Felipe Alvarenga Dinardi Barbosa

PREDIÇÃO DE AMBIENTES DE TRABALHO A PARTIR DE VARIÁVEIS SOCIODEMOGRÁFICAS, UTILIZANDO MACHINE LEARNING

Apoio:

CAPES

CAMPINAS
2024

Felipe Alvarenga Dinardi Barbosa

PREDIÇÃO DE AMBIENTES DE TRABALHO A PARTIR DE VARIÁVEIS SOCIODEMOGRÁFICAS, UTILIZANDO MACHINE LEARNING

Dissertação apresentada ao Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco, Área de Concentração - Avaliação Psicológica, para obtenção do título de Mestre.

ORIENTADOR(A): PROF. DR. NELSON HAUCK FILHO

CAMPINAS
2024

158.6
B197p

Barbosa, Felipe Alvarenga Dinardi
Predição de ambientes de trabalho a partir de variáveis sociodemográficas, utilizando machine learning / Felipe Alvarenga Dinardi. – Campinas, 2024.
70 p.

Dissertação (Mestrado) – Programa de Pós-Graduação *Stricto Sensu* em Psicologia da Universidade São Francisco.
Orientação de: Nelson Hauck Filho

1. Inteligência artificial. 2. Predição. 3. Interesses vocacionais. 4. Análise de dados. 5. Avaliação psicológica. I. Hauck Filho, Nelson. II. Título.

Sistema de Bibliotecas da Universidade São Francisco – SIBUSF
Ficha catalográfica elaborada por: Maria Clara Reginato P. Duarte - CRB-8/040

Educando para a paz

PROGRAMA DE PÓS-GRADUAÇÃO STRICTO SENSU EM PSICOLOGIA

Felipe Alvarenga Dinardi Barbosa defendeu a dissertação **"PREDIÇÃO DE AMBIENTES DE TRABALHO A PARTIR DE VARIÁVEIS SOCIODEMOGRÁFICAS, UTILIZANDO MACHINE LEARNING"** **aprovado** pelo Programa de Pós-Graduação Stricto Sensu em Psicologia da Universidade São Francisco em 24 de junho de 2024 pela Banca Examinadora constituída por:

Prof. Dr. Nelson Hauck Filho
Orientador e Presidente

Prof. Dr. Felipe Valentini
Examinador

Prof. Dr. Rodolfo Augusto Matteo Ambiel
Examinador

Apoio Financeiro

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

Agradecimentos

Agradeço a Deus em primeiro lugar por iluminar meu caminho e por me dar forças para chegar até o final do mestrado. Decidi iniciar o mestrado no início da trágica pandemia da COVID-19, passei por mudanças de emprego e cidade, o que complicaram a continuidade no curso, mas após a retomada, consegui chegar até o final.

À minha avó Elza (*in memorian*) que foi a minha maior fortaleza e meu exemplo de vida, que me acolheu em momentos difíceis junto do meu avô Mario (*in memorian*), que também foi um grande parceiro até os meus 12 anos.

Ao meu pai Mário Lucio e a Lisi pelo companheirismo e por sempre estarem na torcida pelas minhas conquistas.

À minha mãe Cintya pela base inicial da minha educação até os 6 anos.

À minha esposa Larissa pelo amor, carinho, paciência e por me apoiar em momentos turbulentos, além de sempre acreditar nos meus projetos. Conciliar tanto projeto em paralelo só foi possível graças a você.

Ao meu filho Gustavo, que mudou a minha vida e pelo qual tenho buscado ir sempre mais longe buscando fazer o melhor. À minha filha Yasmin, que chegou agora mais para o final do mestrado, me dando mais forças para continuar batalhando pelos meus objetivos.

Às minhas irmãs Nicole, Laura e Júlia, que vibram pela minha trajetória e pelas conquistas.

À minha tia Ana Maria, por ser um exemplo de guerreira e por ser uma referência educacional para mim.

À minha sogra Mônica e meu sogro Mario Flávio pelo suporte e parceria de sempre.

Aos professores da USF e colegas de turma, em especial ao Prof. Makilim Batista, Prof. Evandro Peixoto e Ligia, pela condução da disciplina de Seminários, pois foi a base para a construção deste projeto. Ao Prof. Lucas pela revisão imediata da ficha de pré-análise para a defesa.

Aos amigos Damião, Giovanna e Bryan, que foram grandes parceiros de turma e suporte durante todo o mestrado.

À Luara Carvalho e ao Hugo Sandall, que me deram grande apoio no início do mestrado, apoiando, tirando dúvidas e me orientando dos possíveis caminhos, mas também das dificuldades que iria enfrentar.

Aos meus colegas de início de mestrado, que se tornaram grandes amigos e sócios, Alexandre Jaloto Araê, Gustavo. Juntos iniciamos um grupo com foco inicialmente em estudar CAT (Testagem Adaptativa Computadorizada), mas que logo se tornou o CATvante, Laboratório de Testagem Computadorizada. Obrigado por todo o apoio, suporte, parceria, conselhos e por me ajudarem com a finalização do mestrado. Na reta final, agradeço em especial ao Araê, por ter abdicado de algumas de suas atividades para me apoiar de forma grandiosa.

À Paulinha, Gustavo e Araê pela realização da análise de juízes, pois sei que foi um desafio, sem isso seria impossível finalizar este projeto.

Ao Emanuel pelas contribuições com os modelos de machine learning e o prof. Vithor sobre indicações de modelos para meu projeto.

Ao Roger, por ter sido minha dupla no TCC no curso de Ciência de Computação em 2008 e às psicólogas e amigas Rosiléa e Neidecy, pela contribuição ao nosso trabalho de monografia, este que me motivou a chegar no mestrado, por conta da temática.

Ao meu primeiro orientador, o prof. Rodolfo Ambiel, que abriu as portas antes mesmo da minha decisão de iniciar um mestrado. O mundo é muito “pequeno” e o Rod foi citado em alguns textos do meu TCC na graduação em Ciência da Computação em 2008, sendo que nem o conhecia, mas depois que o conheci, passei a admirar ainda mais o seu trabalho e a pessoa humilde e parceira que ele é.

Aos integrantes da banca examinadora da qualificação, Prof. Ricardo Primi e Prof. Rodolfo, por terem aceitado o convite e pelas contribuições especiais que fizeram.

Por fim ao prof. Nelson Hauck, que me acompanhou como orientador durante a fase final do mestrado, pelas dicas, suporte e por acreditar no projeto. Seus conselhos sempre foram bem objetivos e certeiros, fazendo com que eu lapidasse o trabalho e evoluísse como pesquisador.

Resumo

Dinardi, F. (2024). *Predição de ambientes de trabalho a partir de variáveis sociodemográficas, utilizando machine learning.* Dissertação de Mestrado, Programa de Pós-Graduação Stricto Sensu em Psicologia, Universidade São Francisco, Campinas, São Paulo.

Este estudo investiga o poder preditivo das variáveis sociodemográficas sobre os interesses profissionais, utilizando técnicas de machine learning. A partir de uma análise dos dados do Exame Nacional de Desempenho dos Estudantes (Enade) de 2018 e 2019, foram aplicados modelos de regressão logística multiclasse, redes neurais e árvores de classificação e regressão (CART) para prever os tipos de interesses profissionais segundo o modelo RIASEC. Os interesses profissionais são padrões de preferências, aversão ou indiferença em relação a atividades profissionais, influenciados por fatores como autoconhecimento, experiências de vida, e feedback de pessoas significativas. Os resultados indicaram que as redes neurais apresentaram um desempenho ligeiramente melhor que a regressão logística multiclasse, com uma acurácia de 54%. As classes de interesses mais bem representadas nos dados, como CE, SI e IS, tiveram os melhores índices de precisão, variando de 0,50 a 0,67 no F1-score. Este achado ressalta a importância de uma amostra robusta para a eficácia dos modelos preditivos de machine learning. Entretanto, o estudo também revelou limitações significativas. A predição foi restrita às variáveis sociodemográficas disponíveis no Enade, o que pode não capturar todas as nuances dos fatores que influenciam os interesses profissionais dos indivíduos. Além disso, a baixa performance em classes menos representadas aponta para a necessidade de uma base de dados mais equilibrada e diversificada para futuras pesquisas. Os achados deste estudo sugerem que fatores sociais, econômicos, étnicos, familiares, pessoais e de gênero influenciam significativamente as escolhas profissionais, corroborando com a literatura existente. Para estudos futuros, recomenda-se a inclusão de uma gama mais ampla de variáveis, como dados psicométricos, históricos acadêmicos e características de personalidade, para melhorar a capacidade preditiva dos modelos. Este estudo teve como objetivo principal avaliar o poder preditivo das variáveis sociodemográficas sobre os interesses profissionais, utilizando técnicas de machine learning. Demonstrou-se a eficácia dessas técnicas na interpretação dos fatores que influenciam a formação dos interesses profissionais. O objetivo não é restringir pessoas a determinadas profissões com base em suas características sociodemográficas, mas identificar padrões que possam auxiliar na orientação de carreira e evidenciar a importância dos programas governamentais. Assim, busca-se fornecer insights para aprimorar políticas públicas e iniciativas de apoio à escolha profissional.

Palavras-chave: inteligência artificial, predição, interesses vocacionais, análise de dados, avaliação psicológica

Abstract

Dinardi, F. (2024). *Prediction of Work Environments from Sociodemographic Variables Using Machine Learning*. Master's Dissertation, Stricto Sensu Post-graduate Program in Psychology, São Francisco University, Campinas, São Paulo.

This study investigates the predictive power of sociodemographic variables on professional interests using machine learning techniques. Based on an analysis of data from the National Student Performance Exam (Enade) of 2018 and 2019, models such as multinomial logistic regression, neural networks, and classification and regression trees (CART) were applied to predict types of professional interests according to the RIASEC model. Professional interests are patterns of preferences, aversions, or indifference towards professional activities, influenced by factors such as self-knowledge, life experiences, and feedback from significant individuals. The results indicated that neural networks performed slightly better than multinomial logistic regression, with an accuracy of 54%. The most well-represented interest classes in the data, such as CE, SI, and IS, had the best precision indices, ranging from 0.50 to 0.67 in F1-score. This finding highlights the importance of a robust sample for the effectiveness of machine learning predictive models. However, the study also revealed significant limitations. The prediction was limited to the sociodemographic variables available in Enade, which may not capture all nuances of the factors influencing individuals' professional interests. Furthermore, the low performance in less represented classes points to the need for a more balanced and diversified database for future research. The findings of this study suggest that social, economic, ethnic, family, personal, and gender factors significantly influence professional choices, corroborating existing literature. For future studies, it is recommended to include a broader range of variables, such as psychometric data, academic histories, and personality traits, to improve the predictive capacity of the models. This study's primary objective was to assess the predictive power of sociodemographic variables on professional interests using machine learning techniques. These techniques' effectiveness was demonstrated in interpreting the factors influencing the formation of professional interests. The goal is not to restrict people to specific professions based on their sociodemographic characteristics but to identify patterns that can aid in career guidance and highlight the importance of government programs. Thus, the aim is to provide insights to improve public policies and initiatives supporting professional choice.

Keywords: artificial intelligence, prediction, vocational interests, data analysis, psychological assessment

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Anexos

Introdução

Historicamente, a psicologia sempre lidou com o problema da falta de representatividade e a baixa generalização dos resultados (Franco, 2021). Com o advento do *big data* (grande volume de dados) e da utilização de técnicas de *machine learning (*aprendizagem de máquina), podemos ter resultados melhores na predição do comportamento humano, como por exemplo no domínio da investigação dos interesses profissionais (Bogacheva et al. (2020). Além disso, Yarkoni e Westfall (2017), defensores e promotores da aplicação de métodos de *machine learning* para a psicologia, estas técnicas podem tornar a esta ciência ainda mais preditiva. *Machine Learning* é uma subárea de Inteligência Artificial (IA), que é um campo transdisciplinar e tem raízes na lógica, estatística, psicologia cognitiva, teoria da decisão, neurociência, linguística, cibernética e engenharia da computação (Howard, 2019). Aplicações que usam técnicas de IA estão cada vez mais presentes no cotidiano das pessoas, como é o caso dos sites de busca, *e-commerce*, sistemas de recomendação de serviços, reconhecimento de imagens, sensores, tradução de textos, entre outros. Já no campo da avaliação psicológica, a IA tem sido pouco explorada e para Primi (2018), os eventos científicos e históricos ocorridos nos últimos anos, especialmente no âmbito da IA, apontam para um novo papel da avaliação psicológica no mundo. Portanto, este trabalho visa testar o poder preditivo das variáveis sociodemográficas sobre os interesses profissionais, utilizando técnicas de machine learning. É fundamental destacar que o objetivo não é afirmar que pessoas com certas características sociodemográficas estão restritas a determinados interesses ou profissões, mas sim identificar padrões e tendências que possam auxiliar na orientação de carreira e destacar a importância dos programas governamentais. Essa abordagem busca fornecer insights valiosos para melhorar as políticas públicas e iniciativas de apoio à escolha profissional.

Interesses Profissionais/Vocacionais

Uma frase muito comum que toda criança ouve de pais, amigos e professores é “O que você quer fazer quando você crescer?”, e a resposta a esta questão se torna mais importante à medida que a criança cresce e precisa fazer algumas escolhas educacionais e ocupacionais (Rounds & Su, 2014). Entretanto, estas escolhas podem não ser uma tarefa fácil para algumas pessoas, principalmente quando não se tem muita diferenciação dos principais gostos ou interesses profissionais.

Segundo Lent et al. (1994), os interesses profissionais podem ser definidos como padrões de preferências, aversão ou indiferença acerca de atividades profissionais. Alguns outros autores vão um pouco além desta definição, relatando que os interesses refletem as preferências pessoais por comportamentos, situações, contextos em que as atividades ocorrem e/ou resultados associados com atividades preferidas (Rounds, 1995; Su, Rounds, & Armstrong, 2009). Na visão de Stoll e Trautwein (2017), os interesses gerais começam a se desenvolver na infância, tendo suas primeiras manifestações na preferência por brinquedos, mudando bastante a partir de objetos e atividades, mas também influenciados pela cultura e gênero. Os autores destacam que com a entrada na escola, os interesses começam a se individualizar, sendo reavaliados entre 11 e 13 anos, período que iniciam as diferenciações.

No ensino médio, a partir de uma visão mais realista, os indivíduos começam a decidir de fato o que gostam e o que não gostam, diminuindo assim o nível de interesse em algumas áreas (Stoll & Trautwein, 2017). De acordo com Holland (1997), é nesta fase da adolescência que os interesses começam a se diferenciar, ou seja, o nível de interesse aumenta em algumas atividades e diminui em outras, algo que é fundamental para uma futura escolha profissional. Para uma melhor compreensão de como são formados os interesses profissionais, existem algumas teorias com, com destaque para a Teoria Desenvolvimentista de Super (Super, 1957); Teoria da Personalidade Vocacional e Ambientes de Trabalho de

John Holland (Holland, 1959;1997); Teoria Sócio-Cognitiva do Desenvolvimento de Carreira (TSCDC) (Lent, Brow e Hackett, 1994); e Teoria da Construção de Carreira (Savickas, 2005).

De acordo com a Teoria Desenvolvimentista, existem cinco estágios de vida, sendo o primeiro deles o estágio de crescimento, onde os interesses surgem de forma fantasiosa, com tendência a se estabilizar na adolescência, por meio do autoconhecimento, exploração de papéis, atividades de lazer, entre outros (Super & Jordaan, 1973). Na TSCDC, por meio do envolvimento repetido em certas atividades, da aprendizagem por observação e o recebimento de feedback de pessoas significativas em seu percurso de vida, as pessoas refinam suas habilidades, formando suas crenças de autoeficácia e expectativas de resultados, que influenciam a formação dos interesses profissionais (Lent et al., 1994). Além disso, os mesmos autores consideram que os fatores sociais, físicos, educativos e financeiros podem interferir na escolha de um trabalho, podendo ir no sentido contrário aos interesses. Na Teoria da Construção de Carreira, Savickas (1999), define interesses como esforço de adaptação para usar o ambiente a favor das necessidades e valores, influenciados pela hereditariedade, aprendizagem, crenças de capacidade, autoconceito e identificação com modelos e papéis sociais.

Por fim, na Teoria da Personalidade Vocacional e Ambientes de Trabalho, os interesses profissionais são resultantes de aspectos do desenvolvimento e dos fatores ambientais, podendo ser agrupados em seis tipologias com definições e características próprias, a saber: Realista (R), Investigativo (I), Artístico (A), Social (S), Empreendedor (E) e Convencional (C) (Holland, 1966/1975). Nas próximas duas seções, serão abordadas em maior profundidade as teorias que foram adotadas neste projeto, a TSCDC e a Teoria da Personalidade Vocacional e Ambientes de Trabalho.

Para avaliação dos interesses profissionais, acima conceituados e com base nas teorias citadas, existem alguns inventários, que de acordo com Rounds e Su (2014), os primeiros instrumentos foram desenvolvidos no início do século 20. Os resultados das avaliações dos interesses vocacionais são usados em vários processos de tomada de decisão, como o que cursar em uma universidade, que trabalho seguir e onde trabalhar (Hoff et al., 2019). E foi graças a pesquisa que os inventários foram desenvolvidos, pois, de acordo com Ambiel et al. (2016), houve grande foco na construção e uso dos instrumentos e isso foi observado desde as primeiras publicações. Como exemplo, podemos citar o questionário de autoavaliação para alunos do ensino fundamental criado por Jesse Davis em 1914 e o primeiro inventário padronizado de interesses profissionais, de 1920, publicado pelo Instituto Carnegie (Ambiel et al., 2016). Os mesmos autores relatam que por muito tempo o foco da pesquisa se deu em adultos e estudantes universitários e isso ainda é uma realidade.

Ambiel et al. (2016) ainda relatam que a utilização de instrumentos de interesses profissionais vai além da Orientação Profissional e de Carreira (OPC), uma das únicas áreas de aplicação no Brasil. A OPC também foi definida como um processo que tem como principal objetivo orientar as pessoas com relação as suas carreiras a fim de proporcionar a elas uma maior compreensão das características das profissões; além de um maior autoconhecimento, despertando, assim, potencialidades até o momento desconhecidas (Savickas, 1999). Para Su e Rounds (2014), os interesses predizem prestígio ocupacional e vários indicadores de sucesso acadêmico (i.e., obtenção de diploma, persistência e notas na faculdade).

