NOVA IMS Information Management School

MDSAA

Mestrado em

Data Science and Advanced Analytics

PREVISÃO DE DESEMPENHO DOS CANDIDATOS A FORMAÇÃO PROFISSIONAL

Identificação de medidas para mitigar previsíveis insucessos

Maria Dordio Lobo da Conceição Oliveira

Dissertação

apresentada como requisito parcial para obtenção do grau de Mestre em Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREVISÃO DE DESEMPENHO DOS CANDIDATOS A FORMAÇÃO PROFISSIONAL

Identificação de medidas para mitigar previsíveis insucessos

por

Maria Dordio Lobo da Conceição Oliveira

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre em Advanced Analytics, com especialização em Business Analytics

Orientador: Professor Doutor Vitor Duarte dos Santos

Coorientador*:* Susana Gonçalves

Julho 2023

DECLARAÇÃO DE INTEGRIDADE

Declaro ter realizado o presente trabalho académico com integridade. Confirmo que não recorri à prática de plágio ou de qualquer outra forma de utilização indevida de informação ou de falsificação de resultados durante o processo de elaboração deste trabalho. Declaro ainda que tenho conhecimento das Regras de Conduta e do Código de Honra da NOVA Information Management School.

Maria Dordio Lobo da Conceição Oliveira

Setúbal, Julho 2023

AGRADECIMENTOS

A conclusão desta dissertação marca assim o fim de mais uma etapa na qual consegui comprovar a minha resiliência ao enfrentar tanto desafios profissionais como académicos. No entanto, esta conquista não foi apenas fruto da minha determinação, foi também fruto do apoio incondicional que recebi por parte de todas as pessoas que me acompanharam neste percurso e estiveram sempre presentes até ao fim.

Assim, quero deixar um sincero agradecimento aos meus queridos pais, pela inestimável contribuição e apoio ao longo desta jornada, que se mostraram sempre disponíveis, quando eu mais precisei e mesmo quando não precisei e que, acima de tudo, nunca deixaram de acreditar em mim.

Quero também deixar um grande agradecimento ao meu irmão, que esteve sempre por perto a acompanhar esta etapa e que se mostrou sempre disponível para me ajudar, que me incentivou a ser melhor e a procurar saber mais, que deu sempre os melhores conselhos quando eu precisava e principalmente que acreditou em mim e puxou sempre por mim.

Um especial agradecimento aos meus avós, que estiveram sempre na fila da frente a apoiarem-me e a aplaudir as minhas conquistas sempre de perto.

Aos meus amigos que estiveram sempre lá para mim e que estiveram sempre disponíveis para me ouvirem sempre que precisei.

Quero agradecer ao meu orientador pelo seu contínuo apoio e incentivo durante todo o percurso desta dissertação, pelas suas pertinentes contribuições, bem como pela sua experiência e pelo seu conhecimento que foram cruciais para que este projeto tenha chegado a bom porto.

Por fim, não posso deixar de agradecer a toda a equipa do Citeforma que acompanhou este processo e que se disponibilizou para me ajudar, tanto na recolha dos dados como na sua compreensão. Em particular, à minha coorientadora, pela sua partilha de conhecimento da área que me ajudou a compreender o contexto deste estudo.

RESUMO

A formação profissional tem vindo a ganhar relevância no campo educacional e da qualificação e requalificação da população em idade ativa. É um recurso cada vez mais utilizado pela população que vê nele uma resposta mais rápida para obter qualificações que respondam às necessidades do mercado de trabalho.

Esta situação traduz o investimento em políticas públicas de emprego, em particular, do Instituto do Emprego e Formação Profissional (IEFP), que visam estabelecer um equilíbrio entre a oferta e a procura de emprego, tendo desenvolvido ao longo das últimas décadas diferentes estratégias que procuram responder às necessidades de uma mão de obra mais especializada e qualificada.

Tendo em conta a crescente procura da formação profissional e uma vez que que as ofertas formativas têm um número limitado de vagas é determinante conseguir prever o sucesso que os diferentes candidatos vão ter na formação. É neste contexto que o presente trabalho surge como uma mais-valia para os Centros de Formação, pois, através desta previsão, conseguirão ver as suas ofertas de formação preenchidas com o candidato certo, isto é, aquele que reúne as melhores condições para concluir a sua formação. Assim sendo, a seleção dos candidatos é feita através de testes de aptidão utilizados para avaliar o desempenho, numa determinada formação, com o objetivo de selecionar o candidato mais habilitado a frequentar determinado curso e a concluí-lo com sucesso.

Desta forma, foi feito um estudo em parceria com o Centro de Formação Profissional, Citeforma, que forneceu os dados relativos aos testes de aptidão realizados pelos diferentes candidatos. Este estudo teve como objetivo criar um modelo preditivo, através de técnicas de *machine learning*, que tendo em conta o resultado dos testes de aptidão dos candidatos à formação, permitiu prever o resultado dos formandos nos cursos profissionais a que se candidataram, percebendo também se estes são os mais adequados e, ao mesmo tempo, avaliar que medidas poderão ser tomadas para prevenir possíveis insucessos na formação ou desistências ao longo do seu percurso.

PALAVRAS-CHAVE

Formação Profissional; Testes de aptidão; *Machine Learning*; Redes Neuronais; Sucesso no Ensino e formação

Objetivos do Desenvolvimento Sustentável (ODS):

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

AB	*AdaBoost*
ABI 1	Aptidão Verbal
ABI 2	Aptidão Numérica
ABI 3	Atenção
ABI 4	Série de Números
ABI 5	Codificação
ABI 6	Diagramas
ANN	*Artificial Neural Network*
CE	Centro de Emprego
CET	Cursos de Especialização Tecnológica
CF	*Collaborative Filtering*
CGP	Centros de Gestão Participada
DT	*Decision Tree*
EFA	Educação e Formação de Adultos
FM	Formações Modulares
FPA	Formação Profissional Acelerada
IEFP	Instituto do Emprego e Formação Profissional
IFPA	Instituto de Formação Profissional Acelerada
KNN	*K-Nearest Neighbor*
LinR	*Linear Regression*
LMS	Sistemas de Gestão de Aprendizagem
MAE	Erro médio Absoluto
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
NB	*Naive Bayes*
PMA – F	Fluência Verbal

PMA E Conceção Espacial

PMA N Cálculo Numérico

PMA – R Raciocínio Lógico

PMA V Compreensão Verbal

RF *Random Forest*

RMSE Erro quadrático médio

RS *Recommender Systems*

RVCC Reconhecimento, Validação e Certificação de Competências

SVM *Support Vector Machine*

TIG I Teste de Inteligência Geral

TPD Teste de Perceção de Diferenças

1. INTRODUÇÃO

Esta dissertação tem como objeto de estudo a análise de dados, com base em modelos de *machine learning*, procurando prever qual será o desempenho dos candidatos nos diferentes cursos de formação profissional, identificando medidas que possam mitigar o insucesso na sua conclusão.

Tornou-se relevante fazer uma caracterização da formação profissional através de uma breve referência à sua história no nosso país, bem como do papel que assume no campo da educação, não só daqueles que procuram uma qualificação profissional para uma determinada profissão, mas também de todo um grupo de pessoas que vê nestes cursos a possibilidade de concluir um grau de ensino, ao mesmo tempo que fica habilitado para o exercício de uma atividade profissional.

Neste trabalho de investigação, o objetivo foi identificar quais as variáveis determinantes para o desempenho dos formandos, na conclusão da formação, e se os testes de aptidão são relevantes para determinar se um formando consegue concluir o curso com sucesso ou não. Com o recurso aos modelos de *machine learning*, foi possível perceber qual a importância dos testes de aptidão e, ao mesmo tempo, prever o desempenho de um candidato no curso, otimizando o processo de seleção dos candidatos, de forma a garantir o sucesso da formação profissional.

Para a realização deste estudo, foram utilizados os dados fornecidos pelo Centro de Formação Profissional – Citeforma, mais concretamente, os dados referentes aos cursos de Especialização Tecnológica (CET) e Educação e Formação de Adultos (EFA).

Com o objetivo de tornar o estudo realizado uma ferramenta útil para o centro Citeforma, criou-se uma *dashboard* que permite um acesso dinâmico, interativo e com informação atualizada aos dados referentes à formação existente.

1.1. ENQUADRAMENTO

Tendo em conta a conjuntura do mercado de trabalho atual, existe uma procura, quer por parte dos jovens, mas também por parte das empresas, de ofertas nas áreas profissionais específicas (Barbosa *et al*., 2019). Assim, torna-se essencial dinamizar toda uma oferta formativa que procure dar resposta, não só aos jovens que procuram alternativas ao ensino tradicional, mas também ao mercado de trabalho que procura mão de obra especializada.

Assim, para dar resposta a este problema, uma das soluções poderá estar no ensino profissional. Este tem como objetivo formar e capacitar os formandos, visando não só melhorar as suas habilitações, promover novas competências práticas e o saber fazer numa determinada profissão (Azevedo, 2010; Lourenço, 2015), mas também proporcionar uma maior aproximação entre o ensino e o mercado de trabalho, contribuindo para a diminuição do abandono escolar. Todos estes fatores, promovem o progresso das carreiras profissionais dos formandos, dando resposta às atuais necessidades do mercado de trabalho (Guerreiro, 2014) e, consequentemente, contribuindo para a diminuição da "exclusão social por falta de habilitações académicas e qualificações profissionais", possibilitando uma maior equidade entre a população e uma melhor qualidade de vida (Barbosa *et al.*, 2019; Guerreiro, 2014).

Deste modo, nas últimas décadas, Portugal tem investido cada vez mais na formação profissional, aumentado a sua oferta no ensino, mas tal não foi muito bem recebido pela maioria dos formandos,

nem pela sociedade em geral (Guerreiro, 2014). Isto acontece porque existe todo um preconceito associado aos cursos profissionais e aos formandos que os escolhem, bem como uma sua desvalorização, pois, para muitos, não são vistos como um percurso académico tradicional. Por outro lado, também, a sociedade tem tendência a associar este tipo de formação a jovens com percursos escolares irregulares (formandos que ficam retidos vários anos) (Pereira, 2018) e a contextos socioeconómicos mais desfavorecidos e/ou de famílias de classe operária e de trabalhos não qualificados (Pereira & Carvalho, 2021). Todas estas situações fazem com que os formandos repensem a sua escolha desta vertente académica (Barbosa *et al.*, 2019).

No entanto, desde 2014, os centros de formação e os respetivos cursos profissionais passaram a ser mais valorizados, não só pelo mundo empresarial, mas também pelos formandos e pela sociedade, assistindo-se a um aumento do número de jovens a frequentar este tipo de ensino (Pereira & Carvalho, 2021). Este crescimento mostra que os centros de formação têm cada vez um maior impacto e são uma realidade procurada pela população (Mais Formação, 2015).

Tendo em consideração a crescente procura da formação profissional, é fundamental que os diferentes cursos tenham candidatos adequados às ofertas (Tasnim *et al.*, 2022). Isto é relevante, pois, os formandos ficam com alguma garantia de que é o curso que procuram e pretendem, sendo, por outro lado, também importante que os cursos vejam as suas vagas preenchidas com formandos que os consigam concluir com sucesso.

Para que isto aconteça, ou seja, que os formandos escolhidos consigam concluir com sucesso a formação, é determinante que os métodos utilizados para essa escolha, pelos centros de formação, sejam os mais adequados. Um dos métodos de seleção utilizado para esse efeito, por centros de formação como o Citeforma, são os testes de aptidão (Citeforma, 2021b).

Estes centros têm como intuito aumentar e dar novas competências, através de formação, de modo a melhorarem as qualificações, dos jovens e adultos, indo ao encontro das ofertas do mercado de trabalho atual (Citeforma, 2021a). Assim, os cursos são frequentados por candidatos previamente selecionados, através de um conjunto de provas, nas quais se incluem os testes de aptidão.

Um teste de aptidão é uma forma de avaliação psicométrica e serve para avaliar a capacidade de um indivíduo, numa determinada atividade (Kagan, 2022; Practice: aptitude tests, n. d.). O objetivo da sua realização é perceber quais dos candidatos são mais capacitados para frequentar determinado curso e realizá-lo com sucesso (Conceitos, 2015; Setiawati, 2020). Assim, ao determinar a capacidade dos candidatos para concluírem determinado curso, consegue perceber-se se irão ou não ter um bom desempenho, pois estes testes têm um valor preditivo em termos do sucesso ou insucesso na ação de formação (Practice: aptitude tests, n. d.). É também muito importante que os testes de aptidão sejam adaptados à formação a que o candidato se candidatou, para que o próprio teste, assim como os resultados, sejam o mais próximo possível da realidade (Conceitos, 2015).

Nestes testes, o candidato não precisa de ter nenhum tipo de competência, preparação ou conhecimento prévio para a sua realização, o que permitirá, em seguida, uma comparação dos resultados dos diferentes testes realizados pelos candidatos, para o mesmo curso, sem que haja qualquer discrepância ou vantagem por parte dos formandos. Isto é, estes testes não têm como objetivo comparar as qualificações ou as experiências entre formandos (Practice: aptitude tests, n.d.).

Deste modo, os testes de aptidão mais comuns para avaliar as competências são: raciocínio numérico, desenvolvimento verbal, raciocínio lógico, entre outros (Kagan, 2022; Practice: aptitude tests, n. d.; Setiawati, 2020). Os diferentes testes servem para avaliar as diferentes competências necessárias para cada situação, neste caso, para cada curso. Assim, cada curso valoriza diferentes testes, atendendo ao que mais se adequam a cada situação (Reis, 2018).

1.2. MOTIVAÇÃO

Para que exista sucesso na conclusão dos cursos escolhidos, deve haver uma correspondência entre os resultados obtidos nos testes de aptidão e a formação escolhida pelo candidato. No entanto, apesar de os formandos serem selecionados através de um conjunto de provas, que têm como objetivo garantir que os formandos estão na formação certa, muitas vezes, os resultados obtidos nessas provas poderão, nem sempre, serem suficientes para fazer a seleção certa dos candidatos (Sodhi *et al.*, 2016). Quer isto dizer que os testes de aptidão poderão não ser os mais adequados, podendo levar à escolha errada dos candidatos ou que os próprios testes de aptidão não serão suficientes para se fazer uma boa seleção dos candidatos.

Estas situações verificam-se, por exemplo, quando candidatos com bons resultados nos testes de aptidão, não concluem o curso com sucesso. Pode acontecer também, que os resultados dos testes de aptidão eliminem candidatos que poderiam ter um bom desempenho no curso. Desta forma, o que acontece é que essa correspondência nem sempre se traduz num resultado positivo.

Assim, e com a crescente procura desta formação profissional, é necessário que a escolha dos candidatos seja rigorosa e a mais acertada possível, surgindo por isso então a necessidade de tentar perceber de que modo se poderá aperfeiçoar o método de seleção de candidatos.

Deste modo, as técnicas de *machine learning*, com destaque para o modelo de redes neuronais, surgem como uma mais-valia para analisar, com precisão, os dados disponíveis, previamente recolhidos nos testes de aptidão, para se conseguir perceber se um determinado formando terá ou não sucesso no curso pretendido (Sodhi *et al.*, 2016). Desta forma, o que se procura perceber é: poderão os testes de aptidão prever o desempenho de um formando, num curso profissional?

Complementarmente, pretende-se também responder às seguintes questões de investigação:

- Quais são as características dos candidatos, com boas taxas de sucesso, nos diferentes cursos, que possibilitam formar um padrão de características dos formandos, de um determinado curso?
- Porque é que o mesmo conjunto de candidatos, selecionados para o mesmo curso, têm taxas de sucesso diferentes, tendo em conta que terão tido resultados semelhantes nos testes de aptidão?
- As variáveis que estão a ser tidas em conta serão as mais acertadas? Será necessário considerar e ponderar novas variáveis que possam determinar o sucesso do formando na realização do curso?
- Será que o intervalo do percentil escolhido, para a seleção dos candidatos, é o correto ou precisa de ser ajustado?

1.3. OBJETIVO

O objetivo da investigação é criar um modelo preditivo que, tendo em conta o resultado dos testes de aptidão dos candidatos a formação, permita prever o sucesso dos formandos nos cursos profissionais a que se candidataram e, com base nessa informação, tomar medidas preventivas para mitigar previsíveis insucessos. Tem-se também como meta, implementar uma ferramenta que permita aos utilizadores visualizarem a informação da previsão do desempenho dos formandos, mas que também lhes permita conhecer melhor as características dos formandos que frequentam estes cursos, para assim poderem selecionar os candidatos de forma mais informada.

Para atingir este objetivo, é essencial ter em conta os seguintes objetivos intermédios:

- Estudar a problemática da formação profissional e, em particular, a seleção dos formandos.
- Estudar detalhadamente o estado da arte da análise preditiva.
- Recolher e pré-preparar os dados dos testes de aptidão.
- Definir um modelo concetual de análise preditiva.
- Analisar os resultados e produzir conclusões.

1.4. IMPORTÂNCIA E RELEVÂNCIA DO ESTUDO

O presente estudo pretende melhorar a metodologia usada na seleção dos candidatos, nos centros de formação. Assim, pretende-se identificar quais as variáveis mais relevantes, nos testes de aptidão, de modo a que estes sejam mais eficazes e direcionados às características que cada curso exige, garantindo que os formandos selecionados são os que têm um maior potencial de o terminar com sucesso.

Além do referido atrás, o modelo que se pretende desenvolver também poderá ser utlizado numa vertente financeira, ajudando a uma melhor gestão dos recursos alocados aos centros de formação e aos respetivos cursos.

Todos estes objetivos irão traduzir-se num melhor tratamento dos dados na seleção dos candidatos e, posteriormente, numa melhor escolha. Isto é importante, pois cada vez mais os cursos profissionais são uma mais-valia para a população, promovendo a integração no mercado de trabalho e, consequentemente, uma melhor qualidade de vida (Barbosa *et al.*, 2019).

2. REVISÃO DE LITERATURA

Com o propósito de tomar conhecimento de mais informação, de estudos similares ao que se pretende fazer, foi necessário pesquisar outros trabalhos já feitos na área e ver o que já foi publicado acerca do tema. Esta pesquisa foi fundamental para a elaboração do estudo, tendo por base estudos anteriormente feitos, pois será a partir deles que se terá a oportunidade de saber o que já foi realizado, percebendo quais as metodologias que se deverão seguir e quais as que deverão ser evitadas, mas também, retirar ideias acerca do modo de realizar certos processos (Dorsa, 2020). Assim, posteriormente, serão analisados alguns pontos pertinentes para contextualizar o tema, sendo eles tópicos relacionados com a formação profissional, mas também com a análise preditiva.

2.1. HISTÓRICO DA FORMAÇÃO PROFISSIONAL EM PORTUGAL

A formação profissional em Portugal, desde que surgiu até à atualidade, sofreu diversas transformações, todas elas associadas a flutuações da conjuntura económica. Assim, esta formação teve início no ano de 1852, no século XIX, com os ecos da revolução industrial que criou a necessidade de preparar a mão de obra especializada para satisfazer as novas necessidades do comércio e da indústria. No entanto, quem frequentava este tipo de ensino era direcionado diretamente para o mercado de trabalho, impossibilitando a sua entrada na universidade, acabando por criar alguma distinção social entre quem frequentava e quem não frequentava este tipo de curso. Por conseguinte, observou-se, desde muito cedo, a discriminação associada a este tipo de formação profissional com carácter técnico, pois, quem não frequentava este tipo de ensino teria acesso a um outro tipo de profissões mais prestigiadas e conceituadas, descredibilizando este tipo de formação (Silvestre, 2009).

Em 1960, Portugal viu-se confrontado com uma nova realidade económica e com o progresso tecnológico na indústria, passando a ser necessário um reforço da produtividade nacional. A solução para este problema seria o modelo de formação lançado, na altura, pelo Ministério das Corporações, que foi concebido para preparar rapidamente as pessoas para desempenharem atividades profissionais específicas e minimizar o desemprego sentido na altura. Inicialmente, esta formação era destinada a adultos desempregados, alargando-se, mais tarde, à população mais jovem (IEFP, n. d.; Silvestre, 2009).

Assim, o Ministério das Corporações iniciou o modelo "Formação Profissional Acelerada" (FPA) ou também conhecido como "Formação profissional de adultos" e no mesmo ano foi criado o "Instituto de Formação Profissional Acelerada" (IFPA), com o intuito de requalificar a população, através de uma formação apropriada, aumentando o nível profissional dos trabalhadores (IEFP, n.d.; Silvestre, 2009). Deste modo, é a partir de 1963 que o primeiro centro de formação acelerada entra em funcionamento e o modelo de "Formação Profissional Acelerada" (FPA) é implementado, posteriormente, noutros centros de formação, a partir de 1964 (IEFP, n. d.).

Nos anos 1970, as empresas nacionais viviam da importação de tecnologia e *know how*, e os seus trabalhadores eram muitas vezes incapazes de aplicar os conhecimentos que eram importados, uma vez que havia um desfasamento entre o sistema de ensino e as necessidades de mão de obra qualificada, que se traduzia na incapacidade de refletir os ganhos da aplicação de novas tecnologias na produtividade da economia nacional. Deste modo, a necessidade de mão de obra qualificada para enfrentar os desafios do crescimento industrial e desenvolver a economia nacional, levou à criação do "Instituto do Emprego e Formação Profissional" (IEFP), em 1979, a cargo do Ministério do Trabalho.

