adequadas para uma situação, com o objetivo de maximizar uma recompensa. Para esse tipo de aprendizado não são oferecidos resultados esperados, uma vez que espera-se que o algoritmo os encontre por um processo de tentativa e erro. No entanto, deve-se definir recompensas para as possíveis situações para permitir uma avaliação precisa pelo algoritmo. Tais técnicas são amplamente utilizadas no desenvolvimento de algoritmos para jogos, nos quais existem diversas ações e estados possíveis, sendo o *Q-learning* (Watkins & Dayan, 1992) um exemplo de algoritmo capaz de realizar tal tarefa.

Este trabalho busca aplicar os conceitos de aprendizado de máquina à grafos evolutivos. Com base da teoria discutida anteriormente, pode-se construir um algoritmo que tenta aprender o modo como um grafo evolui, ao analisar seu desenvolvimento em momentos passados. Para um dado vértice em um grafo, por exemplo, o algoritmo pode tentar prever sua probabilidade de gerar uma nova aresta com o outro vértice (o que é chamado de predição de links),

dado que as suas características se assemelham à outras já observadas em algum momento do passado (Lü & Zhou, 2011).

A fonte de experiência para o treinamento do algoritmo se torna o próprio grafo, aliado com informações de suas características topológicas, assim como as informações singulares de cada vértice. É importante ressaltar, no entanto, que informações únicas de um período, como o ano em que amostra é coletada, não são úteis para o aprendizado, uma vez que dados futuros não apresentarão essa informação.

O desempenho de tal algoritmo pode ser mensurado ao se comparar a sua previsão (aquilo que é esperado que aconteça) com o acontecimento real observado em um estado futuro. Neste caso, a medida seria obtida através da comparação do número de vértices previsto com o número real de vértices gerados no intervalo de tempo seguinte. Esta análise pode ser feita de forma local, comparando a quantidade de filhos para cada vértice isoladamente, ou de forma global, ao se olhar para o número total de vértices de uma geração. Note que, como o objetivo principal deste trabalho é o entendimento do crescimento do grafo como um todo, o desempenho do algoritmo em uma análise global é preferível, apesar dos resultados para vértices únicos serem relevantes para garantir que o algoritmo faz previsões confiáveis.

Para este trabalho, busca-se criar um algoritmo que seja capaz de prever, para um dado vértice, a sua propensão à ter um ou mais filhos em um intervalo de tempo futuro. Ao realizar predições para todos os vértices presentes em um grafo em um dado instante, pode-se prever seu crescimento aproximado (dado que previsões erradas podem acontecer) através do surgimento de uma geração futura.

Além disso, busca-se um entendimento sobre as características que ditam a evolução do grafo e sobre os atributos que influenciam na realização da previsão. Desta forma, o algoritmo a ser utilizado deve, na medida do possível, apresentar de forma clara o caminho e os critérios para a obtenção de resultados. Um exemplo de algoritmo que satisfaz estas condições e que será explorado ao longo do trabalho é a árvore de decisão (Safavian & Landgrebe, 1991). A seguir serão discutidos os principais conceitos para a aplicação deste algoritmo, assim como uma visão global sobre seu funcionamento e principais características.

3 Conjunto de dados utilizado

Existem diversas fontes de dados genealógicos publicamente disponíveis que podem ser utilizadas para a geração de um grafo evolutivo de genealogia e para o treinamento de um modelo preditivo. Um número crescente desses dados têm surgido a partir do aumento da disponibilidade de informações e à medida em que novos dados são coletados e adicionados à plataformas já existentes.

Alguns exemplos notáveis de iniciativas voltadas especificamente para o registro da genealogia acadêmica são o *Mathematics Genealogy Project* (Gargiulo *et al.* , 2016) e o *Astronomy Genealogy Project* (Tenn, 2016), ambas plataformas nos quais matemáticos e astrônomos, respectivamente, podem se cadastrar e adicionar suas informações aos repositórios.

É possível, ainda, extrair as informações genealógicas a partir de dados não vocacionados para a genealogia. Dados de co-autoria, nos quais professores e alunos fazem publicações em conjunto, assim como currículos acadêmicos que expõe informações sobre a formação de pesquisadores são exemplos deste tipo de fonte de dados (Damaceno *et al.* , 2017).

Neste trabalho foram utilizados os dados disponibilizados pela Plataforma Acácia[2], que tem como o objetivo documentar relações de orientações acadêmicas utilizando as informações de 6300000 currículos acadêmicos da Plataforma Lattes[3] Damaceno *et al.* (2019). Os dados abrangem um período de 1900 à 2019 e incluem 1272590 pesquisadores e 1404109 relações de orientações acadêmicas, já estruturados na representação de um grafo. Das orientações, 995807 são orientações de mestrado, 371074 de doutorado, e 37228 de pós-doutorado. A informação temporal indicando o ano de início e de término de cada orientação já está presente no conjunto de dados, facilitando o processo de análise de sua evolução ao longo do tempo.

Apesar do grande volume de dados presentes no conjunto, é preciso ter em mente que todos os registros foram preenchidos de forma manual, a partir do cadastro dos próprios pesquisadores na Plataforma Lattes. Por consequência, a integridade e autenticidade de todos os dados não pode ser garantida, e inconsistências nos registros podem existir. Os principais problemas encontrados são a ausência de dados, uma vez que muitos pesquisadores ainda

[2] http://plataforma-acacia.org, último acesso em 27 de novembro de 2019.
[3] http://lattes.cnpq.br, último acesso em 27 de novembro de 2019.

não estão cadastrados, e erros de preenchimento, que geram ruído nos dados a partir de duplicações e incoerências.

3.1 Grafo de genealogia acadêmica da Plataforma Acácia

O grafo obtido através da Plataforma Acácia contém um vértice para cada pesquisador e uma aresta para cada relação de orientação, totalizando 1272590 vértices e 1404109 arestas. Estes números mostram uma leve tendência de crescimento do grafo, dado que existe $1,1$ orientação por pesquisador implicando que, em média, um pesquisador gera mais de um aluno.

Dos vértices, 130266 (10.2%) são raízes, não tendo nenhuma aresta de entrada e, por consequência, nenhum orientador cadastrado. É importante ressaltar que raízes estão relacionadas a erros de cadastro ou falta de registros, dado que um pesquisador apenas surge a partir de uma orientação. Pesquisadores sem nenhum pai são, portanto, pesquisadores não cadastrados na plataforma (que apenas são referenciados por outros), ou que são cadastrados mas não registraram nenhuma orientação recebida.

Por outro lado, a maioria dos pesquisadores têm uma ou duas orientações, correspondendo a 914932 $(65,2\%)$ e 197208 $(14,0\%)$ casos, respectivamente. Outros valores maiores que duas orientações aglomeram o restante dos pesquisadores (30184 casos ou $2,1\%$). Note que existem casos extremos, nos quais pesquisadores acumulam mais de dez orientações recebidas. Estes casos podem estar relacionados aos erros de registro já discutidos. A distribuição de valores de orientações recebidas estão expostos na Figura 3.

Similarmente, a distribuição do número orientações feitas por pesquisador (seu grau de saída) pode também ser observada pela Figura 3. Note que existe um número elevado de pesquisadores sem nenhum filho. Estes pesquisadores, chamados de folhas, totalizam 1042487 pesquisadores $(81,9\%)$, enquanto os que tem apenas um filho são 130048 $(10,2\%)$. Isto demonstra a infrequência com que pesquisadores geram outros pesquisadores, e indica que a maior parte do crescimento do grafo é dado pelos poucos pesquisadores que têm mais de um filho.

A Figura 4 mostra como o grau de saída e o grau de entrada dos vértices do grafo

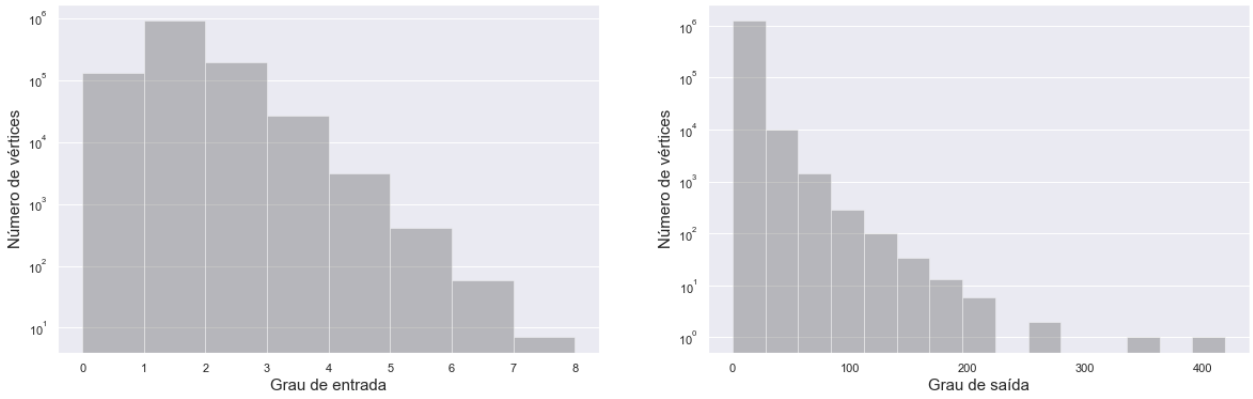

Figura 3: *Distribuição dos graus de entrada (à esquerda) e de saída (à direita) para os vértices do grafo de genealogia acadêmica da Plataforma Acácia, em escala logarítmica.*

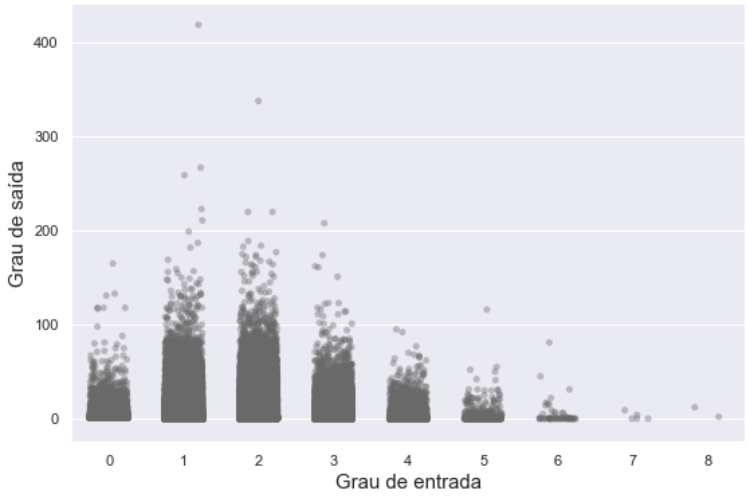

Figura 4: *Gráfico de dispersão entre o grau de entrada e o grau de saída para os vértices do grafo da Plataforma Acácia.*

estão relacionados através de um gráfico de dispersão. Vê-se que os pesquisadores com maior número de arestas de saída têm, normalmente, uma ou duas arestas incidentes, enquanto vértices com um grau de entrada maior costumam efetuar menos orientações. O número de pesquisadores com poucas orientações é grande para todos os valores de grau de entrada. Isso mostra que o número de orientações efetuadas por um pesquisador não depende apenas do seu número de orientações recebidas, indicando que outras medidas são necessárias para uma estimativa mais acurada.

O grafo é composto por uma grande componente conexa que contém 1135347 vértices, ou $89,2\%$. Existem ainda 56263 outras componentes que contém o restante dos vértices. Estas componentes são pequenas se comparadas à componentes principal, tendo a maior delas um total de 86 vértices. Ademais, 46868 componentes contém apenas dois vértices, podendo ser um indício de casos nos quais pesquisadores registraram a si próprios e ao menos um pai ou

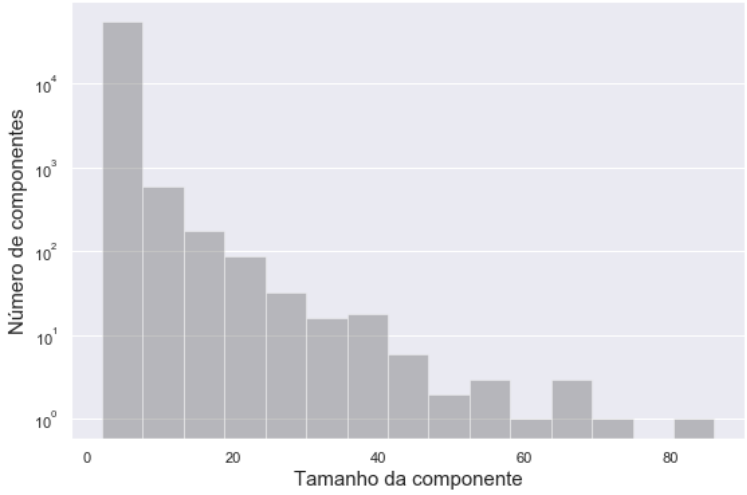

Figura 5: *Distribuição do tamanho das componentes conexas (excluindo a maior delas) do grafo da Plataforma Acácia em escala logarítmica.*

um filho. Nota-se também que cada uma dessas pequenas componentes tem uma ou mais raízes contendo, portanto, ao menos $36,0\%$ de todas as raízes do grafo. A distribuição do tamanho das componentes, excluindo a principal, pode ser observada na Figura 5.

Também pode-se analisar o ano de conclusão das orientações, atributo presente nas arestas do grafo. Note que o ano de conclusão de uma aresta representa também o ano de titulação do pesquisador que recebeu a orientação, mas não indica necessariamente o ano em que um vértice surgiu no grafo (dado que esta pode ter sido a segunda orientação recebida). Ainda assim, pode-se obter uma noção de quando os vértices do grafo surgiram e onde estão localizados no tempo com esta análise.

A Figura 6 apresenta a distribuição dos anos de conclusão das orientações. Note que o gráfico está em escala logarítmica para facilitar a visualização dos dados. Vê-se que o surgimento de novas orientações é exponencial, apresentando duas taxas de crescimento distintas (uma de 1940 à 1970, mais agressiva, e uma mais suave de 1980 à 2020). Este crescimento está de acordo com o esperado, dado que conforme mais pesquisadores são gerados, mais orientações são realizadas. Isto também pode ser uma indicação de que pesquisadores têm se tornado mais férteis conforme o passar do tempo. Finalmente, note que existe uma quantidade pequena e isolada de orientações realizadas no ano de 1900, o que pode indicar que esses são casos onde erros de preenchimento das informações ocorreram.