Entretanto, a partir das mudanças no mercado de trabalho e do atual contexto econômico, os inventários também podem auxiliar as pessoas além da primeira escolha profissional dos adolescentes (Ambiel et al., 2016). Para Harrington e Long (2013), alguns grupos como estudantes do ensino superior, profissionais iniciantes e experientes que estão

desempregados e estudantes que saem da escola básica com déficits em competências para atuação profissional, poderiam ser beneficiados pelas avaliações de interesses profissionais.

Por fim, o modelo RIASEC de Holland (1997) é baseado na suposição de que escolhas profissionais adequadas são uma função de uma correspondência bem-sucedida entre a personalidade vocacional de um indivíduo (por exemplo, Social) e o ambiente de trabalho (por exemplo, alto contato interpessoal). Portanto, o uso do modelo RIASEC na orientação de carreira requer um método para vincular os interesses de um indivíduo, operacionalizados como seu perfil RIASEC, às demandas ocupacionais que também devem ser expressas em códigos RIASEC.

Teoria Social Cognitiva do Desenvolvimento de Carreira

A TSCDC foi desenvolvida com base na Teoria Social Cognitiva (TSC) de Bandura (1986), que diz que o comportamento humano é compreendido como uma expressão da interação constante entre a pessoa e o meio em que está inserida. A TSC privilegia os processos internos e, principalmente, as relações recíprocas entre cognições (crenças e expectativas), comportamento ostensivo (expresso) e ambiente - reciprocidade triádica. Segundo Lent et al. (1994), a TSCDC busca compreender como as pessoas formam seus interesses (acadêmicos e de carreira) e, a partir deles, selecionam opções acadêmicas e/ou profissionais e como ingressam, desempenham e persistem em atividades educacionais e de carreira.

O principal objetivo da TSCDC é enfatizar as capacidades pessoais em relação às circunstâncias ambientais que influenciam as escolhas profissionais ao longo do processo de desenvolvimento de carreira (Lent et al., 1994). Os autores afirmam que a estrutura da TSCDC foi formulada para compreender três etapas do desenvolvimento de carreira: (a)

formação e elaboração de interesses profissionais; (b) escolhas acadêmicas e profissionais; e (c) desempenho e persistência em atividades educacionais e profissionais.

Com base na TSC (Bandura, 1986), a TSCDC considera três mecanismos sociais cognitivos, que são importantes no desenvolvimento de carreira: (a) crenças de autoeficácia, (b) expectativas de resultado e (c) estabelecimento de objetivos/metas. A autoeficácia no contexto de carreira, é concebida como as crenças dos indivíduos em relação à própria capacidade de se envolver/engajar em tarefas relativas à escolha profissional. De acordo com Lent et al. (1994), as fontes de autoeficácia são mecanismos pelos quais as pessoas desenvolvem suas crenças de capacidade. Estas são desenvolvidas pela experiência própria, aprendizagem vicária (observação de outras pessoas), feedbacks sociais, indicadores fisiológicos e emocionais. Os autores reiteram que os indivíduos tendem a evitar tarefas e situações que acreditam exceder sua capacidade e, por outro lado, optam por profissões que envolvem atividades que elas acreditam ser capazes de realizar com sucesso.

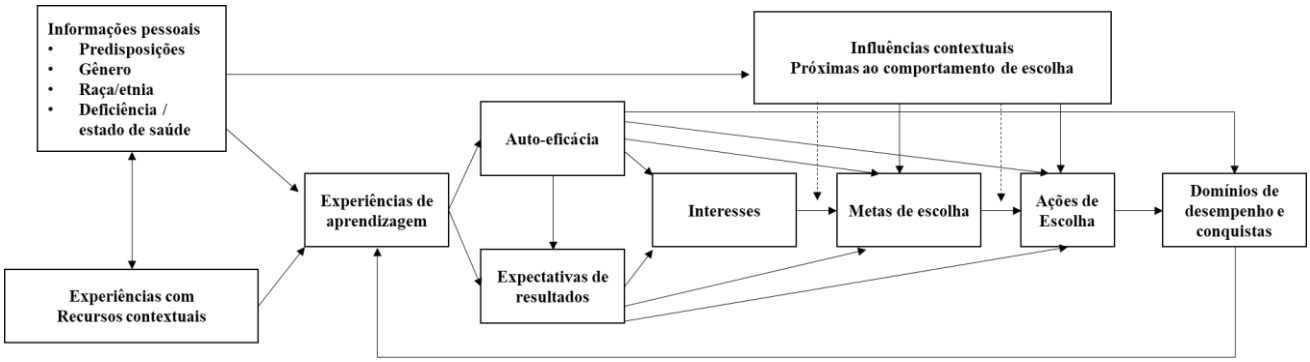
O segundo mecanismo, expectativas de resultado, são as consequências que a pessoas imaginam ao realizar determinados comportamentos, ou seja, as crenças sobre as consequências observáveis e não observáveis das ações (Lent et al., 1994). Isto é, envolve os prováveis resultados das ações, aspectos físicos (recompensas financeiras), sociais (aprovação de colegas ou familiares) ou de autoavaliação (satisfação pessoal). Os autores relatam que o terceiro mecanismo, estabelecimento de metas, ajudam as pessoas a organizarem e orientarem seus comportamentos, sustentá-los por períodos mais longos, mesmo sem reforço externo, aumentando a probabilidade de os resultados serem alcançados.

De acordo com a TSCDC, a formação das crenças de autoeficácia e expectativas de resultado ocorre por meio do envolvimento repetido com certas atividades, da aprendizagem por observação e o recebimento de *feedback* de pessoas significativas em seu percurso de vida, o que refina a habilidade das pessoas (Lent et al., 1994). Além disso, as crenças de

autoeficácia e expectativas de resultado possuem grande influência na formação dos interesses profissionais. Na Figura 1, podemos ver o modelo proposto por Lent et al. (1994).

Figura 1

Adaptado do modelo de fator pessoal, contextual e experiencial que afetam o comportamento das escolhas relacionados à carreira (Lent et al., 1994)



A partir da Figura 1, podemos perceber que os interesses profissionais segundo Lent et al. (1994), têm influência de *inputs* pessoais (predisposição, sexo, raça/etnia, deficiência/status de saúde), características contextuais, experiências de aprendizagem, expectativas de resultado e autoeficácia. Holland (1997) também afirmou que durante os anos de formação, as pessoas geralmente desenvolvem padrões de características de interesses de carreira. Lent et al. (1994) reforçaram ainda que a interação ou observação de membros da família, pares, outras pessoas significantes, a própria cultura e instituições religiosas transmitem muitos valores e padrões pessoais de comportamento e isso influencia os interesses profissionais. Após explorar a TSCDC e como os interesses profissionais se formam, iremos explorar na próxima sessão a Teoria da Personalidade Vocacional e Ambientes de Trabalho,

Teoria de Personalidade Vocacional e Ambientes de Trabalhos

A Teoria da Personalidade Vocacional e Ambientes de Trabalho (Holland, 1959;1997), é uma das teorias mais relevantes no campo da OPC, além de ser uma das mais

estudadas e replicadas em diferentes culturas e línguas da história da Psicologia (Ambiel et al., 2016; Nye et al., 2017). Holland (1966/1975), define interesses profissionais como expressão da personalidade vocacional de uma pessoa por meio das relações estabelecidas com o ambiente laboral e suas preferências vocacionais. As seis tipologias são conhecidas pelo acrônimo RIASEC e estão definidas na Tabela 1.

Tabela 1

Definições dos tipos do RIASEC

Tipos do RIASEC	*Definição*
Realista (R)	Preferências por atividades técnicas ou ocupações ao ar livre, envolvendo o uso de equipamentos e tecnologia, além de exigir habilidades manuais e práticas.
Investigativo (I)	Agrupa interesses em pensar e atividades de pesquisa e vocações, lidando com *construção de teoria, resolução abstrata de problemas e conhecimento de metodologia científica.*
Artístico (A)	Preferências para criar e desenvolver coisas novas, onde a beleza e o design são ingredientes essenciais.
Social (S)	Preferências em relação à interação com pessoas, como educação, treinamento, cuidado e atividades de enfermagem.
Empreendedor (E)	Gosta de agir e fazer, através de atividades

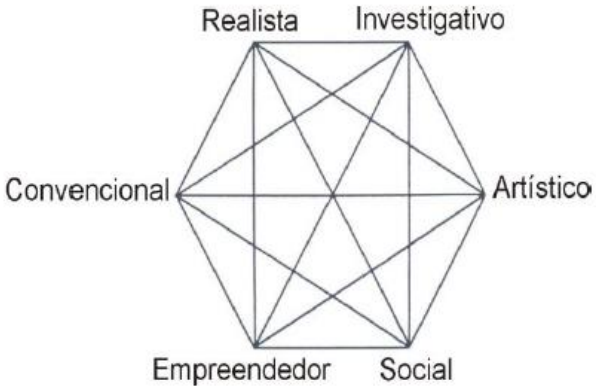
	como a implementação, organização e liderança.
Convencional (C)	Preferências sobre a aplicação correta de regras e padrões, articulados em vocações como contador ou controlador de qualidade.

Nota: Tradução livre de Ambiel et al. (2018)

Ambiel et al. (2018) relatam que os seis tipos têm variados graus de dependência e são representados por uma hexágono, com alguns tipos adjacentes uns aos outros, conforme figura 2. Cada tipo do RIASEC é formado por vários interesses, atividades preferidas, crenças, habilidades, valores e características (Nauta, 2010). Além disso, a teoria possui alguns pressupostos, como Congruência, Consistência, Diferenciação e Identidade.

Figura 2

Modelo hexagonal proposto por Holland (1997)



O conceito de congruência significa o grau de ajuste entre o tipo de personalidade do indivíduo e o tipo de ambiente de trabalho e se divide em três tipos. O primeiro grau de congruência é aquele em que o tipo de personalidade vocacional é idêntico ao ambiente (exemplo: pessoa do tipo Realista trabalhando em ambiente Realista). Já o segundo grau de

congruência é aquele em que o tipo de personalidade dominante atua em um ambiente adjacente lateralmente (exemplo: pessoa do tipo Convencional trabalhando em ambiente Empreendedor). Por último, o terceiro grau é considerado como incongruente, ou seja, quando o tipo de personalidade principal está em um ambiente contrário no hexágono (exemplo: pessoa do tipo Convencional exercendo uma atividade laboral em ambiente Artístico).

Já a consistência significa a manutenção do padrão da personalidade e do padrão do ambiente. Também é verificada pela integração de interesses, competências, valores e percepções, sendo que pessoas com alto grau de consistência são mais previsíveis e resistentes a influências externas. Por outro lado, indivíduos inconsistentes são menos previsíveis por possuírem um repertório comportamental mais extenso por não terem clareza dos interesses e demais variáveis. A consistência em relação ao ambiente é percebida quando há integração das demandas e recompensas dos indivíduos, além de uma pressão para um comportamento compatível. Quando se tem uma gama grande de demandas e recompensas, ocorre a inconsistência ambiental, o que tornam os espaços menos dominantes sob os comportamentos dos sujeitos que o habitam e, que neste caso, são mais desorganizados (Holland, 1975).

Os graus de congruência e consistência são afetados diretamente pelo terceiro pressuposto, conhecido como diferenciação (nitidez), que é identificada por meio da magnitude das pontuações médias nos escores de cada tipologia RIASEC. O ideal é que uma pessoa tenha altas e baixas pontuações nos seis tipos de personalidade vocacional, ou seja, quanto maior a pontuação em uma determinada tipologia em detrimento de outras, mais clareza tem em relação aos interesses profissionais. Portanto, a diferenciação e a indiferenciação da personalidade ou dos padrões ambientais permitem prever a ocorrência de comportamentos desejáveis ou indesejáveis para cada tipo e ambiente (Holland, 1985).

Por último, temos a identidade, que foi um pressuposto adicionado durante a revisão da teoria (Holland, 1985), que, segundo o autor, afeta as interações entre pessoa e ambiente, congruência, consistência e diferenciação. De acordo com cada tipologia, a identidade é verificada de acordo com as características desejáveis e esperadas (exemplo: identidade Artística implica em alta capacidade de relacionamento interpessoal). Em relação ao ambiente, a identidade é determinada de acordo com os comportamentos que são desejáveis nele.

A investigação e entendimento destes pressupostos permitem um melhor ajustamento entre os interesses do indivíduo e as demandas do ambiente de trabalho, porém a visão atual da área de OPC não corrobora com a ideia de ajustamento proposto por Parsons (1909/2005) com o pressuposto "homem certo para o lugar certo". Portanto, o foco da utilização dos instrumentos não deve ser no diagnóstico ou no caráter prescritivo, onde o objetivo era combinar características pessoas com o ambiente, mas sim no apoio da promoção do autoconhecimento (Ambiel, 2019).

Variáveis Sociodemográficas, Econômicas

Para Lent et al. (1994), a escolha de carreira tende a ser consistente com os interesses vocacionais primários (interesses profissionais) – suposição comum entre as teorias de carreira, mas sugerem que esta relação pode ser afetada por importantes recursos contextuais. Os mesmos autores relatam que condições socieconômicas e educacionais ótimas permitem que as pessoas traduzam seus interesses primários em objetivos de carreira correspondentes. Mas nem sempre isso ocorre, os interesses podem estar comprometidos com as necessidades econômicas.

Neste caso, objetivos de carreira e ações podem ser influenciadas menos pelos interesses do que pela disponibilidade de trabalho, pela autoeficácia e pelos resultados esperados (Lent et

al., 1994). Ou seja, além do interesse, fatores sociais, econômicos, étnicos, familiares, pessoais e de gênero, por exemplo, podem afetar as escolhas profissionais (Lima et al., 2017). Além disso, em momentos de tomada de decisão, como escolher um curso ou uma profissão, as variáveis familiares, como a transmissão de conhecimento sobre o trabalho pelos pais, o envolvimento com os filhos, as aspirações e expectativas dos pais e o status socioeconômico e familiar são relevantes (Bryant et al., 2006). Para Lent et al. (1994), de acordo com a TSCDC, as relações entre metas de escolha e metas de ações tendem a ser mais fortes com pessoas que percebem condições ambientais benéficas (presença de amplo suporte, menos barreiras) e menores nas pessoas que percebem condições menos favoráveis.

Por conta dessas diferenças por conta das desigualdades, a literatura recente tem crescido o foco nos elementos sociais e culturais que permeiam o desenvolvimento de carreira, como exemplo, podemos citar, a Teoria da Psicologia do Trabalhar (*Psychology of Working Theory*) de Duffy et al. (2016). Os autores ressaltam que a classe social tem papel fundamental no acesso aos recursos econômicos, bem como ao capital social e cultural, que facilita o desenvolvimento de carreira e acesso ao trabalho decente.

De acordo com o apresentado anteriormente, variáveis sociodemográficas e econômicas podem predizer os interesses profissionais, além de ter grande influência na escolha de uma carreira/profissão, podendo também influenciar na seleção de um curso superior. O uso das avaliações de interesse para fins preditivos é apoiado por décadas de pesquisas que mostram que os interesses são altamente estáveis ao longo do tempo e preveem vários resultados acadêmicos e de carreira importantes (Low et al., 2005; Nye et al., 2012, 2017; Rounds & Su, 2014; Van Iddekinge et al., 2011). Um estudo recente realizado por Bogacheva et al. (2020), revelou que fatores sociodemográficos, incluindo sexo, raça, religião e idade, podem ser considerados fortes preditores de interesses vocacionais/profissionais e que isso pode ter implicações práticas de longo alcance para aconselhamento profissional.

Entretanto, a Teoria da Congruência dos Papéis (Diekman & Eagly, 2008) e a Teoria da Circunscrição do Compromisso (Gottfredson, 1981), sugerem que a expressão dos interesses sofre restrições da ordem social, de forma mais ampla que a própria vontade do indivíduo. Ou seja, gênero, raça/etnia e classe social, por exemplo, pode influenciar na rejeição de uma determinada atividade, por ser considerada além das barreiras do "espaço social" (Su, 2018). Portanto, por mais que uma pessoa tenha o interesse em alguma profissão que seja do tipo Realista, por exemplo, ela pode não manifestar interesse por achar que isso é incompatível com sua identidade de gênero.

Alguns estudos de meta-análise sobre resultados em inventários de interesses profissionais, como diferenças de gênero (Su et al., 2009) e diferenças raciais (Roth et al., 2017), fornecem evidências indiretas para a hipótese de que as pessoas podem expressar interesses alinhados com as expectativas de seu papel social ou dos próprios valores culturais. Ainda sobre o estudo de Su et al. (2009), a partir da análise de 47 inventários de interesse, com uma amostra de 503.188 respondentes, revelou grandes diferenças de gênero, especialmente para o tipo Realista (d = 0.84 maior para homens) e para o Social (d = -0.68 maior para mulheres). Este mesmo estudo demonstra que também existem grandes diferenças nas áreas de *STEM* (ciência, tecnologia, engenharia e matemática) em favor dos homens, sendo para engenharia (d = 1.11), ciência (d = 0.36) e matemática (d = 0.34).

De acordo com outra meta-análise, os autores Ng et al. (2005) encontraram 4 categorias como antecessoras do sucesso profissional, sendo que as variáveis sociodemográficas (por exemplo, sexo, raça, idade e estado civil) podem ser percursores deste "sucesso". Quanto as diferenças raciais (preto-branco, hispânico-branco, branco-asiático), de acordo com os autores Roth et al. (2017), elas são pequenas em interesses vocacionais. No entanto, Su et al. (2009) comentam que alguns grupos minoritários relatam

interesses mais baixos ou mais altos em alguns tipos de trabalho e isso acaba sendo pouco explorado na literatura.