Este novo organismo, visava não só colmatar a falta de mão de obra especializada, mas também tinha como objetivo ajudar a diminuir o desemprego que se fazia sentir na altura (IEFP, n. d.; Silvestre, 2009). Desde a sua criação até à atualidade, o IEFP foi sofrendo algumas alterações, centradas não só na componente de emprego, mas também na sua atividade formativa, de modo a que exista uma complementaridade entre as políticas de emprego e de formação profissional (IEFP, n.d.). Fica pois claro que este novo organismo (IEFP) não só tem como objetivos a formação e a qualificação de mão de obra voltada para as necessidades de mercado, mas também olha para a formação mais formal dessa mesma população (Silvestre, 2009).

Centrado neste novo propósito, o equilíbrio entre as necessidades do mercado e a oferta de formação profissional, o IEFP aposta numa estrutura descentralizada a nível das regiões (delegações regionais), contando também com a participação de parceiros sociais (Silvestre, 2009). As cinco delegações regionais (Norte, Centro, Lisboa e Vale do Tejo, Alentejo e Algarve) trabalham com dois órgãos operativos, os Centros de Emprego (CE) e os Centros de Formação Profissional. Paralelamente, existe a rede de Centros de Gestão Participada (CGP), que são estruturas criadas ao abrigo de protocolos celebrados entre o IEFP e os parceiros sociais e que visam complementar e reforçar a ação das delegações regionais, mas com uma vocação marcadamente setorial, abrangendo vários setores, que vão do comércio, à indústria, aos serviços, entre outros (IEFP, n. d.; Silvestre, 2009). É nesta rede que se enquadra o Citeforma – Centro de Formação Profissional dos Trabalhadores de Escritório, Comércio, Serviços e Novas Tecnologias, a entidade detentora dos dados sobre os quais se baseou este trabalho académico.

Posto isto, o IEFP, através dos Centros de Formação Profissional, continua com o objetivo de potenciar e qualificar a população e, consequentemente, aumentar a produtividade nos diferentes setores do mercado de trabalho, procurando adequar a oferta de formação às necessidades dos diferentes setores, recorrendo às suas unidades mais descentralizadas, os CGP.

2.1.1. Âmbito / Conceito: Importância e benefícios da formação profissional

Entende-se a formação profissional como um mecanismo em que se adquirem qualificações essenciais ao desenvolvimento económico e social, facilitando o acesso ao mercado de trabalho. Além disso, dá a oportunidade de aperfeiçoar as *skills* (competências práticas para o desenvolvimento de uma profissão), proporcionado o desenvolvimento dos indivíduos e a atualização das suas competências ou a aquisição de novos conhecimentos, ficando assim mais aptos ao exercício qualificado da atividade profissional (Neves, 2010).

A formação profissional tona-se importante, uma vez que, com as constantes transformações tecnológicas a acontecer no mercado de trabalho, surge a necessidade de a população ativa adquirir novas competências, de modo a conseguir acompanhar e fazer face às transformações e exigências do mercado atual. Não menos importante é a inovação que trouxe ao sistema educativo português, pois tornou-se uma alternativa ao ensino formal, estabelecendo uma ponte entre a escola e o mercado de trabalho, facilitando a integração. Por outro lado, também, não deve ser descurado o papel que o ensino profissional tem na prevenção do abandono escolar precoce dos jovens, que têm assim a oportunidade de concluir com sucesso a escolaridade obrigatória, permitindo também a requalificação profissional de adultos (Barbosa *et al.*, 2019).

Convém também assinalar o papel que têm no desenvolvimento socioeconómico da população, não só de uma população jovem que procura uma formação mais prática e curta que promova um acesso fácil ao mercado de trabalho e a sua integração social e profissional, mas também de toda a população ativa que procura novas oportunidades e vê na requalificação profissional a possibilidade de crescer a nível profissional (Azevedo, 2010; Barbosa *et al.*, 2019; Choi *et al.*, 2019; Rodrigues, 2010).

Importa também referir que o IEFP, através dos seus CGP, tem dois tipos de formação profissional distintos. A formação profissional inicial, direcionada a toda a população que procura emprego e que quer aumentar as suas qualificações, tem como objetivo dotar os indivíduos de novas competências e conhecimentos, necessários em determinadas profissões, promovendo a sua integração e uma maior facilidade em ingressar no mercado de trabalho; e a formação profissional contínua, que poderá ocorrer no interior das diferentes empresas e é destinada aos seus colaboradores, para que possam atualizar as suas competências e estarem a par dos novos avanços tecnológicos e das mudanças que existem nas suas áreas de trabalho. Ou seja, tem como intuito ajudar a população já empregada a ter a oportunidade de atualizar e aperfeiçoar as suas qualificações, indo ao encontro de novas necessidades do mercado de trabalho (Jamba, 2018; Silvestre, 2009; Sumbo, 2019).

Para além dos diversos tipos de formação dos CGP atrás referidos, importa também destacar a qualificação profissional atribuída pelos cursos de Educação e Formação de Adultos (EFA) e pelos Cursos de Especialização Tecnológica (CET), sendo estes o objeto de estudo sobre o qual recaiu a análise dos dados disponibilizados pelo CGP – Citeforma (Citeforma, 2021a).

Assim, os Cursos de Especialização Tecnológica (CET) são um percurso de formação pós-secundário não superior que visa conferir uma qualificação com base em formação técnica especializada de nível 5 do Quadro Nacional de Qualificações. Os CET foram instituídos pela Portaria nº 989/99 de 3 de novembro, reformulados, em 2006, através do Decreto-Lei nº 88/2006, de 23 de maio, alterados novamente pelo Decreto-Lei nº 39/2022, de 31 de maio. Estes cursos estão organizados em três componentes de formação: uma formação geral, científica e tecnológica, orientada para a aquisição e o desenvolvimento de conhecimentos, uma aptidão e atitudes que têm por base a especialização tecnológica e uma formação em contexto de trabalho onde se procura aplicar e consolidar os conhecimentos da formação anterior em contexto de empresa (Centro Qualifica, n. d.).

Por último, temos os cursos de Educação e Formação de Adultos (EFA) que surgem do trabalho conjunto dos Ministérios do Trabalho e Solidariedade e da Educação, através dos Decretos-Lei nº 1083/2000 de 20 de novembro e nº 650/2001 de 20 de julho, como uma das modalidades de formação com um percurso mais flexível e com uma duração mais variável, bem como um público-alvo mais específico. Surgem com o objetivo claro de aumentar as habilitações escolares da população adulta, ao mesmo tempo que procuram a requalificação profissional, não só de adultos com alguma escolaridade, mas também fazer face a necessidades específicas de adultos com baixas ou muito baixas qualificações, iletrados ou com níveis de literacia reduzidos (Centro Qualifica, n.d.).

Esta formação também tem uma estrutura curricular organizada em diferentes percursos formativos, que vão desde o ensino básico, à dupla certificação ou apenas a percursos direcionados ao desenvolvimento de competências profissionais específicas, que, tal como os CET, se divide em três componentes: a formação de base, a formação tecnológica e a formação em contexto de trabalho. Deste modo, os EFA procuram adaptar o processo formativo do candidato às competências adquiridas ao longo da vida (através do processo de Reconhecimento, Validação e Certificação de Competências

(RVCC)), atribuindo no fim do processo de formação um certificado escolar de nível básico, de nível secundário, de certificação profissional ou ambos, conferindo um nível 1, 2, 3 ou 4 de qualificação do Quadro Nacional de Qualificações (Centro Qualifica, n. d.).

Em conclusão, poderá referir-se que os benefícios da formação profissional são variados e transversais, isto é, não ficam centrados apenas no campo económico, quando surgem como oportunidades de emprego e desenvolvimento de carreira, mas também se focam no campo social, pois trabalham variáveis determinantes para a coesão e justiça social, a equidade e a motivação pessoal, entre outros. É por isso expectável que Portugal continue a apostar neste tipo de ensino, quer como instrumento de qualificação e requalificação profissional, quer também como alternativa ao percurso escolar tradicional (Barbosa *et al.*, 2019).

2.2. PROCESSO DE APRENDIZAGEM

É fundamental perceber quais os fatores que levam um formando a ter sucesso no ensino que frequenta, pois só assim será possível realizar uma previsão de desempenho correta e perceber quais são os indicadores que promovem o sucesso escolar (Lynn & Emanuel, 2021). A identificação destes indicadores possibilitam, posteriormente, distinguir um formando bem sucedido, nas suas aprendizagens, de um com menor sucesso, permitindo encaminhá-lo para um investimento nos indicadores que melhorem o seu desempenho. Para além disso, será determinante ter uma definição concreta daquilo que caracteriza o sucesso de um formando, para que se possam reconhecer as eventuais falhas que existam no sistema de ensino, assim como, perceber onde se deverão investir os recursos, de modo a melhorar o rendimento dos formandos (Araújo, 2017).

O termo sucesso académico, cuja definição está aberta a discussão, por ser uma expressão abrangente, é neste estudo definida da seguinte maneira: "sucesso académico que é composto por seis componentes: realização académica, satisfação, aquisição de aptidões e competências desejadas, persistência, concretização dos objetivos de aprendizagem e sucesso na carreira" (York *et al.*, 2015).

Deste modo, os aspetos que demostram ter impacto no sucesso escolar do formando e que deverão ser considerados, são os seguintes: o rendimento académico, a qualidade do estudo, os fatores psicológicos, a taxa de empregabilidade, a integração dos próprios formandos e o sentimento de bem-estar, o envolvimento do formando com o professor na sala de aula, as expetativas em relação ao curso, a existência de apoio social (Araújo, 2017) e, ainda, o estatuto económico, as características demográficas, o género e as experiências profissionais passadas (Lynn & Emanuel, 2021). Um outro indicador muito importante é a satisfação dos formandos com a sua instituição, pois os formandos satisfeitos têm a tendência a integrar-se melhor socialmente e a sentirem uma maior ligação à instituição que frequentam (Araújo, 2017). É também de considerar que o papel que os professores têm na aprendizagem dos formandos é fundamental para os motivar, assim como a organização curricular dos cursos (Vieira & Azevedo, n. d.).

De notar também a importância das notas obtidas nos testes de avaliação, que são talvez o que define mais o sucesso escolar de um formando. Este fator está também relacionado com o empenho e dedicação que o formando deposita nas suas aprendizagens (Gaspar *et al.*, 2020), mas também com o tempo de estudo que dedica a cada uma delas (Pimenta *et al.*, 2018). A motivação e a oportunidade são também aspetos a ter em conta (Gaspar *et al.*, 2020). Características como a compreensão dos

objetivos de aprendizagem, uma boa orientação escolar, um horário escolar flexível e um maior envolvimento por parte dos pais levam a uma maior motivação (Valeriu, 2014).

Em suma, podemos afirmar que o processo de aprendizagem é algo complexo e que são múltiplos os fatores envolvidos, isto é, são muitas as variáveis a ter em conta e não apenas os testes de aptidão, no entanto, nem sempre é possível contemplar todas estas variáveis no processo de seleção do candidato. Deste modo, embora o papel dos testes de aptidão seja relevante, devido ao seu valor preditivo no sucesso que os formandos irão ter nos seus cursos e, efetivamente, ajudarem a decidir se o perfil do formando se enquadra no curso a que se candidata, na realidade, irão apenas aferir e ter em conta uma parte dos fatores que poderão levar ao sucesso na aprendizagem e à consequente conclusão do curso. Assim, não devemos descurar outros fatores, como, por exemplo, os fatores psicossociais, que sendo mais difíceis de determinar em testes de aptidão, também são importantes para a construção de um perfil do candidato.

2.2.1. Avaliação das aprendizagens / Desempenho profissional

Através da avaliação das aprendizagens, será possível serem avaliadas as competências de um indivíduo, sempre que seja necessário, em qualquer área, seja em contexto escolar ou em contexto de trabalho. Mediante esta avaliação, será possível compreender quais foram os conhecimentos e as competências adquiridas ao longo da aprendizagem, assim como perceber as dificuldades que surgiram ao longo da avaliação, o progresso e os resultados obtidos (Alves, 2016).

Esta avaliação de aprendizagens consiste em atribuir um valor a qualquer coisa que esteja sujeita a análise. No que diz respeito ao contexto escolar, normalmente, a progressão de um formando, no ensino, é a consequência dos resultados obtidos, ou seja, o valor que lhe é atribuído nos momentos em que está a ser avaliado (Leitão, 2013). Assim, avaliar as competências de um formando é muito importante, pois permite ao professor perceber se alcançou todos os objetivos propostos (Alves, 2016), mas também permite aos formandos demonstrarem as suas competências e os seus conhecimentos, de forma que os professores consigam obter algum *feedback*, com a finalidade de poderem aperfeiçoar o seu método de ensino e perceberem quais são as matérias que apresentaram dificuldades de resposta, podendo voltar a incidir nelas e a consolidá-las melhor (Oberg, n. d.).

Por este motivo, existem vários métodos de aferição de competências. Se referirmos os métodos usados no ensino, são sugeridas três metodologias diferentes: são elas, as avaliações diagnósticas, que servem para avaliar previamente os conhecimentos dos formandos, imprescindíveis às novas aprendizagens; as avaliações formativas, que acompanham o percurso do formando, com o objetivo de verificar se estão a progredir e de saber o que precisam de melhorar; e, por fim, as avaliações sumativas, que têm como propósito verificar e atribuir uma nota aos conhecimentos dos formandos, normalmente, através de testes (Leitão, 2013; Zeferino & Passeri, 2007).

No entanto, tal como foi referido anteriormente, não existem só estes testes, nem são feitas apenas avaliações de competências no ensino. Também são necessárias em contexto de trabalho e, por isso, existem outros tipos de teste, tais como os testes de aptidão, que auxiliam as empresas e todas as instituições que deles necessitem, tal como é o caso dos centros de formação, no processo de seleção de candidatos mais acertados para as suas vagas, através da avaliação das suas capacidades (Cândido, 2020).

2.3. A PROBLEMÁTICA DA SELEÇÃO DE FORMANDOS

Define-se seleção como analisar quais os candidatos que se adequam mais ao perfil exigido pela vaga a ser preenchida e que apresentam as melhores condições para ingressar nessa mesma posição. Neste processo de seleção dos candidatos, é feita uma comparação entre os vários candidatos, de modo, a que seja percetível qual deles apresenta as qualificações e características necessárias e exigidas para a posição em questão. Esta seriação é um fator muito importante para qualquer instituição que tenha de escolher um candidato para determinada função, uma vez que tem como finalidade encontrar a pessoa certa para a preencher. Esta escolha poderá ter por base alguns métodos próprios que ajudam estes processos, tal como é o caso dos vários testes que existem atualmente, que contribuem, de alguma forma, para prever se certas competências de um candidato são indicadas ou não para o cargo (Cândido, 2020).

2.3.1. Testes de aptidão dos candidatos: que tipos de testes são usados e a sua finalidade

Tal como foi referido anteriormente, existem diferentes tipos de testes que auxiliam no campo da psicologia: são eles, os testes de inteligência, que são utilizados para avaliar/ obter uma estimativa do quociente intelectual, a inteligência como fator geral; os testes de preferência, que são utilizados para perceber os interesses e as preferências de um indivíduo numa certa área; os testes de personalidade, que são efetuados quando é necessário definir os traços de personalidade, as tendências e os hábitos de uma pessoa, em certos aspetos; e, por fim, os testes de aptidão, que servem para avaliar as capacidades humanas específicas. Neste caso, o teste que demonstra uma maior relevância para este estudo é o teste de aptidão, uma vez que uma das finalidades do estudo é tentar perceber se os testes de aptidão são ou não indicados para avaliar as competências de um indivíduo e direcioná-lo para determinado curso (Setiawati, 2020).

Assim, importa definir no que se traduz ter uma aptidão para algo. Significa ter um potencial e a capacidade de adquirir conhecimentos e competências, numa determinada área de aprendizagem, assim como mostrar um bom desempenho nessa mesma área. Para medir a aptidão, é utilizado um teste de aptidão. Este teste é normalmente procurado e utilizado com o intuito de ajudar na tomada de decisões, mas também com o objetivo de obter informações exatas acerca das capacidades específicas de um indivíduo e perceber qual o seu nível de predisposição para aprender e desempenhar bem uma determinada função. Ou seja, estes testes acabam por funcionar como um indicador de sucesso no futuro, pois tentam, de alguma forma, prever o desempenho que o indivíduo irá ter, numa certa atividade (Ballado *et al*., 2014; Raza & Shah, 2011). De realçar, que estes testes têm em conta que diferentes indivíduos têm diferentes níveis de interesse e competências nas diferentes áreas (Mankar & Chavan, 2013).

No caso deste estudo, e uma vez que se utilizam os dados fornecidos pelo Citeforma, os testes de aptidão que são tidos em conta são 11. Estes 11 testes são constituídos pela bateria ABI que é composta por seis provas (ABI 1, ABI 2, ABI 3, ABI 4, ABI 5 e ABI 6) que avaliam as aptidões verbal, numérica, de raciocínio e de atenção, assim como a capacidade para analisar problemas e procurar soluções. Esta bateria de testes foi desenvolvida para avaliar pessoas que queiram trabalhar ou frequentar algum curso na área da informática. Existe também uma outra bateria de testes que tem como foco a avaliação de aptidões mentais primárias. Esta bateria de testes é constituída por cinco provas diferentes: PMA –R, PMA – F, PMA – V, PMA – E e PMA –N, sendo utilizadas pelo Citeforma, nos seus cursos, apenas as últimas três. Para além destas duas baterias de testes, são ainda

considerados, por este Centro de Formação Profissional, outros dois testes, o TIG -I que é uma prova de inteligência e o Teste de Perceção de Diferenças (TPD) que tem como objetivo avaliar as aptidões percetivas e de atenção (Cruz, 2009; Thurstone & Thurstone, 1984; Thurstone & Yela, 1985; Cattell & Cattell, n.d.).

2.4. ANÁLISE PREDITIVA

2.4.1. Conceitos

Cada vez mais, há um interesse e uma necessidade de se ser capaz de prever comportamentos futuros nas diversas áreas, sendo o desempenho dos formandos um deles. Destacam-se as técnicas de *data mining* que analisam conjuntos de dados e extraem informação a partir deles. Para isto ser possível, existem algumas técnicas que conseguem processar a informação contida nos dados e, consequentemente, fazer previsões, neste caso, prever se o formando irá ter sucesso ou não no curso. Assim, referem-se as principais técnicas: *collaborative filtering* (CF), *recommender systems* (RS), *machine learning* (ML) e *artificial neural network* (ANN) (Rastrollo-Guerrero *et al.*, 2020).

Os sistemas de recomendação têm como função recolher informação acerca das preferências dos utilizadores, para depois conseguirem fornecer previsões e recomendações adequadas. Neste caso, este método poderá ser usado para recolher informação dos formandos, como as suas notas nas avaliações ou os seus comportamentos. Um dos algoritmos usados é a filtragem colaborativa que, dependendo do objetivo do seu uso, poderá ser bem-sucedida ou não (Rastrollo-Guerrero *et al.*, 2020).

"*Machine learning* é um conjunto de técnicas que dá aos computadores a capacidade de aprender sem a intervenção da programação humana", ou seja, as máquinas aprendem automaticamente através da análise de dados e da identificação dos seus padrões de comportamento, permitindo a posterior implementação de modelos complexos utilizados para calcular previsões, facilitando a sua utilização e ajudando nas suas tomadas de decisões, com base em dados futuros (Rastrollo-Guerrero *et al.*, 2020; Rodríguez *et al.*, 2022).

Existem duas abordagens que podem ser implementadas, são elas: a aprendizagem supervisionada e a aprendizagem não supervisionada. A aprendizagem supervisionada contém um conjunto de dados em que a classe a que pertencem os dados é previamente conhecida e o objetivo são os modelos construírem padrões e, a partir deles, preverem a classe a que os dados desconhecidos pertencem. Os métodos que pertencem a este grupo são os de classificação, sendo os tipos de algoritmos de classificação mais comuns: *artificial neural network* (ANN), *decision tree* (DT), *support vector machine* (SVM), *linear regression* e a *logistic regression*. Já a aprendizagem não supervisionada possui um conjunto de dados, mas sem que se tenha alguma informação acerca da métrica a alcançar, sendo então o objetivo do algoritmo tentar encontrar as características e os padrões nos dados, sem que se tenha uma informação prévia acerca da classe a que os dados poderão pertencer, ou seja, esta técnica não envolve qualquer supervisão. Os métodos que fazem parte da aprendizagem não supervisionada são o *clustering* e a *association rules* (Osmanbegovic & Suljic, 2012; Rodríguez *et al.*, 2022).

Deste modo, a previsão poderá ser obtida através de várias técnicas, tais como a regressão, a classificação e o *clustering*. Neste caso, a técnica mais utilizada para prever o desempenho dos

formandos, é a aprendizagem supervisionada, uma vez que produz resultados precisos e consistentes e o método mais comum para resolver este tipo de problemas é a classificação. A aprendizagem não supervisionada não costuma ser muito utilizada neste tema, pois apresenta uma baixa precisão na previsão do desempenho dos formandos (Alsariera *et al.*, 2022).

Para se poder conseguir prever o desempenho académico dos formandos, de maneira a tentar perceber se irá ou não ter sucesso no curso que está a frequentar, tem-se recorrido normalmente a técnicas de *machine learning*, usadas para fazer previsões acerca dos dados em análise, seguindo uma metodologia baseada no CRISP-DM (que irá ser mais detalhada no Capítulo 3), constituída por seis fases, elaboradas nos pontos seguintes.

2.4.2. Metodologia CRISP-DM

2.4.2.1. Recolha de dados

Para estudos nesta área, obtêm-se, normalmente, os dados através de bases de dados de diferentes universidades (Ahmed *et al.*, 2021) ou de qualquer instituição relacionada com a componente de ensino. Recorre-se também a sistemas de gestão de aprendizagem (Maghawry *et al.*, 2022; Yacoub *et al.*, 2022), podendo a recolha de dados também ser feita a partir de questionários feitos aos próprios formandos de vários estabelecimentos de ensino (Osmanbegovic & Suljic, 2012).

