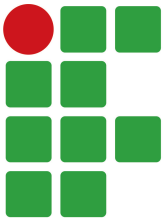


INSTITUTO FEDERAL
Goiás
Câmpus Anápolis

Flauberth Duarte de Maria Santos

Predição de Informações Acadêmicas em Dados de Séries Temporais

Anápolis-GO

2021

Flauberth Duarte de Maria Santos

Predição de Informações Acadêmicas em Dados de Séries Temporais

Projeto de Pesquisa apresentado ao curso de Bacharelado em Ciências da Computação, como requisito para obtenção do grau final na disciplina de Projeto Final de Curso.

Instituto Federal de Goiás - Câmpus Anápolis

Orientador: Prof. Dr. Sérgio Daniel Carvalho Canuto

Anápolis-GO

2021

FICHA CATALOGRÁFICA

R696e	SANTOS, Flauberth Duarte de Maria Predição de informações acadêmicas em dados de séries temporais. / Flauberth Duarte de Maria Santos – – Anápolis: IFG, 2021. 46 p. : il. color. Orientador: Prof. Dr. Sérgio Daniel Carvalho Canuto. Trabalho de Conclusão do Curso de Ciência da Computação, Instituto Federal de Goiás, Campus Anápolis, 2021. 1. Inteligência artificial. 2. *Machine learning*. 3. *Time series*. 4. Computação. I. CANUTO, Sérgio Daniel Carvalho orien.. II. Título.
	CDD 005.1

Ficha catalográfica elaborada pelo Bibliotecário Matheus Rocha Piacenti CRB1/2992
Biblioteca Clarice Lispector, Campus Anápolis
Instituto Federal de Educação, Ciência e Tecnologia de Goiás

INSTITUTO FEDERAL
Goiás

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
SISTEMA INTEGRADO DE BIBLIOTECAS

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

[ ] Tese	[ ] Artigo Científico
[ ] Dissertação	[ ] Capítulo de Livro
[ ] Monografia – Especialização	[ ] Livro
[X] TCC - Graduação	[ ] Trabalho Apresentado em Evento

[ ] Produto Técnico e Educacional - Tipo: _____

Nome Completo do Autor:
Matrícula:
Título do Trabalho:

Autorização - Marque uma das opções

1. (X) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. ( ) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data ___/___/_____ (Embargo);
3. ( ) Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

( ) O documento está sujeito a registro de patente.
( ) O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
( ) Outra justificativa: _____

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

i. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
ii. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
iii. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Anapolis, 21/03/22.
Local Data

Flavietta Duarte da Maia Santos
Assinatura do Autor e/ou Detentor dos Direitos Autorais

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
CÂMPUS ANÁPOLIS

ATA DA SESSÃO PÚBLICA DE APRESENTAÇÃO E DEFESA DO TRABALHO DE CONCLUSÃO DE CURSO

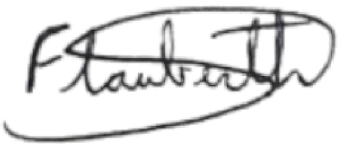
Aos **sete** dias do mês de **fevereiro** de **2022**, às **08h00**, em sessão pública por meio de webconferência via Google Meet, foi realizada a sessão pública de apresentação e qualificação do Trabalho de Conclusão de Curso do Graduando **Flauberth Duarte de Maria Santos (matrícula 20181060140188)** do curso de Bacharelado em Ciência da Computação. A banca foi composta pelos seguintes membros: **Dr. Sérgio Daniel Carvalho Canuto, Dr. Daniel Xavier Sousa e Dr. Raphael de Aquino Gomes**, sob a presidência do primeiro. O trabalho de conclusão de curso tem como título **Predição de Informações Acadêmicas em Dados de Séries Temporais**, sob orientação de **Sérgio Daniel Carvalho Canuto**. Após a apresentação, tendo sido o autor arguido pela Banca Examinadora, a nota obtida foi **9,3** e o Trabalho de Conclusão de Curso foi considerado **APROVADO** pela banca examinadora.

Encerra-se a presente sessão às **09** horas e **00** minutos. Eu, **Sérgio Daniel Carvalho Canuto**, dato e assino a presente ata que segue assinada por todos os membros da Banca e pelo graduando.

Dr. Sérgio Daniel Carvalho Canuto
(Assinado Eletronicamente)

Dr. Daniel Xavier Sousa
(Assinado Eletronicamente)

Dr. Raphael de Aquino Gomes
(Assinado Eletronicamente)



Flauberth Duarte de Maria Santos
(Graduando)

Documento assinado eletronicamente por:

- **Raphael de Aquino Gomes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 07/03/2022 11:19:30.
- **Daniel Xavier de Sousa**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 07/03/2022 09:15:01.
- **Sergio Daniel Carvalho Canuto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 07/03/2022 07:53:20.

Este documento foi emitido pelo SUAP em 07/02/2022. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifg.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 243292
Código de Autenticação: 9fd6d0280b

Lista de ilustrações

Lista de tabelas

Sumário

Resumo

Produções acadêmicas são um dos elementos essenciais na carreira de pesquisadores. De fato, o número de publicações recentes e de qualidade associado a pesquisadores ou grupos de pesquisa pode ser utilizado como critério para avaliação de projetos de pesquisa, obtenção de recursos financeiros para pesquisa, bem como o potencial de inovação vinculado ao pesquisador, grupo de pesquisa e cursos de pós-graduação. Dada a importância da produtividade acadêmica como elemento capaz de auxiliar na tomada de decisões administrativas, faz-se necessária uma análise de fatores associados à predição de publicações futuras por parte dos pesquisadores.

Este trabalho visa investigar fatores que podem estar relacionados à predição de novas publicações de qualidade, bem como estratégias para predição de publicações acadêmicas. Particularmente, este trabalho se concentra no tratamento da informação contida em diversos dados associados aos pesquisadores do Instituto Federal de Goiás, considerando fatores como produções acadêmicas, perfil do pesquisador, publicações anteriores, atuações no ensino e em grupos de pesquisa. A partir de tais informações, foi analisada a evolução temporal da produção acadêmica dos pesquisadores.

Como resultados, os fatores como publicações anteriores, atividades vinculadas à pesquisa e disciplinas ministradas se apresentaram como mais importantes para explicar a variação da evolução das pesquisas. A predição de publicações com a ferramenta Random Forests se mostrou mais efetiva que a regressão de Poisson, previamente empregada para essa tarefa, e até 10x mais eficiente, em tempo de execução, que Redes Neurais Recorrentes (RNN e LSTM).

Palavras-chave: Time Series, Machine Learning, Inteligência Artificial

Abstract

Academic publications are one of the essential elements in the career of researchers. In fact, the quality and quantity of publications of researchers and research groups are a criterion to evaluate research projects by funding agencies, as well as the innovation potential of researchers, research groups and postgraduate courses. Given the importance of academic productivity for administrative decision-making, it is important to analyse the factors associated with the prediction of future publications.

This work aims to investigate factors that may be related to the prediction of new quality publications, as well as strategies for predicting academic publications. Particularly, this work focuses on the treatment of information of researchers at Federal Institute of Goiás, considering factors such as academic production, researcher profile, previous publications, activities in teaching and in academic production groups. From such information, there was a temporal evolution of the production of each researcher.

As a result, factors such as previous publications, activities related to research and subjects taught proved to be the most important to explain the evolution of research. The prediction of academic productivity using Random Forests is more effective than the previously employed Poisson Regression. Random Forests is also 10 times more efficient than Recurring Neural Networks (RNN and LSTM).

Introdução

O número de publicações afeta diversos aspectos da carreira de um pesquisador. De fato, saber o quanto um pesquisador ou um grupo de pesquisa pode produzir e quais fatores estão associados à produção acadêmica é uma questão que atrai não somente a curiosidade em instituições de ensino, mas também pode ser utilizada para avaliação de projetos de pesquisa, obtenção de recursos financeiros para pesquisa, bem como o potencial de inovação vinculado ao pesquisador ou grupo de pesquisa, muito útil para auxiliar na tomada de decisões administrativas. Para prever automaticamente o potencial de contribuições acadêmicas do pesquisador em anos futuros, faz-se necessária a exploração de diversos possíveis fatores que podem aumentar ou diminuir o potencial de contribuição.

Os dados presentes no currículo do pesquisador podem fornecer uma série de informações que podem auxiliar na predição de produções acadêmicas futuras. Dentre elas, a quantidade de produções bibliográficas do ano corrente e dos anos anteriores e a nota Qualis associada aos veículos das publicações dos pesquisador também oferecem indícios de que no futuro o pesquisador poderá manter a quantidade/qualidade semelhante de publicações para o próximo ano. Entretanto, tentar prever futuras contribuições acadêmicas com base apenas nas produções pode não ser o suficiente para uma boa predição, dada a complexidade e a quantidade de outros fatores que podem estar associadas às atividades de pesquisa. A complexidade dessa predição se deve à grande diversidade de perfis de pesquisador nas condições de pesquisa oferecidas em diferentes períodos de tempo (WAY et al., 2017; XIE, 2020).

A utilização de poucas informações sobre o pesquisador pode não ser suficiente para a compreensão dos diversos perfis que existem. Portanto, este trabalho visa investigar a informação contida em diversos dados associados ao pesquisador, explorando não somente se as produções anteriores foram publicadas em revistas de prestígio ou anais de conferência, mas também outras condições associadas, como a quantidade e tipos de disciplinas ministradas pelo pesquisador docente em um determinado período, outras produções técnicas, participação de grupos de pesquisa, ambiente/estrutura acadêmica, projetos em andamento, idade e outros fatores que podem estar relacionados à predição de novas publicações de qualidade.

A obtenção e tratamento do conjunto de informações acadêmicas torna o problema ainda mais complexo pois esse estudo requer um conjunto de dados de alta qualidade com um longo intervalo de tempo. Nesse sentido, a utilização da base de dados fornecida pelo IFG Produz[1] que é a plataforma de dados de pesquisas realizadas no IFG que permite

[1] <https://ifgproduz.ifg.edu.br>

fazer buscas em informações retiradas de bancos de dados nacionais como o currículo lattes e grupos de pesquisa cadastrados no CNPq, fornece um conjunto de dados rico em informações que possui o potencial de contribuir para seleção e coleta de dados, por disponibilizar essas informações a plataforma IFG Produz tem grande importância para a realização deste trabalho. Tais informações podem ser divididas em subconjuntos de dados contendo as informações de cada pesquisador a cada ano, e enriquecidas com outras fontes de dados. Além disso, este trabalho apresenta o pré-processamento, limpeza, engenharia de atributos a partir dos dados e tratamento do desbalanceamento entre as classes que torna o problema mais difícil onde há mais registros de uma classe do que de outra, existe o desafio de diminuir o desbalanceamento entre os dados para viabilizar as predições, a principal metodologia para tratamento de desbalanceamento adotada neste trabalho foi o nosso método de sub-amostragem de dados proposto para remoção de exemplos pertencentes à classe majoritária onde se adiciona ao conjunto de treinamento os registros mais antigos da classe minoritária, bem como a posterior aplicação de estratégias de aprendizado de máquina visando a geração de modelos de predição capazes de considerar uma ampla gama de informações que podem ser obtidas sobre os pesquisadores do Instituto Federal de Goiás. A utilização desse conjunto de informações tem o potencial para aprimorar resultados anteriores na área de predição de produções acadêmicas, haja vista que mais informações sobre o perfil do pesquisador podem ser usadas na modelagem da grande diversidade de perfis de pesquisadores.

Embora toda a base de dados do trabalho e as etapas de tratamento dos dados foram aplicadas em cima dos dados do IFG por intermédio dos dados providos pela plataforma IFG Produz e o Portal de Dados Abertos, este trabalho não se limita apenas a esse caso e sua metodologia pode ser aplicada a outras instituições desde que tenha acesso às fontes de dados descritos na Seção 4.1 para o contexto da instituição onde o trabalho será aplicado.

1 Justificativa

Sem dúvidas, a predição de possíveis contribuições acadêmicas de pesquisadores com artigos publicados em revistas de qualidade é interessante para as instituições e para os grupos de pesquisa. O conhecimento prévio sobre o potencial de contribuições acadêmicas pode auxiliar o gestor na avaliação de grupos de pesquisa, destacar o potencial de produtividade dos pesquisadores no âmbito acadêmico e auxiliar na avaliação dos fatores envolvidos na produtividade.

Embora não seja uma tarefa simples ou fácil, dada a diversidade de perfis de pesquisador, este trabalho visa auxiliar nas previsões de publicações de artigos dos pesquisadores do IFG e seus grupos de pesquisa. Para isso, uma das contribuições deste trabalho consiste na metodologia para a predição de informações a partir do tratamento da evolução de diversos dados acadêmicos em diferentes contextos/janelas temporais.

Em suma, considerando a base de dados de pesquisadores do IFG, a importância desse trabalho se justifica pela busca de respostas às seguintes questões de pesquisa:

- Qual o impacto de fatores como atividades vinculadas à pesquisa, perfil do pesquisador, publicações anteriores e atividades de ensino na predição de publicações futuras?
- Qual o impacto do tratamento do desbalanceamento de dados no contexto de predição de publicações dentre pesquisadores do IFG?
- Qual a eficiência e eficácia na tarefa de predição de publicações de pesquisadores em diferentes métodos de inteligência artificial, considerando os dados de pesquisadores do IFG e estratégias previamente utilizados na literatura?

2 Objetivos

2.1 Objetivo Geral

O principal objetivo deste trabalho consiste na avaliação de fatores e estratégias para predição da evolução temporal da produção acadêmica dos pesquisadores do IFG.