Conforme apresentado acima, existem alguns estudos que tentam relacionar as variáveis sociodemográficas aos interesses profissionais, avaliando desde as diferenças entre sexo, raça, idade, estado civil, classe social, entre outros e qual a influência estas características dos indivíduos têm na escolha de uma profissão e/ou curso superior. Cabe ressaltar, que diante do contexto de desigualdade social no Brasil, que foi escancarada por conta da pandemia do coronavírus (Nascimento & Massi, 2021), esta avaliação de predição em relação as variáveis sociodemográficas são de grande importância para o contexto brasileiro. Na próxima seção, será apresentado o Exame Nacional de Desempenho dos Estudantes (Enade), uma fonte de dados aberta, com coleta padronizada, amostra heterogênea, com grande número de sujeitos, que engloba diferentes características sociodemográficas, voltadas tanto para características socioeconômicas quanto para motivos de ingresso e características do curso.

Enade

Conforme consta no site do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2021), o Enade tem o objetivo de: (a) avaliar o rendimento dos concluintes dos cursos de graduação em relação aos conteúdos programáticos previstos nas diretrizes curriculares dos cursos, (b) o desenvolvimento de competências e habilidades necessárias ao aprofundamento da formação geral e profissional, e (c) o nível de atualização dos estudantes com relação à realidade brasileira e mundial. Ele integra o Sistema Nacional de Avaliação da Educação Superior (Sinaes), composto também pela Avaliação de cursos de graduação e pela Avaliação institucional. Juntos eles formam o tripé avaliativo que permite conhecer a qualidade dos cursos e instituições de educação superior brasileiras. Os resultados

do Enade, aliados às respostas do Questionário do Estudante, são insumos para o cálculo dos Indicadores de Qualidade da Educação Superior.

A inscrição é obrigatória para estudantes ingressantes e concluintes habilitados de cursos de bacharelado e superiores de tecnologia vinculados às áreas de avaliação da edição. A situação de regularidade do estudante é registrada no histórico escolar. Além disso, os alunos que fazem a prova do Enade respondem um questionário socioeconômico, que traz bastante insumos para pesquisas como este projeto de dissertação (INEP, 2021).

Alguns estudos já foram realizados utilizando os dados do Enade e de acordo com uma revisão sistemática de Lima et al. (2019) no Google Acadêmico, entre 2005 e 2016, foram encontradas 40 pesquisas. Os objetivos destes estudos foram: (a) avaliar o conhecimento e conteúdo da prova (4); (b) avaliar questões administrativas (4); (c) o desempenho/rendimento dos estudantes no exame (19); (d) testar ou desenvolver ferramentas para avaliar o exame (5); (e) investigar a estrutura do exame (3); e (f) analisar a formação do docente e relacioná-la com o desempenho dos estudantes nas provas (5) (Lima et al., 2019). Em relação ao tipo de dados usados, 16 trabalhos usaram a nota do Enade, 10 o questionário socioeconômico, 8 o conteúdo da prova e o conceito do Enade, e apenas 1 utilizou o questionário do coordenador.

Ainda de acordo com os autores, para a análise de dados, grande parte dos estudos utilizou a estatística descritiva, além de 5 com regressão linear, 3 com análise fatorial e apenas 2 com *data mining* (Lima et al., 2019). De acordo com Shafique e Qaiser (2014), *data mining* tem o objetivo de analisar e extrair informação de grandes quantidades de dados, com diferentes propósitos. Para realizar a tarefa de *data mining*, existem alguns modelos de processos, como o *Knowledge Discovery Databases*, que segundo Gimenes (2000), envolve várias disciplinas de áreas relativas a aprendizado de máquina (*machine learning*), reconhecimento de padrões, bases de dados, estatística e matemática, aquisição de

conhecimento para sistemas especialistas e visualização de dados. Na próxima seção, iremos apresentar com mais detalhes a técnica de *machine learning*, por ser a utilizada neste projeto.

Além dos estudos citados acima, um outro mais recente, conduzido por (Nascimento & Massi, 2021), analisou dados de 861 mil estudantes que realizaram o Enade entre 2016 e 2018, para avaliar a relação dos motivos que levam à escolha de um determinado curso de graduação. Neste mesmo estudo, foram identificadas que as classes sociais C, D e E optam pela graduação principalmente para se inserirem no mercado de trabalho, além de escolherem mais os cursos à distância. Já as pessoas de classe média escolhem cursos de baixa concorrência, mas que tenham mais valorização social. Por último, a motivação da escolha pela classe alta se dá por influência familiar ou pela vocação.

Machine Learning

O termo IA foi usado pela primeira vez em um *workshop* de verão em 1956 no *Dartmouth College* em *New Hampshire* (Howard, 2019). Nesta ocasião, McCarthy et al. (1955) definiram o "problema da inteligência artificial" como o de "fazer a máquina se comportar de maneiras que ela poderia ser chamada de inteligente se um humano se comportasse dessa maneira". A IA é composta por algumas subáreas, mas a área considerada essencial é a de *machine learning* (Dasgupta, & Nath, 2016). Esta área tem como objetivo a aprendizagem dos computadores a partir dos dados (Howard, 2019), para ajudar na geração de *insights* úteis, fazer predições e ajudar na tomada de decisão (Jordan & Mitchell, 2015).

Machine Learning possui categorias de algoritmos específicos, que são utilizados de acordo com o problema a ser resolvido, tais como: aprendizado supervisionado, aprendizado não-supervisionado, aprendizado por reforço e aprendizado semi-supervisionado (Zahour et al., 2020). No caso do aprendizado supervisionado, os rótulos dos dados são conhecidos e o que se espera de saída (resultado) também, por exemplo, identificar câncer de pulmão a partir

de imagens de raio-x (Choy et al., 2018); no aprendizado não-supervisionado, os rótulos de dados e as saídas esperadas não são conhecidas, por exemplo, ajudar especialistas de marketing no processo de segmentação (Sánchez-Hernández et al., 2013); no aprendizado por reforço, existem regras baseadas em recompensa, como na robótica, onde os robôs precisam ter a capacidade de aprender, melhorar, adaptar e reproduzir tarefas (Kormushev & Caldwell, 2013); o aprendizado semi-supervisionado é formado por técnicas do aprendizado supervisionado e não-supervisionado.

Além disso, de acordo com Zahour et al. (2020), cada categoria de algoritmos pode conter múltiplos sub-algoritmos e tipos, conforme Tabela 2. Quanto aos algoritmos de aprendizado supervisionado apresentados pelos autores, cabe acrescentar a Regressão logística multiclasse, Árvores de Classificação e Regressão (CART) e *Gradient Boosting*. Os autores também comentam que os algoritmos podem ser paramétricos, onde se conhece a informação da população, o que não ocorre com os algoritmos não-paramétricos.

Tabela 2

Algoritmos de machine learning

Aprendizado supervisionado	Aprendizado não-supervisionado	Aprendizado por reforço
Redes neurais artificiais	Redes neurais artificiais	Aprendizagem-Q
Estatística bayesiana	Aprendizagem por regras de associação	Autômatos de aprendizagem
Autômatos de aprendizagem	Clusterização por particionamento	
Aprendizagem baseada em instância		
Classificadores lineares		
Árvores de decisão		
Redes bayesianas		
Modelos de Markov Ocultos		

Nota: Traduzido de Zahour et al. (2020)

A partir dos algoritmos apresentados na Tabela 2, vamos focar no aprendizado supervisionado, que podem ser utilizados em tarefas de classificação e regressão linear. A diferença entre as duas é que a regressão linear é utilizada quando as saídas são variáveis contínuas, já a classificação, quando os resultados são categóricos (Sen et al., 2020). De acordo com os mesmos autores, a classificação pode ser binária ou multiclasse (*multi-label*). A classificação é composta por 2 etapas, a construção do modelo e a classificação dos códigos RIASEC ou registros (Gorade et al., 2017), conforme representadas nas Figuras 3 e 4.

Figura 3

Etapa 1 – Construção do modelo, adaptado de Gorade et al. (2017)

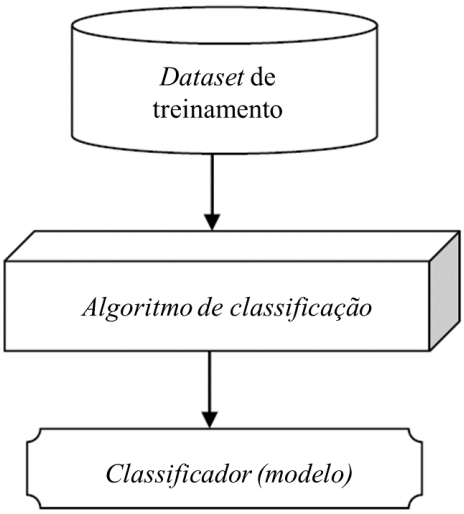
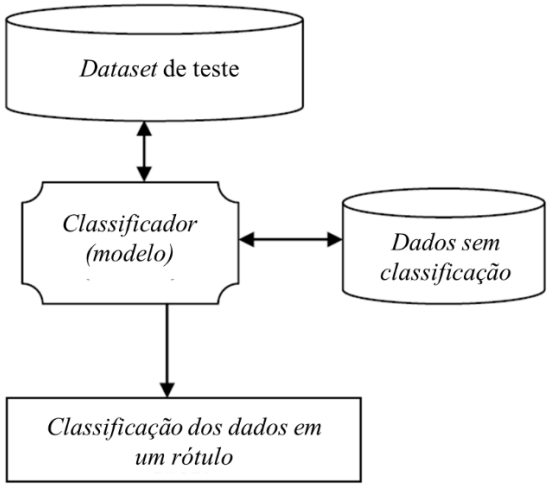


Figura 4

Etapa 2 – Modelo usado para classificação de registros desconhecidos, adaptado de Gorade et al. (2017)



Além das duas etapas demonstradas, os classificadores possuem algumas características importantes, a saber: (a) correção, (b) tempo, (c) força, (d) tamanho dos dados e (e) extensibilidade (Gorate et al., 2017). Os autores definem correção como a precisão na classificação dos dados, ou seja, avalia se os registros são classificados corretamente ou incorretamente. Já o tempo, tem a ver com os custos computacionais na tarefa de classificação. A força tem a ver com a capacidade de classificar os dados corretamente, mesmo que os mesmos tenham ruídos (valores ausentes ou incorretos). Os classificadores devem ser independentes e escalonáveis, independentemente do tamanho do banco de dados. Por fim, a extensibilidade, que tem a ver com a adição de novos recursos sempre que necessário, mas trata-se de um recurso de difícil implementação (Gorate et al., 2017).

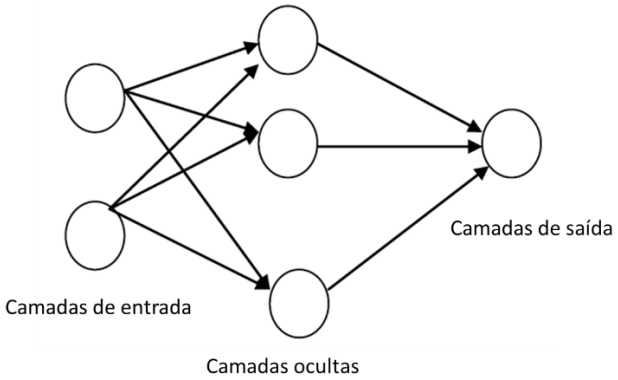
De acordo com Sen et al (2020), na classificação binária, podem existir apenas dois resultados possíveis, como por exemplo, prever se irá chover ou não, ou prever se um e-mail pode ser classificado como spam ou não. Já na classificação multiclasse, tem-se a possibilidade de ter além de dois resultados possíveis, como exemplo, classificar o

desempenho dos alunos entre bom, médio ou ruim (Sen et al., 2020). Um outro exemplo é classificar os interesses profissionais dos indivíduos de acordo com a letra dominante de cada perfil tipológico do RIASEC (exemplo, determinada pessoa tem o perfil R, mas poderia ser I, A, S, E ou C). Existem alguns algoritmos para resolução dos problemas de classificação multiclasse, como exemplo temos as Redes Neurais multi-classe, CART e Regressão logística multiclasse.

Os algoritmos de redes neurais são baseados no sistema nervoso biológico, com vários elementos inter-relacionados, conhecidos como neurônios e têm como objetivo o reconhecimento de padrões (Gorade et al., 2017). Os autores comentam ainda que um modelo de rede neural é composto por 3 camadas, as de entrada, as ocultas e as de saída, como exemplificado na figura 3. Além disso, eles relatam que as camadas de entrada e saída estão conectadas por gráficos acíclicos, que possuem bordas ponderadas e nós. Para calcular a saída da rede neural (por exemplo R, I, A, S, E ou C), de acordo com entradas específicas (variáveis sociodemográficas), cada nó da camada oculta e da camada de saída, tem um valor calculado (Zahour et al., 2020). Este valor, segundo os mesmos autores, é a soma ponderada dos valores dos nós da camada anterior, além de ter uma função de ativação a essa soma.

Figura 5

Exemplo de uma rede neural, adaptado de Gorade et al. (2017)



Já o algoritmo de Regressão logística multiclasse, tem o objetivo de prever a probabilidade de um evento ocorrer, ajustando os dados para uma função logística e pode ser usado na predição de múltiplos resultados (Zahour et al., 2020). Os autores relatam ainda que ambos os modelos são treinados a partir de dados já rotulados, ou seja, conhecidos. Após a escolha do modelo, as novas entradas de dados, até então desconhecidas, podem ser preditas.

O modelo CART, pode ser usado tanto para classificação quanto para regresso. Neste modelo, é feito um particionamento nos dados e dentro de cada partição são feitas previsões simples, que podem ser representadas por uma árvore de decisão. Para problemas de classificação, as árvores são projetadas para variáveis dependentes que tem um número finito de valores não ordenados (Loh, 2011).

Por fim, o modelo *Gradient Boosting* é uma técnica de aprendizado de máquina que combina vários modelos fracos, geralmente árvores de decisão, para formar um modelo forte e robusto. A cada iteração, o algoritmo ajusta o modelo atual corrigindo os erros das previsões anteriores, minimizando uma função de perda através de uma abordagem de gradiente descendente (Friedman, 2001). O gradiente descendente é um algoritmo de otimização amplamente utilizado em aprendizado de máquina e inteligência artificial. Ele busca minimizar (ou maximizar) uma função de custo ajustando iterativamente os parâmetros do modelo.

Diante do exposto, vamos apresentar alguns estudos que utilizaram técnicas de *machine learning* com foco nas questões acadêmicas e de orientação vocacional. Zahour et al. (2020), fizeram um estudo comparativo para a tarefa de classificação de questões de orientação escolar baseado nos seis tipos do RIASEC, de John Holland e identificaram melhores resultados a partir dos algoritmos de redes neurais. Os autores encontraram uma acurácia de 81.8% usando redes neurais multiclasse e 79.5% usando regressão logística multiclasse.

No estudo de Kiselev et al. (2020), os autores enfatizam a importância da utilização de *machine learning* no desenvolvimento de sistemas de orientação profissional, assistidos pelo computador. Os autores predisseram os interesses profissionais baseados no perfil dos indivíduos na rede social russa *VKontakte,* de acordo com seus posts. Em outro estudo, mencionado anteriormente, os autores Bogacheva et al. (2020), utilizaram regressão logística multiclasse, a partir de uma amostra de 145.828 indivíduos, para prever os interesses vocacionais a partir de fatores sociodemográficos.

Já no estudo de Belyanova et al. (2019), os autores usaram redes neurais convolucionais para preverem os tipos do RIASEC necessários para determinados trabalhos, de acordo com o texto da *job description* (descrição do trabalho). Outro estudo com o mesmo objetivo de prever a personalidade vocacional a partir de *job posts* (posts de trabalho), os colegas Silva et al. (2020), utilizaram 217.874 posts/vagas de trabalho de Singapura na tarefa de predição. Cabe ressaltar que os posts de trabalho, conforme os autores relatam, não possuem perfis de tipos de personalidade, possuindo somente o título do trabalho, descrição das habilidades necessárias e a classificação ocupacional padrão de Singapura. Salientamos ainda que os trabalhos de Belyanova et al. (2019) e Silva et al. (2020) se basearam o O*NET, um site americano que possui diversas ocupações classificadas de acordo com os tipos do RIASEC. No presente trabalho, como as categorias dos dados são conhecidas (variáveis sociodemográficas) e as saídas desejadas (interesses profissionais baseados no curso superior), foram utilizadas técnicas de aprendizado supervisionado, através de algoritmos de classificação.

Objetivos

O objetivo geral deste estudo foi testar o poder preditivo das variáveis sociodemográficas sobre os tipos do RIASEC, com utilização de técnicas de *machine learning*. Para alcançar este objetivo geral, temos os seguintes objetivos específicos: Classificar os cursos avaliados

no Enade 2018 e 2019 de acordo com as letras dominantes do RIASEC, formando códigos RIASEC com 2 letras; identificar o algoritmo de *machine learning* com maior acurácia para estabelecer a relação preditiva das variáveis sociodemográficas sobre os tipos do RIASEC.

Hipóteses

Espera-se que (H1) as variáveis sociodemográficas sejam capazes de predizer em alguma medida os códigos RIASEC e (H2); espera-se que o modelo de redes neurais tenha maior acurácia que o modelo de regressão logística multiclasse e CART (Hoff et al., 2019; Zahour et al., 2020).

Método

Foi desenvolvido para este projeto um estudo com 2 etapas, sendo a Etapa 1 com foco na classificação dos dois códigos dominantes dos tipos do RIASEC para os cursos avaliados no Enade 2018 e 2019, a Etapa 2 para identificar o algoritmo de *machine learning* com maior acurácia para estabelecer a relação preditiva das variáveis sociodemográficas sobre os tipos do RIASEC e para verificar quais variáveis sociodemográficas são mais preditoras dos códigos RIASEC.

Etapa 1 – Classificação dos cursos avaliados pelo Enade 2018 e 2019 de acordo com os dois principais códigos do RIASEC

Esta etapa tem o objetivo de classificar os cursos avaliados pelo Enade 2018 e 2019, de acordo com os tipos do RIASEC, conforme primeiro objetivo específico. O Enade 2018 e 2019 avaliou estudantes concluintes de 56 cursos que conferem diploma, conforme mostra a tabela 2. Escolhemos os dois anos do Enade para termos representação de cursos que tenham ao menos uma das letras dos tipos do RIASEC.

