INSTITUTO FEDERAL DO ESPÍRITO SANTO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

JOÃO MARCOS MARETO CALADO

OBSERVATÓRIO AUTOMÁTICO DE EGRESSOS VIA REDES SOCIAIS

Serra
2021

JOÃO MARCOS MARETO CALADO

OBSERVATÓRIO AUTOMÁTICO DE EGRESSOS VIA REDES SOCIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Instituto Federal do Espírito Santo, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Orientadora: Prof[a]. Dr[a]. Karin Satie Komati

Orientador: Prof. Dr. Jefferson Oliveira Andrade

Serra
2021

Dados Internacionais de Catalogação na Publicação (CIP)

C141o 2021

Calado, João Marcos Mareto
Observatório automático de egressos via redes sociais / João Marcos Mareto Calado. - 2021.
78 f.; il.; 30 cm

Orientadora: Profª. Dra. Karin Satie Komati.
Coorientador: Prof. Dr. Jefferson Oliveira Andrade.
Dissertação (mestrado) - Instituto Federal do Espírito Santo, Programa de Pós-graduação em Computação Aplicada, 2021.

1. Aprendizado do computador. 2. Redes sociais on-line. 3. Instituto Federal de Educação, Ciência e Tecnologia do Espírito Santo. Campus Serra - Ex-alunos. I. Komati, Karin Satie. III. Andrade, Jefferson Oliveira. III. Instituto Federal do Espírito Santo. IV. Título.

CDD 006.31

Bibliotecária Rogeria Gomes Belchior - CRB6/ES 417

JOÃO MARCOS MARETO CALADO

OBSERVATÓRIO AUTOMÁTICO DE EGRESSOS VIA REDES SOCIAIS

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Instituto Federal do Espírito Santo, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em 02 de setembro de 2021

COMISSÃO EXAMINADORA

_____

Profª Drª Karin Satie Komati
Instituto Federal do Espírito Santo
Campus Serra

_____

Prof. Dr. Jefferson Oliveira Andrade
Instituto Federal do Espírito Santo
Campus Serra

_____

Prof. Dr. Mateus Conrad Barcellos da Costa
Instituto Federal do Espírito Santo
Campus Serra

_____

Prof. Dr. Luciano de Oliveira Toledo
Instituto Federal do Espírito Santo
Campus Santa Teresa

Emitido em 02/09/2021

FOLHA DE APROVAÇÃO-TCC Nº 4/2021 - CMPCA (11.02.32.01.07.08)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 13/09/2021 17:27)
JEFFERSON OLIVEIRA ANDRADE
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO
SER-CCTI (11.02.32.01.08.02.06)
Matrícula: 1208144

(Assinado digitalmente em 13/09/2021 10:01)
KARIN SATIE KOMATI
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO
CMPCA (11.02.32.01.07.08)
Matrícula: 2324453

(Assinado digitalmente em 13/09/2021 10:10)
LUCIANO DE OLIVEIRA TOLEDO
PRO-REITOR(A) - TITULAR
REI-PRODI (11.02.37.12)
Matrícula: 1545289

(Assinado digitalmente em 13/09/2021 13:55)
MATEUS CONRAD BARCELLOS DA COSTA
PROFESSOR DO ENSINO BASICO TECNICO E TECNOLOGICO
SER-CCSI (11.02.32.01.08.02.04)
Matrícula: 1182338

Para verificar a autenticidade deste documento entre em https://sipac.ifes.edu.br/documentos/ informando seu número: **4**, ano: **2021**, tipo: **FOLHA DE APROVAÇÃO-TCC**, data de emissão: **13/09/2021** e o código de verificação: **b0afb11d4f**

À Marcilene.
Uma mulher brilhante que vive comigo há 7 anos e me atura desde então.

AGRADECIMENTOS

Agradeço à minha família, amigos, professores, e pessoas que ajudaram na realização deste trabalho. Sou imensamente grato pela paciência e incentivo.

Em primeiro lugar agradeço a Deus, por ter me dado, em algum grau, a capacidade de fazer e terminar esta pós-graduação, e por ter me dado a perseverança para não desistir mesmo nos momentos mais difíceis.

Agradeço imensamente minha orientadora, Professora Doutora Karin Satie Komati, que me acompanhou desde o início do curso. Sem sua dedicação e paciência, este trabalho não existiria. Muito brigado, professora.

Agradeço também ao Professor Doutor Jefferson Oliveira Andrade, que como co-orientador, sempre revisou e deu grandes contribuições a este trabalho.

Também agradeço à Banca pela avaliação do trabalho e principalmente pelas contribuições de melhoria.

Não poderia deixar de agradecer a todos os docentes que me deram aula durante o curso, vocês foram importantíssimos nesta conclusão. Sem o conhecimento repassado por vocês, eu não teria condições de sequer iniciar este trabalho.

Quero também mencionar o Jackson, um rapaz muito inteligente que me auxiliou bastante na etapa de programação deste trabalho. Fizemos uma publicação juntos e não poderia deixar de mencioná-lo.

Por fim, mas não menos importante, agradeço a minha família, meus amigos e as pessoas do meu trabalho. Obrigado. Sem dúvidas, se não fosse todo o apoio de vocês, eu não chegaria aqui.

Estudar é o maior ato de rebeldia contra o sistema.

RESUMO

Todas as instituições de ensino devem acompanhar seus egressos para medir a inserção de seus alunos no mercado de trabalho e assim, avaliar se há compatibilidade de sua atuação com a formação recebida, visando subsidiar melhorias em suas matrizes curriculares ou métodos de ensino. Além disso, algumas instituições desejam monitorar se os egressos continuaram a sua vida acadêmica. Tipicamente, este acompanhamento é feito através de pesquisas por meio de questionários enviados aos egressos. Porém, além de ser um processo muito laborioso, frequentemente não atinge o nível de engajamento desejado por parte dos egressos. Nesse contexto, as redes sociais virtuais abrem a possibilidade de usar os dados públicos dos usuários para a pesquisa de egressos de uma instituição de ensino para acompanhamento de sua carreira no mercado ou da continuação dos seus estudos. Porém, diferentes perfis de um mesmo usuário em redes sociais distintas, frequentemente apresentam inconsistências entre campos, tais como não apresentarem o mesmo nome ou o mesmo endereço, o que torna a identificação cruzada difícil. Este trabalho avaliou a viabilidade da construção de um observatório de egressos do Campus Serra do Ifes através da extração automática de dados de redes sociais e pelo uso de modelos de aprendizado de máquina para a identificação cruzada de perfis de egressos do Campus Serra do Ifes. Foram realizados testes preliminares com uma base de dados pública e anotada, a "GT Dataset", com perfis das redes sociais do Google+ e do Twitter. Usando a abordagem de aprendizado de máquina para o problema de ligação de perfis de usuários em diferentes redes sociais, até o momento conseguiu-se replicar, e superar em alguns casos, os resultados relatados na literatura, apresentando acurácia de 0,96 no melhor caso com os classificadores AdaBoost e XGBoost. Após os testes iniciais, foram elaborados uma aplicação Web e um programa coletor. O programa coletor buscou dados do LinkedIn de modo a criar uma base de dados dos egressos do Ifes Campus Serra. A fim de complementar informações e traçar um perfil mais completo, também foram pesquisados currículos Lattes dos egressos. Os resultados mostraram que a metodologia e técnicas empregadas obtiveram sucesso no correto pareamento dos perfis e que este trabalho pode contribuir para a melhora do processo de obtenção de dados dos egressos. A partir dos dados coletados foi possível realizar análises acadêmicas e profissionais dos ex-alunos dos cursos do Ifes Campus Serra.

Palavras-chave: Identificação em sistemas cruzados. Redes sociais. Classificadores. Acompanhamento de egressos. *Web scraping*.

ABSTRACT

All educational institutions must monitor their alumni to assess the insertion of their students into the labor market and thus assess whether their performance is compatible with the training received, aiming to support improvements in their curricula or teaching methods. In addition, some institutions wish to monitor whether alumni continued their academic life. Usually, this monitoring is done through surveys through questionnaires sent to alumni. However, in addition to being a very laborious process, it often does not reach the desired level of engagement on the part of the alumni. In this context, virtual social networks open up the possibility of using the public data of users for the research of alumni of an institution of teaching to monitor both their career in the market and the continuation of their studies. However, different profiles of the same user in different social networks often present inconsistencies between fields, such as not having the same name or the same address, which makes cross-identification difficult. This work evaluated the feasibility of building an observatory for the alumni of the Campus Serra of Ifes by automatically extracting data from social networks and using machine learning models for cross-identification of alumni profiles. Preliminary tests were carried out with a public and annotated database, the "GT Dataset", with profiles from Google+ and Twitter social networks. Using the machine learning approach to the problem of linking user profiles in different social networks, so far, it has been possible to replicate, and even surpass in some cases, the results reported in the literature, with an accuracy of 0.96 in the best case with AdaBoost and XGBoost classifiers. After the initial tests, a web application and a collector program were developed. The collector program sought data from LinkedIn in order to create a database of alumni from the Campus Serra of Ifes. In order to complement information and draw a more complete profile, Lattes curricula of alumni were also queried. The results showed that the methodology and techniques used were successful in correctly matching the profiles and that this work can contribute to the improvement of the process of obtaining data from alumni. From the collected data, it was possible to carry out academic and professional analyzes of the ex-students of the Serra do Ifes Campus courses.

Keywords: Cross-System Personalization. Social Networks. Classifiers. Alumni Monitoring. *Web scraping*.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

A educação superior, definida pela Lei de Diretrizes e Bases da Educação Nacional (BRASIL, 1996), tem como uma de suas finalidades, formar alunos diplomados nas diferentes áreas de conhecimento, de forma que estejam aptos para ingressarem no mercado de trabalho e para que contribuam no desenvolvimento da sociedade brasileira. Além disso, é dito que a educação superior deve colaborar na formação contínua do estudante.

O Sistema Nacional de Avaliação da Educação Superior (Sinaes) tem como objetivo assegurar o processo nacional de avaliação das Instituições de Ensino Superior (IES), dos cursos de graduação e do desempenho acadêmico de seus estudantes. Essa avaliação, tem por finalidade a melhoria da qualidade da educação superior, de forma que oriente a oferta de cursos e a eficácia institucional (BRASIL, 2004). É o Sinaes que define o processo de avaliação, e é o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) que o coordena (BRASIL, 1997). O Inep define o Instrumento de Avaliação Institucional Externa (BRASIL, 2017), que contém os critérios tanto o de credenciamento quanto o de recredenciamento institucional e de cursos. O processo é composto por diversas etapas que, ao final, geram o Conceito Institucional (CI), graduado em cinco níveis variando de 1 a 5, cujos valores iguais ou superiores a três indicam qualidade satisfatória (INEP, 2017).

O instrumento de avaliação externa contém cinco eixos, (1) Planejamento e Avaliação Institucional, (2) Desenvolvimento Institucional, (3) Políticas Acadêmicas, (4) Políticas de Gestão e (5) Infraestrutura. Estes eixos recebem notas de 1 a 5, sendo atribuídos pesos diferentes para cada um dos cinco eixos no cálculo utilizado para obtenção do CI, para o credenciamento e para o recredenciamento: os eixos 1 e 3 possuem peso 10, enquanto os eixos 2 e 5 têm peso 30 e o eixo 4 possui peso 20 (INEP, 2017). No contexto de avaliação da educação superior, este trabalho visa o indicador 3.7 do Eixo 3 de Políticas Acadêmicas da IES. Este item pontua a instituição conforme suas ações referentes à política institucional de acompanhamento dos egressos, a seguir listam-se os conceitos (nota) e sua descrição associada, os grifos estão iguais ao texto original.

- Conceito 1: **não** há política institucional de acompanhamento dos egressos.
- Conceito 2: a política institucional **não** garante mecanismo de acompanhamento de egressos.
- Conceito 3: a política institucional **garante** mecanismo de acompanhamento de egressos **e** a atualização sistemática de informações a respeito da continuidade na vida acadêmica ou da inserção profissional.
- Conceito 4: a política institucional **garante** mecanismo de acompanhamento de egressos, a atualização sistemática de informações a respeito da continuidade na vida acadêmica

ou da inserção profissional **e** estudo comparativo entre a atuação do egresso e a formação recebida, **subsidiando** ações de melhoria relacionadas às demandas da sociedade e do mundo do trabalho.

- Conceito 5: a política institucional **garante** mecanismo de acompanhamento de egressos, a atualização sistemática de informações a respeito da continuidade na vida acadêmica ou da inserção profissional, estudo comparativo entre a atuação do egresso e a formação recebida, **subsidiando** ações de melhoria relacionadas às demandas da sociedade e do mundo do trabalho, e **promove** outras ações reconhecidamente exitosas ou inovadoras.

Muitas IES usam como processo de trabalho, ligação telefônica direta ao egresso ou a solicitação aos egressos para que respondam à um formulário *online* com perguntas sobre a continuidade na vida acadêmica além de informações profissionais, processo este, que foi realizado no Ifes em 2016 para os egressos dos cursos técnicos (IFES, 2016c), e o resultado está público (IFES, 2016a). No entanto, somente uma pequena parte dos egressos responderam, e com isso nenhum dos campi conseguiu obter o alcance de 95% de confiança da pesquisa[1], pois não alcançaram o quantitativo mínimo de respostas em relação ao número de egressos totais. O campus Barra de São Francisco, por exemplo, teve apenas 5 respostas, com um cálculo de mínimo de 66 respostas para taxa de 95% de confiança de um total de 79 egressos. Na época, o campus Serra contava com 1.635 egressos de nível técnico com amostra mínima de 311, mas apenas 43 egressos responderam ao questionário, gerando nível de confiança atualizado em 49,4%.

A hipótese deste trabalho é que um sistema de coleta de informações que seja independente da interação com o usuário possa aumentar a taxa de amostragem. Se esta hipótese se confirmar, isso significará um aumento na confiança estatística em relação ao modelo atual de aplicação de questionário.

1.1 A PROPOSTA

Para sistematizar os dados do acompanhamento de egressos, a proposta deste trabalho é que a coleta de dados seja feita de forma automática via redes sociais utilizando técnicas de identificação e associação de perfis, alcançando o Conceito 3 do indicador de acompanhamento de egressos. Relembrando que o Conceito 3 é que a política institucional garanta mecanismo de acompanhamento de egressos e a atualização sistemática de informações a respeito da continuidade na vida acadêmica ou da inserção profissional.

Uma das dificuldades da solução é conseguir identificar informações correspondentes à pessoa que está sendo buscada dada a quantidade de homônimos encontrados em redes

[1] O índice de nível de confiança representa a probabilidade de uma pesquisa ter os mesmos resultados se for aplicada com um outro grupo de pessoas, dentro do mesmo perfil de amostra e com a mesma margem de erro.

sociais. Este é um problema que possui várias diferentes denominações, tais como, a identificação de usuário em sistemas cruzados (em inglês *Cross-system Personalisation*) (CARMAGNOLA; CENA, 2009) (ESFANDYARI et al., 2018), ou como *Social Link Identification* (ZHANG et al., 2019) (SHU et al., 2017) ou *Disambiguate Identity References* (ROWE, 2009) (DIGIAMPIETRI; LINDEN; BARBOSA, 2015), dentre outros.

O problema de identificação de usuário em sistemas cruzados é ilustrado na Figura 1, que apresenta os diferentes campos correspondentes entre dois perfis da mesma pessoa em dois *sites* diferentes: do Twitter e da plataforma Lattes. Na figura é possível verificar que existem dados em comum em posições diferentes, e provavelmente com nomes de campos diferentes entre os sistemas[2]. Os campos correspondentes estão marcados por retângulos unidos por linhas entre a página à esquerda e à direita, que são o nome, a descrição da graduação e o estado-país de residência. É importante reiterar que o nome completo está idêntico, mas em muitas redes sociais, as pessoas escolhem por indicar nomes com abreviações ou partes do nome ou mesmo um nome de usuário totalmente diferente do nome real.

Figura 1 – Dois perfis da mesma pessoa em páginas diferentes, Twitter e plataforma Lattes, evidenciando informações comuns entre os diferentes perfis.

Fonte: Elaborado pelo autor (2020)

De acordo com Esfandyari et al. (2018), a identificação de usuário em sistemas cruzados é um problema difícil de ser resolvido dada a não estruturação das informações, além da falta

[2] A figura é meramente ilustrativa, assim, poderiam ser outras páginas que não a do Lattes e a do Twitter.

de garantia na veracidade das informações preenchidas. Corroborando com Esfandyari et al. (2018), Shu et al. (2017) complementa as dificuldades e desafios da tarefa de identificar os perfis em diferentes redes sociais. São elencados dois motivos pelos quais existe essa dificuldade. O primeiro é que embora usuários tenham contas em diferentes redes, a informação de uma mesma pessoa no mundo real pode ser desigual entre as redes, e o segundo motivo é que as informações da identidade dos usuários é ruidosa, incompleta e altamente não estruturada.

Para tentar resolver esses desafios, diversas pesquisas vêm sendo realizadas utilizando abordagens diferentes (SHU et al., 2017). As abordagens podem ser agrupadas (ESFANDYARI et al., 2018) em (i) identificação baseada em nome de usuários, (ii) identificação baseada em perfis de usuários e a (iii) identificação baseada em conteúdo e rede de amigos. Este trabalho utilizará a abordagem de identificação baseada em perfis, que consiste em selecionar um determinado conjunto de atributos que compõe o perfil do usuário da rede social, tendo em vista, serem as técnicas cujos resultados se mostraram mais promissores.

O problema de identificação de usuário em sistemas cruzados, é definido como: dados dois perfis P^{s1} e P^{s2} de duas redes sociais diferentes $s1$ e $s2$, é possível especificar se pertencem à mesma pessoa. Isto corresponde à aprender uma função de identificação $f(P^{s1}, P^{s2})$, tal que:

$$f(P^{s1}, P^{s2}) = \begin{cases} 1 & \text{se } P^{s1} \text{ e } P^{s2}\text{, pertencem à mesma pessoa} \\ 0 & \text{caso contrário} \end{cases} \tag{1}$$

Para realizar a identificação será utilizada a técnica de classificação, composta de duas fases principais, a saber: (i) extração de características; e (ii) construção de modelo de classificação (SHU et al., 2017). Neste tipo de problema definido pela Equação 1, a extração de características se diferencia da extração de características de um texto ou de uma imagem, em que o vetor de características de uma única amostra é repassado diretamente ao classificador. Neste caso, os perfis devem ser "pareados" através de métricas de distância ou similaridade, sendo criado um vetor com estas métricas a partir de dois perfis. É este vetor de similaridade/distância entre os campos de dois perfis que será a entrada para algoritmos de aprendizado de máquina, tanto no treinamento quanto na classificação. Na abordagem de aprendizado de máquina, na fase de treinamento, cada vetor de características transformado terá um rótulo "positivo" ou "negativo", indicando respectivamente se o par de perfis correspondem à mesma pessoa (instância "positiva") ou não (instância "negativo"). Na fase de testes, o modelo do classificador responderá positivo ou negativo.

A avaliação desta proposta é dividida em dois experimentos: a comparação entre técnicas

de aprendizado de máquina e uma solução de acompanhamento de egressos para o Ifes. A metodologia será avaliar modelos de classificadores para a identificação de usuário em uma base de dados pública e anotada, e aplicar os modelos em dados reais coletados de egressos do estudo de caso, do Campus Serra do Ifes.

1.1.1 Experimento 1: Comparação entre técnicas de aprendizado de máquina

Para avaliar os modelos de classificação para este problema, comparamos 8 (oito) algoritmos de classificação:

- Regressão Logística (LI; JAIN, 1998);
- LDA (do inglês *Linear Discriminant Analysis*) (SCHÜTZE; HULL; PEDERSEN, 1995);
- Naïve Bayes (RISH et al., 2001);
- KNN (do inglês *k-nearest neighbors*) (COVER; HART, 1967);
- Árvore de decisão (SAFAVIAN; LANDGREBE, 1991);
- SVM (do inglês *Support Vector Machine*) (BOSER; GUYON; VAPNIK, 1992);
- AdaBoost (FREUND; SCHAPIRE; ABE, 1999); e
- XGBoost (CHEN; GUESTRIN, 2016).

A avaliação de desempenho é baseada em métricas de precisão, sensibilidade e acurácia. Será usado um conjunto de características a partir do perfil dos usuários proposto por Esfandyari e colegas (ESFANDYARI et al., 2018), que após extraídas, serão usadas como conjunto de entrada para os classificadores, além da mesma base de dados denominada "GT dataset", que contém 10.571 pares de perfis corretamente rotulados de usuários do *Google+* e *Twitter*. A diferenciação entre este trabalho e o de Esfandyari e colegas (ESFANDYARI et al., 2018) é que aquele trabalho comparava os resultados dos classificadores Floresta Aleatória e MLP (*Multilayer Perceptron*), enquanto este trabalho estende e faz a comparação usando outros oito algoritmos de classificação já mencionados.

1.1.2 Experimento 2: Estudo de caso do campus Serra do Ifes

O Instituto Federal do Espírito Santo (Ifes) é uma Instituição de Ensino Superior (IES) (BRASIL, 2008) e portanto deve obedecer à legislação vigente, implementando as formas de avaliação obrigatórios definidas pelo Sinaes, para que seus cursos tenham o credenciamento ou recredenciamento. Deste modo, faz-se necessário acompanhar os discentes diplomados, se estão empregados no mercado de trabalho, se deram continuidade aos estudos, optando pela carreira acadêmica, ou ainda se o currículo ministrado pela IES está compatível com as expectativas do mercado de trabalho.

É um tema tão importante que o Ifes definiu em seu Plano Diretor Institucional vigente, o observatório de egressos como um dos oito projetos prioritários (IFES, 2014). Dessa forma, além de ser uma demanda externa para o recredenciamento, é uma demanda da própria instituição, conhecer o público formado, conforme pode ser visto na justificativa do projeto.

> "Faz-se necessário compreender a inserção do egresso da instituição na dinâmica social, econômica, geográfica, ambiental e cultural dos municípios, mesorregiões e microrregiões que compõem área de influência do Ifes, com o objetivo de cumprir a sua missão institucional de contribuir para o desenvolvimento sustentável do Estado. Neste contexto, a implantação do Acompanhamento de Egressos tem importância relevante, pois as informações disponibilizadas serão utilizadas para avaliar o impacto das políticas institucionais em educação profissional e tecnológica e amparar a proposição de estratégias e a tomada de decisões para a melhoria da eficácia e efetividade dos programas e projetos de educação profissional e tecnológica do Ifes (IFES, 2016b)."