2.2 Objetivos Específicos

i) Aplicar modelos baseados em séries temporais para predição da existência de novos artigos acadêmicos dos pesquisadores do IFG.

ii) Propor conjuntos atributos a partir das informações contidas nos dados explorados, a fim de viabilizar a utilização de técnicas de predição com informações temporais.

iii) Avaliar, de forma quantitativa, fatores que impactam na produção acadêmica dos pesquisadores.

iv) Comparar a estratégia proposta com outras estratégias da literatura.

3 Referencial Teórico

Há três aspectos principais relacionados à tarefa de predição automática das contribuições acadêmicas de pesquisadores tratadas nesse trabalho: os dados em séries temporais (CHATFIELD, 2004; EDELWEISS, 2001), as etapas da estratégia de mineração de dados com predição automática a partir desses dados (CASTRO; FERRARI, 2017; BURMAN; SOM, 2019; JAYAPRAKASH; KRISHNAN; JAIGANESH, 2020) e os trabalhos relacionados à aplicação de métodos de predição em janelas temporais em dados de pesquisadores (XIE, 2020; LAURANCE et al., 2013; WAY et al., 2017).

Neste capítulo são apresentados os principais conceitos e técnicas aplicadas neste trabalho. Na Seção 3.1 são apresentados conceitos de dados em janelas temporais. Na Seção 3.2 é dada a definição de Mineração de Dados e suas etapas, considerando as técnicas de coleta, pré-processamento, classificação e avaliação utilizadas, sendo esses os principais focos de interesse desse trabalho. De forma mais específica, a predição em janelas temporais é descrita na Seção 3.4. Finalmente na Seção 3.4 são apresentados os trabalhos relacionados.

3.1 Dados em Séries Temporais

De forma geral, o tratamento convencional de dados em bases de dados é feito a partir do armazenamento apenas das informações no estado presente dos dados, muitas vezes ignorando a informação temporal (SERRA; ZáRATE, 2015). Por exemplo, para uma dada entidade “autor”, e um atributo “número de publicações”, o atributo em questão em um sistema de gerenciamento de banco de dados armazena o estado atual da quantidade de publicações desse autor. Para o tratamento de informações temporais, é necessário considerar a dimensão temporal para cada atributo relacionado a cada entidade. Dessa forma, para se armazenar informações temporais, é necessário um tratamento e armazenamento de atributos adicionais que monitoram a questão temporal de interesse para cada entidade.

Segundo (EDELWEISS, 2001), o tratamento de dados temporais depende do conceito de tipo de dados temporais a serem considerados, podendo se enquadrar em três aspectos, nomeados como “instante”, “intervalo (ou janela)” e “duração”:

- Instante: o momento em que ocorre um determinado evento. Pode ser contínuo (um ponto no tempo tem duração infinitesimal), ou discreto (para dois pontos consecutivos, não existe um ponto intermediário).

- Intervalo/janela temporal: eventos que ocorrem entre dois instantes. Pode ser visto como um subconjunto de pontos no eixo temporal, representado pelos dois instantes que os delimitam.
- Duração: tempo decorrido entre instantes, podendo ser fixa ou variável. Uma duração fixa independe do contexto (ex.: 60 minutos). Uma duração variável, por outro lado, depende do contexto (ex.: 1 mês pode ser 28, 29, 30 ou 31 dias).

A definição a priori do tipo de dados temporais é fundamental para o tratamento de qualquer dado temporal (EDELWEISS, 2001). De fato, modelar problemas que envolvem séries temporais necessita da definição dos dados e de suas características ordinais, haja vista que uma série temporal é um conjunto de observações feitas em sequência ao longo do tempo (CHATFIELD, 2004). Nesse sentido, além de observar que os dados devem estar ordenados em relação ao tempo, as definições de instante, janela temporal (intervalo) e duração são fundamentais. Por exemplo, neste trabalho, foram considerados instantes discretos (finais de ano) com duração de um ano entre eles. Da mesma forma, dados na janela temporal (intervalo) de 5 anos podem ser utilizados para análise de tendências sobre a evolução dos dados a partir de técnicas de aprendizado de máquina.

Dados em séries temporais podem ser tratados para descrição de dados, explicação de dados e predição de informações temporais (CHATFIELD, 2004). A descrição é normalmente o primeiro passo para análise de dados, que podem ser simplesmente plotados em relação ao tempo para obtenção de uma simples estratégia descritiva. Tal estratégia pode ser usada para verificação de efeitos sazonais, *outliers* e presença de tendências na evolução dos dados, sendo indicada como um primeiro passo metodológico.

A utilização das séries temporais também pode ser direcionada para explicação de variáveis, em especial, para a comparação da evolução de variáveis. Usualmente, estratégias de correlação entre variáveis (ex.: Person, Spearman (MADSEN, 2007)) e modelos de regressão podem ser usadas como estratégias capazes de explicar variações dos dados. Apesar de estratégias simples de correlação e regressão linear serem tradicionalmente empregadas na explicação do comportamento temporal, o emprego de estratégias estatísticas mais sofisticadas pode ser necessário para o tratamento de padrões complexos de evolução temporal (MADSEN, 2007).

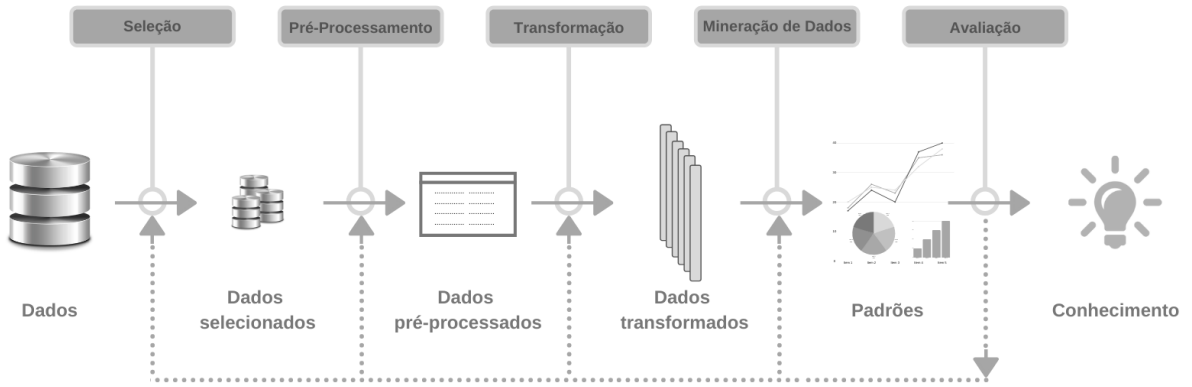
A predição em séries temporais consiste em utilizar uma dada sequência de dados observada para predizer valores futuros da série. A predição com dados temporais é um dos problemas importantes, com diversas aplicações, como por exemplo, para tarefas de predição da evolução do preço de ações, predição de vendas, clima e diversas outras tarefas. Em todas elas, a identificação e tratamento de atributos informativos/relevantes para predição é um passo necessário para aplicação de métodos de aprendizado de máquina (G. BEN TAIEB S., 2013).

O tratamento de dados, especialmente para técnicas de predição em séries temporais, exige diversas etapas, como obtenção, limpeza, pré-processamento, engenharia de atributos e emprego de técnicas de classificação ou regressão (CHATFIELD, 2004). As etapas de mineração de dados, descritas na Seção 3.2, portanto, são indispensáveis para a aplicação prática dos dados com o componente temporal.

3.2 Mineração de Dados

As etapas da descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases* – KDD) são mecanismos necessários para sistematização do tratamento do conhecimento em dados brutos coletados de uma fonte de dados. Tais mecanismos podem ser sintetizadas, segundo (CASTRO; FERRARI, 2017) em quatro etapas, como simplificado na Figura 1. A primeira consiste na obtenção da base de dados, onde os dados são obtidos no seu nível mais básico. No contexto do presente trabalho, a primeira etapa consiste na obtenção dos dados brutos a partir do currículos dos pesquisadores, grupos de pesquisa e outras atividades relacionadas à docência sendo desenvolvidas pelos pesquisadores.

Figura 1 – Processo KDD.



Fonte: adaptado de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

A segunda etapa consiste no pré-processamento e transformação dos dados, onde é feita a engenharia de atributos relevantes à tarefa sendo desenvolvida, a limpeza de ruídos e dados inconsistentes, a integração de bases de dados e a transformação/consolidação dos dados obtidos de forma a torná-los prontos para estratégia de predição a ser usada. Nesse trabalho, esta etapa consiste na integração de bases de currículos, atividades de docência e de grupos de pesquisa para geração/engenharia de atributos.

A mineração de dados é a quarta etapa do fluxo, onde conhecimentos são extraídos do conjunto de atributos da etapa anterior. Algoritmos de predição são um exemplo de estratégia para extração de conhecimento. Existem diversos algoritmos disponíveis para predição automática, como Random Forests (JAYAPRAKASH; KRISHNAN; JAIGANESH,

2020), Support Vector Machines (BURMAN; SOM, 2019) e Redes Neurais (XIE, 2020). Todos esses algoritmos são capazes de extrair padrões que relacionam os atributos dos dados a uma variável resposta, a qual se deseja prever. Portanto, dada uma base de treinamento, o objetivo geral dessas técnicas consiste no aprendizado de padrões em dados de treinamento para predições em dados ainda não vistos. Apesar do mesmo objetivo, cada algoritmo possui suas especificidades. Por exemplo, Random Forests se utiliza de um comitê de árvores de decisão, sendo normalmente pouco sensível à parametrização e capaz de tratar diferentes tipos de atributos (JAYAPRAKASH; KRISHNAN; JAIGANESH, 2020). O SVM visa encontrar a margem máxima de separação a partir de um hiperplano separador, e redes neurais se utilizam de uma arquitetura de neurônios artificiais cujos pesos que os conectam formam o modelo de predição.

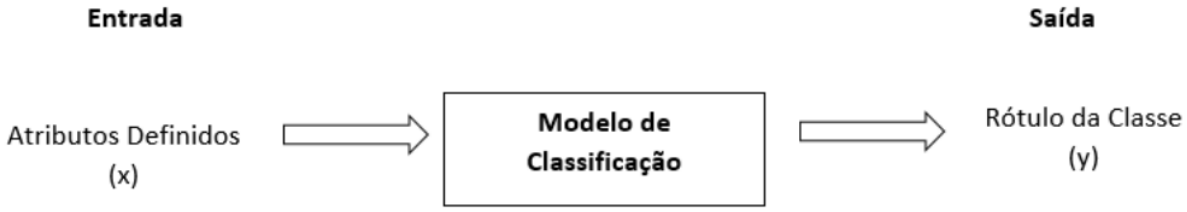
Por fim, a última etapa consiste na análise de performance. Na análise de performance, medidas como Acurácia, Macro/MicroF1 (CASTRO; FERRARI, 2017) são utilizados para fornecer resultados quantitativos sobre a performance dos modelos de predição. A estratégia de validação é dependente da tarefa de predição utilizada. Tanto as medidas de avaliação quanto a forma de separar os dados (entre dados de treinamento do modelo e dados de validação/teste do modelo) devem ser escolhidas de acordo com a tarefa. Por exemplo, em tarefas onde não se considera o caráter temporal, é comum que os dados sejam ordenados de forma aleatória e posteriormente particionados para validação cruzada (BREIMAN; SPECTOR, 1992). Em dados temporais, entretanto, esse tipo de validação não pode ser utilizada, dado que a ordem dos dados não pode ocorrer de forma aleatória. Nesse caso, são utilizadas janelas temporais, onde dados anteriores a um determinado período são utilizados como treinamento e dados posteriores para avaliação da predição dos modelos (XIE, 2020).

3.3 Classificação e Regressão em Janelas Temporais

Classificação e regressão são processos clássicos para predição de uma variável de resposta a partir de um conjunto de dados. De forma geral, o processo de classificação é a tarefa de atribuir uma classe ou categoria predefinida aos objetos de uma determinada entrada. Essa classificação leva em consideração os atributos e características deste objeto, gerando um modelo que permite obter a classe de saída. A Figura 2 mostra este procedimento.

Em classificação, a variável de resposta (rótulo da classe - y) é uma dentre um conjunto de categorias pré-definidas. Por exemplo, dados atributos sobre um pesquisador (ex.: publicações anteriores, número de orientações, disciplinas ministradas, etc) um modelo de classificação utiliza tais atributos para prever se o pesquisador terá (y=1) ou não (y=0) uma publicação acadêmica no próximo ano (isso é, duas possíveis categorias de resposta

Figura 2 – Processo de Classificação mapeando um conjunto de features para uma variável resposta (classe) .



Fonte: adaptado de (TAN; STEINBACH; KUMAR, 2006).

do modelo).

Por outro lado, em modelos de regressão, a variável de resposta y deixa de ser o rótulo da classe com categorias pré-definidas e passa a ser um valor real. Seguindo o exemplo anterior, o modelo de regressão poderia prever a quantidade de artigos publicados no próximo ano.

Tanto modelos de classificação quanto de regressão são aprendidos a partir de um conjunto de exemplos de treino previamente rotulados. Por exemplo, a Figura 3 representa um conjunto de treinamento para o aprendizado de um modelo de classificação sobre a existência de produção acadêmica de pesquisadores no próximo ano, onde cada pesquisador é um exemplo de treino. Os atributos do pesquisador (preditores) e as rotulações são utilizados para o aprendizado do modelo de predição a partir de técnicas de aprendizado de máquina.