Tabela 3

Cursos avaliados no Enade 2018 e 2019

Cursos de Bacharelado	Cursos Tecnológicos
Administração	Tecnologia em agronegócios
Administração pública	Tecnologia em comércio exterior
Agronomia	Tecnologia em design de interiores
Arquitetura e urbanismo	Tecnologia em design de moda
Biomedicina	Tecnologia em design gráfico
Ciências contábeis	Tecnologia em estética e cosmética
Ciências econômicas	Tecnologia em gastronomia
Comunicação social – jornalismo	Tecnologia em gestão ambiental

Comunicação social - publicidade e propaganda	Tecnologia em gestão comercial
Design	Tecnologia em gestão da qualidade
Direito	Tecnologia em gestão de recursos humanos
Educação física (bacharelado)	
Enfermagem	Tecnologia em gestão financeira
Engenharia ambiental	Tecnologia em gestão hospitalar
Engenharia civil	Tecnologia em gestão pública
Engenharia da computação	Tecnologia em logística
Engenharia de alimentos	Tecnologia em marketing
Engenharia de controle e automação	Tecnologia em processos gerenciais
Engenharia de produção	Tecnologia em radiologia
Engenharia elétrica	Tecnologia em segurança no trabalho
Engenharia florestal	
Engenharia mecânica	
Engenharia química	
Farmácia	
Fisioterapia	
Fonoaudiologia	
Medicina	
Medicina veterinária	
Nutrição	
Odontologia	
Psicologia	
Relações internacionais	
Secretariado executivo	
Serviço social	
Teologia	

Turismo
Zootecnia

Nota: Manual do usuário do Enade (2018; 2019)

Participantes

Participaram desta etapa três pesquisadores especialistas da área de Avaliação Psicológica, em Orientação Profissional e de Carreira e com conhecimento da Teoria de Personalidade Vocacional e Ambientes de Trabalho. Trata-se de uma amostra não-aleatória de conveniência.

Fonte de Dados

Planilha com cursos avaliados no Enade 2018 e 2019

Foi criada uma planilha em Excel com uma lista dos 56 cursos que foram avaliados no Enade 2018 e 2019.

Base de Dados O*NET

A base de dados O*NET *(Occupational Information Network)* é uma fonte abrangente de informações sobre ocupações nos Estados Unidos. Ela foi desenvolvida pelo Departamento de Trabalho dos EUA para fornecer dados atualizados e detalhados sobre as habilidades, conhecimentos, e características necessárias para diversas ocupações, além de informações sobre o mercado de trabalho e as condições de trabalho (*U.S. Department of Labor*, 2024). O*NET é amplamente utilizado por profissionais de carreira, empregadores, pesquisadores e educadores para orientar decisões de carreira, desenvolvimento de currículos e pesquisas sobre o mercado de trabalho, inclusive no Brasil, por não existir uma base tão completa para classificação de profissões em códigos RIASEC.

Procedimentos

Os 56 cursos foram classificados de acordo com os dois códigos dominantes dos perfis tipológicos RIASEC pelos quatro juízes e de acordo com a base ONET (fazendo o papel de um quarto juiz). A participação dos pesquisadores convidados foi condicionada ao seu aceite e assinatura de um Termo de Consentimento Livre e Esclarecido TCLE, que foi enviado junto ao e-mail de convite. Após aceitação dos juízes, a planilha foi enviada por e-mail.

Análise de dados e Resultados

Foi realizada uma análise de concordância entre os três juízes e o resultado da consulta na base ONET, avaliando a classificação de cada um dos 56 cursos avaliados pelo Enade 2018 e 2019, para determinar os dois códigos dominantes do RIASEC. Por fim, foram excluídos da base do Enade todos os cursos que não tiveram concordância entre ao menos três juízes (incluindo O*NET), ou seja, aqueles cursos que ao menos três juízes não classificaram com os mesmos tipos do RIASEC. Vale ressaltar que a ordem dos códigos RIASEC influenciou no critério de classificação, de tal forma que um se um juiz classificasse um curso como IR e outro juiz como RI, o curso seria removido da base de dados do Enade. Por fim, restaram na base 23 cursos (concordância maior ou igual a 75%), classificados em seu perfil RIASEC, conforme tabela 4.

Tabela 4

Análise de concordância dos juízes na classificação dos cursos nos tipos RIASEC

Curso	Juiz 1	Juiz 2	Juiz 3	O*NET	Concordância (>= 75%)	Final
ADMINISTRAÇÃO	CE	CE	CE	EC	75%	CE
ADMINISTRAÇÃO PÚBLICA	ES	CE	CR	EC	-	
AGRONOMIA	RI	RI	RC	RI	75%	RI
ARQUITETURA E URBANISMO	AI	AR	AI	AR	-	
BIOMEDICINA	IS	IR	IR	IR	75%	IR
CIÊNCIAS CONTÁBEIS	CI	CE	AE	CI	-	
CIÊNCIAS ECONÔMICAS	IC	IC	CE	IC	75%	IC
COMUNICAÇÃO SOCIAL – JORNALISMO	SE	AE	AE	AS	-	

COMUNICAÇÃO SOCIAL - PUBLICIDADE E PROPAGANDA	AE	AE	EC	AE	75%	AE
DESIGN	A	AE	AE	AR	-	
DIREITO	EI	EC	EI	EC	-	
EDUCAÇÃO FÍSICA (BACHARELADO)	SR	RS	EC	SR	-	
ENFERMAGEM	SI	SI	SI	SI	100%	SI
ENGENHARIA AMBIENTAL	IR	IR	IR	RI	75%	IR
ENGENHARIA CIVIL	RE	RI	RI	RC	-	
ENGENHARIA DA COMPUTAÇÃO	IR	IR	IC	RI	-	
ENGENHARIA DE ALIMENTOS	IR	IR	IR	IR	100%	IR
ENGENHARIA DE CONTROLE E AUTOMAÇÃO	IC	RI	RI	RI	75%	RI
ENGENHARIA DE PRODUÇÃO	IR	EI	RI	RC	-	
ENGENHARIA ELÉTRICA	IR	IR	RI	RI	-	
ENGENHARIA FLORESTAL	IR	IR	RI	RI	-	
ENGENHARIA MECÂNICA	IR	RI	RI	RI	75%	RI
ENGENHARIA QUÍMICA	IR	IR	IR	RI	75%	IR
FARMÁCIA	IC	IC	CR	IS	-	
FISIOTERAPIA	SI	SI	SI	SI	100%	SI
FONOAUDIOLOGIA	SI	SA	IS	SI	-	
MEDICINA	SI	IS	IS	IS	75%	IS
MEDICINA VETERINÁRIA	IR	IS	IS	RS	-	
NUTRIÇÃO	IS	IS	IS	SI	75%	IS
ODONTOLOGIA	IS	IS	IS	SR	75%	IS
PSICOLOGIA	SI	SI	SI	SI	100%	SI
RELAÇÕES INTERNACIONAIS	IE	ES	EC	ES	-	
SECRETARIADO EXECUTIVO	CE	CE	CE	CS	75%	CE
SERVIÇO SOCIAL	SI	SC	SI	SE	-	
TEOLOGIA	SI	SA	SI	SA	-	
TURISMO	SE	SE	ER	ES	-	
ZOOTECNIA	RI	RC	IS	RI	-	
TECNOLOGIA EM AGRONEGÓCIOS	ER	ER	RC	RE	-	
TECNOLOGIA EM COMÉRCIO EXTERIOR	ES	ES	CE	EC	-	
TECNOLOGIA EM DESIGN DE INTERIORES	AR	AE	AE	AE	75%	AE
TECNOLOGIA EM DESIGN DE MODA	AR	AC	AE	AE	-	
TECNOLOGIA EM DESIGN GRÁFICO	AR	AR	AR	AE	75%	AR
TECNOLOGIA EM ESTÉTICA E COSMÉTICA	AE	SE	AE	SA	-	
TECNOLOGIA EM GASTRONOMIA	ER	AE	ER	RA	-	
TECNOLOGIA EM GESTÃO AMBIENTAL	EI	ES	IR	RI	-	
TECNOLOGIA EM GESTÃO COMERCIAL	EC	ES	CE	EC	-	
TECNOLOGIA EM GESTÃO DA QUALIDADE	EI	EC	RI	CR	-	
TECNOLOGIA EM GESTÃO DE RECURSOS HUMANOS	ES	ES	EC	EC	-	
TECNOLOGIA EM GESTÃO FINANCEIRA	EC	EC	CE	CE	-	
TECNOLOGIA EM GESTÃO HOSPITALAR	EC	ES	CS	ES	-	
TECNOLOGIA EM GESTÃO PÚBLICA	EC	EC	CR	EC	75%	EC
TECNOLOGIA EM LOGÍSTICA	CE	RE	CE	CE	75%	CE
TECNOLOGIA EM MARKETING	EA	EA	EC	EA	75%	EA
TECNOLOGIA EM PROCESSOS GERENCIAIS	CE	EC	CE	CE	75%	CE
TECNOLOGIA EM RADIOLOGIA	IR	RS	CR	IR	-	

TECNOLOGIA EM SEGURANÇA NO TRABALHO RS EC CR RC -

Etapa 2 – Identificação do algoritmo com maior acurácia e verificação das variáveis sociodemográficas mais preditoras

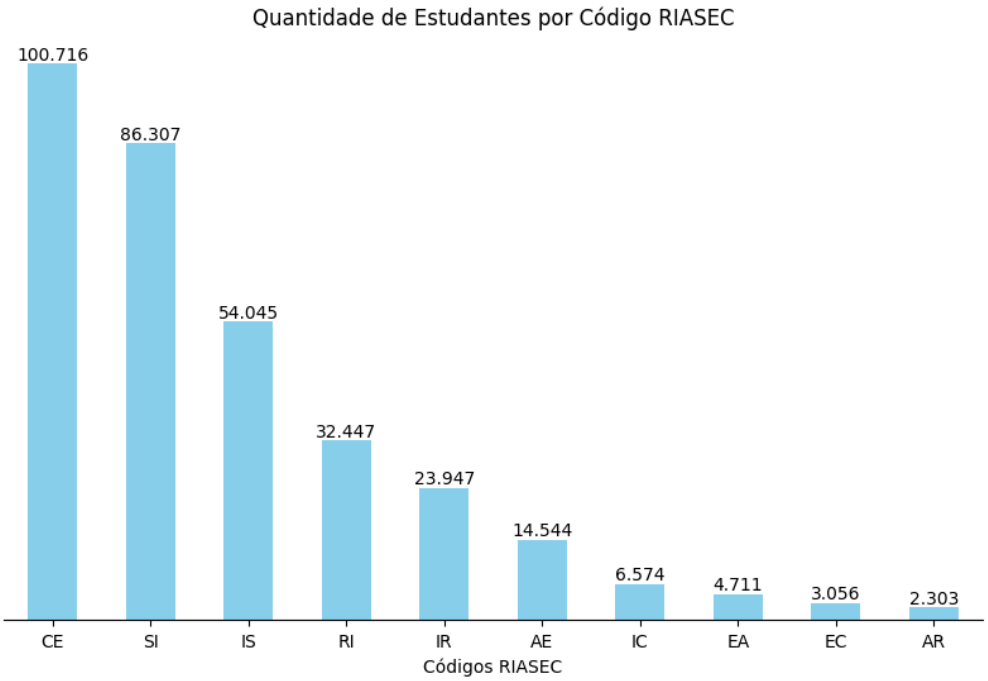
Esta etapa teve o objetivo de identificar o algoritmo de *machine learning* com maior acurácia para estabelecer a relação preditiva das variáveis sociodemográficas sobre os tipos do RIASEC e para verificar quais variáveis sociodemográficas são mais preditoras de cada um dos tipos do RIASEC.

Fonte de dados

A amostra inicial desta etapa foi composta pelos estudantes que realizaram o Enade 2018 e 2019, que são brasileiros, com presença válida na prova do Enade e com respostas em todas as provas (formação geral e conhecimento específico), totalizando 740.381 estudantes, sendo 394.541 do Enade 2018 e 345.840 do Enade 2019. Após limpeza nos dados e filtragem dos 23 cursos que tiveram concordância entre ao menos 3 juízes referente a etapa 1, a amostra final foi composta por 328.650 estudantes, distribuídos entre os códigos RIASEC conforme figura 6.

Figura 6

Distribuição da amostra por código RIASEC



Não haverá a necessidade de aprovação em comitê de ética CEP/CONEP, conforme Resolução CNS 510/2016, por se adequar como pesquisa que utiliza informações de acesso público, nos termos da Lei no 12.527, de 18 de novembro de 2011.

Instrumentos

Nesta pesquisa foram utilizadas as informações referentes ao Enade 2018 e Enade 2019. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) disponibiliza em seu portal os arquivos de microdados do Enade, que contêm, dentre outras, informações sobre os participantes e os instrumentos. São disponibilizados, por exemplo, os dados das inscrições dos participantes e suas respostas aos testes de desempenho e ao questionário do estudante (esse conjunto denominaremos de “Questionário dos Participantes”).

Questionário dos Participantes

O arquivo com os dados dos participantes contém variáveis que estão agrupadas da seguinte maneira: (a) dados da instituição de ensino superior e do curso (por exemplo, código da área de enquadramento do curso no Enade, código da modalidade do ensino), (b) dados do estudante (por exemplo, idade, sexo, ano de conclusão do ensino médio, ano de início da graduação), (c) dados da formação geral e componente específico (por exemplo, as respostas aos itens dos testes de desempenho, se compareceu no dia da prova ou não), (d) dados do questionário de percepção da prova (por exemplo, grau de dificuldade da prova na formação geral), (e) dados do questionário do estudante (por exemplo, estado civil, cor ou raça, motivo para a escolha do curso) e (f) dados do grau de concordância do aluno (por exemplo, contribuição das disciplinas para a formação integral, como cidadão e profissional). Nesta pesquisa, foram utilizadas variáveis dos dados da instituição superior e do curso, dados do estudante e dados do questionário do estudante.

A amostragem foi composta por variáveis de interesse desse estudo, conforme dicionário dos microdados do Enade 2018 e 2019. Foram consideradas (a) informações da Instituição de Ensino Superior e Curso, (b) informações sobre ano de conclusão do ensino médio e ano de início da graduação, (c) turno da graduação e (d) informações sobre respostas ao questionário sociodemográfico do Enade 2018 e 2019. Este questionário contém questões de múltipla escolha, onde o estudante deve responder apenas uma opção, além de informações adicionais retiradas do Questionário do Participante. As variáveis que foram utilizadas no modelo constam na Tabela 5.

Tabela 5

Variáveis do questionário sociodemográfico e do participante

Variáveis
Idade
Alguém da família com curso superior?
Bolsa acadêmica
Bolsa de estudos/financiamento do curso
Bolsa de Permanência
Escolaridade Mãe
Escolaridade Pai
Estado Civil
Grupo determinante que ajudou durante o curso
Motivo escolha curso
Onde e com quem mora
Pessoas na residência
Políticas de ação afirmativa ou inclusão social
Quem incentivou graduação
Raça
Renda Familiar
Sexo
Situação de Trabalho
Situação Financeira
Tipo de Escola Ensino médio
Idade

Alguém da família com curso superior?
Bolsa acadêmica
Bolsa de estudos/financiamento do curso
Bolsa de Permanência
Escolaridade Mãe

Nota: Enade (2018; 2019)

Procedimentos

Todas as informações foram obtidas a partir dos microdados do Enade, disponíveis no portal do Inep em um arquivo compactado. A obtenção dos arquivos dos microdados foi feita em maio de 2021 por meio do link https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados. É importante destacar que os dados dos participantes não estão identificados, pois seu número de inscrição não está disponível nos microdados. Além disso, as informações identificáveis (como CPF e nome) estão ausentes. Dentre os arquivos que compõem os microdados está o banco de dados dos respondentes, com variáveis como Tipo de presença no Enade (TP_PRES), Tipo de presença na prova (TP_PR_GER), Tipo de presença na parte objetiva na formação geral (TP_PR_OB_FG), Tipo de presença na parte discursiva na formação geral (TP_PR_DI_FG), Tipo de presença na parte objetiva no componente específico (TP_PR_OB_CE), Tipo de presença na parte discursiva no componente específico (TP_PR_DI_CE) e nacionalidade (QE_I03).

Para a composição da amostra, foram selecionados os estudantes com presença válida no Enade (TP_PRES = 555), com resultados considerados válidos em todas as provas (TP_PRES = 555; TP_PR_GER = 555; TP_PR_OB_FG = 555; TP_PR_DI_FG = 555; TP_PR_OB_CE = 555; TP_PR_DI_CE = 555) e que sejam brasileiros (QE_I03 = A).

Análise de dados

Os microdados do Enade 2018 e 2019 foram importados no software R Studio (R Core Team, 2021), para que seja realizada, dentre outras coisas, a filtragem dos objetos da pesquisa e exclusão dos dados que não foram utilizados.

Para a tarefa de predição, baseado nas respostas ao questionário sociodemográfico e classificação, foram aplicadas técnicas de Regressão logística multiclasse, CART e Redes Neurais. Classificação Regressão logística multiclasse requer um *dataset* com categorias identificadas e possui um modelo de treino, que pode ser usado para predizer os valores de múltiplas saídas a parir de novos *inputs* de dados (Zahour et al., 2020); para modelo CART é utilizado tanto para classificação quanto para regressão, realizando particionamento dos dados e fazendo previsões simples em cada partição, representadas por uma árvore de decisão (Loh, 2011). Já a técnica de redes neurais, trabalha com um conjunto de camadas interconectadas, onde as entradas são a primeira camada e estão conectados a uma camada de saída por um gráfico acíclico composto de pesos e nós (Zahour et al., 2020).

As análises seguintes foram realizadas em ambiente Python (Van Rossum & Drake, 2009). Foram criadas novas variáveis com todas as alternativas das 20 variáveis sociodemográficas, com valores entre 0 (caso o item não tenha sido escolhido) e 1 (caso o item tenha sido escolhido), utilizando a técnica de *dummy* do pacote *pandas*. Por exemplo, para a variável sexo, teremos duas variáveis, uma para o sexo masculino (1 quando for sexo masculino e 0 quando for sexo feminino) e outra para o sexo feminino (1 quando for sexo feminino e 0 quando for sexo masculino). Após isso, foram elaborados e executados os modelos, conforme a seguinte ordem: Regressão logística, utilizando o pacote *scikit-learn* (Pedregosa et al., 2011), para o modelo CART utilizaremos também o *scikit-learn* e para Redes neurais utilizaremos o pacote *tensorflow* (Abadi et al., 2016).