Assim, planeja-se usar o mesmo processo de extração de características e técnicas de aprendizado de máquina do Experimento 1 e aplicá-los em uma base de dados de egressos do Ifes Campus Serra. A partir da lista de egressos extraída do sistema acadêmico do Ifes, foi feita a coleta automática de dados da rede social LinkedIn. Via informações do LinkedIn buscou-se extrair a sua situação empregatícia atual, considerando a empresa na qual a pessoa trabalha, o cargo que ocupa, quanto tempo está na mesma empresa e a posição geográfica da instituição na qual trabalha. A seleção da rede social LinkedIn, é devido ao fato de ser a maior rede social voltada ao mercado profissional (HERRMAN, 2019), e espera-se com isso um elevado grau de seriedade de preenchimento das informações.

O resultado do LinkedIn será complementado pelos dados coletados da plataforma Lattes. No entanto, diferente do LinkedIn, a busca na plataforma Lattes será feita pelo CPF. Serão extraídas informações a respeito de continuidade de estudos e produção acadêmica. No sistema Lattes também é possível extrair a informação de vínculo empregatício atual, com indicação de tempo de trabalho.

A escolha pelo Campus Serra se deve ao fato de se ter relatórios recentes de acompanhamento de egressos em 2019[3] e 2020[4]. O que possibilita comparar os resultados automáticos com o modo de operação manual de coleta de dados.

1.2 LIMITAÇÕES DA PROPOSTA

Listam-se algumas limitações da proposta deste trabalho:

[3] https://www.serra.ifes.edu.br/noticias/campus-serra-divulga-pesquisa-com-egressos
[4] https://www.serra.ifes.edu.br/noticias/campus-serra-divulga-pesquisa-de-egressos-realizada-em-2020

- De acordo com o Sinaes, o egresso é todo discente que tenha frequentado algum curso na IES tendo concluído ou não os estudos (INEP, 2017). Porém, neste trabalho serão considerados apenas os egressos que foram diplomados, isto é, aqueles que concluíram todas as etapas do curso e obtiveram o diploma, que é o processo já realizado pelo Ifes.
- Ao mesmo tempo em que espera-se mapear uma quantidade maior de egressos, há a desvantagem de que nesta abordagem perde-se a percepção/opinião do egresso em relação a satisfação com o curso feito no Ifes, levando em conta suas perspectivas e expectativas.
- Por ser um sistema automático, o mesmo poderá ser executado periodicamente para extração de informações atualizadas. Mas a confiabilidade das informações varia de acordo com a rede social usada.

1.3 OBJETIVO GERAL

Este trabalho tem por objetivo avaliar a viabilidade da construção de um observatório de egressos do Campus Serra do Ifes por meio da extração automática de dados de redes sociais.

1.3.1 Objetivos Específicos

Os objetivos específicos identificados para se atingir o objetivo geral proposto são:

- Construção de uma base de dados com os perfis coletados da rede social LinkedIn.
- Construção de uma ferramenta de extração de dados que permita a criação da base de dados de perfis da rede social LinkedIn à partir de uma lista de egressos.
- Pesquisar e selecionar as métricas de distância/similaridade de textos para extração de características;
- Implementar e comparar classificadores para identificação de pares correlatos de perfis de egressos.
- Caracterizar o perfil atual dos egressos do Ifes:
  - Desenvolver metodologias de extração de informações profissionais dos egressos à partir dos perfis de redes sociais. Assim, retratar a atual situação de inserção profissional dos egressos.
  - Desenvolver metodologias de extração de informações de formação acadêmica dos egressos à partir de informações de redes sociais. Assim, identificar se houve continuidade dos estudos.
  - Identificar a mudança de estado/país do local de trabalho de cada egresso.

- Construir mecanismos de complementação dos dados extraídos de redes sociais à partir de outros perfis públicos dos egressos na internet, e.g., Currículo Lattes.

1.4 CONTRIBUIÇÕES DO TRABALHO

Como contribuições, dois artigos sobre a pesquisa foram publicados, sendo um deles na Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação (REABTIC), publicado na primeira edição de 2021 (CALADO; ANDRADE; KOMATI, 2021). O artigo da revista REABTIC é uma contribuição incremental ao trabalho produzido por Esfandyari et al. (2018), documentando o Experimento 1, com a comparação de mais algoritmos classificadores na tarefa de ligação de perfis de usuários em redes sociais.

O segundo artigo foi publicado nos anais do XXVII Simpósio de Engenharia de Produção (SIMPEP 2020), intitulado "Um sistema de acompanhamento de egressos usando dados do site Escavador" (CALADO et al., 2020). No artigo em questão é elaborado um sistema capaz de obter informações de alunos egressos de instituições de ensino a partir de uma lista de perfis, buscando as informações no site Escavador. O desenvolvimento de tal sistema foi um passo intermediário para o módulo de coleta de dados do LinkedIn.

1.5 ORGANIZAÇÃO DO TRABALHO

Este texto está organizado da seguinte maneira: no Capítulo 2 são apresentados trabalhos que abordam as diferentes técnicas já apresentadas a respeito da identificação cruzada de perfis de usuários em redes sociais. Além desses, são apresentados alguns trabalhos a respeito do acompanhamento de egressos, com destaque para dois trabalhos nacionais.

No Capítulo 3 é apresentada a metodologia do trabalho, consistindo na apresentação e descrição da base de dados utilizada para estudo e replicação da técnica de pareamento de perfis, das técnicas de extrações de características selecionadas e dos métodos de classificação supervisionada de aprendizado de máquina que serão utilizados. Também é descrito o módulo coletor para construção de uma base de dados de egressos do Ifes campus Serra e aplicação das técnicas e algoritmos para identificação dos perfis coletados. Além da descrição dos algoritmos de classificação e das métricas de avaliação.

No Capítulo 4 são apresentados os resultados dos experimentos e a complementação de informações com a plataforma Lattes. Além disso, é feita uma comparação do método proposto para coleta e identificação dos perfis de egressos, com as pesquisas de egressos conduzidas pelo Ifes campus Serra nos anos de 2019 e 2020. No Capítulo 5 são discutidas as conclusões e possibilidades de trabalhos futuros.

2 REVISÃO BIBLIOGRÁFICA

As redes sociais vêm sendo objeto de estudos desde bem antes do advento das redes sociais digitais, como visto em Wasserman e Galaskiewicz (1994). O autor ainda cita questões como influência, popularidade, interação entre pessoas, e mesmo propagação de doenças já eram estudados como fenômenos sociais desde antes do surgimento da Internet.

No ano de 1997 foi lançado o SixDegrees.com, o primeiro *site* do tipo rede social, conforme informado no trabalho de Boyd e Ellison (2007). Porém, foi somente a partir de 2003 que os maiores *sites* em termos de quantidade de usuários começaram a ser lançados, como Orkut, Linkedin e MySpace em 2003, Youtube em 2005 e Facebook aberto a todos em 2006. O Orkut, apesar de não estar mais ativo, até 2011, era o maior site de rede social no Brasil, ultrapassando 30 milhões de usuários ativos. Dados de outubro de 2018 mostram que as maiores redes sociais eram o Facebook com 2,6 bilhões de usuários em suas plataformas; o YouTube com 1,9 bilhões e o Instagram com 1 bilhão de pessoas (VALENTE, 2018). Dados atualizados de 2021 retirados do site Statista (https://www.statista.com/), mostram que o facebook ganhou cerca de 290 milhões de usuários, passando a ter 2.89 bilhões de usuários ativos mensais, enquanto o Youtube atingiu a marca de 2.1 bilhões de usuários e o Instagram permaneceu estagnado em 1 bilhão de usuários.

Veldman (VELDMAN, 2009) evidencia que os usuários costumam criar contas em diversas redes sociais online e passar bastante tempo nestas atividades. Ter a capacidade de ligar os perfis entre diversas bases de dados poderia levar a um maior entendimento sobre o comportamento e costume dos usuários, permitindo a melhora na provisão e customização de serviços, além de recomendações melhores (CARMAGNOLA; CENA, 2009).

Uma das dificuldades encontradas em sites de redes sociais é conseguir identificar a pessoa que buscamos dado a quantidade de homônimos existentes. A identificação de que dois perfis em diferentes redes sociais pertencem à mesma pessoa no mundo real é um problema difícil dada a não estruturação das informações, além da falta de garantia na veracidade das informações preenchidas (ESFANDYARI et al., 2018). São elencados dois motivos pelos quais existe esse desafio (SHU et al., 2017): o primeiro é que embora usuários tenham contas em diferentes redes, a informação de uma mesma pessoa no mundo real pode ser diversa entre as redes, e o segundo motivo é que as informações da identidade dos usuários é ruidosa, incompleta e altamente não estruturada.

Tanto Shu et al. (2017) quanto Esfandyari et al. (2018) conduziram seus trabalhos de forma parecida, começando pela definição do conjunto de características e posterior extração a partir de uma base de dados. Posteriormente as características extraídas são utilizadas como entrada para a etapa de construção de um modelo aplicando os algoritmos de classificação. Finalmente, o modelo treinado é aplicado à uma base de dados real com a

finalidade de prever se dois perfis de usuários correspondem entre si.

O trabalho conduzido por Shu et al. (2017) detalha a extração de característica dos usuários em três formas a saber: (i) características de perfil, (ii) características de conteúdo e (iii) características de rede de amizades. Esfandyari et al. (2018) por sua vez, agrupa as abordagens da seguinte forma: (i) baseada em nome do usuário, baseada em atributos de perfil e, (iii) baseada em conteúdo e rede de amizades. Já Deng et al. (2019) dividem as soluções em (i) baseada em atributos de perfil, (ii) baseada em estrutura de redes de relacionamentos, (iii) baseada na geração de dados, como posts.

Nas próximas subseções serão detalhadas as diferentes abordagens segundo Esfandyari et al. (2018), as baseadas em nome de usuário e baseadas em atributos de perfil, que serão utilizadas nesse trabalho. Ao final, há ainda uma seção sobre trabalhos de acompanhamento de egressos.

2.1 IDENTIFICAÇÃO BASEADA EM NOME

Na identificação baseada em nome de usuário, as soluções levam em consideração apenas o nome de usuário. Consequentemente, os diferentes métodos dependem apenas das características extraídas das *strings* que compõem os nomes.

O trabalho de Zafarani e Liu (2009) demonstrou a possibilidade de identificar perfis correspondentes em doze comunidades, utilizando nomes de usuários e um motor de busca, no caso, o Google. A técnica começa pesquisando pelo nome de um determinado usuário no Google tentando encontrar um conjunto de palavras-chave que podem representar possíveis nomes de usuários na redes sociais alvo. Então este conjunto é estendido, adicionando ou removendo prefixos e sufixos comuns de seus membros. O método propõe considerar uma das doze redes sociais como base e a partir dessa definição, pesquisar os nomes dos usuários dessa rede nas outras onze redes alvo. Esse processo foi repetido até que todas as doze redes sociais alvos fossem consideradas como base da pesquisa. Os resultados mostraram que na média, o método proposto conseguiu desempenho de 66% de acerto, enquanto o melhor cenário foi de 92% de acerto.

Perito et al. (2011) também exploraram a possibilidade de ligar perfis de usuários apenas olhando seus nomes de usuários. Para isto, eles estimaram o quão único é um nome utilizando teoria de modelo de linguagem e Cadeias de Markov. Para cada nome, o classificador checa todos os possíveis nomes numa lista de similaridades, o que torna essa abordagem difícil de usar em larga escala.

O trabalho de Zafarani e Liu (2013) propôs uma metodologia denominada “Modeling Behavior for Identifying Users” (MOBIUS) baseada nos padrões comportamentais dos usuários quando estes selecionam seus nomes de usuários, que podem ser cadeias de

caracteres alfanuméricas ou *e-mails*, e demonstraram que o ambiente, a personalidade e as limitações humanas resultam em escolhas de nomes redundantes.

A técnica MOBIUS superou os trabalhos anteriores e três métodos mais comuns: *Exact Username Match* (comparação exata do nome do usuário); *Substring Matching* (casamento de *substring*) e *Patterns in Letters* (padrões de letras). O Quadro 1 apresenta os resultados de precisão do experimento. Para o experimento, foi construída uma base de dados contendo nomes de usuários de 32 sites diferentes, tais como Flickr, Reddit, StumbleUpon e Youtube.

Quadro 1 – Performance da metodologia MOBIUS

Técnica	**Precisão**
MOBIUS (Naive Bayes)	91,38%
Método de Zafarani et al (2009).	66,00%
Método de Perito et al (2011).	77,59%
Método b1: *Exact Username Match*	77,00%
Método b2: *Substring Matching*	63,12%
Método b3: *Patterns in Letters*	49,25%

Fonte: Zafarani e Liu (2013).

No trabalho de Li e outros (LI et al., 2017), é mostrada uma solução para composição de uma série de características extraídas dos nomes de usuários. Para os experimentos foram utilizadas três bases de dados, sendo uma com perfis do Facebook-Twitter com 67.826 perfis, outra do Facebook-Foursquare com 288 480 perfis e outra do Foursquare-Twitter com 102 315 perfis. O trabalho utilizou sete classificadores: Gaussian Naïve Bayes, Bernoulli Naïve Bayes, regressão logística, regressão logística com validação cruzada embarcada, SVM, árvore de decisão e floresta aleatória, via algoritmos da biblioteca scikit-learn com parametrização padrão (PEDREGOSA et al., 2011). A solução proposta apresentou medida-F1 atingindo 96,24 %, 92,49 %, e 90,68 % em três conjuntos de dados reais diferentes, respectivamente.

2.2 IDENTIFICAÇÃO BASEADA EM PERFIL

A identificação baseada em atributos de perfil leva em consideração o conjunto de atributos do perfil de usuários disponíveis publicamente nos sites de redes sociais. Características de perfil podem ser utilizadas de diferentes maneiras de modo a decidir se duas identidades virtuais pertencem à mesma pessoa do mundo real. As abordagens empregadas podem ser categorizadas em: baseadas em distância e baseadas em frequência (SHU et al., 2017).

Para os métodos que consideram a distância, a similaridade entre campos de perfil de duas identidades de usuários podem ser medidas comparando a distância entre eles. Essa "distância" pode ser medida por meio de métricas como *Jaro-Winkler distance*, *Jaccard similarity* e *Levenshtein distance* (SHU et al., 2017). Essas métricas são conhecidas como

medidas de similaridade e, de acordo com Witten e Frank (2005), informam a distância entre diferentes valores para os atributos de determinada entidade. Essa distância é então utilizada como indicativo da semelhança entre os valores. Os trabalhos envolvendo ligação de perfis de redes sociais lidam basicamente com dados que são cadeia de caracteres ou nominais, que por sua vez não possuem uma métrica implícita para indicar a semelhança. Deste modo, a utilização de medidas de similaridade se tornam indicadas para campos com este tipo de dado (WITTEN; FRANK, 2005).

Para os métodos que consideram a frequência, ao invés de calcular a distância entre os valores de cada atributo dos perfis envolvidos no processo de identificação, deve-se investigar o padrão de frequência. Neste modelo, o texto é separado utilizando a técnica *bag of words* e/ou *TF-IDF* (SHU et al., 2017).

Alguns autores, no entanto, abordaram este problema de uma forma diferente. Carmagnola e Cena (2009), por exemplo, realizaram análises definindo um conjunto de propriedades dos perfis, e fatores de importância específicos para cada propriedade do perfil. O algoritmo proposto por eles, compara os atributos de perfil levando em consideração o fator de importância. Apesar do teste ter sido feito numa base de dados com 80 usuários, os resultados foram bem sucedidos, com 59 de um total de 64 casos sendo corretamente identificados, e 2 de um total de 16 casos sendo incorretamente marcados como identificados, quando na verdade deveria ser 16 de um total de 16 casos marcados como não identificados.

Vosecky, Hong e Shen (2009) usaram uma abordagem baseada no estabelecimento de valores que indiquem pesos para os campos dos perfis envolvidos na comparação. Dessa forma, é possível controlar a influência que cada propriedade tem no processo de classificação de similaridade dos perfis, e para comparar os atributos, foram utilizadas técnicas como *exact*, *partial* e *fuzzy match*. A base de dados envolveu mil usuários do Facebook e do StudiVZ, e ao final do processo, os autores conseguiram uma taxa de sucesso de 83%.

2.3 TRABALHOS SOBRE ACOMPANHAMENTO DE EGRESSOS

Na literatura, são encontrados trabalhos a respeito do acompanhamento de egressos usando sistemas com diferentes níveis de automação. A primeira subseção trata de trabalhos bem similares à proposta deste trabalho, em que a coleta independe de interação com o egresso e os dados são coletados da rede social LinkedIn. Na segunda subseção, a coleta dos dados dos egressos tem o propósito de avaliar o ranqueamento da universidade e não o perfil dos egressos. A terceira subseção apresenta trabalhos que versam sobre a interação dos egressos com a instituição. Um desses trabalhos solicita que o egresso indique o endereço do perfil na rede LinkedIn. Deve-se considerar que essa é uma solução que facilita a coleta de dados, mas que é temporária, pois redes sociais são dinâmicas, podendo ser desativadas a qualquer momento. Assim, uma nova rede social de perfil profissional pode surgir, bem

como o LinkedIn pode vir a ser desativado. A proposta deste trabalho é mais genérica e poderia ser aplicada à outras redes sociais.

2.3.1 Coleta de egressos via LinkedIn

Nesta subseção, serão descritos de forma mais detalhada sobre os trabalhos de Gonçalves et al. (2014) e Almeida (2018), pois são a respeito de acompanhamento de egressos coletados da rede LinkedIn. Ambos os trabalhos se concentram na coleta e na análise dos perfis, mas não usam de técnicas de inteligência artificial para determinar de fato o perfil correto do egresso.

No trabalho de Gonçalves et al. (2014), é proposto um sistema cuja arquitetura é composta de três módulos a saber: (i) Buscador; (ii) Filtro; (iii) Extração. Na metodologia proposta, o módulo Buscador utiliza a API do Google de nome CSE (*Custom Search Engine*) para encontrar as informações e conteúdo na rede social LinkedIn. Foram usadas 5 listas de egressos como entrada: 1.542 egressos do curso de Ciência da Computação da UFMG, 1.579 do curso de engenharia metalúrgica da UFOP, 1.259 do curso de Ciência da Computação da USP, 900 do curso de química da USP e 812 do curso de Ciência da Computação da PUC-PR. Para cada egresso da lista, o programa gera uma combinação de nomes possíveis, pois os egressos podem não usar o nome completo ao se registrar nas redes sociais.

O módulo denominado Filtro tem o objetivo de determinar a significância de uma página candidata recuperada pelo módulo Buscador, filtrando as páginas que julgar não pertencer a um egresso de fato. Os atributos são separados em três grupos: nome do curso de graduação, instituição e titulação. A similaridade é calculada pela função de similaridade de cosseno. Escolhe-se a página candidata de menor valor da função de similaridade. É feita uma comparação com o resultado da técnica Naive Bayes, avaliando que o método proposto de menor valor de similaridade por cosseno é o suficiente no estudo de caso.

Por fim, o módulo Extração é responsável por recuperar informações acadêmicas, profissionais e pessoais a partir dos perfis filtrados pelo módulo Filtro. Esta extração é baseada na construção de expressões regulares manualmente definidas. A partir da extração, essas informações são armazenadas num banco de dados para posterior análise. O trabalho não faz uso de classificadores ou de heurísticas para determinar a relevância de uma página ser ou não de um egresso.

Ao final dos experimentos, os autores relatam ter conseguido obter em média 7.5% de egressos em programas com mais de 1000 egressos. Além desse resultado, citam um caso específico de obtenção de 12,2% dos egressos para o programa de Ciência da Computação da USP, comparando esses resultados contra uma taxa de 6% para utilização de métodos convencionais onde os egressos precisam responder questionários ou entrar em contato com os programas. Também foi realizada uma análise de localização da empresa na qual o

egresso trabalha correntemente e perceberam uma alta concentração de egressos na cidade de origem de seu curso de graduação, com exceção dos egressos em Ciência da Computação da USP.

O trabalho de Almeida (2018), por sua vez, apresenta uma ferramenta para coleta de dados de perfis de usuários da rede social LinkedIn. Concentrando no desenvolvimento de um robô como solução para capturar dados, utiliza a ferramenta *Selenium*, um programa comumente utilizado para testes de interface, contorna o problema de autenticação e automatiza o processo de coleta de dados a partir do site do LinkedIn.

No caso, o *Selenium* foi utilizado devido a capacidade de carregar uma janela do navegador e preencher formulários e fazer a navegação Web de forma autônoma. Com a utilização do *Selenium*, o autor construiu um programa robô que utiliza o próprio site do LinkedIn para realizar as buscas. A partir da coleta do robô, o autor armazena informações pessoais, profissionais e acadêmicas em tabelas específicas num banco de dados. Almeida (2018) ainda incluiu a extração de dados dos perfis do sistema de currículo Lattes para compor a base de dados, mas não foram tratadas as incoerências entre as informações do Lattes com o LinkedIn.

No trabalho, o autor construiu um módulo Web que permite ao usuário informar um nome e as informações são recuperadas do banco de dados e exibidas na forma de linha de tempo. O estudo de caso não recebe como entrada uma lista específica referente aos egressos. A lista inicial foi composta de 1.813 nomes aleatórios simples (tais como José, Maria e Gustavo) e a busca foi por perfis já haviam apontado ligação com o perfil da PUC-Rio do LinkedIn ou que estavam com um nível de conexão com algum perfil já ligado à este perfil da PUC-Rio. Assim, não foram usadas técnicas de inteligência artificial para determinação do correto perfil do egresso, pois não havia uma lista específica. O resultado da coleta consistiu em 57.901 perfis do LinkedIn. A partir da coleta dos perfis e do respectivo processamento, o trabalho responde uma série de perguntas tais como:

- Quais são os cursos e/ou áreas mais procurados da PUC-Rio? A resposta foram os cursos de economia e administração.
- Quais empresas mais contratam da PUC-Rio? A resposta foi a PUC-Rio e a Infotec Petrobrás.
- Qual é a média do tempo gasto em graduação da PUC-Rio? A resposta foi de 4 anos.
- Qual o tempo médio de um emprego por empresa da PUC-Rio? A resposta foi de 4,82 anos.
- Quanto tempo leva um alumnus de graduação PUC-Rio a assumir um cargo de relevância? A resposta foi de um tempo médio de 9,3 anos.