Por fim, vale analisar a posição relativa destes vértices no grafo a fim de obter uma noção

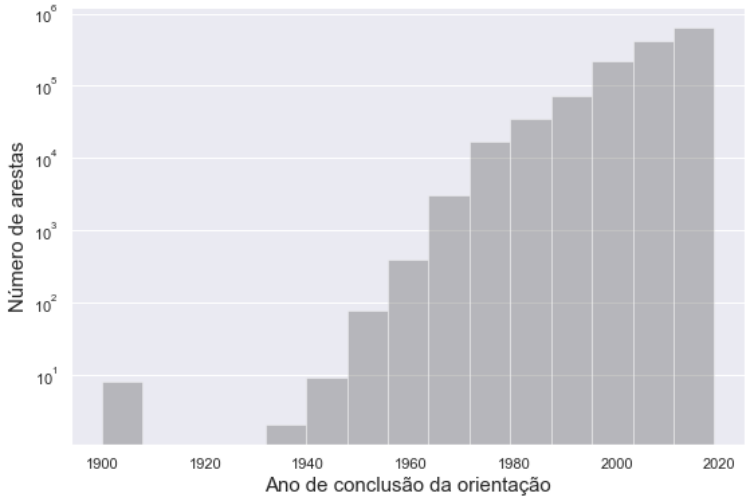

Figura 6: *Distribuição dos anos de conclusão das orientações do grafo da Plataforma Acácia em escala logarítmica.*

do potencial de seu potencial de geração de novos pesquisadores, além de explorar o formato e o modo como o grafo evolui como um todo. Para tanto, foram calculadas a maior distância até uma folha e a maior distância até uma raiz de cada vértice do grafo. Para os casos nos quais o vértice é uma raiz a distância à raiz foi definida como zero. De forma análoga, a distância de uma folha à folha mais distante também foi definida como zero. Essas medidas são complementares mas não idênticas, pois o número de descendentes de um vértice não é necessariamente ditado pelo seu número de ascendentes (como já discutido anteriormente). É importante ressaltar que a maioria dos pesquisadores pertence à mesma componente conexa, o que torna a análise a seguir viável.

Primeiro, vale notar que a maior distância entre dois vértices do grafo é dez, um indício da história recente que este grafo representa. Ao analisar a distribuição de distâncias até as raízes dos pesquisadores (Figura 7), nota-se que existe uma quantidade elevada daqueles com valores de dois à cinco, atingindo um pico em três. Além disso, a média obtida para esta medida é de $2,97$. Isso evidencia que a maior parte dos pesquisadores têm, no máximo, um tetravô ($84,2\%$ de pesquisadores com distância até quatro).

Em contrapartida, o gráfico que representa a distância dos pesquisadores às folhas exibe caráter decrescente ao longo de todo o eixo das abscissas. Se desconsiderarmos os valores nulos (correspondentes às folhas, as quais têm um grande número de ocorrências), vê-se que a distância à uma folha da maioria dos pesquisadores é de até dois ($86,5\%$), correspondente a pesquisadores que são, no máximo, avós. Desta forma, percebe-se que a maior quantidade de

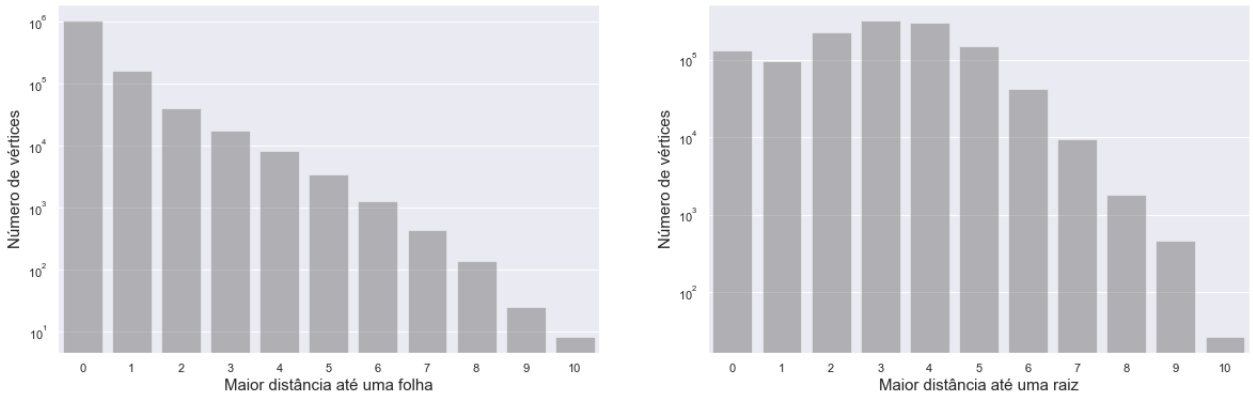

Figura 7: *Distribuição de maiores distâncias até uma folha (à direita) e a uma raiz (à esquerda) para os vértices do grafo da Plataforma Acácia, em escala logarítmica.*

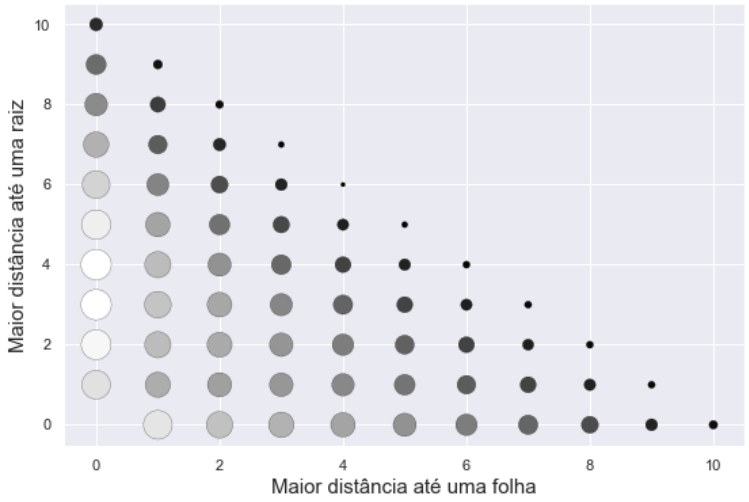

Figura 8: *Gráfico de dispersão relacionando a superficialidade com a profundidade dos vértices do grafo da Plataforma Acácia. Pontos de maior tamanho e cor mais clara indicam uma maior concentração de casos naquela coordenada.*

pesquisadores está localizada nos níveis mais próximos às folhas dos grafos, enquanto estão mais distantes de suas raízes, indicando que ainda podem estar em atividade e podem ter grande potencial de geração de novos pesquisadores.

Finalmente, pode-se ver pelo gráfico de dispersão da Figura 8 que existe uma grande concentração de pesquisadores com baixa distância a uma folha mas média distância a uma raiz. Além disso, as distâncias até as raízes decrescem conforme aumenta-se a distância às folhas e vice-versa. Esse fenômeno é mais um indício de que muitos pesquisadores se encontram mais próximos às folhas que às raízes, corroborando a hipótese anterior. Vale ressaltar, no entanto, que casos em que ambas as distâncias apresentam valores baixos (zero, um e dois) sofrem grande influência das componentes conexas com poucos pesquisadores.

4 Procedimento metodológico

Esta seção visa demonstrar todos os passos realizados para a obtenção dos resultados expostos na Seção 5. Os procedimentos metodológicos aqui discutidos foram divididos em duas etapas distintas: pré-processamento dos dados e treinamento de modelos preditivos. Cada uma destas etapas pode ser novamente dividida em tarefas que facilitam o entendimento e a organização do trabalho.

Todo o procedimento tem como principais resultados um conjunto de dados que expõe de forma clara e facilmente explorável o crescimento e a evolução do grafo ao longo do tempo (permitindo o estudo do passado), e um algoritmo treinado utilizando técnicas de aprendizado de máquina com base neste conjunto que tenta prever possíveis cenários do grafo no futuro. Estes resultados são obtidos a partir dos dados provenientes de um conjunto de dados pré-existente, que é tido como a entrada de todo este procedimento

A seguir cada uma das etapas e suas tarefas será discutida de forma detalhada em suas respectivas seções.

4.1 Pré-processamento dos dados

A primeira etapa do procedimento, chamada de pré-processamento dos dados, envolve a manipulação e limpeza dos dados oriundos de um conjunto de dados, visando facilitar o seu uso e aumentar a sua qualidade para o estudo específico que deseja-se realizar posteriormente (ao converte-lo a um formato mais apropriado e remover inconsistências, por exemplo). Esta fase é composta de três tarefas realizadas sequencialmente que recebem os dados em sua forma bruta (sem alterações à forma original) e transformam-nos em novos dados prontos para serem utilizados na próxima etapa.

A tarefa de “manipulação do grafo” inicia esta etapa recebendo um conjunto de dados e, após representá-lo no formato de um grafo, realiza alterações tomando proveito de sua estrutura. Note que esta tarefa não tange a conversão de um conjunto de dados para a estrutura de um grafo, dado que assume que esse já contém as informações necessárias para que seja representado no formato desejado. Esta tarefa deve ser capaz de, portanto, ler e

interpretar as informações recebidas e, então, realizar as alterações desejadas. Sendo assim, os principais trabalhos realizados por esta tarefa são a remoção de vértices e arestas indesejados, a alteração de atributos de vértices e arestas e a resolução de inconsistências na estrutura do grafo.

Já a segunda tarefa, "extração da informação temporal do grafo", utiliza o grafo advindo da tarefa anterior e busca converter a informação temporal que ele carrega para um formato no qual ela pode ser melhor utilizada. Para tanto, o grafo é analisado em diferentes momentos ao longo do tempo e atributos de seus vértices são extraídos em cada períodos observado. Além disso, esta tarefa também realiza o cálculo de medidas variadas para cada vértice. Desta forma, ao agrupar os dados obtidos por este processo ao longo de todo o intervalo de tempo examinado, a forma como cada um desses atributos e métricas cresce conforme o avanço do tempo pode ser explorada. Ao final de todos os processos um novo conjunto de dados é gerado contendo todos os dados coletados em todos os períodos observados.

Por fim, a terceira tarefa visa realizar uma última manipulação dos dados obtidos pela tarefa anterior. O principal objetivo neste passo é gerar novos atributos através dos dados brutos que foram extraídos diretamente do grafo em seus distintos momentos do tempo. Além disso, a informação de crescimento do grafo pode ser finalmente transformada em um dado mais palpável, ao compararmos a quantidade de novos pesquisadores entre dois intervalos de tempo diferentes. Mais especificamente, novos dados são gerados para responder a seguinte pergunta: "quantos novos filhos um pesquisador tem se compararmos um momento passado com um momento futuro?".

4.1.1 Manipulação do grafo

Partindo do conjunto de dados fornecido como entrada, o primeiro passo que deve ser realizado antes das manipulações é a sua leitura e representação no formato de um grafo. Apesar do conjunto de dados já conter as informações necessárias para a construção do grafo (como discutido anteriormente), ainda se faz necessário o carregamento dessas informações e da construção de sua estrutura em um formato que pode ser percorrido e manipulado. Tipicamente, os dados para a construção de um grafo se apresentam em duas partes distintas, descrevendo os vértices e as arestas separadamente. A Tabela 1 mostra uma possível

Vértice	Arestas (vértice, tempo)
A	(C, 0)
B	(D, 1), (K, 2)
C	(D, 0), (E, 0), (F, 1)
D	(F, 0), (G, 1), (E, 2), (L, 2)
E	(H, 2)
F	(I, 2)
G	(M, 2), (I, 2)
H	(I, 1), (J, 1)
I	(J, 2)
J	
K	
L	
M	

Tabela 1: *Tabela definindo os vértices e arestas de um grafo evolutivo fictício. As arestas estão definidas como uma tupla contendo os vértices que são apontados pelo vértice indicado na primeira coluna e o instante de tempo em que elas ocorrem. As células em branco indicam que não existe nenhuma aresta partindo do vértice correspondente.*

representação do grafo evolutivo apresentado na Figura 2. Este grafo será utilizado ao longo dessas seções como exemplo para o melhor entendimento do processo aqui descrito.

Sugere-se, portanto, que o grafo seja gerado a partir da criação de vértices como entidades (podendo ser objetos de uma classe, para o caso da programação orientada à objetos), seguido pela representação de suas interações como referências a essas entidades armazenadas nos próprios vértices (indicando uma lista dos identificadores das entidades relacionadas, por exemplo) ou como suas próprias entidades que armazenam ambos os vértices envolvidos na relação. Além disso, nota-se que para os grafos aqui utilizados as arestas são direcionadas e isso deve ser contemplado pela representação escolhida. Desta forma, é importante haver uma convenção que claramente indica o vértice que origina a aresta e aquele que a recebe. Indicar com rótulos os dois tipos distintos de vértices, ou convencionar que o vértice que recebe uma aresta é sempre referenciado (ao passo que o que a origina é quem o referencia), são alternativas para atender esse requisito.

Dado que neste trabalho as informações utilizadas são puramente topológicas, os dados necessários para a geração do grafo são apenas a definição dos vértices e da maneira como estão ligados. Além disso, também faz-se necessária a presença de informações temporais que indicam o momento de surgimento de cada um dos vértices e arestas, podendo ser expressas como um marcador temporal (uma data ou unidade de tempo) presente em um atributo

de uma dessas duas entidades. Portanto, outras informações do conjunto de dados são dispensáveis, o que torna o método agnóstico aos dados utilizados desde que satisfaçam ambas as condições anteriores.

Após a construção do grafo, pode-se começar a sua manipulação. As alterações realizadas no grafo são focadas na remoção de arestas e vértices que não serão utilizados nas análises futuras, assim como na eliminação de inconsistências nos dados.

Inicialmente, deseja-se focar apenas no momento de surgimento de vértices. Apesar do surgimento de arestas ser também relevante para uma análise de evolução do grafo, esse fenômeno não colabora para o seu crescimento. Note que arestas não podem ser criadas sem estarem ligadas a dois vértices distintos, implicando que as novas arestas ou unem vértices já existentes, ou surgem em conjunto com ao menos um desses vértices.

Desta maneira, retiram-se as arestas geradas entre dois vértices já existentes (chamadas de arestas secundárias), mantendo aquelas que foram introduzidas ao grafo em conjunto com novos vértices. Para tanto, escolhe-se a aresta incidente mais antiga de um vértice (aquela que aconteceu primeiro), enquanto todas as outras arestas incidentes são removidas. Este processo, além de retirar arestas que não são de interesse para esta análise, também simplifica a estrutura do grafo para a realização dos próximos passos. Isto se deve, principalmente, ao fato de que cada vértice terá apenas uma ou zero arestas incidentes após a remoção. A partir disso, um vértice ou se torna uma raiz, ou passa a ter uma única raiz como seu ascendente mais longínquo e existe um único caminho direto entre todos ambos. Além disso, a remoção de arestas secundárias faz com que componentes conexas grandes tendam a ser quebradas caso sejam compostas por componentes menores interligadas por este tipo de aresta. Deste modo, o grafo resultante pode ser entendido como uma floresta de árvores geradoras direcionadas, implicando também a não existência de ciclos. Este grafo é ideal para a análise do crescimento do grafo original, dado que apresenta caminhos que podem ser percorridos em uma ordem bem definida, mantendo toda a informação de quando os vértices foram gerados ao longo do tempo.