Para a execução dos modelos, a base foi segmentada em 70% para treinamento e 30% para teste, com a utilização do *train_test_split* do *scikit-learn*, que também foi usado para a técnica de *confusion matrix* (matriz de confusão). A técnica de matriz de confusão é utilizada quando o objetivo é avaliar o desempenho de um modelo, verificando o número de classificações corretas versus as classificações preditas para cada classe para verificar acertos e erros. Adicionalmente, usando o *scikit-learn* e *matplotlib*, foi feita a análise da Curva ROC (*Receiver Operating Characteristic*), que é um gráfico que ilustra a capacidade de um sistema classificador binário ao variar seu limiar de discriminação, plotando a taxa de verdadeiros positivos (sensibilidade) contra a taxa de falsos positivos (1-especificidade). A área sob a curva ROC (AUC) fornece um valor único para resumir o desempenho do classificador, onde AUC de 1 indica classificação perfeita e AUC de 0,5 sugere nenhuma capacidade discriminativa (Fawcett, 2006).

Para comparação entre modelos, foram utilizados os índices de precisão, sensibilidade e *F1-score*. Precisão é uma medida de proporção de verdadeiros positivos em relação ao total de positivos preditos pelo modelo; a sensibilidade mede a capacidade do modelo identificar corretamente todas as instâncias positivas; já o *F1-score* é a média harmônica entre precisão e sensibilidade. Para cada modelo, foi avaliada também a especificidade, que é a medida da eficácia de um teste ou modelo de classificação em identificar corretamente os casos negativos. Em termos simples, ela representa a proporção de verdadeiros negativos entre todos os indivíduos que não possuem a condição testada, evitando falsos positivos. A utilização destas técnicas contribuiu para a escolha do modelo com maior desempenho sobre diversas óticas.

Por último, para a análise de importância das variáveis, foi utilizado o modelo de *Gradient Boosting*, através do pacote *scikit-learn.* Apesar de terem sido aplicados modelos de Redes Neurais, CART e Regressão Logística Multiclasse para a predição dos interesses

profissionais, e também termos considerado o método de *Permutation Importance*, o uso do *Gradient Boosting* foi devido à sua capacidade de fornecer interpretações mais claras e diretas das contribuições de cada variável. O Permutation Importance, apesar de ser um método robusto, apresenta uma interpretação mais complexa e um custo computacional mais elevado. Dessa forma, o *Gradient Boosting* se mostrou mais eficiente e adequado para as análises.

Resultados

A fim de verificar se as variáveis sociodemográficas são capazes de predizer em alguma medida o perfil de interesses dos cursos (H1), foram realizadas as análises preditivas tendo como Variáveis Independentes as características sociodemográficas dos respondentes do Enade e como Variável Dependente, o código RIASEC composto por duas letras. O primeiro modelo apresentado será a Regressão Logística Multiclasse, o segundo será o CART e o terceiro será a Rede Neural. Para cada modelo serão apresentadas: 1) tabela contendo a precisão do modelo (a proporção de verdadeiros positivos sobre positivos preditos), a sensibilidade do modelo (a quantidade de positivos verdadeiros sobre os positivos reais), o *F1-score* (a média harmônica da precisão e da sensibilidade), a especificidade (proporção de verdadeiros negativos entre todos os indivíduos que não possuem a condição testada, evitando falsos positivos), o suporte (quantidade de exemplos dos códigos RIASEC específicos), a acurácia (a proporção de previsões corretas sobre o total de previsões), a média macro (a média das métricas - precisão, sensibilidade, *F1-score* - considerando cada código igualmente) e a média ponderada (a média das métricas ponderada pelo suporte de cada código RIASEC); 2) a Curva ROC; e 3) a matriz de confusão para cada modelo.

Os resultados da análise de regressão logística multiclasse, realizada com 1000 iterações, demonstram um desempenho variado entre os diferentes códigos do RIASEC. A regressão logística multiclasse apresentou acurácia geral de 0,54. Isso significa que o modelo está correto em 54% das predições. Os códigos "SI" E “CE” destacam-se com os melhores desempenhos, apresentando *F1-scores* de 0,58 e 0,68, respectivamente. Esses códigos possuem alta sensibilidade, 0,65 para "SI" e 0,79 para "CE", indicando que o modelo é eficaz em identificar corretamente esses códigos. Por outro lado, códigos como "EA" e "EC" apresentaram desempenho insatisfatório, com *F1-scores* de 0,00 e 0,00, respectivamente, refletindo baixa precisão e sensibilidade. Isso sugere que o modelo tem dificuldade em prever

corretamente esses códigos. A Tabela 6 apresenta os índices de precisão, sensibilidade, *F1-score* e suporte para todos os códigos, a figura 7 a Curva ROC e a figura 8 apresenta a matriz de confusão do modelo.

Tabela 6

Índice de precisão, sensibilidade, F1-score, especificidade e suporte para o modelo de regressão logística multiclasse

Código RIASEC	Precisão	Sensibilidade	*F1-score*	*Especificidade*	Suporte
AE	0,37	0,11	0,16	0,99	4.389
AR	0,23	0,00	0,01	1,00	661
CE	0,59	0,79	0,68	0,76	30.232
EA	0,00	0,00	0,00	1,00	1.406
EC	0,29	0,00	0,00	1,00	875
IC	0,25	0,01	0,02	1,00	1.976
IR	0,35	0,08	0,12	0,99	7.175
IS	0,52	0,49	0,51	0,91	16.200
RI	0,44	0,35	0,39	0,95	9.901
SI	0,53	0,65	0,58	0,79	25.780
Acurácia			0,54		
Média Macro	0,36	0,25	0,25	0,94	98.595
Média Ponderada	0,5	0,54	0,5	0,85	98.595

Figura 7

Curva ROC modelo Regressão Logística Multiclasse

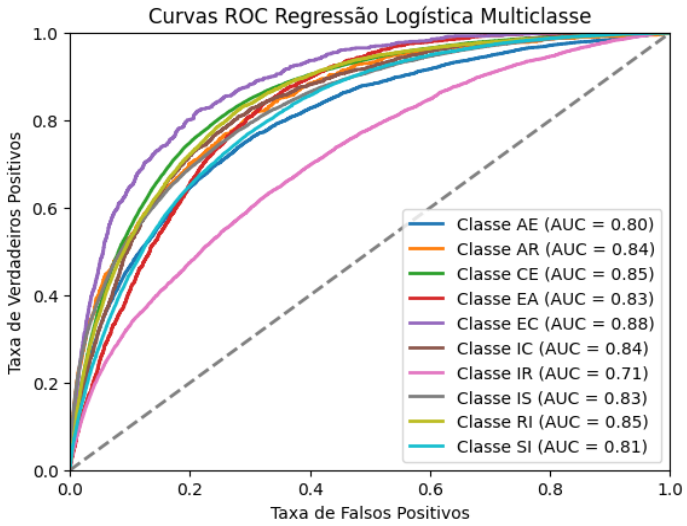
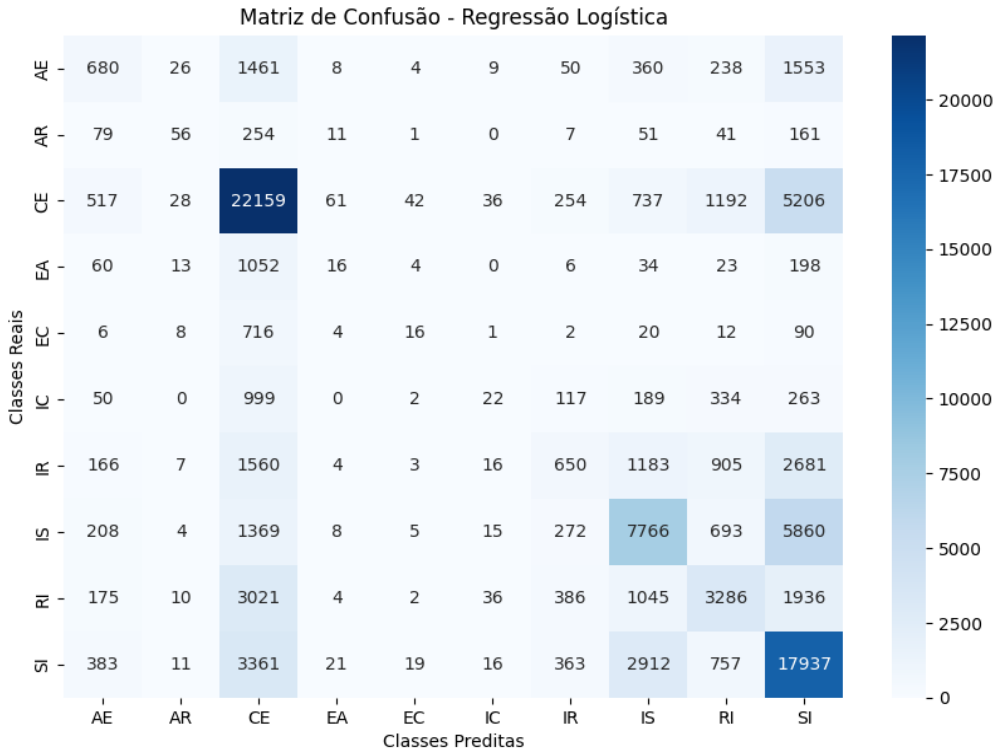


Figura 8

Matriz de confusão para o modelo de regressão logística multiclasse



Os resultados da análise do modelo CART (*Classification and Regression Tree*) demonstraram um desempenho variado entre os diferentes códigos do modelo RIASEC. A análise aponta que a acurácia geral do modelo foi de 0,40, indicando que o modelo está correto em 40% das predições. Os códigos "SI" e "CE" novamente destacam-se com os melhores desempenhos, apresentando *F1-scores* de 0,45 e 0,55, respectivamente. Esses resultados mostraram que o modelo CART é razoavelmente eficaz em identificar corretamente esses códigos. Por outro lado, códigos como "EA", "EC" e "IC" apresentaram desempenho insatisfatório, com *F1-scores* de 0,04, 0,06 e 0,06, respectivamente. Isso sugere que o modelo CART tem dificuldade em prever corretamente esses códigos, refletindo baixa precisão e sensibilidade. A Tabela 7 apresenta os índices de precisão, sensibilidade, *F1-score* e suporte para todos os códigos, a figura 9 a Curva ROC e a figura 10 apresenta a matriz de confusão do modelo.

Tabela 7

Índice de precisão, sensibilidade, F1-score, Sensibilidade e suporte para o modelo CART

Código RIASEC	Precisão	Sensibilidade	*F1-score*	*Especificidade*	Suporte
AE	0,14	0,13	0,13	0,96	4.389
AR	0,09	0,09	0,09	0,99	661
CE	0,55	0,55	0,55	0,80	30.232
EA	0,04	0,04	0,04	0,99	1.406
EC	0,06	0,07	0,06	0,99	875
IC	0,06	0,06	0,06	0,98	1.976
IR	0,14	0,14	0,14	0,93	7.175
IS	0,39	0,40	0,39	0,88	16.200
RI	0,28	0,28	0,28	0,92	9.901
SI	0,45	0,45	0,45	0,81	25.780
Acurácia			0,40		
Média Macro	0,22	0,22	0,22	0,93	98.595
Média Ponderada	0,4	0,4	0,4	0,85	98.595

Figura 9

Curva ROC modelo CART

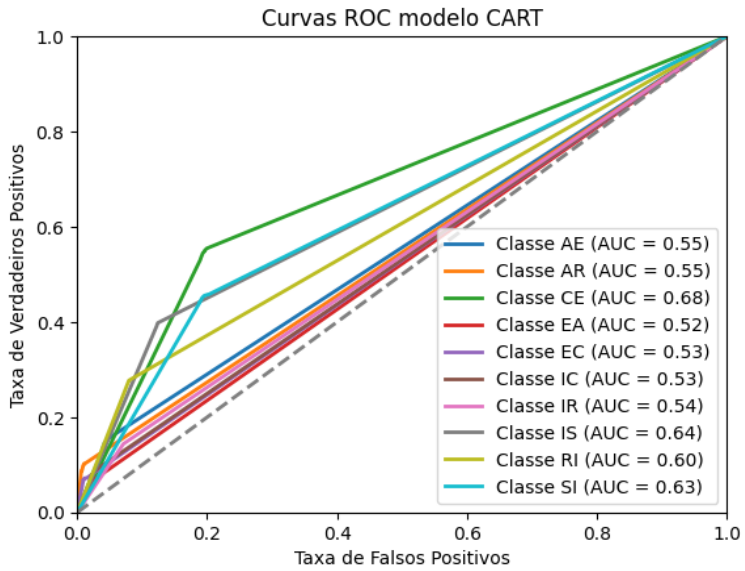
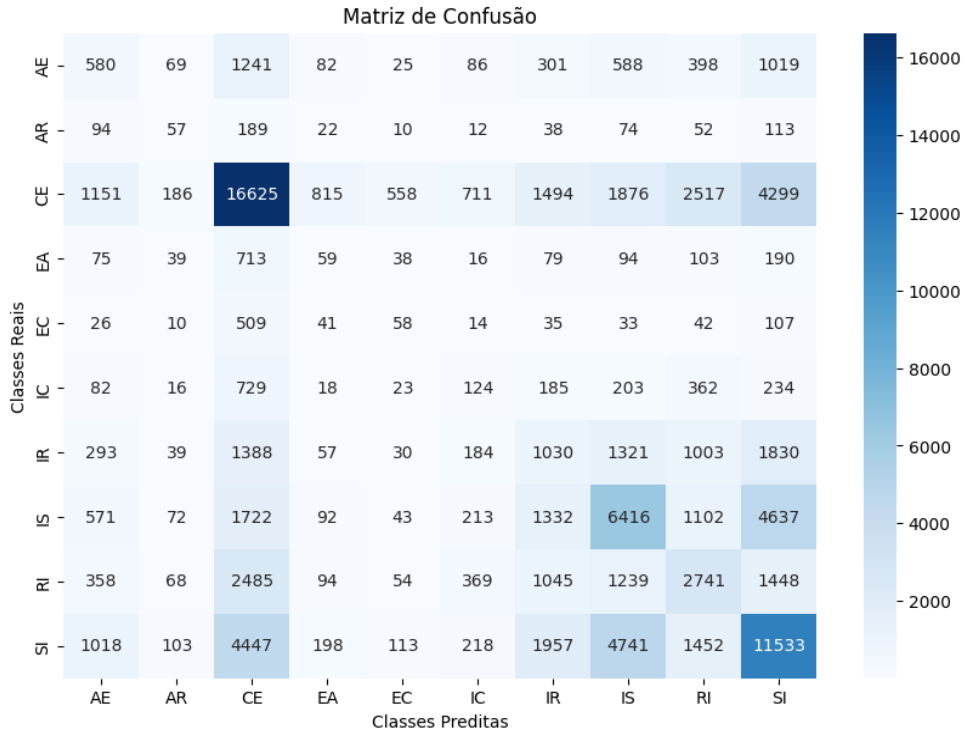


Figura 10

Matriz de confusão para o modelo CART



Os resultados da análise do modelo de Redes Neurais demonstram um desempenho variado entre os diferentes códigos do modelo RIASEC. Vale ressaltar que a eficácia de um modelo de Redes Neurais depende da quantidade de camadas e nódulos configurados. Para este estudo foi utilizado o padrão da biblioteca *tensorflow*, com duas camadas (a primeira com 128 nódulos e a segunda com 64 nódulos). Os resultados apontam para acurácia geral do modelo de 0,54. Os códigos "SI" e "CE" destacam-se com os melhores desempenhos, apresentando *F1-scores* de 0,58 e 0,67, respectivamente. Esses resultados mostram que o modelo de Redes Neurais é eficaz em identificar corretamente essas categorias, refletindo alta sensibilidade de 0,64 para "SI" e 0,79 para "CE". Por outro lado, códigos como "EA" e "IC" apresentam desempenho insatisfatório, com *F1-scores* de 0,03 e 0,03, respectivamente. Isso sugere que o modelo de Redes Neurais tem dificuldade em prever corretamente esses códigos, refletindo baixa precisão e sensibilidade. A Tabela 8 apresenta os índices de

precisão, sensibilidade, *F1-score*, especificidade e suporte para todos os códigos, a figura 11 a curva ROC e a figura 12 apresenta a matriz de confusão do modelo.