Previsores classe

Nome	atr1	atr2	atr3	atr4	classe
Pesquisador 1	0,7	1	1	18	1
Pesquisador 2	0,9	2	0	19	0
Pesquisador 3	0,7	3	2	24	1
Pesquisador 4	0,01	4	1	44	0
Pesquisador 5	0,1	9	1	52	1
Pesquisador 6	0,5	4	0	34	1
Pesquisador 7	0,54	8	2	39	0
Pesquisador 8	0,63	9	0	24	0
Pesquisador 9	0,19	1	2	25	0
Pesquisador 10	1	0	2	20	1

Figura 3 – Conjunto de treinamento tradicionalmente usado em técnicas de aprendizado de máquina.

O cenário da Figura 3 apresenta uma coleção estática, que ignora a dimensão temporal dos dados. Entretanto ao considerar a dimensão temporal dos dados, torna-se necessário o tratamento da criação de atributos e exemplos de treino para levar em consideração as mudanças dos atributos em relação ao tempo (G. BEN TAIEB S., 2013).

Particularmente, considerando que uma entidade pode evoluir em relação ao tempo, cada uma de suas evoluções em diferentes janelas (intervalos) de tempo podem ser tratadas como novos exemplos de treinamento. Nesse cenário, os atributos do exemplo de treino são geradas a partir de uma janela de tempo, e cada nova janela corresponde a um novo exemplo (G. BEN TAIEB S., 2013).

Por exemplo, na Figura 4 o mesmo pesquisador pode contribuir com diversos exemplos de treinamento. Considerando uma janela de 5 anos, atributos (como média de publicações dos 5 anos, orientações nos últimos 5 anos, etc.) são gerados em diferentes janelas de 5 anos. Nesse caso, podemos gerar um exemplo de treino com atributos gerados considerando uma janela entre 2010 e 2015 e a classe 1 indicando que ele publicou em 2016. Dessa forma, múltiplos exemplos de treinamento podem ser gerados considerando diferentes períodos de tempo (G. BEN TAIEB S., 2013).

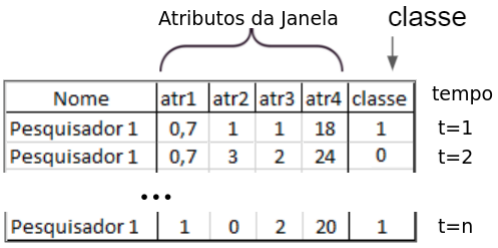


Figura 4 – Replicação do mesmo exemplo de treinamento em diferentes janelas temporais.

A engenharia de atributos a serem utilizados têm um papel central na qualidade da predição em dados de treino que consideram a dimensão temporal, sendo eles altamente dependentes do contexto. Nesse sentido, estatísticas que avaliam a distribuição de dados dentro da janela, técnicas de redução de dimensionalidade e normalização podem auxiliar no aprimoramento da efetividade (BHUIYAN et al., 2020). A prática de criação de diferentes atributos específicos para o contexto dos dados temporais a serem utilizados é bastante difundida, em especial para estratégias de predição no mercado de ações (MA; HAN; FU, 2019), onde indicadores de compra e venda são utilizados como atributos.

3.4 Predição da Produtividade Acadêmica em Janelas Temporais

Na literatura, há uma quantidade limitada de trabalhos que tratam especificamente de modelos e preditores para produtividade acadêmica no decorrer da carreira do pesquisador. A maioria dos trabalhos focam em aspectos/preditores específicos, como a relação (pelo coeficiente de Pearson) entre o sucesso da produtividade acadêmica e o número de publicações antes do doutorado (LAURANCE et al., 2013), a relação entre a produtividade acadêmica e idade do pesquisador (LEHMAN, 2017; TOMASSINI; LUTHI, 2007), onde

conclui-se que a produtividade tende a aumentar rapidamente, atingir um pico e então reduzir lentamente.

Análises de dados mais recentes demonstraram a complexidade do problema, onde foi encontrada uma grande diversidade de padrões de produtividade em relação ao tempo, o que dificulta a predição da produtividade do pesquisador (WAY et al., 2017). Neste trabalho, a análise da diversidade foi obtida a partir de regressões lineares considerando o tempo e o número de publicações dos pesquisadores. A diversidade de padrões é elencada como uma potencial causa para existência de poucos trabalhos na literatura que tratam o tema da predição de produtividade (XIE, 2020).

Mais recentemente, o trabalho (XIE, 2020) se utiliza da regressão Poisson como modelo de predição do número de publicações de pesquisadores (e grupos de pesquisadores) com o tempo. A modelagem proposta, quando comparada a redes neurais recorrentes, foi capaz de apresentar melhores resultados para predições de grupos de pesquisadores.

Apesar do esforço de predição automática da literatura considerando a produtividade em diferentes janelas temporais, tais predições são focadas em poucas informações disponíveis para predição (em geral, o número de publicações e o tempo). A análise de preditores que se utilizam de múltiplos dados, incluindo a ocupação docente do pesquisador, a nota (qualis) dos artigos e os grupos de pesquisa do pesquisador não foram avaliadas na literatura. Este trabalho visa suprir essa deficiência a partir da coleta, processamento, predição e avaliação desses múltiplos dados na tarefa de predição de produtividade.

4 Metodologia

Este trabalho se utiliza da metodologia exploratória, para caracterização e exploração inicial dos dados, bem como a pesquisa quantitativa, que visa apresentar evidências empíricas sobre a capacidade preditiva dos dados em janelas temporais. Neste capítulo, será descrita a estratégia de aplicação da metodologia para descoberta de conhecimento em bases de dados (Knowledge Discovery in Databases) (CASTRO; FERRARI, 2017), seguindo as etapas descritas a seguir. Na Seção 4.1, serão descritas as estratégias coleta de dados relevantes para a tarefa, bem como as características da coleção. Na Seção 4.2, a estratégia de pré-processamento de dados necessária para o problema em questão será apresentada. Na Seção 4.3, serão descritos os métodos de predição a serem avaliados para descoberta de padrões nos dados pré-processados. Por fim, a estratégia de validação da predição será descrito na Seção 4.4.

4.1 Seleção/Coleta dos Dados

Neste trabalho, foram utilizadas diversas fontes de dados para obtenção das informações dos pesquisadores do IFG. Foi utilizado o banco de dados do IFG Produz[1], dados capturados do currículo lattes[2] dos pesquisadores, dos espelhos de grupos de pesquisa cadastrados no CNPq[3], do Portal Brasileiro de Dados Abertos[4] e da plataforma de dados aberta Sucupira[5].

Especificamente, os dados dos currículos extraídos da plataforma Lattes foram obtidos a partir da plataforma IFG Produz. As coleções das disciplinas dos docentes foram obtidas a partir do Portal Brasileiro de Dados Abertos. Já para as informações sobre os grupos de pesquisa IFG, foi necessária a coleta a partir do crawler (selenium) que utiliza o navegador (de forma headless) para obter informações atualizadas sobre os grupos de pesquisa, o mesmo processo de coleta foi utilizado na plataforma sucupira para vincular o qualis de uma revista ao artigo publicado nela pelo pesquisador. Todas essas informações foram integradas à base de dados do IFG Produz.

Logo, foram obtidas informações sobre pesquisadores, grupos de pesquisa, artigos em conferências ou revistas, e informações sobre atuação em docência. Atualmente, a base de dados conta com 1834 pesquisadores, 76 grupos de pesquisas e 23241 produções bibliográficas. Os grupos de pesquisa coletados dos espelhos de grupos cadastrados no

[1] <https://ifgproduz.ifg.edu.br>
[2] <https://lattes.cnpq.br>
[3] <http://dgp.cnpq.br/dgp/faces/consulta/consulta_parametrizada.jsf>
[4] <https://dados.gov.br/organization>
[5] <https://sucupira.capes.gov.br>

CNPq foram obtidos os valores atuais se o pesquisador faz parte ou não de um grupo no momento atual, esse aspecto da participação nos grupos não foi possível gerar de forma temporal pela falta de informação da passagem de pesquisadores nos grupos levando em consideração o tempo. Para gerar o passo posterior de pré-processamento de dados, foram utilizado atributos extraídos da modelagem de banco de dados apresentada na Figura 5. Nesta figura podemos ver algumas tabelas como produção bibliográfica que podem ser do tipo artigos publicados, resumos, trabalhos, livros publicados ou capítulos de livros dos quais o tipo artigo é o tipo de publicação que tem os qualis que será alvo das predições enquanto a tabela produão técnica tem produções do tipo: relatório técnico, organizações de seminários ou eventos, curadoria entre outros tipos.

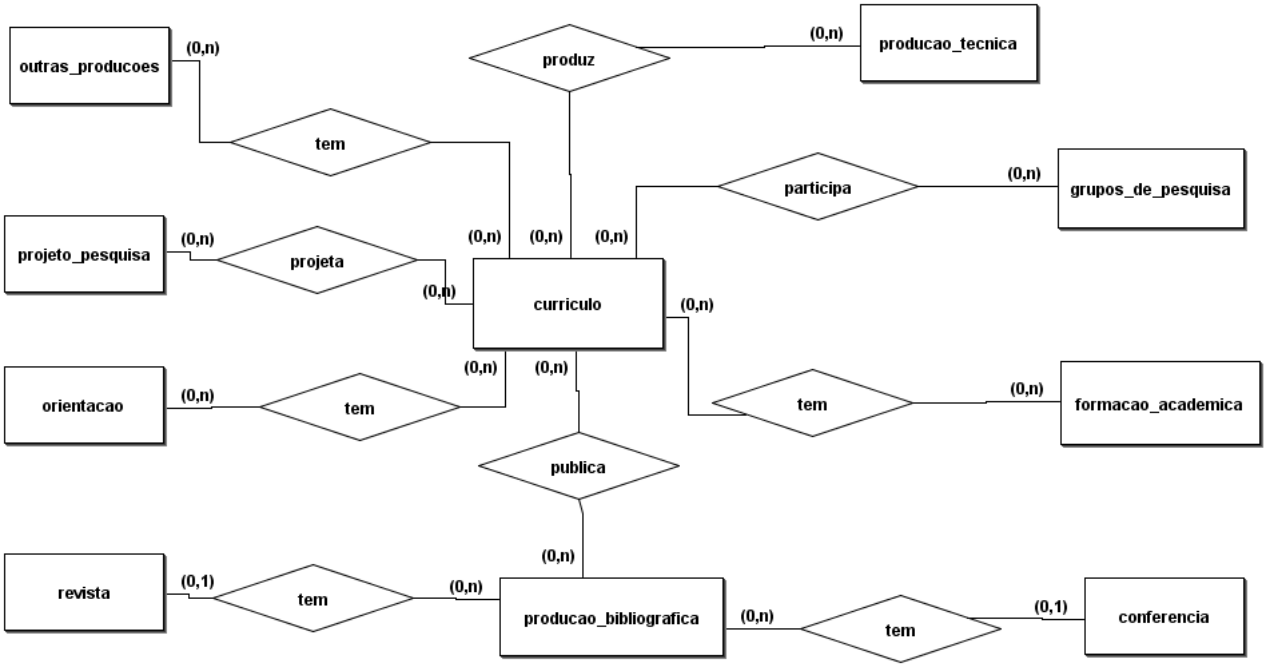


Figura 5 – Modelagem da base de dados do IFG Produz.

4.2 Pré-processamento dos Dados

Durante a etapa de engenharia de atributos, diversas estratégias foram utilizadas para representar atributos a partir dos dados coletados. Todos atributos foram gerados para avaliar um período de tempo de uma janela de 5 anos, que representa as produções acadêmicas e informações recentes sobre os docentes. Tal período de tempo foi escolhido devido ao fato que em exames de seleção, critérios para obtenção de bolsas e avaliação docente, usualmente são consideradas apenas as atividades docentes e de pesquisa dos últimos 5 anos [6].

[6] <https://www.gov.br/capes/pt-br/centrais-de-conteudo/formulario-pro-administracao-doc>

Dentre as informações utilizadas, a quantidade, média e a moda dos artigos produzidos se enquadram nesse cenário. Foram também considerados o número de orientações, projetos de pesquisa e produção técnica são outras atividades docentes que podem estar vinculadas à pesquisa. Da mesma forma, o número de disciplinas que o pesquisador atualmente ministra e a idade do pesquisador são informações que podem ter efeito sobre a pesquisa. Portanto, a análise de tais informações podem auxiliar na atividade de traçar um perfil de pesquisador.

Diversos grupos de atributos foram criados para categorizar os tipos de informações obtidas na base de dados. A Seção 4.2.1 apresenta detalhes sobre os atributos gerados nessa fase de pré-processamento de dados. Seguindo as estratégias de pré-processamento, a rotulação e tratamento do desbalanceamento de dados rotulados é apresentada na Seção 4.2.2. Por fim, uma breve caracterização de atributos criados durante a fase de pré-processamento é apresentada na Seção 4.2.3.

4.2.1 Grupos de atributos

Os atributos utilizados neste trabalho foram divididos em 5 grupos, que se referem aos diferentes tipos de informação obtidos da base de dados. Os grupos de atributos são divididos em:

- Atributos para atividades vinculadas a pesquisa.
- Atributos do perfil do pesquisador.
- Atributos sobre demais produções.
- Atributos sobre publicações acadêmicas.
- Atributos sobre disciplinas ministradas.