2.4.2.2. Pré-processamento dos dados

De modo a tratar os dados, é feito um seu pré-processamento. Este passo é essencial, uma vez que transforma os dados originais em dados adequados e preparados para serem aplicados nos modelos. Esta etapa inclui a limpeza dos dados e a sua transformação, ou seja, a remoção de valores duplicados e o tratamento dos valores em falta, podendo este passo ser dado através da remoção desses valores ou do seu preenchimento, apresentando-se como a melhor solução simplesmente descartá-los. Para além disso, quando for necessário, poderá ter de ser feita uma conversão dos dados nominais para numéricos, sendo o método de discretização não supervisionada de divisão automática equidistante o mais comum para o fazer e também uma reformulação dos dados para estarem dentro de um determinado intervalo, usando o método de normalização dos dados, como o *min-max*. Quando os dados não estão balanceados, devemos usar um dos seguintes três métodos: *oversampling*, *undersampling* ou *hybrid methods*, para solucionar o problema. Para detetar *outliers*, poderá ser utilizado o método de *clustering DBSCAN* (Maghawry *et al.*, 2022; Xiao *et al.*, 2021; Yacoub *et al.*, 2022).

2.4.2.3. Seleção de variáveis e divisão de amostras

A etapa de seleção de variáveis é muito importante, pois é a partir dela que se selecionam as variáveis mais relevantes e influentes no conjunto de dados, mas também as que mais contribuem para a variável-alvo. Para fazer esta seleção, uma das técnicas que poderá ser utilizada é a *SelectKBest* que seleciona as primeiras características k com a maior pontuação baseada no teste Qui-Quadrado, para comparar os resultados reais e previstos, como uma função de pontuação (Maghawry *et al.*, 2022).

Algumas das variáveis que costumam ser, na maior parte das vezes, selecionadas e que apresentam maior impacto na previsão do desempenho de um formando são as dos dados pessoais, tais como a idade, o estatuto dos pais, a satisfação com o curso, o rendimento, a nacionalidade e o género do

formando, sendo esta última muito significante, pois formandos do género masculino e feminino apresentam estilos diferentes nas suas aprendizagens, tendo, os do género masculino, um estilo mais otimista (Alsariera *et al.*, 2022).

Após todo o processo de tratamento de dados e antes de partir para a aplicação dos algoritmos, será então necessário dividir os dados, de forma a criar dois conjuntos, um de teste e outro de treino. Este processo poderá ser feito, normalmente, através do método de *n-fold cross validation* (Xiao *et al.*, 2021).

2.4.2.4. Algoritmos de previsão

Assim, de entre todos os algoritmos de classificação que existem, os que costumam ser mais aplicados neste estudo são: *artificial neural network* (ANN), *decision tree* (DT), *random forest* (RF), *adaBoost* (AB), *support vector machine* (SVM), *k-nearest neighbor* (KNN), *naive bayes* (NB), *multilayer perceptron* (MLP) e *linear regression* (LinR).

Normalmente, as *decision tree* (DT) são utilizadas frequentemente, pois são bastante simples e fáceis de implementar e, o mais importante, fáceis de interpretar. Para além disso, são bastante fáceis de compreender, pois são construídas sobre as regras *if-then* (Alsariera *et al.*, 2022; Osmanbegovic & Suljic, 2012). Já o modelo *support sector machine* é usado tanto para os problemas de classificação como de regressão. "O SVM é construído sobre o conceito de construção de um hiperplano que divide de forma ótima o conjunto de dados em dois grupos" (Ahmed *et al.*, 2021). A regressão linear "é o melhor modelo de previsão para testar a causa do efeito de uma variável dependente sobre uma ou mais variáveis independentes. Além disso, a abordagem de regressão linear é bastante fácil e rápida de processamento para conjuntos de dados de grandes dimensões" (IEEE Staff, 2017). O modelo *artificial neural network* (ANN) "é constituído por um conjunto de entidades altamente interligadas, chamadas Elementos de Processamento. Esta estrutura e função da rede é inspirada no sistema nervoso central biológico, particularmente, no cérebro. Cada Elemento de Processamento é concebido para imitar a sua contraparte biológica, o neurónio, que aceita um conjunto ponderado de entradas e responde com a saída correspondente" (Rastrollo-Guerrero *et al.*, 2020). Este algoritmo, apesar de apresentar muito bons resultados nas suas previsões, tem a desvantagem de ser muito difícil de interpretar e de compreender (Osmanbegovic & Suljic, 2012).

Os algoritmos que apresentam os melhores resultados para prever o sucesso dos formandos são: *decision tree* (DT) (Lynn & Emanuel, 2021), *support vector machine*, *artificial neural network* (Alsariera *et al.*, 2022) e *multilayer perceptron* (Ramesh *et al.*, 2013; Widyahastuti & Tjhin, 2017), pois forneceram as previsões mais fiáveis. Para melhorar ainda mais o desempenho dos modelos, poderão ainda utilizar-se *ensemble methods*, tais como o *bagging* e o *boosting*, em conjunto com os modelos acima referidos (Amrieh *et al.*, 2016).

2.4.2.5. Apreciação de resultados

Os indicadores de desempenho mais comuns, para avaliar os resultados dos modelos de previsão, são: *accuracy*, *precision*, *recall* e *F-measure*. Mas também existem outras métricas, tais como o coeficiente de correlação que descreve o grau de ligação entre o valor real e o valor previsto. O intervalo do coeficiente de correlação situa-se entre -1 e 1; assim, se o valor do coeficiente for 0, significa que não existe correlação e caso seja próximo de 1, significa que existe uma relação positiva. Existe também o

erro médio absoluto (MAE) que é definido como a quantidade utilizada para medir quão próximas são as previsões e o erro quadrático médio (RMSE) que é utilizado para medir as diferenças entre os valores previstos por um modelo e os valores reais observados. (Maghawry *et al.*, 2022; Widyahastuti & Tjhin, 2017; Xiao *et al.*, 2021).

2.5. SÍNTESE

Deste modo, concluímos que a formação profissional é algo que já existe há muitos anos e que tem sofrido bastantes alterações, do seu começo à atualidade, com o objetivo de conseguir acompanhar as mudanças que se verificam na sociedade e no mercado de trabalho. Assim, este tipo de formação torna-se relevante, uma vez que permite aos formandos poderem especializar-se em novas áreas de trabalho, mas também atualizar-se, de forma a melhorarem as suas competências, indo ao encontro das novas tendências de mercado. Consequentemente, e com o aumento da procura de candidatos para este tipo de formação, vê-se a necessidade de otimizar o processo de seleção de candidatos para frequentarem estes cursos, quer isto dizer, que é importante que essa escolha seja a mais acertada, para que, de alguma forma, haja a garantia de que quem a for frequentar irá tirar o melhor proveito e concluí-la com sucesso. Para isso, poderá recorrer-se a técnicas de *machine learning* que, atráves de modelos de classificação, incluídos na aprendizagem supervisionada, ajudam a prever o desempenho académico dos formandos, sendo este estudo vantajoso, pois seria assim possível ter informações prévias acerca de saber se os candidatos conseguiriam, ou não, concluir a formação, para a qual se inscreveram, com sucesso.

Com base nos testes de aptidão, é comum, nos estudos que os utilizam, prever comportamentos futuros, em que o algoritmo *multiple linear regression* é o que obtém o melhor resultado. Assim, um teste de aptidão ajuda a medir as capacidades e o desempenho de uma pessoa em diversos programas de ensino, sendo os testes verbais e numéricos os que apresentam uma maior influência na previsão do sucesso académico (Setiawati, 2020).

3. METODOLOGIA

O presente estudo seguiu uma metodologia formada por três fases, em que, cada uma delas, foi constituída por diferentes etapas, tal como pode ser observado na Figura 3.1. Assim, a primeira fase incidiu na parte da investigação, a segunda fase consistiu na análise dos dados recolhidos e a última fase foi dedicada à conclusão.

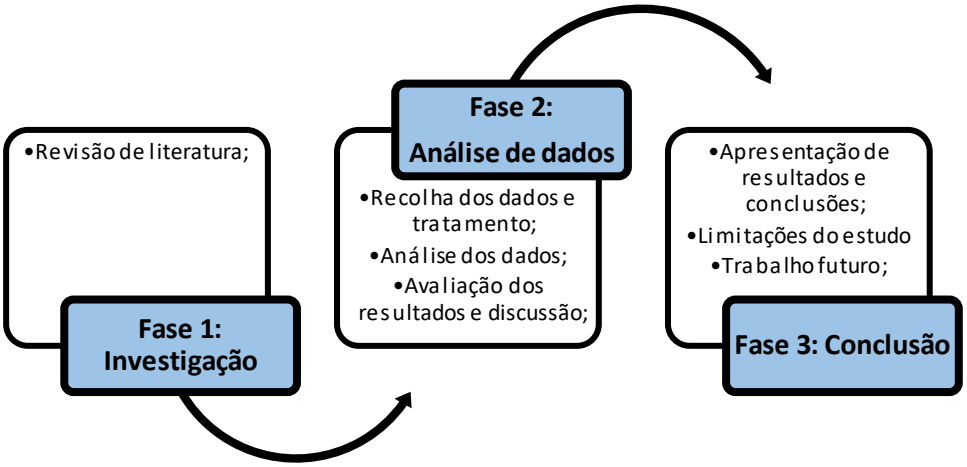


Figura 3.1 – Esquema da metodologia a aplicar

Posto isto, na fase de investigação, realizou-se a revisão de literatura. Esta etapa teve como objetivo poder identificar trabalhos anteriores que sejam similares ao tema a ser estudado ou trabalhos que contenham conteúdos necessários e úteis, contribuindo para um melhor desenvolvimento do estudo. Para além disso, esta pesquisa foi também essencial para perceber as vantagens e desvantagens dos diferentes modelos que existem, com o intuito de perceber quais serão os mais adequados aos objetivos pretendidos e ser possível aplicar as técnicas de *machine learning*.

A segunda fase da metodologia diz respeito à análise de dados, em que foram recolhidos e posteriormente tratados, analisados e interpretados os resultados obtidos. De referir que na primeira etapa desta fase, os dados foram fornecidos pelo Centro de Formação Profissional – Citeforma, consistindo, de uma forma geral, em resultados de testes de aptidão realizados por diferentes candidatos. Relativamente à etapa tratamento de dados, esta foi realizada, tendo em conta a metodologia *CRISP-DM*. Esta metodologia, tal como mostra a Figura 3.2, divide o ciclo de vida de um projeto de *data mining* em seis fases e tem como intuito ajudar na sua orientação (Watson *et al.*, 2000).

A primeira fase diz respeito ao *business understanding* que tem como intuito determinar os objetivos do trabalho, ficando-se a conhecer melhor o projeto a realizar, assim como, a delinear um plano para si. A segunda etapa, *data understanding*, em que são recolhidos os dados, tem como objetivo a sua compreensão, bem como fazer a descrição dos dados, de modo a se ficar a conhecer melhor as suas características e a perceber o que poderá ser feito. O terceiro ponto é relativo à preparação dos dados, em que se irão preparar os dados, para depois serem aplicados aos modelos. Nesta etapa, está então

incluída a limpeza dos dados, a sua transformação e a seleção de variáveis. Em seguida, é a fase em que se aplicam os modelos; aqui, são escolhidos os algoritmos a usar, assim como os parâmetros. Em penúltimo lugar, é feita a avaliação dos resultados dos modelos, em que são revistas as etapas anteriores, de modo a verificar se está tudo correto e, por fim, escolher o modelo que mais se adequa aos objetivos previstos e que apresenta melhores resultados. Por fim, a última etapa é a implementação, em que é aplicado o resultado da análise. Este processo consiste em elaborar um relatório e uma apresentação com os resultados obtidos, para dar resposta ao problema proposto, mas também criar um documento com toda a informação necessária, de modo a perceber o que foi feito (Chapman *et al.*, 2000)

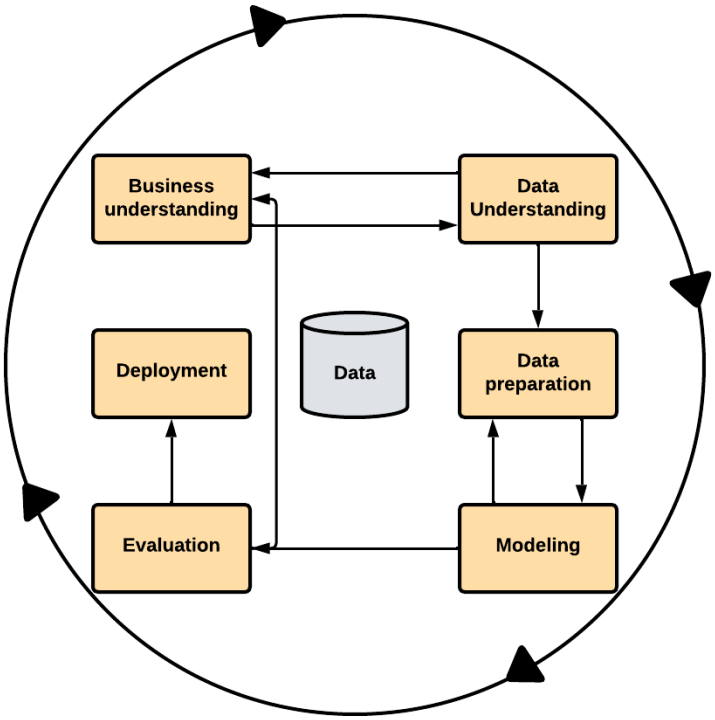


Figura 3.2 – Fases do modelo de referência CRISP-DM (Watson *et al.*, 2000)

Na última fase, a conclusão, foram mencionadas as conclusões obtidas pelo estudo, em que serão apresentadas as respostas às questões de investigação, propostas anteriormente, assim como as limitações que foram identificadas e algumas sugestões para trabalhos futuros.

4. Modelo preditivo de aptidão dos candidatos a formação

Para se proceder à previsão da aptidão que os formandos irão ter na sua formação foi necessário passar por diferentes etapas para chegar ao resultado final. Para tal, neste estudo, foi usada a metodologia CRISP-DM, já referida anteriormente, no Capítulo 3. Assim, de acordo com esta estrutura, a primeira fase começa por identificar o *business understanding*, com o intuito de identificar os objetivos do estudo, de seguida, inicia-se o processo de *data mining*, composto por diferentes passos necessários à exploração e ao tratamento dos dados. Uma vez os dados preparados, são testados por vários algoritmos de *machine learning*, com a finalidade de prever o resultado pretendido. Por fim, são avaliados os resultados obtidos.

4.1. ANÁLISE DO NEGÓCIO

O principal objetivo deste estudo é tentar ajudar os Centros de Formação Profissional a perceber o impacto e a necessidade dos testes de aptidão, bem como, se um determinado formando irá ter sucesso ou não no curso que escolheu frequentar, com o intuito de canalizar as pessoas com o perfil mais adequado para cada formação e, consequentemente, reduzir o número de formandos reprovados no curso, por falta de compatibilidade. Assim, ao identificar que um formando não irá ter sucesso num determinado curso, a equipa que seleciona os candidatos poderá fazer uma seleção mais acertada e preencher as vagas disponíveis, para cada curso, com os candidatos que irão ter um melhor aproveitamento no curso.

Neste caso, o Centro de Formação Profissional – Citeforma, utilizado para este estudo, tem uma oferta formativa constituída por 19 cursos diferentes, que dizem respeito apenas a formação em cursos EFA e CET. Os cursos de técnico especialista em desenvolvimento de produtos multimédia (CET_Multimédia), técnico especialista de gestão da qualidade, ambiente e segurança (CET_Qualidade) e técnico especialista de gestão de turismo (CET_Turismo) parecem apresentar mais formandos reprovados, com uma taxa de reprovação média de sete formandos por turma.

Do ponto de vista das técnicas de *machine learning*, o principal objetivo é aplicar os modelos de classificação para tentar prever o desempenho de um formando, num curso profissional, tendo em conta os testes de aptidão e outras componentes, tais como as entrevistas que os formandos fazem e as suas habilitações literárias, tidas em conta no processo de seleção dos candidatos.

4.2. ANÁLISE DOS DADOS

4.2.1. Descrição dos dados

O conjunto de dados em análise apresenta uma coleção de 1331 registos de formandos, sendo estes registos constituídos pela informação de diferentes candidatos a estes cursos e um conjunto de 30 propriedades diferentes. Os dados apresentam informação de 19 cursos diferentes, iniciados e concluídos no período entre 2017 e 2022, tendo os cursos em questão, em média, a duração de um ano.

Os dados são constituídos não só por características pessoais dos candidatos, como a idade, o género, as habilitações literárias e a nacionalidade, mas também por dados como que curso escolheram para

frequentar e que nota obtiveram no final. Para além disso, os dados são constituídos pelos diferentes conjuntos de provas existentes no processo de seleção, realizados por todos os formandos. Estas provas são constituídas por diversos testes de aptidão; dos 11 testes existentes, o formando apenas realiza o conjunto de testes exigido para o curso a que se está a candidatar. Complementarmente, existem ainda duas entrevistas, uma realizada em conjunto com os vários candidatos ao mesmo curso e uma apenas com o coordenador responsável pelo processo de seleção. Adicionalmente, existe também uma prova de conhecimentos, no entanto, esta prova não condiciona o processo de seleção, pois nem sempre é tida em conta.

Uma vez apresentada a constituição do conjunto de dados e a fim de proporcionar uma melhor compreensão das variáveis que compõem o conjunto de registos e a sua interpretação, apresenta-se a seguir uma tabela onde se descrevem as variáveis.

Tabela 4.1 – Descrição de variáveis

Variáveis	**Descrição**
Formando	Número atribuído ao formando.
Idade	Idade do formando.
Género	Género do formando.
Curso	Diferentes cursos que existem no Citeforma.
Ano de início	Ano de início de cada curso.
Ano de fim	Ano de conclusão de cada curso.
Habilitações literárias	Habilitações literárias dos formandos.
TIG-I	Prova de inteligência que proporciona uma avaliação geral do fator "g". Pertence a um conjunto de provas que se baseiam em "dominós". Este teste tem como objetivo avaliar a capacidade dos sujeitos para conceptualizar e aplicar o raciocínio sistemático a certos problemas.
TPD	Teste de perceção de diferenças. Permite avaliar as aptidões necessárias para perceber, rápida e corretamente, semelhanças e diferenças em modelos estimulantes, espacialmente ordenados.
PMA-V	Teste de compreensão verbal. Testa a capacidade para captar ideias expressas através da linguagem, tanto em forma escrita como oral.
PMA-N	Este teste serve para avaliar o cálculo numérico e tem como objetivo perceber a capacidade do individuo para resolver, rápida e acertadamente, problemas quantitativos simples.
PMA-E	Teste de conceção espacial. Testa a capacidade para imaginar e conceber estruturas espaciais e compará-las entre si.
ABI-1	Prova de compreensão verbal do tipo "sinónimos".
ABI-2	Teste que avalia a capacidade para manipular símbolos matemáticos e para resolver problemas numéricos apresentados de forma parcialmente verbal.
ABI-3	Teste que avalia a atenção concentrada, mediante uma tarefa de deteção de erros num contexto que tenta aproximar-se das listagens de programas.
ABI-4	Prova de raciocínio lógico, em que os sujeitos deverão encontrar, entre as soluções possíveis, o número que continuaria a série proposta.

Variáveis	Descrição
ABI-5	Prova que mede um aspeto específico da atenção: a capacidade para localizar elementos que estão misturados com outros e assinalar o código correspondente, tendo em conta determinadas características.
ABI-6	Avalia a capacidade para analisar um problema e para organizar soluções numa série de etapas lógicas.
Habilitações	É atribuído ao formando uma pontuação, consoante a sua habilitação literária. É atribuída a pontuação 1 aos formandos que têm mestrado; 2, aos formandos que tenham licenciatura e 3 aos que apenas tenham o secundário concluído.[1]
Qualificação profissional	É atribuído ao formando uma pontuação consoante a sua qualificação profissional. É atribuída a pontuação 1 aos formandos sem qualificação na AEF; a pontuação 2 aos com qualificação de nível IV em área afim e pontuação 3 aos com qualificação de nível IV na AEF.
Resultados testes	É atribuído ao formando uma pontuação consoante o valor que obteve nos testes psicométricos. É atribuída a pontuação 1 aos formandos em que a média do percentil do resultado dos testes se localiza no 1º tercil; a pontuação 2 aos formandos que se localizam no 2º tercil e pontuação 3 aos formandos que se situam no 3º tercil do grupo de selecionados.
Resultado Prova de conhecimentos	É atribuído ao formando uma pontuação consoante o valor que obteve na prova de conhecimentos. É atribuída a pontuação 1 aos formandos que obtiveram um resultado entre 50% e 64%; a pontuação 2 aos resultados entre 65% e 80% e a pontuação 3 aos resultados superiores a 81%.
Entrevista grupo pequeno	É atribuído ao formando uma pontuação consoante o desempenho que teve na entrevista de grupo pequeno. É atribuída a pontuação 0 aos formandos que tiveram um desempenho desfavorável; a pontuação 2 aos formandos que apresentaram um desempenho favorável com reservas e a pontuação 5 a um formando com um desempenho favorável.
Entrevista coordenador	É atribuído ao formando uma pontuação consoante o desempenho que teve na entrevista com o coordenador. Assim, o coordenador poderá optar por atribuir uma bonificação de até 3 pontos ou atribuir uma penalização de até 3 pontos aos formandos.
Classificação final	Soma das colunas de pontuação consideradas para a classificação final obtida pelo formando no processo de seleção. As colunas são as seguintes: habilitações, qualificação profissional, resultados dos testes, resultado da prova de conhecimentos, entrevista de grupo pequeno e entrevista com o coordenador. O resultado da prova de conhecimentos e a qualificação profissional nem sempre são considerados na classificação final, dependendo do curso.
Prova dos conhecimentos	Resultado obtido, pelo formando, na prova de conhecimentos.
Código	Código atribuído a cada curso
Nota do candidato	Nota obtida pelo formando no final do curso. Ou seja, se foi aprovado, reprovado, se desistiu do curso, se desistiu na primeira semana experimental do curso, se não foi chamado para frequentar o curso ou se o curso não se realizou ou transitou para outro ano.