Tabela 8

Índice de precisão, sensibilidade, F1-score e suporte para o modelo de Redes Neurais

Código RIASEC	Precisão	Sensibilidade	*F1-score*	*Especificidade*	Suporte
AE	0,32	0,13	0,18	0,99	4.389
AR	0,34	0,10	0,15	1,00	661
CE	0,59	0,79	0,67	0,75	30.232
EA	0,26	0,01	0,03	1,00	1.406
EC	0,05	0,00	0,00	1,00	875
IC	0,16	0,01	0,03	1,00	1.976
IR	0,34	0,05	0,09	0,99	7.175
IS	0,56	0,45	0,50	0,93	16.200
RI	0,40	0,42	0,41	0,93	9.901
SI	0,53	0,64	0,58	0,80	25.780
Acurácia			0,54		
Média Macro	0,35	0,26	0,26	0,94	98595
Média Ponderada	0,50	0,54	0,50	0,85	98595

Figura 11

Curva ROC modelo de Redes Neurais

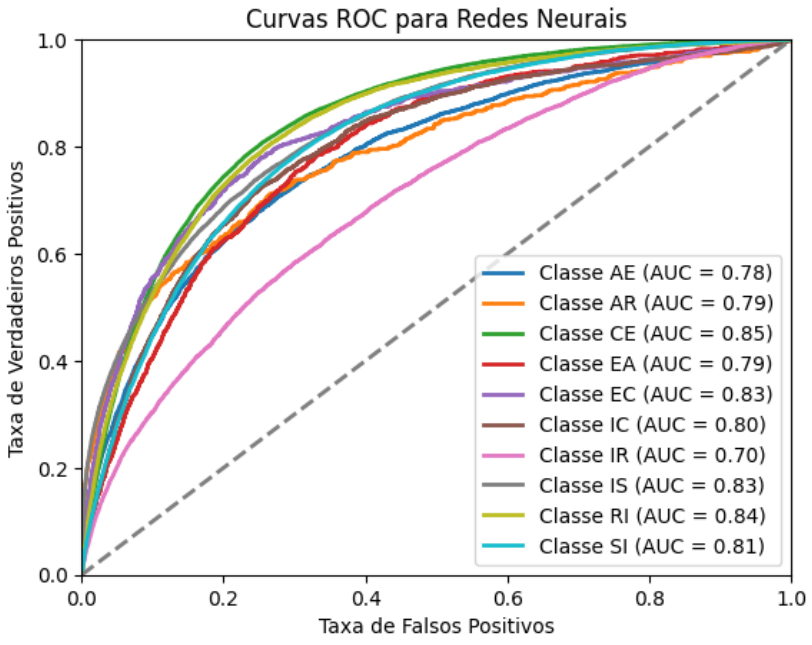
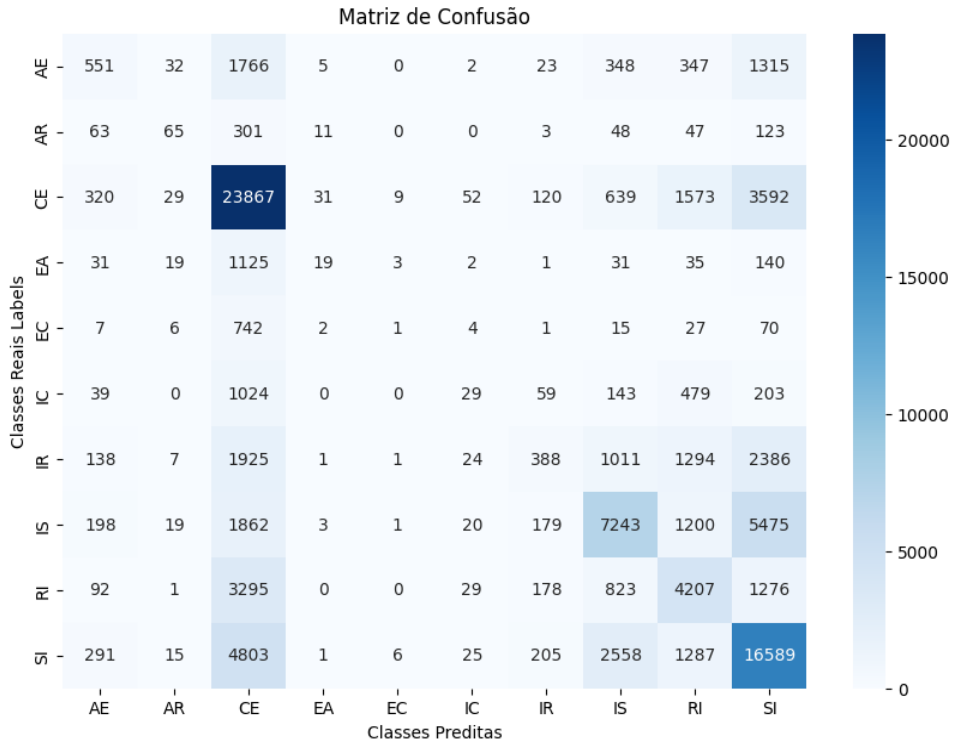


Figura 12

Matriz de confusão para o modelo de Redes Neurais



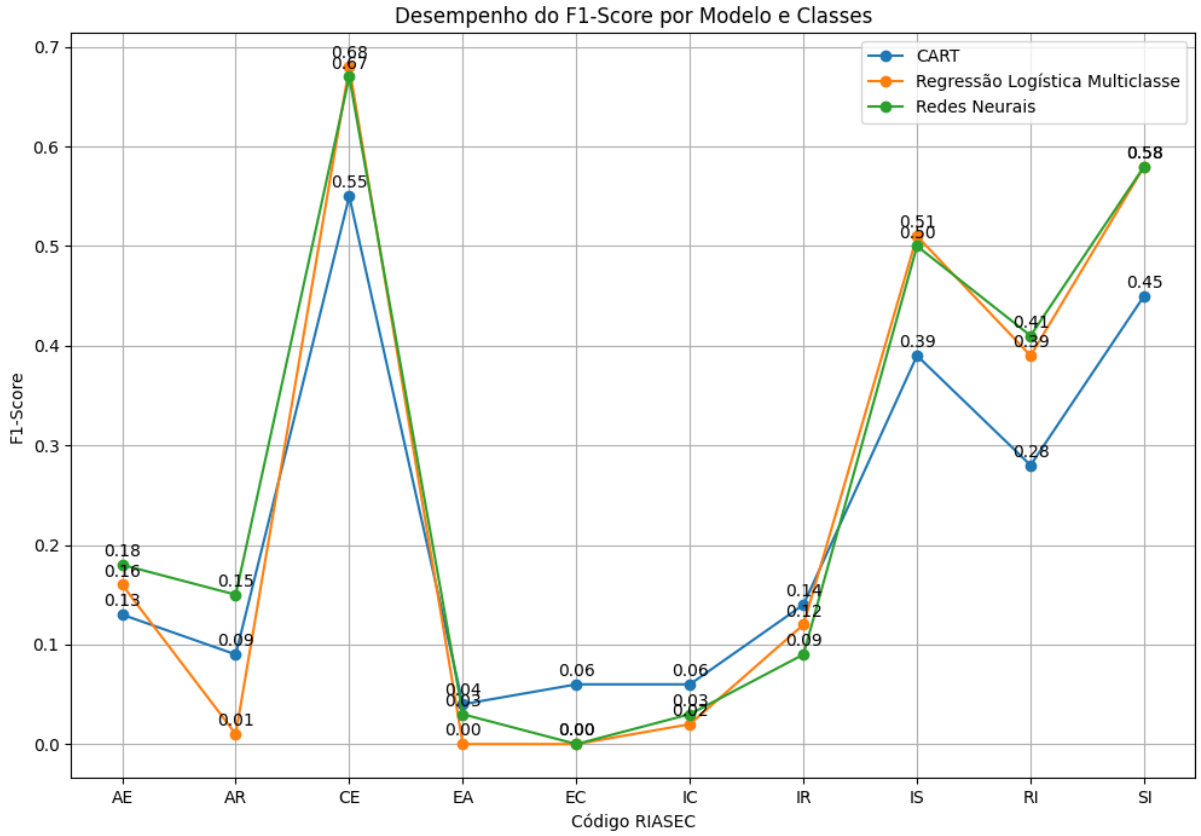
Buscando compreender se o modelo de redes neurais tenha maior acurácia que o modelo de regressão logística multiclasse e CART (H2), foram realizadas comparações entre os índices encontrados em cada um dos modelos. Para comparação geral entre modelos, foi utilizada a Acurácia, e para a comparação entre códigos específicos foi utilizado o *F1-score*, pois ele pondera a precisão do modelo com base na quantidade de códigos disponíveis na base de dados.

O modelo de Redes Neurais apresentou a mesma acurácia que o modelo de regressão logística, 0,54. Ambos tiveram acurácia consideravelmente maior do que o modelo CART (0,40). Além disso, o modelo de Redes Neurais teve desempenho igual ao Regressão logísticas (ambos foram melhores em 5 classes que o outro), mas comparando com o modelo CART, o modelo de redes neurais foi melhor em 6 de 10, empatando em duas classe, EC (0,00) e SI (0,58). O modelo de regressão logística multiclasse foi melhor nas classes CE

(0,68), SI (0,58), IS (0,51), sendo que o de redes neurais melhor para as classes AE (0,18), AR (0,15), EA (0,03), IC (0,03) e RI (0,41). O CART apresentou resultados superiores para AR (0,09), EA (0,04), EC (0,06), IC (0,06) e IR (0,14), mas com índices muito baixos (0,04 a 0,14, respectivamente), indicando que nenhum dos modelos performam bem para classificação desses códigos RIASEC. A figura 13 representa o desempenho de cada modelo, considerando o F1-score de cada código RIASEC.

Figura 13

Desempenho considerando F1-score de cada código RIASEC para os modelos regressão logística multiclasse, CART e redes neurais

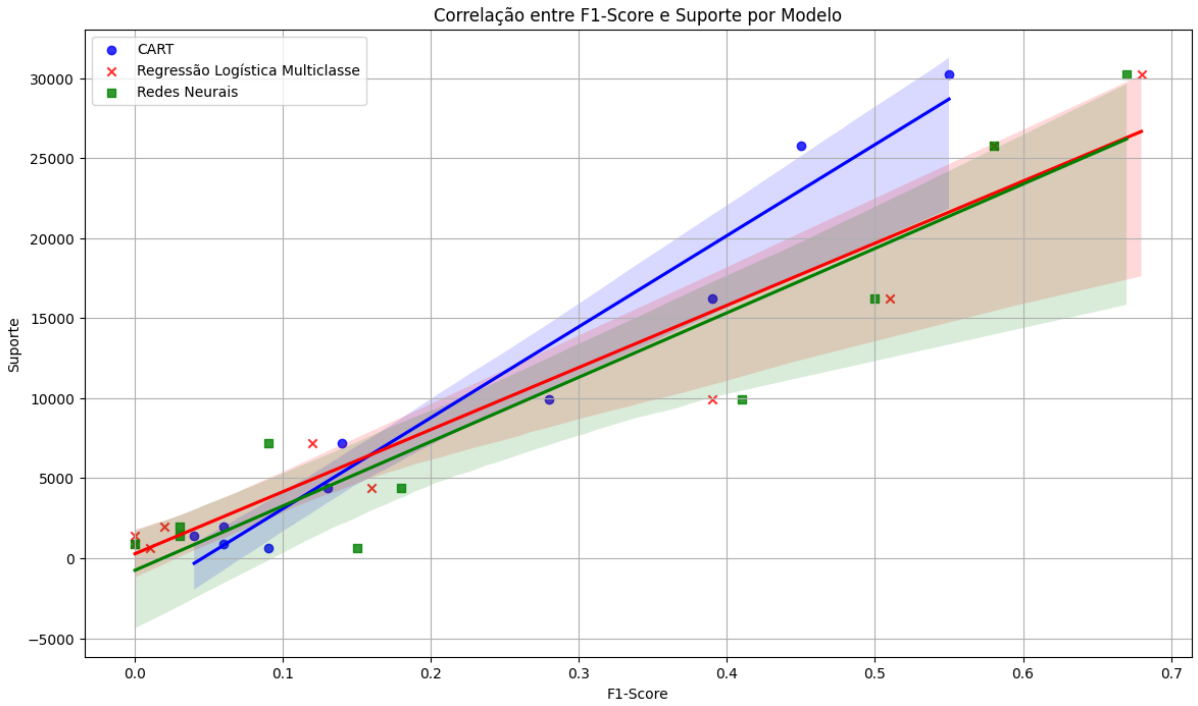


Além disso, foi calculada a correlação entre suporte e o *F1-score* de cada modelo, a fim de investigar a relação entre a eficácia dos modelos e a quantidade de códigos disponíveis na base de dados. O método utilizado foi a correlação de Pearson. Os resultados sugerem que o desempenho dos modelos é melhor em classes que são mais representadas no banco de

dados, sendo Regressão Logística (r = 0,70, p = 0,021), CART (r = 0,73, p = 0,015) e Redes Neurais (r = 0,72, p = 0,017).

Figura 14

Correlação entre F1-score e suporte

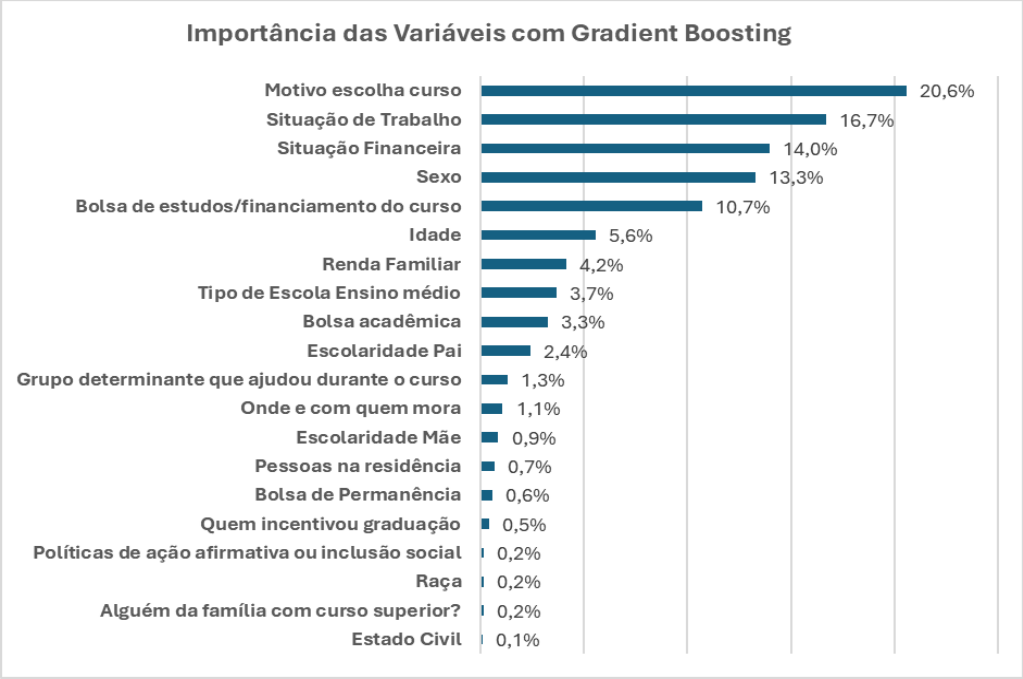


Por fim, foi feita a análise de importância das variáveis com o modelo *Gradient Boosting*, conforme figura 15, por ser um modelo que consegue com insights significativos sobre os fatores que influenciam os interesses profissionais dos estudantes. A acurácia geral do modelo *Gradient Boosting* também foi de 0,54%. O motivo da escolha do curso emergiu como o fator mais influente, contribuindo com 20,6% da importância total do modelo. Em seguida, a situação de trabalho foi identificada como um determinante crucial, representando 16,7% da importância, seguida pela situação financeira dos estudantes, que contribuiu com 14,0%. O sexo dos respondentes também se destacou, com uma contribuição de 13,3%, indicando diferenças notáveis de interesses entre os gêneros. Além disso, o apoio financeiro, na forma de bolsas de estudos ou financiamentos, foi responsável por 10,7% da variabilidade explicada pelo modelo.

Outros fatores importantes incluíram a idade (5,6%), a renda familiar (4,2%), o tipo de escola de ensino médio frequentada (3,7%), e a obtenção de bolsas acadêmicas (3,3%). A escolaridade dos pais também mostrou relevância, com a escolaridade do pai contribuindo com 2,4% e a da mãe com 0,9%. Esses achados sublinham a complexidade dos determinantes dos interesses profissionais, sugerindo que uma combinação de fatores acadêmicos, financeiros e pessoais desempenha um papel crucial na definição das trajetórias de carreira dos estudantes. Adicionalmente, as figuras 16, 17 e 18 representam a análise de importância dos 3 códigos RIASEC com melhor desempenho nos modelos baseado nas variáveis originais (sexo, por exemplo), CE, IS e SI, respectivamente. Foram geradas também análises de acordo com as variáveis *dummies* (sexo_m ou sexo_f), o que normalmente gera resultados distintos por conta da: distribuição da importância (A importância das variáveis originais pode ser redistribuída nas variáveis *dummy*); interação entre as variáveis (*dummies* capturam interações diferentes); Regularização (Regularização afeta *dummies* e variáveis originais de formas distintas); e Multicolinearidade (*dummies* podem introduzir multicolinearidade, alterando importâncias).

Figura 15

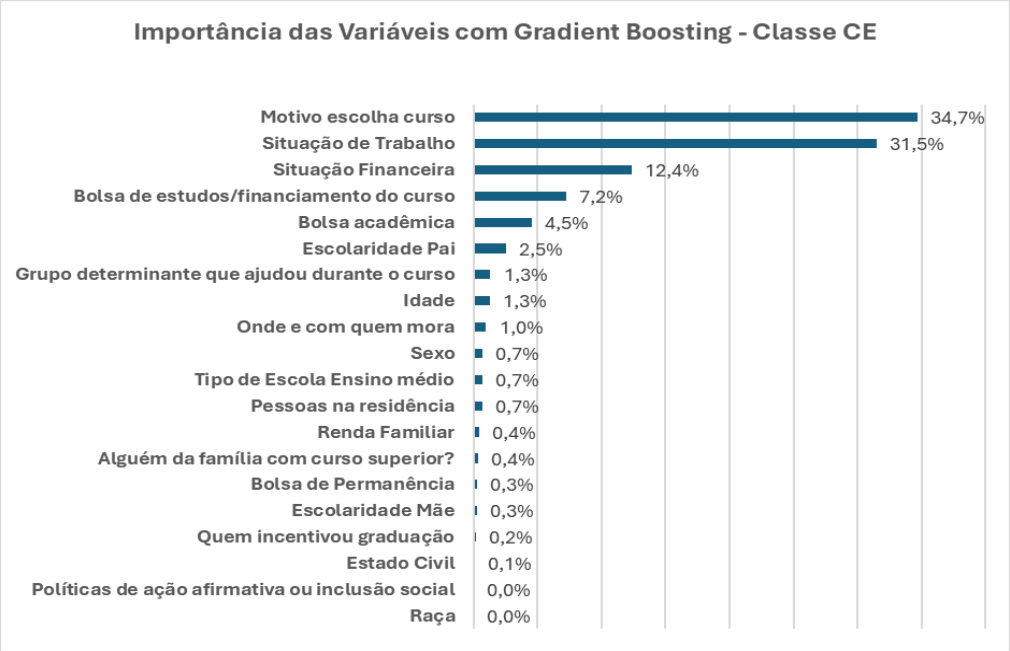
Importância das variáveis sociodemográficas na predição das classes RIASEC



Para o código CE, as três principais variáveis que representam 78,5% da explicação da variabilidade do modelo foram, motivo escolha curso com 34,7%, situação de trabalho com 31,5% e Situação financeira com 12,4%. A classe CE é composta pelos cursos: Administração, Secretariado Executivo, Tecnologia em Logística e Tecnologia em Processos Gerenciais. Para as variáveis *dummies*, 66,8% foram explicados por 3 variáveis, sendo 34,4% situação de trabalho, alternativa E - Trabalho 40 horas semanais ou mais, 22,6% pelo motivo escolha curso, alternativa E – Vocação e 9,8% também pelo motivo escolha curso, mas alternativa H – Outro motivo.