Para facilitar o transcorrer dos próximos passos, uma pequena alteração deve ser feita nos atributos dos vértices e arestas caso necessário: se a informação temporal existente no conjunto de dados está contida apenas nas arestas (indicando o momento em que surgem),

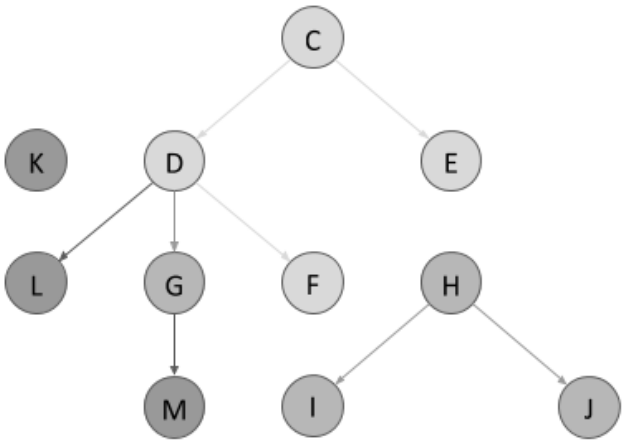

Figura 9: *Grafo evolutivo descrito pela Tabela 1 após a realização os procedimentos descritos na Seção 4.1.1. As cores dos vértices e das arestas indicam o momento de surgimento de cada uma delas, sendo cores mais escuras correspondentes a momentos mais avançados no tempo.*

essa informação deve ser passada para os vértices. Essa mudança é importante pois, como o crescimento do grafo é dado pelo surgimento de vértices, deve-se saber quando isso ocorre. Para isso, define-se o momento de surgimento de um vértice como aquele de sua primeira aresta incidente. Isso faz sentido para grafos de genealogia, uma vez que novos pesquisadores são introduzidos ao grafo com o passar dos anos e apenas caso tenham recebido uma orientação.

Contudo, existe um problema associado à definição anterior. Como o ano de surgimento dos vértices passa a ser definido por suas arestas incidentes, vértices que não recebem nenhuma aresta não possuem essa informação. Perceba que esse tipo de vértice é definido como uma raiz do grafo, as quais normalmente correspondem a um número pequeno de casos. Para resolver esta situação retiram-se as raízes do grafo, deixando apenas os vértices que tem arestas incidentes. O grafo resultante ainda contém raízes, no entanto, criadas a partir dos vértices que recebiam arestas das raízes originais.

Por fim, o último passo dessa etapa foca na resolução de conflitos e inconsistências nas informações contidas no grafo. Em grafos de genealogia, particularmente, existe o requisito de que vértices devem ou ser raízes, ou ter um vértice mais antigo que o originou através de uma orientação. Porém, ainda existem casos em que vértices antigos recebem orientações de vértices mais novos. Isso pode ocorrer por diversos motivos distintos, sendo dois deles (i) o preenchimento de dados temporais incorretos, tais como a inversão de caracteres em datas como 1990 e 1909, e (ii) a falta de cadastros relativos à vértices antigos, tal como o caso em

que um pesquisador cadastra sua orientação de doutorado, mas não a de mestrado que foi a primeira a o originar.

A resolução destes conflitos se baseia na seguinte ideia: caso um vértice A aponte para n vértices e apenas um deles (vértice B) é inconsistente, isso é um indício de que B tenha seu cadastro incorreto e deve ter suas informações alteradas para que surja após A. No entanto, se esta alteração gerar novos conflitos com vértices apontados por B, temos um indício de que a informação deste vértice é de fato verdadeira e que a informação de A é incorreta. Podemos, portanto, alterar a informação de A para que surja antes de B. Note que essa alteração não causa novos conflitos com outros vértices apontados por A mas, se A é apontado por um vértice C, um novo conflito pode ser gerado. Neste caso considera-se que a aresta entre A e B é a causa dos problemas e, portanto, removê-la resolve todos os conflitos encontrados para este vértice. No caso em que mais de um vértice apontado por A é inconsistente, a primeira etapa do processo pode ser pulada e as etapas ocorrem a partir da alteração das informações de A.

A Figura 9 mostra o grafo obtido pela aplicação do processo descrito ao grafo definido pela Tabela 1. Note que, comparado ao grafo original (exposto pela figura Figura 2 no tempo $t = 3$), este grafo exibe um número inferior de vértices, que foram excluídos por serem raízes, e arestas, correspondentes àquelas que foram retiradas por incidirem sobre vértices já existentes no tempo $t = 2$ e $t = 3$. Além disso, existem duas novas componentes do grafo, geradas pela remoção de arestas secundárias e inconsistentes (tais como a aresta entre o vértice E e o vértice H), cada uma com uma raiz. Por fim, vale observar que o grafo é uma floresta de árvores geradoras direcionadas, não apresentando ciclos e com um caminho direto partindo de todas as raízes para todos os vértices pertencentes à sua componente.

Após a realização de todos estes passos, o grafo está pronto para ser processado pelas próximas tarefas.

4.1.2 Extração da informação temporal do grafo

Com o grafo processado e alterado para satisfazer as condições necessárias, a próxima tarefa da etapa de pré-processamento dos dados é a extração da informação temporal do

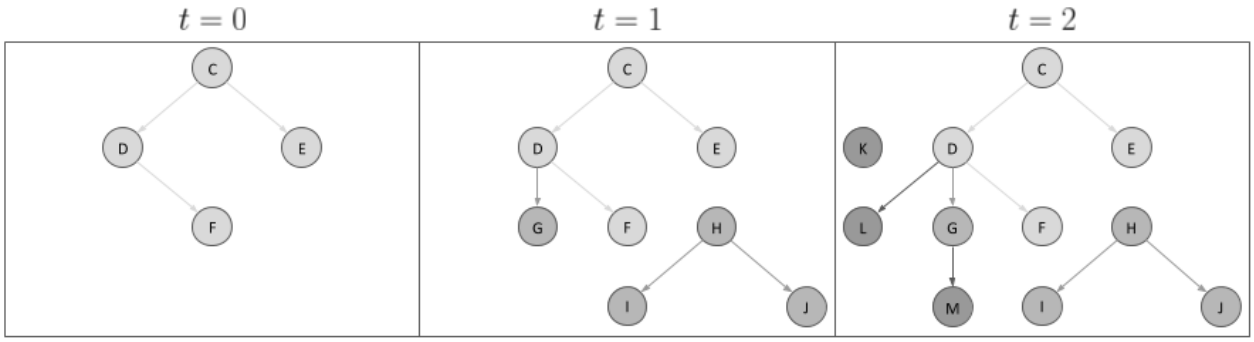

Figura 10: *Evolução temporal do grafo da Figura 9, obtido pelo processamento do grafo original da Figura 2 seguindo os passos descritos na Seção 4.1.1.*

grafo. Como já indicado anteriormente, os dados que indicam o momento de surgimento de um vértice e de uma aresta já estão presentes no grafo. No entanto, esta informação não pode ser observada a partir da representação padrão do grafo, que normalmente o expõe apenas no momento mais recente possível. Nesta visão, o grafo contêm todos os vértices e todas as arestas, independentemente do momento em que surgiram.

O crescimento do grafo pode ser notado, contudo, quando comparamos a sua representação em dois momentos distintos. Se no momento mais recente tem-se todos os vértices e arestas, então em um momento passado algumas destas entidades não estarão presentes, dado que não haviam surgido então. Ao analisar o grafo em diferentes momentos, portanto, tem-se uma visão geral de sua evolução.

A extração da informação temporal é dada, portanto, pela observação do grafo ao longo de toda sua existência, de modo que apenas os vértices e arestas já existentes em um dado período são selecionados para a análise. Note que, como o grafo utilizado nesta tarefa é representado por uma floresta de árvores geradoras mínimas, nas quais os vértices mais antigos estão mais próximos às raízes, enquanto vértices mais jovens estão mais próximos às folhas além de surgirem conjunto de novas arestas (conceitos já explorados na Seção 4.1.1), a evolução do grafo sempre é dada pelo crescimento das árvores seguindo o fluxo de suas arestas. Observe também que vértices e arestas são estáticas a partir do momento que surgem, indicando que não voltarão a desaparecer. Dito isso, os grafos gerados a cada intervalo de tempo passado são subgrafos de todos os grafos mais recentes que o seguem no tempo. Ao final do processo, um conjunto de grafos é obtido, representando cada estado passado do grafo. O número de grafos depende do intervalo escolhido entre observações.

Vértice	Tempo	Pai	Idade	Filhos	Irmãos	Descendentes
C	0	-	0	2	0	3
D	0	C	0	1	1	1
E	0	C	0	0	1	0
F	0	D	0	0	0	0
C	1	-	1	2	0	4
D	1	C	1	2	1	2
E	1	C	1	0	1	0
F	1	D	1	0	1	0
G	1	D	0	0	1	0
H	1	-	0	2	0	2
I	1	H	0	0	1	0
J	1	H	0	0	1	0
C	2	-	2	2	0	6
D	2	C	2	3	1	4
E	2	C	2	0	1	0
F	2	D	2	0	2	0
G	2	D	1	1	2	1
H	2	-	1	2	0	2
I	2	H	1	0	1	0
J	2	H	1	0	1	0
K	2	-	0	0	0	0
L	2	D	0	0	2	0
M	2	G	0	0	0	0

Tabela 2: *Tabela com algumas das medidas extraídas para os grafos expostos na Figura 11.*

A análise comparativa dos diversos grafos gerados é dada pela comparação de suas características. Entretanto, em vez de utilizar sua estrutura representada pela interligação de vértices, cada grafo passa a ser representado por um conjunto de medidas que são capazes de capturar suas particularidades. Assim, analisando as mesmas medidas, todos os grafos podem ser comparados simultaneamente e de um modo facilmente escalável. Além disso, ao armazenar os dados obtidos em formato tabular, este método permite que as informações possam ser facilmente manipuladas e analisadas de acordo com técnicas de aprendizado de máquina. Por fim, dado que a representação da estrutura do grafo não será mais utilizada nas etapas seguintes, não existe uma grande perda acarretada pela transformação da forma em que os dados são representados.

As medidas extraídas de cada grafo podem ser tanto globais (caracterizando o grafo como um todo), ou locais (caracterizando vértices individuais). Neste trabalho, decidiu-se utilizar apenas medidas locais para a realização das futuras análises, dado que considera-se que estas

podem dar indícios do crescimento do grafo com uma confiabilidade similar se não melhor do que dados globais. Isto se deve por diversos motivos a análise de vértices individuais pode indicar crescimentos heterogêneos no grafo que variam de acordo com regiões distintas, os quais seriam obscurecidos pelo crescimento homogêneo apresentado quando analisa-se o grafo de forma global. Além disso, vértices com características distintas podem apresentar comportamentos semelhantes, o que pode facilitar o entendimento de algum fenômeno que afeta o comportamento de um grupo isolado de vértices. Esta análise permite, assim, identificar quais regiões e vértices colaboram para a evolução do grafo e quais as características que ditam esse crescimento. Por fim, o uso de dados exclusivos para cada vértice permite a multiplicação do número de dados disponíveis para a análise e posterior treinamento por um algoritmo de aprendizado de máquina, dado que cada vértice do grafo apresentará um conjunto de informações para cada ano, em oposição a apenas uma informação para o grafo completo.

As informações extraídas de cada vértice em cada período são puramente topológicas ou temporais. Também decidiu-se utilizar medidas que indiquem características específicas de cada vértice e de sua região do grafo, em oposição a medidas relacionadas a topologia do grafo como um todo. Além disso, também são armazenadas informações referentes ao estado do vértice e sua origem, tais como o seu rótulo (identificação que o caracteriza como único), o rótulo do vértice que o originou (que é único, caso exista) e o ano de observação. Para um vértice são coletadas, portanto, as seguintes informações[4]:

- Tempo
- Rótulo
- Rótulo do pai
- Idade do vértice
- Número de filhos
- Número de irmãos

[4]Para facilitar o entendimento das medidas coletadas, a nomenclatura utilizada segue a convenção descrita na Seção 2.1, na qual relações entre vértices são denominadas de acordo com seu paralelo com as relações de uma árvore genealógica familiar.

- Número de primos
- Número de ascendentes
- Número de descendentes
- Maior distância até uma folha

Vale notar que, apesar de poucas informações serem extraídas, essas podem ser facilmente manipuladas para a geração de outras medidas. Note também que, devido a estrutura de árvores geradoras direcionadas que os grafos apresentam, diversas outras medidas não utilizadas já estão contempladas pelos dados extraídos, tais como o grau de entrada e o grau de saída do vértice, que já estão indicados pelo número de ascendentes e pelo número de filhos de um vértice, respectivamente. Finalmente, apenas o rótulo do pai do vértice é armazenado, dado que permite rastrear a origem do vértice e das medidas referentes a ele, ao passo que armazenar todos os rótulos dos filhos seria inviável.

Ao final da extração destes dados, serão obtidas um número de tabelas igual ao número de grafos observados. Estas tabelas são compostas por exatamente as mesmas colunas, apesar de conterem um número distinto de linhas, dado que a quantidade de vértices aumenta de um instante para outro. Essas tabelas são concatenadas, de forma a gerar uma nova tabela contendo todas as informações de todos os grafos para todos os instantes observados.

4.1.3 Criação de novos atributos

A partir da tabela obtida pelo extração da informação temporal na tarefa anterior, esta tarefa visa manipular os dados de modo a extrair novas medidas e atributos que possam ser úteis na análise do crescimento do grafo.

Uma primeira abordagem para o enriquecimento do conjunto de dados é gerar novos atributos para um vértice baseados nas informações do mesmo período de observação do vértice que o deu origem. Pode-se, por exemplo, gerar atributos tais como “número de primos do pai” e “número de descendentes do pai” que poderão, futuramente, indicar algum comportamento de crescimento do filho.

Além disso, atributos podem ser criados a partir da combinação de duas ou mais medidas através de cálculos matemáticos. Por exemplo, pode-se calcular o número médio de filhos gerado por ano por um vértice, ou o número de descendentes de um vértice dividido pelo número de descendentes do pai. Essas medidas, apesar de serem promissoras para proporcionar um ganho com relação à capacidade preditiva dos modelos de aprendizado de máquina que serão treinados na próxima etapa, diminuem a interpretabilidade dos dados e, uma vez que o entendimento do motivo do crescimento observado é um dos principais objetivos deste trabalho, este tipo de manipulação foi evitada.