Atributos para atividades vinculadas a pesquisa. Este grupo de atributos se refere a informações sobre projetos de pesquisa, participação grupo de pesquisa e quantidade de orientações, como descritos na Tabela 4.2.1. Tais atributos capturam informações sobre o envolvimento do pesquisador em projetos e orientações na pesquisa, algo que normalmente resulta na produção de conhecimento acadêmico. Em suma, utilizamos estatísticas que representam as atividades de pesquisa de cada pesquisador em uma janela de 5 anos.

Atributos do perfil do pesquisador. Os atributos "idade, formação e cursando doutorado", descritos na Tabela 4.2.1, tratam das informações e estatísticas que representam o perfil do pesquisador. Tais atributos nos expressam o fato de um docente estar cursando o doutorado, a idade do docente, e o grau de formação do mesmo. Tais atributos se mostraram importantes para determinar o potencial de produção acadêmica (LEHMAN, 2017; LAURANCE et al., 2013).

Atividades vinculadas a pesquisa	
Nome	Descrição
Projetos de Pesquisa (Agrupado)	5 atributos referentes à quantidade de projetos de pesquisa desenvolvidos a cada ano nos últimos 5 anos. Esta feature está agrupada e se refere a 5 features chamadas: Projeto de pesquisa ano 5 que é do ano atual da janela temporal, em uma janela temporal de 2013-2017, o ano atual é 2017, projeto de pesquisa ano 4 são os projetos do ano anterior: 2016, projetos de pesquisa do ano 3 são os do ano antepenultimo 2015 e projetos de pesquisa ano 2 e 1.
Participação Grupo de Pesquisa	Se o pesquisador participa de um grupo de pesquisa, atributo obtido do momento atual do grupo de pesquisa devido a dificuldade de obter essa informação de forma temporal
Quant. Orientações 5 anos	Quantidade de orientações do pesquisador gerado dentro dos 5 anos da janela

Tabela 1 – Atividades vinculadas a pesquisa

Atributos vinculadas ao perfil do pesquisador	
Nome	Descrição
Idade	Idade do pesquisador baseado no último ano da janela de 5 anos, por exemplo, numa janela de 2013-2017 a idade do pesquisador é calculada com o ano de 2017
Formação	Grau de formação do pesquisador, quanto mais alto maior o número
Cursando Doutorado	Boleano se o pesquisador está cursando doutorado durante o período dentro da janela temporal

Tabela 2 – Atributos do perfil do pesquisador

Grupo de atributos demais produções. Os atributos "produção técnica e quantidade de outras produções", descritos na Tabela 4.2.1 tratam das informações que representam os demais tipos de produções de um pesquisador como: artes cênicas, produções musicais e culturais que não são do tipo artigo mas que ocupam o tempo de produtividade de um pesquisador.

Atributos demais produções	
Nome	Descrição
Producão Técnica	Quantidade de produções técnicas dos 5 anos da janela
Quant. Outras Produções	Quantidade de outras produções dos 5 anos da janela, gerada a partir de produções classificadas com outras no currículo lattes como: artes cênicas, produções musical e cultural

Tabela 3 – Atributos demais produções

Grupo de atributos publicações bibliográficas. O grupo de atributos: (média artigos A1-A2, média de artigos B1-B3, média de artigos B4-B5, minimo A1-A2, máximo A1-

A2, mínimo B1-B3, máximo B1-B3, mínimo B4-B5, máximo B4-B5, moda artigos A, moda artigos B1-B3 e moda artigos B4-B5), descritos na Tabela 4.2.1 tratam das informações que representam as publicações bibliográficas do tipo artigo de um pesquisador. A divisão da produção acadêmica de um pesquisador em diferentes estatísticas sobre os qualis dos artigos publicados tem como objetivo discriminar os diferentes tipos de informação sobre a qualidade dos artigos publicados.

Atributos publicações bibliográficas	
Nome	Descrição
Média artigos A1-A2	Média da quantidade de artigos de qualis A1 e A2
Média artigos B1-B3	Média da quantidade de artigos de qualis B1, B2 e B3
Média artigos B4-B5	Média da quantidade de artigos de qualis B4 e B5
Min. A1-A2	Valor mínimo de artigos A1 ou A2 publicados em 5 anos
Max. A1-A2	Valor máximo de artigos A1 ou A2 publicados em 5 anos
Min. B1-B3	Valor mínimo de artigos B1, B2 ou B3 publicados em 5 anos
Max. B1-B3	Valor máximo de artigos B1, B2 ou B3 publicados em 5 anos
Min. B4-B5	Valor mínimo de artigos B4 ou B5 publicados em 5 anos
Max. B4-B5	Valor máximo de artigos B4 ou B5 publicados em 5 anos
Moda artigos A	Valor mais frequente da quantidade de artigos de qualis A1 e A2
Moda artigos B1-B3	Valor mais frequente da quantidade de artigos de qualis B1, B2 e B3
Moda artigos B4-B5	Valor mais frequente da quantidade de artigos de qualis B4 e B5

Tabela 4 – Atributos publicações bibliográficas

Grupo de atributos disciplinas. O grupo de atributos: (bacharelado, integrado, licenciatura, mestrado, tecnólogo, ensino médio, superior), descritos na Tabela 4.2.1 tratam das informações que representam as disciplinas ministradas por um pesquisador. Tais informações apresentam a dedicação do professor para disciplinas em diferentes níveis de formação acadêmica.

4.2.2 Rotulação e Tratamento do Desbalanceamento

Na predição, a rotulação ou variável de resposta é importante para delimitar o objetivo do modelo. Nesse sentido, para a classificação, a variável de resposta, no contexto desse trabalho, é um atributo binário que representa o que desejamos prever. Nesse sentido, para tarefa de classificação, desejamos responder à pergunta "se um pesquisador terá uma publicação B2 (ou qualis superior) ou não" no ano seguinte.

Na regressão a variável de resposta que queremos prever é a quantidade de publicações de qualis B2 (ou qualis superior) um pesquisador terá no ano seguinte ao da janela temporal. A escolha do estrato qualis B2 foi feita seguindo critérios usualmente

Atributos Disciplinas	
Nome	Descrição
Bacharelado	Quantidade de disciplinas de bacharelado que o pesquisador ministra
Integrado	Quantidade de disciplinas da modalidade integrado que o pesquisador ministra
Licenciatura	Quantidade de disciplinas da modalidade licenciatura que o pesquisador ministra
Mestrado	Quantidade de disciplinas da modalidade mestrado que o pesquisador ministra
Tecnólogo	Quantidade de disciplinas da modalidade tecnólogo que o pesquisador ministra
Ensino Médio	Quantidade de disciplinas do nível ensino médio que o pesquisador ministra
Superior	Quantidade de disciplinas do nível superior que o pesquisador ministra

Tabela 5 – Atributos disciplinas

traçados por programas de pós-graduação para qualificar o estrato mínimo das publicações aceitas [7].

A partir da variável de resposta, outra estratégia de pré-processamento dos dados consiste no tratamento do desbalanceamento. Por exemplo, assumindo a classe/variável de resposta correspondente à existência (ou não) de publicações no próximo ano da janela temporal, é possível que a grande maioria dos pesquisadores possam estar vinculados à não existência de publicações. Nesse caso, pode ser necessário o tratamento do desbalanceamento a partir da manipulação dos dados que serão utilizados para o treinamento de técnicas de aprendizado de máquina.

Neste trabalho, foram utilizadas duas técnicas para tratar o desbalanceamento:

1. SMOTE (Synthetic Minority Over-sampling Technique) (CHAWLA K. W. BOWYER, 2002).

2. Sub-amostragem de dados antigos (proposta nesse trabalho).

SMOTE. Uma das estratégias mais usadas para tratar o desbalanceamento é denominada SMOTE – *Synthetic Minority Over-sampling Technique* (Técnica de Super-Amostragem Sintética da Minoria). A SMOTE cria observações sintéticas da classe minoritária (no nosso caso, docentes com publicação em um determinado ano). Mais precisamente, a técnica SMOTE executa os seguintes passos:

[7] https://ppgcg.ufsc.br/emissao-de-diploma/

1. Busca os k vizinhos mais próximos dos exemplos da classe minoritária (isto é, buscando observações similares).

2. Escolhe aleatoriamente um dos k vizinhos mais próximos e o utiliza para criar novas observações modificadas aleatoriamente.

Em outras palavras, o SMOTE (CHAWLA K. W. BOWYER, 2002) aumenta a quantidade de exemplos da classe minoritária baseado em registros mais próximos a partir de novas instâncias sintéticas para diminuir o desbalanceamento entre as classes.

Sub-amostragem de dados antigos (proposta). Neste trabalho é proposta uma nova metodologia para remoção de exemplos pertencentes à classe majoritária. No nosso cenário, exemplos de treinamento podem ser obtidos de diferentes janelas temporais. Portanto assumindo que exemplos recentes sejam mais importantes para predições futuras, a estratégia proposta pode ser facilmente descrita em dois passos:

1. Ordene os exemplos dos mais antigos para os mais novos.

2. Exclua os exemplos mais antigos da classe majoritária, com exceção dos exemplos da janela temporal mais recente.

Na proposta, as janelas temporais são empilhadas e no passo 1 elas serão ordenadas dos valores mais antigos para os mais novo e então no passo 2 é excluído os registros da classe majoritária das janelas mais antigas permanecendo apenas os registros da classe minoritária, essa abordagem parte do princípio que os dados mais antigos são menos importantes para o aprendizado de padrões por meio de técnicas de aprendizado de máquina. Apesar disso, são preservados os exemplos da classe majoritária mais recentes (isso é, gerados com a janela temporal mais recente), ainda que os exemplos da classe majoritária recentes sejam mais numerosos que o conjunto de exemplos da classe minoritária.

4.2.3 Exemplos de Análise de Atributos

Durante o pré-processamento dos dados para a construção dos grupos de atributos previamente apresentados, faz-se necessária uma análise preliminar dos dados transformados em atributos para tarefa de mineração (CASTRO; FERRARI, 2017). Portanto, objetivando uma melhor compreensão da distribuição e correlação dos atributos obtidos da coleção, foram feitas análises preliminares dos atributos em questão.

A Figura 6 apresenta um mapa de calor com a correlação (com Spearman) de todas as features do grupo de features de atividades vinculadas à pesquisa. Na figura, são correlacionadas todas as features desse grupo: a existência de projetos de pesquisa do primeiro ao quinto ano de uma janela temporal, a participação em grupo de pesquisa, a quantidade de orientações e o número de produções. É possível observar que os projetos

de pesquisa têm grande correlação com as orientações, pois quem orienta normalmente tem projetos de pesquisa em andamento. Da mesma forma, a maior correlação pode ser observada entre orientações e grupos de pesquisa. Tal correlação indica que pesquisadores que orientam normalmente fazem parte de grupos de pesquisa. As maiores correlações com o número de produções futuras podem ser observadas com orientações e participação de grupos de pesquisa, haja vista que estar num grupo de pesquisa e orientar promover subsídios para próximas produções de artigos de um pesquisador.

O gráfico da Figura 7 mostra a vinculação entre a quantidade de artigos publicados e o nível de formação do pesquisador. Considerando o intervalo de 5 anos, a grande maioria dos docentes do IFG estão com 0 artigos publicados. Dos pesquisadores que publicaram 1 artigo, a maioria são doutores, seguidos de mestres. Como há poucos pós-doutores na instituição em relação ao número de mestres e doutores, o gráfico apresenta uma pequena quantidade de pós-doutores com uma publicação. Tal distribuição se mantém com dois ou mais artigos: Dos docentes que publicaram 2 artigos, a maioria também são doutores e depois mestres, e os que publicaram 3 artigos ou mais são doutores.

A Figura 8 mostra o gráfico de pesquisadores por idade. Nele, podemos observar que a maioria dos pesquisadores estão na faixa de 30 a 40 anos de idade. Tal informação apresenta um quadro docente relativamente jovem, e com grande potencial para envolvimento em produções acadêmicas, haja vista a grande quantidade de mestres e doutores presentes na instituição.