2.3.2 Ranqueamento de universidades

Além dos trabalhos citados anteriormente que retratam a coleta de perfis de egressos, há outros trabalhos que abordam questões diferentes, como formas de se ranquear universidades utilizando informações diversas a respeito de seus alunos egressos. São detalhados os trabalhos de (KOZITSINA et al., 2020) e de Moreno-Delgado, Orduña-Malea e Repiso (2020).

O artigo de (KOZITSINA et al., 2020) foi publicado no site arXiv. O trabalho foi desenvolvido por um grupo de pesquisa russo e discute uma forma de avaliar a influência das universidades na sociedade. A proposta é utilizar como indicador de impacto da universidade, o número de visualizações das páginas dos egressos na Wikipedia, comparando com outros ranqueamentos já estabelecidos. Os autores definem que o sucesso nas carreiras dos egressos representam a qualidade dos programas e sistemas educacionais da universidade e, que ao mesmo tempo, a significância das universidades estão relacionadas à influência de seus egressos. Os autores então discutem sobre alguns ranqueamentos internacionais já existentes, como o *Ranking* QS (Quacquarelli Symonds) e *Ranking* ARWU (Academic Ranking of World Universities), sendo o último também conhecido como o *Ranking* de Xangai.

Um detalhe é que os autores informam os números de universidades e egressos que tiveram os dados coletados, mas não mencionam a metodologia utilizada para esta coleta, apenas que a coleta foi feita na Wikipedia. Reconhecem ainda que o método proposto possui algumas limitações como a influência que certos eventos ou notícias podem ter sobre o número de visualizações de certas pessoas famosas. No entanto, esse fato também pode indicar popularidade, e que o papel do egresso na sociedade (tais como presidente, cantor, cientista) também pode fazer com que os números de visualizações nas páginas cresçam em ritmos diferentes.

Apesar de ficar demonstrado alguma correlação dos resultados do método proposto com alguns dos ranqueamentos existentes, os autores ressaltam a forte correlação de 0,83 entre a popularidade dos egressos e das universidades, assumindo que a popularidade da universidade é um reflexo da popularidade dos formados. Outro ponto é que o interesse dos futuros alunos pode variar pelo conhecimento dos sucessos recentes, ao tentar uma vaga em determinada universidade. Por fim, consideram que o indicador não deve ser utilizado de forma individual, mas de forma complementar aos outros ranqueamento já existentes.

Outro trabalho que busca ranquear universidade de forma baseada no sucesso de seus egressos é o trabalho conduzido por Moreno-Delgado, Orduña-Malea e Repiso (2020), cujo objetivo geral é determinar a viabilidade de ranquear universidades considerando o número

de graduados empregados nas companhias listadas no índice IBEX35 da bolsa espanhola. O IBEX35 é um índice que lista as 35 empresas com maior volume de *trade* em euros nos 6 meses anteriores. Para essa atividade, formularam as seguintes perguntas: (P1) De acordo com LinkedIn, quais universidades possuem maior número de egressos trabalhando nas empresas listadas na bolsa espanhola? (P2) A longevidade, modo de ensino ou distância da universidade e a empresa, estão relacionados com o número de egressos trabalhando nessas empresas? (P3) De acordo com o LinkedIn quais companhias contratam o maior número de egressos?

Após coleta de dados, os autores passam para análise dos dados e estabelecem o que chamam de "indicador IBEX35", um indicador que é a porcentagem de egressos de uma universidade que trabalham em empresas listadas no índice IBEX35. Respondendo então à primeira pergunta (P1), informam que a "Universidad Complutense de Madrid" possui 10.043 egressos trabalhando nas empresas listadas no índice IBEX35, seguida da "Universidad Politécnica de Madrid", com 8.410 egressos. Porém, informam que ao analisar o indicador IBEX35, duas instituições pequenas (Menéndez Pelayo and Pontificia Comillas) ficam com as pontuações mais altas, complementando que as instituições de grande porte ficaram com notas baixas no indicador formulado.

Em resposta à segunda pergunta (P2), os autores encontraram que o número de egressos que trabalham nas empresas do índice IBEX35 está relacionado com a longevidade e o tamanho das universidades, e que os resultados da análise realizada no estudo indicam que as empresas tendem a contratar egressos que estudaram em universidades situadas nas regiões onde essas companhias estão localizadas. E explicam que esse comportamento pode estar relacionado às especializações das universidades e pela existência de relações entre as universidades e essas empresas.

Como resposta a terceira pergunta (P3), os autores informam que as empresas Telefónica, Banco Bilbao Vizcaya Argentaria, e Indra são as que empregam mais egressos das universidades espanholas. Em termos de diversidade, Telefónica, Aena, Indra, Dia, Cellnex, e Colonial são as que recrutam egressos de um maior número de diferentes universidades.

Por fim, os autores concluíram que o indicador proposto não tem acurácia suficiente para ranquear as universidades, mas que apesar das limitações se mostra como uma ferramenta que contém informações de interesse para classificar as instituições, consistindo em um recurso a ser avaliado quando estiver escolhendo uma universidade para se inscrever.

2.3.3 Análise de egressos com formas de interação

Nesta subseção são detalhados os trabalhos de Sasikumar et al. (2020) e de Moreno-Delgado, Malea e Repiso (2020). O primeiro propõe que as instituições de ensino adotem a utilização de um portal de egressos, onde os alunos e ex-alunos possam ter acesso a

uma série de funcionalidades e que informem dados para que a instituição possa fazer o acompanhamento de egressos. O segundo solicita que o egresso informe o endereço do perfil na rede LinkedIn para que a instituição possa fazer o acompanhamento de egressos.

No caso do trabalho de Sasikumar et al. (2020), o conjunto de funcionalidades se limita basicamente a eventos relacionados ao departamento do egresso e vagas de trabalho ou estágio postadas por outros egressos e que os estudantes possam se comunicar instantaneamente com os egressos e alunos da instituição. Um ponto que os autores mencionaram foi a utilização do algoritmo SVM para filtragem dos eventos relacionados ao departamento dos egressos, porém não deram detalhes a respeito da performance obtida na filtragem ou se esse filtro poderia ter sido feito utilizando outro tipo de mecanismo. Ao final, o artigo conclui que o portal do egresso pode contribuir para a melhora do relacionamento entre estudantes e egressos, aumentando a comunicação entre eles, inclusive melhorando a relação com a instituição. E que os departamentos podem obter as informações cadastradas pelos estudantes e egressos.

O trabalho realizado por Moreno-Delgado, Malea e Repiso (2020) utiliza dados de 80 universidades espanholas para aprofundar o entendimento sobre as relações entre universidades e empresas, estudando como as companhias listadas no índice IBEX35 contratam egressos dessas universidades e de que forma as contratações são afetadas pelas localizações tanto das companhias quanto das universidades.

A partir do LinkedIn, dados de 3.716.720 perfis de egressos de 80 universidades espanholas, dos quais 97.748 são de usuários que trabalham de fato nas empresas listadas no IBEX35. Durante a análise, foram identificados o número de graduados de cada universidade que afirmaram trabalhar em cada umas das empresas do IBEX35. Posteriormente, foi realizada uma análise similar com as empresas do IBEX35, catalogando o nome, a localização e o número de empregados com perfil LinkedIn.

Além dessas informações, os autores elaboraram outras análises estatísticas como matrizes de similaridade com as informações obtidas. Os autores concluem que existe uma forte correlação entre a localização da universidade em relação ao número de egressos empregados em empresas do IBEX 35 da mesma região. Isto é, empresas que possuem a sede na mesma localização que a universidade, tendem a contratar egressos dessa universidade.

Ao final, os autores mencionam que os resultados, apesar de se mostrarem bons, devem ser analisados com precaução, pois os dados preenchidos pelos usuários não foram verificados. Também há a consideração de que algumas empresas e universidades façam campanhas para incentivar que seus empregados ou graduados criem perfis, enquanto outras que não fazem este tipo de campanha acabaram tendo resultados piores nas análises.

3 MATERIAIS E MÉTODOS

Neste capítulo serão descritos os materiais utilizados para pesquisa e também os métodos, a extração de características, os classificadores e as métricas de avaliação a serem aplicados para a identificação dos perfis das redes sociais. São dois experimentos a serem realizados, um experimento usando uma base de dados pública e anotada, a "GT dataset", e outra com os dados reais dos egressos dos cursos superiores do Campus Serra Ifes. A base de dados "GT dataset" contém pares de perfis de usuários o Google+ e do Twitter, enquanto a base de dados de egressos faz um pareamento com dados do sistema acadêmico do Ifes e com dados do LinkedIn. A coleta da base de dados de egressos é parte do projeto do presente estudo. A extração de características e a composição do vetor de características é diferente para cada experimento, enquanto os algoritmos de classificação e métricas de avaliação são os mesmos. O objetivo do primeiro experimento é o de validar a arquitetura de extração de características e classificadores, e o objetivo do segundo experimento é aplicar a arquitetura do primeiro experimento com dados reais dos egressos coletados em uma rede social de cunho profissional, avaliando a eficiência da arquitetura.

Como já dito, a etapa de extração de características tem como entrada dois perfis pareados e, como saída, um único vetor de características. Para cada par de atributo, um de cada perfil de entrada, é calculada a distância entre os valores, gerando um vetor transformado. Na fase de treinamento, cada vetor transformado terá um rótulo positivo ou negativo, indicando respectivamente se o par é da mesma pessoa ou não. Na fase de testes, dado um vetor transformado, o modelo do classificador responderá se é positivo ou negativo.

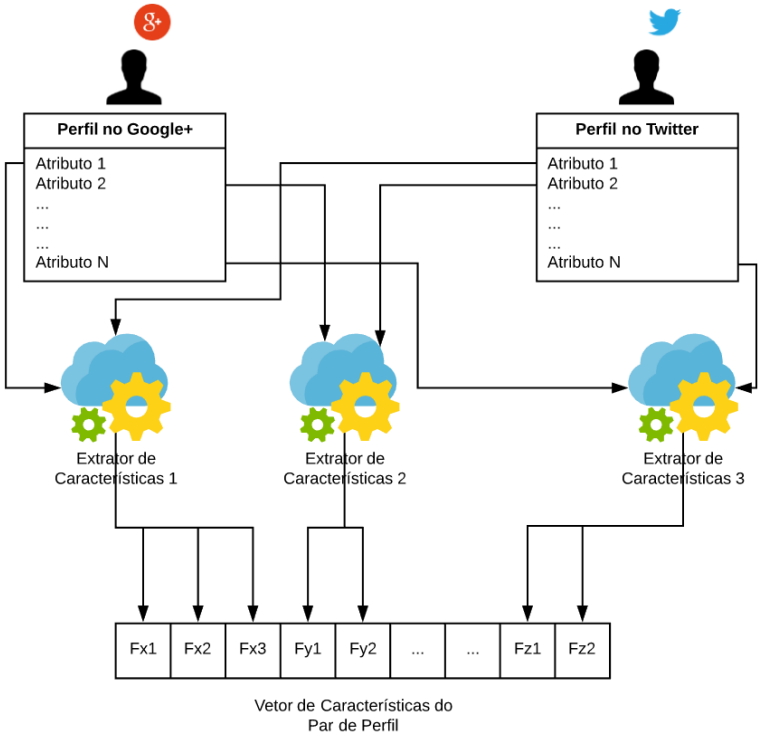
Cada par de atributos correspondentes, um de cada perfil de entrada, passa por um processo de extração de características. A Figura 2 ilustra um exemplo em que dois perfis, um do Google+ e outro do Twitter, possuem um conjunto de atributos: "Atributo 1", "Atributo 2" até "Atributo N". A quantidade de atributos em cada perfil pode ser diferente, de modo que é possível que alguns atributos não sejam usados para a extração de características. Para cada par de atributos selecionados, é calculada a distância/similaridade entre os valores, gerando uma ou mais posições do vetor de características. Na Figura 2, por exemplo, o "Extrator de Características 1" extrai 3 características (Fx1, Fx2 e Fx3) do par "Atributo 1" do Google+ e "Atributo 1" do Twitter, enquanto o "Extrator de Características 2" extrai 2 (Fy1 e Fy2), e assim por diante até que o vetor de características esteja completo.

3.1 BASES DE DADOS

Para o experimento inicial, foi utilizada uma base de dados pública, disponível no portal do Laboratório de Protocolo de Redes e Tecnologias (NPTLab) da Universidade de Milão[1]. A

[1] No momento da pesquisa, o site do Laboratório de Protocolo de Redes e Tecnologias da Universidade de Milão passou por uma alteração e os *links* de *download* da base de dados deixaram de funcionar,

Figura 2 – Arquitetura da Extração de Características dos Perfis.



Fonte: Elaborado pelo autor (2020).

base de dados, denominada "GT dataset", contém pares de perfis corretamente rotulados de usuários do *Google+* e *Twitter*, a mesma usada por Esfandyari et al. (2018).

3.1.1 Base GT *dataset*

A base de dados "GT Dataset" é composta de apenas 1 arquivo em formato JSON, sendo este arquivo uma lista de registros que são compostos de atributos de perfis do Google+ junto de atributos de perfis do Twitter. Nesta base, cada registro possui sempre 15 atributos, sendo 6 do perfil do Google+, 8 do perfil do Twitter e 1 atributo de identificação do registro na base de dados. Os atributos são listados no Quadro 2, atributos iniciados com a letra 'G' são do Google+ e com a letra 'T' do Twitter.

porém ainda é possível acessar os arquivos diretamente pela URL http://nptlab.di.unimi.it/wp-content/uploads/datasetUserIdentification.zip

Quadro 2 – Atributos da base GT Dataset

#	Nome do atributo	Descrição
1	_id	identificador do registro na base de dados
2	Gid	identificador do perfil no Google+
3	G_Firstname	representa o primeiro nome do usuário no Google+
4	G_Lastname	Atributo que representa o último nome do usuário no Google+
5	G_Displayname	representa o nome de usuário no Google+
6	G_Location	representa a localização do usuário no Google+
7	G_aboutme	contém uma descrição a respeito do usuário no Google+
8	Tid	identificador do usuário no Twitter
9	T_Fullname	representa o nome completo do usuário no Twitter
10	T_ScreenName	representa o nome de usuário no Twitter
11	T_Location	representa a localização do usuário no Twitter
12	T_Description	contém uma descrição a respeito do usuário no Twitter
13	T_Time_Zone	representa a zona de horário do usuário no Twitter
14	T_StatusText	representa um texto breve a respeito do estado do usuário no Twitter
15	T_Language	representa a língua do usuário no Twitter

Fonte: Elaborado pelo autor (2020), de acordo com (ESFANDYARI et al., 2018).

Apesar da base de dados conter 15 atributos, apenas os atributos em comum, ou seja, que existem na rede social *Google+* e *Twitter* serão utilizados para a extração de características e consequente associação de perfis. Os atributos que serão pareados são:

- concatenação do G_Firstname e G_Lastname com um espaço em branco entre eles pareado com o campo T_Fullname;
- G_Displayname pareado com T_ScreenName;
- G_Location pareado com T_Location;
- G_aboutme pareado com T_Description.

Assim, os três atributos de identificadores não serão usados: _id, Gid e Tid. Nem os atributos específicos do Twitter: T_Time_Zone, T_StatusText e T_Language.

Como todas as instâncias da base "GT dataset" são positivas, isto é, instâncias cujos pares de perfis foram corretamente identificados como pertencentes ao mesmo indivíduo, as instâncias negativas foram criadas pelo método descrito em (ESFANDYARI et al., 2018). A técnica envolve a criação aleatória de pares P_i^{s1}, P_j^{s2} tal que P_i^{s1} seja o perfil do usuário i na rede social $s1$ de uma instância positiva, e o (P_j^{s2}) seja o perfil do usuário j na rede social $s2$ de uma outra instância positiva, sendo que $i = j$.

Os autores da base elaboraram três conjuntos de treinos (Treino 1, Treino 2 e Treino 3), cada um com um nível de dificuldade diferente para os classificadores. Todos os conjuntos de treino são balanceados com 50% das instâncias positivas e 50% das instâncias negativas, diferenciando na forma como as negativas foram selecionadas:

- Treino 1: as instâncias negativas são selecionadas de forma randômica. O tamanho deste conjunto de dados é de 3.500 registros;
- Treino 2: a construção de instâncias negativas teve como objetivo obter um nível de dificuldade maior, 50% das instâncias negativas foram obtidas de forma aleatória e 50% são construídas para que cada par negativo tenha valores similares em ao menos um atributo. O tamanho deste conjunto de dados é de 3.540 registros;
- Treino 3: foi elaborado a fim de se obter um conjunto de treino mais difícil. Nele, todas as instâncias negativas são construídas de forma que cada par negativo tenha valores similares em ao menos um atributo. O tamanho deste conjunto de dados é de 3.550 registros.

Também elaboraram dois conjuntos de testes:

- Teste 1: inclui 50% de instâncias positivas e 50% das instâncias negativas construídas de forma aleatória. Contém 870 registros;
- Teste 2: inclui instâncias positivas que possuem ao menos 1 atributo diferente, enquanto todas as instâncias negativas são construídas de forma a ter valores iguais em ao menos um atributo, assim como no conjunto de treino 2. Contém 663 registros, sendo 375 positivas (representando 56,5%) e 288 negativas (43,5%).

Estes conjuntos de treino e teste, que estão disponíveis para *download*[2] serão utilizados para a extração de características e posterior análise por meio de algoritmos classificadores.

3.1.2 Base de egressos do Ifes Campus Serra - Egressos Dataset

Como o propósito deste trabalho é criar um sistema automático de acompanhamento de egressos do Ifes Campus Serra, um dos subsistemas é o de criar uma base de dados contendo informações pessoais, profissionais e acadêmicas dos egressos. Para tanto, foi desenvolvido um programa de coleta de dados inspirado nos trabalhos de Almeida (2018) e Gonçalves et al. (2014).

Foi desenvolvido um sistema composto de uma aplicação Web e um módulo Coletor. A aplicação Web possui uma interface gráfica e as funcionalidades de controle de acesso de usuários (*login*), cadastro de instituições, cadastro de cursos, cadastro de fonte de dados, e importação de planilha com os dados de egressos a serem coletados. Tal sistema foi construído utilizando a linguagem de programação Python. O cadastro de instituições e cursos só estão presentes no sistema para futura expansão, e atualmente contém apenas os nomes das instituições e cursos.

[2] O *link* para *download* é <http://nptlab.di.unimi.it/wp-content/uploads/GoogleTwitterTrainTest.zip>

O usuário deve preencher as informações de fonte de dados, isto é, as plataformas que servirão de fonte de dados. No caso do LinkedIn, por exemplo, deve-se informar as credenciais de acesso (*login*) nesta plataforma. Após este primeiro cadastro, o sistema permitirá que o operador faça a importação de uma planilha contendo informações de egressos. As informações dos egressos são: nome, instituição, nome do curso, data de ingresso, data de conclusão, data de nascimento, sexo, matrícula, CPF, naturalidade, e cidade.

O módulo Coletor usa as informações de: nome, instituição, nome do curso, data de ingresso e data de conclusão. Os campos data de nascimento e sexo apenas são úteis para fins de completude de informação, mas não são utilizados pelo módulo Coletor. O campo matrícula é importante para identificar vínculos diferentes de uma mesma pessoa dentro da instituição, por exemplo, dado que uma pessoa pode ter mais de um vínculo na instituição, mas também não é usado no módulo Coletor. O campo CPF não é usado para a coleta do LinkedIn, mas é utilizado na busca na plataforma Lattes. Os campos naturalidade e cidade auxiliam no mapeamento de mudança de local de residência do egresso, a naturalidade refere-se à cidade de nascimento, enquanto o campo cidade, refere-se à cidade de residência no ato de matrícula no curso do Ifes.

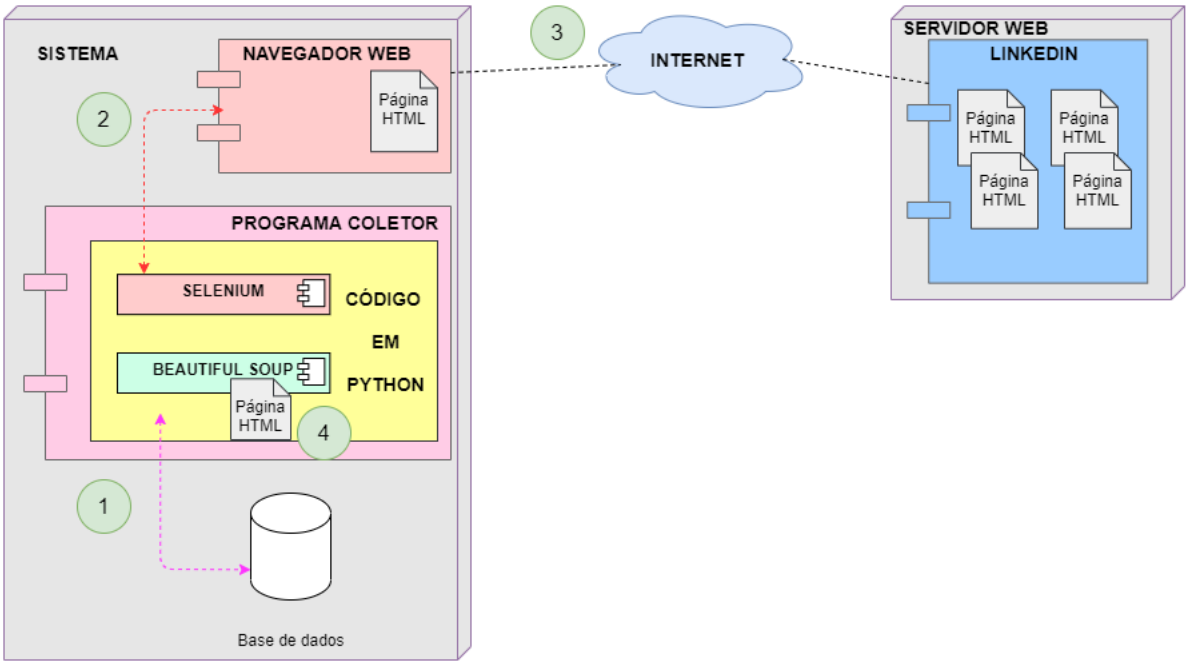
Durante o processamento da importação da lista de egressos, o sistema gera as possíveis variações de nomes para cada egresso na base de dados. Assim, gera-se todas as combinações de nomes dos egressos considerando todas as supressões e abreviações possíveis, mantendo o primeiro nome e respeitando a ordem dos sobrenomes. Por exemplo, para o nome "João Pereira da Silva" as combinações são: João Pereira da Silva, João P. da Silva, João P. Silva, João Pereira, João da Silva e João Silva. Este passo é necessário, pois verificou-se que uma pessoa pode não incluir o nome completo no LinkedIn.