Por fim, para analisar o impacto de cada vértice individual no crescimento do grafo, o número de novos vértices originados é armazenado. O cálculo utilizado para obter tal informação é simplesmente a diferença do número de filhos de um vértice em dois instantes de observação distintos. O intervalo entre os dois momentos pode ser arbitrariamente escolhido, de modo a obter uma informação sobre um futuro arbitrariamente distante. Contudo, esta informação não poderá ser obtida para todos os vértices em todos os instantes observados, uma vez que grafos mais recentes não contêm a informação sobre os acontecimentos mais distantes no futuro.

4.2 Treinamento de modelos preditivos

Os dados obtidos pelo procedimento descrito na Seção 4.1, ainda que possam ser utilizados para analisar o crescimento do grafo em diversos instantes do passado, não permite o estudo de sua evolução para um momento futuro. A próxima etapa do procedimento metodológica descrita nesta seção visa, portanto, extrair regras de crescimentos a partir dos dados do passado para a realização de previsões para o futuro. Para tanto, modelos de aprendizado de máquina (Seção 2.2) serão utilizados, buscando aprender as principais características de vértices que os fazem gerar novos vértices.

Esta etapa, assim como a anterior, pode ser divididas em tarefas que, em conjunto, utilizam os dados advindos do grafo para treinar o modelo. A primeira tarefa, chamada de "Definição do problema e variável de saída" explora as possíveis abordagens a serem seguidas para o treinamento de um modelo, incluindo o tratamentos dos dados para sua posterior

utilização e a criação de variáveis de saída para o modelo que responderão a perguntas específicas sobre o futuro. Já segunda tarefa trata das técnicas a serem utilizadas durante o treinamento, abordando temas como a quantidade e divisão dos dados e a avaliação dos resultados do modelo.

4.2.1 Definição do problema

Existem diversas abordagens distintas que podem ser exploradas partindo do conjunto de dados obtido pela Seção 4.1.3 e, antes do início do treinamento de um modelo preditivo, decisões devem ser tomadas com relação a qual percurso seguir. A primeira questão a ser respondida neste cenário está relacionada ao problema que está sendo analisado.

Sabendo que o crescimento do grafo está sendo estudado sob o ponto de vista de surgimento de novos vértices, fica evidente que o modelo deve responder quantos vértices serão gerados em um momento futuro. Ainda, como o conjunto de dados que será utilizado durante o treinamento do modelo é composto por amostras que representam um vértice em um momento específico do tempo, esta previsão será dada pelo número de novos vértices gerados a partir de cada um dos vértices existentes. A qualidade das previsões pode ser medida ao se comparar estes resultados obtidos pelo algoritmo com o número real observado de vértices gerados.

Um vértice pode, no entanto, gerar um número novos vértices em um intervalo com limite inferior zero e limite superior é inexistente, caracterizando uma tarefa de aprendizado supervisionado para regressão. Este comportamento, apesar de ser esperado e observado em situações reais, dificulta a tarefa de previsão do modelo, uma vez que aumenta a quantidade de possíveis valores os quais a variável de saída pode tomar. Esse fato também se torna um problema quando o número de vértices gerados é pequeno (na ordem de poucas dezenas), uma vez que um erro na previsão pode significar uma variação muito grande do acontecimento real.

Uma tarefa alternativa, que pode facilitar o treinamento de um modelo e aumentar o seu desempenho, é estimar se um vértice gerará, ou não, um ou mais novos vértices. Esta abordagem, apesar de similar a anterior, tem como variável de resposta apenas um valor

binário (sim ou não), caracterizando um problema de aprendizado supervisionado para a classificação. A realização desta tarefa não requer nenhuma alteração ao conjunto de dados a ser utilizado para a regressão, com exceção da variável de saída que deverá ser corrigida para valores indicando a ausência ou presença de novos filhos. Perceba que este novo problema mitiga as dificuldades indicadas anteriormente ao reduzir o espaço de valores possíveis de resposta. Em contraponto, a tarefa de classificação resulta em respostas mais genéricas que convêm menos informação que a tarefa de regressão, não informando um número preciso de novos vértices gerados.

4.2.2 Abordagens para o treinamento de modelos

A partir do problema definido na seção anterior, pode-se iniciar o treinamento com o conjunto de dados escolhido. Entretanto, existem considerações a serem feitas com relação a quais dados serão efetivamente utilizados para treinar o modelo.

Caso todo o conjunto de dados seja utilizado para o treino, o modelo aprenderá perfeitamente as informações contidas nele e terá um ótimo desempenho quando realizando previsões sobre os dados contidos nele. Entretanto, se um conjunto de novas amostras, distintas de todas aquelas presentes no conjunto de treino, forem apresentadas para o modelo para a obtenção de novas previsões, existe a possibilidade do desempenho ser muito inferior àquele observado inicialmente. Este fenômeno, conhecido como sobre-ajuste, pode inviabilizar o uso do modelo para as aplicações desejadas, devido à baixa confiança na qualidade de seus resultados. Como discutido na Seção 2.2, um modelo preditivo deve ser capaz de generalizar o conhecimento disponibilizado pelo conjunto de dados para dados novos, oferecendo resultados confiáveis mesmo para amostras nunca antes vistas.

Para obter uma estimativa mais confiável do comportamento do modelo para dados não observados anteriormente, o uso de um conjunto de teste é recomendado. O conjunto de teste, o qual é um subconjunto do conjunto de dados original, não é utilizado pelo modelo durante o treinamento, servindo apenas para avaliar seu desempenho ao final do treinamento. Ao obter previsões sobre este conjunto e compará-las com os resultados reais conhecidos, é possível obter uma medida confiável do desempenho do modelo. É importante, assim, que nenhuma informação prévia do conjunto de teste seja conhecida antes do final do treinamento, de modo

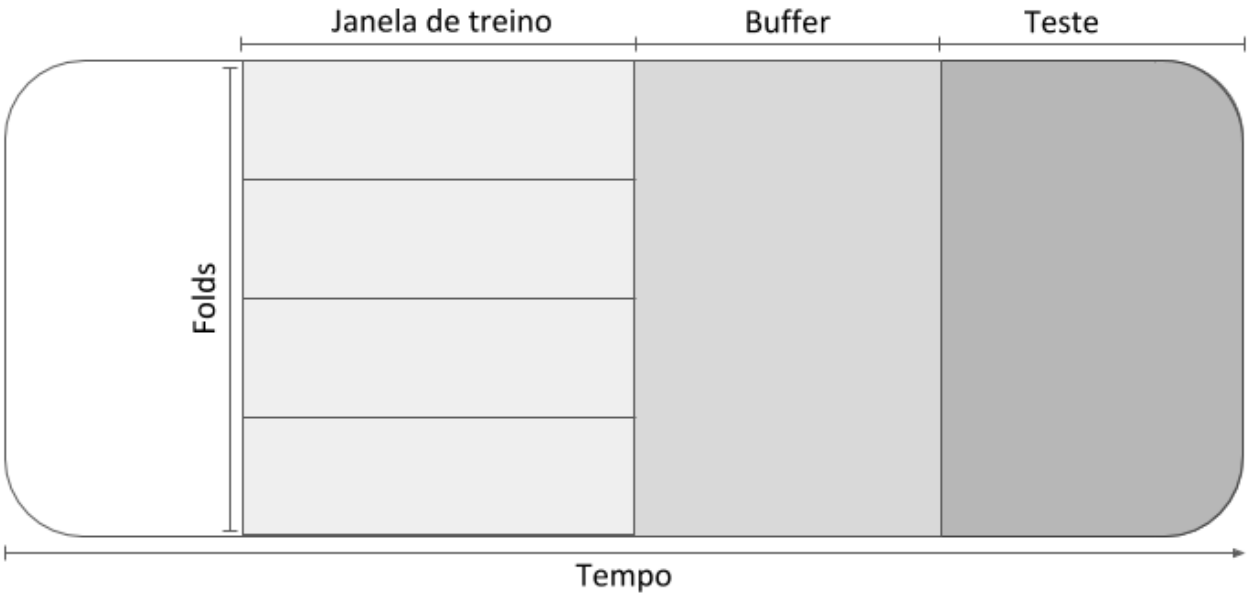

Figura 11: *Demonstração da forma como os dados são divididos para o treinamento do modelo.*

a evitar que o modelo aprenda indiretamente suas características e volte a realizar previsões incorretas.

Para o problema de previsão de novos vértices em um grafo, o conjunto de testes pode ser extraído do conjunto de dados de duas maneiras distintas: (i) através de uma divisão espacial, ou (ii) através de uma divisão temporal. A divisão espacial é dada pela separação de amostras de um mesmo período de tempo e permite analisar se um modelo é capaz de aprender as características de amostras desconhecidas. A divisão temporal, por outro lado, é importante para estimar o quanto os dados do passado conseguem prever o futuro, indicando qualquer mudança que possa ter ocorrido com a distribuição das amostras com o passar do tempo. O uso destes dois tipos de teste permite uma visão combinada destes dois fatores distintos e este é o método recomendado para este trabalho.

O uso de um conjunto de testes temporal tem como desvantagem o fato que parte dos dados não poderão ser utilizados para o treino nem para o teste. Considerando que os dados do conjunto de teste iniciam-se a partir de um determinado momento, os dados do conjunto de treino que descrevem um período de tempo anterior mas suficientemente próximo a este podem conter informações sobre as características do conjunto de teste, o que fere o seu princípio. Assim, deve-se utilizar uma folga entre os dados de treino e de teste, de modo a evitar este vazamento de informações e o enviesamento dos resultados das predições.

O conjunto de testes com divisão espacial já não oferece este problema e pode ser escolhido

arbitrariamente. Para este método, escolheu-se utilizar a técnica de validação cruzada para avaliar a capacidade de generalização do modelo para dados não pertencentes ao conjunto de treinamento. Divide-se, assim, o conjunto de treinamento em um número k de subconjuntos, chamados de *folds*. Para o treinamento $k - 1$ subconjuntos são utilizados, sendo o k-ésimo subconjunto selecionado para validar os resultados obtidos. Este processo se repete para cada um dos *folds*, garantindo o entendimento do desempenho do modelo para diferentes tipos de conjuntos de treino e teste. Baseando-se nesta técnica é possível ajustar hiperparâmetros do modelo, os quais podem afetar a qualidade das predições, ao analisar o efeito das alterações nos resultados.

Por fim, uma "janela de treinamento" foi definida para limitar a quantidade de dados utilizados pelo modelo para o treinamento. Essa janela se inicia na data mais recente dos dados do modelo e se estende para o passado, por uma quantidade arbitrária de tempo. Este procedimento permite analisar como o desempenho do modelo cresce de acordo com a quantidade de dados e pode colaborar para o entendimento de como os dados do passado estão relacionados com o futuro.

5 Resultados

Esta seção discute os resultados obtidos a partir da aplicação dos métodos descritos na Seção 4 sobre o grafo da Seção 3. Para a obtenção dos resultados a linguagem de programação *Python* foi utilizada em conjunto com bibliotecas de manipulação e processamento de dados (*Pandas*, *NumPy* e *SciPy*) e da biblioteca de análise preditiva de dados *scikit-learn*.

Os resultados serão apresentados seguindo o fluxo definido durante a exposição da metodologia. Primeiro, exploram-se as informações obtidas a partir do processamento do grafo e da extração de suas características temporais. Em seguida, serão discutidos o treinamento do modelo preditivo e os resultados de suas previsões para o futuro.

5.1 Informações temporais do grafo Acácia

O grafo de genealogia da Plataforma Acácia foi processado de acordo com o método indicado nas Seções 4.1.1 e 4.1.2. Seguindo a ordem das tarefas definidas durante estas seções, o método de processamento se inicia com a remoção de arestas secundárias. A remoção de tais arestas resultou na remoção de 261785 arestas, deixando o grafo com um total de 1142324 arestas. Além disso, o número de componentes conexas mais que dobrou, partindo de 56263 para um total de 130266 (igual ao número de raízes), sendo a maior componente do grafo original quebrada em diversas outras. Das novas componentes, 101404 são compostas por dois ou menos vértices, contendo um total de 173549 vértices, enquanto as dez mil maiores componentes correspondem a 76% de todos os vértices.

Em seguida, dado que a informação temporal do grafo original está contida apenas nas arestas, as raízes do grafo original foram retiradas, uma vez que não se tem informações sobre elas. A remoção das raízes elimina componentes conexas formadas por apenas um vértice, além de remover arestas entre vértices ligados à raízes, o que pode gerar novas componentes. Esta operação também gera novas raízes caso os vértices tenham sido gerados diretamente por uma raiz. Ao final da remoção, restaram 1037337 vértices e 1007132 arestas, que equivalem a 82% e 72% dos valores do grafo original, respectivamente. Além disso, 30205 novas raízes foram geradas, que originam um número igual de componentes conexas. O número de componentes pequenas reduziu consideravelmente, sendo 57% delas compostas por dez ou

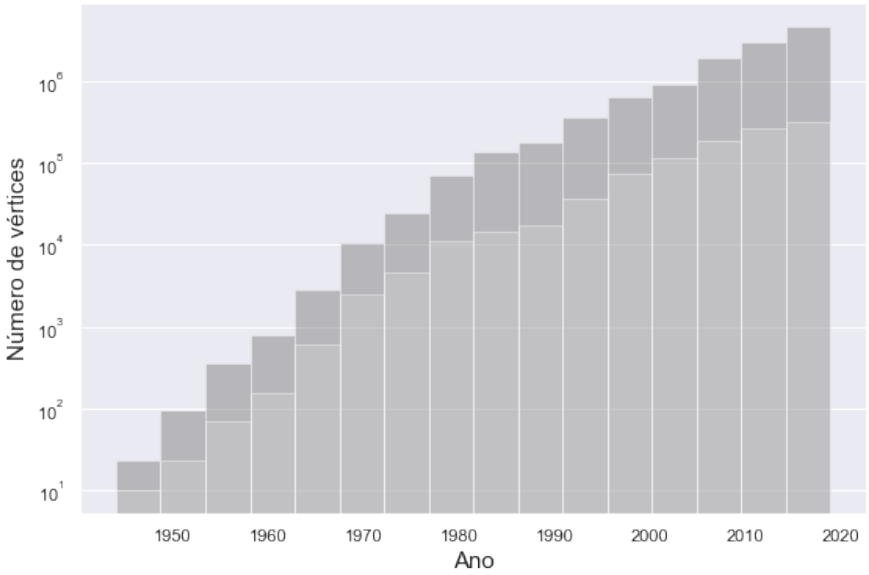

Figura 12: *Número de vértices por ano para o grafo de genealogia acadêmica da Plataforma Acácia. A cor mais escura indica o número total de vértices, enquanto a cor mais clara indica a quantidade de novos vértices.*

mais vértices.