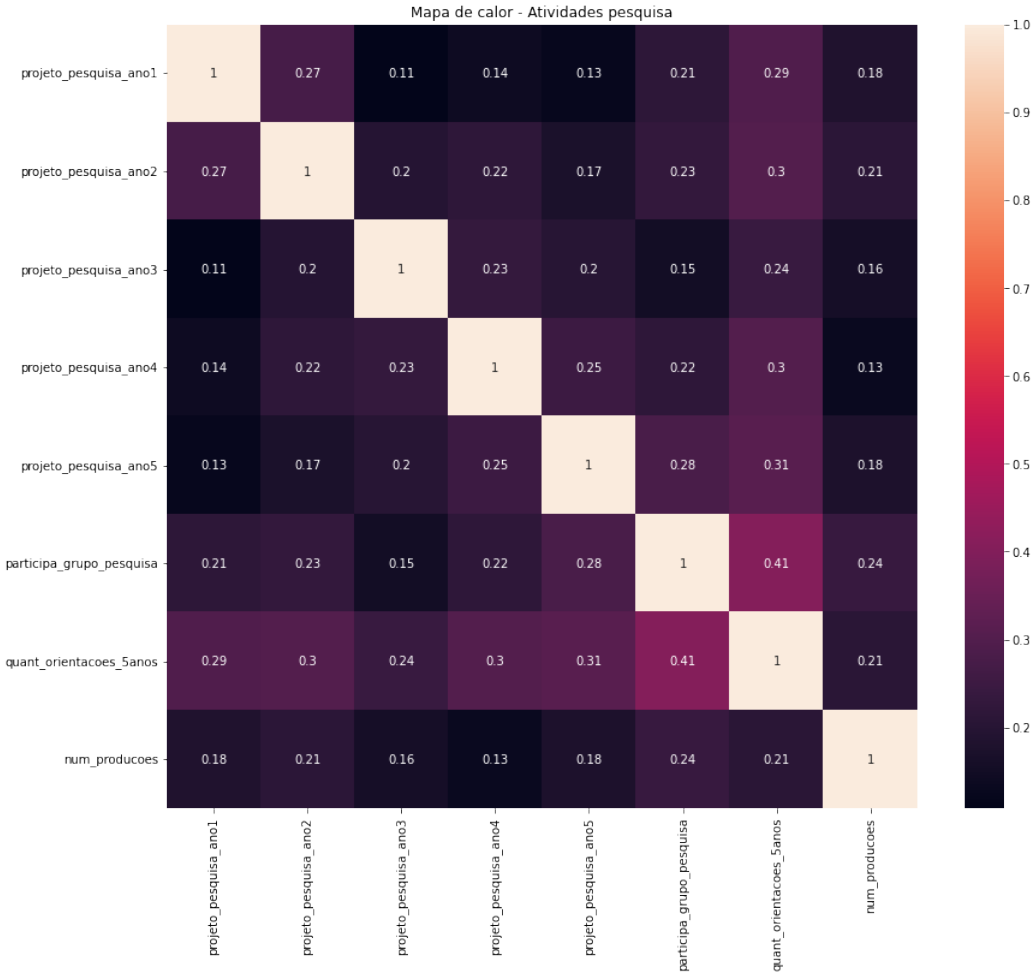


Figura 6 – Mapa de calor - correlação de Spearman

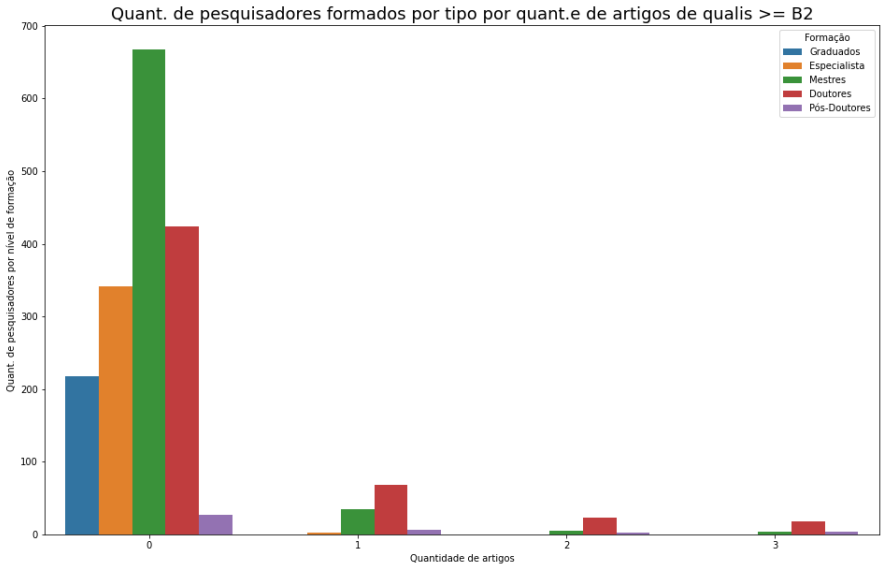


Figura 7 – Quant. de pesquisadores formados pelo tipo por Quant. de publicações de qualis >= B2

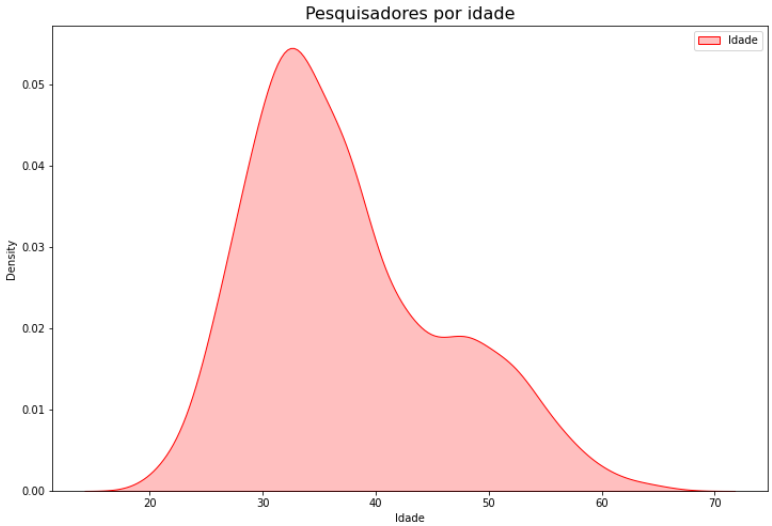


Figura 8 – Pesquisadores por idade

O Gráfico da Figura 9 apresenta a idade dos docentes que estão cursando doutorado, neste gráfico foi considerado os pesquisadores que estão cursando doutorado independente de já terem um, ou seja, o gráfico pode conter pesquisadores que já são doutores. Por se tratar de um conjunto de profissionais relativamente jovem, há grande potencial para que os profissionais se engajem em cursos de doutorado, o que eleva a produtividade acadêmica (LAURANCE et al., 2013). Como esperado, a faixa de idade onde se encontram a maioria dos docentes também é a faixa onde se encontra a maioria que está cursando doutorado, embora o engajamento no doutorado seja proporcionalmente maior entre os docentes na faixa dos 40 anos.

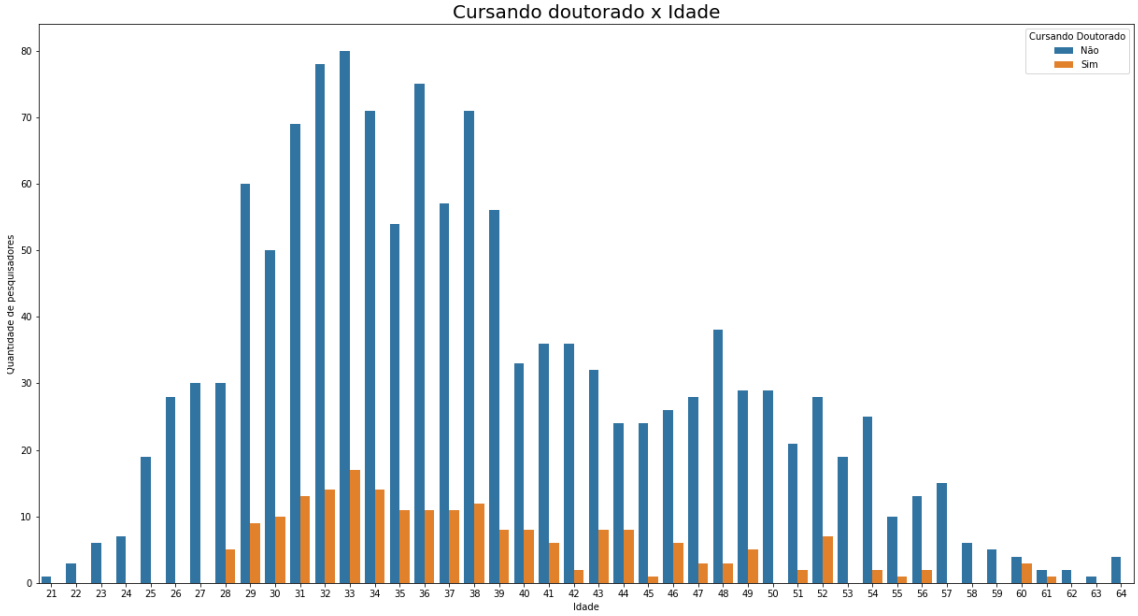


Figura 9 – Cursando doutorado por idade

4.3 Estratégias de Predição

Um dos aspectos centrais desse trabalho consiste na predição das publicações acadêmicas de pesquisadores do IFG a partir das informações pré-processadas nas etapas anteriores. Em outras palavras, dado um conjunto de informações sobre um pesquisador, o pesquisador publicará no próximo ano? (classificação binária). Quantos artigos ele publicará? (regressão).

Dado o caráter temporal, este trabalho utilizará técnicas de aprendizado de máquina usualmente utilizadas para tratamento de séries temporais. Recentemente, estratégias de aprendizado de máquina como Random Forests e LSTM se mostraram dentre as mais efetivas para tratar esse tipo de problema no contexto de predição de ações (MA; HAN; FU, 2019). No contexto de predição de performance acadêmica, a regressão Poisson foi recentemente proposta como estratégia capaz de modelar efetivamente o comportamento de pesquisadores (XIE, 2020).

4.3.1 Random Forests

Nos experimentos, seguimos a parametrização da implementação disponibilizada pela biblioteca sklearn[8] na maioria dos parâmetros: para a amostragem de dados utilizados por construir cada árvore, foi adotada amostragem, com repetição, de todos os exemplos de treino. Para a amostragem das f features disponíveis, foi utilizada a raiz quadrada no número de features $\sqrt{f}$ por ser um atributo padrão do algoritmo. Como sugerido por (BREIMAN, 2001), não foi utilizada poda e o número de árvores foi setado para 3000, dado que (BREIMAN, 2001) sugere no mínimo 100 árvores, sendo que mais árvores podem aprimorar a efetividade em problemas complexos.

4.3.2 Redes Neurais Recorrentes - LSTM

Redes recorrentes são estratégias baseadas em redes neurais para tratar problemas que envolvem sequências temporais (CHOLLET, 2017). Diferentemente das redes diretas (como multilayer perceptron), redes recorrentes possuem um conjunto de parâmetros dedicado para armazenar e processar a memória da rede a cada nova entrada (instante) de uma série temporal.

Nesse trabalho, utilizamos duas variações de redes recorrentes: Redes Recorrentes Simples e LSTM:

Redes recorrentes simples (SimpleRNN). A SimpleRNN se utiliza de um conjunto de neurônios que processa, além da entrada (o ano corrente), o resultado interno do processamento de iterações anteriores (memória). Tal estratégia foi adotada em (XIE,

[8] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

2020) para predição do número de publicações de pesquisadores no próximo ano a partir do número de publicações em anos anteriores. Nos experimentos apresentados, foi adotada a utilização de 30 neurônios e uma janela tem 5 anos, sendo essa a quantidade de neurônios descritas na estratágia de parametrização de (XIE, 2020).

Redes recorrentes Long-Short Term Memory (LSTM). Redes LSTM podem ser vistas como estensões das redes SimpleRNN. A LSTM é capazes de filtrar as informações temporais a partir de uma arquitetura baseada em um conjunto de pontes (gates) que indicam quais informações devem ser esquecidas/lembradas em uma série temporal. O objetivo da manipulação de dados temporais com arquitetura LSTM é torná-la resiliente a ruídos e informações irrelevantes. Nesse trabalho, foi adotada a mesma parametrização utilizada na SimpleRNN descrita anteriormente.

4.3.3 Regressão Poisson

A regressão Poisson é uma estratégia de predição que assume que os dados seguem uma distribuição conhecida como Poisson (XIE; OUYANG; LI, 2016). No contexto do trabalho, a distribuição poisson expressa a probabilidade de um determinado número de ocorrências publicações em um determinado instante tempo. Tal escolha (poisson) se deve à possibilidade de parametrizar a distribuição Poisson para se encaixar na curva de publicações observada na evolução temporal das publicações dos autores (XIE, 2020). Nesse trabalho, seguimos a estratégia do emprego do modelo linear generalizado com distribuição Poisson para o treinamento dos modelos, como descrito em (XIE, 2020), e janela de tempo de 5 anos para o treinamento dos modelos.

4.3.4 Ferramentas Utilizadas

Para a construção da base de dados, análise dos atributos e modelos de predição foram utilizados as bibliotecas apresentadas na Tabela 6.

Nome	Descrição
Pandas	Biblioteca usada para manipulação e análise de dados
Seaborn	visualização de dados estatísticos
Matplotlib	Visualização de dados e gráficos em geral
Numpy	Usada para manipular matrizes e arrays
Sklearn	biblioteca de aprendizado de máquina

Tabela 6 – Bibliotecas utilizadas

4.4 Estratégias de Validação e Separação dos Dados

Para validação dos modelos, é necessária a separação dos dados em conjuntos de treinamento, onde os modelos serão aprendidos, e teste, onde os modelos serão avaliados. Nessa seção, serão detalhadas as estratégias de separação de dados e validação de modelos utilizadas nesse trabalho.

4.4.1 Separação de Janelas Temporais

Cada janela temporal pode ser vista como um dataframe com os atributos utilizados para realizar as predições (previsores) e outro atributo para variável de resposta (classe). A Figura 10 apresenta um dataframe, com os atributos derivados de uma janela de tempo (2009-2013) como colunas que representam as informações (atributos) sobre cada pesquisador obtidas nesse período. Os atributos podem ser, por exemplo, a quantidade de publicações que o pesquisador obteve entre os anos de 2013 e 2017, entre outros. Os atributos foram gerados segundo a descrição da Seção 4.2.1. No exemplo da figura, a classe, extraída do ano 2014, representa o fato de um pesquisador ter publicado ou não um artigo em 2014. A estratégia de rotulação para tarefas de classificação e regressão segue a descrição da Seção 4.2.2. Deste ponto em diante, adotaremos a representação simplificada da janela temporal representada na parte azul da Figura 4.2.1, que simplifica a descrição dos atributos e classe de uma janela temporal.