O módulo Coletor também foi feito utilizando-se a linguagem de programação Python e usou a ferramenta *Selenium*[3], configurada para execução de um *driver* de navegador sem interface gráfica, de modo a possibilitar seu uso em sistemas operacionais servidores, pois estes tipicamente não possuem interface gráfica. O *Selenium* é uma ferramenta que torna possível simular ações humanas de interação no navegador web como, por exemplo, abrir o navegador, preencher campo de busca, clicar em links ou algum botão visível na página. Além disso, com o *Selenium* também é possível capturar, em forma de texto, o conteúdo da página atual apresentada no navegador. Para facilitar a busca e a análise de informações nesse tipo de texto, estruturado em HTML, foi utilizada a biblioteca *Beautiful Soup*[4], capaz de fazer a correta interpretação do HTML dos perfis recuperados pelo Coletor, ao invés do uso de expressões regulares.

[3] <https://selenium-python.readthedocs.io/>
[4] <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

Na Figura 3, o passo 1 (sinalizado pelo número 1 na figura), indica o início do funcionamento do Programa Coletor, que busca as combinações de nomes na base de dados. Para cada combinação de nome, o programa faz com que o *Selenium* abra o navegador (passo 2) e realize a pesquisa no LinkedIn. Na sequência, o site LinkedIn retorna páginas de resultados contendo os perfis relacionados ao nome buscado (passo 3). O Programa Coletor utiliza o *Beautiful Soup* para realizar o processamento do conteúdo HTML de cada página retornada como resultado da busca pela combinação de nome e também salvar na base de dados o resultado do processamento (passo 4).

Figura 3 – Arquitetura do módulo Coletor.



Fonte: Elaborado pelo autor (2021).

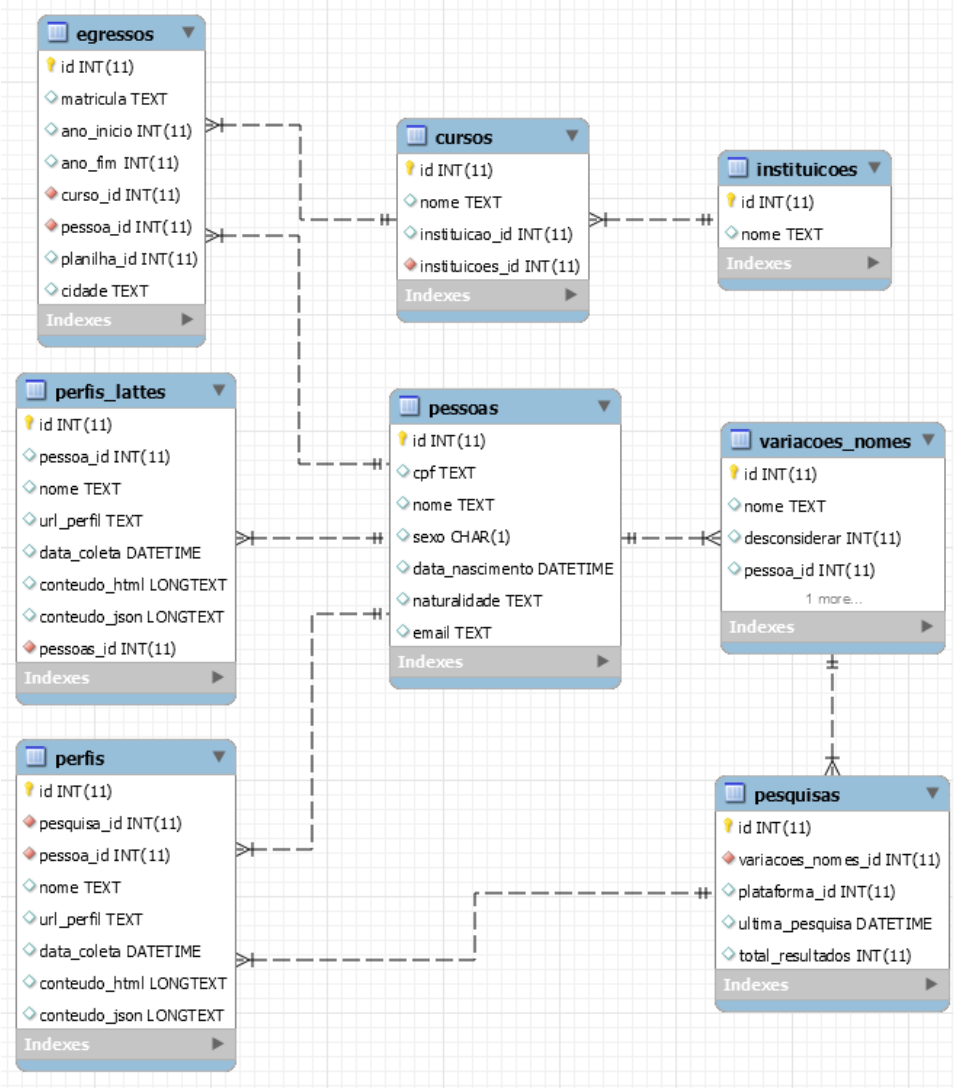
Com o passar do tempo, o LinkedIn tem introduzido mecanismos capazes de detectar a utilização de robôs (*bots*), e por este motivo, durante a elaboração desse trabalho, o módulo Coletor precisou ser ajustado algumas vezes para que durante a coleta houvesse tempos de espera entre a navegação nos resultados da busca, e também no tempo esperado para se realizar uma nova busca. Além disso, foi necessário definir um limite arbitrário de 60 perfis coletados diariamente, de modo que o LinkedIn não adicionasse uma página de autenticação contendo um captcha a fim de impedir a utilização automatizada.

O modelo Entidade-Relacionamento (modelo ER) da Figura 4 exibe as principais entidades que são utilizadas pelo módulo Coletor. Não são apresentadas algumas entidades que são utilizadas somente pela aplicação Web. Todo o código-fonte, incluindo o roteiro (*script*) de criação do banco de dados está disponível publicamente no Github[5]. No modelo ER, a lógica pretendida é que cada pessoa esteja associada a um ou mais egressos, que por

[5] <https://github.com/joaomarcosmareto/observatorio-egressos>

sua vez, estão associados a um curso de uma instituição. Cada pessoa, está associada a uma ou mais combinações de nome (variacoes_nome). Cada pesquisa por sua vez, está associada a uma variação de nome, que é o valor de entrada para a pesquisa nas fontes de dados, como neste caso, o LinkedIn (a entidade fonte de nome não é apresentada no modelo). Cada pesquisa, por sua vez, retorna zero ou mais resultados (entidade perfis).

Figura 4 – Modelo Entidade-Relacionamento das principais entidades do módulo Coletor.



Fonte: Elaborado pelo autor (2021).

Um detalhe a ser observado, é a presença dos campos “conteudo_html” e “conteudo_json” nas tabelas “perfis”. Estes campos representam respectivamente o conteúdo original obtido da fonte de dados, seja o currículo Lattes tal como foi obtido ou a página do LinkedIn, tal como obtida pelo coletor, e o resultado do processamento desse registro obtido. Na coluna “conteudo_json” do LinkedIn, são armazenados o nome, o sobrenome e a localização na seção principal da página obtida. Além disso, para cada experiência acadêmica, são armazenados o nome do curso, a instituição em que foi cursado, o título, e os anos de início e fim do curso. Para cada experiência profissional, são armazenados o vínculo, o nome da organização, os anos de início e fim do vínculo, e a localização da empresa.

Esta base também contém registros obtidos na Plataforma Lattes, obtidos de forma manual, tendo como entrada para pesquisa a mesma lista de egressos do Ifes que foi utilizada para a pesquisa automatizada no LinkedIn. No entanto, não seguirá o mesmo fluxo do LinkedIn, pois a pesquisa na Plataforma Lattes é feita via informação de CPF do registro pretendido, ou seja, é uma busca direta em que a chave de busca identifica univocamente o registro, não sendo necessário fazer pesquisa por combinações de nome. A tabela "perfis_lattes", armazena os registros dos currículos Lattes, tem os campos: nome (do perfil do Lattes), a *url* do currículo (url_perfil) e a data de coleta. Além desses, a coluna "conteudo_html" armazena a página HTML do Lattes e na coluna "conteudo_json", por sua vez, são armazenados o nome em citações bibliográficas, a data de atualização, a identificação do Lattes (idLattes), o texto resumo, o endereço profissional do dono do currículo, dentre outros. Além disso, são armazenadas informações sobre os registros de cursos acadêmicos realizados. Para cada curso são obtidas informações do tipo do curso, ordem em que foi feito, título do trabalho apresentado, nome da instituição, nome do curso, situação, ano de início e ano de fim do curso. Ainda são salvos os registros de atuações profissionais do dono do currículo.

3.2 EXTRAÇÃO DE CARACTERÍSTICAS

Na etapa de extração de características, foram empregadas as seguintes medidas de similaridade/distância:

- *Exact Match* (EM): comparação exata dos dois valores de entrada;
- *Longest Common Substring* (LCS): a cadeia mais longa em comum. Em geral, este valor é normalizado, dividindo-se pela média do tamanhos das duas *strings* de entrada;
- *Longest Common Sub-Sequence* (LCSS): medida parecida com a LCS, porém de forma que a sequência não precise ser contígua. Novamente o valor de retorno é normalizado pela média do tamanho das duas strings originais;
- *Levenshtein Distance* (LD): o algoritmo de Levenshtein calcula o número mínimo de operações de edição que são necessárias para modificar uma *string* de forma à obter outra *string*;
- *Jaccard Similarity* (JS): é o cálculo do tamanho da interseção de termos (por exemplo, palavras) dividida pelo tamanho da união dos conjuntos dos termos das entradas e;
- *Cosine Similarity* (CS) *with* TF-IDF *weights*: Esta é uma técnica bem conhecida de recuperação de informação, que mede a similaridade entre dois conjuntos de textos. Primeiramente são calculados os pesos TF-IDF e posteriormente esses pesos servem de entrada para a similaridade de cosseno. Os termos TF e IDF vêm do inglês e significam respectivamente frequência do termo e inverso da frequência nos documentos. Enquanto TF mede o número de vezes que um termo (palavra) aparece em cada texto, IDF tenta

dar importância aos termos com base na frequência em que aparecem nos textos de entrada e é calculado com base no número de textos e no número de textos que contém o termo a ser pesquisado. Após os pesos serem calculados para cada termo dos dois textos, eles são multiplicados e os resultados são armazenados em dois vetores, um vetor dedicado a cada texto de entrada. A similaridade entre os vetores então é dada pelo produtos desses dois vetores, medindo o cosseno do ângulo entre eles no espaço vetorial (TATA; PATEL, 2007).

3.2.1 GT Dataset

A etapa de extração de características tem como entrada dois perfis pareados e tem como saída um único vetor de características. Conforme descrito sobre a "GT dataset", um perfil é da rede social *Google+* e outro perfil é do *Twitter*. Cada perfil possui vários atributos. O que se deseja é verificar se há dois perfis que correspondem à mesma pessoa.

Para os atributos que são correspondentes, cada par de atributo, um de cada perfil de entrada passa por um processo de extração de características. Para cada par de atributos selecionados, é calculada a similaridade/distância entre os valores, gerando uma ou mais posições do vetor de característica. Seguindo a metologia proposta por Esfandyari et al. (2018), as métricas foram aplicadas da seguinte forma, resultando em um vetor com 14 características (Figura 5):

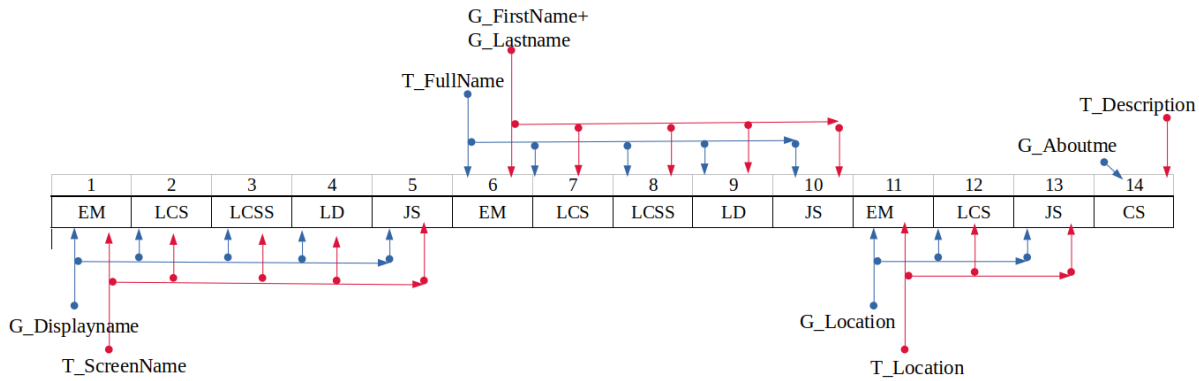
- 5 características, aplicando EM, LCS, LCSS, LD e JS ao par de campos *G_Displayname* e *T_ScreenName*;
- 5 características, aplicando EM, LCS, LCSS, LD e JS ao par de campos *T_FullName* e concatenação dos campos *G_Firstname* e *G_Lastname* com um espaço em branco;
- 3 características, aplicando as métricas EM, LCS, e JS entre os campos *G_Location* e *T_Location*;
- 1 característica pela métrica CS entre os campos *G_Aboutme* e *T_Description* dos registros.

Na fase de treinamento, cada vetor de características transformado terá um rótulo "positivo" ou "negativo", indicando respectivamente se o par de perfis é da mesma pessoa ("positivo") ou não ("negativo"). Na fase de testes, deve-se entregar um vetor transformado, cujo modelo do classificador responderá positivo ou negativo.

3.2.2 Egressos Dataset

Na base Egressos Dataset, a extração de características segue o mesmo processo da base GT Dataset. Dois perfis são pareados, sendo um do sistema de informações acadêmicas do

Figura 5 – Vetor de características com 14 posições.



Fonte: Elaborado pelo autor (2020).

Ifes, e um outro da rede social LinkedIn. Apesar dos dados do Sistema Acadêmico do Ifes não serem exatamente iguais aos dados disponíveis nos perfis do LinkedIn, espera-se que as características aplicadas ao conjunto de perfis consigam identificar quando um perfil do LinkedIn pertence a um egresso do Ifes, com a mesma assertividade do modelo construído a partir da base GT Dataset. Nesta base, os atributos foram pareados conforme cada linha do Quadro 3, sendo que o único atributo com nome diferente é o "Nome do curso" do sistema acadêmico com o "Título" do perfil do LinkedIn.

Quadro 3 – Pareamento dos atributos na base Egressos Dataset

Sistema Acadêmico	**Linkedin**
Nome completo	Nome completo
Nome do Curso	Título
Instituição	Instituição
Ano de início do curso	Ano de início do curso
Ano de finalização do curso	Ano de finalização do curso

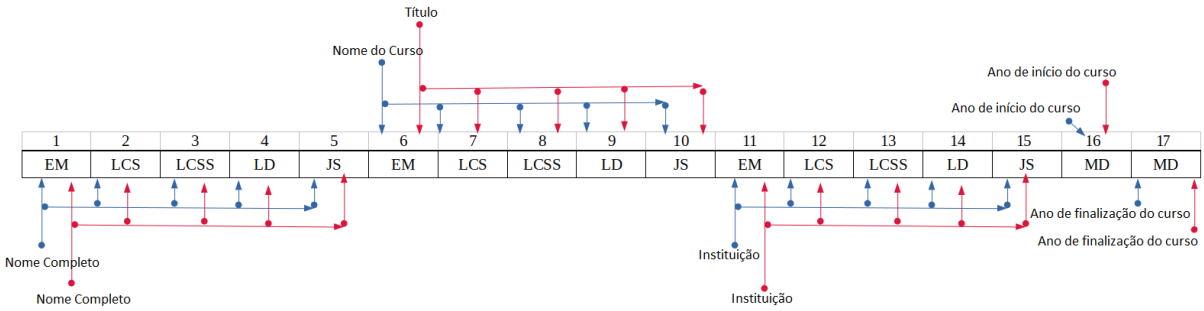
Fonte: Elaborado pelo autor (2021).

Foram usadas as métricas de distância/similaridade: Exact Match (EM), Longest Commom Substring (LCS), Longest Commom Sub-Sequence (LCSS), Levenshtein Distance (LD), Jaccard Similarity (JS) e Diferença Simples. As métricas foram aplicadas da seguinte forma, resultando em um vetor com 17 características (Figura 6):

- 5 características, aplicando EM, LCS, LCSS, LD e JS ao par de campos "Nome completo" do Sistema Acadêmico e LinkedIn;
- 5 características, aplicando EM, LCS, LCSS, LD e JS ao par de campos "Nome do Curso" do Sistema Acadêmico e "Título" do LinkedIn;
- 5 características, aplicando EM, LCS, LCSS, LD e JS ao par de campos "Instituição" do Sistema Acadêmico e LinkedIn;
- 1 característica, sendo o módulo do resultado da diferença (MD) entre o campo "Ano de início do curso" no Sistema Acadêmico e do LinkedIn;

- 1 característica, sendo o módulo do resultado da diferença (MD) entre o campo “Ano de finalização do curso” no Sistema Acadêmico e do LinkedIn.

Figura 6 – Vetor de características com 17 posições.



Fonte: Elaborado pelo autor (2021).

3.3 ALGORITMOS DE CLASSIFICAÇÃO

Os modelos preditivos clássicos podem ser baseados em distância, probabilístico, otimização e procura (FACELI et al., 2000). Assim, foi selecionado pelo menos um de cada tipo:

- três baseados em modelo probabilístico, a Regressão Logística, o LDA e o o Naïve Bayes;
- o KNN que é baseado em distância;
- a árvore de decisão baseado em procura;
- o SVM baseado em otimização;
- e ainda foram selecionados dois métodos do tipo *ensemble*: o AdaBoost e o XGBoost.

Todo o código foi desenvolvido na linguagem Python 3.6 e, com exceção do classificador XGBoost, os classificadores utilizados neste experimento são provenientes do pacote scikit-learn[6], enquanto o XGBoost, foi obtido a partir do site do Github[7]. A seguir há uma breve descrição dos algoritmos e dos valores utilizados nos parâmetros em cada algoritmo classificador. Não houve etapa de redução de dimensionalidade.

3.3.1 Regressão logística

A regressão logística é uma técnica estatística que tem como objetivo produzir um modelo que permita a predição de valores de uma variável, tipicamente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias (LI; JAIN, 1998).

[6] <https://scikit-learn.org/stable/>
[7] <https://github.com/dmlc/xgboost>

A regressão logística pode ser vista como uma melhoria da regressão linear. Enquanto a regressão linear usa uma reta que melhor se ajusta aos dados, na regressão logística é usada uma curva em formato de 'S', sendo comum o uso da função sigmoide ou a função logit. Algumas vantagens da técnica de regressão logística são: a facilidade em lidar com variáveis independentes categóricas, fornece resultados em termos de probabilidade, a facilidade de classificação de indivíduos em categorias e requer pequeno número de suposições.

Na regressão logística o parâmetro C controla o inverso da força de regularização, variando entre um limite de decisão suave e a classificação correta dos dados de treinamento. Aumentar o valor de C pode levar ao sobre-ajuste, enquanto um valor pequeno pode levar a um modelo com subajuste. O valor utilizado foi de 1. O algoritmo "resolvedor", i.e., que calcula os coeficientes, utilizado foi o Broyden–Fletcher–Goldfarb–Shanno com memória limitada como resolvedor (FLETCHER, 1987, Cap. 3). Além disso, o número máximo de iterações que o algoritmo resolvedor poderá executar para convergir para a solução da classificação foi definido como 100 e o erro de critério de parada da convergência é de 0,0001.

3.3.2 Linear Discriminant Analysis

O *Linear Discriminant Analysis* (LDA) usa o método *Singular Value Decomposition* para realizar a classificação. Tal como a técnica regressão logística, ambas se enquadram na classe de métodos estatísticos multivariados, pois relacionam um conjunto de variáveis independentes com uma variável dependente categórica (HAIR; ANDERSON, 2009). Ambas são técnicas utilizadas para classificação e discriminação de grupos, quando se deseja separar duas classes de objetos e ainda alocar um novo objeto em uma dessas classes; ambas procuram encontrar uma função ou um conjunto de funções que discrimine os grupos definidos pela variável categórica visando minimizar erros de classificação.

A diferença entre os métodos é que o modelo logístico pode ser utilizado de uma forma geral, pois não faz suposições quanto a distribuição das variáveis independentes. O LDA é ótima (porque minimiza os erros de classificação) em um contexto onde o conjunto de variáveis independentes possui um comportamento probabilístico de normalidade multivariada. Assim, o LDA deve ser preferida quando as distribuições são não-normais (VENABLES; RIPLEY, 1994).

3.3.3 Naïve Bayes

O Naïve Bayes é um algoritmo de classificação probabilístico baseado na aplicação do Teorema de Bayes (RISH et al., 2001). Tal teorema descreve a probabilidade de um evento A ocorrer, dado a ocorrência de um evento B. Este teorema é expresso na seguinte

Equação 2.