Por fim, as inconsistências nos anos de surgimento dos vértices foram endereçadas. Antes de tratados, o total filhos que surgiram antes de seus pais era de 1019 e, a partir da execução do algoritmo descrito na Seção 4.1.1, todos estes casos foram tratados. Além disso, dados aberrantes, tais como anos de surgimento inviáveis, foram tratados manualmente. Ao fim deste processo todos os vértices passaram a ter anos de surgimento num intervalo entre 1944 e 2019.

As informações temporais foram extraídas deste grafo resultante seguindo o método indicado na Seção 4.1.2. Para cada ano no intervalo de anos de surgimento, informações sobre cada um de seus vértices foram colhidas, partindo de um único vértice no primeiro ano até o número total de vértices no último ano. O processo resultou em 76 conjuntos de dados distintos, os quais foram unidos para gerar um único conjunto de dados que pode ser manipulado.

Novos atributos foram gerados pela manipulação dos dados do conjunto. Atributos foram calculados para o pai e o avô de cada vértice, sendo definidos como zero ou nulo caso esses não existam. Em especial, a variação no número de filhos foi calculada para um número arbitrário de anos no futuro (δ), através da diferença entre o número de filhos de um vértice para um grafo no ano k e para um grafo futuro no ano $k + \delta$. Neste caso, foi utilizado um valor de δ de um a dez anos.

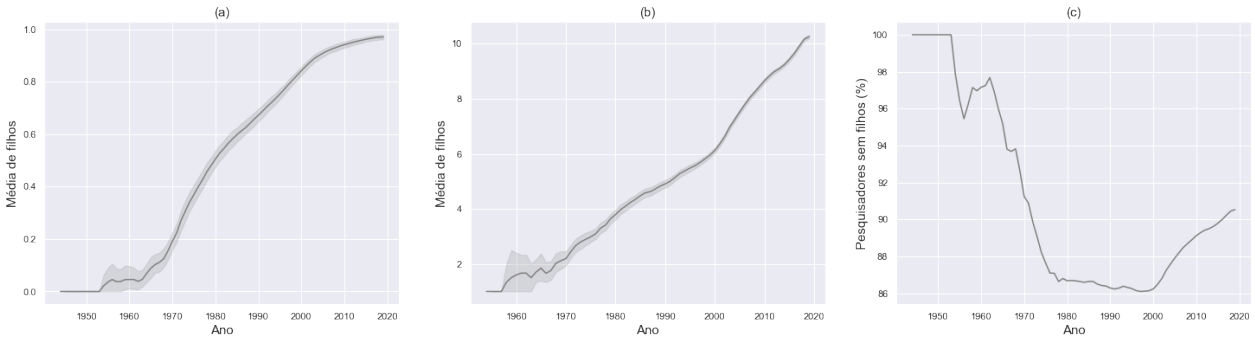

Figura 13: *(a) Média de novos filhos por pesquisador por ano, (b) média de novos filhos por pesquisador com ao menos um filho por ano e (c) porcentagem de pesquisadores sem nenhum filho por ano. A área sombreada representa o intervalo de confiança das curvas.*

Este novo conjunto de dados contém 11928501 amostras, estando distribuídas conforme mostra a Figura 12. A figura indica o também o número de novos vértices surgidos a cada ano no grafo. Note que ambos o número de vértices total e o número de novos vértices crescem de maneira exponencial para este grafo, apresentando uma forma similar com o crescimento observado à Figura 6[5]. Além disso, ao comparar a diferença entre o número total de vértices e o número de novos vértices, vê-se que existe um distanciamento cada vez maior dos dois números: em 1978, vértices novos correspondiam a 17% do número total deste vértice, caindo para 11% em 1998 e 7% em 2018. Este fenômeno pode indicar uma diminuição no ritmo de formação de novos pesquisadores, dado que um número maior de pesquisadores produz uma quantidade proporcionalmente menor.

Para explorar mais a fundo esta ideia, a Figura 13 (a) indica o número médios de filhos por pesquisador por ano. Nota-se que existe uma tendência crescente para o número de filhos por pesquisador até o final dos anos 1990. Esta tendência passa a diminuir a partir do início dos anos 2000, se estabilizando nos anos seguintes. No entanto, ao se excluir os dados dos pesquisadores sem nenhum filho, selecionando aqueles que tiveram ao menos um, percebe-se que este fenômeno de redução na produção não é observado (Figura 13 (b)). Esta diminuição pode estar relacionada, portanto, ao número crescente de pesquisadores que não produziram nenhum filho nos últimos anos. A Figura 13 (c) demonstra este efeito, apresentando crescimento da quantidade de pesquisadores sem nenhum filho a partir dos anos 2000.

[5]Como discutido na Seção 3, o crescimento observado na Figura 6 não é o mesmo do crescimento aqui apresentado, já que essa mostra o número de arestas ao invés do número de vértices.

Figura 14: *(a) Média de novos filhos por pesquisador por ano e (b) porcentagem de pesquisadores sem nenhum filho por ano, para janelas de tempo futuro (em anos) de tamanho variado*

Também foram analisadas as produções de novos filhos para os pesquisadores do grafo para cada ano. A Figura 14 (a) indica os valores da média de novos filhos no futuro por pesquisador (para os pesquisadores que tiveram ao menos um filho) a cada ano, enquanto a Figura 14 (b) apresenta a porcentagem de pesquisadores sem nenhum filho por ano. Como estes dados se referem ao futuro e esse futuro pode ser observado para um tempo arbitrariamente distante, foram selecionadas janelas de tempo de tamanho distinto para a análise. Assim, para cada ano, pode-se ver quais os valores calculados para essas duas medidas para cinco momentos do futuro diferentes (um, três, cinco, sete e dez anos). Perceba que as curvas acabam em anos distintos, uma vez que dados completos estão disponíveis apenas para anos anteriores ao ano mais recente subtraído do tamanho da janela.

Primeiro, observa-se que o número médio de vértices para os pesquisadores que têm ao menos um filho está crescendo com o passar do tempo para todas as janelas de tempo futuro. Apesar de sofrer uma leve queda da metade para o final dos anos 1970, percebe-se que houve um grande crescimento a partir do começo dos anos 1980. Este número se estabilizou no final dos anos 1990 e voltou a crescer na metade dos anos 2000.

Além disso, o número de filhos para a janela de um ano no futuro apresenta um crescimento leve quando comparado com as outras curvas. Isso indica que nem todos os pesquisadores têm filhos acadêmicos todo ano mas que os mesmos pesquisadores costumam ter filhos quando observados por um período mais extenso. Ou seja, dado que um pesquisador tem um filho, existe uma tendência a ele ter mais filhos.

Por outro lado, o efeito inverso é observado na Figura 14 (b). O número percentual

de pesquisadores sem nenhum novo filho sofre uma grande queda nos anos 1970 e mostra um comportamento crescente desde então, indicando proporcionalmente que cada vez mais pesquisadores não estão gerando outros. Além disso, apesar de mais de 90% dos pesquisadores não formarem nenhum filho quando consideramos um intervalo de um ano, este número cai para em torno de 85% quando considera-se um intervalo de dez anos. Assim, conforme observamos um número maior de anos para o futuro, o número de pesquisadores sem novos filhos também é menor ou, em outras palavras, aproximadamente 15% dos pesquisadores tem novos filhos em até dez anos.

Vale ressaltar também que todas as figuras apresentam uma variação brusca para anos antigos. Essa variação ocorre devido à pouca quantidade de dados presentes no conjunto de dados para anos anteriores a 1970, o que aumenta a incerteza das medidas calculadas para essas datas. Existe também uma alteração acentuada para anos mais recentes, podendo esta ser atribuída ao atraso na atualização das informações pelos pesquisadores, sendo necessário um tempo maior para que estes dados se estabilizem.

Todos estes fatos corroboram a ideia de que a produção acadêmica tem diminuído ao longo dos últimos, a partir do começo dos anos 2000. Ainda assim, existem ressalvas a serem feitas com relação a estes resultados, dado que as informações contidas na base de dados podem estar enviesadas. Este fenômeno pode ser causado, por exemplo, por baixa taxa de cadastramento de pesquisadores tanto atualmente, devido ao cadastro atrasado daqueles que se formam, quanto no passado, pela falta de dados históricos e pouco engajamento de pesquisadores mais antigos.

Por fim, a Figura 15 (a) apresenta a média de filhos por idade acadêmica dos pesquisadores. Os dados estão novamente divididos em janelas de tempo, que indicam o número de anos observados no futuro. Note que todas as curvas apresentam uma elevação para as idades acadêmicas entre 20 e 30 anos, indicando que esse é o momento de maior produção dos pesquisadores. Já a Figura 15 (b) mostra o número de filhos associado à idade acadêmica dos pesquisadores. De fato, este número passa a ser expressivo a partir dos 20 anos de idade acadêmica e mostra um crescimento estável ao longo dos anos. Além disso, o número de filhos para pesquisadores com menos de dez anos de idade acadêmica é muito próximo a zero, fenômeno causado pela grande quantidade de pesquisadores recém-formados e que ainda não

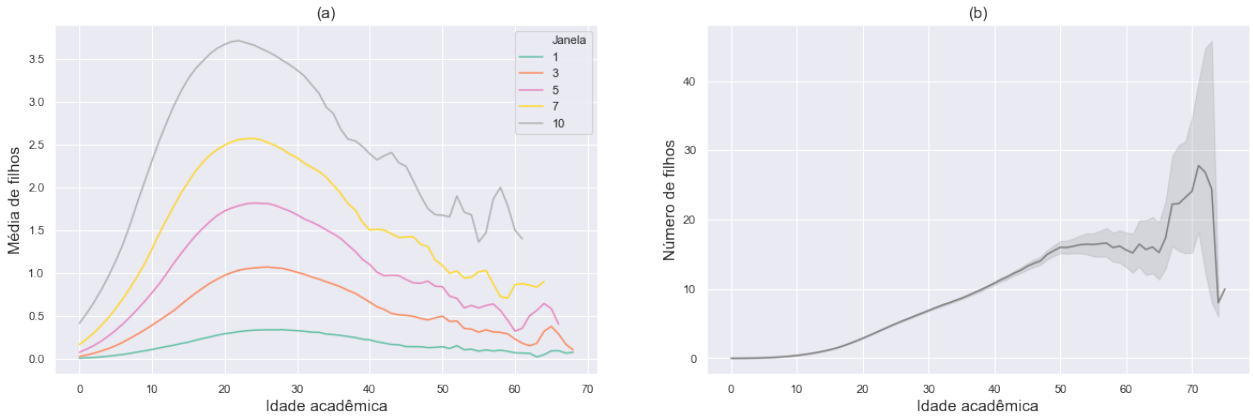

Figura 15: *(a) Média de novos filhos por pesquisador por idade acadêmica para janelas de tempo futuro (em anos) de tamanho variado e (b) número médio de filhos por pesquisador por idade acadêmica.*

tiveram oportunidade de orientação. Isto pode indicar que, apesar de haver uma grande quantidade de pesquisadores sem nenhum filho, é possível que sua produção comece a crescer nos próximos anos, se igualando àquela dos pesquisadores mais velhos.

5.2 Modelo de previsão de crescimento

Utilizando os dados analisados na Seção 5.1, modelos de aprendizado de máquina foram treinados para prever o crescimento do grafo em um momento arbitrariamente futuro. Para a geração desses modelos, seguiu-se os passos indicados na Seção 4.2.

A primeira decisão a ser tomada para iniciar a modelagem é o tipo de problema a ser abordado. Como discutido na Seção 4.2.1, a previsão do crescimento do grafo pode ser dada tanto como um problema de regressão, tentando prever o número exato de vértices que serão gerados, quanto como um problema de classificação, buscando indicar os vértices que gerarão ou não um novo pesquisador. Para esta análise o problema de classificação foi escolhido.

A obtenção de uma variável de saída que pudesse ser utilizada para o treinamento supervisionado de um modelo de classificação foi dada pelo processamento das variáveis que indicam quantos novos pesquisadores são gerados por vértice para anos futuros distintos. A variável foi definida como zero, caso o pesquisador não tivesse nenhum filho acadêmico, ou um, caso tivesse um ou mais filhos acadêmicos. Esta variável permite, assim, dividir os pesquisadores em duas categorias que podem ser utilizadas para a classificação. Ao final do processo foram geradas dez novas variáveis, uma para cada janela no futuro observada. Pesquisadores da

categoria “1” correspondem a 5% dos casos, para a janela de um ano no futuro, até 24% para uma janela de dez anos no futuro. Essas categorias, também chamadas de classes, foram nomeadas “tem filho” e “não tem filho”.

Para o treinamento de modelos foi utilizada a variável de saída correspondentes à três anos no futuro. O conjunto de treinamento é completado por outras 24 variáveis que descrevem as amostras, correspondendo à características topológicas de cada vértice para uma vizinhança próxima.

Neste trabalho utilizam-se árvores de decisão para a obtenção dos resultados. Isso se deve pois busca-se uma explicação das características que levam ao surgimento de novos vértices, e a perda de desempenho nas predições pelo uso de um algoritmo simples não é visto como um problema.

O treinamento dos modelos se iniciou com uma etapa de estimação de hiperparâmetros. Para isso, a técnica de busca aleatória foi utilizada, em que diversos conjuntos de hiperparâmetros são testados aleatoriamente para obter o melhor conjunto possível deles. Para evitar o enviesamento do modelo por sobre-ajuste, a otimização dos hiperparâmetros foi realizada em conjunto com uma validação cruzada.

A otimização de hiperparâmetros foi feita com todo o conjunto de treino para o qual a variável de saída é conhecida. Para o caso da variável de três anos e sendo a data mais recente do grafo 2019, a data máxima utilizada pra o treinamento é 2016. Os parâmetros testados foram a profundidade máxima da árvore, limitada a um máximo de seis, o número mínimo de amostras por folha, limitado ao máximo de 20%, e o número máximo de folhas, limitado ao máximo de 20. Os limites foram definidos para que não se perdesse a interpretabilidade do modelo. Os melhores parâmetros encontrados foram profundidade máxima de seis, um número máximo de folhas de quinze e um número mínimo de amostras por folha de 2% da base.

A avaliação dos modelos foi feita utilizando a medida F1 e a área abaixo da curva ROC. Para o conjunto de dados utilizado a acurácia do modelo não é a melhor medida de desempenho, devido ao grande desbalanceamento dos dados. Uma combinação de precisão e revocação é, portanto, desejada, de modo que possamos melhor avaliar as capacidades do modelo.