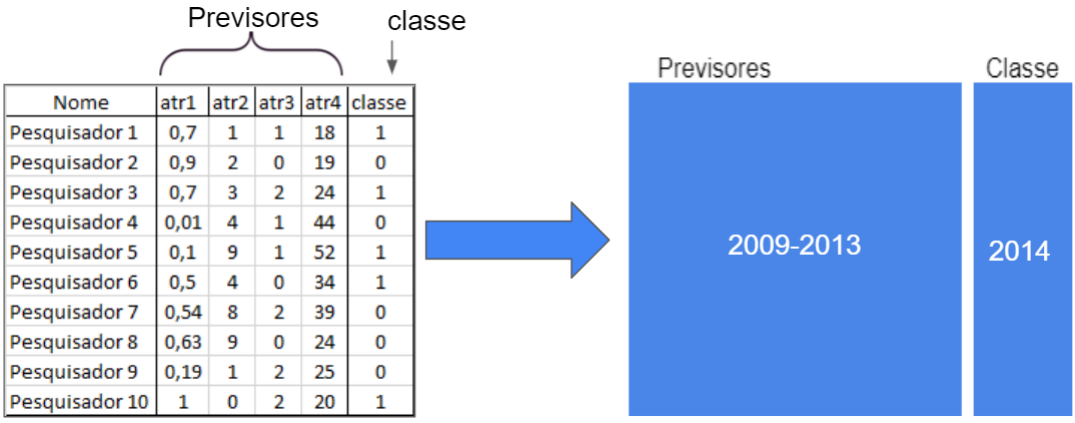


Figura 10 – Exemplificação dos dados (dataframe) de uma janela temporal.

Ao representar os dados de todos os pesquisadores em diferentes janelas temporais, é possível utilizar tais dados como exemplos de treinamento e exemplos de teste. Por exemplo, a Figura 11 representa em vermelho uma janela temporal cujos registros podem ser todos utilizados para geração de exemplos de treino (com features e rotulação derivadas de dados entre 2013 e 2017 e rotulação derivada de 2018). Da mesma forma, a Figura 12, em azul, representa dados gerados em outra janela temporal, de 2014 à 2018 e a rotulação derivada de 2019. Os atributos derivados dessa janela podem ser utilizados como exemplos

de teste. Em uma situação prática, por exemplo, o modelo aprenderia os padrões de predição a partir da janela em vermelho e faria a predição da janela em azul, assumindo que não são conhecidos os rótulos para o ano de 2019.

Figura 11 – 2013-2017 Treinamento

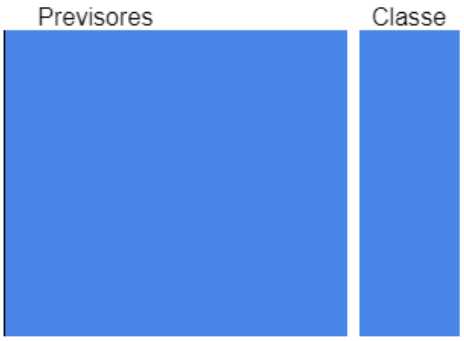


Figura 12 – 2014-2018 Teste

4.4.2 Empilhamento de Janelas Temporais

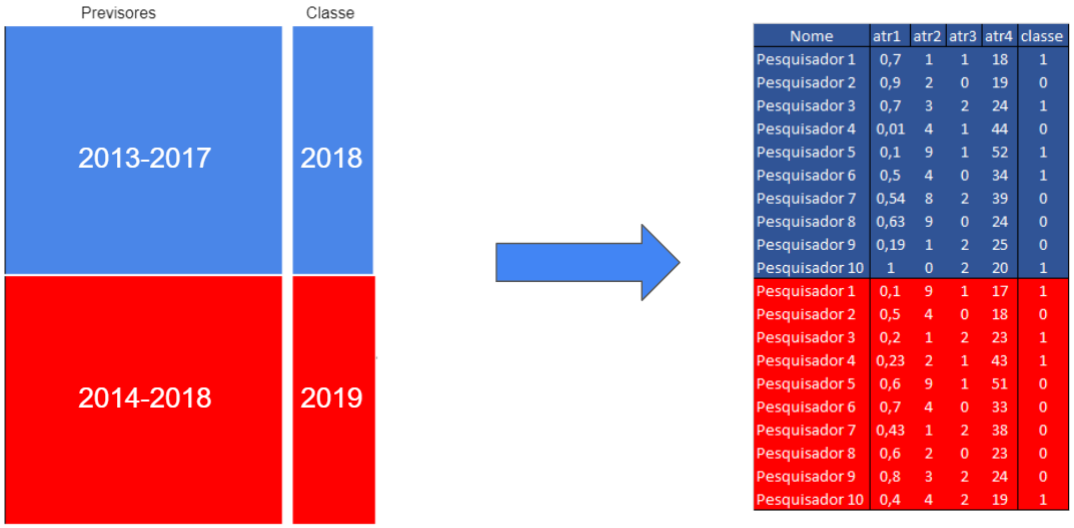
Como discutido na seção anterior, cada pesquisador dentro de uma janela temporal pode corresponder a um exemplo de treino. Entretanto, é possível expandir a quantidade de exemplos de treinamento com janelas temporais anteriores, como descrito na Seção 3.4. Tal expansão, denominada aqui como empilhamento de janelas, se utiliza do mesmo princípio utilizado na geração de exemplo de treino com séries temporais. Em tal cenário, basta expandir os exemplos de treino com novos exemplos de treino extraídos da janela temporal anterior.

Por exemplo, considere o Pesquisador 1 da Figura 13. Nesse cenário o pesquisador 1 aparece representado duas vezes na coleção de treinamento. A primeira ocorrência do pesquisador 1 corresponde ao exemplo de treino cujos atributos foram extraídos da janela temporal entre 2013-2017, e a segunda ocorrência corresponde ao exemplo de treino cujos atributos foram extraídos da janela temporal entre 2014-2018.

O empilhamento garante uma maior quantidade de exemplos de treinamento, incluindo padrões mais antigos (e potencialmente informativos) da evolução temporal. É importante observar que o empilhamento ocorre apenas com exemplos de treinamento. Os exemplos da janela temporal de teste não são empilhados durante a avaliação.

4.4.3 Validação com Janelas Temporais Deslizantes

Com o objetivo de garantir a replicação e confiabilidade de diferentes resultados experimentais em diferentes janelas de tempo, faz-se necessária a avaliação de diferentes conjuntos de teste obtidos em diferentes janelas temporais. Para tal avaliação, utilizamos a metodologia de Validação com Janelas Temporais Deslizantes (HASTIE; TIBSHIRANI;



Nome	atr1	atr2	atr3	atr4	classe
Pesquisador 1	0,7	1	1	18	1
Pesquisador 2	0,9	2	0	19	0
Pesquisador 3	0,7	3	2	24	1
Pesquisador 4	0,01	4	1	44	0
Pesquisador 5	0,1	9	1	52	1
Pesquisador 6	0,5	4	0	34	1
Pesquisador 7	0,54	8	2	39	0
Pesquisador 8	0,63	9	0	24	0
Pesquisador 9	0,19	1	2	25	0
Pesquisador 10	1	0	2	20	1
Pesquisador 1	0,1	9	1	17	1
Pesquisador 2	0,5	4	0	18	0
Pesquisador 3	0,2	1	2	23	1
Pesquisador 4	0,23	2	1	43	1
Pesquisador 5	0,6	9	1	51	0
Pesquisador 6	0,7	4	0	33	0
Pesquisador 7	0,43	1	2	38	0
Pesquisador 8	0,6	2	0	23	0
Pesquisador 9	0,8	3	2	24	0
Pesquisador 10	0,4	4	2	19	1

Figura 13 – Empilhamento das duas janelas temporais com detalhes de como os registros de pesquisadores se repetem mas com informações diferentes

FRIEDMAN, 2001), em que diferentes janelas de dados utilizadas para treinamento e teste respeitando a ordem temporal, evitando que padrões futuros sejam considerados durante a etapa de treinamento dos modelos.

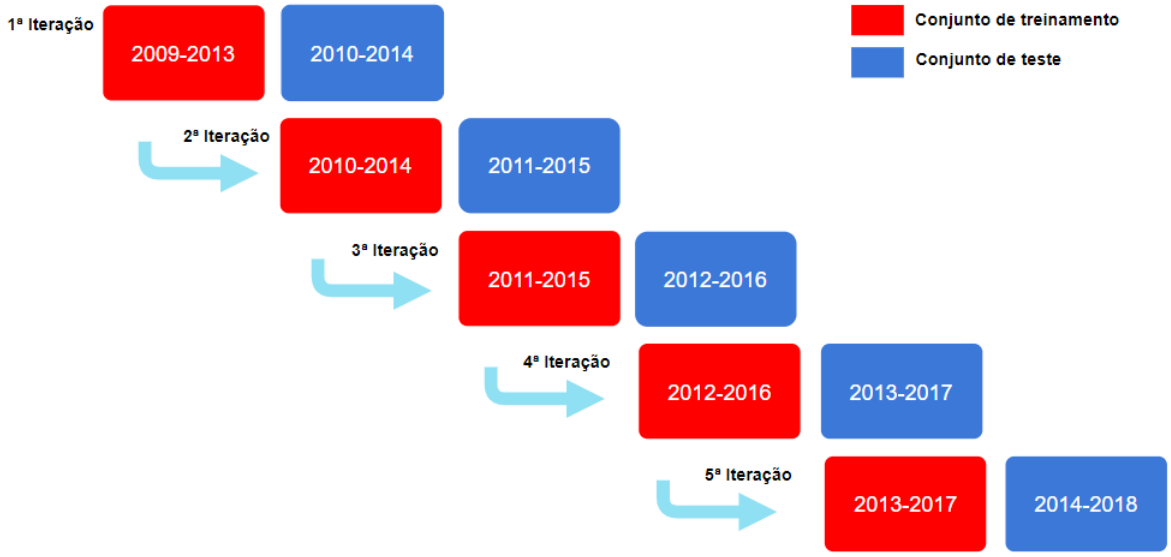


Figura 14 – Janela temporal deslizante

A Figura 14 exemplifica o processo de separação de 5 conjuntos de treino/teste com a metodologia de validação com janelas temporais deslizantes a partir de dados entre 2009 e 2018. Os dados são divididos de forma que um conjunto de treinamento contenha somente dados pertencentes à janela anterior aos que serão avaliados. Nesse cenário, as janelas temporais são geradas em forma que os atributos de cada pesquisador são coletados de acordo com cada janela de tempo.

Por exemplo, no primeiro par treino/teste (primeira iteração) da Figura 14, uma janela temporal de 2009-2013, utilizada para treino, é formado pelos atributos dos pesquisadores obtidos dentro desse espaço de 5 anos, tendo a classe alvo a sua produção acadêmica no ano de 2014. A próxima janela de 5 anos será então usada como conjunto de teste, gerado a partir do ano 2010 até 2014 e tendo com classe de publicações do ano 2015. Com isso, dados futuros (cuja classe é obtida em 2015) não são usados para o treinamento do modelo. O mesmo procedimento é repetido para iterações posteriores.

O objetivo final da utilização dessa técnica é a execução de 5 experimentos com 5 janelas temporais diferentes, o que nos permite avaliar a qualidade dos modelos de predição em diferentes períodos de tempo. Nos experimentos, utilizamos a janela temporal deslizante onde seguindo a figura 14 Dado os 5 resultados é então obtida a medida de efetividade média de predição do modelo (ex.: MacroF1) das iterações.

4.4.4 Estratégias e Métricas de Avaliação

Neste trabalho, os resultados são apresentados como a média de 5 experimentos, sendo eles obtidos nas diferentes janelas temporais obtidas entre 2009 e 2018, como apresentado na Figura 14. A comparação entre experimentos foi feita utilizando o teste-t pareado com 95% de significância estatística. Os resultados experimentais são apresentados a partir de duas estratégias específicas para classificação e regressão:

Classificação. Em experimentos de classificação, cujo objetivo é verificar se um pesquisador publica ou não no próximo ano, foi adotada a medida Macro-F1, devido ao fato que tal medida dá o mesmo peso para avaliação de diferentes classes, sendo indicada como estratégia de avaliação em cenários de desbalanceamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2001) (como o caso deste trabalho). Para duas categorais, a MacroF1 pode ser descrita como a simples média:

$$MacroF1 = \frac{F1_{classe0} + F1_{classe1}}{2}$$

Onde F1 é a média harmônica entre precisão e revocação obtidas para os exemplos de teste da classe0 e os exemplos de teste da classe1.

Regressão. Para tarefas de regressão, cujo objetivo é verificar quantos artigos um pesquisador publica, a avaliação da predição será feita a partir da métrica o MSE (mean squared error) que é uma função que calcula as diferenças elevadas ao quadrado penalizando os erros. Tal medida pode ser descrita como:

$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$, onde e_t é a subtração entre o valor predito pelo modelo e o valor real do exemplo t.

Foi também utilizada a visualização da matriz de confusão para avaliação de resultados de classificação. Matriz de confusão é uma tabela com duas linhas e duas

colunas que relata o número de falsos positivos (erro ao falar que um registro pertence a classe positiva), falsos negativos (erro ao falar que um registro pertence a classe negativa), verdadeiros positivos (quando acerta em falar que o registro pertence a classe positiva) e verdadeiros negativos (quando acerta em falar que um registro pertence a classe negativa). Isso permite uma análise mais detalhada do que a mera proporção de classificações corretas (precisão). Tal análise inicial elucidou a pouca quantidade de informação para no treinamento do modelo (como a falta de features relevantes ou o reduzido número de exemplos de treinamento), bem como o grande desbalanceamento nos dados.