Figura 16

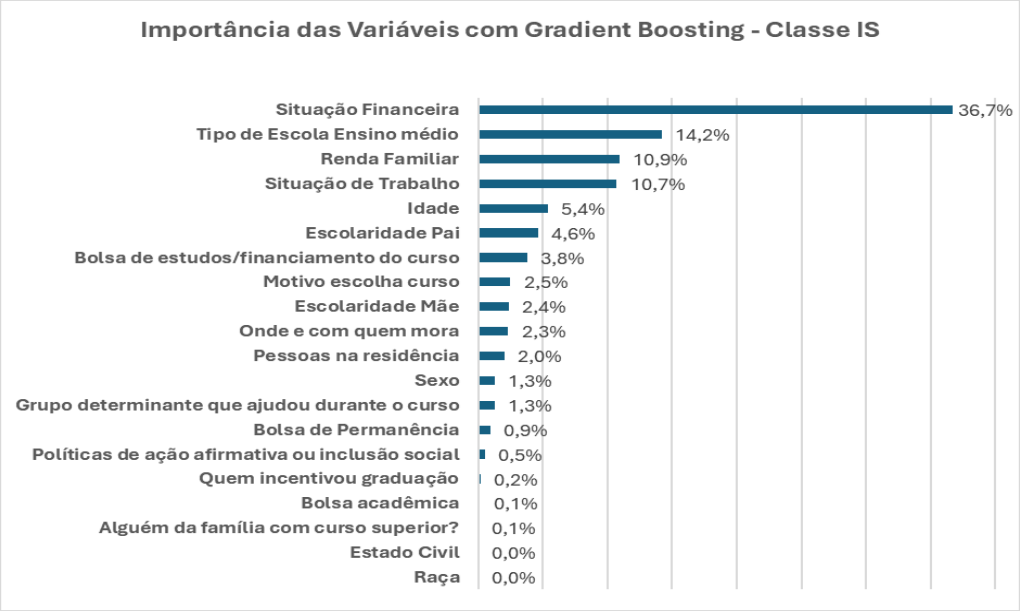
Importância das variáveis sociodemográficas na predição da classe CE



Para o código IS, as quatros principais variáveis que representam 72,6% da explicação da variabilidade modelo foram, situação financeira com 36,7%, tipo de escola ensino médio com 14,2%, renda familiar com 10,9% e situação de trabalho com 10,7%. A classe IS é composta pelos cursos: Medicina, Nutrição e Odontologia. Para as variáveis *dummies*, 65,7% foram explicados por 3 variáveis, sendo 43,3% situação financeira, alternativa B - Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas, 15,9% pelo tipo de escola ensino médio, também alternativa B - Todo em escola privada (particular) e 6,5% por bolsa de estudos/financiamento do curso, alternativa E - FIES, apenas.

Figura 17

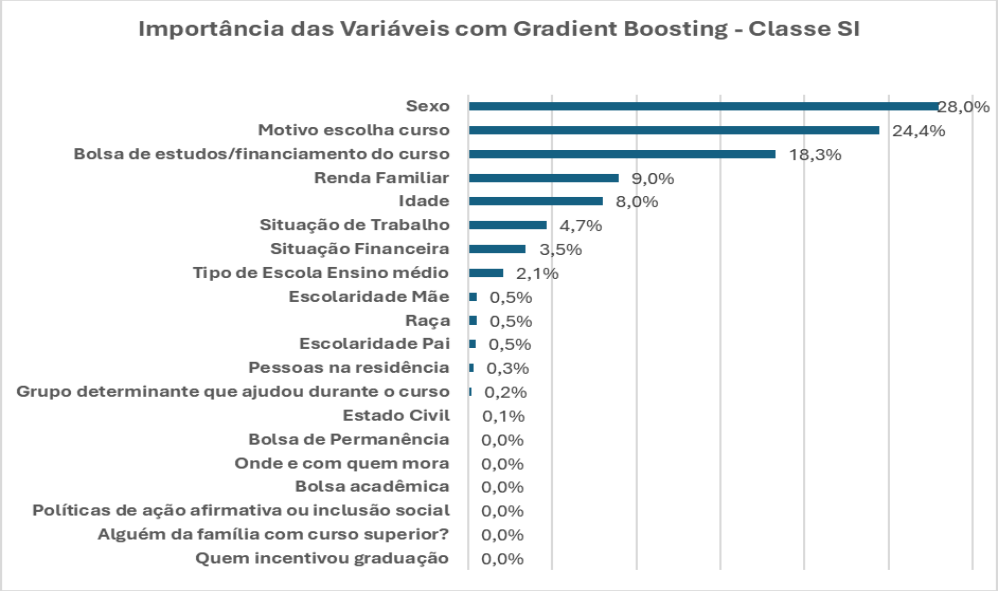
Importância das variáveis sociodemográficas na predição da classe IS



Para o código SI, as três principais variáveis que representam 70,7% da explicação da variabilidade do modelo foram, sexo com 28%, motivo escolha curso com 24,4% e bolsa de estudos/financiamento do curso com 18,3%. A classe IS é composta pelos cursos: Enfermagem, Fisioterapia e Psicologia. Para as variáveis *dummies*, 61,1% da variabilidade do modelo foi explicada por 3 variáveis, sendo 29,3% sexo masculino, 17,1% por bolsa de estudos/financiamento do curso, alternativa E - FIES, apenas e 14,7% pelo motivo escolha curso, alternativa E – Vocação. A diferença entre a variável original e as variáveis *dummies* pode ocorrer porque as variáveis *dummies* capturam detalhes específicos que as variáveis originais não conseguem isolar. No caso do sexo, por exemplo, a presença masculina é tão notável nos cursos predominantemente femininos acima, que se torna um fator preditivo forte.

Figura 18

Importância das variáveis sociodemográficas na predição da classe SI



Discussão

O primeiro objetivo e objetivo geral deste estudo foi testar o poder preditivo das variáveis sociodemográficas sobre a classificação dos cursos por códigos dos tipos do RIASEC, com utilização de técnicas de *machine learning*. É possível dizer que esse objetivo foi atingido, mas é preciso ponderar sobre os resultados e considerar as limitações do estudo.

De maneira geral, os códigos RIASEC dos cursos – ou seja, os cursos classificados a partir do modelo RIASEC – mais bem representadas na base de dados, também foram os que tiveram melhor desempenho nos modelos preditivos. Os códigos CE, IS e IS foram os que apresentaram maiores índices de precisão (considerando o *F1-score* do modelo de Redes Neurais, que apresentou os melhores resultados gerais), com índices variando de 0,50 a 0,67. Juntos, esses códigos representaram 73,4% da base de dados. Nenhum código com *F1-score* < 0,50 ocupava mais de 10% da base. A correlação forte entre o suporte e o *F1-score* nos 3 modelos, Regressão Logística (r = 0,70, p = 0,021), CART (r = 0,73, p = 0,015) e Redes Neurais (r = 0,72, p = 0,017), corrobora esse fato, demonstrando que códigos mais representados também são os que possuem melhores índices de precisão. Ademais, os resultados corroboram com os princípios do *machine learning*, em especial quando considerado o aprendizado de máquina supervisionado, uma vez que um grande volume de dados é essencial para a qualidade dos modelos de predição (Sen et al., 2020; Howard, 2019; Gorate et al., 2017).

Quando considerados somente os cursos cujos códigos de interesses profissionais estavam bem representados na base de dados (CE com 30,6% e SI com 26,3%), os índices de precisão dos modelos variaram entre 0,55 e 0,67. Em termos práticos, e considerando a média ponderada de precisão desses códigos para o modelo de redes neurais, isso significa que somente variáveis sociodemográficas mapeadas pelo Enade podem ser utilizadas para explicar o comportamento de escolha de curso de em média 50% dos estudantes brasileiros.

Idealmente, a escolha de carreira deveria ser consistente com os interesses profissionais do sujeito (Lent et al, 1994). No entanto, os resultados sugerem que fatores sociais, econômicos, étnicos, familiares, pessoais e de gênero, por exemplo, podem afetar as escolhas profissionais, assim como encontrado na literatura (Lima et al., 2017). Resta saber se os resultados dos modelos de predição seriam melhores caso estivessem a disposição variáveis que indicassem os interesses profissionais de cada sujeito na base de dados.

É preciso considerar a limitação da interpretação dos resultados. A grande quantidade de variáveis independentes que alimentaram o modelo, potencializada pela estratégia de dicotomização das variáveis, a escolha das bibliotecas utilizadas para realizar as análises e o alto custo computacional envolvido nas análises, dificultou a investigação da contribuição única de cada uma das variáveis para cada modelo preditivo. Inicialmente, não foi possível saber qual variável teve mais influência para a precisão de cada um dos modelos. No entanto, a limitação foi contornada ao utilizar a técnica de *Gradient Boosting*, conforme descrito por Friedman (2001), que facilitou a análise da importância das variáveis de forma mais eficiente e com menor custo computacional. A aplicação do *Gradient Boosting* permitiu investigar a influência das variáveis na acurácia geral de cada modelo e na contribuição para a avaliação de cada código RIASEC. Dessa forma, foi possível identificar as características sociodemográficas que mais impactam na categorização de cada código.

Ao realizar a análise de importância das variáveis utilizando o *Gradient Boosting*, observou-se variações na explicação da variabilidade do modelo ao comparar variáveis originais e *dummies*. As análises foram realizadas com os códigos com índice F1-score mais altos, CE, IS e SI.

Para a classe CE, composta pelos cursos de Administração, Secretariado Executivo, Tecnologia em Logística e Tecnologia em Processos Gerenciais, as variáveis mais influentes na predição foram o motivo da escolha do curso (34,7%), a situação de trabalho (31,5%) e a

situação financeira (12,4%), explicando 78,6% da variabilidade do modelo. Esses resultados refletem as motivações profissionais, as condições de empregabilidade e as necessidades financeiras dos estudantes. Quando as variáveis foram desmembradas em *dummies*, 66,8% da variabilidade foi explicada por três principais variáveis: situação de trabalho, alternativa E - Trabalho 40 horas semanais ou mais (34,4%), motivo da escolha do curso, alternativa E - Vocação (22,6%) e motivo da escolha do curso, alternativa H - Outro motivo (9,8%). Esses resultados revelam que as condições de trabalho intensivas e as motivações "vocacionais" são fatores determinantes para os estudantes da classe CE. Cabe ressaltar que no Brasil o termo vocação não é comumente usado entre especialistas de orientação profissional e de carreira, apesar de ter sido usado no ENADE.

Para a classe IS, composta pelos cursos de Medicina, Nutrição e Odontologia, as principais variáveis foram a situação financeira (36,7%), o tipo de escola de ensino médio (14,2%), a renda familiar (10,9%) e a situação de trabalho (10,7%), explicando 72,6% da variabilidade do modelo. Esses achados indicam que as condições financeiras e o tipo de formação educacional são determinantes significativos para os estudantes dessa classe. Na análise com variáveis *dummies*, 65,7% da variabilidade foi explicada por três variáveis: situação financeira, alternativa B - Não tenho renda e meus gastos são financiados pela minha família ou por outras pessoas (43,3%), tipo de escola de ensino médio, alternativa B - Todo em escola privada (particular) (15,9%) e bolsa de estudos/financiamento do curso, alternativa E - FIES, apenas (6,5%). A análise *dummies* destacou como a falta de renda própria e o suporte familiar são fatores críticos para os estudantes da classe IS.

Para a classe SI, composta pelos cursos de Enfermagem, Fisioterapia e Psicologia, as variáveis mais importantes foram o sexo (28%), o motivo da escolha do curso (24,4%) e a bolsa de estudos/financiamento do curso (18,3%), explicando 70,7% da variabilidade do modelo. Esses resultados destacam como o gênero, as motivações vocacionais e o suporte

financeiro são fatores preponderantes para os estudantes na classe SI. Na análise com variáveis *dummies*, 61,1% da variabilidade foi explicada por três variáveis: sexo masculino (29,3%), bolsa de estudos/financiamento do curso, alternativa E - FIES, apenas (17,1%) e motivo da escolha do curso, alternativa E - Vocação (14,7%). Esse resultado é particularmente interessante, pois a presença masculina nos cursos predominantemente femininos torna-se um fator preditivo significativo. A análise com *dummies* permitiu identificar essas nuances específicas que não seriam visíveis ao considerar apenas as variáveis originais, proporcionando insights valiosos sobre as influências sociodemográficas para a classe SI.

Para estudos futuros, sugere-se que sejam utilizadas técnicas como o *Permutation Importance* ou SHAP *(SHapley Additive exPlanations*), que também são capazes de identificar a contribuição única de características sociodemográficas para a acurácia de cada modelo de forma detalhada e interpretável.

O segundo objetivo deste estudo foi verificar qual modelo preditivo (Regressão Logística Multiclasse, CART e Redes Neurais) era melhor para predizer códigos RIASEC a partir de variáveis sociodemográficas. Esse resultado também foi atingido, mas com algumas limitações.

O algoritmo de Regressão Logística Multiclasse é projetado para prever a probabilidade de ocorrência de um evento, ajustando os dados a uma função logística. Esses modelos são treinados utilizando dados que já foram previamente classificados, ou seja, dados com rótulos conhecidos (Zahour et al., 2020). Para esse estudo, o algoritmo de Regressão Logística Multiclasse teve um bom desempenho e se mostrou equiparável a modelos mais atuais e computacionalmente mais potentes (acurácia de 0,54 versus 0,54, em comparação a Redes Neurais), apesar do *F1-score* médio ponderado de 0,50 comparado ao *F1-score* de 0,54 das Redes Neurais. Mesmo com a chegada de novos métodos estatísticos,

mais potentes e mais maleáveis, a utilização de regressão logística multiclasse para tarefas de classificação como a apresentada nesse estudo ainda pode ser considerada adequada.

O modelo CART pode ser empregado tanto para tarefas de classificação quanto de regressão. Este modelo funciona particionando os dados e, dentro de cada partição, são realizadas previsões simples, representadas por uma árvore de decisão. Em problemas de classificação, as árvores são projetadas para variáveis dependentes com um número finito de valores não ordenados (Loh, 2011). No entanto, os resultados para esse modelo foram consideravelmente piores em comparação com os outros modelos testados nesse estudo. A maioria dos índices de acurácia (*F1-score*), não passou de 0,45, exceto para CE, com 0,55. Sua utilização para esse tipo de tarefa precisa ser melhor investigada em estudo posteriores.

Ambos os modelos de Regressão Logística Multiclasse e Redes Neurais apresentaram a acurácia de 0,54 e um *F1-score* médio ponderado de 0,50, demonstrando desempenho semelhante na predição dos códigos RIASEC a partir de variáveis sociodemográficas. No entanto, houve variações significativas em termos de precisão, sensibilidade e especificidade para diferentes códigos RIASEC. O modelo de Regressão Logística Multiclasse obteve uma precisão notavelmente alta para o código CE (0,59) e uma sensibilidade de 0,79, resultando em um F1-score de 0,68. Em contraste, o modelo de Redes Neurais apresentou uma precisão e sensibilidade igualmente altas para o código CE, com um *F1-score* de 0,67, e também mostrou bom desempenho para o código SI, com um *F1-score* de 0,58.

Embora a acurácia geral dos modelos tenha sido igual (0,54), a especificidade foi consistentemente alta em ambos os modelos, indicando a capacidade dos algoritmos em identificar corretamente as categorias negativas. No entanto, a precisão para códigos como EC e IC foi baixa em ambos os modelos, o que sugere a necessidade de melhorias na modelagem desses códigos específicos.

Por fim, cabe ressaltar que os algoritmos de redes neurais são mais atuais, e são estruturados em diferentes camadas, de entrada, ocultas e de saída (Gorade et al., 2017). Mesmo sem a otimização das camadas ocultas, o modelo apresentou resultados muito bons quando comparados com seus pares. Para estudos posteriores, sugere-se que sejam testadas estruturas diferentes e otimizadas para melhor classificação dos códigos RIASEC.

Considerações Finais

Este estudo teve como objetivo principal testar o poder preditivo das variáveis sociodemográficas sobre os tipos de interesses profissionais, utilizando técnicas de machine learning, a fim de interpretar os impactos que essas variáveis possuem na formação dos interesses profissionais, mesmo que os mesmos não estejam explícitos na base de dados. Os resultados obtidos indicaram que as variáveis sociodemográficas são capazes de predizer, em certa medida, os perfis de interesses dos cursos superiores classificados de acordo com o modelo RIASEC.

Entre os modelos analisados, as redes neurais e a regressão logística multiclasse apresentaram resultados iguais, com pequenas diferenças superando a regressão logística multiclasse e o modelo CART. As redes neurais e a regressão logística multiclasse tiveram um F1-score médio ponderado de 0,50. Isso demonstra que ambos os modelos são igualmente eficazes na predição dos perfis RIASEC, com as redes neurais mostrando ligeira vantagem em algumas métricas individuais.

Os resultados revelaram que os códigos com maior representatividade na base de dados, como CE, SI e IS, apresentaram os melhores índices de precisão. Este achado reforça a importância de uma amostra robusta para a eficácia dos modelos preditivos de *machine learning*. Além disso, a correlação positiva significativa entre o suporte das classes e *o F1-score* sugere que a representatividade das classes influencia diretamente a acurácia dos modelos.

Inicialmente, a interpretação da importância das variáveis foi limitada devido à grande quantidade de variáveis independentes e à estratégia de dicotomização. No entanto, essa limitação foi contornada ao utilizar a técnica de *Gradient Boosting*, conforme descrito por Friedman (2001). A aplicação do *Gradient Boosting* permitiu identificar as características

sociodemográficas que mais impactam na categorização de cada código RIASEC, proporcionando uma análise mais eficiente e com menor custo computacional.