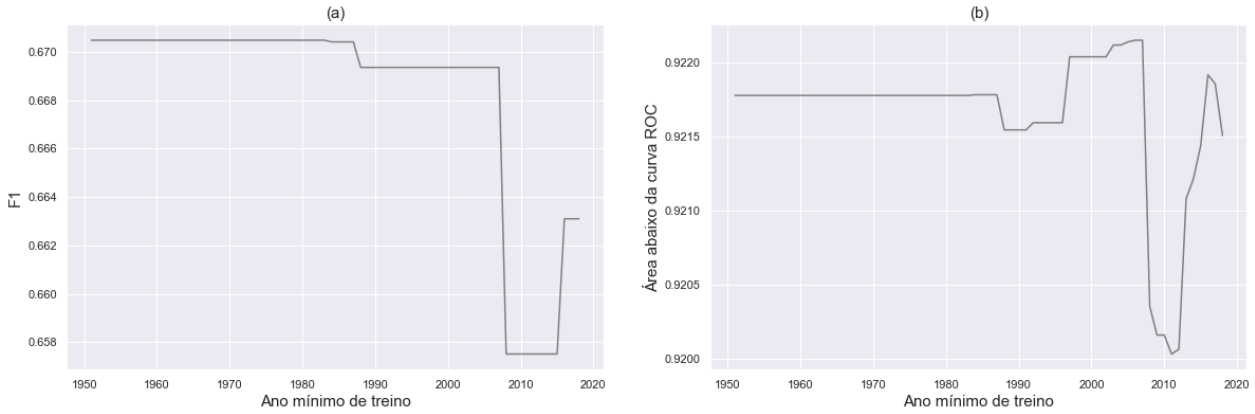

Figura 16: *Medidas de desempenho do modelo de classificação (a) F1 e (b) área abaixo da curva ROC, para diferentes conjuntos de dados com variação no ano mínimo de treino utilizado.*

Antes de decidir por fim qual modelo utilizar como o foco das análises, um último parâmetro do treinamento deve ser definido. Como utiliza-se dados do passado para treinar o modelo, é necessário saber se esses dados de fato colaboram para o aprendizado. Para isso, analisou-se o desempenho do modelo ao utilizar dados de um passado cada vez mais distante.

A Figura 16 mostra o resultado das medidas de desempenho F1 e da área abaixo da curva ROC para diferentes conjuntos de dado. A medida F1 foi calculada considerando o melhor limiar de decisão possível. O ano representado no eixo das abscissas indica o mais antigos dos dados utilizados para o treinamento de cada modelo. Nota-se que, após uma queda inicial do desempenho para os anos iniciais, esse volta a crescer conforme dados mais antigos são considerados. A medida da curva ROC volta a cair em meados dos anos 2000, mas se estabiliza em meados de 1980. A estabilização pode ser atribuída à pequena quantidade de dados presentes em anos mais antigos (vide Figura 12) o que diminui sua contribuição para os resultados.

Assim, podemos afirmar que o uso de dados do passado pode sim colaborar para resultados melhores. Utilizamos, desta forma, a base de dados inteira (considerando todos os anos) para o treinamento de um modelo final. No entanto, note que a variação percentual desta medida é pequena e o uso destes dados não altera os resultados de forma tão significativa.

O modelo final foi treinado, portanto, com os parâmetros anteriormente descritos e o conjunto de dados completo, considerando todos os anos possíveis. O conjunto de treino foi gerado a partir de dados de 1944 a 2012, contendo um total de 5906297 amostras. Já o

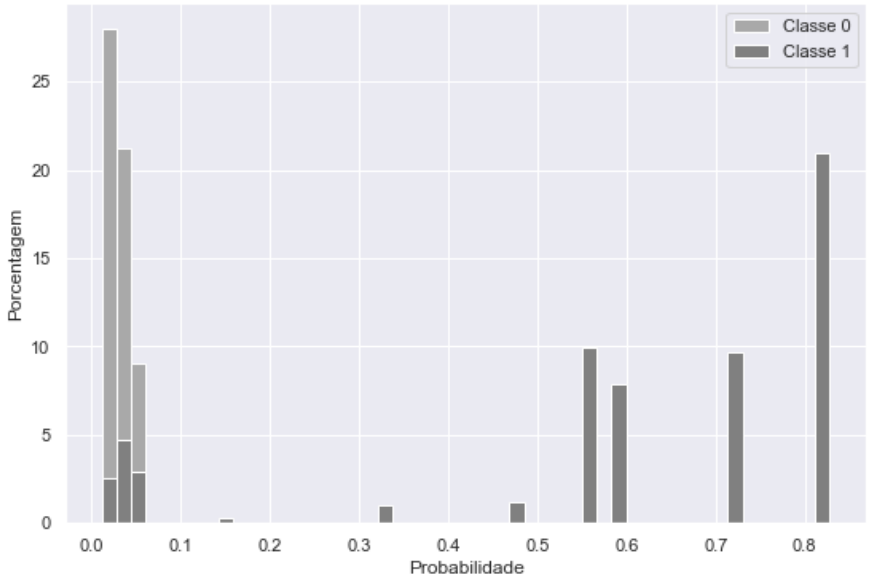

Figura 17: *Distribuição das classes reais "tem filho" (em cinza claro) e "não tem filho" (cinza escuro) pela probabilidade de pertencer à classe "tem filho" prevista pelo modelo classificador.*

conjunto de testes é composto de dados do ano de 2016, totalizando 865712 amostras ou $12,8\%$ do total. A base de treino contém $11,5\%$ de amostras da classe "tem filho", enquanto a base de treino contém $6,8\%$.

O modelo obteve uma medida de área abaixo da curva ROC de $0,92$, indicando que as distribuições de classes positivas e negativas estão com um grande grau de separação, e as classes podem ser distinguidas com grande grau de certeza. Este fato pode ser visto pela Figura 17, a qual mostra a probabilidade de uma amostra pertencer às classes "não tem filho" e "tem filho" que foi determinada pelo classificador, e como esta se compara à classe real das amostras. De fato, pode-se ver que as classes são facilmente separáveis caso um limiar seja traçado em pontos superiores a $0,1$. A única fonte de incerteza uma pequena quantidades de casos da classe "tem filho" que têm baixa probabilidade e que se misturam com as classes "não tem filho".

O valor utilizado como o limiar de classificação foi definido como 0.57, indicando que todas as amostras com probabilidade maior a essa serão definidas como pertencendo à classe "tem filho" e todas às outras à classe "não tem filho". Com este limiar o modelo consegue prever corretamente as classes de $94,7\%$ das amostras. No entanto, boa parte dessa acurácia é dada pelo acerto da classe majoritária ("não tem filho").

A Figura 18 mostra esse efeito através de uma matriz de confusão. Veja que o número

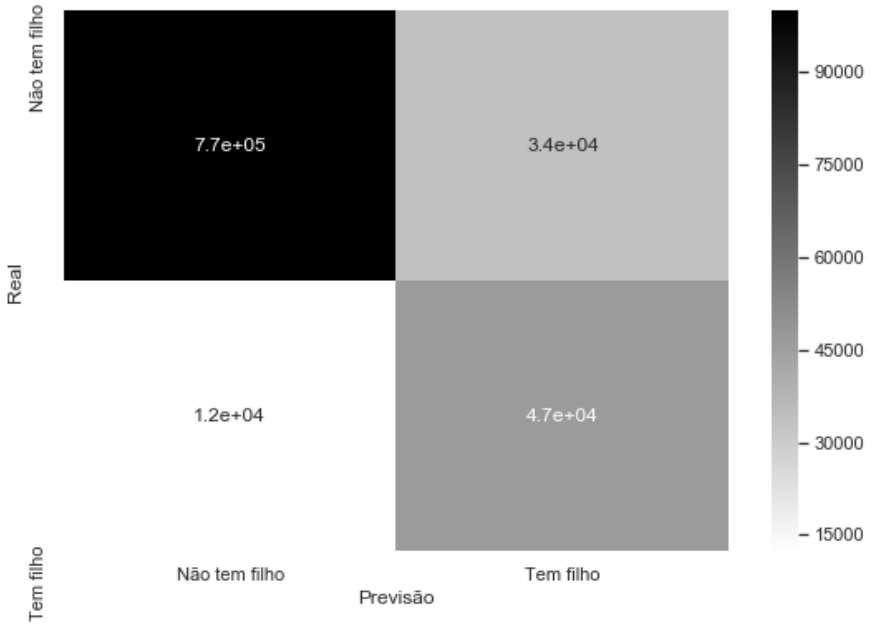

Figura 18: *Matriz de confusão para as previsões do modelo de classificação. O eixo das abscissas indica a classe prevista pelo classificador, enquanto o eixo das ordenadas mostra a classe real das amostras. Cores mais claras indicam uma menor concentração de amostras.*

de amostras da classe "não tem filho" classificados corretamente é muito superior à todos os outros. Por outro lado, o número de casos em que essa classe é classificada erroneamente é comparável às classificações corretas da classe "tem filho". No entanto, para os casos em que esta classe é classificada de forma errada (caso em que é atribuído um valor errado para a classe real), o número de ocorrências é consideravelmente menor.

Uma análise mais detalhada pode ser feita observando a precisão e a revocação do modelo, as quais medem a porcentagem de amostras corretamente classificadas como "tem filho" e a porcentagem de amostras "tem filho" que foram classificadas corretamente. Como esta é a menor classe, sua análise ajuda a estimar o desempenho do modelo para a predição da classe minoritária.

Para este modelo, o valor da medida de precisão foi de $0,58$, indicando que 58% daquelas amostras indicadas como "tem filho" são de fato desta classe. Por outro lado, a medida de revocação encontrada foi de 0.79, o que confirma que 79% dos casos em que um pesquisador é da classe "tem filho" foram corretamente encontrados. Quando calculadas para a classe oposta ("não tem filho"), essas medidas apresentam o bom desempenho do modelo para a classificação desses casos, sendo iguais à 0.98 para a precisão e 0.96 para a revocação.

Assim, podemos considerar que este modelo pode fielmente dizer se um pesquisador não

vai gerar um filho, indicando talvez a grande distinção destes pesquisadores para outros. No entanto, quando se trata das predições para a geração de um filho, as variáveis utilizadas podem não indicar corretamente características que ajudam na identificação deste tipo de pesquisador. Isso pode ser atribuído tanto a um problema do tipo de problema abordado, quanto a uma limitação das medidas topológicas utilizadas.

5.3 Características de pais acadêmicos

Utilizando-se das características das árvores de decisão, é possível analisar todo o caminho percorrido para a tomada de uma decisão baseada nos atributos de uma amostra. A Figura 21 apresenta a árvore obtida pelo treinamento descrito anteriormente. Cada quadrado representa uma decisão a ser tomada, com a condição indicada em sua primeira linha. Os valores indicados por "Amostras" representam a porcentagem das 5906297 amostras utilizadas para o treinamento do modelo. Já os valores associados à "Proporção" indicam a proporção das classes "não tem filho" e "tem filho", respectivamente. As cores também mostram esta proporção de acordo com sua intensidade, sendo a cor laranja associada à primeira classe e a cor azul à segunda classe.

Pela análise desta árvore, podemos extrair todas as regras utilizadas para classificar se um dado pesquisador terá ou não um filho. A principal característica que difere um pesquisador que tem um filho daquele que não tem é a presença ou não de outros filhos. Este atributo se encontra na raiz da árvore de decisão e divide as amostras em um grupo de pesquisadores majoritariamente da classe "não tem filho", com 88% das amostras, e outro com pesquisadores da classe "tem filho", sendo composto por 12% das amostras. O uso deste atributo sozinho já torna possível uma boa classificação da amostra. Ainda assim, existem outras características que diferem os pesquisadores, sendo uma das principais delas a idade acadêmica do pesquisador, que aparece em um total de seis regras.

O maior número de regras se concentra nos pesquisadores que não tem filhos. No entanto, apesar de ocorrerem diversas quebras, a variação na quantidade de pesquisadores de cada classe e no poder de classificação da árvore não é grande. Existe, porém, uma quebra relevante à esquerda da árvore correspondente à pesquisadores sem ascendentes, os quais tem um

Filhos	Ascendentes	Idade	Sobrinhos	Irmãos	% do total	% classe “tem filho”
>5	-	≤25	-	-	3.3	82.8
>2, ≤5	-	≤25	-	-	2.4	72.1
≤2	-	-	-	>0	2.0	58.6
>2	-	>25	-	-	2.2	58.2
≤2	-	-	-	0	2.0	47.2
0	0	>7	-	-	2.6	32.6
0	0	≤7	-	-	3.5	14.3
0	>0	>4, ≤19	>0	≤27	11.7	5.5
0	>0	>6, ≤19	0	-	12.7	3.9
0	>0	>1, ≤4	>0	-	5.7	3.8
0	>0	>4, ≤19	>0	>27	6.0	3.6
0	>0	>4, ≤6	0	-	7.3	2.6
0	>0	≤1	>0	-	3.7	2.3
0	>0	>19	-	-	4	1.6
0	>0	≤4	0	-	30.8	1.2

Tabela 3: *Regras para a decisão de amostras para a árvore de decisão.*

número maior de casos de pesquisadores que geram um filho.

Já para os pesquisadores que tem ao menos um filho, a árvore apresenta regras que melhor separam as amostras. Por exemplo, existem grupos de pesquisadores deste conjunto que contêm mais de 40% de casos da classe “não tem filho”, enquanto outros apresentam menos de 20% deste tipo. Isso evidencia a importância dessas regras para a classificação.

A Tabela 3 mostra em mais detalhes as regras utilizadas para a tomada das decisões. Existem no total 15 conjuntos de regras diferentes, representando as folhas da árvore e as possibilidades de classificação. Perceba que, apesar das regras serem compostas por apenas cinco atributos, existe uma grande variação entre os seus valores. Além disso, estas regras podem ter interpretações diversas, como no caso em que a ausência de ascendentes pode significar a ausência de um pai cadastrado na Plataforma Acácia, ou no caso em que a ausência de sobrinhos significa também a ausência de netos para o pai.

Vale notar que os conjuntos de pesquisadores com maior quantidade da classe “tem filho” também corresponde à vasta minoria das amostras. O maior grupo de todos, que também é aquele com menor quantidade da classe “tem filho”, é maior que a soma de todos os grupos compostos por mais de 10% desta classe.