[1] Nota: Nesta componente, a pontuação mais pequena é atribuída aos formandos com maior nível de qualificação, uma vez que o que se pretende é beneficiar os formandos que tenham menos qualificações para frequentar estes cursos.

Variáveis	Descrição
Nacionalidade	Nacionalidade do formando.
Estado	Classifica se o formando foi selecionado para frequentar o curso, se ficou como suplente, se foi eliminado do processo de seleção ou se o próprio formando desistiu.

4.2.2. Exploração dos dados

Depois de apresentadas todas as variáveis e com o objetivo de se poder fazer um melhor tratamento dos dados, foi necessário compreender melhor os dados, as variáveis disponíveis e as características dos formandos que se candidatam, para frequentar um destes cursos e assim realizou-se a seguinte análise dos dados:

- **Número de formandos por faixa etária:** após a representação gráfica de quantos formandos existem por grupo etário, observou-se que o grupo etário com mais candidatos é o dos jovens adultos, seguido do dos adultos e dos de meia-idade e, por fim, dos candidatos mais velhos.
- **Número de formandos por género:** mais de metade dos candidatos são do sexo feminino (739 pessoas) enquanto os restantes (592 pessoas) são do sexo masculino.
- **Número de formandos por género que existem por faixa etária:** concluiu-se que a maior parte dos candidatos de sexo feminino pertencem ao grupo etário dos adultos, seguido do dos jovens adultos, dos de meia-idade e, por fim, dos candidatos mais velhos, enquanto no sexo masculino o grupo etário com mais presença é o dos jovens adultos, seguido do dos adultos, dos de meia-idade e, por fim, dos candidatos mais velhos.
- **Número de formandos por curso:** o curso que apresenta maior procura é o CET_Contabilidade, seguido do CET_Redes e o curso com menor procura é o EFA_Rececionista.
- **Número de formandos que iniciaram o curso em cada ano:** o ano com mais procura pelos candidatos foi o de 2020, seguido do ano de 2019 e o ano com menos candidatos foi o de 2021.
- **Número de formandos por habilitações académicas:** candidatos com apenas o secundário parecem ser os mais frequentes.
- **Número de formandos por classificação:** a maior parte dos formandos apresenta uma classificação final de "aprovado". No entanto, se formos comprar a percentagem do número de formandos que apresenta uma classificação final de "desistentes" conclui-se que existe um maior número de formandos neste categoria do que na categoria da classificação final de "reprovados".
- **Classificação dos formandos por faixa etária:** os candidatos aprovados costumam estar mais presentes no grupo etário dos jovens adultos. Em relação aos candidatos reprovados, são também os jovens que aparecem em maior quantidade.
- **Classificação dos formandos por género:** os candidatos do sexo feminino costumam pertencer em maior número aos aprovados, enquanto os candidatos do sexo masculino têm um maior peso nos reprovados.
- **Classificação dos formandos por habilitações académicas:** os candidatos com a habilitação académica do secundário são sempre os que aparecem com maior frequência em todas as diferentes notas finais.

- **Classificação dos formandos por curso:** é notável que, na maior parte dos cursos, a classificação final mais frequente é a aprovação do formando.
- **Género dos formandos por curso:** os cursos de multimédia e programação apresentam um maior número de candidatos do sexo masculino, enquanto os outros cursos, tais como o de contabilidade, o administrativo ou o de rececionista parecem apresentar um maior número de candidatos do sexo feminino.

É de ter em conta que, depois de concluída a fase de tratamento de dados, esta análise é novamente feita, no ponto 4.3.2, com o intuito de observar as alterações que possam existir nas variáveis. Esta análise irá ser depois complementada com o auxílio das respetivas representações gráficas.

4.3. PREPARAÇÃO DOS DADOS

Neste ponto, são abordadas todas as etapas realizadas para preparar os dados que irão ser utilizados posteriormente nos modelos de *machine learning*. Estas etapas consistiram na remoção de valores duplicados, no tratamento de dados em falta, na verificação se os dados estão todos coerentes, na análise se existiam *outliers* (valores que se diferenciam de todos os outros) e na transformação de variáveis, de acordo com as necessidades.

4.3.1. Limpeza dos dados

4.3.1.1. Tratamento dos valores duplicados e em falta

Numa primeira análise para este tópico, verificou-se se existiam valores duplicados e em falta no conjunto de dados. Em relação aos valores duplicados, não foi detetado nenhum caso. No que diz respeito aos valores em falta, foram identificadas várias variáveis nessa situação, no entanto, nem em todas as variáveis será necessário tratar os valores em falta, como, por exemplo, nas variáveis dos testes de aptidão (TIG-I, TPD, PMA-V, PMA-N, PMA-E, ABI-1, ABI-2, ABI-3, ABI-4, ABI-5 e ABI-6), pois, tal como foi referido anteriormente, os formandos não realizaram todas as provas existentes, mas apenas aquelas que correspondem ao curso a que se candidataram e, por isso, para cada formando, nem todas as colunas dos testes estão preenchidas.

Contudo, existem variáveis em que foi necessário tratar os valores em falta, no entanto, cada uma dessas variáveis tinham as suas particularidades e exigiam um método diferente de procedimento. Posto isto, algumas das variáveis com valores em falta não foram tratadas neste ponto, mas sim no ponto 4.3.1.3, pois era onde melhor se enquadrava, uma vez que, para resolver esses problemas, teriam de se proceder a alterações nessas variáveis.

Em relação às variáveis que foram tratadas nesta fase do estudo, tais como a "idade", o "código" e as "habilitações literárias", não exigiram a eliminação de quaisquer tipos de registos e, por isso, foram preenchidos através da média ou moda, conforme se pode observar com mais detalhe de seguida.

- A variável "idade" apresentava seis valores em falta e, sendo poucos dados, estes foram substituídos pela média da coluna das idades.
- A variável "código" tinha 19 valores em falta que foram substituídos por 0, uma vez que esse curso sofreu uma alteração na data de conclusão e transitou para o ano de 2023, por isso não lhe tendo sido atribuído um código.

- Por fim, na variável "habilitações literárias", o objetivo era substituir os valores em falta pela habilitação literária mais frequente nessa coluna, no entanto, para isso, foi necessário restruturar essa coluna, uma vez que que apresentava várias formas escritas diferentes das palavras secundário, licenciatura e mestrado. Deste modo, o primeiro passo a realizar foi uniformizar essa coluna, criando uma função que substituísse as diferentes formas de cada palavra para uma única. Depois desse passo concluído, foi possível perceber qual das habilitações literárias era mais frequente, neste caso, o secundário e, em seguida, substituir então os valores em falta dessa coluna por essa habilitação mais frequente.

4.3.1.2. Verificação da coerência dos dados

Resolvida a questão dos valores duplicados e em falta, fomos verificar se existiam incoerências nos dados. Averiguou-se se existiam idades acima dos 110 anos, o que não se verificou, foi também averiguado se alguns dos 11 testes de aptidão apresentava pontuações acima do valor máximo de pontuação, neste caso, 100, também não tendo havido ocorrências. Nas seguintes cinco incoerências analisadas, averiguou-se se as colunas "habilitações", "qualificação profissional", "resultado dos testes", "entrevista grupo pequeno" e "entrevista coordenador" continham apenas as pontuações possíveis ou se tinham pontuações diferentes do expectável, tendo todas as colunas apresentado incoerências: 101, 625, 132, 87 e 83 valores, respetivamente. Estes valores foram posteriormente tratados na secção 4.3.1.4. Por fim, foram eliminadas os registos que não tinham nenhum dos campos dos testes de aptidão preenchidos e que deveriam estar.

4.3.1.3. Tratamento de *outliers*

Neste tópico, foi verificada a existência de *outliers* para as variáveis numéricas, através do uso de vários histogramas e *box plots*, de cada uma das variáveis em questão. No entanto, após a observação dos gráficos, foi detetada a presença de alguns *outliers*, contudo, a maioria estava presente nas variáveis dos testes de aptidão e esses valores menos usuais foram considerados importantes para análises posteriores, tendo sido tomada a decisão de não retirar nenhum dos *outliers* existentes.

4.3.1.4. Transformação de variáveis

Nesta etapa, foram tratadas as restantes variáveis com valores em falta, pois tal como foi referido anteriormente, essas variáveis precisavam de ser tratadas de maneira diferente, uma vez que se tinha acesso a informação acerca do modo de preencher esses campos corretamente, sem simplesmente eliminar esses registos ou substituí-los pela média, moda ou mediana da coluna. Desta maneira, foram feitas alterações às variáveis, de modo a torná-las apropriadas para as análises futuras. Deste modo, os parâmetros considerados para estas alterações são os seguintes: "habilitações", "qualificação profissional", "resultados testes", "entrevista grupo pequeno" e "entrevista coordenador".

No que respeita à variável "habilitações", para resolvermos o problema dos 101 valores em falta nessa coluna, detetado tanto quando se verificaram os valores em faltas das variáveis, mas também na etapa das incoerências, usámos os valores da coluna "habilitações literárias" para fazer correspondências com os valores da coluna "habilitações", uma vez que os valores das pontuações atribuídas de 1, 2 ou 3, na coluna habilitações, tinham como base os valores da coluna habilitações literárias, tal como foi explicado na Tabela 4.1. Ou seja, os valores da coluna "habilitações literárias" com o nome secundário ou outras qualificações correspondiam a uma pontuação de 3 na coluna "habilitações"; o nome

licenciatura da coluna "habilitações literárias" correspondia a 2 na coluna "habilitações" e o nome mestrado da coluna "habilitações literárias" correspondia a 1 na coluna "habilitações", por fim, os valores que não tinham correspondência retornavam ao valor 0. Essas novas alterações ficaram guardadas numa nova coluna, com o nome "new_habilitações".

Em seguida, procedeu-se ao tratamento da questão relativa aos 625 valores em falta na coluna "qualificação profissional". Estes valores apresentavam os campos por preencher propositadamente, visto que apenas os cursos classificados como CET tinham de completar este critério, tal como foi explicado na Tabela 4.1. No entanto, e para que na etapa da previsão dos modelos não tivesse de se fazer uma separação de base de dados em cursos CET e cursos EFA, que depois, por consequência, resultaria em duas bases de dados muito pequenas, pois os dados não existem em grande volume, decidiu-se preencher a restante coluna. Assim, assumiu-se que este campo também era preenchido pelos cursos EFA e, por isso, foi atribuído o valor mais frequente desta coluna, neste caso, o valor 1, aos valores em falta dessa coluna, resolvendo assim esta questão.

Também a coluna "resultados testes" apresentava alguns valores em falta que foram então preenchidos pela fórmula correta que é usada para atribuir a pontuação a esta variável. Assim, a fórmula usada para completar a coluna baseia-se nas médias das notas que os candidatos tiveram nos testes de aptidão, para um determinado curso, posteriormente, essas médias são ordenadas por ordem decrescente e esse grupo de médias é então dividido em três partes, aplicando-se, neste caso, a regra dos tercis. Uma vez os grupos divididos, o primeiro tercil representa os formandos que tiveram as médias mais altas nas provas e recebe uma pontuação de três pontos, no segundo tercil, os formandos recebem uma pontuação de 2 e no último tercil, em que se encontram o grupo de formandos com as médias mais baixas, é atribuída uma pontuação de 1. Depois de aplicada a fórmula aos valores em falta, e de toda a coluna estar bem preenchida, os valores foram guardados numa nova coluna com o nome "resultados_testes".

Em seguida e também tal como foi verificado anteriormente, a coluna "entrevista grupo pequeno" continha algumas incoerências, uma delas onde constavam quatro valores "0/3" que depois foram substituídos por 0. Também a coluna "entrevista coordenador" apresentava alguns valores não esperados, pois só deviam existir pontuações com os valores 1, 2 e 3 e havia pontuações com valores decimais, tais como 0,5, 1,5 ou 2,5, e mesmo palavras, como em alguns destes exemplos: "não compareceu" ou "faltou". Desta maneira, foram arredondados os valores decimais por excesso e atribuída a pontuação de 0 para substituir essas palavras, uma vez que os conteúdos das palavras indicavam que o formando nunca teve nenhuma pontuação atribuída nesta componente por falta de comparência. Para além destas incoerências, estas colunas também apresentavam alguns valores em falta, no entanto, estes não foram substituídos pelos métodos tradicionalmente usados, uma vez que não fazia sentido atribuir um valor com base na média, moda ou mediana a uma variável em que os valores são atribuídos por uma pessoa a outra pessoa, tendo em conta o seu desempenho na entrevista. Contudo, estes registos também não foram eliminados nesta etapa, uma vez que, para as análises posteriores, existiam diferentes bases de dados que englobavam um conjunto de variáveis diferentes. Posto isto, a base de dados que irá conter as colunas dos testes de aptidão, mas não as colunas das pontuações[2] (onde estas duas colunas, "entrevista grupo pequeno" e "entrevista

[2] Na "coluna das pontuações" são consideradas as seguintes variáveis: habilitações, qualificação profissional, resultados testes, resultado prova conhecimentos, entrevista grupo pequeno e entrevista coordenador.

coordenador", estão inseridas), poderá incluir essas linhas que têm dados para as restantes colunas da base de dados, contudo, na base de dados que utiliza estas colunas, estas linhas serão depois eliminadas para não conterem os valores em falta e não afetarem a análise.

Ainda nesta secção, mas sem estar diretamente relacionado com as questões a resolver acerca dos valores em falta e as incoerências detetadas, verificou-se que não faria sentido manter alguns dos valores presentes na coluna "nota do candidato". Deste modo, aferiu-se que o Curso CET_Qualidade, iniciado em 2021, não ficou concluído no ano de conclusão previsto, 2022, tendo sofrido uma transição para o ano de 2023, aparecendo como resultado final na coluna "nota do candidato", "transitou para 2023". Assim, uma vez que o objetivo do estudo é tentar prever a nota do candidato no final do curso e considerando que estes formandos ainda não tinham essa informação, não fazia sentido considerar estes dados para a análise e, por isso, foram eliminadas estas 19 linhas da base de dados. Na mesma ordem de ideias, o curso EFA_Logística, com início no ano de 2018, e o curso EFA_Vitrinismo, com início no ano de 2019, não se chegaram a realizar e, por isso, foram também eliminadas as linhas em que na coluna "nota do candidato" constava "não se realizou". Por último, nesta coluna existiam ainda formandos classificados como "não foi chamado", o que também não fazia sentido permanecer associado a estes formandos, pois não chegaram a frequentar nenhum curso, tendo assim sido eliminadas as 137 linhas com esta designação.

Posteriormente a estas alterações, trabalhou-se a coluna "género" que também foi adaptada, com o objetivo de transformar a coluna de categórica em numérica, pois, numa análise posterior, os algoritmos de *machine learning* necessitam normalmente de variáveis numéricas. Assim, esta variável foi alterada de "M" para "0" e de "F" para "1". De forma semelhante, as colunas "nota do candidato" e "estado" foram transformadas, de modo a que, partindo dessas variáveis, originassem duas colunas com as mesmas características, mas agora numéricas. Assim, criou-se a coluna "nota_final", onde foi feita a correspondência de "aprovado" para "1", "reprovado" para "2", "desistente" para "3" e "desistiu no processo de seleção" para "4" e, por fim, a variável "estado" de "selecionado" para "1", "suplente" para "2" e "não selecionado" para "3".

Para finalizar o tópico da transformação de variáveis, as colunas "habilitações" e "resultados testes" são eliminadas, pois a informação correta está agora nas novas colunas criadas, "new_habilitações" e "resultados_testes". Também a coluna "nacionalidade" foi eliminada, por falta da maior parte dos dados, pois de 1331 dados da base só 165 têm a informação preenchida.

Igualmente, as colunas "prova de conhecimentos" e "resultado prova conhecimentos" são eliminadas, uma vez que a coluna "prova de conhecimentos" tem aproximadamente 400 valores em falta de 1331, o que corresponde, aproximadamente, aos mesmos valores em falta na coluna "resultado prova conhecimentos". Dado que a coluna prova de conhecimentos é que contêm a nota que o candidato obteve nessa prova e que, depois, a essa nota é atribuída uma pontuação específica na coluna "resultado prova conhecimento", sem esses valores não se consegue preencher os valores em falta dessa coluna e, por isso, com tantos espaços vazios, o mais adequado foi eliminar essas duas colunas. Esta decisão também não teve muito impacto, uma vez que estas colunas nem sempre eram consideradas para a seleção dos candidatos.

Por fim, a coluna "classificação final" também foi eliminada, pois devido às alterações feitas nas colunas das pontuações que dão origem ao resultado que aparece nessa coluna, esse resultado poderia já não estar correto e, por isso, optou-se por a eliminar e criar uma nova. Essa nova coluna,

agora com o nome de "pontuação_final", foi criada a partir da soma das novas colunas preenchidas corretamente: "new_habilitações", "qualificação profissional", "resultados_testes", "entrevista grupo pequeno" e "entrevista coordenador". É de relembrar que, nesta nova coluna, a coluna "resultado prova conhecimentos" já não foi contabilizada, uma vez que foi eliminada, tal como foi referido anteriormente.

4.3.2. Construção da base de dados final

Depois de todos os dados devidamente tratados e prontos para a próxima fase dos modelos, foi necessário ajustar e criar as bases de dados finais para serem utilizadas nos modelos. Uma vez que o objetivo não é utilizar uma única base de dados, mas sim diferentes bases de dados, com diferentes características, de modo a experimentar diferentes abordagens e a identificar qual a mais eficaz para dar resposta ao problema, foram criadas quatro bases de dados.

A primeira base de dados diz respeito à base de dados original, apenas com as alterações feitas às variáveis mencionadas anteriormente. Esta primeira base de dados, com a designação "df", tem como objetivo tentar prever a nota final, utilizando as colunas dos testes de aptidão, não tendo em conta as colunas das pontuações atribuídas. A segunda base de dados, com a denominação "df_points", baseia-se na base de dados "df", mas como o objetivo é usar apenas as colunas das pontuações, e não as colunas de cada teste de aptidão, foram retiradas desta base de dados as linhas com os valores em falta das colunas "entrevista grupo pequeno" e "entrevista coordenador". É de referir que estas duas bases de dados têm em consideração que a nota final engloba não só os formandos aprovados e reprovados, mas também os desistentes e desistentes no processo de seleção.

Por fim, a terceira e a quarta base de dados, respetivamente com os nomes "df_aprovado_reprovado" e "df_aprovado_reprovado_sem_entrevistas", seguem a mesma ordem de ideias que a primeira e a segunda, mas com uma alteração na coluna da nota final, em que só são tidos em conta os formandos aprovados e os reprovados. Ou seja, nestas duas bases de dados não consideramos os desistentes no processo de seleção, uma vez que estes formandos não chegaram a frequentar o curso e, apesar de terem sido selecionados para o frequentar e terem os registos de todos os resultados das provas que fizeram, nunca chegaram a formalizar a sua candidatura. Para além disso, poderia não fazer sentido tentar prever se o formando iria desistir, pois essa desistência poderia estar relacionada com fatores externos. Também os formandos considerados desistentes, formandos esses que desistiram durante o curso e não chegaram a concluí-lo, sofreram uma alteração, tendo-se decidido que seriam agora considerados na categoria dos formandos reprovados, o que poderá fazer sentido, pois ambas as categorias não concluíram o curso com sucesso, embora, não eliminar também estas linhas trouxesse a vantagem de não serem eliminadas mais 137 linhas, que seriam importantes para a análise. Posto isto e tal como foi referido anteriormente, estas duas bases de dados seguiram a mesma ordem de ideias das duas primeiras, sendo relevante referir que esta quarta base de dados (similar à segunda base de dados) também não considera as linhas vazias correspondentes às colunas "entrevista grupo pequeno" e "entrevista coordenador".