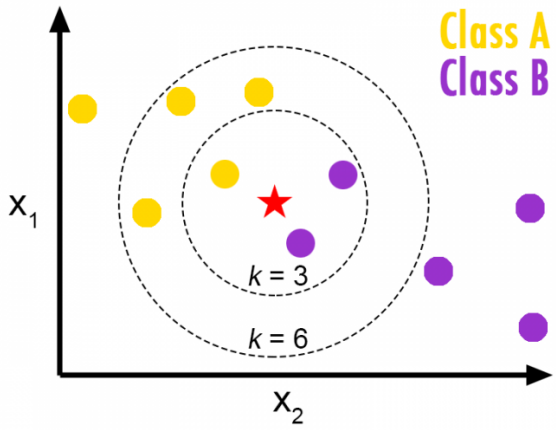
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

A principal característica deste classificador é o fato de ele desconsiderar a correlação entre variáveis sendo analisada. Logo, ele assume que cada uma das variáveis são condicionalmente independentes. Por isso o nome do classificador possui o termo *Naïve* tradução na língua inglesa para a palavra "ingênuo". Outra suposição bastante típica é que, dentro de cada classe, os valores numéricos seguem a distribuição normal - também chamada de distribuição gaussiana, distribuição de Gauss ou distribuição de Laplace-Gauss). Ou seja, descrevem uma distribuição simétrica em formato de sino. Os parâmetros do método são estimados automaticamente utilizando a máxima verossimilhança (em inglês, *maximum likelihood*). Naïve-Bayes oferece bons resultados quando se tem disponível um conjunto de treinamento médio ou grande.

3.3.4 KNN

K-vizinhos mais próximos (KNN ou k-NN ou kNN, do inglês *K-Nearest Neighbors*) é uma técnica não-linear formulada por Cover e Hart (1967). Baseado em distância, KNN efetua a classificação a partir de uma votação por maioria simples dos vizinhos mais próximos de cada ponto. Classificar um ponto, utilizando o KNN, pode ser resumido em três passos, segundo Neto, Jr. e Souza (2017), que são: (i) o cálculo da distância entre o exemplo que não é conhecido, com os demais exemplos do conjunto de treinamento, (ii) a identificação dos K vizinhos mais próximos e (iii) a utilização do rótulo da classe de vizinhos mais próximos para determinar o rótulo do exemplo desconhecido, usando um sistema de votação. Alguns sistemas de votação são peso maior ao voto de acordo com a distância do objeto de teste.

Figura 7 – Exemplo de classificação binária usando KNN.



Fonte: Bay (2015).

A Figura 7 é um exemplo de classificação pelo KNN. Há duas classes já rotuladas, os círculos amarelos (Class A) e os círculos roxos (Class B), e o objeto de teste é a estrela vermelha. É feito o cálculo da distância do objeto de testes com os outros elementos já rotulados. Quando é considerado $K = 6$, objeto de teste é associado a Class A, pois a votação dos 6 vizinhos mais próximos, que estão dentro do menor círculo tracejado, são de 2 votos para a Class B e 4 voto para a Class A. No entanto, se for considerando $K = 3$, a distância para conter os 3 vizinhos mais próximos diminui e é dado pelo círculo tracejado menor, a votação agora é vencida pela Class B (1 objetos da Class A contra 2 objetos da Class B). Logo, são parâmetros do método, o número de vizinhos, a forma de cálculo de distância e o sistema de votação.

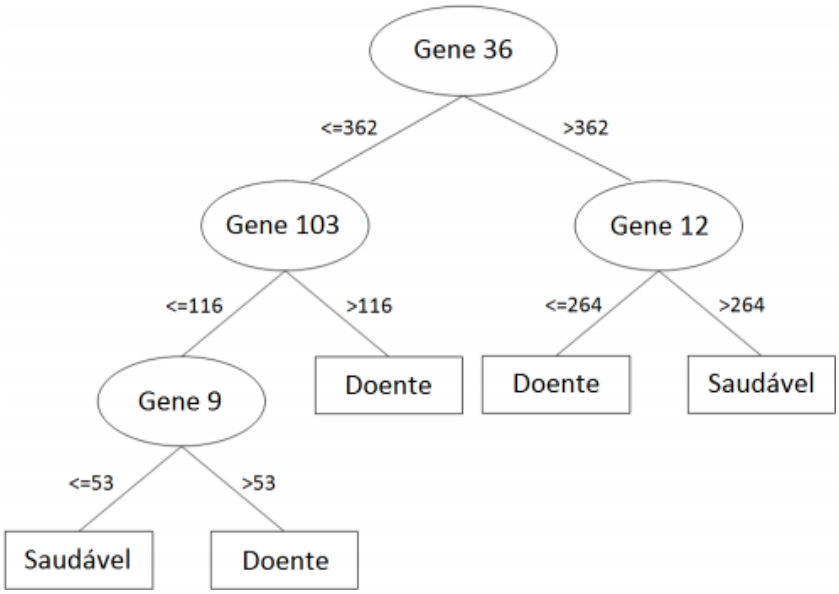
Nos experimentos, foram usados os seguintes parâmetros: 5 vizinhos ($K = 5$); todos os 5 vizinhos tem o mesmo poder de voto, isto é, não contém pesos diferentes, independente da distância; a distância usada foi a de Minkowski com potência 2 (que é o mesmo que a distância euclidiana); o parâmetro que controla o algoritmo a ser utilizado pelo classificador foi setado como "auto", o que significa que o classificador vai determinar se utiliza o algoritmo *BallItree*, *KDTree* ou força bruta de acordo com a entrada do método fit; o parâmetro *leaf_size*, por sua vez, foi setado com o valor padrão de 30. De acordo com Pedregosa et al. (2011) este valor é passado para os algoritmos *BallItree*, *KDTree*, caso sejam utilizados, e impacta na quantidade de memória utilizada, no tempo de uma consulta e no tempo de construção da árvore, porém não altera o resultado de uma consulta.

3.3.5 Árvore de decisão

A Árvore de Decisão é um classificador que faz suas deduções utilizando a estratégia dividir para conquistar. Um problema complexo é resolvido dividindo-o em outros mais simples. E nestes problemas mais simples, a mesma estratégia é aplicada. A solução obtida de cada um destes subproblemas são combinadas e tomam o formato de uma árvore (RODRIGUES et al., 2018). Uma vantagem deste modelo é a facilidade em entender as regras do modelo, bem como dos atributos mais relevantes.

As árvores são representadas pelos seguintes elementos: nós de divisão, ramos e nós folhas. Cada nó de divisão de uma árvore representa um teste em um atributo. Cada uma das possíveis repostas deste teste geram ramos. Este ramo pode se conectar a novos nós de divisão ou então a nós folha. O nó folha (o terminal) contém uma variável de saída que é utilizada para fazer a predição. Um pequeno exemplo de uma árvore de decisão pode ser visualizado na Figura 8. No exemplo, deseja-se descobrir se um determinado sujeito está doente por meio dos valores de expressão de alguns genes, se o Gene 36 for maior que 362 e o Gene 12 for maior que 264, então a pessoa é saudável, e assim, por diante. Desse modo os nós folhas representam a classificação a ser encontrada, e os nós de divisão representam as regras que são utilizadas para se chegar à tal decisão.

Figura 8 – Exemplo de uma árvore de decisão



Fonte: Oshiro (2013).

O processo de construção de uma árvore de decisão se chama "indução". Segundo Murthy (1998), a maioria das estratégias de indução de uma árvore procedem de forma gulosa, ou seja, buscam a melhor solução em cada fase (ótimo local) esperando que conduza a um ótimo global. Começando com uma árvore vazia e o conjunto de treinamento, o seguinte conjunto de passos é seguido:

1. Se todos os objetos do conjunto de treinamento no nó t pertencem a categoria c, então crie um nó folha com a classe c.
2. Se não, avalie cada um dos possíveis testes pertencentes ao conjunto de possíveis testes **S**, usando uma heurística.
3. Escolha o teste s que melhor divide o conjunto como o teste do nó atual.
4. Crie nós filhos de acordo com os diferentes resultados obtidos a partir deste teste e particione o conjunto de dados usando o teste s.
5. Um nó é tido como puro se todo o conjunto de dados no nó t pertencem a mesma classe. Repita os passos anteriores em todos os nós impuros.

Deve-se tomar cuidado quanto a profundidade da árvore gerada, pois se ela atingir sua profundidade máxima, poderá ocorrer o efeito *overfitting* e a predição ser comprometida ao aplicá-la em novos dados. Assim, é comum que as funções prontas de construção de árvore de decisão tenham uma opção de poda, que define a altura máxima da árvore. Outra forma de poda acontece no Passo 1 do algoritmo, em que se indica o número mínimo de amostra requerido para um nó ser um nó folha. E no Passo 3 e 4, para que o nó seja

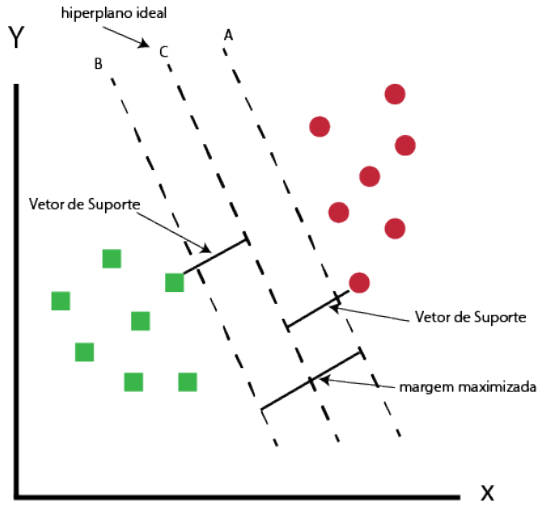
dividido é possível impor restrição de um número mínimo de amostras ou que tenha um número máximo de características.

Não houve limitação quanto à profundidade da árvore, também não houve limitação do número máximo de características a serem consideradas no processo de divisão, nem o número máximo de nós folhas. O número mínimo de amostras em um nó folha foi de 1 e o número mínimo de amostras necessárias para separar um nó foi de 2.

3.3.6 SVM

A máquina de vetor de suporte (SVM, do inglês *Support Vector Machine*) é um algoritmo de classificação binária e linear que foi introduzido por Boser, Guyon e Vapnik (1992). Baseado em otimização, o SVM constrói hiperplanos para separar duas classes no espaço. Esse algoritmo tem como vantagem ser eficaz em dados com muitas dimensões, mesmo que haja um número de instâncias menor do que o total de dimensões (CERAVOLO; BRASIL; KOMATI, 2019).

Figura 9 – Exemplo classificação binária usando SVM.



Fonte: Cavalcanti (2019).

Considere um vetor de características de duas dimensões, x e y, como no exemplo da Figura 9. O SVM busca por uma reta (hiperplano em duas dimensões) que divide o conjunto de objetos em regiões nas quais os objetos sejam da mesma classe. Existem infinitas possibilidades de hiperplanos. O SVM encontra os pontos mais próximos de ambas as classes, esses pontos são chamados vetores de suporte. O objetivo é maximizar a margem, a distância entre o hiperplano e os vetores de suporte. O hiperplano para o qual a margem é máxima é o hiperplano ideal ou ótimo.

Boser, Guyon e Vapnik (1992) ainda informa que embora o algoritmo proposto opere sobre dados linearmente separados, pode-se aplicar uma função de transformação (*kernel*) sobre os dados para que eles fiquem num formato mais adequado para a definição das retas.

Neste trabalho foi usado o SVM Classifier com o kernel função de base radial (do inglês, *Radial Basis Function* – RBF). O kernel utilizado necessita que os parâmetros gamma e C tenham os valores especificados. O valor padrão de gamma foi inicializado como scale, significando que depende dos dados de entrada, ao invés de ser um valor fixo arbitrário. Este parâmetro é um valor escalar e acaba determinando a influência que os pontos um dos exemplos de treinamento possui. O parâmetro C foi inicializado com o valor de 1.0. Valores altos de C podem levar ao superajuste enquanto valores muito baixos podem levar ao subajuste.

3.3.7 AdaBoost

De acordo com Chaves (2012), AdaBoost (*Adaptive Boosting*) já foi o algoritmo mais influente e popular da família dos algoritmos de *Boosting*. O algoritmo AdaBoost é um método *ensemble* (que tendem a apresentar um menor sobreajuste) que combina os classificadores individuais em série. Sendo definida por Freund, Schapire et al. (1996) como um método para aprimorar o desempenho de classificadores fracos de modo a transformá-los em classificadores fortes, a técnica de *boosting* faz a combinação de classificadores gerados a partir de um mesmo classificador base, e essa combinação ocorre de modo iterativo, onde o funcionamento do classificador da iteração é ajustado de acordo com os erros cometidos pelo classificador anterior (CHAVES, 2012). Deste modo, ao final, é gerado um classificador forte composto dos classificadores fracos de cada iteração.

Um ponto de atenção é em relação aos possíveis ruídos nos dados, uma vez que ao utilizar a técnica de *boosting*, o algoritmo acaba dando mais importância aos erros, os ruídos podem acabar prejudicando os resultados de classificação final. A técnica apresenta algumas vantagens como:

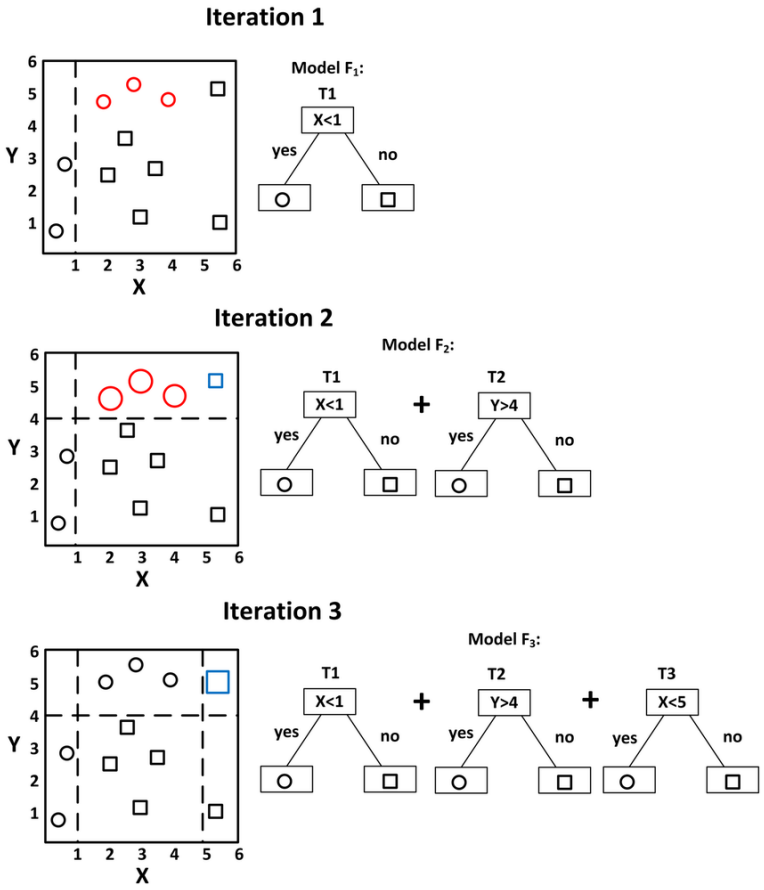
- Os valores de margem de classificação podem se apresentar mais precisos quando comparada com SVM (FREUND; SCHAPIRE; ABE, 1999)
- Corresponde a um programa linear
- Simplicidade de implementação
- Flexível para ser utilizado em diferentes áreas do conhecimento
- Capacidade de ser utilizado com outros classificadores base

Neste trabalho, o classificador *AdaBoost* foi implementado tendo como classificador fraco base, a Árvore de Decisão com profundidade máxima igual a 1. 50 foi o número máximo de árvores utilizadas pelo *boosting*, todas com peso iguais a 1, significando que todas contribuirão da mesma forma com a iteração seguinte do algoritmo.

3.3.8 XGBoost

XGBoost (acrônimo de e**X**treme **G**radient **Boost**ing) é uma técnica que tem se destacado pelo bom desempenho em competições envolvendo classificação de dados (CHEN; GUESTRIN, 2016). Tal como o AdaBoost, é do tipo *boosting*, isto é, um método *ensemble* que combina vários classificadores em série. No caso, o XGBoost se baseia na técnica de *Gradient Boosting*, que normalmente é implementado via árvores de decisão de tamanho fixo como modelos de aprendizagem individuais. A Figura 10 apresenta um exemplo simples de *Gradient Boosting* via árvore de decisão, em que se quer classificar os círculos dos quadrados. A primeira iteração (Iteration 1) tem uma única árvore que divide o espaço pela variável de entrada X com a regra $X < 1$ (em caso positivo é circulo, caso contrário é quadrado), e segunda iteração (Iteration 2) acrescenta uma segunda árvore, que acrescenta regra $Y > 4$ e, na terceira iteração (Iteration 3) a regra e $X < 5$. Ao final foram usadas 3 árvores de profundidade 1.

Figura 10 – Exemplo simples de *Gradient Boosting* via árvore de decisão.



Fonte: Zhang et al. (2018).

O algoritmo *Gradient Boosting* existe desde 1999 (MASON et al., 1999), no entanto, o tempo de treinamento das várias implementações eram custosas em termos de tempo e processamento. O procedimento sequencial resulta em modelos com grande capacidade preditiva, mas pode ser muito lento para treinar quando centenas ou milhares de árvores precisam ser criadas a partir de grandes conjuntos de dados. Implementações ingênuas

são lentas, porque o algoritmo requer que uma árvore seja criada por vez para tentar corrigir os erros de todas as árvores anteriores no modelo. Foi a implementação otimizada do *framework* XGBoost, que diminui o tempo de treinamento, além do fato de ter sido disponibilizada de forma pública e gratuita, que fez com que a técnica se tornasse popular (BROWNLEE, 2016).

Nesta implementação, o modelo foi baseado em árvore com profundidade máxima de 3, usando no máximo 100 árvores (parâmetros padrões do pacote *python* utilizado).

3.4 MÉTRICAS DE AVALIAÇÃO

Nesta seção, detalha-se a matriz de confusão e as medidas de acurácia, precisão e sensibilidade que são utilizadas para avaliação do resultado da classificação dos perfis corretos (POWERS, 2020).

No campo da aprendizagem de máquina, uma matriz de confusão, também conhecida como matriz de erro, é uma tabela que permite a visualização do desempenho de um algoritmo de classificação (STEHMAN, 1997). Cada linha da matriz representa as instâncias em uma classe prevista, enquanto cada coluna representa as instâncias em uma classe real (ou vice-versa) (POWERS, 2020).

Quadro 4 – Exemplo de uma matriz de confusão

		REAIS		
		POSITIVO	NEGATIVO	
PREDITAS	POSITIVO	VP	FP	precisão = (VP)/(VP+FP)
	NEGATIVO	FN	VN	
		sensibilidade = (VP)/(VP+FN)		acurácia = (VP+VN)/(VP+FP+FN+VN)

Fonte: Elaborado pelo autor (2021).

O exemplo de uma matriz de confusão é mostrado no Quadro 4, onde foi utilizado como exemplo, a classificação binária entre POSITIVO e NEGATIVO, conceito presente neste trabalho. Onde positivo é quando os dois perfis pareados são da mesma pessoa e negativo quando não são. Os valores desejados estão na diagonal principal da matriz, onde a predição é realizada de forma correta para ambos os casos: se a instância for positiva e classificada como positiva, é contada como um Verdadeiro Positivo (VP) e se a instância for negativa e classificada como negativa, ela será contada como um Verdadeiro Negativo (VN).

Os erros estão fora da diagonal principal, se a instância for POSITIVO e for predita como NEGATIVO, é contada como um Falso Negativo (FN) e; se a instância for NEGATIVO, mas for predita como POSITIVO, é contada como Falso Positivo (FP). De outra forma, a sensibilidade responde á pergunta “de todas as amostras que são positivas, quantas foram

classificados corretamente como positivas?", ou é o quanto se identificou corretamente os pares positivos dentre os pares positivos reais. E a precisão é a resposta à pergunta "dos exemplos classificados como positivos, quantos realmente são positivos?" o quanto, ou é o quanto se identificou corretamente do total de pares preditos como positivos.

A acurácia consiste na proporção de classificação corretas, tanto em casos positivos quanto em negativos, isto é, a soma de verdadeiros positivos (VP) mais verdadeiros negativos (VN) dividido pela soma das amostras, conforme é mostrado na Equação 3. Essa métrica irá demonstrar o percentual total de acertos em ambos os rótulos, tanto positivo quanto negativo.

$$\text{acurácia} = \frac{\text{VP} + \text{VN}}{(\text{VP} + \text{FN}) + (\text{VN} + \text{FP})} \tag{3}$$

A sensibilidade, também conhecido em algumas literaturas como revocação, consiste na proporção de verdadeiros positivos frente à todos os casos positivos existentes, isto é, a soma da quantidade de verdadeiros positivos (VP), divididos pela soma dos verdadeiros positivos (VP) mais os falsos negativos (FN), conforme mostrado na Equação 4.

$$\text{sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}} \tag{4}$$

A precisão nos dá a informação de quantas classificações verdadeiramente positivas o classificador predisse dentre todas as classificações positivas realizadas. Consiste na proporção de verdadeiros positivos (VP) divididos pela soma de verdadeiros positivos (VP) e falsos positivos (FP), conforme mostrado na Equação 5.

$$\text{precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}} \tag{5}$$

Outra métrica considerada particularmente útil para conjuntos de dados com classes desbalanceadas é a *medida F1* ou *escore F1*. O escore F1 é uma medida de "acurácia" calculada à partir da precisão e da sensibilidade. Como é possível ver pela Equação 3, a medida de **acurácia** é significativamente afetada por um grande número de acertos em qualquer uma das duas classes. Deste modo, se uma das duas classes está representada em número muito maior do que a outra, erros na classe subrepresentada podem não se

refletir na acurácia. Neste tipo de situação o escore F1 que procura um equilíbrio entre precisão e sensibilidade provavelmente será uma medida mais informativa.

$$F_1 = 2 \cdot \frac{\text{precisão} \times \text{sensibilidade}}{\text{precisão} + \text{sensibilidade}} \tag{6}$$

A Equação 6 apresenta a fórmula para o cálculo da medida F1 que é a média harmônica entre a precisão e a sensibilidade.

4 RESULTADOS E DISCUSSÃO

Neste capítulo serão apresentados os resultados do Experimento 1, utilizando a base de dados GT Dataset, e do Experimento 2, com a base de dados de egressos do Campus Serra do Ifes. A base de dados criada a partir do Módulo Coletor, será denominada a partir de agora, apenas de base de dados "Egressos Dataset". Ainda são apresentados os dados com a complementação de informações da Plataforma Lattes e ao final, comparação com os relatórios de egressos do Campus Serra do Ifes realizados em 2019 e 2020 pelo REC.