As regras também corroboram os resultados encontrados na Seção 5.1, uma vez que pesquisadores muito novos ou com idade avançada têm produções menores que os outros. Também

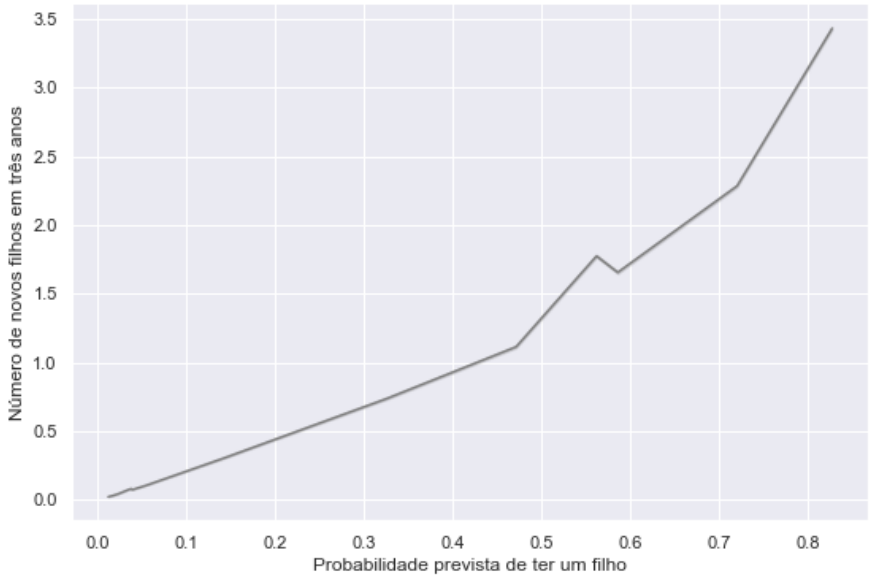

Figura 19: *Relação entre a quantidade de filhos que um pesquisador vai gerar nos próximos três anos e a probabilidade de ele ter um filho prevista pelo classificador.*

é possível notar que outro fator que influencia a geração de novos pesquisadores é a presença ou não de ascendentes.

Finalmente, os pesquisadores com maiores chances de terem um filho são aqueles com mais de cinco filhos e menos de 25 anos de idade. Estes pesquisadores podem ser interpretados como pesquisadores jovens e produtivos, que estão em pontos altos de suas carreiras. Em contraponto, aqueles que não tem filhos, têm ao menos um ascendente, têm idade menor que quatro e não tem sobrinhos apresentam a menor chance de terem um filho. Estes são pesquisadores recém-formados, cujos pais também são jovens (por não possuírem netos), o que pode indicar que sua vizinhança está em desenvolvimento.

5.4 Cenário de crescimento do grafo

Para analisar um possível cenário futuro para o grafo, as regras anteriores foram aplicadas sobre os dados do ano de 2019 para obter uma predição sobre o ano de 2022.

A partir da aplicação das regras e da obtenção de uma probabilidade de produção de um filho para cada vértice, pode-se supor quais vértices terão ou não um filho através de uma amostragem aleatória. Assim, escolhem-se vértices para gerarem um novo filho de acordo com sua probabilidade prevista. Seguindo este método, $9,1\%$ ou aproximadamente 95 mil pesquisadores geram novos filhos em 2022.

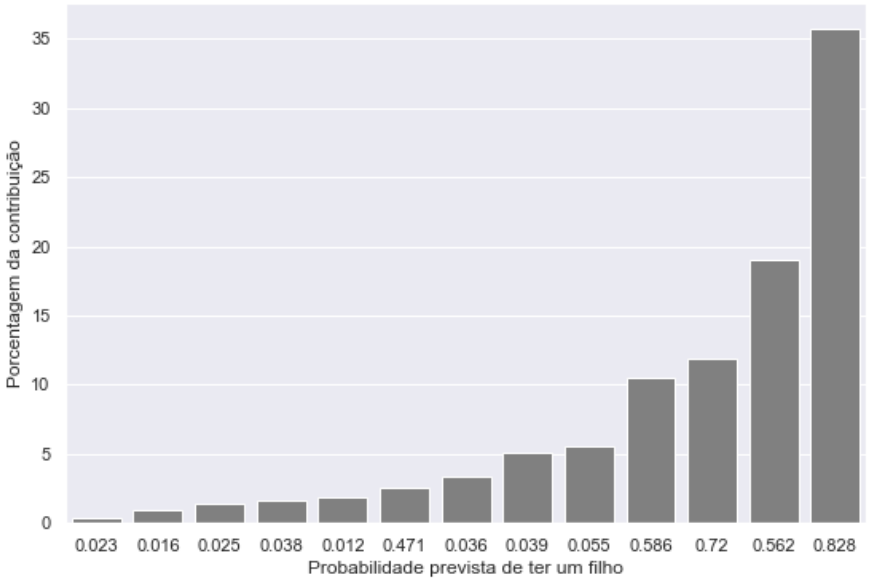

Figura 20: *Distribuição da contribuição de cada faixa de probabilidade para a formação de novos pesquisadores.*

Ainda que este método permita identificar quais pesquisadores gerarão novos filhos, o número de filhos que serão gerados não pode ser estimado utilizando o modelo, dado que utilizou-se uma técnica de aprendizado de máquina para a classificação (Seção 4.2.1). Porém, uma boa aproximação pode ser feita utilizando a informação do número filhos em três anos para cada conjunto de pesquisadores classificado pelas regras.

Calcula-se, assim, a média de filhos em um futuro distante três anos para pesquisadores que foram classificados para grupos com a mesma probabilidade de gerar um filho. Esta média pode, então, ser multiplicada pela probabilidade de ter um filho de cada pesquisador, obtendo uma previsão aproximada do número de filhos que ele terá no futuro.

A Figura 19 mostra os valores obtidos para esta relação entre a probabilidade prevista de gerar um filho e o seu número real de filhos em três anos. Como pesquisadores com maior probabilidade geram mais filhos, o número total de novos pesquisadores deve ser consideravelmente maior que o número de pesquisadores que foram classificados como geradores de filhos. É esperado, ainda, que a contribuição dos pesquisadores com maior probabilidade para o número total de pesquisadores gerados seja maior que a de pesquisadores com pouca probabilidade. Apesar de sua menor quantidade, as maiores probabilidade e médias de filhos compensam a desigualdade no número de ocorrências.

Com a aplicação do procedimento, obteve-se o resultado de que 280 mil pesquisadores

serão formados até 2022. Isso equivale a um crescimento de 27% com relação ao número de pesquisadores de 2019. A Figura 20 mostra quais foram as faixas de probabilidade que mais contribuíram para a formação de novos pesquisadores segundo este método.

Perceba que a faixa que mais contribuiu foi, de fato, a dos pesquisadores com maior probabilidade de gerarem um novo filho. No entanto, a ordem das contribuições não segue as probabilidades de forma decrescente como seria o esperado, mas apresenta inversões. Por exemplo, note que pesquisadores com 47% de chance de gerarem um novo pesquisador contribuíram menos que aqueles que tinham apenas $3,6\%$ de chance. Isto evidencia que, apesar de terem uma probabilidade menor, os pesquisadores pertencentes a estes grupo ainda são relevantes para o crescimento.

Conclui-se, assim, que o cenário da ciência modelado pelo grafo da Plataforma Acácia apresenta de fato um crescimento. Este crescimento está ligado principalmente à pesquisadores que estão no auge de suas carreiras (como discutido anteriormente) e que produzem mais filhos que todas as outras faixas. Ainda assim a geração de filhos para probabilidades ainda tem um papel importante e devem ser considerados, principalmente quando se considera o enorme volume que representam no grafo.

6 Cronograma de atividades

O cronograma seguido durante a realização do trabalho está apresentado na Tabela 4. A cada atividade foram alocados um número determinado de meses indicando o momento em que a atividade foi durante o período de um ano do projeto. Atividades como foram planejadas estão representadas por células preenchidas em cinza escuro, enquanto células cinza claro indicam o cronograma que foi efetivamente realizado. Devido ao atraso na realização das atividades 1, 2 e 3 outras atividades tiveram que ser reduzidas para que pudessem ser realizadas dentro do período de um ano indicado para o projeto.

Atividade		Quadrimestre 1				Quadrimestre 2				Quadrimestre 3			
		01	02	03	04	05	06	07	08	09	10	11	12
1. Definição do problema e análise de viabilidade	P												
	R												
2. Estudo das bases de dados e de grafos de genealogia	P												
	R												
3. Definição do método de extração de informação temporal	P												
	R												
4. Geração do conjunto de dados de treinamento	P												
	R												
5. Aplicação de técnicas de modelagem	P												
	R												
6. Análise de resultados	P												
	R												
7. Escrita de relatórios	P												
	R												

Tabela 4: *Cronograma das atividades a serem realizadas ao longo do Projeto de Graduação em Computação ao longo de um ano. Cada intervalo da tabela equivale a um período de um mês dentro do quadrimestre indicado. A letra "P" indica o planejamento das atividades, enquanto a letra "R" indica o período em que foram realizadas.*

7 Considerações finais

Existem diversas formas de se explorar a forma como o conhecimento é criado e difundido pela comunidade acadêmica. Abordagens envolvendo a análise de citações, de colaboração acadêmica por meio da co-autoria de trabalhos científicos e de genealogia acadêmica já são conhecidas e demonstram resultados promissores para a expansão do entendimento das áreas de cientometria e bibliometria.

Considerando as diversas relações existentes no meio acadêmico, o uso de grafos surge como uma boa opção para a realização de análises mais detalhadas sobre o funcionamento de problemas de disseminação de conhecimento. Este método se mostra bastante promissor para a modelagem das relações de orientação acadêmica que ocorrem entre professores e alunos de mestrado e doutorado. Ainda que muitos estudos tenham se utilizado desse método para obter resultados promissores sobre a genealogia acadêmica, são poucos os que se aproveitam da sua característica evolutiva ao longo do tempo.

Utilizando as informações temporais presentes no conjunto de dados da Plataforma Acácia, definiu-se um método para a extração de dados de crescimento deste grafo e seu posterior uso para o treinamento de modelos de aprendizado de máquina. A aplicação de tais técnicas tornou possível a descoberta de novas relações e de padrões não-óbvios que possam trouxeram um novo entendimento sobre a produção acadêmica de pesquisadores em relação à recursos humanos.

Os resultados obtidos neste projeto indicam padrões de crescimento do grafo de genealogia acadêmica da Plataforma Acácia. É notável, principalmente, o fenômeno da redução do ritmo da produção de novos pesquisadores ao longo dos anos, causado majoritariamente pelo número crescente de pesquisadores que não tiveram filhos acadêmicos. Isso pode ser consequência do atraso na orientação de novos pesquisadores, dado que a idade acadêmica de maior produção é em torno de 25 anos e a maior parte dos pesquisadores sem filhos foi formada após os anos 2000.

O modelo utilizado para a previsão do crescimento apresenta bons resultados sendo capaz de classificar 94% das amostras corretamente. Ainda que nem todos os pesquisadores geradores de filhos tenham sido identificados de forma correta, a grande maioria daqueles que não

geram foram classificados corretamente.

Acima de tudo, as regras obtidas pelo classificador são capazes de separar pesquisadores de forma convincente. Em especial, a análise destas regras mostrou que pesquisadores sem nenhum filho costumam não gerar nenhum novo pesquisador e correspondem à grande maioria dos casos (88%). Já a maior probabilidade de geração está relacionada aos pesquisadores com menos de 25 anos de idade acadêmica e que já tem pelo menos um filho, apesar de estes corresponderem a um número muito pequeno de pesquisadores (aproximadamente 2%).

Por fim, a aplicação das regras para os dados obtidos no ano de 2019 mostrou que o cenário da ciência brasileira está evoluindo e é esperado que cresça até 27% em três anos de acordo com o modelo proposto, sendo equivalente a 280 mil novos pesquisadores. Estes novos pesquisadores serão formados por aproximadamente 95 mil pesquisadores. Boa parte da contribuição pela formação de novos pesquisadores está relacionada aos pesquisadores com maior probabilidade de gerar um vértice, apesar de os pesquisadores com baixa probabilidade ainda terem relevância expressiva devido a sua grande quantidade.

A partir de tudo o que foi discutido, este projeto trouxe um novo olhar sobre grafos de genealogia acadêmica. As análises realizadas sobre a comunidade acadêmica podem ajudar a entender fenômenos que aconteceram no passado, assim como o impacto de novas alterações ao ecossistema do ensino superior realizadas em um momento contemporâneo. Este estudo também pode ser útil para o gerenciamento deste crescimento, ao levar à uma melhor utilização de recursos e à criação de políticas de incentivo à formação acadêmica onde necessário.

É importante ressaltar, no entanto, que o método utilizado é agnóstico à base de dados e pode ser aplicado em diversos tipos de grafo que apresentam a mesma estrutura hierárquica e evolutiva. Alguns exemplos de uso destas técnicas são para (i) a previsão de crescimento populacional, (ii) a análise de comportamento de conteúdo viral em meios digitais, e (iii) crescimento de redes de indicação de serviços. Desta forma, este método tem o potencial de realizar previsões confiáveis sobre o crescimento de um grafo, permitindo obter um bom indicativo de como será um possível estado futuro a partir de sua história passada.

7.1 Limitações e pesquisas futuras

Apesar dos métodos obtidos por este trabalho apresentarem bons resultados e colaborarem para a ampliação do entendimento sobre como grafos de genealogia acadêmica crescem e evoluem com o passar do tempo, existem diversas abordagens não exploradas que se mostram promissoras para a continuidade desta linha de pesquisa. A seguir estão expostas algumas possibilidades de estudos a serem realizados para melhorar os resultados obtidos neste trabalho e explorar outros aspectos do problema aqui apresentado.

1. A remoção de arestas secundárias (Seção 4.1.1), apesar de tornar o processamento do grafo mais simples e facilitar o andamento de outras etapas estipuladas no método desenvolvido, remove informações importantes com relação à interação entre os pesquisadores e seus alunos. Acima de tudo, esta modificação na estrutura do grafo altera as métricas calculadas durante a etapa de extração de informações temporais descrita na Seção 4.1.2. O uso destas arestas pode, portanto, alterar significativamente o resultado final dos estudos realizados, apresentando um ganho considerável no desempenho das predições ao modelar com maior fidelidade os fenômenos do mundo real.

2. Neste trabalho considerou-se que todos os tipos de orientações são idênticos. Para melhor modelar a interação entre pesquisadores as arestas poderiam ser pesadas de acordo com sua duração, seu número ou o tipo de orientação que representam. Esta abordagem se torna ainda mais interessante para o caso onde arestas secundárias não são removidas, dado que casos onde pesquisadores apresentam mais de um tipo de aresta incidente de diferentes tipos passam a ocorrer.