Os resultados sugerem que fatores sociais, econômicos, étnicos, familiares, pessoais e de sexo podem influenciar significativamente as escolhas profissionais, corroborando com a literatura existente. Para estudos futuros, recomenda-se a inclusão de uma gama mais ampla de variáveis, como dados psicométricos, históricos acadêmicos e características de personalidade, para melhorar a capacidade preditiva dos modelos.

Em conclusão, este estudo aprofundou a compreensão do impacto das variáveis sociodemográficas na predição dos interesses profissionais, demonstrando a eficácia das técnicas de *machine learning*, especialmente das redes neurais e regressão logística multiclasse, na avaliação psicológica. Os resultados mostraram que as variáveis sociodemográficas podem, em certa medida, predizer os interesses profissionais, com as redes neurais apresentando um desempenho melhor que os modelos testados, mas muito próximo das métricas do modelo de regressão logística multiclasse. Essas descobertas ressaltam a importância de incorporar dados sociodemográficos detalhados em análises preditivas para melhorar a precisão das avaliações relativas aos interesses profissionais. Além disso, o estudo identificou a necessidade de uma base de dados mais equilibrada e diversificada para aprimorar a capacidade preditiva dos modelos, sugerindo direções para futuras pesquisas nesta área.

Referências

Ambiel, R. A. M. (2019). Avaliação psicológica aplicada aos processos de escolha e transição de carreira. Em Compêndio de Avaliação Psicológica (pp. 262-272). Editora Vozes.

Ambiel, R. A. M., Lamas, K. C. A., & Melo-Silva, L. L. (2016). Avaliação dos interesses profissionais no Brasil: revisão da produção científica. Avaliação Psicológica, 15 (especial), 1-9. doi:10.15689/ap.2016.15ee.01

Ambiel, R. A. M., Noronha, A. P. P., & Martins, G. H. (2020). Manual Técnico Brasileiro da 5ª Edição do Questionário de Busca Autodirigida. Relatório técnico não publicado.

Bandura, A. (1986). *The Explanatory and Predictive Scope of Self-Efficacy Theory. Journal of Social and Clinical Psychology,* 4(3), 359–373.

Belyanova, M., Chernobrovkin, S., I Latkin, I. & Samarev, R. (2019). Comparison of convolutional neural networks and search based approaches for extracting psychological characteristics from job description. Proceedings of the International Symposium on Neural Networks, 491–500.https://doi.org/0.1521/jscp.1986.4.3.359

Bogacheva, Eugenia & Tatarenko, Filipp & Smetannikov, Ivan. (2020). *Predicting Vocational Personality Type from Socio-demographic Features Using Machine Learning Methods.* 93-98. 10.1145/3437802.3437819.

Bryant, B.K., Zvonkovic, A.M., & Reynolds, P. (2006). *Parenting in relation to child and adolescent vocational development. Journal of Vocational Behavior*, 69, 149–175. doi:10.1016/j.jvb.2006.02.004

Choy, G., Khalilzadeh, O., Michalski, M.H., Do, S., Samir, A., Pianykh, O., Geis, J.R., Pandharipande, P., Brink, J., & Dreyer, K. (2018). *Current Applications and Future Impact of Machine Learning in Radiology. Radiology,* 288 2, 318-328.

Dasgupta, A., & Nath, A. (2016). *Classification of machine learning algorithms. International Journal of Innovative Research in Advanced Engineering* (IJIRAE), 3(3), 6-11.

Dawis, R. V. (2005). *The Minnesota theory of work adjustment. Career Development,* 1.

Diekman, A. B., & Eagly, A. H. (2008). Of women, men, and motivation: A role congruity account. In J. Y. Shah, & W. L. Gardner (Eds.), Handbook of motivation science *(pp. 434-447). Guilford Press.*

Duffy, R. D., Blustein, D. L., Diemer, M. A., & Autin, K. L. (2016). *The Psychology of Working Theory. Journal of Counseling Psychology,* 63(2), 127–148. https://doi.org/10.1037/cou0000140

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Franco, Víthor Rosa. (2021). Aprendizado de Máquina e Psicometria: Inovações Analíticas na Avaliação Psicológica. Avaliação Psicológica, 20(3), a-c. https://dx.doi.org/10.15689/ap.2021.2003.ed

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29*(5), 1189-1232. doi:10.1214/aos/1013203451

Fritsch, S., Guenther, F. and Wright, M. N. (2019). neuralnet: Training of Neural Networks. *R* package version 1.44.2. https://CRAN.R-project.org/package=neuralnet

GIMENES, E. *DATA MINING – DATA WAREHOUSE.* A importância da mineração de dados em tomadas de decisões. Taquaritinga: 2000.

Gottfredson, L. S. (1981). Circumscription and compromise: A developmental theory of occupational aspirations. Journal of Counseling Psychology, 28(6), 545-579.

Gorade, S. M., Deo, A., & Purohit, P. (2017). *A study of some data mining classification techniques. International Research J. of Engineering and Technology* (IRJET), 4.

Harrington, T, & Long, J. (2013). *The history of interest inventories and career assessments in career counseling. The Career Development Quarterly*, 61, 83-92.

Hoff K.A., Perlus J.G., Rounds J. (2019) *Vocational Interests: Revisiting Assumptions About Their Development and What They Predict.* In: Athanasou J., Perera H. (eds) *International Handbook of Career Guidance. Springer*, Cham. https://doi.org/10.1007/978-3-030-25153-6_31

Holland, J. L. (1959). *A theory of vocational choice. Journal of Counseling Psychology*, 6(1), 35-45. http://dx.doi.org/10.1037/h0040767

Holland, J. L. (1966/1975). *Técnica de la eleción vocacional: Tipos de personalidad y modelos ambientales* (F. P. López, Trad.). México, D.F: Trillas.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Howard, J. *Artificial intelligence: Implications for the future of work*. Am J Ind Med. 2019; 62: 917– 926. https://doi.org/10.1002/ajim23037

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. (2021). Apresentação do Enade. Brasília: MEC/INEP. Retirado de https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade

Jordan, M. I., & Mitchell, T. M. (2015). *Machine learning: Trends, perspectives, and prospects. Science* (New York, N.Y.), 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kormushev, P., Calinon, S., & Caldwell, D. (2013). *Reinforcement Learning in Robotics: Applications and Real-World Challenges. Robotics*, 2(3), 122–148. doi:10.3390/robotics2030122

Kuhn, Max (2021). *caret: Classification and Regression Training. R package version* 6.0-88. https://CRAN.R-project.org/package=caret

Lent, R. W., Brow, S. D., & Hackett, G. (1994). *Towards a unifying social cognitive theory of career and academic interests, choice and performance. Journal of Vocational Behavior*, 45(1), 79-122. doi:10.1006/jvbe.1994.1027

Liao, H.Y, Armstrong, P.I., & Rounds, J. (2008). Development and initial validation of public domain Basic Interest Markers. Journal of Vocational Behavior. 73. 159-183. [doi: 10.1016/j.jvb.2007.12.002].

Lima, F. I. A. de, Voig, A. E. G. T., Feijó, M. R., Camargo, M. L., & Cardoso, H. F. (2017). A influência da construção de papeis sociais de gênero na escolha profissional. DOXA: Revista Brasileira De Psicologia E Educação, 19(1), 33–50. https://doi.org/10.30715/rbpe.v19.n1.2017.10818

LIMA, Priscila da Silva Neves et al. Análise de dados do Enade e Enem: uma revisão sistemática da literatura. Avaliação, v. 24, n. 1, p. 89-107, 2019

Loh, W.-Y. (2011), Classification and regression trees. WIREs Data Mining Knowl Discov, 1: 14-23. https://doi.org/10.1002/widm.8

Low, K. S. D., Yoon, M., Roberts, B. W., & Rounds, J. (2005). *The stability of vocational interests from early adolescence to middle adulthood: A quantitative review of longitudinal studies. Psychological Bulletin,* 131(5), 713-737. http://dx.doi.org/10.1037/0033-2909.131.5.713

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Friedrich Leisch (2021). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly*: *E1071*), TU Wien. R *package version* 1.7-7. https://CRAN.R-project.org/package=e1071

McCarthy J, Minsky, M.L., Rochester, N. & Shannon, C.E. *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* 1955. http://jmc.stanford.edu/articles/dartmouth.html.

Monard, M. C., & Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. Sistemas inteligentes-Fundamentos e aplicações, 1(1), 32.

Ng, T. W., Eby, L. T., Sorensen, K. L., & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. Personnel psychology, 58(2), 367-408.

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2012). *Vocational interests and performance: A quantitative summary of over 60 years of research. Perspectives on Psychological Science,* 7(4), 384-403. doi: 10.1177/1745691612449021

Nye, C. D., Su, R., Rounds, J., & Drasgow, F. (2017). *Interest congruence and performance: Revisiting recent meta-analytic findings. Journal of Vocational Behavior*, 98, 138-151. doi:10.1016/j.jvb.2016.11.002

Primi, Ricardo. (2018). Avaliação Psicológica no Século XXI: de Onde Viemos e para Onde Vamos. Psicologia: Ciência e Profissão, 38(spe), 87-97. https://doi.org/10.1590/1982-3703000209814

R Core Team (2021). R: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

Reis, M., Camacho, I., Ramiro, L., Tomé, G., Gomes, P., Gaspar, T., …Matos, M. G. (2015). A escola e a transição para a universidade: Idades transicionais e o seu impacto na saúde - notas a partir do estudo HBSC/OMS. Journal of Child and Adolescent Psychology, 6(2), 77-92. Recuperado em https://www.researchgate.net/publication/301354028

Revelle, W. (2020) psych: *Procedures for Personality and Psychological Research*, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.3.

Rounds, J., & Su, R. (2014). *The nature and power of interests. Current Directions in he Nature and Power of Interests. Association for Psychological Science*, 23(2), 98-103. doi: 10.1177/0963721414522812

Roth, P. L., Van Iddekinge, C. H., DeOrtentiis, P. S., Hackney, K. J., Zhang, L., & Buster, M. A. (2017). Hispanic and Asian performance on selection procedures: A narrative and meta-analytic review of 12 common predictors. Journal of Applied Psychology, 102(8), 1178-1202.

Sánchez-Hernández, G., Chiclana, F., Agell, N., & Aguado, J. C. (2013*). Ranking and selection of unsupervised learning marketing segmentation. Knowledge-based systems*, 44, 20-33.

Savickas, M. L. (1995). *Examining the personal meaning of inventoried interests during career counseling. Journal of Career Assessment*, 3(2), 188-201.

Savickas, M. L. (1999). The Psychology of Interests. In M. L. Savickas & A. R. Spokane (Eds.), *Vocational Interests: Meaning, measurement and counseling use* (pp. 19-56). Palo Alto, CA: Davies-Black.

Savickas, M. L. (2005). *The theory and practice of career construction*. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 42–70). Hoboken, NJ: Wiley

Savickas, M. L., Nota, L., Rossier, J., Dauwalder, J. P., Duarte, M. E., & Guicharad, J., ... VanVianen, A. E. M. (2009). Life designing: A paradigm for career construction in the 21st century. *Journal of Vocational Behavior*, 75, 239–250. doi: 10.1016/j.jvb.2009.04.004

Sen P.C., Hajra M., Ghosh M. (2020) Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal J., Bhattacharya D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_11

Shafique, U & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research. 12. 2351-8014.

Silva, A., Lo, P.C. & Lim, E.P. (2020). Jplink: on linking jobs to vocational interest types. Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 220–232.

Stoll, G., & Trautwein, U. (2017). *Vocational interests as personality traits. Personality Development Across the Lifespan, 401–417.* doi:10.1016/b978-0-12-804674-6.00025-9

Su, R. The three faces of interests: An integrative review of interest research in vocational, organizational, and educational psychology. Yjvbe (2018), doi.org/10.1016/j.jvb.2018.10.016

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: a meta-analysis of sex diferences in interests. Psychological Bulletin, 135(6), 859-884.

Super, D. E. (1957). The Psychology of Careers. New York: Harper and Row.

Therneau, T. & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees. R package

version 4.1-15. https://CRAN.R-project.org/package=rpart

U.S. Department of Labor, Employment and Training Administration. (2024). O*NET Resource Center. Retrieved from https://www.onetcenter.org

Usama F. Paul, S. Data mining and KDD: Promise and challenges, Future Generation Computer Systems, Volume 13, Issues 2–3, 1997, Pages 99-115, https://doi.org/10.1016/S0167-739X(97)00015-0.

Van Iddekinge, C. H., Putka, D. J., & Campbell, J. P. (2011). *Reconsidering vocational interests for personnel selection: The validity of an interest-based selection test in relation to job knowledge, job performance, and continuance intentions. Journal of Applied Psychology*, 96(1), 13-33. doi: 10.1037/a0021193

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition*. Springer, New York. ISBN 0-387-95457-0

Yarkoni, T., & Westfall, J. (2017). *Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science*, 12(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zahour, O., Benlahmar, E., Eddaouim, A., & Hourrane, O. (2020). *A Comparative Study of Machine Learning Methods for Automatic Classification of Academic and Vocational Guidance Questions. International Journal Of Interactive Mobile Technologies* (IJIM), 14(08), pp. 43-60. doi:http://dx.doi.org/10.3991/ijim.v14i08.13005

Anexos

Anexo A - Ficha de avaliação dos juízes

ESCOLHA DO TIPO RIASEC DOMINANTE DOS CURSOS AVALIADOS PELO ENADE 2018 E 2019

Estou desenvolvendo um projeto de pesquisa, cujo objetivo é testar o poder preditivo das variáveis sociodemográficas sobre os ambientes profissionais, com utilização de técnicas de *machine learning.* A partir do objetivo da presente pesquisa, iremos utilizar os microdados do Enade 2018 e 2019, por ter um questionário sociodemográfico amplo, além de ter a identificação do curso superior do participante. Assim, solicito sua colaboração, quanto a identificação dos dois perfis tipológicos RIASEC que mais caracterizam cada um dos 56 cursos. Estarei disponível para esclarecer qualquer dúvida.
Atenciosamente,

Felipe Alvarenga Dinardi Barbosa
(Discente de mestrado do PPG USF, autor do estudo)
Prof. Dr. Prof. Dr. Nelson Hauck Filho
(Pesquisador, docente do PPG USF, orientador do estudo)

Documento 1: Avaliação dos juízes
Instrução: Favor identificar os 2 perfis tipológicos RIASEC que mais caracterizam cada curso do ensino superior na lista abaixo. Caso tenha algum comentário, fique à vontade.
Exemplo: Curso X - AS; Curso Y - EC; Curso Z - IR.

Bacharelado/Tecnólogo	Curso	RIASEC	Comentários
Bacharelado	ADMINISTRAÇÃO		
Bacharelado	ADMINISTRAÇÃO PÚBLICA		
Bacharelado	AGRONOMIA		
Bacharelado	ARQUITETURA E URBANISMO		
Bacharelado	BIOMEDICINA		
Bacharelado	CIÊNCIAS CONTÁBEIS		
Bacharelado	CIÊNCIAS ECONÔMICAS		
Bacharelado	COMUNICAÇÃO SOCIAL – JORNALISMO		
Bacharelado	COMUNICAÇÃO SOCIAL - PUBLICIDADE E PROPAGANDA		
Bacharelado	DESIGN		
Bacharelado	DIREITO		

Bacharelado	EDUCAÇÃO FÍSICA (BACHARELADO)		
Bacharelado	ENFERMAGEM		
Bacharelado	ENGENHARIA AMBIENTAL		
Bacharelado	ENGENHARIA CIVIL		
Bacharelado	ENGENHARIA DA COMPUTAÇÃO		
Bacharelado	ENGENHARIA DE ALIMENTOS		
Bacharelado	ENGENHARIA DE CONTROLE E AUTOMAÇÃO		
Bacharelado	ENGENHARIA DE PRODUÇÃO		
Bacharelado	ENGENHARIA ELÉTRICA		
Bacharelado	ENGENHARIA FLORESTAL		
Bacharelado	ENGENHARIA MECÂNICA		
Bacharelado	ENGENHARIA QUÍMICA		
Bacharelado	FARMÁCIA		
Bacharelado	FISIOTERAPIA		
Bacharelado	FONOAUDIOLOGIA		
Bacharelado	MEDICINA		
Bacharelado	MEDICINA VETERINÁRIA		
Bacharelado	NUTRIÇÃO		
Bacharelado	ODONTOLOGIA		
Bacharelado	PSICOLOGIA		
Bacharelado	RELAÇÕES INTERNACIONAIS		
Bacharelado	SECRETARIADO EXECUTIVO		
Bacharelado	SERVIÇO SOCIAL		
Bacharelado	TEOLOGIA		
Bacharelado	TURISMO		
Bacharelado	ZOOTECNIA		
Tecnólogo	TECNOLOGIA EM AGRONEGÓCIOS		
Tecnólogo	TECNOLOGIA EM COMÉRCIO EXTERIOR		
Tecnólogo	TECNOLOGIA EM DESIGN DE INTERIORES		
Tecnólogo	TECNOLOGIA EM DESIGN DE MODA		
Tecnólogo	TECNOLOGIA EM DESIGN GRÁFICO		
Tecnólogo	TECNOLOGIA EM ESTÉTICA E COSMÉTICA		
Tecnólogo	TECNOLOGIA EM GASTRONOMIA		
Tecnólogo	TECNOLOGIA EM GESTÃO		

	AMBIENTAL		
Tecnólogo	TECNOLOGIA EM GESTÃO COMERCIAL		
Tecnólogo	TECNOLOGIA EM GESTÃO DA QUALIDADE		
Tecnólogo	TECNOLOGIA EM GESTÃO DE RECURSOS HUMANOS		
Tecnólogo	TECNOLOGIA EM GESTÃO FINANCEIRA		
Tecnólogo	TECNOLOGIA EM GESTÃO HOSPITALAR		
Tecnólogo	TECNOLOGIA EM GESTÃO PÚBLICA		
Tecnólogo	TECNOLOGIA EM LOGÍSTICA		
Tecnólogo	TECNOLOGIA EM MARKETING		
Tecnólogo	TECNOLOGIA EM PROCESSOS GERENCIAIS		
Tecnólogo	TECNOLOGIA EM RADIOLOGIA		
Tecnólogo	TECNOLOGIA EM SEGURANÇA NO TRABALHO		