4.1 EXPERIMENTOS NA BASE GT DATASET

De modo a testar a metodologia proposta por Esfandyari et al. (2018), o experimento proposto pelos autores da base GT Dataset foi replicado estendendo à outros classificadores. Testa-se o desempenho dos classificadores usando a extração de características baseada em similaridade de atributos de perfil do tipo texto (*string*), na resolução do problema de associação de perfis em redes sociais da base "GT Dataset". Seguiu-se a metodologia proposta por Esfandyari et al. (2018), de forma que fosse possível comparar os resultados. Cada um dos três conjuntos de treino foi utilizado para treinar os classificadores. Em seguida, cada modelo de classificador foi testado em dois conjuntos de testes diferentes, gerando ao final, 48 experimentos. Foi usada a validação cruzada com k-*fold*, com $k = 10$.

Os experimentos foram executados em um computador com sistema operacional Microsoft Windows 10 de 64 bits, processador Ryzen 7 1700 e 8 gigabytes de memória RAM. Para comparação dos resultados, foram utilizadas as métricas acurácia, precisão, sensibilidade e o tempo de execução dos testes. Os resultados estão resumidos no Quadro 5, com o tempo de execução e o valor da acurácia de cada experimento. Em todos os quadros de resultados, foram usadas duas casas decimais de modo a facilitar a leitura. Os valores em negrito são os melhores valores da métrica avaliada (considerando diferença de 0,1) de um determinado Teste/Treino.

Como esperado, os modelos testados nos conjuntos de dados com o menor nível de dificuldade (Treino 1/Teste 1) obtiveram os melhores resultados. Resultado compatível com a metodologia proposta. Os melhores resultados de acurácia foram do AdaBoost e do XGBoost, resultado esperado, pois métodos *ensemble*, em geral, têm melhores resultados que modelos individuais. Com relação ao tempo de execução, o AdaBoost leva praticamente o dobro do tempo de resposta que o XGBoost. Os classificadores KNN e SVM apresentam tempo de resposta bem maior que todos os outros e também os piores valores de acurácia.

O Quadro 6 apresenta as métricas de precisão e sensibilidade. Assim como no quadro anterior, os melhores classificadores foram AdaBoost e XGBoost, em todos os casos de treino e teste. Os valores em negrito marcam os melhores resultados. Os resultados

da regressão logística e do LDA também são considerados bons. No geral, os piores resultados são dos classificadores KNN e SVM. É interessante observar que dentre os métodos individuais, os baseados em teoria probabilística, a regressão logística e o LDA, apresentam resultados melhores que os outros.

Quadro 5 – Comparativo de desempenho dos classificadores na GT Dataset, onde T (ms) é o tempo e Acc é a acurácia.

	Treino 1				Treino 2				Treino 3			
	Teste 1		Teste 2		Teste 1		Teste 2		Teste 1		Teste 2	
Classificadores	T	Acc	T	Acc	T	Acc	T	Acc	T	Acc	T	Acc
Regressão Logística	2	0,94	2	0,82	2	0,92	1	0,88	2	**0,92**	1	0,88
LDA	2	0,90	1	**0,84**	2	0,90	1	0,86	2	0,89	1	0,85
Naïve Bayes	2	0,94	2	0,81	3	0,87	2	0,83	2	0,86	1	0,81
KNN	35	0,90	27	0,78	34	0,85	30	0,81	34	0,76	26	0,71
Árvore de decisão	2	0,93	1	0,82	2	0,89	2	0,88	2	0,83	1	0,84
SVM	23	0,89	21	0,76	28	0,84	29	0,79	28	0,83	21	0,77
AdaBoost	13	**0,96**	12	**0,84**	13	**0,93**	11	**0,90**	12	**0,93**	12	**0,91**
XGBoost	7	**0,96**	6	**0,85**	7	**0,94**	5	**0,91**	8	**0,93**	6	**0,91**

Fonte: Elaborado pelo autor (2021).

Quadro 6 – Comparativo de desempenho dos classificadores na GT Dataset, onde P é Precisão e S é Sensibilidade.

	Treino 1				Treino 2				Treino 3			
	Teste 1		Teste 2		Teste 1		Teste 2		Teste 1		Teste 2	
Classificadores	P	S	P	S	P	S	P	S	P	S	P	S
Regressão Logística	0,95	0,94	0,82	0,81	**0,93**	0,92	0,88	0,89	**0,93**	**0,91**	0,88	0,89
LDA	0,91	0,90	**0,85**	**0,85**	0,92	0,90	0,87	0,87	0,91	0,89	0,86	0,86
Naïve Bayes	0,94	0,94	0,81	0,81	0,90	0,87	0,85	0,84	0,89	0,86	0,85	0,83
KNN	0,91	0,90	0,78	0,78	0,87	0,85	0,83	0,83	0,77	0,76	0,71	0,71
Árvore de decisão	0,93	0,93	0,83	0,80	0,89	0,89	0,88	0,89	0,83	0,83	0,84	0,84
SVM	0,91	0,89	0,76	0,76	0,88	0,84	0,83	0,81	0,87	0,83	0,82	0,80
AdaBoost	**0,96**	**0,96**	**0,85**	**0,83**	**0,93**	**0,93**	**0,90**	**0,91**	**0,93**	**0,93**	**0,91**	**0,92**
XGBoost	**0,97**	**0,96**	**0,86**	**0,83**	**0,94**	**0,94**	**0,91**	**0,92**	**0,93**	**0,93**	**0,91**	**0,92**

Fonte: Elaborado pelo autor (2021).

Os resultados obtidos neste trabalho foram comparados com os resultados obtidos por Esfandyari et al. (2018) e apresentados nos Quadros 7, 8 e 9. Em termos de acurácia, o melhor resultado do XGBoost foi no cenário Treino 3 e Teste 1, apresentando valor de 0,93 de acurácia contra 0,90 do trabalho de Esfandyari et al. (2018). Não é perceptível a diferença de acurácias nos outros cenários. Praticamente não há diferenças em questão da métrica de precisão entre os trabalhos. Em termos de sensibilidade, os resultados deste trabalho, AdaBoost e XGBoost, foram melhores no cenário Treino 3 e Teste 1 e foram piores no cenário Treino 1 Teste 2.

Quadro 7 – Comparação de Acurácia entre este trabalho e o de Esfandyari e colegas (ESFANDYARI et al., 2018).

		Treino 1		Treino 2		Treino 3	
	Classificadores	Teste 1	Teste 2	Teste 1	Teste 2	Teste 1	Teste 2
(ESFANDYARI et al., 2018)	MLP	0,96	0,84	0,92	0,91	0,90	0,91
	Random Forest	0,95	0,85	0,93	0,91	0,90	0,92
Este trabalho	AdaBoost	0,96	0,84	0,93	0,90	0,93	0,91
	XGBoost	0,96	0,85	0,94	0,91	0,93	0,91

Fonte: Elaborado pelo autor (2021).

Quadro 8 – Comparação de Precisão entre este trabalho e o de Esfandyari e colegas (ESFANDYARI et al., 2018).

		Treino 1		Treino 2		Treino 3	
	Classificadores	Teste 1	Teste 2	Teste 1	Teste 2	Teste 1	Teste 2
(ESFANDYARI et al., 2018)	MLP	0,96	0,85	0,92	0,91	0,90	0,91
	Random Forest	0,96	0,85	0,94	0,91	0,90	0,93
Este trabalho	AdaBoost	0,96	0,85	0,93	0,90	0,93	0,91
	XGBoost	0,97	0,86	0,94	0,91	0,93	0,91

Fonte: Elaborado pelo autor (2021).

Quadro 9 – Comparação de Sensibilidade entre este trabalho e o de Esfandyari e colegas (ESFANDYARI et al., 2018).

		Treino 1		Treino 2		Treino 3	
	Classificadores	Teste 1	Teste 2	Teste 1	Teste 2	Teste 1	Teste 2
(ESFANDYARI et al., 2018)	MLP	0,96	0,85	0,92	0,91	0,90	0,91
	Random Forest	0,95	0,85	0,93	0,91	0,90	0,92
Este trabalho	AdaBoost	0,96	0,83	0,93	0,91	0,93	0,92
	XGBoost	0,96	0,83	0,94	0,92	0,93	0,92

Fonte: Elaborado pelo autor (2021).

4.2 EXPERIMENTOS NA BASE EGRESSOS DATASET

Após os experimentos com a base GT Dataset, foi possível comprovar a replicação dos resultados do artigo de Esfandyari et al. (2018) e compará-los com mais métodos de classificação. Assim, serão usadas as mesmas técnicas para o Experimento 2, com a base de dados criada de egressos do Campus Serra do Ifes. Uma das diferenças entre os experimentos deu-se no conjunto de dados, devido ao fato dos atributos disponíveis na base Egressos Dataset serem diferentes dos atributos disponíveis na base criada por Esfandyari et al. (2018). Outra diferença foi a ausência de conjuntos de treino e testes separados e balanceados.

A base de dados "Egressos Dataset" contém perfis coletados a partir das informações de alunos egressos dos cursos superiores e FIC (Formação Inicial e Continuada) do Ifes campus Serra até o mês de junho de 2020, segundo sistema acadêmico do Ifes. A lista inicial dos egressos é composta por 454 registros. Os 454 nomes de egressos resultaram em 3.542 combinações de nomes. Essas combinações de nomes, por sua vez, permitiram que 7.765 perfis do LinkedIn fossem coletados e salvos na base de dados. Para o pareamento, os perfis do LinkedIn que possuem mais de 1 registro de curso acadêmico, foram separados

em tantos registros quanto cursos, resultando em 11.252 perfis do LinkedIn. Esses perfis foram confrontados com os dados dos egressos do Ifes, e por anotação manual, chegou-se a 197 pares corretamente identificados como verdadeiros, enquanto 11.055 registros do LinkedIn não pertencem aos egressos.

A Tabela 1 apresenta o nível de confiança da pesquisa por curso dos perfis encontrados no LinkedIn (AGRANONIK; HIRAKATA, 2011). A coluna "População" é a quantidade de egressos da lista por curso. A "Amostra mínima" é a quantidade de egressos que deveriam responder ser encontrados para se ter um nível de confiança de 95%. A coluna "LinkedIn" é a quantidade de egressos encontrados, conferidos manualmente. A coluna "Nível de confiança" é o novo índice, calculado a partir da quantidade de egressos encontrados no LinkedIn. Para um total de 454 egressos, a amostra mínima referente à 95% de confiança é de 208 egressos. Com o conjunto identificado é de 197 egressos no LinkedIn, o nível de confiança atualizado é de 93,8%. Analisando os cursos de forma individual, o resultado foi insuficiente com nível de confiança variando entre 52,9% e 70,8%.

Tabela 1 – Nível de confiança de pesquisa por curso da base Egressos Dataset pela coleta do LinkedIn.

Curso	**População**	**Amostra mínima**	**LinkedIn**	**Nível de confiança**
Superior de Tecnologia em Redes de Computadores	94	76	37	56,3%
Tecnologia em Análise e Desenvolvimento de Sistemas	55	48	37	52,9%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	107	84	37	54,6%
Sistemas de Informação	81	67	43	64,7%
Engenharia de Controle e Automação	104	82	54	70,8%
cursos FIC	13	13	0	0,0%
Soma	454	208	197	93,8%
Soma sem FIC	441	206	197	94,1%

Fonte: Elaborado pelo autor (2021).

Como na base Egressos Dataset há poucos pares verdadeiros, optou-se pela separação dos dados em apenas um conjunto de treino e testes, de tal modo que 80% dos pares verdadeiros e falsos pertencessem ao conjunto de treino, enquanto 20% dos pares verdadeiros e falsos pertencessem ao conjunto de testes. Em números absolutos, o conjunto de treino conta com 157 pares verdadeiros e 8.844 pares falsos, enquanto o conjunto de testes conta com 40 pares verdadeiros e 2.211 pares falsos. Este resultado desbalanceado de classes é um desafio na predição de classificadores (LEEVY et al., 2018).

Assim como nos experimentos já relatados na seção anterior, cada algoritmo classificador foi executado utilizando a validação cruzada com *k-fold*, sendo $k = 10$. Os resultados estão discriminados adiante, no Quadro 10. Neste quadro podemos ver, além da acurácia, da precisão e da sensibilidade, também o *Escore F1*, que é uma medida de acurácia formada pela média harmônica entre a precisão e a sensibilidade. Diferente do Experimento 1, como os resultados dos experimentos diferenciaram-se, na maioria das vezes, na terceira casa decimal, optou-se por exibir o resultado com 6 casas decimais, de modo a facilitar a diferenciação. Além disso, os valores em negrito são os resultados abaixo do do valor 0,95. Tal como nos resultados obtidos a partir dos experimentos executados na base GT Dataset, os melhores resultados obtidos nos experimentos na base Egressos Dataset foram com os algoritmos classificadores do tipo *ensemble*, enquanto o pior resultado novamente foi obtido na execução do algoritmo classificador SVM.

Quadro 10 – Comparativo de desempenho dos classificadores na Egressos Dataset.

	Resultados do Teste			
Classificadores	Acurácia	Precisão	Sensibilidade	Escore F1
Regressão Logística	0.996002	0.983045	**0.899774**	**0.93956810**
LDA	0.995558	**0.912590**	0.973191	**0.94191677**
Naïve Bayes	0.984896	**0.770604**	0.980037	**0.86279304**
KNN	0.997335	0.984982	**0.937274**	0.96053597
Árvore de Decisão	0.998667	0.999322	0.961369	0.97997817
SVM	0.982230	**0.491115**	**0.500000**	**0.49551767**
Ada Boost	0.999556	0.999774	0.987500	0.99359910
XGBoost	0.999556	0.999774	0.987500	0.99359910

Fonte: Elaborado pelo autor (2021).

Quadro 11 – Comparativo das matrizes de confusão dos classificadores na base Egressos Dataset.

	Resultados do Teste			
Classificadores	VN	FP	FN	VP
Regressão Logística	2210	1	8	32
LDA	2203	8	2	38
Naïve Bayes	2178	33	1	39
KNN	2210	1	5	35
Árvore de Decisão	2211	0	3	37
SVM	2211	0	40	0
AdaBoost	2211	0	1	39
XGBoost	2211	0	1	39

Fonte: Elaborado pelo autor (2021).

No geral, todos os classificadores apresentaram bons resultados de acurácia para o conjunto de dados criado, enquanto as maiores diferenças se dão nas métricas de precisão e sensibilidade, típico de bases de dados desbalanceados. Para a métrica de precisão, os piores resultados foram SVM (0,491115) e o Naïve Bayes (0,770604), e para a métrica de sensibilidade, os piores resultados foram o SVM (0,500000) e a Regressão Logística

(0,899774). Para a métrica de escore F1, os piores resultados foram SVM (0,49551767) e o Naïve Bayes (0,86279304).

Quando há um grande desbalanceamento da base de dados, é interessante avaliar o resultado da matriz de confusão. No Quadro 11, é exibida uma lista contendo os classificadores e os respectivos valores da matriz de confusão dos resultados. No quadro, VN significa Verdadeiro Negativo, FP significa Falso Positivo, FN significa Falso Negativo e VP significa Verdadeiro Positivo. O pior resultado foi do SVM, em que todos os registros foram classificados não sendo de egressos. Ou seja, o classificador não conseguiu traçar a separação entre as classes. O fato do LDA ter apresentado resultados piores que a regressão logística, pressupõe que as variáveis de entrada tinham distribuição não-normais.

Quadro 12 – Pareamento dos atributos do registro falso negativo do AdaBoost e do XGBoost

Sistema Acadêmico	**Linkedin**
Nome completo: Renato Pescinalli Morati	Nome completo: Renato Pescinalli Morati
Nome do Curso: Superior de Tecnologia em Redes de Computadores	Título: Graduação
Instituição: Instituto Federal do Espírito Santo Campus Serra	Instituição: Centro Federal de Educação Tecnológica do Espírito Santo
Ano de início do curso: 2007	Ano de início do curso: 2006
Ano de finalização do curso: 2010	Ano de finalização do curso: 2010

Fonte: Elaborado pelo autor (2021).

Nos resultados do AdaBoost e do XGBoost, apenas um único registro foi incorreto (1 falso negativo). O falso negativo significa que o perfil realmente era de egresso, mas o classificador não o considerou como um. Ao se verificar o registro, foi encontrado exatamente o mesmo nos dois classificadores, cujos atributos são apresentados no Quadro 12. Este par de registro apresenta valores referentes ao – Nome do curso/Título, instituição, ano de inicio do curso – diferentes entre o perfil do LinkedIn e do sistema acadêmico do Ifes. Essas diferenças foram o suficiente para que o classificador não considerasse que era a mesma pessoa, mesmo que o nome completo seja o mesmo.

Ao final, os resultados dos classificadores AdaBoost e XGBoost foram muito motivadores para a implantação do sistema em ambiente de produção. Importante salientar que os classificadores não encontraram qualquer falso positivo, isto é, encontrado um perfil no LinkedIn que não era egresso.

4.2.1 Análise de egressos segundo perfil do LinkedIn

Além dos resultados já discutidos para construção do modelo classificador, outro ponto importante é o processamento das informações coletadas do LinkedIn. O Quadro 13 apresenta um exemplo do resultado da coleta do perfil do LinkedIn de um egresso. Em

termos gerais, têm-se o nome da pessoa, um texto resumo (sobre) e a sua localização (onde a pessoa mora). É preciso notar que os campos acadêmico e profissional, são do tipo lista, permitindo que seja traçada uma evolução temporal do dono do perfil. No exemplo, tem-se que a lista de registros acadêmicos possui 2 elementos, sendo possível identificar o nome do curso realizado, a instituição onde foi cursado, o tipo do curso (campo título), e os anos de início e fim do curso. Com relação às experiências profissionais, há 6 registros, onde 4 não estão visíveis. Os campos obtidos são o vínculo com a organização (que é o cargo), a organização (empresa onde trabalha), os anos de início e fim do vínculo e a localização da organização. Com isso, é possível traçar não só a evolução de carreira desta pessoa, mas também as experiências acadêmicas e movimentação geográfica.

Quadro 13 – Exemplo de um perfil obtido a partir da coleta no LinkedIn

```
    "nome": "João Marcos Mareto Calado",
    "sobre": "Conhecimentos e Habilidades:Desenvolvimento Móvel: Programaç...",
    "localizacao": "Vitória, Espírito Santo, Brasil",
    "academico": [
        {
            "curso": "Engenharia de Sistemas",
            "instituicao": "ESAB - Escola Superior Aberta do Brasil",
            "titulo": "Pós-graduação Lato Sensu - Especialização",
            "ano_inicio": "2014",
            "ano_fim": "2016"
        },
        {
            "curso": "Análise e Desenvolvimento de Sistemas",
            "instituicao": "Ifes - Instituto Federal do Espírito Santo",
            "titulo": "Tecnólogo",
            "ano_inicio": "2006", "ano_fim": "2013"
        }
    ],
    "profissional": [
        {
            "vinculo": "Coordenador de equipe",
            "organizacao": "Ifes - Instituto Federal do Espírito Santo",
            "ano_inicio": "11/2019", "ano_fim": "o momento",
            "localizacao": "Vitória, Espírito Santo, Brasil"
        },
        {...}
        {
            "vinculo": "Analista de TI",
            "organizacao": "Ifes - Instituto Federal do Espírito Santo",
            "ano_inicio": "04/2013", "ano_fim": "06/2016",
            "localizacao": "Vitória, Espírito Santo, Brasil"
        },
    ]
```

Fonte: elaborado pelo autor (2021).

A partir dos dados coletados, é possível fazer a análise de inserção profissional dos egressos

mapeados. Dos 197 egressos, 172 perfis contém registro de vínculo empregatício sem data de fim, assim, considerou-se que estão empregados atualmente, correspondendo à 87,30% dos registros coletados. A Tabela 2 apresenta os dados da inserção profissional por curso. Na tabela, o valor da coluna percentual (%curso) indica o percentual dos egressos empregados pela quantidade de perfis coletados, e coluna (%total) é calculado em relação ao total geral de 197. Mesmo tendo o cargo à qual o egresso está empregado ou já foi empregado, é complexo avaliar se o emprego é ou não da área de seu curso. A variedade de títulos diferentes para um cargo depende muito da empresa e não há um padrão. Para tal análise seria necessário se fazer uma lista extensa de nomes de cargos possíveis de acordo com o curso.

Tabela 2 – Inserção profissional por curso, de acordo com a coleta do LinkedIn.

Curso	**Empregados**	**Coletados**	**%curso**	**%total**
Tecnologia em Redes de Computadores	34	37	91,9%	17,25%
Tecnologia em Análise e Desenvolvimento de Sistemas	25	27	92,6%	12,69%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	33	37	89,2%	16,75%
Sistemas de Informação	38	42	90,5%	19,29%
Engenharia de Controle e Automação	42	54	77,7%	21,32%
Cursos FIC	0	0	0	0
Soma	172	197	—	87,30%

Fonte: Elaborado pelo autor (2021).

Durante a análise foi possível identificar que os 197 egressos que tiveram o perfil pareado com o LinkedIn, possuem mais de 872 vínculos empregatícios, pois consideram-se todas as mudanças de emprego na carreira. Após remoção de registros que indicavam mudança de cargo numa mesma empresa para uma mesma pessoa, restaram 733 registros. Após agrupamento por empresa, foi possível identificar as empresas que mais empregaram egressos de nível superior do Ifes Campus Serra. O Quadro 14 apresenta as empresas que tiveram pelo menos 5 indicações de egressos, independente do ano. Um detalhe importante é que nem sempre é possível identificar é quando um vínculo refere-se à um estágio ou bolsa. O LEDS, por exemplo, apesar de ter sido citado por 8 egressos como vínculos empregatícios, trata-se de um laboratório de extensão do Ifes Campus Serra, e portanto não configurando como vínculo empregatício formal. Nota-se que o Ifes é o maior empregador dentre todas as instituições presentes nos vínculos coletados.