3. As variáveis utilizadas para realizar predições foram puramente baseadas em medidas topológicas das vizinhanças dos vértices do grafo. Isso torna o método agnóstico à base de dados, mas também desperdiça muitas informações que poderiam ser úteis para facilitar a realização de previsões mais corretas. No entanto, utilizar informações que não podem ser retiradas diretamente do grafo (tais como a área de atuação do pesquisador e a instituição ao qual está vinculado) é uma tarefa árdua, dado que previsões sobre esses valores para novos vértices que surgem devem também ser feitas.

4. Ainda que algoritmo de aprendizado de máquina utilizado para aprender as informações

do conjunto de dados e realizar previsões tenha permitido um fácil entendimento das características que levam ao crescimento do grafo e a extração de regras simples para a caracterização e classificação de professores, esse apresenta um desempenho inferior à outros métodos já amplamente utilizados na literatura (Caruana & Niculescu-Mizil, 2006). Assim, o uso de um algoritmo que não ofereça tal entendimento, mas que tenha um desempenho melhor, pode ser um passo interessante para a melhoria das previsões e a obtenção de uma visão mais correta do futuro da comunidade acadêmica representada pelo grafo de genealogia.

5. A tarefa de aprendizado de máquina proposta foi a de aprendizado supervisionado para classificação. Inicialmente, esta pesquisa focou em desenvolver um método para saber se um pesquisador gerará ao menos um filho no futuro. Uma melhoria em relação a esta tarefa seria o uso de um algoritmo de aprendizado de máquina para regressão, de modo a realizar previsões com relação ao número de novos filhos de um pesquisador em um momento futuro. Ainda que o conjunto de dados utilizado para realizar este tipo de previsão seja o mesmo que o utilizado para a classificação com a diferença de que a variável e saída é contínua, isto demanda um esforço e cuidado adicional de análise e calibração dos resultados, apesar de possibilitarem a extração de melhores resultados.

6. A variável de saída utilizada para a predição e o treinamento do modelo foi explorada para apenas um valor no futuro. Existem diversos outros valores que poderiam ser utilizados, podendo trazer resultados distintos e até melhores que os aqui apresentados. O valor de três anos foi escolhido de forma arbitrária, mais medidas poderiam ter sido utilizadas para realizar uma decisão mais abrangente.Por exemplo, o tempo médio de formação de um pesquisador de mestrado ou de doutorado poderia ter sido utilizado, já que é relevante saber se existirão novos pesquisadores deste tipo no futuro.

7. A visão do futuro obtida pelo método apresentado está limitada pelo tempo máximo das previsões do algoritmo de aprendizado de máquina utilizado. Neste trabalho, o algoritmo foi treinado para identificar pesquisadores que gerariam um novo filho acadêmico em até dez anos a partir da última data de coleta dos dados. Esta data acrescida de dez anos caracteriza, portanto, o momento futuro máximo do qual temos alguma

visão. Para explorar datas futuras mais distantes, seguindo os limites do conjunto de dados utilizado, um método para iterar sobre as previsões do modelo, de forma a simular o crescimento do grafo pela adição dos vértices previstos poderia ser utilizado. Desta maneira, a evolução do grafo poderia ser modelada para o futuro de forma indefinida, proporcionando uma visão de momentos futuros arbitrariamente longínquos. Ressalta-se, no entanto, que cada iteração poderia adicionar uma porcentagem de erros ao crescimento do grafo, o que será propagada ao longo do tempo, tornando as previsões cada vez mais imprecisas.

Referências Bibliográficas

Aggarwal, Charu, & Subbian, Karthik. 2014. Evolutionary network analysis: A survey. *ACM Computing Surveys (CSUR)*, **47**(1), 10.

Alsabti, Khaled, Ranka, Sanjay, & Singh, Vineet. 1997. An efficient k-means clustering algorithm.

Basu, Sudipta, & Waymire, Gregory B. 2006. Recordkeeping and human evolution. *Accounting Horizons*, **20**(3), 201–229.

Batagelj, Vladimir. 2003. Efficient algorithms for citation network analysis. *arXiv preprint cs/0309023*.

Berlingerio, Michele, Bonchi, Francesco, Bringmann, Björn, & Gionis, Aristides. 2009. Mining graph evolution rules. *Pages 115–130 of: joint European conference on machine learning and knowledge discovery in databases*. Springer.

Bishop, Chris, Bishop, Christopher M, *et al.* . 1995. *Neural networks for pattern recognition*. Oxford university press.

Boaventura, Michel, Boson, Karina, da Silva, Ana Paula Couto, Veloso, Adriano, & Junior, Wagner Meira. 2014. Caracterização temporal das redes de colaboração científica nas universidades brasileiras: anos 2000-2013. *Pages 9–20 of: Anais do III Brazilian Workshop on Social Network Analysis and Mining*. SBC.

Bondy, John Adrian, Murty, Uppaluri Siva Ramachandra, *et al.* . 1976. *Graph theory with applications*. Vol. 290. Macmillan London.

Breunig, Markus M, Kriegel, Hans-Peter, Ng, Raymond T, & Sander, Jörg. 2000. LOF: identifying density-based local outliers. *Pages 93–104 of: ACM sigmod record*, vol. 29. ACM.

Butler, Pierce. 2011. *An introduction to library science*. Read Books Ltd.

Caruana, Rich, & Niculescu-Mizil, Alexandru. 2006. An empirical comparison of supervised learning algorithms. *Pages 161–168 of: Proceedings of the 23rd international conference on Machine learning*. ACM.

Chen, Chenyi, Seff, Ari, Kornhauser, Alain, & Xiao, Jianxiong. 2015. Deepdriving: Learning affordance for direct perception in autonomous driving. *Pages 2722–2730 of: Proceedings of the IEEE International Conference on Computer Vision*.

Collobert, Ronan, & Weston, Jason. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. *Pages 160–167 of: Proceedings of the 25th international conference on Machine learning*. ACM.

Damaceno, R., Rossi, L., & Mena-Chalco, J. P. 2017. Identificação do grafo de genealogia acadêmica de pesquisadores: Uma abordagem baseada na Plataforma Lattes. *Pages 76–87 of: SBBD*.

Damaceno, Rafael JP, Rossi, Luciano, Mugnaini, Rogério, & Mena-Chalco, Jesús P. 2019. The Brazilian academic genealogy: evidence of advisor–advisee relationships through quantitative analysis. *Scientometrics*, **119**(1), 303–333.

David, Stephen V, & Hayden, Benjamin Y. 2012. Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. *PloS one*, **7**(10), e46608.

Doreian, Patrick, & Stokman, Frans N. 2013. The dynamics and evolution of social networks. *Pages 9–26 of: Evolution of social networks*. Routledge.

El Naqa, Issam, & Murphy, Martin J. 2015. What is machine learning? *Pages 3–11 of: Machine Learning in Radiation Oncology*. Springer.

Franklin, James. 2005. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, **27**(2), 83–85.

Gargiulo, F., Caen, A., Lambiotte, R., & Carletti, T. 2016. The classical origin of modern mathematics. *EPJ Data Science*.

Grácio, Maria Cláudia Cabrini, & Oliveira, Ely Francina Tannuri de. 2014. Estudos de Análise de Cocitação de Autores: uma abordagem teórico-metodológica para a compreensão de um domínio. *Tendencias da Pesquisa brasileira em Ciência da Informação*, 1–22.

Henriksen, Dorte. 2016. The rise in co-authorship in the social sciences (1980–2013). *Scientometrics*, **107**(2), 455–476.

Hicks, Diana, Wouters, Paul, Waltman, Ludo, De Rijcke, Sarah, & Rafols, Ismael. 2015. Bibliometrics: the Leiden Manifesto for research metrics. *Nature News*, **520**(7548), 429.

Hosmer Jr, David W, Lemeshow, Stanley, & Sturdivant, Rodney X. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.

Hummon, Norman P, & Dereian, Patrick. 1989. Connectivity in a citation network: The development of DNA theory. *Social networks*, **11**(1), 39–63.

Kogan, Maurice. 2000. Higher education communities and academic identity. *Higher Education Quarterly*, **54**(3), 207–216.

Kononenko, Igor. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, **23**(1), 89–109.

Kostakos, Vassilis. 2009. Temporal graphs. *Physica A: Statistical Mechanics and its Applications*, **388**(6), 1007–1023.

Leydesdorff, Loet, & Milojević, Staša. 2012. Scientometrics. *arXiv preprint arXiv:1208.4566*.

Lü, Linyuan, & Zhou, Tao. 2011. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, **390**(6), 1150–1170.

Marrou, Henri Irénée, & Marrou, Henri Irénée. 1982. *A history of education in antiquity*. Univ of Wisconsin Press.

Mena-Chalco, Jesús P, Digiampietri, Luciano A, & Cesar-Jr, Roberto M. 2012. Caracterizando as redes de coautoria de currículos Lattes. *Pages 1–12 of: Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Mena-Chalco, Jesús Pascual, & Junior, Cesar. 2013. Prospecção de dados acadêmicos de currículos Lattes através de Scriptlattes. *Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João*.

Mitchell, Tom M, *et al.* . 1997. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, **45**(37), 870–877.

Moed, Henk F. 2006. *Citation analysis in research evaluation*. Vol. 9. Springer Science & Business Media.

Newman, Mark EJ. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, **101**(suppl 1), 5200–5205.

Papadimitriou, Panagiotis, Dasdan, Ali, & Garcia-Molina, Hector. 2010. Web graph similarity for anomaly detection. *Journal of Internet Services and Applications*, **1**(1), 19–30.

Priem, Jason, & Hemminger, Bradely H. 2010. Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First monday*, **15**(7).

Rish, Irina, *et al.* . 2001. An empirical study of the naive Bayes classifier. *Pages 41–46 of: IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3.

Rossi, Luciano, Damaceno, Rafael JP, Freire, Igor L, Bechara, Etelvino JH, & Mena-Chalco, Jesús P. 2018. Topological metrics in academic genealogy graphs. *Journal of Informetrics*, **12**(4), 1042–1058.

Safavian, S Rasoul, & Landgrebe, David. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, **21**(3), 660–674.

Sengupta, IN. 1992. Bibliometrics, informetrics, scientometrics and librametrics: an overview. *Libri*, **42**(2), 75–98.

Sugimoto, C. R. 2014. Academic Genealogy. *Pages 365–382 of:* Cronin, B, & Sugimoto, C R (eds), *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, first edn. MIT Press.

Sugimoto, Cassidy R, Ni, Chaoqun, Russell, Terrell G, & Bychowski, Brenna. 2011. Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in Library and Information Science. *Journal of the American Society for Information Science and Technology*, **62**(9), 1808–1828.

Suykens, Johan AK, & Vandewalle, Joos. 1999. Least squares support vector machine classifiers. *Neural processing letters*, **9**(3), 293–300.

Tenn, J. S. 2016. Introducing AstroGen: the Astronomy Genealogy Project. *arXiv preprint arXiv:1612.08908*.

Thelwall, Mike. 2008. Bibliometrics to webometrics. *Journal of information science*, **34**(4), 605–621.

Vázquez, Alexei, Flammini, Alessandro, Maritan, Amos, & Vespignani, Alessandro. 2003. Modeling of protein interaction networks. *Complexus*, **1**(1), 38–44.

Vinkler, Peter. 1993. Research contribution, authorship and team cooperativeness. *Scientometrics*, **26**(1), 213–230.

Watkins, Christopher JCH, & Dayan, Peter. 1992. Q-learning. *Machine learning*, **8**(3-4), 279–292.

Witten, Ian H, Frank, Eibe, Hall, Mark A, & Pal, Christopher J. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yan, Erjia, & Ding, Ying. 2009. Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, **60**(10), 2107–2118.

A Árvore de decisão

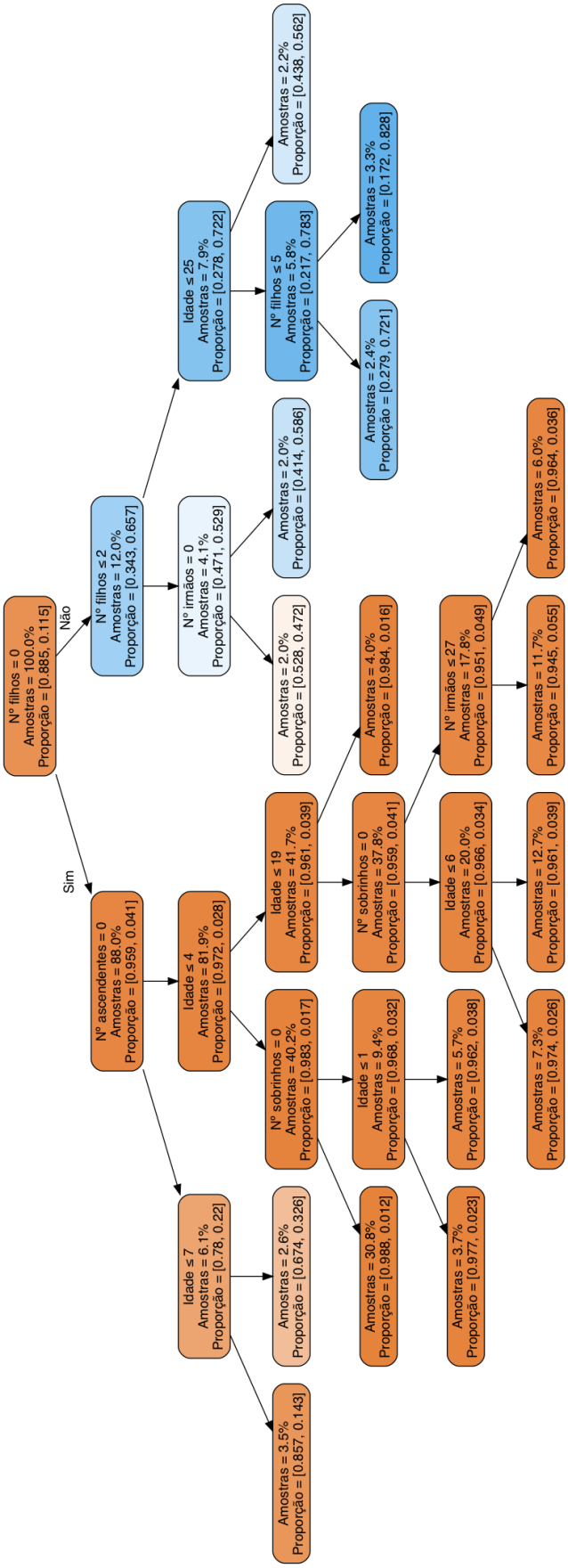

Figura 21: *Árvore de decisão para a predição da geração de novos pesquisadores para o grafo de genealogia da Plataforma Acácia.*