5 Resultados Experimentais

5.1 Empilhamento de Dados e Tratamento do Desbalanceamento

Os dados utilizados neste trabalho podem ser considerados como extremamente desbalanceados. De fato, a Figura 15 apresenta o cenário de desbalanceamento considerando a quantidade de exemplos de treino positivos (docentes com ao menos uma publicação no próximo ano) e negativos (docentes que não publicam). Esse problema tende a prejudicar estratégias de predição, que acabam focando na classe majoritária durante o aprendizado.

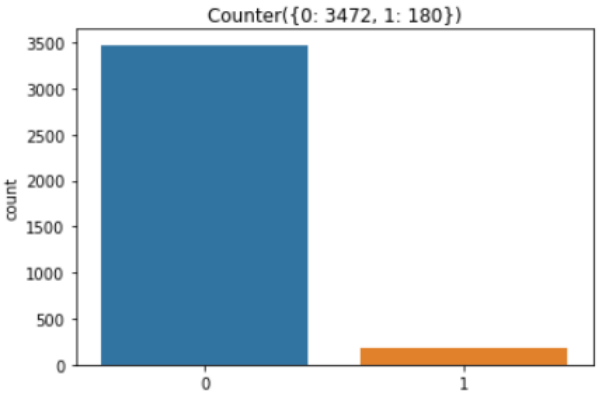


Figura 15 – Desbalanceamento das classes. Classe 0 corresponde à quantidade de docentes sem publicação e classe 1 são docentes com publicação.

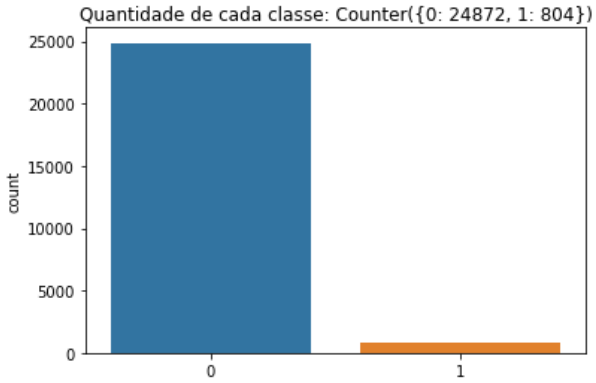


Figura 16 – Desbalanceamento das classes com empilhamento.

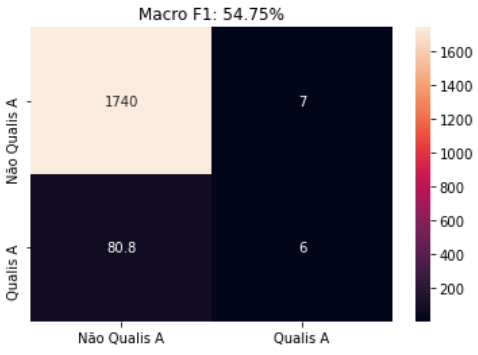


Figura 17 – Resultado do teste com empilhamento.

Tal problema tende a se agravar com a estratégia de empilhamento da Seção 4.4.2. De fato, a inclusão de exemplos de treino de janelas temporais mais antigas prejudica ainda mais o desbalanceamento, como descrito na Figura 16. Tal desbalanceamento faz com que um modelo de predição (no caso, o random forests) tenda a classificar incorretamente a

maioria dos exemplos, tendendo a selecionar a classe majoritária, como mostra a Tabela de Confusão descrita na Figura 17.

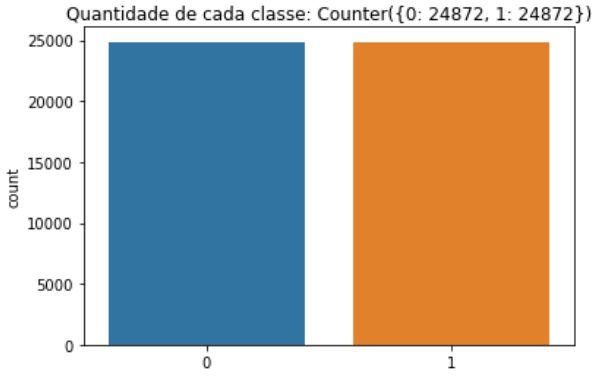


Figura 18 – Balanceamento das classes com SMOTE.

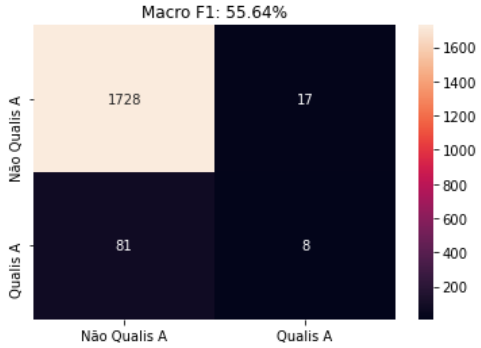


Figura 19 – Resultado do teste com SMOTE.

Balanceamento com SMOTE. Aplicando o método de sobreamostragem SMOTE sobre os dados empilhados, é possível observar, na Figura 18, que não há mais desbalanceamento de dados. Entretanto, a geração de exemplos sintéticos com o método SMOTE não foi capaz de aprimorar significativamente os resultados da classificação com o random forests, como mostrado na tabela de confusão da Figura 19.

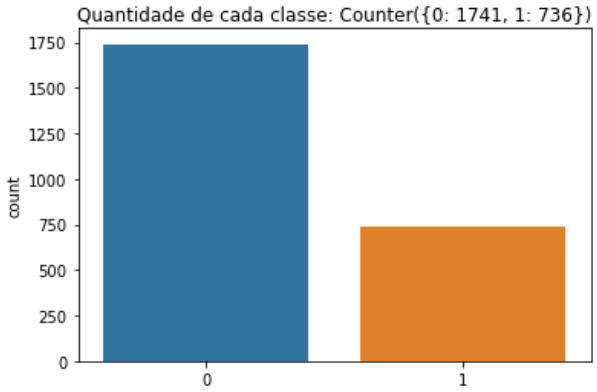


Figura 20 – Balanceamento das classes proposto

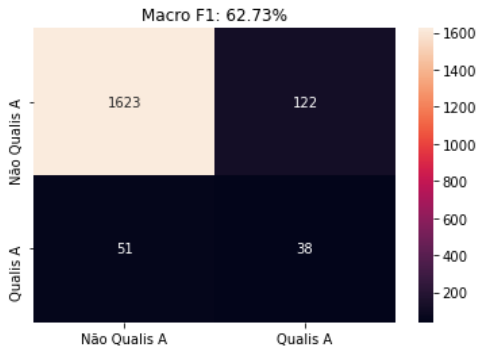


Figura 21 – Resultado com balanceamento proposto

Balanceamento com o método proposto. O método proposto para desbalanceamento (descrito na Seção 4.2.2), consiste em remover exemplos de treino da classe majoritária, desde que os mesmos se encontrem em janelas temporais mais antigas. De forma simples, pode ser vista como uma manipulação da estratégia de empilhamento, onde são mantidos todos os exemplos mais recentes (da última janela temporal) e empilhados apenas exemplos da classe minoritária. A Figura 20 apresenta o efeito do desbalanceamento proposto. Embora ainda haja desbalanceamento, pois são mantidos os exemplos recentes da classe majoritária, a efetividade aumentou substancialmente em relação à estratégia SMOTE (Figura 21)

5.1.1 Sumário dos Resultados com Empilhamento e Desbalanceamento

Em suma, considerando os efeitos do empilhamento e balanceamento são apresentados na Tabela 7. De forma geral, é possível concluir que empilhar os dados e tratar o balanceamento com a estratégia proposta está associado aos melhores resultados de classificação dentre as abordagens avaliadas. Além disso, o resultado apresentado na divisão sem empilhamento se mostrou o pior, dado que não há a inclusão de dados mais antigos, que podem ser informativos para estratégia de classificação. Na divisão com empilhamento, apesar dos novos exemplos de treino mais antigos, é possível observar que não há melhoria significativa dos resultados em relação à divisão sem empilhamento devido ao problema de desbalanceamento.

Tabela 7 – Resultados dos diferentes experimentos executados.

Estratégia	Resultado (MacroF1)
Divisão sem empilhamento (Seção 4.4.1)	53,35%
Divisão com empilhamento (Seção 4.4.2)	54,75%
Divisão com empilhamento+SMOTE	55,65%
Divisão com empilhamento+Proposta de Balanceamento)	62,73%

5.2 Comparação entre Abordagens para Predição de Publicações

Neste trabalho, é feita a comparação entre diferentes estratégias para predição de publicações. Nos experimentos desta seção, além da comparação de abordagens de predição, utilizaremos diferentes combinações de grupos de atributos. A Tabela 8 apresenta a nomenclatura adotada para denotar os atributos utilizados nos experimentos. Com os grupos de atributos definidos na subseção de grupos de atributos, para simplificar cada grupo vai corresponder a um número: 1 - grupo de atributos produções bibliográficas, 2 - grupo de atributos perfil do pesquisador, 3 - atributos demais produções, 4 - atributos atividades vinculadas à pesquisa e 5 - atributos de disciplinas.

Tabela 8 – Grupos de atributos

Letra	**Nome do Grupo**
A	Produções Bibliográficas
B	Perfil do Pesquisador
C	Demais Produções
D	Atividades Pesquisa
E	Disciplinas

5.2.1 Resultados da Regressão

Para fins de comparação da estratégia de Regressão Poisson adotada por (XIE, 2020), assim como no trabalho, foi utilizado apenas o atributo de quantidade de produções

acadêmicas para as predições de futuras publicações. Na Tabela 9, é possível observar que o resultado da regressão Poisson, utilizando apenas um atributo (histórico de publicações), obteve o pior resultado dentre os métodos avaliados. De fato, dada a complexidade dos dados, modelos mais sofisticados, como RF, RNN e LSTM tendem a aprender padrões mais sofisticados, o que os leva a melhores resultados.

Regressão	
Abordagem de predição	**Efetividade (MSE)**
RF	**0.222 (0.031)**
Poisson (XIE, 2020)	0.262 (0.042)
RNN	**0.220 (0.025)**
LSTM (LIU, 2021)	**0.225 (0.027)**

Tabela 9 – Comparação de efetividade (MSE) para a tarefa de predição de número de publicações. Foram utilizados todos os atributos propostos em todos os métodos, com exceção do método da literatura (Poisson (XIE, 2020)), onde utilizamos apenas os atributos utilizados pelo trabalho proposto. As melhores abordagens (em negrito) são estatisticamente superiores com 95% de confiança em relação às duas variações do método Poisson.

Na Tabela 10 podemos ver na comparação entre o Random Forest (RF) com a regressão de Poisson, Rede Neural Recorrente (RNN) e Rede Neural LSTM onde na coluna "Efetividade com todos atributos" vemos o melhor resultado com cada regressor utilizando todas as features: A+B+C+D+E comparado o melhor resultado de cada regressor com determinados grupos de features, coluna: "Efetividade com melhores atributos". Embora os resultados de Redes Neurais como RNN e LSTM estejam com menor MSE: RNN: 0.215 e LSTM: 0.215. comparado com o MSE do RF: 0.217, utilizando o t-test pareado foi comprovado que não há significância estatística entre utilizar algum dos 3, a comparação com o melhor resultado dos 3 ganham apenas se comparado com o resultado de Poisson: 0.242.

Porém outra análise que podemos fazer desses regressores é que se levarmos em consideração o tempo de execução de cada um, Tabela 11, o tempo da abordagem RF é substancialmente inferior ao das demais abordagens.

Regressão			
Abordagem de predição	**Efetividade com todos atributos (MSE)**	**Efetividade com Melhores Atributos (MSE)**	**Melhores Atributos (grupos)**
RF	0.2229 (0.036)	0.217 (0.031)	A+B+D+E
Poisson	0.3516 (0.211)	0.242 (0.04)	A+B+D+E
RNN	0.2205 (0.030)	0.215 (0.025)	A+B+D
LSTM	0.2259 (0.040)	0.215 (0.027)	A+B+D

Tabela 10 – Comparação de efetividade (MSE) dos métodos de predição utilizando todos atributos em relação à efetividade com os melhores atributos selecionados. Apesar das pequenas diferenças entre a segunda e terceira coluna, não houve significância estatística com a seleção de atributos.

Regressão		
Abordagem de predição	**Tempo (segundos)**	**MSE**
RF	4.132	0.217 (0.031)
Poisson (XIE, 2020)	0.041	0.242 (0.04)
RNN	44.41	0.215 (0.025)
LSTM (LIU, 2021)	79.85	0.215 (0.027)

Tabela 11 – Poisson é estatisticamente inferior as demais abordagens apesar de ter o tempo inferior, A diferença entre os resultados do random forest, RNN e LSTM são próximos e sem significância estatística. Entretanto, o tempo de execução da abordagem random forest é substancialmente inferior ao das demais abordagens.

5.3 Importância de Fatores Associados à Produtividade Acadêmica com Projeto Fatorial $2^k r$

A melhor forma de avaliar as interações entre os 5 grupos de features é fazendo a análise de todos eles com 2^k possibilidades de combinações de grupos para cada dataset. Aplicando o experimento com cada possibilidade de combinações (usando a validação cruzada de 5 partes em janelas deslizantes da Seção 4.4.3), nós podemos também avaliar os efeitos de fatores externos que não podem ser controlados. Seguimos a padronização quantitativa chamada 2^k fatorial para analisar os efeitos individuais de grupos de features, também a efetividade produzida pelas suas interações.