Ao visualizar apenas vínculos empregatícios localizados no Brasil, tem-se que 35 egressos possuíram vínculos em estados que não sejam o Espírito Santo, sendo que desses, 5 chegaram a trabalhar para empresas de outros países e 1 dos egressos chegou a ter vínculos

Quadro 14 – Empresas que mais contrataram egressos pela coleta do LinkedIn

Empresa	**Egressos**	**% do Total**
Instituto Federal do Espírito Santo (Ifes)	40	23.26%
ArcelorMittal	11	6.40%
PicPay	11	6.40%
Universidade Federal do Espírito Santo (UFES)	10	5.81%
Petrobras	9	5.23%
Secretaria de Educação do Estado do Espírito Santo (SEDU)	9	5.23%
Spassu Tecnologia e Serviços	9	5.23%
Accenture	8	4.65%
CSI - Solução & Tecnologia	8	4.65%
LEDS (Laboratório de Extensão em Desenvolvimento de Sistemas)	8	4.65%
PD Case Informática Ltda	8	4.65%
Vale	8	4.65%
Vixteam Consultoria & Sistemas	8	4.65%
Mogai Tecnologia de Informação	7	4.07%
Nexa Tecnologia	7	4.07%
Banestes S/A – Banco do Estado do Espírito Santo	6	3.49%
Etaure TI & Automação	6	3.49%
Prefeitura Municipal de Vitória (PMV)	6	3.49%
Prosperi Tecnologia	6	3.49%
E&L Produções de Software	5	2.91%
EDP	5	2.91%
pag!	5	2.91%
Senac Espírito Santo	5	2.91%

Fonte: Elaborado pelo autor (2021).

no Chile e na Suécia. No Quadro 15 é possível visualizar os estados do Brasil que mais contrataram egressos do Ifes Campus Serra. O destaque fica para os estados pertencentes à Região Sudeste pois foram os que mais contrataram. No quadro fica evidente a busca dos egressos por oportunidades em São Paulo, estado que apresenta quase o dobro da soma de todos os outros estados.

Quadro 15 – Estados, fora o Espírito Santo, que mais contrataram egressos de acordo com a coleta do LinkedIn

Estados	**Egressos**	**% do Total**
São Paulo	22	12.79%
Rio de Janeiro	5	2.91%
Minas Gerais	3	1.74%
Santa Catarina	2	1.16%
Paraná	1	0.58%
Rio Grande do Sul	1	0.58%

Fonte: Elaborado pelo autor (2021).

Continuando com a análise dos perfis, pode-se observar que 21 egressos preencheram informações de forma que fosse possível identificar vínculos empregatícios internacionais, isto é, fora do Brasil, sendo que desses 21 egressos, 2 passaram por 2 países. No Quadro 16 tem-se a lista dos países que tiveram empresas que contrataram egressos do Ifes campus Serra. Pelos dados, podemos observar que os destinos mais procurados pelos egressos são

Europa (somando 12 egressos de Portugal, Reino Unido, Suécia, Alemanha, Finlândia, Holanda e Irlanda) e América do Norte (somando 8 egressos do Canadá e EUA). Também é possível visualizar que tivemos apenas 1 vínculo empregatício para a América do Sul (Chile), 1 vínculo na Austrália e outro para o continente asiático, sendo a Coreia do Sul o único país da Ásia a recepcionar profissionais do Ifes Campus Serra.

Quadro 16 – Países, fora o Brasil, que mais contrataram egressos de acordo com a coleta do LinkedIn

País	**Egressos**	% do Total
Canadá	6	3.49%
Portugal	3	1.74%
Estados Unidos da América	2	1.16%
Reino Unido	2	1.16%
Suécia	2	1.16%
Alemanha	1	0.58%
Austrália	1	0.58%
Chile	1	0.58%
Coreia do Sul	1	0.58%
Finlândia	1	0.58%
Holanda	1	0.58%
Irlanda	1	0.58%
Polônia	1	0.58%

Fonte: Elaborado pelo autor (2021).

Apenas com as informações dos vínculos não é possível determinar que o egresso tenha mudado de estado ou país, dado que o trabalho pode ser na modalidade *home office*. Acredita-se que se o valor do campo localização do perfil for semelhante ou próximo do campo localização do vínculo empregatício atual, pode ser considerado um forte indício de que o egresso esteja de fato naquela localização. As informações precisam ser analisadas com cautela de modo a evitar conclusões incorretas. Por este motivo, foram feitas análises apenas a respeito da localização dos vínculos e não da localização dos egressos.

Quanto à continuidade de estudos, a Tabela 3 apresenta quantos alunos fizeram pós-graduação *lato sensu*, mestrado e doutorado após a realização do curso de graduação no Ifes. Caso um egresso tenha feito mais de uma pós-graduação *lato sensu* foram contabilizados todas as vezes, ou seja, um mesmo egresso que tenha feito 4 pós-graduação *lato sensu*, somou-se o valor 4 para o curso do egresso no Ifes. Um mesmo aluno pode ter feito mestrado e doutorado, e a contagem foi feita independentemente. Não foram contabilizados cursos que não tinha a data de conclusão, mas foram contabilizados todos os cursos que indicavam data de conclusão posterior à 2021. Também não foram contabilizados cursos técnicos e/ou superiores realizados após a conclusão do curso superior no Ifes.

Tabela 3 – Continuidade de estudos de acordo com a coleta do LinkedIn.

Curso	**Especialização**	**%curso**	**Mestrado**	**%curso**	**Doutorado**	**%curso**
Tecnologia em Redes de Computadores	22	59,5%	13	35,1%	4	10,8%
Tecnologia em Análise e Desenvolvimento de Sistemas	14	51,8%	8	29,6%	1	3,7%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	19	51,3%	1	2,7%	0	0,0%
Sistemas de Informação	2	47,6%	3	7,1%	2	4,7%
Engenharia de Controle e Automação	11	20,3%	16	3,0%	0	0,0%
Cursos FIC	0	0,0%	0	0,0%	0	0,0%
Soma	68	—	41	—	7	—

Fonte: Elaborado pelo autor (2021).

É possível observar que em torno de metade dos perfis de egressos dos cursos da área de computação concluíram um curso de especialização, e cerca de um terço dos egressos de tecnologia presenciais finalizam um curso de mestrado.

4.3 COMPLEMENTAÇÃO COM DADOS DA PLATAFORMA LATTES

Diferente da coleta e processamento automático de perfil do LinkedIn, a coleta a partir da Plataforma Lattes foi realizada de forma direta, de acordo com o CPF do egresso. O objetivo desta parte é a complementação de informações do que foi coletado pelo LinkedIn. A entrada de dados foi a mesma lista de egressos do Ifes que foi utilizada para a pesquisa automatizada no LinkedIn. Os currículos foram então processados pela aplicação Web desenvolvida neste trabalho, que ao final, salvou as informações no banco de dados, na tabela "`perfis_lattes`".

No Quadro 17 é exibido um perfil do Lattes, obtido de um egresso do Ifes. Nele podemos observar a presença de algumas informações importantes como o campo *url* que contém o endereço completo do currículo na plataforma Lattes, a presença do campo da data da última atualização (*atualizacaoCV*), o nome completo do dono do perfil, nome em citações, o texto resumo do Lattes, o endereço profissional, além de uma lista de experiências de ensino, que no exemplo, podem ser vistos 2 registros, um do tipo graduação e outro do tipo mestrado profissionalizante, bem como o nome dos cursos, e respectivas datas de início e conclusão.

Quadro 17 – Exemplo de um perfil obtido a partir da coleta no Lattes

```
    "idLattes":"1379254257583609",
    "atualizacaoCV":"16022021",
    "url":"http://lattes.cnpq.br/1379254257583609",
    "nomeCompleto":"João Marcos Mareto Calado",
    "nomeEmCitacoesBibliograficas":"CALADO, J. M. M.",
    "textoResumo":"...",
    "enderecoProfissional":"...",
    "lista-ensino":[
        {
            "tipo":"graduacao",
            "sequencia-formacao":"1",
            "titulo-do-trabalho":"FERRAMENTA PARA CRIAÇÃO AUTOMÁTICA DE CÓDIGO FONTE
            ↪  BASEADA EM TEMPLATES",
            "nome-instituicao":"Instituto Federal de Educação, Ciência e Tecnologia
            ↪  do Espírito Santo",
            "nome-curso":"Análise e Desenvolvimento de Sistemas",
            "status-do-curso":"CONCLUIDO",
            "ano-de-inicio":"2009",
            "ano-de-conclusao":"2013",
            "ano-de-obtencao-do-titulo":null
        },
        {
            "tipo":"mestrado-profissionalizante",
            "sequencia-formacao":"2",
            "titulo-do-trabalho":null,
            "nome-instituicao":"Instituto Federal de Educação, Ciência e Tecnologia
            ↪  do Espírito Santo",
            "nome-curso":"COMPUTAÇÃO APLICADA",
            "status-do-curso":"EM_ANDAMENTO",
            "ano-de-inicio":"2019",
            "ano-de-conclusao":"",
            "ano-de-obtencao-do-titulo":""
        }
    ],
    "lista-atuacoes":[
        {
            "nome-instituicao":"Instituto Federal de Educação, Ciência e Tecnologia
            ↪  do Espírito Santo",
            "sequencia-atividade":"1",
            "vinculos":[
                { ... }
            ]
        }
    ]
```

Fonte: elaborado pelo autor (2021).

No Lattes, foram encontrados 286 currículos a partir da lista de entrada dos 454 egressos. A Tabela 4 apresenta o nível de confiança da pesquisa por curso dos perfis coletados pela Plataforma Lattes. Segue a mesma ordem da Tabela 1, exceto pela coluna “Lattes”. O novo nível de confiança foi recalculado de acordo com a quantidade de perfis Lattes

encontrados. A quantidade total de currículos Lattes, no geral (286), foi maior que a amostra mínima de 208. Analisando cada curso individualmente, os cursos de Sistemas de Informação e Engenharia de Controle de Automação apresentam quantidade de perfis muito próximos da quantidade de amostra mínima.

Tabela 4 – Nível de confiança de pesquisa por curso da coleta da Plataforma Lattes.

Curso	**População**	**Amostra mínima**	**Lattes**	**Nível de confiança**
Tecnologia em Redes de Computadores	94	76	58	77,9%
Tecnologia em Análise e Desenvolvimento de Sistemas	55	48	36	68,8%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	107	84	45	62,0%
Sistemas de Informação	81	67	66	93,9%
Engenharia de Controle e Automação	104	82	80	93,6%
Cursos FIC	13	13	1	8,0%
Soma	454	208	286	99,5%
Soma sem FIC	441	206	285	99,5%

Fonte: Elaborado pelo autor (2021).

Tabela 5 – Inserção profissional por curso, de acordo com a coleta do Lattes.

Curso	**Empregados**	**Coletados**	**%curso**	**%total**
Tecnologia em Redes de Computadores	35	58	60,3%	12,24%
Tecnologia em Análise e Desenvolvimento de Sistemas	14	36	38,8%	4,89%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	26	45	57,7%	9,09%
Sistemas de Informação	19	66	28,8%	6,64
Engenharia de Controle e Automação	32	80	40,0%	11,19%
Cursos FIC	1	1	100%	0,35%
Soma	127	286	—	44,40%

Fonte: Elaborado pelo autor (2021).

A partir dos dados coletados, é possível fazer a análise de inserção profissional dos egressos mapeados. Dos 286 egressos coletados no Lattes, 127 perfis contém registro de vínculo empregatício sem data de fim, assim, considerou-se que estão empregados atualmente, correspondendo à 44,40% dos registros coletados. A Tabela 5 apresenta os dados da inserção profissional por curso, na mesma ordem e semântica da tabela apresentada no resultado LinkedIn.

Durante a análise foi possível identificar que os 286 egressos que tiveram o perfil pareado com o Lattes, possuem 596 vínculos empregatícios. Após remoções das mudanças de cargo numa mesma organização e após agrupamento por empresa, foi possível identificar as empresas que mais empregaram egressos de nível superior do Ifes Campus Serra. O Quadro 18 apresenta as empresas que tiveram pelo menos 5 indicações de egressos, independente do ano.

Quadro 18 – Empresas que mais contrataram egressos pela Coleta do Lattes

Empresa	**Egressos**	**% do Total**
Instituto Federal do Espírito Santo (Ifes)	61	35.47%
Secretaria de Educação do Estado do Espírito Santo (SEDU)	32	18.60%
Universidade Federal do Espírito Santo (UFES)	11	6.40%
Vale S.A.	8	4.65%
Prodest - ES	7	4.07%
SENAI - Departamento Regional do Espírito Santo	7	4.07%
SENAC Espírito Santo	6	3.49%
Prefeitura Municipal de Vitória (PMV)	6	3.49%
ArcelorMittal	5	2.91%

Fonte: Elaborado pelo autor (2021).

Diferente do LinkedIn, no Lattes não existe um campo para localização da organização empregadora, assim análise de estados e países empregadores fica prejudicada, dependendo de operação manual de pesquisa na internet do endereço dessas organizações. Por este motivo, a partir do Lattes não foram feitas análises a respeito da localização dos vínculos empregatícios dos egressos.

Tabela 6 – Continuidade de estudos de acordo com a coleta do Lattes.

Curso	**Especialização**	**%curso**	**Mestrado**	**%curso**	**Doutorado**	**%curso**
Tecnologia em Redes de Computadores	31	53,4%	17	29,3%	3	5,0%
Tecnologia em Análise e Desenvolvimento de Sistemas	9	25,0%	9	25,0%	0	0,0%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	28	62,2%	2	4,4%	0	0,0%
Sistemas de Informação	2	3,0%	9	13,6%	0	0,0%
Engenharia de Controle e Automação	5	6,2%	9	11,5%	0	0,0%
Cursos FIC	1	100,0%	1	100,0%	0	0,0%
Soma	76		47		3	

Fonte: Elaborado pelo autor (2021).

Quanto à continuidade de estudos, a Tabela 6 apresenta quantos alunos fizeram pós-graduação *lato sensu*, mestrado e doutorado após a realização do curso de graduação no Ifes. A contagem seguiu as mesmas regras adotadas para a coleta do LinkedIn. Ao comparar as informações de continuidade dos estudos obtidas a partir do Lattes e do

LinkedIn, é possível perceber que nos dois casos, o curso de Tecnologia em Redes de Computadores foi o que mais teve continuidade dos estudos por parte dos egressos.

4.4 COMPARAÇÃO DO ACOMPANHAMENTO DE EGRESSOS PELO REC DO CAMPUS SERRA

A fim de comparar os resultados do trabalho desenvolvido, são apresentadas informações de duas pesquisas de egressos conduzidas pelo campus Serra do Ifes. Uma pesquisa foi feita em divulgada em 2019[1] enquanto a outra em 2020[2]. O objetivo destas pesquisas foram elaboradas visando a obtenção de informações relativas à vida acadêmica e profissional dos participantes.

A pesquisa 2019 conduzida por Ifes (2019) teve início em outubro de 2018 e contou com uma relação de alunos formados entre os anos de 2015 a 2018, totalizando 422 alunos dos cursos: (i) Engenharia de Controle e Automação, (ii) Bacharelado em Sistemas de Informação, (iii) Redes de Computadores, (iv) Tecnologia em Análise e Desenvolvimento de Sistemas (TADS), (v) Técnico em Automação Industrial, (vi) Técnico em Informática, (vii) Técnico em Manutenção e Suporte em Informática. Importante avaliar que há egressos dos cursos técnicos. É mencionado que aproximadamente 105 (52,3%) dos 201 participantes da pesquisa de 2019, são de curso técnico, restando 96 participações para egressos de cursos superiores. Destacam-se os seguintes resultados da pesquisa:

- A pesquisa teve a participação de 201 egressos, o que corresponde a 47,6% do total de contatos realizados. A amostra mínima para se ter nível de confiança de 95% é de 201, logo, a pesquisa atingiu à quantidade mínima.
- Com relação as informações profissionais, a pesquisa mostrou que: 38,8% trabalham (78 egressos), 34,8% trabalham e estudam (70 egressos), 17,4% apenas estudam (35 egressos), 8,5% não está trabalhando nem estudando (17 egressos).
- A pesquisa mostrou que o trabalho de 73,6% (148 egressos) dos egressos fica localizado na Grande Vitória, 14,9% (30 egressos) dos postos de trabalho estão localizados em outra região do estado do ES e apenas 11,5% (23 egressos) dos entrevistados informaram trabalhar em outro Estado.
- 142 participantes da pesquisa informaram sobre a continuidade dos estudos, sendo que: 53,5% (76 egressos) cursou uma graduação, 34,5% (49 egressos) cursou especialização, 11,3% (16 egressos) cursou mestrado e 0,7% (10 egressos) cursou doutorado;

Como a pesquisa teve o contato com os egressos, foi possível ter um retorno quanto à opinião sobre se a sua remuneração mensal está ou não acima da média do mercado, qual

[1] https://www.serra.ifes.edu.br/noticias/campus-serra-divulga-pesquisa-com-egressos
[2] https://www.serra.ifes.edu.br/noticias/campus-serra-divulga-pesquisa-de-egressos-realizada-em-2020

o grau de satisfação em relação à atividade profissional, percepção do egresso quanto à oferta de vagas de trabalho, qual a opinião sobre o curso realizado no Ifes, como avaliam o conhecimento teórico e prático durante a sua formação no Ifes. Além de opiniões abertas, onde os egressos poderiam escrever em um campo de texto longo.

A pesquisa 2020 conduzida e divulgada em Ifes (2020), contou com uma relação de alunos formados entre os anos de 2014 a 2019, totalizando 648 egressos. Além dos 7 (sete) cursos da edição 2019, foram acrescidos: (i) qualificação profissional em eletricista predial – PROEJA, (ii) Técnico em Automação Industrial Integrado ao Ensino Médio (iii) Técnico em Informática para Internet Integrado ao Ensino Médio e (iv) Mestrado Profissional em Engenharia de Controle e Automação. É mencionado que aproximadamente 105 (52,3%) dos 201 participantes da pesquisa de 2019, são de curso técnico, restando 96 participações para egressos de cursos superiores. Destacam-se os seguintes resultados da pesquisa:

- A pesquisa teve a participação de 213 egressos, o que corresponde a 32,87% do total de contatos realizados. A amostra mínima para se ter nível de confiança de 95% é de 241, logo, a pesquisa não atingiu à quantidade mínima. Com 213 respostas, o nível de confiança é de 92,5%.
- Com relação as informações profissionais, a pesquisa mostrou que: 49,3% trabalham (105 egressos), 33,3% trabalham e estudam (71 egressos), 12,7% apenas estudam (27 egressos), 4,7% não está trabalhando nem estudando (10 egressos).
- A pesquisa mostrou que o trabalho de 74,6% (159 egressos) dos egressos fica localizado na Grande Vitória, 13,1% (28 egressos) dos postos de trabalho estão localizados em outra região do estado do ES e apenas 12,2% (26 egressos) dos entrevistados informaram trabalhar em outro Estado.
- 140 participantes da pesquisa informaram sobre a continuidade dos estudos, sendo que: 36,4% (51 egressos) cursou uma graduação, 28,6% (40 egressos) cursou especialização, 17,9% (25 egressos) cursou mestrado e 0,7% (10 egressos) cursou doutorado;

A pesquisa de 2020 ampliou mais as informações, avaliando qual o tipo de vínculo empregatício (empregados com carteira assinada (CLT), servidor público concursado, autônomo/prestador de serviço/empreendedor, contrato temporário e empregados sem carteira assinada). Também identificou 4 egressos que estão trabalhando em empresas fora do Brasil, e teve perguntas específicas sobre o setor de estágio do campus.

4.5 CONSIDERAÇÕES SOBRE OS RESULTADOS

Foi possível validar a metodologia proposta por Esfandyari et al. (2018) sobre a base de dados coletada na rede social LinkedIn, a base Egressos Dataset. Mesmo tendo-se uma

base de dados desbalanceada, característica diferenciada com relação à base GT Dataset, os classificadores AdaBoost e XGBoost apresentaram bons resultados, apresentando apenas um único falso negativo. Com a base de dados coletada automaticamente, foi possível chegar aos índices quanto à inserção profissional, a continuidade do estudo e a localização das empresas nas quais os egressos trabalham. O nível de confiança da coleta de perfis do LinkedIn (93,8%) ficou bem próximo do que o Ifes define em seu planejamento, que é de 95%. Acredita-se que a quantidade de egressos em redes sociais do tipo LinkedIn aumente com o tempo. Ao mesmo tempo, este resultado é um estudo de caso bem delimitado, dado o fato de que os cursos superiores são da área de computação e engenharia, perfis de profissionais que conjecturamos que usem mais a rede social LinkedIn que perfis da área de artes ou humanas.

Este trabalho conseguiu identificar 143 egressos que possuem tanto o LinkedIn e o Lattes devidamente identificados. Um dos exemplos é o perfil do LinkedIn apresentado no Quadro 13 e o perfil Lattes apresentado no Quadro 17, que pertencem à mesma pessoa. E com isso, é possível perceber que no perfil Lattes há o registro do curso mestrado profissionalizante, que não aparece no perfil do LinkedIn, e ao contrário no LinkedIn há o curso de pós-graduação *lato sensu* e no Lattes não há. Assim, demonstra-se a importância da união de informações coletadas por fontes diversas. No exemplo citado, é possível unir as informações, pois entende-se que não há conflito na informação de ter cursado uma pós-graduação *lato sensu* e em seguida um mestrado.

No entanto, a união de perfis não é uma tarefa simples, pois podem ocorrer conflitos de informações que devem ser resolvidos. Um exemplo foi uma informação que um usuário declarou estar fazendo um doutorado em uma universidade no Brasil no Lattes, mas simultaneamente estava numa universidade no exterior pelo LinkedIn. Uma possibilidade é que a pessoa estava fazendo um doutorado sanduíche, e informou de forma diferente em cada perfil. Outras questões como nomes de cargos diferentes na mesma empresa na mesma época, incompatibilidade de datas de início e fim na mesma empresa também foram detectados. Além disso, deve-se considerar a data da última atualização do perfil para avaliar qual é o perfil que tem o potencial de ter informações mais recentes.

Na Tabela 7 é possível visualizar a quantidade de egressos agrupados por curso que possuem perfis em ambas as plataformas utilizadas como fonte para coleta dos dados (na coluna “Em comum”), re-apresentando a quantidade de perfis identificados do LinkedIn e do Lattes, a soma de perfis já desconsiderando os valores duplicados em comum (coluna “Soma”), e população total da lista de entrada (Pop.) e o nível de confiança calculado sobre essa quantidade de perfis das duas plataformas. O novo nível de confiança (100%) com todos os perfis das duas plataformas é um motivador para que em trabalhos futuros seja desenvolvido um sistema de integração de dados, resolvendo todas as incoerências dos

dados.

Tabela 7 – Egressos com perfis comuns, em ambas as plataformas.