Assim, de modo a tornar mais percetíveis as bases de dados que foram construídas, segue-se uma tabela com um resumo da informação:

Tabela 4.2 – Resumo das bases de dados finais

Base de dados utilizada	Variáveis utilizadas
Base de dados 1 (df): Base de dados original, depois do tratamento dos dados. Uma vez que se vão utilizar os dados dos testes de aptidão, não se retiraram as linhas por preencher das entrevistas. Base de dados que é constituída pela classificação final de: aprovado, reprovado, desistente e desistiu no processo de seleção.	Idade, género, ano de início, ano de fim, TIG-I; TPD; PMA-V, PMA-N; PMA-E; ABI-1; ABI-2; ABI-3; ABI-4; ABI-5; ABI-6, código, new_habilitações, estado
Base de dados 2 (df_points): Base de dados original, depois do tratamento dos dados. Onde não vão ser utilizados os dados dos testes de aptidão e, por isso, se retiram as linhas por preencher das entrevistas. Base de dados que é constituída pela classificação final de: aprovado, reprovado, desistente e desistiu no processo de seleção.	Idade, género, ano de início, ano de fim, código, new_habilitações, estado, qualificação profissional, entrevista grupo pequeno, entrevista coordenador
Base de dados 3 (df_aprovado_reprovado): Base de dados original, depois do tratamento dos dados. Uma vez que se vão utilizar os dados dos testes de aptidão não se retiraram as linhas por preencher das entrevistas. Base de dados que é constituída pela classificação final de: aprovado, reprovado.	Idade, género, ano de início, ano de fim, TIG-I; TPD; PMA-V, PMA-N; PMA-E; ABI-1; ABI-2; ABI-3; ABI-4; ABI-5; ABI-6, código, new_habilitações, estado
Base de dados 4 (df_aprovado_reprovado_sem _entrevistas): Base de dados original, depois do tratamento dos dados. Onde não vão ser utilizados os dados dos testes de aptidão e, por isso, se retiram as linhas por preencher das entrevistas. Base de dados que é constituída pela classificação final de: aprovado, reprovado.	Idade, género, ano de início, ano de fim, código, new_habilitações, estado, qualificação profissional, entrevista grupo pequeno, entrevista coordenador

Assim e tal como foi referido anteriormente, segue novamente a análise das variáveis, mas agora depois das suas transformações e tendo como referência a base de dados "df". É de referir que, no final, a base de dados ficou com 30 variáveis e 1142 dados de formandos diferentes. Nas figuras seguintes apresenta-se graficamente o resultado das transformações operadas.

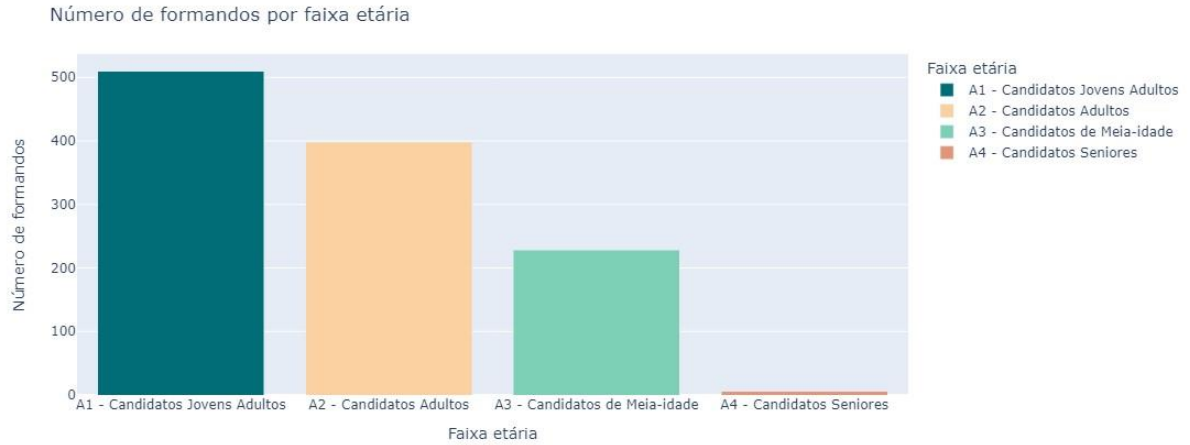


Figura 4.1 – Número de formandos por faixa etária

Após a representação gráfica de quantos formandos existem, por grupo etário, observa-se que o grupo etário com mais candidatos é o dos jovens adultos, seguido do grupo dos adultos e dos de meia-idade e, por fim, dos candidatos mais velhos.

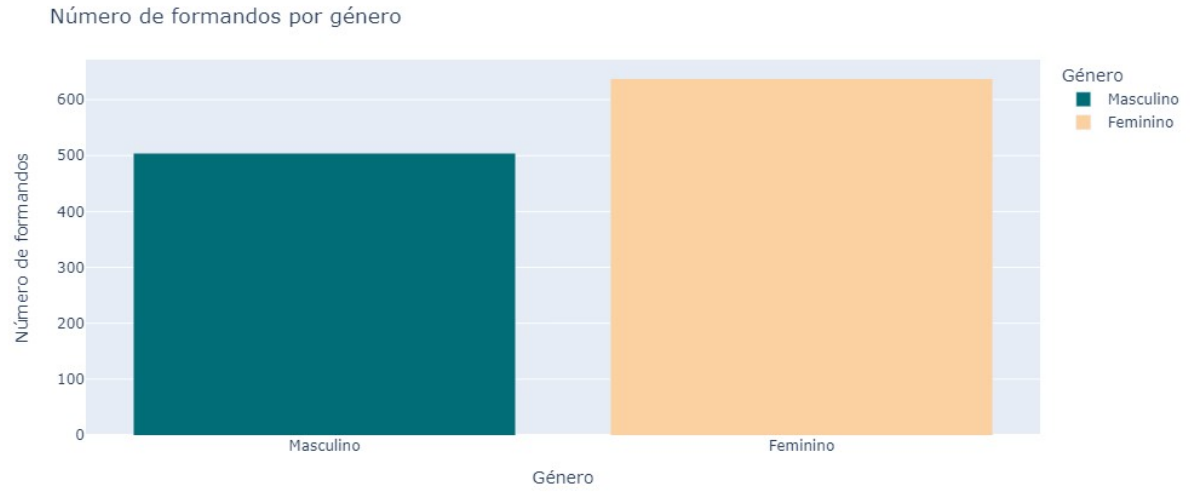


Figura 4.2 – Número de formandos por género

Relativamente ao gráfico da Figura 4.2, é percetível que mais de metade dos candidatos são do sexo feminino (710 pessoas), enquanto as restantes 570 são do sexo masculino.

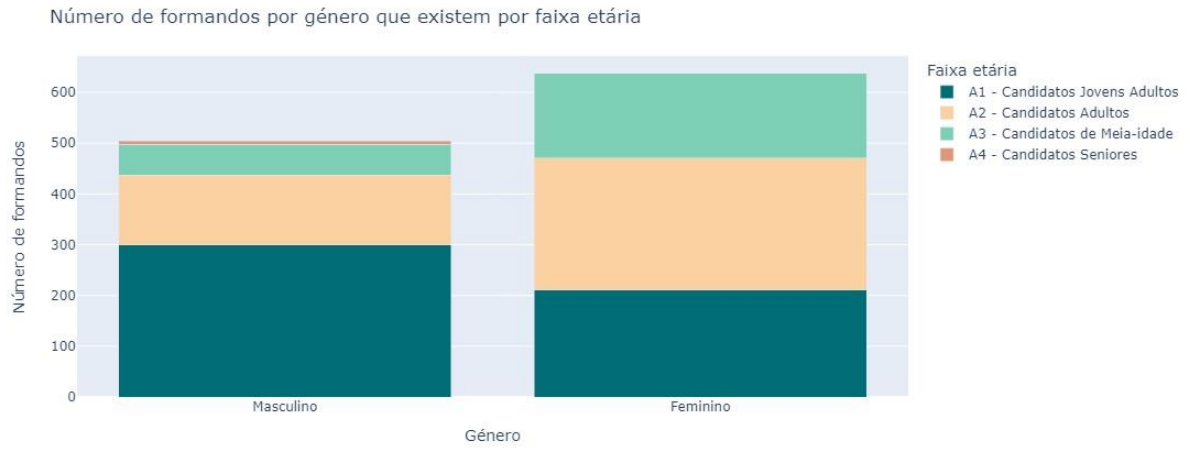


Figura 4.3 – Número de formandos por género que existem por faixa etária

Em relação ao gráfico da figura 4.3, conclui-se que a maior parte dos candidatos do sexo feminino pertencem ao grupo etário dos adultos, seguindo-se o dos jovens adultos, o da meia-idade e, por fim, ao dos candidatos mais velhos, enquanto no sexo masculino o grupo etário com mais presença é o dos jovens adultos, seguindo-se o dos adultos, da meia-idade e, por fim, o dos candidatos mais velhos.

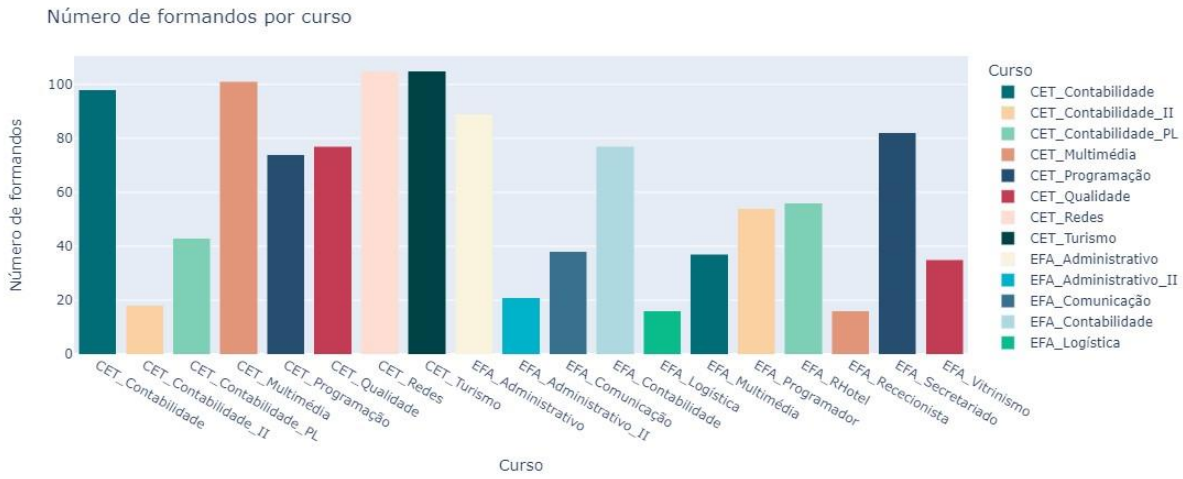


Figura 4.4 – Número de formandos por curso

No que se refere à Figura 4.4, o curso que apresenta uma maior procura é o CET_Contabilidade, seguido do CET_Redes e o curso com menor procura é o EFA_Rececionista.

Figura 4.5 – Número de formandos que iniciaram o curso em cada ano

Na Figura 4.5, é possível observar que o ano com mais procura pelos candidatos foi o de 2020, seguido do de 2019 e o ano com menos candidatos foi o de 2021.

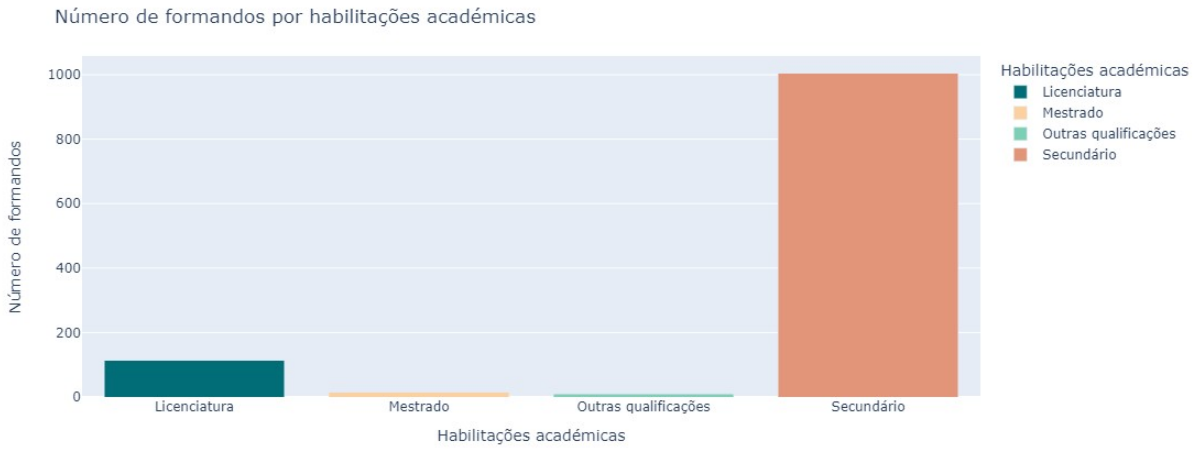


Figura 4.6 – Número de formandos por habilitações académicas

No que diz respeito à Figura 4.6, é notável perceber que a habilitação académica mais frequente deste grupo de formandos é o secundário, seguido da licenciatura, do mestrado e de outras qualificações.

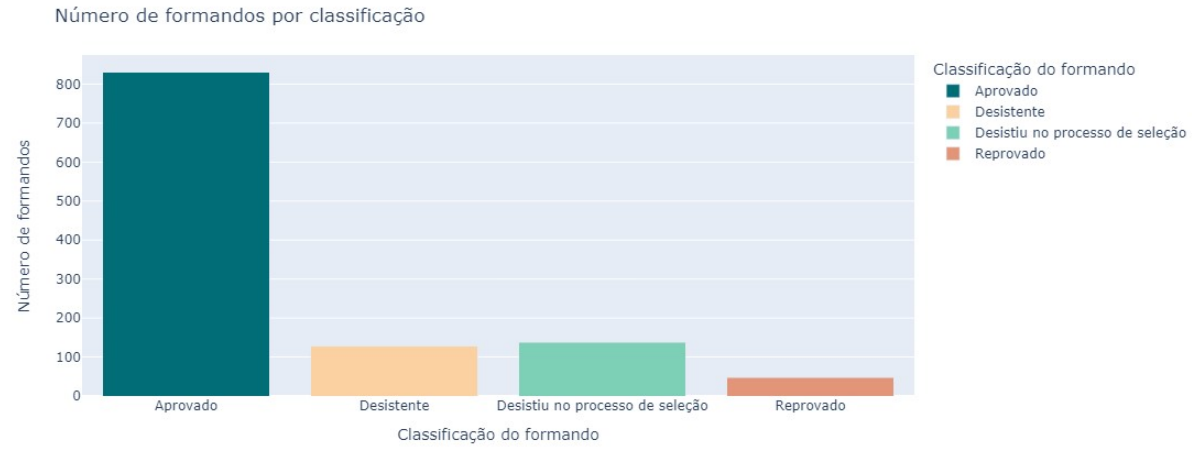


Figura 4.7 – Número de formandos por classificação

Quanto à Figura 4.7, a maior parte dos formandos apresenta sucesso no final do curso e, por isso, a classificação final "aprovados "é a que apresenta um maior número de formandos, no entanto, quando vamos comparar os formandos com classificação final "desistentes" com os formandos classificados como reprovados, conclui-se que existe um maior número de formandos nesta categoria do que na categoria da classificação final "reprovados".

Figura 4.8 – Classificação dos formandos por faixa etária

Relativamente à Figura 4.8, os formandos que apresentam uma classificação final de "aprovado" costumam ter uma presença mais significativa na faixa etária dos jovens adultos. Em relação aos formandos reprovados, são também os jovens que aparecem em maior quantidade nesta categoria.

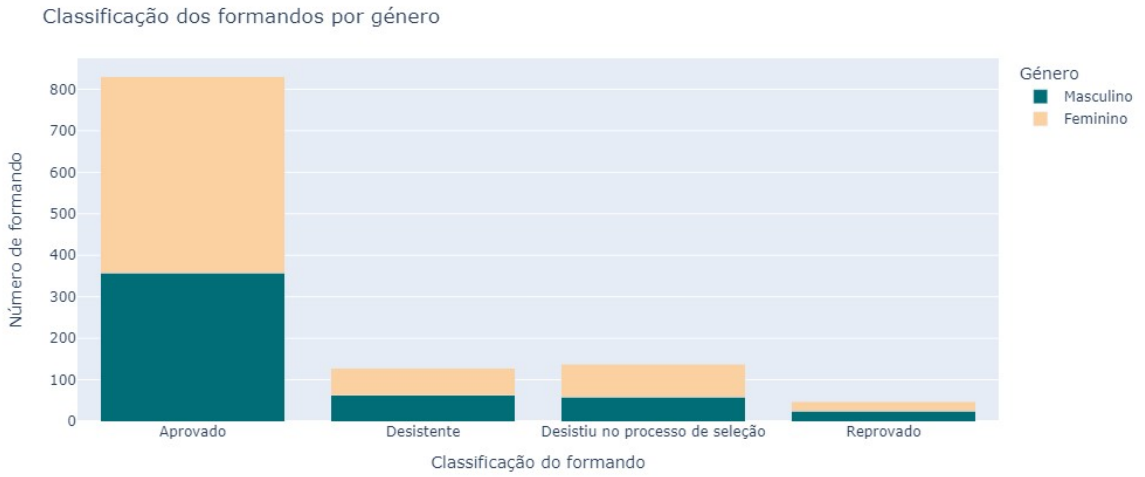


Figura 4.9 – Classificação dos formandos por género

No que diz respeito à Figura 4.9, o maior número de candidatos com a classificação final de "aprovado" costuma pertencer ao sexo feminino, enquanto no que diz respeito aos formandos reprovados é o sexo masculino que tem maior peso.

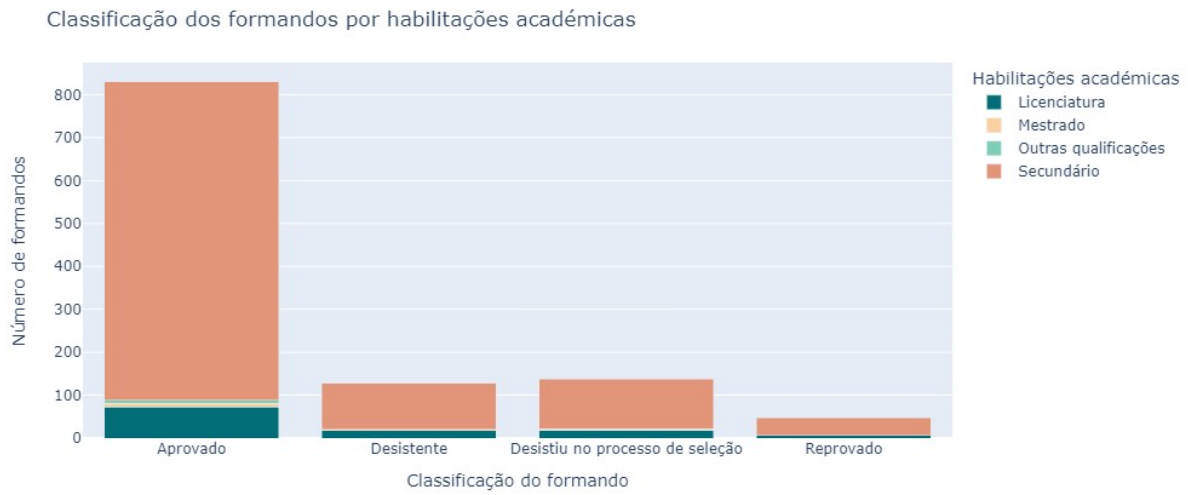


Figura 4.10 – Classificação dos formandos por habilitações académicas

Em relação à Figura 4.10, é notável perceber que, independentemente da classificação final do formando, a habilitação académica mais frequentada por parte dos formandos é o secundário.

Figura 4.11 – Classificação dos formandos por curso

No que se refere à Figura 4.11, é percetível que, na maior parte dos cursos, a classificação final mais frequente é a aprovação por parte do formando, no entanto, no caso do curso EFA_Logística, o curso apresenta uma maior taxa de formandos desistentes do que de formandos aprovados. Para além disso, pode concluir-se que os cursos CET_Contabilidade_II, EFA_Logística, EFA_Administrativo, EFA_Administrativo_II, EFA_RHotel, EFA_Rececionista e EFA_Vitrinismo não apresentam nenhum formando reprovado.

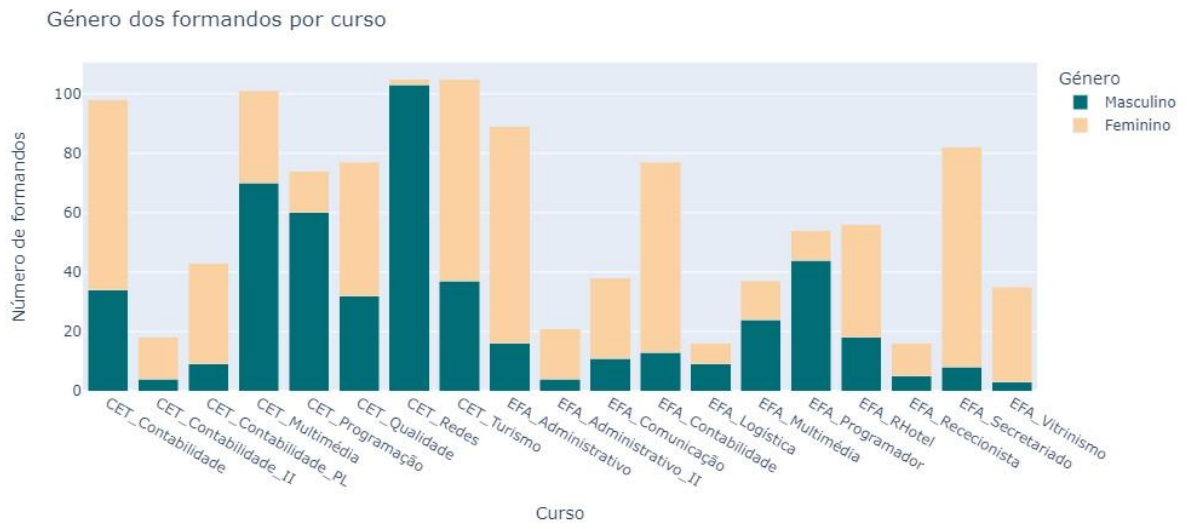


Figura 4.12 – Género dos formandos por curso

Quanto à Figura 4.12, os cursos de multimédia, programação e redes apresentam um maior número de formandos do sexo masculino, enquanto os outros cursos, tais como contabilidade, administrativo, rececionista, parecem apresentar um maior número de formandos do sexo feminino.