O primeiro passo para realizar o 2^k fatorial é definir o fator binário que afeta a variável de resposta (MacroF1 score). No nosso caso, o fator corresponde à presença ou ausência de um grupo de features. Nomeamos a presença de cada grupo de features de acordo com a Tabela 8.

Para cada dataset, foi computado a porcentagem da variação dos resultados que pode ser explicado por cada grupo individual de features considerando cada possível

combinação. Sumarizamos o efeito de cada combinação em diferentes datasets mostrando a média de 5 execuções na Tabela 13.

5.3.1 Resultados da Classificação

Classificação			
Combinação	**Macro F1**	**Combinação**	**Macro F1**
A	0.6211 (0.0318)	A+B+C	0.6091 (0.0315)
B	0.5588 (0.0559)	A+B+D	0.6271 (0.0304)
C	0.4835 (0.0074)	A+B+E	0.6447 (0.0416)
D	0.5805 (0.0215)	A+C+D	0.6139 (0.0264)
E	0.557 (0.0154)	A+C+E	0.6305 (0.0286)
A+B	0.6229 (0.0217)	A+D+E	0.6418 (0.0257)
A+C	0.6264 (0.0317)	B+C+D	0.6046 (0.0188)
A+D	0.6116 (0.0272)	B+C+E	0.6005 (0.0237)
A+E	0.6187 (0.035)	B+D+E	0.6184 (0.0245)
B+C	0.5784 (0.0393)	C+D+E	0.6009 (0.0198)
B+D	0.6007 (0.0263)	A+B+C+D	0.6315 (0.0336)
B+E	0.5931 (0.0205)	A+B+C+E	0.6329 (0.0379)
C+D	0.5925 (0.0191)	A+B+D+E	0.6513 (0.0285)
C+E	0.5645 (0.0176)	A+C+D+E	0.6372 (0.036)
D+E	0.6016 (0.0219)	B+C+D+E	0.6168 (0.0296)
		A+B+C+D+E	0.6432 (0.0261)

Tabela 12 – Resultados da Macro F1 para classificação entre todas as combinações dos grupos de features utilizando Random Forest

De acordo com a Tabela 12 o melhor resultado na classificação entre todas as combinações de grupos de features foi a combinação dos grupos: A+B+E, pois de acordo com a Tabela 13 isso ocorre ao se retirar o grupo de atributos C que não tem grande importância para predição de publicações e pode acabar atrapalhando o aprendizado do modelo.

5.3.2 Resultado Projeto Fatorial

De acordo com o projeto fatorial 13, o fator mais importante é o A - produções bibliográficas anteriores, pois esse fator, de forma isolada, explica 25% da variação dos resultados de predição considerando as demais combinações. Outros fatores também apresentaram alta relevância de forma isolada. O fator D - atividades vinculadas a pesquisa, tem o segundo maior resultado do experimento fatorial, com 8.65%. Como esperado, tal resultado indica que as atividades vinculadas à pesquisa influenciam diretamente na predição de futuras publicações de um pesquisador. O grupo E - Disciplinas, apresentou um resultado de 7.38%, o que nos mostra que as disciplinas que um pesquisador ministra são o terceiro fator mais importante para inferir futuras publicações desse pesquisador. O fator B - Perfil do pesquisador, é o quarto fator mais importante, explicando 6.28% da variação dos resultados do projeto fatorial.

Considerando fatores de forma isolada, apenas o fator C (Demais produções) não apresenta importância para predição de publicações de um pesquisador, pois apresentou apenas 0.45% (sem significância estatística). Isso indica que as demais produções (ex.: músicas, maquetes, lançamentos de CDs) não são suficientemente relevantes para predição publicações acadêmicas.

Além dos fatores individuais, a interação entre fatores também apresentou importância para explicar a predição de publicações. A interação A+D se apresentou como sendo a mais significativa interação entre fatores, explicando 5.66% da variação dos resultados. Tal interação se explica devido ao fato que produções bibliográficas anteriores (fator A) estão relacionadas ao fato do pesquisador estar participando de atividades relacionadas ao grupo de pesquisa como projetos e orientações (fator D).

Da mesma forma, a interação A+B entre o perfil do pesquisador (fator B) e suas publicações anteriores (fator A) também são importantes para predição de futuras publicações, explicando em até 3.36% a variação dos resultados. De fato, como mostrado anteriormente na Figura 7, há uma relação direta entre publicações dos pesquisadores e sua formação dos mesmos.

Outras interações importantes, como A+D+E; A+B+E; A+B+D; A+B+D+E consideram diferentes relações entre publicações anteriores, perfil do pesquisador, atividades de pesquisa e disciplinas. Tais interações, explicam (somadas) 8,33% da variação dos resultados.

Segundo o resultado do projeto fatorial, 25% da variação dos resultados não pode ser explicada (resíduos), indicando a presença de variáveis externas às modeladas pelo experimento.

Tabela Projeto Fatorial	
Combinação	**Porcentagem**
A	**25.31%**
B	**6.28%**
A+B	**3.36%**
C	0.45%
A+C	**0.68%**
B+C	0.45%
A+B+C	0.03%
D	**8.65%**
A+D	**5.66%**
B+D	**1.59%**
A+B+D	**2.73%**
C+D	0.34%
A+C+D	0.38%
B+C+D	0.31%
A+B+C+D	0.01%
E	**7.38%**
A+E	**1.48%**
B+E	**0.77%**
A+B+E	**1.36%**
C+E	0.45%
A+C+E	0.30%
B+C+E	0.13%
A+B+C+E	0.22%
D+E	**1.18%**
A+D+E	**2.39%**
B+D+E	0.47%
A+B+D+E	**1.85%**
C+D+E	0.07%
A+C+D+E	0.32%
B+C+D+E	0.12%
A+B+C+D	0.11%
Resíduos	**25.17%**

Tabela 13 – Resultado do projeto fatorial com todas as combinações de grupos de features. As porcentagens indicam a importância de cada fator e cada combinação de fatores. Os fatores seguem a notação A - produções bibliográficas, B - Perfil do pesquisador, C - Demais produções, D - Atividades vinculadas a pesquisa e E - Disciplinas. Os resultados em negrito apresentaram significância estatística segundo o teste do projeto fatorial 2kr.

6 Conclusões e Trabalhos Futuros

Neste trabalho, foi tratado o problema de previsão de publicações acadêmicas considerando a evolução temporal dos dados, bem como diversas informações, como produções bibliográficas anteriores, perfil do pesquisador, outras produções do pesquisador e disciplinas ministradas pelo pesquisador. Dentre os trabalhos pesquisados, esse é o primeiro a fazer uma análise de predição de publicações considerando todas essas informações em conjunto. Para lidar com o enorme conjunto de dados, a metodologia KDD foi amplamente empregada, onde problemas práticos para o tratamento de dados foram investigados durante todo o desenvolvimento do trabalho.

Este trabalho não é uma metodologia para o IFG em específico mas utilizou o IFG como fonte de dados para desenvolvê-lo, o processo descrito pode ser replicado para outra instituição, contanto que seja utilizado informações sobre pesquisadores assim como mesmo utiliza para viabilizar as predições.

A partir de todas as informações obtidas, foi possível verificar o grande impacto de fatores como publicações anteriores, atividades vinculadas à pesquisa e disciplinas ministradas na importância para explicar a variação da evolução das pesquisas. A predição de publicações com a ferramenta Random Forests se mostrou mais efetiva que a regressão de Poisson, previamente empregada para essa tarefa, e até 10x mais eficiente, em tempo de execução, que Redes Neurais Recorrentes (RNN e LSTM).

O tratamento do desbalanceamento se mostrou importante para o problema de predição de publicações na coleção avaliada, dado o enorme desbalanceamento observado nos dados. Nesse sentido, uma análise mais profunda de outras estratégias de balanceamento em relação à técnica de balanceamento proposta é um caminho futuro de desenvolvimento importante para o tratamento de dados para predição.

No mesmo sentido, seria possível fazer um estudo para verificar se com esses fatores o resultado se manteria avaliando o conjunto de pesquisadores por grandes áreas do conhecimento, assim como grupos de pesquisa pode ser um caminho interessante para futuros trabalhos. Tanto a análise desses diferentes cenários quanto a aplicação de métodos de predição em cada um deles, pode trazer novas informações sobre o comportamento de diferentes grupos de docentes, outra linha a se seguir é fazer a mesma análise em produções técnicas assim como foi feito nas produções bibliográficas pois produções técnicas são muito importantes para um pesquisador e incluir a classificação atual do qualis que agora tem a categoria A3 e A4 que não foram exploradas neste trabalho.

Referências

BHUIYAN, R. A. et al. A robust feature extraction model for human activity characterization using 3-axis accelerometer and gyroscope data. *Sensors*, v. 20, n. 23, 2020. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/20/23/6990>. Citado na página 17.

BREIMAN, L. Random forests. *Machine Learning*, Kluwer Academic Publishers, v. 45, n. 1, p. 5–32, 2001. ISSN 0885-6125. Disponível em: <http://dx.doi.org/10.1023/A%3A1010933404324>. Citado na página 29.

BREIMAN, L.; SPECTOR, P. Submodel selection and evaluation in regression – the x-random case. *International Statistical Review*, v. 60, p. 291–319, 1992. Citado na página 15.

BURMAN, I.; SOM, S. Predicting students academic performance using support vector machine. In: *2019 Amity International Conference on Artificial Intelligence (AICAI)*. [S.l.: s.n.], 2019. p. 756–759. Citado 2 vezes nas páginas 12 e 15.

CASTRO, L. D.; FERRARI, D. *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*. [S.l.: s.n.], 2017. 658 p. ISBN 9788547200985. Citado 5 vezes nas páginas 12, 14, 15, 19 e 25.

CHATFIELD, C. *The analysis of time series: an introduction*. 6th. ed. Florida, US: CRC Press, 2004. Citado 3 vezes nas páginas 12, 13 e 14.

CHAWLA K. W. BOWYER, L. O. H. W. P. K. N. V. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 30, 2002. Citado 2 vezes nas páginas 24 e 25.

CHOLLET, F. *Deep Learning with Python*. [S.l.]: Manning, 2017. ISBN 9781617294433. Citado na página 29.

EDELWEISS, N. *Bancos de Dados Temporais: Teoria e Pratica*. [S.l.], 2001. Citado 2 vezes nas páginas 12 e 13.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>. Citado na página 14.

G. BEN TAIEB S., L. B. Y. B. Machine learning strategies for time series forecasting. *Business Intelligence.*, v. 138, jan 2013. Citado 3 vezes nas páginas 13, 16 e 17.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001. (Springer Series in Statistics). Citado 2 vezes nas páginas 33 e 34.

JAYAPRAKASH, S.; KRISHNAN, S.; JAIGANESH, V. Predicting students academic performance using an improved random forest classifier. In: *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*. [S.l.: s.n.], 2020. p. 238–243. Citado 2 vezes nas páginas 12 e 15.

LAURANCE, W. F. et al. Predicting Publication Success for Biologists. *BioScience*, v. 63, n. 10, p. 817–823, 10 2013. ISSN 0006-3568. Disponível em: <https://doi.org/10.1525/bio.2013.63.10.9>. Citado 4 vezes nas páginas 12, 17, 21 e 28.

LEHMAN, H. *Age and achievement*. [S.l.: s.n.], 2017. 1-359 p. Citado 2 vezes nas páginas 17 e 21.

LIU, R. W. J. Research Performance Prediction for Scientists Using LSTM Neural Networks. In: *Workshop on AI + Informetrics*. [S.l.: s.n.], 2021. Citado 2 vezes nas páginas 39 e 40.

MA, Y.; HAN, R.; FU, X. Stock prediction based on random forest and lstm neural network. In: *2019 19th International Conference on Control, Automation and Systems (ICCAS)*. [S.l.: s.n.], 2019. p. 126–130. Citado 2 vezes nas páginas 17 e 29.

MADSEN, H. *Time Series Analysis*. 1th. ed. [S.l.]: Chapman and Hall/CRC., 2007. Citado na página 13.

SERRA, A. P.; ZáRATE, L. E. Characterization of time series for analyzing of the evolution of time series clusters. *Expert Syst. Appl.*, Pergamon Press, Inc., USA, v. 42, n. 1, p. 596–611, jan 2015. ISSN 0957-4174. Disponível em: <https://doi.org/10.1016/j.eswa.2014.08.012>. Citado na página 12.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. [S.l.]: Pearson Education, 2006. Citado na página 16.

TOMASSINI, M.; LUTHI, L. Empirical analysis of the evolution of a scientific collaboration network. *Physica A: Statistical Mechanics and its Applications*, v. 385, n. 2, p. 750–764, 2007. ISSN 0378-4371. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0378437107007960>. Citado na página 17.

WAY, S. F. et al. The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 114, n. 44, p. E9216–E9223, 2017. ISSN 0027-8424. Disponível em: <https://www.pnas.org/content/114/44/E9216>. Citado 3 vezes nas páginas 8, 12 e 18.

XIE, Z. Predicting publication productivity for researchers: A piecewise poisson model. *Journal of Informetrics*, v. 14, n. 3, p. 101065, 2020. ISSN 1751-1577. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157719302676>. Citado 10 vezes nas páginas 3, 8, 12, 15, 18, 29, 30, 38, 39 e 40.

XIE, Z.; OUYANG, Z.; LI, J. A geometric graph model for coauthorship networks. *Journal of Informetrics*, v. 10, n. 1, p. 299–311, 2016. ISSN 1751-1577. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1751157715300912>. Citado na página 30.