Curso	Em comum	LinkedIn	Lattes	Soma	Pop.	Confiança
Tecnologia em Redes de Computadores	25	37	58	70	94	90%
Tecnologia em Análise e Desenvolvimento de Sistemas	19	27	36	44	55	86%
Tecnologia em Análise e Desenvolvimento de Sistemas EaD	18	37	45	64	107	79%
Sistemas de Informação	36	42	66	72	81	99%
Engenharia de Controle e Automação	45	54	80	89	104	99%
Cursos FIC	0	0	1	1	13	8%
Total	143	197	286	340	454	100%

Fonte: Elaborado pelo autor (2021).

Para automatizar o processo de coleta das informações do Lattes, enviamos pedido à Diretoria de Tecnologia de TI do Ifes, requisitado acesso às informações do Ifes ao CNPq, pois esta última instituição expõe um serviço web autorizado por faixa de endereço IP. Assim, espera-se que no futuro seja possível obter as informações de todos os currículos Lattes que possuem como instituição de ensino o Ifes.

Durante o desenvolvimento dos experimentos, foi realizado um esforço para ter acesso à outras fontes de dados. Preenchemos formulário para "Solicitar bases de dados identificados da RAIS e do CAGED para fins estatísticos"[3]. De acordo com informações do site, qualquer pessoa poderia ter acesso às bases de dados identificados contidos na Relação Anual de Informações Sociais (RAIS) e no Cadastro Geral de Empregados e Desempregados (CAGED). No entanto, a resposta obtida foi a de fazer o *download* dos microdados já disponível[4] de forma pública. Uma nova tentativa, via e-mail, também se mostrou improdutiva. Da mesma forma, os dados da receita federal disponíveis[5] sobre empresas e seus sócios, contém apenas o nome completo do sócio e parte do CPF. Para avaliar egressos que são servidores públicos federais, da mesma forma, o portal de transparência não fornece do CPF completo dos servidores. A análise de servidores estaduais e municipais ainda se depara com a questão da não padronização e descentralização dos dados.

A comparação quantitativa entre os resultados deste trabalho e das pesquisas de egressos realizadas pelo REC é complexa, pois as pesquisas partiram de listas de egressos diferentes. Mesmo que a lista de entrada fosse a mesma, deve-se considerar que todos os resultados,

[3] <https://www.gov.br/pt-br/servicos/solicitar-acesso-aos-dados-identificados-rais-e-caged>
[4] <http://pdet.mte.gov.br/microdados-rais-e-caged>
[5] <https://www.gov.br/receitafederal/pt-br/assuntos/orientacao-tributaria/cadastros/consultas/dados-publicos-cnpj>

tanto o da coleta do LinkedIn quanto o das pesquisas pelo REC, representam uma foto, um instantâneo de um momento. A mudança de empregos dos egressos é dinâmica, logo, também é dinâmica a localização da empresa. Verifica-se o benefício de que um sistema de coleta automática das informações mantenha informações atualizadas, permitindo esse acompanhamento dinâmico.

O relatório 2020 do REC informa que a coleta de dados foi do dia 07/04/2020 até o dia 22/06/2020. Houve grande esforço de recursos humanos, pois além do envio dos e-mails sobre a pesquisa, os egressos que não responderam à pesquisa foram contatados via telefone, além da tabulação dos dados. Espera-se que depois que o sistema automático entre em funcionamento, o esforço de coleta diminuirá, pois uma vez que o classificador identifique o perfil, não será mais necessário gerar todas as combinações de nomes e pesquisar por elas nas redes sociais. Este trabalho só será necessário no caso de novos egressos ou no caso de remoção do perfil. Assim, estima-se que uma nova pesquisa de dados seja bem mais rápida que a primeira, que foi utilizada para construção da base de dados e do modelo classificador.

5 CONCLUSÃO

Ao final, atingimos o objetivo deste trabalho, que foi o de avaliar a viabilidade da construção de um observatório de egressos do Campus Serra do Ifes através da extração automática de dados de redes sociais. O acompanhamento de egressos é um problema importante para uma instituição de ensino, pois de acordo com o SINAES[1], a política institucional deve garantir mecanismo de acompanhamento de egressos e a atualização sistemática de informações a respeito da continuidade na vida acadêmica ou da inserção profissional.

Para alcançar o objetivo, foi realizado um estudo de trabalhos correlatos recentes sobre o tema de identificação cruzada entre diferentes redes sociais baseada na abordagem de atributos de perfil. Nas pesquisas realizadas, o trabalho de Esfandyari et al. (2018) foi o que demonstrou o método com o melhor resultado na tarefa de identificação de perfis de um mesmo indivíduo e foi o único que disponibilizou a base de dados usada de forma pública. Desta forma foi possível testar mais classificadores e compará-los com os resultados do trabalho base.

Para a coleta de informações da rede social LinkedIn foram utilizadas as metodologias propostas nos trabalhos de Almeida (2018) e Gonçalves et al. (2014). O resultado foi uma base de dados desbalanceada. Mesmo com essa dificuldade de desbalanceamento, a metodologia de Esfandyari et al. (2018), de usar pareamento de perfis, com métricas de distância/similaridade de atributos como entrada de classificadores, demonstrou bom resultado sobre a base de dados com os egressos do Campus Serra na rede social LinkedIn, apresentando um único falso positivo para os classificadores AdaBoost e XGBoost.

Com os resultados da coleta, foi possível mapear dados de inserção profissional, continuidade de estudos e localização das empresas nas quais os egressos trabalham. Além disso, também foi avaliado este mesmo mapeamento com o coleta de perfis da Plataforma Lattes. Chega-se à conclusão que a diversidade de informações de diferentes fontes é interessante, mas que mais estudos seriam necessários para a integração coesa de dados.

Quanto à hipótese enunciada deste trabalho, de que um sistema de coleta de informações que seja independente da interação com o usuário possa aumentar a taxa de amostragem, entende-se que a mesma não foi confirmada nem rejeitada. Pode-se dizer que a confiança estatística em relação ao modelo atual de aplicação de questionário ficaram equiparáveis. Considera-se que uma campanha de cadastro dos currículos Lattes dos alunos que estejam próximos da colação de grau, e orientação aos alunos sobre a importância da confiabilidade e fidedignidade das informações de seus perfis das redes sociais profissionais, poderia melhorar a quantidade de perfis coletados.

[1] <http://download.inep.gov.br/educacao_superior/avaliacao_institucional/instrumentos/2017/IES_recredenciamento.pdf>

5.1 TRABALHOS FUTUROS

Como trabalhos futuros, há melhorias e investigações que podem ser feitas para ampliar ainda mais o presente trabalho:

- Na literatura existem diversos algoritmos classificadores documentados e este trabalho apresenta o estudo e realiza experimentos utilizando apenas parte dos classificadores disponíveis, assim uma contribuição importante seria a aplicação e estudo de outros algoritmos classificadores aplicados a esta base de dados, inclusive com aplicação de redes neurais.
- Outro estudo seria a aplicação de técnicas para extração de características que não foram utilizadas neste trabalho. Há vários trabalhos na literatura, tal como o de Shu et al. (2017), que apresentam técnicas diferentes aplicadas ao contexto de pareamento de perfis em redes sociais.
- Como evolução do coletor e modelo criados neste trabalho, sugere-se a implementação de reconhecimento facial das redes sociais. Esperamos que com tal funcionalidade, a coleta de informações possa se tornar mais eficiente, assim como o correto pareamento de perfis.
- Outro ponto de melhoria no coletor, é a criação de uma rotina automática de coleta e atualização sistemática das informações do currículo Lattes dos egressos do Ifes.
- Especificação e implementação de um conjunto mínimo de funcionalidades para o sistema Web, de modo que passe a contar com interface gráfica, tornando o sistema completamente funcional e que torne possível sua adoção no Ifes.
- Implementar gráficos para apresentação dos dados de perfis, facilitando a visualização dos dados. Todos os resultados apresentados neste texto foram feitos de forma tabular.
- Ampliar a inserção de novas fontes de dados de forma que seja possível automatizar a obtenção de ainda mais informações, enriquecendo o conjunto de dados e tornando o sistema mais útil de forma geral.
- Investigar métodos para integração de dados incompatíveis oriundos de perfis de diferentes redes sociais.

REFERÊNCIAS

AGRANONIK, Marilyn; HIRAKATA, Vânia Naomi. Cálculo de tamanho de amostra: proporções. *Clinical & Biomedical Research*, v. 31, n. 3, 2011.

ALMEIDA, Luis Gustavo. *Recuperação de dados pessoais na Web em redes sociais autenticadas*. 2018. 123 f. Dissertação (Mestrado em Informática) — PUC-Rio, Rio de Janeiro, 2018.

BAY, Tháng. *Scikit-learn: K-nearest neighbors*. 2015. Disponível em: <https://ongxuanhong.wordpress.com/2015/07/28/scikit-learn-k-nearest-neighbors/>. Acesso em: 06 ago. 2021.

BOSER, Bernhard E.; GUYON, Isabelle M.; VAPNIK, Vladimir N. A training algorithm for optimal margin classifiers. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152. ISBN 0-89791-497-X. Disponível em: <http://doi.acm.org/10.1145/130385.130401>. Acesso em: 16 jul. 2017.

BOYD, Danah M; ELLISON, Nicole B. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, Wiley Online Library, v. 13, n. 1, p. 210–230, 2007.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996. estabelece as diretrizes e bases da educação nacional. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 23 dez. 1996. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&pagina=1&data=23/12/1996&totalArquivos=289>. Acesso em: 08 jun. 2019.

BRASIL. Lei nº 9.448, de 14 de março de 1997. transforma o instituto nacional de estudos e pesquisas educacionais - inep em autarquia federal, e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 15 mar. 1997. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=15/03/1997&jornal=1000&pagina=7&totalArquivos=12>. Acesso em: 08 jun. 2019.

BRASIL. Lei nº 10.861, de 14 de abril de 2004. institui o sistema nacional de avaliação da educação superior - sinaes e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 14 abr. 2004. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=15/04/2004&jornal=1&pagina=3&totalArquivos=160>. Acesso em: 08 jun. 2019.

BRASIL. Lei nº 11.892, de 29 de dezembro de 2008. institui a rede federal de educação profissional, científica e tecnológica, cria os institutos federais de educação, ciência e tecnologia, e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 30 dez. 2008. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?jornal=1&pagina=1&data=30/12/2008&totalArquivos=120>. Acesso em: 08 jun. 2019.

BRASIL. Decreto nº 9.235, de 15 de dezembro de 2017. dispõe sobre o exercício das funções de regulação, supervisão e avaliação das instituições de educação superior e dos cursos superiores de graduação e de pós-graduação no sistema federal de ensino. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 30 dez. 2017. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=18/12/2017&jornal=515&pagina=2&totalArquivos=208>. Acesso em: 08 jun. 2019.

BROWNLEE, Jason. *XGBoost With Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. [S.l.]: Machine Learning Mastery, 2016.

CALADO, João Marcos Mareto et al. Um sistema de acompanhamento de egressos usando dados do site escavador. In: UNESP. *Anais do XXVII Simpósio de Engenharia de Produção (SIMPEP 2020)*. [S.l.], 2020.

CALADO, João Marcos Mareto; ANDRADE, Jefferson Oliveira; KOMATI, Karin Satie. Comparação de classificadores para o problema de cross-system personalization em redes sociais. *Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação*, v. 1, n. 13, 2021.

CARMAGNOLA, Francesca; CENA, Federica. User identification for cross-system personalisation. *Information Sciences*, v. 179, n. 1, p. 16–32, 2009. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0020025508003551>.

CAVALCANTI, Toti. *Aula 08 – Scikit-Learn – Support Vector Machine ou máquina de vetores de suporte*. 2019. Disponível em: <https://www.codigofluente.com.br/aula-08-scikit-learn-maquina-de-vetores-de-suporte/>. Acesso em: 16 jul. 2019.

CERAVOLO, Isabella; BRASIL, Antonio Alexandre; KOMATI, Karin. Classifying readers with dyslexia from eye movements using machine learning and wavelets. In: *ENIAC 2019*. [S.l.: s.n.], 2019.

CHAVES, Bruno Butilhão. *Estudo do algoritmo AdaBoost de aprendizagem de máquina aplicado a sensores e sistemas embarcados.* 2012. Tese (Doutorado) — Universidade de São Paulo, 2012.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

COVER, Thomas; HART, Peter. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967.

DENG, Kaikai et al. A user identification algorithm based on user behavior analysis in social networks. *IEEE Access*, IEEE, v. 7, p. 47114–47123, 2019.

DIGIAMPIETRI, Luciano; LINDEN, Ricardo; BARBOSA, Lenin. Desambiguação de nomes em redes sociais acadêmicas: Um estudo de caso usando dblp. In: *Anais do IV Brazilian Workshop on Social Network Analysis and Mining*. Porto Alegre, RS, Brasil: SBC, 2015. ISSN 2595-6094. Disponível em: <https://sol.sbc.org.br/index.php/brasnam/article/view/6788>.

ESFANDYARI, Azadeh et al. User identification across online social networks in practice: Pitfalls and solutions. *Journal of Information Science*, SAGE Publications Sage UK: London, England, v. 44, n. 3, p. 377–391, 2018.

FACELI, Katti et al. *Inteligência Artificial*: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2000.

FLETCHER, Roger. *Practical methods of optimization*. [S.l.]: John Wiley & Sons, 1987.

FREUND, Yoav; SCHAPIRE, Robert; ABE, Naoki. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, JAPANESE SOC ARTIFICIAL INTELL, v. 14, n. 771-780, p. 1612, 1999.

FREUND, Yoav; SCHAPIRE, Robert E et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156.

GONÇALVES, Gabriel Resende et al. Gathering alumni information from a web social network. In: IEEE. *2014 9th Latin American Web Congress*. [S.l.], 2014. p. 100–108.

HAIR, Joseph F.; ANDERSON, Barry J. Babin Rolph C. *Multivariate data analysis*. [S.l.]: Pearson New International Edition, 2009.

HERRMAN, John. *Redes em crise, exceto uma: por que ninguém fala sobre o LinkedIn?* São Paulo, SP, 2019. Disponível em: <https://exame.com/negocios/por-que-ninguem-fala-sobre-o-linkedin/>, Acesso em 06 set. 2020.

INSTITUTO FEDERAL DO ESPÍRITO SANTO. *Plano de Desenvolvimento Institucional*: 2014/2 – 2019/1. Espírito Santo, ES, 2014. 205 p. Disponível em: <https://ifes.edu.br/images/stories/files/documentos_institucionais/pdi_2-08-16.pdf>. Acesso em: 08 jun. 2019.

INSTITUTO FEDERAL DO ESPÍRITO SANTO. *Indicadores dos egressos técnicos do Ifes*. Espírito Santo, ES, 2016. Disponível em: <https://prodi.ifes.edu.br/images/stories/Prodi/DPLA/EGPP/Portal_projetos/Observat\%C3\%B3rio_de_Egressos/Indicadores_Egressos_Inicial_a_18_10_2016.xlsx>. Acesso em: 08 jun. 2019.

INSTITUTO FEDERAL DO ESPÍRITO SANTO. *Observatório de Egressos do Ifes*. Espírito Santo, ES, 2016. 3 p. Disponível em: <https://prodi.ifes.edu.br/component/content/article/2-uncategorised/16279>. Acesso em: 08 jun. 2019.

INSTITUTO FEDERAL DO ESPÍRITO SANTO. *Questionário de Egressos de Nível Técnico*. Espírito Santo, ES, 2016. Disponível em: <https://questionario.ifes.edu.br/index.php/564646/lang-pt-BR>. Acesso em: 08 jun. 2019.

INSTITUTO FEDERAL DO ESPÍRITO SANTO. *Pesquisa com Egressos*. Espírito Santo, ES, 2019. Disponível em: <https://serra.ifes.edu.br/images/stories/Not\%C3\%ADcias/2019/Pesquisa_com_Egressos_-_Abril_2019rev01.pdf>. Acesso em: 24 jul. 2021.

INSTITUTO FEDERAL DO ESPÍRITO SANTO. *Pesquisa com Egressos*. Espírito Santo, ES, 2020. Disponível em: <https://serra.ifes.edu.br/images/stories/Editais_do_Campus_Serra/2020/RELAT\%C3\%93RIO_DA_PESQUISA_DE_EGRESSOS_2020_COM_ANEXOS.pdf>. Acesso em: 24 jul. 2021.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. *Instrumento de Avaliação Institucional Externa Presencial e a distância*: Credenciamento. Brasília, DF, 2017. 46 p. Disponível em: <http://download.inep.gov.br/educacao_superior/avaliacao_institucional/instrumentos/2017/IES_recredenciamento.pdf>. Acesso em: 08 jun. 2019.

KOZITSINA, Tatiana et al. Measuring university impact: Wikipedia approach. *CoRR*, abs/2012.13980, 2020. Disponível em: <https://arxiv.org/abs/2012.13980>.

LEEVY, Joffrey L et al. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, Springer, v. 5, n. 1, p. 1–30, 2018.

LI, Yongjun et al. User identification based on display names across online social networks. *IEEE Access*, IEEE, v. 5, p. 17342–17353, 2017.

LI, Yong H; JAIN, Anil K. Classification of text documents. *The Computer Journal*, Oxford University Press, v. 41, n. 8, p. 537–546, 1998.

MASON, Llew et al. Boosting algorithms as gradient descent in function space. In: *Proc. NIPS*. [S.l.: s.n.], 1999. v. 12, p. 512–518.

MORENO-DELGADO, Alicia; MALEA, Enrique Orduña; REPISO, Rafael. Relevancia de la ubicación en la relación universidad-empresa: análisis de la procedencia de los egresados de universidades españolas en empresas del ibex35. *Revista General de Informacion y Documentacion*, Universidad Complutense de Madrid, v. 30, n. 1, p. 297–312, 2020.

MORENO-DELGADO, Alicia; ORDUÑA-MALEA, Enrique; REPISO, Rafael. LinkedIn como fonte de dados para classificar as universidades de acordo com a empregabilidade dos licenciados em empresas de topo. *Transinformação*, SciELO Brasil, v. 32, 2020.

MURTHY, Sreerama K. Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, v. 2, p. 345–389, 1998.

NETO, Wilson Borba da Rocha; JR., José Maria Pires de Menezes; SOUZA, Rhulio Victor Luz. Análise de dados obtidos através de um sistema de telemetria automotivo utilizando k-nn. *XIV Encontro Nacional de Inteligência Artificial e Computacional*, p. 960–971, 2017.

OSHIRO, Thais Mayumi. *Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica*. 2013. 101 f. Dissertação (Mestrado em Bioinformática) — Universidade de São Paulo, São Paulo, 2013.

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

PERITO, Daniele et al. How unique and traceable are usernames? In: SPRINGER. *International Symposium on Privacy Enhancing Technologies Symposium*. [S.l.], 2011. p. 1–17.

POWERS, David Martin W. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061, 2020. Disponível em: <https://arxiv.org/abs/2010.16061>.

RISH, Irina et al. An empirical study of the naive bayes classifier. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46.

RODRIGUES, Diego S. et al. A comparative analysis of loan requests classification algorithms in a peer-to-peer lending platform. In: *Proceedings of the XIV Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2018. (SBSI'18). ISBN 9781450365598. Disponível em: <https://doi.org/10.1145/3229345.3229390>.

ROWE, Matthew. Applying semantic social graphs to disambiguate identity references. In: SPRINGER. *European Semantic Web Conference*. [S.l.], 2009. p. 461–475.

SAFAVIAN, S Rasoul; LANDGREBE, David. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, IEEE, v. 21, n. 3, p. 660–674, 1991.

SASIKUMAR, R et al. Alumni info-com management with distinct classification of data. *International Research Journal of Multidisciplinary Technovation*, p. 42–50, 2020.

SCHÜTZE, Hinrich; HULL, David A; PEDERSEN, Jan O. A comparison of classifiers and document representations for the routing problem. In: *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 1995. p. 229–237.

SHU, Kai et al. User identity linkage across online social networks: A review. *ACM SIGKDD Explorations Newsletter*, ACM, v. 18, n. 2, p. 5–17, 2017.

STEHMAN, Stephen V. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, Elsevier, v. 62, n. 1, p. 77–89, 1997.

TATA, Sandeep; PATEL, Jignesh M. Estimating the selectivity of tf-idf based cosine similarity predicates. *ACM Sigmod Record*, ACM New York, NY, USA, v. 36, n. 2, p. 7–12, 2007.

VALENTE, Jonas. *Facebook chega a 2,6 bilhões de usuários no mundo com suas plataformas*. Brasília, DF, 2018. Disponível em: <http://agenciabrasil.ebc.com.br/geral/noticia/2018-10/facebook-chega-26-bilhoes-de-usuarios-no-mundo-com-suas-plataformas>. Acesso em: 22 jun. 2019.

VELDMAN, Irma. *Matching Profiles from Social Network Sites*. Netherlands: University of Twente, 2009.

VENABLES, W. N.; RIPLEY, B. D. Multivariate analysis. In: _____. *Modern Applied Statistics with S-Plus*. New York, NY: Springer New York, 1994. p. 301–328. ISBN 978-1-4899-2819-1. Disponível em: <https://doi.org/10.1007/978-1-4899-2819-1_12>.

VOSECKY, Jan; HONG, Dan; SHEN, Vincent Y. User identification across multiple social networks. In: IEEE. *2009 first international conference on networked digital technologies*. [S.l.], 2009. p. 360–365.

WASSERMAN, Stanley; GALASKIEWICZ, Joseph. *Advances in social network analysis*: Research in the social and behavioral sciences. Sage publications. Califórnia, CA: Sage, 1994.

WITTEN, Ian H.; FRANK, Eibe. *Data Mining*: Practical machine learning tools and techniques. Elsevier. San Francisco, CA: Morgan Kaufmann, 2005.

ZAFARANI, Reza; LIU, Huan. Connecting corresponding identities across communities. In: AAAI. *Third International AAAI Conference on Weblogs and Social Media*. [S.l.], 2009. p. 354–357.

ZAFARANI, Reza; LIU, Huan. Connecting users across social media sites: a behavioral-modeling approach. In: ACM. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2013. p. 41–49.

ZHANG, Yuxiang et al. A local expansion propagation algorithm for social link identification. *Knowledge and Information Systems*, Springer, v. 60, n. 1, p. 545–568, 2019.

ZHANG, Zhongxing et al. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from european narcolepsy network database with machine learning. *Scientific reports*, Nature Publishing Group, v. 8, n. 1, p. 1–11, 2018.