Pode concluir-se que apesar das alterações feitas à base de dados, as conclusões retiradas nos gráficos mantêm-se praticamente iguais às conclusões retiradas anteriormente.

4.4. DIVISÃO DA BASE DE DADOS

Após os dados terem sido todos tratados, seguiram-se os próximos passos, em que se começou pela divisão da base de dados e na definição das variáveis-objetivo. Tendo em conta que o objetivo do estudo é prever se um formando que seja selecionado para frequentar um curso irá ter um desempenho positivo, sendo aprovado no final, ou se irá reprovar, a variável nota final foi definida como variável-objetivo. Assim, para um melhor desempenho dos algoritmos de m*achine learning*, desenvolvidos na etapa seguinte, as quatro bases de dados foram todas divididas, de igual modo, em duas bases: a base de treino, constituída por 70% da base original e os restantes 30% da base de testes. Deste modo, os modelos serão treinados na base de treino, de modo a otimizar o desempenho dos modelos e, posteriormente, serão testados na base de testes.

4.5. NORMALIZAÇÃO DOS DADOS

Em seguida, uma vez que os valores que constituem cada variável da base de dados possui uma escala de valores bastante diferente, como, por exemplo: os dados dos testes de aptidão têm uma escala de 0 a 100, outra variáveis, tais como as colunas das pontuações, têm uma escala de valores entre 0 e 10, não faria sentido utilizar variáveis com escalas diferentes nos modelos, pois tal iria afetar o seu desempenho. Assim, foi necessário colocar todos os dados na mesma escala antes dos modelos, para evitar erros no cálculo das previsões.

Deste modo e uma vez os dados separados, no passo anterior, procedeu-se à normalização das variáveis, tanto da base de treino como na base de testes, com o objetivo de todas as variáveis estarem na mesma escala. Para se chegar a este resultado, foi aplicado o StandardScaler, convertendo todos os dados em valores numa escala entre 0 e 1.

4.6. SELEÇÃO DE VARIÁVEIS

Uma outra etapa também importante para melhorar o desempenho dos modelos foi perceber quais as variáveis mais significativas e que teriam maior impacto no desempenho dos algoritmos e as que não seriam muito relevantes. Para o efeito, foram usados quatro métodos distintos. Foi utilizado um método de filtragem, ANOVA, um método *embedded*, LASSO, um método de *intrinsic*, TREE, e um método de *wrapper*, RFE. Neste caso, não foi necessário usar métodos específicos para as variáveis categóricas, uma vez que, neste momento, na base de dados, as variáveis existentes eram todas numéricas.

Uma vez que de cada método resulta uma escolha de variáveis diferentes, decidiu-se que a seleção de variáveis a manter iria ser feita com base no critério da soma do número de variáveis que cada método apontava ser para descartar. É de referir que o conjunto de variáveis obtido difere consoante cada base de dados utilizada, mas, as variáveis "idade", "ano de início, "ano de fim" e "código" foram selecionadas em todos os casos e as variáveis "género", "estado" e "new_habilitações" também foram selecionadas na maior parte dos casos. Depois, as restantes variáveis, dos testes de aptidão e das colunas das pontuações, foram sempre selecionadas na base de dados a que diziam respeito.

4.7. BALANCEAMENTO DOS DADOS

Por último e antes da serem aplicados os modelos, foi preciso ter em conta que a variável que se pretende prever conta com uma grande desproporcionalidade nos dados, uma vez que a maioria dos dados apresenta o valor "1", aprovado e que o valor "2", que corresponde aos formandos reprovados, está em minoria. Assim, recorreu-se a um método de balanceamento dos dados, o SMOTE. Esta técnica gera exemplos sintéticos para a classe de dados que está em minoria, conseguindo, assim, equilibrar a amostra de dados.

4.8. MODELOS

Após todos estes passos e de os dados estarem finalmente prontos, iniciou-se a fase dos modelos. Nesta etapa, são escolhidos os modelos de *machine learning* mais apropriados, consoante o problema e as características dos dados, uma vez que cada algoritmo apresenta vantagens e desvantagens diferentes e nem todos são adequados ao objetivo. Para este estudo, decidiu-se usar sete modelos diferentes de aprendizagem supervisionada para classificação, de modo a ser possível, posteriormente, comparar-se resultados e concluir qual deles se ajusta melhor a cada caso. Assim, os modelos escolhidos foram *decision tree*, *support vector machine*, *neural network*, *random forest*, *GaussianNB*, *LogisticRegression*, *LinearDiscriminanAnalysis*, *AdaBoostClassifier*, *GradientBoostingClassifier* e *ExtraTreesClassifier*.

Para medir o desempenho dos modelos e, posteriormente, poderem comparar-se os resultados, usou-se um sistema de *K-fold cross validation*. Este método consiste na divisão do conjunto de dados de treino em "K" subconjuntos, de igual tamanho, correspondendo estes números de subconjuntos resultantes aos "*folds*". Após a divisão, o modelo é experimentado, utilizando K-1 subconjuntos, sendo estes subconjuntos o que representa o conjunto de treino e o restante subconjunto, usados para dados de teste, tal como se pode verificar no exemplo da Figura 4.13. Este método é vantajoso para conjuntos de dados que são desproporcionados na variável-objetivo, tal como é o caso dos dados deste estudo, pois os valores que aparecem em minoria poderão ficar situados mais no conjunto de dados de teste do que nos de treino. Assim, é possível fazer uma boa previsão dos dados recorrendo a este sistema,

onde se vão usando vários subconjuntos de dados diferentes aumentando, assim, a probabilidade de se terem em conta os valores em minoria. Para além disso, o método de *K-fold cross validation* ajuda a resolver o problema de *overfiting* (Berrar, 2018).

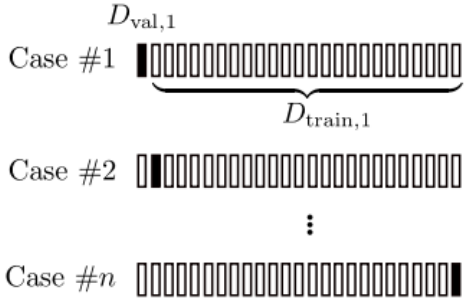


Figura 4.13 – *K-fold cross validation* (Berrar, 2018)

Para além disso, e de modo a que os modelos fiquem otimizados o máximo possível, usou-se a técnica de *grid search*. Esta técnica é usada para otimizar os hiperparâmetros, sendo introduzido um conjunto de parâmetros possíveis para esse algoritmo, fazendo o modelo uma pesquisa exaustiva sobre esses parâmetros, para, no fim, conseguir concluir quais são os melhores parâmetros desse conjunto que lhe foi dado, para se usar nesse algoritmo. No entanto, esta técnica tem uma desvantagem que é ser um processo, muitas vezes, lento a executar e apresentar os resultados (Liashchynskyi & Liashchynskyi, 2020).

Após a escolha dos melhores parâmetros e do melhor modelo a usar para fazer a previsão da variável-objetivo, os dados de teste são então passados ao modelo. É de realçar que todos estes passos são feitos para cada um dos quatro conjuntos de dados usados. Para além da existência de quatro base de dados diferentes, tal como já foi explicado anteriormente, cada base de dados poderá usar um de dois conjuntos de colunas existentes. Ou seja, uma vez que é necessário perceber se os testes de aptidão são suficientes para prever o desempenho do formando, utiliza-se a primeira e a terceira base de dados para perceber o peso dos testes da aptidão, e a segunda e a quarta base de dados, sem os testes, para perceber realmente se estas variáveis são significativas e fazem diferença no valor da previsão.

4.9. AVALIAÇÃO

Tendo em conta que os Centros de Formação possuem inúmeros testes de aptidão, mas que os candidatos apenas realizam um grupo de testes de aptidão necessários ao curso a que se candidatam e não efetuam todos os existentes, esta situação cria um problema. Este problema surge, uma vez que cada candidato só preenche o resultado dos testes que realizou e para os testes que não fez, esses campos ficam em branco. Visto que nos modelos não convém utilizar dados por preencher, pois podem induzir em erro os modelos, contornou-se esta situação verificando que grupos de testes existiam, ou seja, que testes eram realizados ao mesmo tempo, e os modelos foram testados para cada grupo de testes e não para toda a base de dados, em geral. É preciso não esquecer que existem quatro bases de dados, mas que apenas a primeira e a terceira é que são constituídas pelas colunas dos testes de aptidão e são só estas que necessitam desta divisão do grupo de testes.

Assim, observou-se a existência de nove grupos de testes diferentes: ('PMA-N', 'TIG-I' e 'TPD'), ('ABI-1', 'ABI-4', 'ABI-5' e 'ABI-6'), ('PMA-E', 'PMA-V' e 'TIG-I'), ('ABI-1', 'ABI-2', 'ABI-4', 'ABI-5' e 'ABI-6'), ('PMA-V', 'TIG-I' e 'TPD'), ('ABI-1', 'ABI-5', 'ABI-6' e 'TIG-I'), ('ABI-1', 'ABI-2', 'ABI-4' e 'ABI-6'), ('PMA-E', 'TIG-I' e 'TPD'), ('ABI-2', 'ABI-4', 'ABI-5' e 'ABI-6').

Posto isto, para cada grupo de testes foram feitos todos os mesmos passos (*train test split*, *stadartization*, *feature enginering*, *imbalance classification* e *moddling*) e selecionado o melhor modelo para cada um deles. Após ter sido selecionado o melhor modelo, verificou-se qual foi o valor da sua previsão. De reforçar que todas as bases de dados são constituídas por um conjunto de colunas fixas que estão presentes em todas as bases de dados, sendo elas "idade", "género", "new_habilitações", "ano de Início", "ano de fim", "código" e "estado", e, depois, as outras colunas dos testes de aptidão e das pontuações diferem consoante a base de dados utilizada.

Assim, para a base de dados que tenta prever se um formando vai ser aprovado ou reprovado, considerando apenas os testes de aptidão, foram obtidos os seguintes resultados:

Tabela 4.3 – Resultado do desempenho dos modelos para a base "df_aprovado_reprovado"

Base de dados utilizada: df_aprovado_reprovado		
Testes	**Melhor modelo**	***Accuracy***
('PMA-N', 'TIG-I' e 'TPD')	RFC	85,44%
('ABI-1', 'ABI-4', 'ABI-5' e 'ABI-6')	NN	71,43%
('PMA-E', 'PMA-V' e 'TIG-I')	SVC	90,91%
('ABI-1', 'ABI-2', 'ABI-4', 'ABI-5' e 'ABI-6')	NN	76,67%
('PMA-V', 'TIG-I' e 'TPD')	RFC	75,23%
('ABI-1', 'ABI-5', 'ABI-6' e 'TIG-I')	NB	94,44%
('ABI-1', 'ABI-2', 'ABI-4' e 'ABI-6')	RFC	64,71%
('PMA-E', 'TIG-I' e 'TPD')	SVC	82,35%
('ABI-2', 'ABI-4', 'ABI-5' e 'ABI-6')	NN	75,00%

Tal como se pode observar, os valores obtidos sofrem de *overfiting*, principalmente se repararmos no grupo de testes que obteve o valor mais elevado da previsão (identificado a vermelho). Estes valores podem ter sofrido esta condição por bastarem poucos dados para contribuir para esse fator.

Na base de dados que tenta prever se um formando irá ser aprovado, reprovado ou desistir, considerando também apenas os testes de aptidão, foram obtidos os seguintes resultados:

Tabela 4.4 – Resultado do desempenho dos modelos para a base "df"

Base de dados utilizada: df		
Testes	**Melhor modelo**	***Accuracy***
('PMA-N', 'TIG-I' e 'TPD')	SVC	71,90%
('ABI-1', 'ABI-4', 'ABI-5' e 'ABI-6')	SVC	69,23%
('PMA-E', 'PMA-V' e 'TIG-I')	SVC	83,33%
('ABI-1', 'ABI-2', 'ABI-4', 'ABI-5' e 'ABI-6')	RFC	67,65%
('PMA-V', 'TIG-I' e 'TPD')	NN	43,09%
('ABI-1', 'ABI-5', 'ABI-6' e 'TIG-I')	SVC	85,00%
('ABI-1', 'ABI-2', 'ABI-4' e 'ABI-6')	SVC	71,79%
('PMA-E', 'TIG-I' e 'TPD')	RFC	55,56%
('ABI-2', 'ABI-4', 'ABI-5' e 'ABI-6')	SVC	60,00%

Nesta base de dados, pode também perceber-se que o melhor grupo de testes foi o mesmo que o da Tabela 4.3 "('ABI-1', 'ABI-5', 'ABI-6' e 'TIG-I')", mas, neste caso, não se nota tanto a questão dos valores com *overfiting*.

Após terem sido analisadas estas duas bases de dados, poderá concluir-se que os modelos que mais se destacaram, entre os 10 utilizados, foram o RFC, com três grupos de testes a utilizarem-no como melhor modelo, com a base de dados "df_aprovado_reprovado" e dois grupos de testes na base "df"; o SVC, com dois e seis grupos de testes a utilizarem-no, respetivamente, nas bases "df_aprovado_reprovado" e "df"; o modelo NN, com três e um grupos de testes a utilizarem-no, respetivamente, nas bases "df_aprovado_reprovado" e "df" e, por fim, a base de dados "df_aprovado_reprovado" ainda tem um grupo de testes a utilizar como melhor modelo, o NB.

De um modo geral, pode perceber-se que a utilização dos testes de aptidão, na maior parte dos grupos de testes, para calcular a previsão do desempenho dos formandos no final do curso, será positiva, apesar de sofrerem algum *overfiting* que poderá ser justificado por as bases de dados que constituem cada grupo de testes serem pequenas, pois, ao dividirmos a base de dados original em diferentes bases de dados para cada grupo de testes, daremos origem a nove bases de dados bem mais pequenas, o que poderá afetar, ainda mais, os resultados.

No entanto, para contrariar o problema de ter nove base de dados pequenas criadas pelos diferentes grupos de testes, foi testada a utilização, não da pontuação obtida em cada teste individual, mas sim a da coluna final, da pontuação atribuída ao resultado dos testes ("resultados_testes"), para assim minimizar o problema das bases de dados serem pequenas e perceber se seria um fator que estaria realmente a afetar a previsão dos modelos. Assim, usou-se uma base de dados em que não foram considerados os testes individuais, mas sim a coluna "resultados_testes" e as outras três colunas tidas em conta no processo de seleção ("qualificação profissional", "entrevista grupo pequeno", "entrevista coordenador").

Assim, para a base de dados em que são considerados os formandos aprovados, reprovados e desistentes, obtiveram-se os seguintes resultados:

Tabela 4.5 – Resultado do desempenho dos modelos para a base "df"

Base de dados utilizada: df		
Testes	**Melhor modelo**	***Accuracy***
Colunas das pontuações com a pontuação dos testes	RFC	63,41%

E para a base de dados que só tem em conta os formandos aprovados e reprovados:

Tabela 4.6 – Resultado do desempenho dos modelos para a base "df_aprovado_reprovado"

Base de dados utilizada: df_aprovado_reprovado		
Testes	**Melhor modelo**	***Accuracy***
Colunas das pontuações com a pontuação dos testes	RFC	81,03%

Pode concluir-se que utilizar-se as nove bases de dados para cada grupo de testes ou usar diretamente a coluna da pontuação final dos testes não afeta muito os valores obtidos.

Por outro lado, testando agora efetivamente se os testes de aptidão têm impacto na previsão do desempenho dos formandos, foi feita a previsão através da utilização de outras variáveis tidas em conta na seleção dos candidatos ("qualificação profissional", "entrevista grupo pequeno", "entrevista coordenador"), para além das fixas ("idade", "género", "new_habilitações", "ano de início", "ano de fim", "código", "estado"), sem ter em conta o resultado obtido nos testes de aptidão.

Deste modo, obtiveram-se os seguintes resultados na base de dados em que consideramos, na variável objetivo, os formandos aprovados, reprovados e desistentes:

Tabela 4.7 – Resultado do desempenho dos modelos para a base "df_points"

Base de dados utilizada: df_points		
Testes	**Melhor modelo**	***Accuracy***
Colunas das pontuações sem os testes	RFC	59,45%

Pode concluir-se que, efetivamente, não foram obtidos muito bons resultados para esta base de dados, considerando estas variáveis e esta variável-objetivo.

No entanto, na base de dados em que apenas consideramos os formandos aprovados e reprovados foram obtidos os seguintes resultados:

Tabela 4.8 – Resultado do desempenho dos modelos para a base "df_aprovado_reprovado_sem _entrevistas"

Base de dados utilizada: df_aprovado_reprovado_sem _entrevistas		
Testes	**Melhor modelo**	***Accuracy***
Colunas das pontuações sem os testes	RFC	80,00%

Neste caso, os resultados já se mostram mais promissores, tendo esta base de dados obtido um valor muito positivo, não considerando os testes de aptidão e apenas considerando outras colunas.

Assim e apesar dos resultados obtidos, por exemplo, de 85% na Tabela 4.4 e de 81,03% na Tabela 4.6, quando utilizamos os testes de aptidão para fazer a previsão, a diferença acaba por não ser muito significativa em relação aos resultados das previsões, quando não são utilizado os testes de aptidão, 80%, e, por isso, é possível concluir que os testes de aptidão não trazem uma diferença considerável aos resultados, uma vez que as diferenças percentuais entre os valores das diferentes situações são mínimas.

Este resultado poderá significar que os testes de aptidão ajudam a validar o resultado que os outros parâmetros já haviam confirmado, isto é, garantir a conclusão da formação; no entanto, enquanto um parâmetro de seleção isolado, os testes não são determinantes para prever o desempenho de um formando, mas sim, poderão servir como fator eliminatório. Por exemplo, dois formandos que tenham a mesma pontuação final na soma de todos critérios, quando já só reste uma vaga, o valor dos testes de aptidão poderá servir como fator eliminatório para a escolha do candidato a selecionar.

5. Construção de uma *Dashboard*

5.1. OBJETIVOS

Tendo como base o objetivo deste estudo, de tentar prever o desempenho dos formandos no final do curso que vão frequentar e uma vez atingido esse objetivo para este grupo de dados, viu-se que também seria benéfico tornar este objetivo palpável e interativo. Assim, este estudo completa-se com a implementação de uma ferramenta que possa ser usada diariamente por todas as pessoas que precisem de ter acesso a esta informação, para que, de alguma forma, torne a tomada de decisões mais fácil e informada, mas também que se possa fazer esta previsão, não só para estes dados, mas também para dados novos que se queira introduzir na base de dados.

Para além desta funcionalidade na *dashboard*, seria também importante aproveitar para completá-la com informação acerca dos candidatos que frequentam o curso (tal como a idade, o número de formandos a frequentar cada curso, etc.), de modo a que quem precise tenha a informação atualizada, correta e de fácil acesso, numa visão geral do panorama dos cursos e candidatos a frequentar o Citeforma. Assim, esta *dashboard* permitirá, não só que sejam conhecidas as características dos seus candidatos, como também o seu desempenho.

5.2. ESTRUTURA

Assim sendo e com um objetivo estabelecido, foi necessário criar uma *dashboard* que correspondesse a todas as necessidades, mas que também fosse fácil de usar, interativa e personalizada. Desta maneira, foi criado um *layout* de raiz, tendo por base componentes de html, de modo a tornar os gráficos interativos e de chegar à informação de forma intuitiva, através da ajuda de botões ou de listas de valores.

Desta forma, do lado esquerdo, observa-se uma lista com os anos de início de cada curso, neste caso, entre 2017 e 2021, por baixo, uma lista de todos os cursos que existiram nesse ano e uma lista com o desempenho do formando no final do curso, sendo esta lista inserida numa caixa que percorre a folha de cima a abaixo, de modo a acompanhar sempre a informação que se visualiza do lado direito. Do outro lado da página (lado direito), tem-se acesso a várias informações, dependendo do(s) ano(s) e curso(s) selecionado(s). Assim, no topo direito, encontram-se quatro caixas com informação rápida acerca da idade, do número de pessoas a frequentar os cursos e do número de cursos ativos. Por baixo, encontra-se uma variedade de gráficos, com informação variada acerca da habilitação literária dos formandos, do grupo etário, do género e das notas finais de curso. Assim, procura-se que o utilizador consiga recolher, de forma rápida e fácil, a informação necessária acerca dos seus candidatos. Por último e por baixo de toda esta informação, encontra-se uma secção onde se poderá calcular a previsão do desempenho do candidato no final do curso, através do preenchimento da informação necessária do candidato em relação ao qual se pretende saber o desempenho futuro no curso. É também possível verificar que nível de precisão teve o modelo, de um modo geral, daquela previsão fornecida para aquele candidato.

5.3. VISUALIZAÇÕES

5.3.1. Lista dos anos, cursos e desempenho

Tal como foi referido anteriormente, o lado esquerdo da *dashboard* é constituído por uma lista com todos os anos em que se iniciaram cursos, por uma lista de cursos que existiram em cada ano e o desempenho do formando no final do curso, encontrando-se aí a nota do formando no fim do curso. Nas três listas, será possível escolher apenas um ano, um curso ou um desempenho ou escolher vários anos, cursos e classificações finais, conforme o desejado. Dependendo do que for escolhido, esta informação irá depois afetar toda a informação presente do lado direito da *dashboard*. Esta visualização poderá ser consultada no Apêndice A.

5.3.2. Resumo das principais características dos dados

Numa primeira caixa, é possível verificar-se a média das idades dos formandos que realizaram aquele(s) curso(s), naquele(s) ano(s), na segunda caixa, observa-se o número de pessoas por curso(s), na terceira caixa, o número de pessoas por género e, por fim, na última caixa, obtém-se a informação acerca de quantos cursos se realizaram em cada ano. Esta visualização poderá ser consultada no Apêndice B.

5.3.3. Perfil do candidato

Na secção do perfil do candidato tem-se acesso a quatro gráficos, com informação acerca dos formandos, esta informação muda consoante o(s) ano(s) e o(s) curso(s) selecionado(s), tal como foi referido anteriormente. Assim, no gráfico "Número de pessoa por curso", observa-se o número de candidatos que frequentou aquele curso em determinado ano, através de um gráfico de barras. No segundo gráfico, "Número de pessoas por habilitação literária", também de barras, verifica-se quantos formandos iniciaram a sua candidatura com o secundário, a licenciatura e o mestrado. No terceiro gráfico, "Número de pessoas por grupo etário", mostra-se, através de um gráfico circular, quantas pessoas existem em cada grupo etário a frequentar os cursos no(s) ano(s) selecionado(s). Por fim e novamente num gráfico de barras, apresenta-se o "Número de pessoas por curso e por género", em que o objetivo é perceber o número de pessoas por género a frequentar determinado curso. Estas visualizações poderão ser consultadas nos Apêndices C, D e E.

5.3.4. Análises do desempenho dos candidatos

Esta secção é constituída por três gráficos de barras. No primeiro gráfico, observa-se o "Número de pessoas por nota", ou seja, pode conhecer-se a informação de quantas pessoas foram aprovadas, reprovadas ou desistentes no(s) ano(s) e curso(s) selecionado(s), no segundo gráfico, "Número de pessoas por nota e por género", retira-se a mesma informação que do anterior, mas agora com a particularidade de se saber também o número de pessoas por género. Na mesma ordem de ideias, no terceiro gráfico, "Número de pessoas por nota e por idade" poderá retirar-se a informação do número de candidatos que foram aprovados, reprovados ou desistentes no(s) ano(s) e curso(s) selecionado(s), mas por grupo etário. Estas visualizações poderão ser consultadas nos Apêndices F e G.

5.3.5. Previsão

A última secção da *dashboard* tem como objetivo permitir ao utilizador criar modelos que poderão ser experimentados, recorrendo aos vários agrupamentos de dados existentes e, posteriormente, postos à prova, de uma maneira em que seja possível ao utilizador fornecer novos dados. Isto poderá acontecer quando, por exemplo, o utilizador insere novos dados, referentes a um novo candidato, para assim obter a previsão do desempenho do futuro formando, que advém daquilo que o modelo infere, com base nas características fornecidas e os dados que usou para "aprender". Estas visualizações poderão ser consultadas nos Apêndices H e I.

5.4. DISCUSSÃO DA *DASHBOARD*

Esta *dashboard* consegue assim ser bastante útil, pois fornece, de maneira rápida e prática, bastante informação necessária aos seus utilizadores. Com esta *dashboard*, o utilizador consegue assim fazer, de forma geral, uma avaliação do estado dos seus cursos, assim como dos formandos que os frequentam. A informação recolhida abrange desde os dados acerca das características do formando, tais como as médias de idade, as habilitações académicas, as classificações finais que estão a obter no final do curso, mas também dados mais detalhados acerca dos cursos, tal como é o caso do número de formandos por curso, entre outras opções disponíveis. Tudo isto, poderá ajudar, de alguma forma, os utilizadores, pois conseguem ter toda a informação de que necessitam agregada num só sítio e de forma sintetizada, o que consequentemente poderá ajudar à tomada de decisões, caso seja necessário retificar e ajustar a oferta ao tipo de procura e ao perfil do candidato.

Para além disso, esta *dashboard* oferece a possibilidade de o utilizador poder fazer a previsão da classificação final de novos candidatos que surjam no Citeforma, com o intuito de ser uma ferramenta adicional que ajude a uma seleção de candidatos mais eficaz.

No entanto, é necessário ter em conta que esta *dashboard* tem algumas limitações e que deve crescer e ser adaptada, tendo em conta as necessidades do utilizador.

6. Conclusão

6.1. Síntese do trabalho desenvolvido

Em conclusão, esta dissertação teve dois pontos fulcrais como objetivo: perceber se os testes de aptidão poderão ser usados para prever o desempenho de um formando e, em paralelo, prever efetivamente o desempenho de um formando no final do curso. Adicionalmente, através das questões de investigação enunciadas no início do estudo, foi possível retirar outras conclusões, como ter a possibilidade de caracterizar um formando com bom aproveitamento escolar e perceber se as variáveis tidas em conta são as mais corretas.

Uma vez definidos os objetivos, foi feita uma revisão de literatura, em que foi possível compreender o que já tinha sido feito e abordado em estudos similares, com o intuito de perceber qual é o caminho a seguir. Através desta pesquisa, foi possível concluir e confirmar a importância dos centros de formação e o tipo de formação que oferecem na atualidade, uma vez que permitem aos formandos a oportunidade de se especializarem em novas áreas de trabalho, mas também, atualizarem-se e melhorarem as suas competências, permitindo-lhes acompanhar as transformações e as tendências do mercado de trabalho atual. Para além disso e numa perspetiva mais a pensar no modo de fazer a previsão do desempenho dos formandos, foi possível compreender quais as técnicas de *machine learning* mais utilizadas, assim como os modelos utilizados nesta temática.

Assim, a previsão do desempenho dos formandos foi feita com base numa análise de dados, constituída por 1331 registos de diferentes formandos que frequentaram os 19 cursos ativos, ao longos dos anos de 2018 a 2022 e constituídos por 30 variáveis diferentes. Deste modo, esta análise revelou que, apesar dos testes de aptidão serem determinantes e uma componente importante no processo de seleção dos formandos, não são as variáveis que parecem ter um maior impacto no resultado final da previsão do desempenho, pois o estudo comprovou que, tendo em conta apenas outras variáveis, tais como as diferentes entrevistas realizadas, a idade e as habilitações literárias consegue-se alcançar um resultado bastante positivo e similar ao das previsões feitas quando são aplicado os testes de aptidão. No entanto, deve ter-se em conta que estas conclusões são apenas baseadas no que foi possível concluir, através dos dados disponibilizados para este estudo, e que, tal como irá ser mencionado mais detalhadamente a seguir, é necessário ter em consideração que outras variáveis que não foram consideradas para este estudo também têm uma grande repercussão no desempenho do formando no curso. Paralelamente, uma vez que os dados facultados não foram em grande volume, os resultados poderão não ser os mais precisos, pois os modelos de *machine learning* exigem uma grande quantidade de dados para aprender de forma mais eficaz e retirar conclusões mais exatas. Assim, será necessário ter em consideração estes fatores quando os resultados apontam para que os testes de aptidão não são as variáveis mais determinantes no processo de seleção.

Contudo, não foram apenas estas as conclusões retiradas, pois, através das questões das investigações, foi possível compreender que, de uma forma geral, um formando que normalmente conclui a sua formação com sucesso é constituído por um conjunto de características, tais como, uma idade a rondar os 28 anos; em relação às notas obtidas nos diferentes testes, os testes de aptidão TIG-I e TPD obtêm uma classificação, em média, de 54 pontos, o teste PMA-V, de 65 pontos, o PMA-N, de 61, o PMA-E, de 69, o ABI-1, de 53, o ABI-2, de 41, o ABI-3, de 43, o ABI-4 e o ABI-5, de 59 e o ABI-6, de 77 pontos, todos em média. Em relação à qualificação profissional, um formando tem normalmente em média um ponto, nas habilitações literárias os formandos que são aprovados no final do curso têm,

normalmente, em média, o secundário como habilitação literária e, no que diz respeito às entrevistas, na entrevista com o coordenador, o formando costuma geralmente obter em média a pontuação de 1,65 e na entrevista com o grupo pequeno uma pontuação de 4. Deste modo, de uma forma genérica e aproximada, é possível traçar o perfil de um candidato que tenha normalmente sucesso no final do curso e assim é possível ter um exemplo de que valores, em média, se deverão ter em conta quando se consideram para estes aspetos.

Quanto à questão, "Porque é que o mesmo conjunto de candidatos, selecionados para o mesmo curso, têm taxas de sucesso diferentes, tendo em conta que terão tido resultados semelhantes nos testes de aptidão?", tal poderá dever-se aos fatores externos que não são tidos em conta, tais como, por exemplo, um formando arranjar um trabalho durante o ano em que está a frequentar o curso. Uma vez que esta formação profissional tem normalmente a duração de um ano e se realiza durante o dia, com uma carga horaria elevada, implica uma disponibilidade quase total do formando para poder frequentar o curso e o concluir com sucesso. Para além disso, verifica-se, muitas vezes, que os formandos saem do curso por razões de cariz pessoal e de circunstâncias inesperadas, sendo estas ocorrências difíceis de prever e, por isso, não são consideradas no momento da seleção dos candidatos.

Relativamente à questão, "As variáveis que estão a ser tidas em conta serão as mais acertadas? Será necessário considerar e ponderar novas variáveis que possam determinar o sucesso do formando na realização do curso?", deverá ter-se em conta o referido no ponto 2.2 da revisão de literatura, onde se faz referência ao papel que as variáveis assumem no processo de aprendizagem. Estas devem também ser tidas em conta, assumindo um carácter de complementaridade aos resultados dos testes de aptidão, pois, como método de seleção, estes não são suficientes, tal como se pôde comprovar nos resultados obtidos neste estudo. Conclui-se que as variáveis que já são tidas em conta deverão continuar a sê-lo, no entanto, deverão ser consideradas outras, relativas a características pessoais, para que, de alguma forma, seja possível estabelecer o perfil do candidato, prevendo se será um candidato aprovado, reprovado ou desistente.

Deste modo, variáveis/ campos como o número de filhos, se é empregado ou desempregado, se pretende trabalhar durante o decorrer do curso, o local onde mora e a nacionalidade (apesar de já existir a variável neste estudo, a maior parte dos formandos não tem a informação preenchida), entre outros, deverão também ser tidos em conta. Tal como se pôde observar, nos resultados obtidos, aquando da aplicação dos modelos de *machine learning*, obteve-se uma boa previsão, complementando-se os testes de aptidão com um conjunto mais alargado de variáveis, tais como os mencionadas em cima, alcançando melhores resultados nas previsões dos modelos e na seleção dos candidatos.

No que diz respeito à questão, "Será que o intervalo do percentil escolhido, para a seleção dos candidatos, é o correto ou precisa de ser ajustado?", esta deve-se à forma como é atribuída a pontuação a um formando, após realizar os testes de aptidão necessários ao curso a que se candidatou. Tal como foi referido no capítulo anterior, após os candidatos do mesmo curso realizarem as provas de aptidão, é feita uma média dessas notas e essas médias são, posteriormente, ordenadas por ordem decrescente e esse grupo de formandos desse curso é dividido em três, pois é considerado um percentil e ao primeiro tercil é atribuída uma pontuação de 3 pontos, uma vez que neste grupo se encontram os formandos com média mais elevada nas provas, no segundo tercil, uma pontuação de 2 e no último, uma pontuação de 1.

No entanto, não se pode considerar se um percentil será o mais correto ou não, pois, segundo se apurou, depende do peso que se queira dar às variáveis, ou seja, se o objetivo for que os testes de aptidão não sejam uma métrica penalizadora, então, deverão continuar a utilizar-se os tercis, pois este método acaba por englobar um grande grupo de candidatos no mesmo grupo, criado dentro do tercil, ou seja, uma pessoa que tenha uma média de 80 nos testes, poderá pertencer ao mesmo grupo de uma pessoa que teve uma média de 69 nos testes, acabando por beneficiar as notas mais baixas dos testes de aptidão com pontuações mais altas, enquanto que, se considerarmos um quartil ou um quintil, formandos que tenham notas piores em testes de aptidão deixam de ser tão beneficiados e passa a dar-se maior valor a melhores notas, pois os quartis forçam a uma maior demarcação entre as notas dos formandos, deixando apenas a maior pontuação para quem realmente teve as melhores notas.

Assim, caso se queira que os testes de aptidão sejam um fator penalizador na seleção dos candidatos, dever-se-á então considerar um quartil ou um quintil. Para além disso, segundo a análise realizada para diferentes quantis, o que se verificou é que, realmente, quando a pontuação final (soma de todas as colunas das pontuações utilizadas para fazer a seleção dos candidatos) foi verificada, alguns formandos ficavam em posições diferentes, ou seja, um formando que teria sido selecionado ficou como suplente e um formando que iria ser suplente ficou como selecionado, mas este acontecimento depende do que se pretende valorizar ou penalizar, não havendo uma maneira exata de fazer a escolha dos percentis, pois dependendo da escolha, são obtidos resultados diferentes para as diferentes finalidades desejadas. Ou seja, a maneira como filtramos os dados afeta o resultado; um exemplo, é o caso seguinte: se tivesse sido usado o método do tercil, um formando teria ficando como suplente e se tivesse sido usado, por exemplo, o método do quartil ou do quintil, teria ficado selecionado.

Tabela 6.1 – Comparação entre percentis e o seu impacto

	Habil.	**Qual. Prof**	**Entr. Gru. P.**	**Entr. Coord.**	**Res. Tes.**	**Total**	**Est.**
Tercil	3	1	3	0	3	10	Suplente
Quartil	3	1	3	0	4	11	Selecionado
Quintil	3	1	3	0	5	12	Selecionado

Legenda: Habil.: Habilitações; Qual. Prof.: Qualificação profissional; Entr. Gru. P.: Entrevista grupo pequeno; Entr. Coord.: Entrevista coordenador; Res. Tes.: Resultado testes; Est.: Estado

Por fim, ao criar-se uma ferramenta que permite saber se um formando irá ou não concluir o seu curso com sucesso em tempo real, o estudo acaba por fornecer informações valiosas para as pessoas que trabalham na seleção dos candidatos, ao poderem tomar uma decisão mais informada e segura. Assim, ter o auxílio desta previsão poderá ajudar a colmatar as falhas na seleção de potenciais formandos para frequentar um curso que iriam concluí-lo com pouco aproveitamento escolar. Para além do mais, esta ferramenta permite, também de forma atualizada, ter acesso a informação detalhada acerca dos cursos e dos seus desenvolvimentos.

6.2. LIMITAÇÕES

Deste modo, pode dizer-se que a maior limitação deste estudo foi a falta de dados, uma vez que para realizar uma previsão com resultados mais seguros e concretos é necessário ter uma base de dados mais robusta e com um maior número de registos de formandos, pois só assim será possível ter mais certezas acerca das variáveis que têm um maior impacto no resultado. Dado que os resultados são provenientes dos modelos de *machine learning*, dependem de uma quantidade significativa de dados para estabelecer padrões e quando tal não acontece, poderá haver dificuldade em generalizar corretamente as informações. Também um volume limitado de dados, poderá não representar todos os casos existentes, de forma suficiente, e formar um conjunto de dados desequilibrado, tal como foi o caso deste estudo que tinha poucos dados em relação aos formandos reprovados, o que torna a aprendizagem dos modelos mais difícil, ao identificar padrões menos comuns e, consequentemente, resultados menos precisos. Assim, será preciso continuar a alimentar a base de dados e a inserir novos dados para os resultados serem cada vez mais precisos e realistas.

6.3. RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Sugere-se para trabalhos futuros que, se possível, se acrescentem mais variáveis e dados à base de dados, por forma a ser possível obter uma análise ainda mais profunda. Para além disso, seria interessante melhorar ainda mais o conteúdo disponibilizado na *dashboard*, facilitando as visualizações, de acordo com as necessidades do utilizador e acrescentar mais parâmetros ou dados necessários. Também seria importante que a introdução de novos dados e de novas turmas na base de dados fosse cada vez mais eficaz, tornando o processo o mais automatizado possível, de modo a que seja possível mostrar as previsões, não só para um formando novo de cada vez, mas, por exemplo, para uma turma inteira de um determinado curso. A constante introdução de novos dados na base de dados, também possibilitaria que os resultados das previsões fossem cada vez mais precisos e fiáveis, pois a adição de novos dados faz com que os modelos de *machine learning* aprendam de forma mais acertada e com mais conhecimento, sendo esta questão também algo a melhorar em trabalhos futuros, nesta área de estudo.

Referências Bibliográficas

Ahmed, D. M., Abdulazeez, A. M., Zeebaree, D. Q. & Ahmed, F. Y. H. (2021). Predicting University's Students Performance Based on Machine Learning Techniques. *2021 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2021 – Proceedings*, pp. 276-281. https://doi.org/10.1109/I2CACIS52118.2021.9495862

Alsariera, Y. A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A. A. & Ali, N. (2022). Assessment and Evaluation of Different Machine Learning Algorithms for Predicting Student Performance. *Computational Intelligence and Neuroscience* (Vol. 2022). Hindawi Limited. https://doi.org/10.1155/2022/4151487

Alves, M. (2016). *Avaliação de Desempenho*. http://hdl.handle.net/10284/5797

Amrieh, E. A., Hamtini, T. & Aljarah, I. (2016). Mining Educational Data to Predict Student's Academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*, 9(8), pp. 119-136. https://doi.org/10.14257/ijdta.2016.9.8.13

Araújo, A. M. (2017). Sucesso no Ensino Superior: Uma revisão e conceptualização | | Success in Higher Education: A review and conceptualization. *Revista de Estudios e Investigación en Psicología y Educación*, 4(2), pp. 132-141. https://doi.org/10.17979/reipe.2017.4.2.3207

Azevedo, J. (2010). Escolas Profissionais: Uma história de sucesso escrita por todos. http://hdl.handle.net/10400.14/4698

Ballado, R. S., Morales, R. A. & Ortiz, R. M. (2014). Development and Validation of a Teacher Education Aptitude Test. *International Journal of Interdisciplinary Research and Innovations* (Vol. 2). www.researchpublish.com

Barbosa, B., Melo, A., Rodrigues, C., Amaral, C., Fernando, S., Gonçalo, C., Dias, P., Filipe, S., Traqueia, A. & Nogueira, S. (2019). Caracterização do Ensino e Formação Profissional em Portugal. https://www.edulog.pt/storage/app/uploads/public/5ee/94a/b74/5ee94ab7440cb365019630.pdf

Berrar, D. (2018). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* (Vols. 1-3, pp. 542-545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Cândido, A. R. (2020). A Influência do LinkedIn como Ferramenta Profissional no Recrutamento: Estudo do perfil dos utilizadores no setor do Turismo. http://hdl.handle.net/10174/28985

Cattell, A. & Cattell, R. (n. d.). Fator "G"

Centro Qualifica. (n.d.). Retrieved June 25, 2023, from https://www.qualifica.gov.pt/#/modalidades

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Choi, S. J., Jeong, J. C. & Kim, S. N. (2019). Impact of vocational education and training on adult skills and employment: An applied multilevel analysis. *International Journal of Educational Development*, 66, pp. 129-138. https://doi.org/10.1016/j.ijedudev.2018.09.007

Citeforma. (2021a). Apresentação. https://www.citeforma.pt/apresentacao.

Citeforma. (2021b). O que é? https://www.citeforma.pt/geral/o-que-e-centro-qualifica

Conceitos. (2015, October). Conceito de Teste de Aptidão. https://conceitos.com/teste-de-aptidao/

Cruz, M. (2009). Aptidões Básicas para informática

Dorsa, A. (2020). O Papel da Revisão da Literatura na Escrita de Artigos Científicos. *Interações (Campo Grande)*, pp. 681-684. https://doi.org/10.20435/inter.v21i4.3203

Gaspar, T., Tomé, G., Ramiro, L., Almeida, A. & Matos, M. G. (2020). Learning and Well-Being Ecosystems: Factors that Influence School Success. *Psicologia, Saúde & Doença*, 21(02), pp. 462-481. https://doi.org/10.15309/20psd210221

Guerreiro, N. M. dos R. F. (2014). O Impacto da Formação Profissional na Vida de Adultos com Baixa Escolaridade. http://hdl.handle.net/10316/26606

IEFP (n. d.). História. https://www.iefp.pt/historia

Jamba, I. (2018). Políticas e Práticas de Formação Profissional Continua: O Caso de um grupo de empresas de consultoria e engenharia. http://hdl.handle.net/10400.26/20873

Kagan, J. (2022, August 12). Aptitude Test: Definition, How It's Used, Types, and How to Pass. https://www.investopedia.com/terms/a/aptitude-test.asp

Leitão, I. (2013). Os Diferentes Tipos de Avaliação: Avaliação Formativa e Avaliação Sumativa. http://hdl.handle.net/10362/13803

Liashchynskyi, P. & Liashchynskyi, P. (2020). Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS. http://arxiv.org/abs/1912.06059

Lourenço, T. (2015). A Importância da Formação Profissional enquanto Investimento em Capital Humano. https://core.ac.uk/download/pdf/43583746.pdf

Lynn, N. D. & Emanuel, A. W. R. (2021). Using Data Mining Techniques to Predict Students' Performance. A Review. *IOP Conference Series: Materials Science and Engineering*, 1096(1), 012083. https://doi.org/10.1088/1757-899x/1096/1/012083

Maghawry, H., Yacoub, M. F., Helal, N. A., Gharib, T. F. & Ventura, S. (2022). An Enhanced Predictive Approach for Students' Performance. *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 13, Issue 4). www.ijacsa.thesai.org

Mais Formação (2015). Evolução dos Centros de Formação em Portugal. https://www.maisformacao.pt/evolucao-dos-centros-de-formacao-em-portugal/.

Mankar, J. & Chavan, D. (2013). Differential Aptitude Testing of Youth. *International Journal of Scientific and Research Publications* (Vol. 3, Issue 7). www.ijsrp.org

Neves, C. (2010). O Desempenho dos Profissionais da Formação Profissional: Um Estudo de Âmbito Regional.

Oberg, C. (n. d.). Guiding Classroom Instruction Through Performance Assessment. https://www.aabri.com/manuscripts/09257.pdf

Osmanbegovic, E. & Suljic, M. (2012). Data Mining Approach for Predicting Student Performance. *Economic Review: Journal of Economics and Business* (Vol. 10, Issue 1). http://hdl.handle.net/10419/193806

Pereira, J. (2018). Ensino Profissional: Escolha vocacional ou subterfúgio para jovens e adultos? http://hdl.handle.net/10400.13/1990

Pereira, J. & Carvalho, R. (2021). Ensino profissional: Escolha vocacional ou escapatória para jovens em transição? *Psychologica*, 64(1), pp. 49-67. https://doi.org/10.14195/1647-8606_64-1_3

Pimenta, C., Ribeiro, R., Sá, V., Paulo Belfo, F. & Politécnico de Coimbra, I. (2018). Fatores que Influenciam o Sucesso Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa (Factors Influencing School Success in Undergraduate Programs at a Portuguese Higher Education Institution). https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1015&context=capsi2018

Practice: Aptitude tests (n. d.). What Is An Aptitude Test? https://www.practiceaptitudetests.com/what-is-an-aptitude-test/.

Ramesh, V., Parkavi, P. & Ramar, K. (2013). Predicting Student Performance : A Statistical and Data Mining Approach. *International Journal of Computer Applications* (Vol. 63, Issue 8).

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A. & Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences (Switzerland)* (Vol. 10, Issue 3). MDPI AG. https://doi.org/10.3390/app10031042

Raza, M. & Shah, A. (2011). Impact of Favourite Subject towards the Scientific Aptitude of the Students at Elementary Level. *Pakistan Journal of Social Sciences (PJSS)* (Vol. 31, Issue 1).

Reis, C. (2018, March 4). Teste de aptidão: O que é e para que serve. https://www.e-konomista.pt/teste-de-aptidao/

Rodrigues, L. (2010). O Ensino Técnico-Profissional em Portugal. *Revista da Faculdade de Educação*, ano VIII, nº 2 (Vol. 14).

Rodríguez, D., Silva, J. A. & Bravo, L. E. (2022). Revisión sobre la Predicción del Rendimiento Académico Mediante Métodos de Ensamble. *Ingeniería Solidaria*, 18(2), pp. 1-28. https://doi.org/10.16925/2357-6014.2022.02.01

Setiawati, F. A. (2020). Aptitude Test's Predictive Ability for Academic Success in Psychology Student. *Psychological Research and Intervention*, 3(1), pp. 1-12. http://journal.uny.ac.id/index.php/pri

Silvestre, A. (2009). O IEFP e as Políticas de Formação Profissional: Passado e Futuro. http://hdl.handle.net/10773/3415

Sodhi, J. S., Dutta, M. & Aggarwal, N. (2016). Efficacy of Artificial Neural Network based Decision Support System for Career Counseling. *Indian Journal of Science and Technology*, 9(32). https://doi.org/10.17485/ijst/2016/v9i32/100738

Sumbo, H. B. (2019). Formação Profissional Contínua nas Organizações: Relevância do Diagnóstico de Necessidades de Formação. http://hdl.handle.net/10451/41332

Tasnim, N. E., Parvin, N., Barua, R., Rahman, M. M., Asrafe, M. H., Rahman, M. M. & Alam, K. K. (2022). Students' View about Attitude and Aptitude Test of Students in the Selection Process of MBBS Course in Bangladesh. *Bangladesh Journal of Medical Education*, 13(2), pp. 3-12. https://doi.org/10.3329/bjme.v13i2.60940

Thurstone, L. & Yela, M. (1985). Teste de perceção de diferenças

Thurstone, T. & Thurstone, L. (1984). Aptidões mentais primárias

Valeriu, D. (2015). Factors Generating of Positive Attitudes Towards Learning of the Pupils. *Procedia – Social and Behavioral Sciences*. Vol. 180, 5 May 2015, pp. 554-558. https://doi.org/10.1016/j.sbspro.2015.02.159

Vieira, M. & Azevedo, J. (n. d.). Fatores Que Promovem o Sucesso Educativo nas Escolas Profissionais. https://orcid.org/0000-0002-4986-7153

Watson, H., Hanlon, N. & Barquin, R. (2000). Journal of Data Warehousing. https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Widyahastuti, F. & Tjhin, V. (2017). Predicting Students Performance in Final Examination Using Linear Regression and Multilayer Perceptron. IEEE. *2017 10th International Conference on Human System Interactions (HSI)*, 188-192. doi: 10.1109/HSI.2017.8005026.

Xiao, W., Ji, P. & Hu, J. (2021). A Survey on Educational Data Mining Methods Used for Predicting Students' Performance. John Wiley and Sons Inc. https://doi.org/10.1002/eng2.12482

Yacoub, M., Maghawry, H., Helal, N., Soto, S. & Gharib, T. (2022). Predicting Students' Performance Using an Enhanced Aggregation Strategy for Supervised Multiclass Classification. *International Journal of Intelligent Computing and Information Sciences*, 22(3), pp. 124-137. https://doi.org/10.21608/ijicis.2022.146420.1195

York, T. T., Gibson, C., Rankin, S., York, T. T. & Gibson, C. (2015). Defining and Measuring Academic Success. *Practical Assessment, Research, and Evaluation*, 20. https://doi.org/10.7275/hz5x-tx03

Zeferino, A. & Passeri, S. (2007). Avaliação da Aprendizagem do Estudante (Vol. 3). https://files.cercomp.ufg.br/weby/up/148/o/AVALIACAO_DA_APRENDIZAGEM.pdf

Apêndice A

Lista dos anos, cursos e desempenho

Anos

- [ ] 2017
- [ ] 2018
- [ ] 2019
- [ ] 2020
- [ ] 2021

Cursos

- [ ] CET_Contabilidade
- [ ] CET_Redes
- [ ] CET_Multimédia
- [ ] CET_Turismo
- [ ] CET_Qualidade
- [ ] EFA_Programador
- [ ] EFA_Secretariado
- [ ] EFA_Logística
- [ ] EFA_Administrativo
- [ ] EFA_RHotel
- [ ] EFA_Vitrinismo
- [ ] CET_Programação
- [ ] EFA_Contabilidade
- [ ] CET_Contabilidade_II
- [ ] EFA_Administrativo_II
- [ ] CET_Contabilidade_PL
- [ ] EFA_Rececionista
- [ ] EFA_Comunicação
- [ ] EFA_Multimédia

Desempenho final de curso

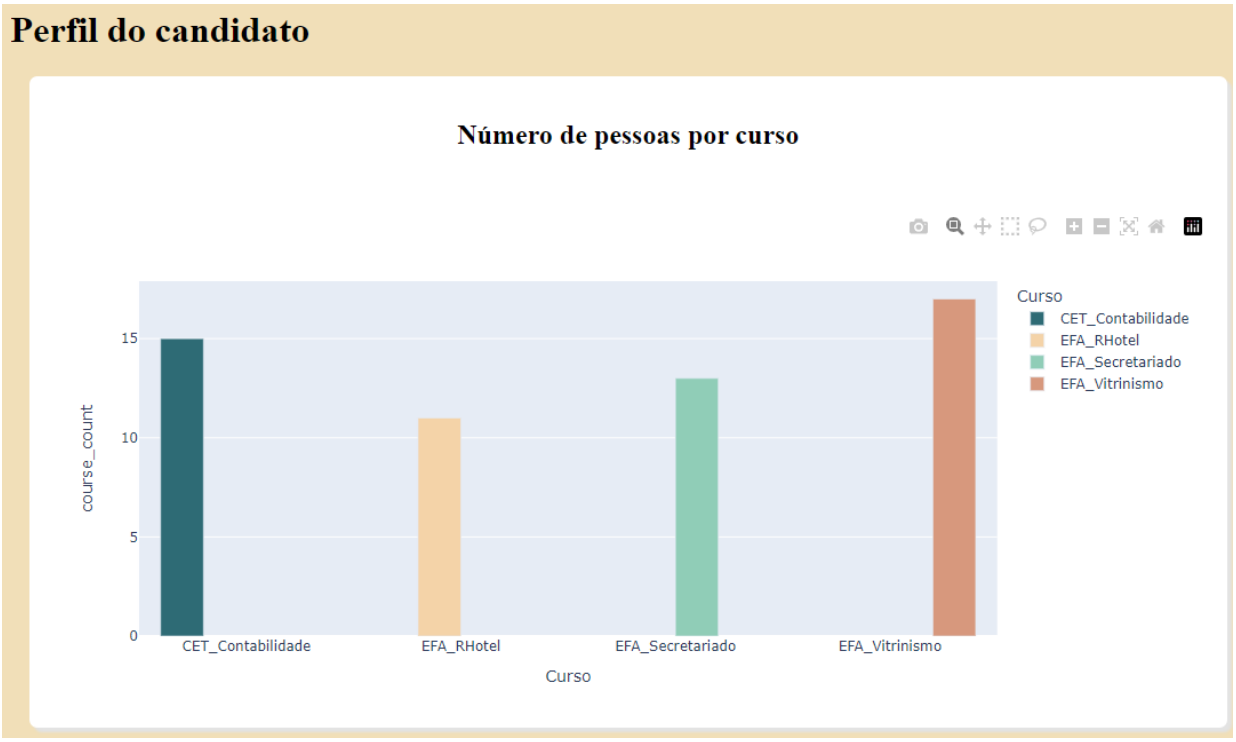
- [ ] Aprovado
- [ ] Desistente
- [ ] Reprovado
- [ ] Desistiu no processo de seleção

Apêndice B

Resumo das principias características dos dados

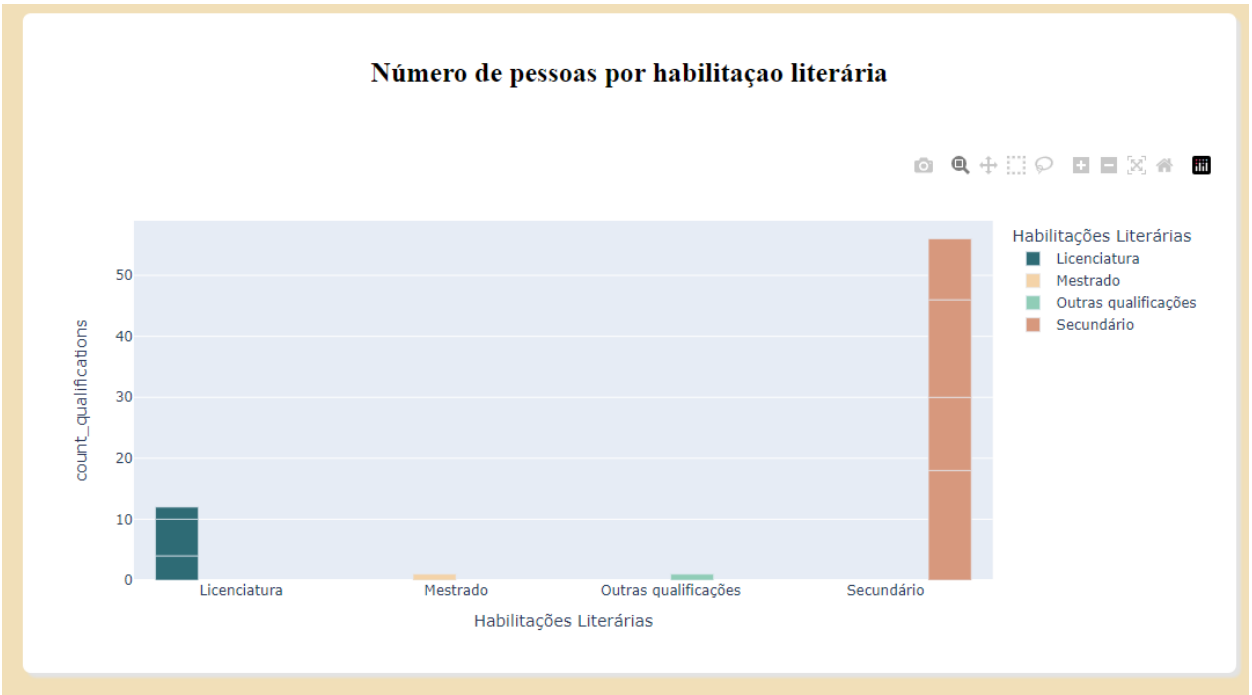
Apêndice C

Perfil do candidato



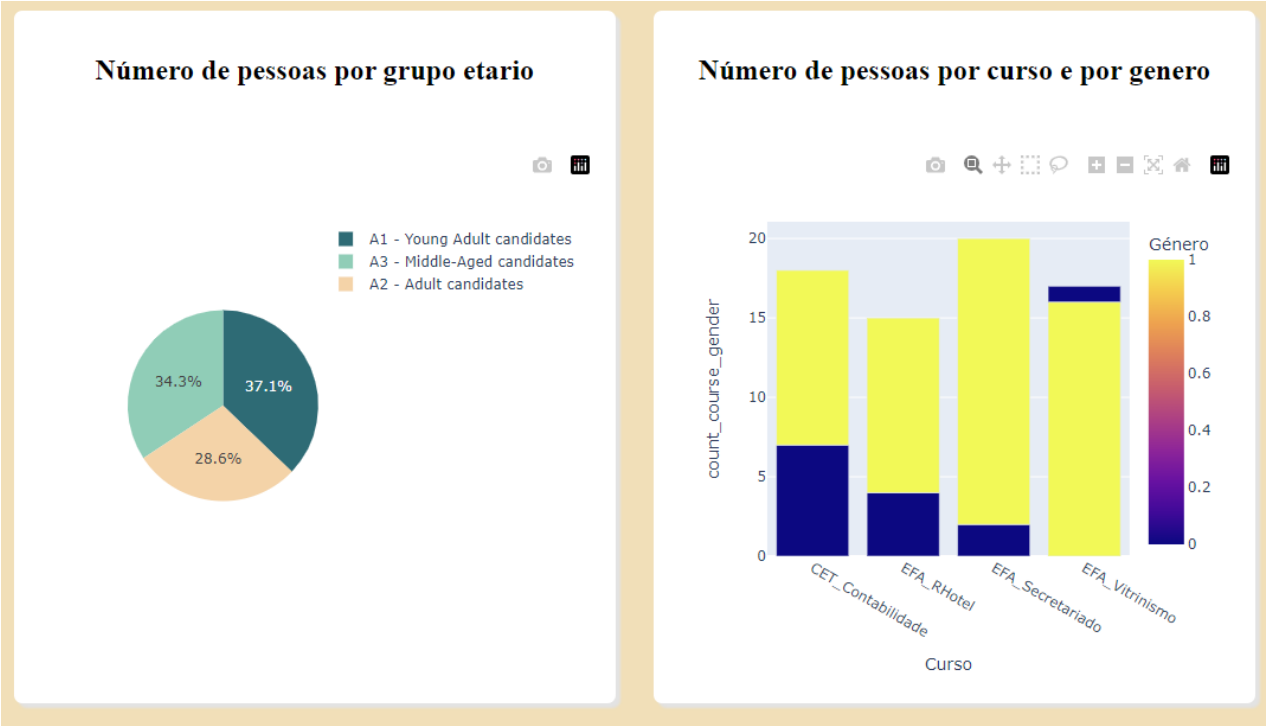
Apêndice D

Perfil do candidato



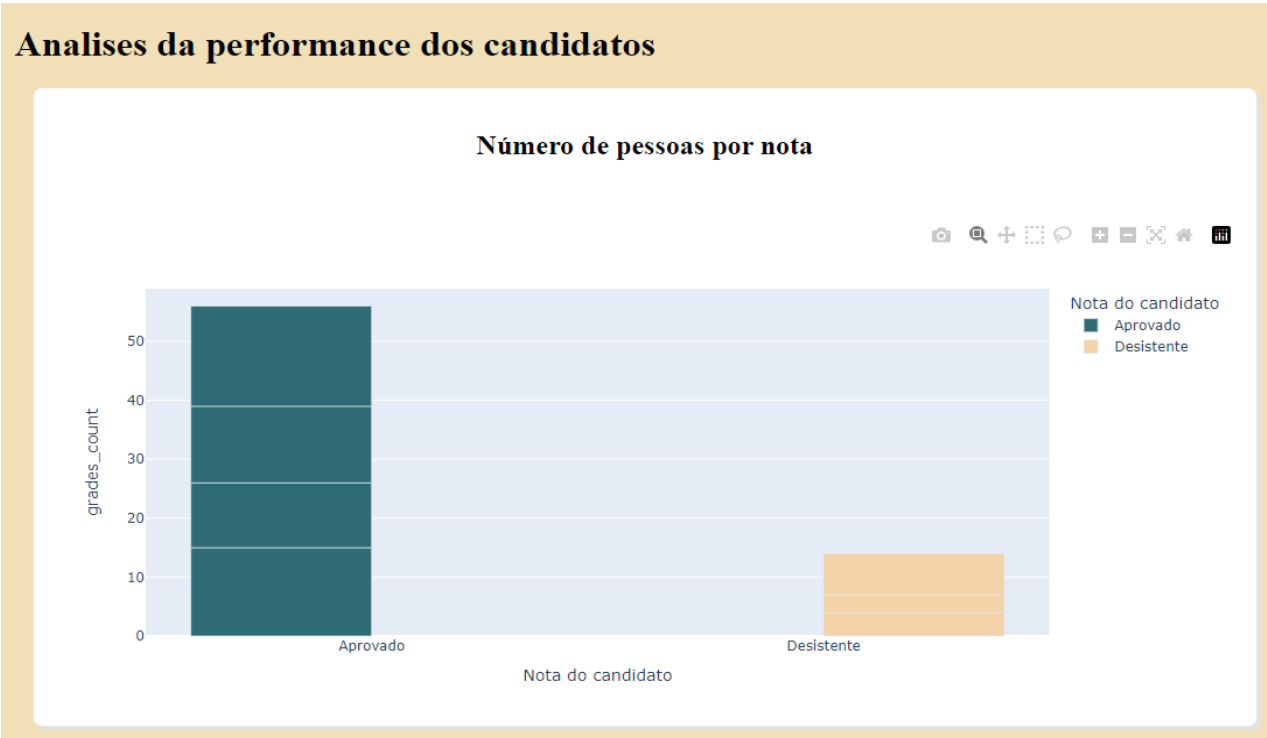
Apêndice E

Perfil do candidato



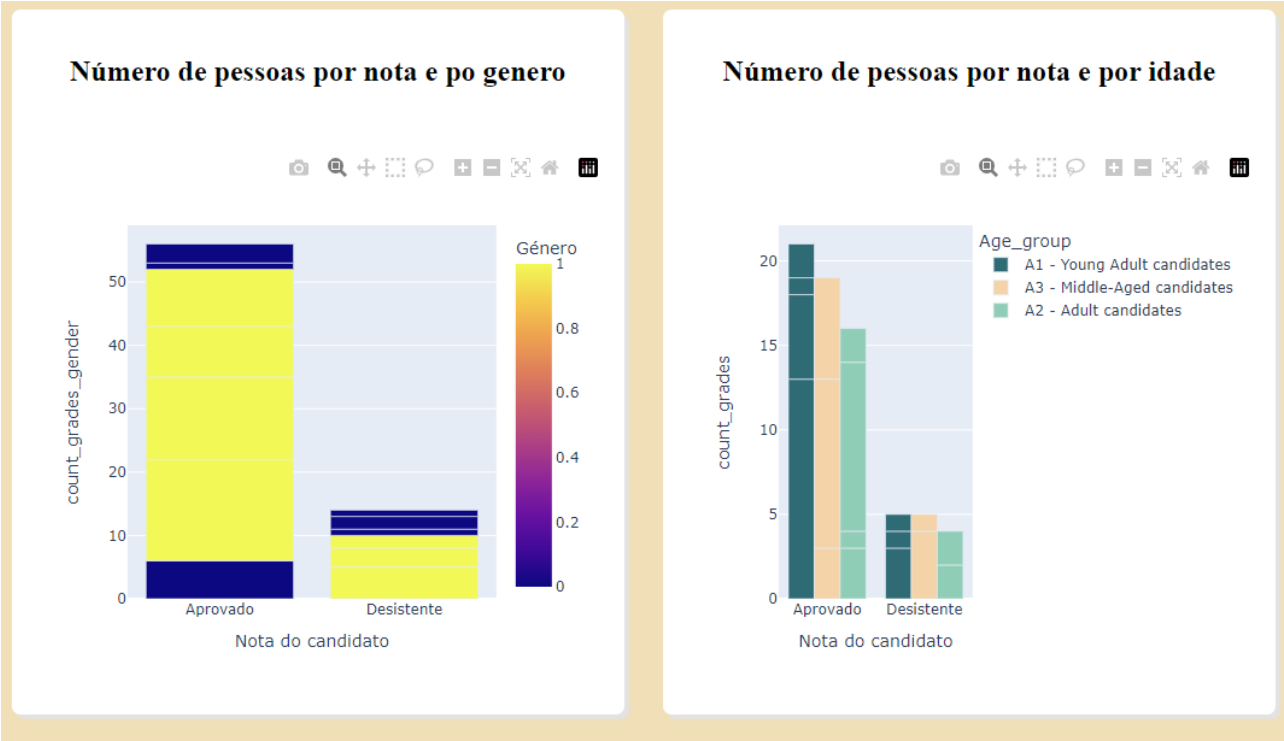
Apêndice F

Análises do desempenho dos candidatos



Apêndice G

Análises do desempenho dos candidatos



Apêndice H

Previsão

Apêndice I

Previsão

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa