INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Previsão de anulação de projetos financiados por fundos públicos

Alberto Neto Vilas

Mestrado em Gestão de Sistemas de Informação

Orientador:
Doutor Luís Miguel Martins Nunes, Professor Associado,
ISCTE-IUL

Coorientadora:
Doutora Ana Maria Carvalho de Almeida, Professora Associada,
ISCTE-IUL

outubro, 2021

TECNOLOGIAS
E ARQUITETURA

Departamento de Ciências e Tecnologias da Informação

Previsão de anulação de projetos financiados por fundos públicos

Alberto Neto Vilas

Mestrado em Gestão de Sistemas de Informação

Orientador:
Doutor Luís Miguel Martins Nunes, Professor Associado,
ISCTE-IUL

Coorientadora:
Doutora Ana Maria Carvalho de Almeida, Professora Associada,
ISCTE-IUL

outubro, 2021

Agradecimentos

A realização deste trabalho foi parcialmente financiada por fundos nacionais através da FCT - Fundação para a Ciência e Tecnologia, I.P. no âmbito dos projetos 01/SAMA2020/2019. (IA-Incentivos) e UIDB/04466/2020 (ISTAR).

Esta investigação contou com incondicionais apoios que tornaram possível a sua realização:

Ao meu orientador, Professor Luis Nunes, agradeço toda a sua orientação e dedicação ao longo deste percurso. Sempre se mostrou disponível para me ajudar com qualquer dúvida, fornecendo-me sempre os melhores conselhos para tornar esta investigação o mais rigorosa possível.

Á minha co-orientadora, Professora Ana de Almeida, pela sua prestimiosa ajuda e apoio.

À minha equipa de projeto, um sincero obrigado. Em especial, ao Professor Raul Laureano, por me ter dado a conhecer sobre a existência deste projeto e me ter motivado para fazer parte dele.

Aos meus pais por me terem proporcionado sempre o melhor que podiam e pelas ferramentas que me ensinaram e me ajudaram a tornar o homem que sou hoje.

Aos meus avós por todas as palavras de conforto e o carinho que me deram a vida toda.

Á minha namorada por ser a minha melhor amiga. Agradeço por todas as palavras encorajadoras que me ajudaram a nunca desistir e pela companhia ao longo deste percurso.

A todos os que mencionei, um sincero obrigado.

Resumo

A atribuição de fundos públicos é de extrema importância para o aumento de competitividade das empresas. A justiça na sua aplicação e a garantia da sua boa e completa utilização é uma preocupação constante das agências competentes, IAPMEI e AICEP. A análise dos dados das candidaturas por sistemas automáticos, pode ajudar a focar as fiscalizações em projetos de maior risco. O problema que será minimizado com este estudo será a previsão de projetos anulados no momento da candidatura. Os projetos anulados são aqueles que durante a sua execução são cancelados, podendo dar lugar a devoluções de valores monetário. Todos os projetos cativam o valor elegível, desde que são aceites até ao fim ou cancelamento da sua execução.

Esta dissertação apresenta um estudo usando os dados das candidaturas de projetos do IAPMEI. O objetivo foi criar modelos de previsão de anulação de projetos e identificação das principais características envolvidas nesta tarefa. Para além disto, está englobada toda a tarefa de extração e transformação de dados relativos a todos os ficheiros de um ciclo de vida de um projeto.

Por fim, os modelos de previsão das anulações que resultaram deste estudo foram integrados num protótipo que visa automatizar a tarefa de classificação dos projetos no momento da candidatura.

Através desta dissertação, as instituições que gerem os fundos públicos serão capazes de gerir melhor os fundos disponíveis de forma a otimizar a sua aplicação e criar mais oportunidades e maior eficiência para as empresas que usufruem dos mesmos.

Palavras-Chave: ETL, Big Data, Data Mining

Abstract

The allocation of public funds is extremely important to increase the competitiveness of companies. Fairness in its application and the guarantee of its good and complete use is a constant concern of the competent agencies, IAPMEI and AICEP. The analysis of application data by automated systems can help focus inspections on higher risk projects. The problem that will be minimized with this study will be the forecast of canceled projects at the time of application. Canceled projects are those that are canceled during their execution, which may give rise to refunds of monetary values. All projects captivate the eligible value, as long as they are accepted until the end or cancellation of their execution.

This dissertation presents a study using data from IAPMEI project applications. The objective was to create models to forecast cancellation of projects and identify the main characteristics involved in this task. In addition, the entire task of extracting and transforming data relating to all files in a project's lifecycle is included.

Finally, the cancellation prediction models that resulted from this study were integrated into a prototype that aims to automate the task of classifying projects at the time of application.

Through this dissertation, the institutions that manage public funds will be able to better manage the available funds in order to optimize their application and create more opportunities and greater efficiency for the companies that use them.

Keywords: ETL, Big Data, Data Mining

Índice Geral

Índice de Tabelas

Índice de Figuras

Glossário de Abreviaturas e Siglas

AICEP – Agência para o Investimento e Comércio Externo de Portugal

ANN – Feedforward Artificial Neural Network

AUC – Area Under the Curve

BPN – Back propagation networks

CRISP-DM – Cross-Industry Standard Process for Data Mining

CSV – Comma-separated values

DT – Decision Tree

EUA – Estados Unidos da América

IA – Inteligência Artificial

IAPMEI – Instituto de Apoio às Pequenas e Médias Empresas e à Inovação

IES – Informação Empresarial Simplificada

I&D – Investigação e Desenvolvimento

IoT – Internet of Things

KNN – K-nearest Neighbors

LR – Logistic Regression

MLP – Multilayer Perceptron

PME – Pequenas e Médias Empresas

RF – Random Forests

RFE – Recursive Feature Elimination

ROC – Receiver Characteristic Operator

SEMMA – Sample, Explore, Modify, Model, Assess

SI – Sistema Incentivos

SVM – Support Vector Machines

TI – Tecnologias de Informação

UE – União Europeia

XGBoost – Extreme Gradient Boosting

XML – Extensible Markup Language

Capítulo 1 – Introdução

1.1. Enquadramento do tema

Aumentar a competitividade das empresas portuguesas é um objetivo que tem cruzado várias décadas e vários governos. Para se obter esse aumento de competitividade é importante existir um apoio financeiro, que pode ser proporcionado por Sistemas de Incentivos geridos por instituições como o IAPMEI, Instituto de Apoio às Pequenas e Médias Empresas e à Inovação, ou a AICEP, Agência para o Investimento e Comércio Externo de Portugal. São estas as instituições a que as empresas recorrem para obterem apoios para o financiamento dos seus projetos.

Estes apoios são dados depois de uma análise criteriosa das candidaturas. Assim sendo, é importante indicar as fases que compõem o ciclo de vida de um projeto, sendo elas: candidatura, análise de candidatura, pedido de pagamento, análise de pedido de pagamento e fecho de um projeto.

Esta dissertação incidiu nas duas primeiras fases do ciclo de um projeto (candidatura e na análise de candidatura) utilizando-se os dados do IAPMEI.

Através do site do IAPMEI, é possível ver que existem três domínios de investimento que assentam nos respetivos domínios de desenvolvimento empresarial, cada um destes com objetivos distintos:

- Inovação e Empreendedorismo: promoção do empreendedorismo e incentivo à criação de novos negócios;
- Qualificação e Internacionalização de PME: promoção da competitividade e a promoção do aumento do desenvolvimento das PME para uma presença internacional;
- I&D: promoção da investigação e interoperabilidade entre empresas e o mundo científico (IAPMEI, 2020).

Com vista a criar um sistema automático de análise de projetos propostos a este sistema de incentivos, foi necessário fazer o tratamento dos dados fornecidos pelo IAPMEI.

Apesar dos dados estarem estruturados, foi preciso uniformizar os dados e tratar os ficheiros de forma a colocá-los no formato mais adequado para realizar o trabalho. De seguida, procedeu-se à sua extração. A terceira tarefa esteve relacionada com o

processamento de dados. Foi feito um tratamento com base nos dicionários de dados e nos requisitos que os próprios modelos exigem. Durante a quarta tarefa foram construídas variáveis complexas, isto é, variáveis que foram construídas apartir de operações entre uma ou mais variáveis simples, com vista a aumentar a probabilidade de obtenção de bons resultados. Para a escolha de quais abordagens utilizar para a construção dos modelos, foi importante produzir algumas estatísticas acerca das variáveis simples e complexas. Esta análise permitiu otimizar a tarefa de seleção de *features* para a modelação.

A quinta tarefa correspondeu à construção dos modelos com os dados previamente tratados, testando vários algoritmos e levantando as características mais importantes. Por fim, a sexta tarefa consistiu na avaliação dos modelos utilizando diversas métricas. Existiu uma sétima tarefa de construção do protótipo que não se encontra descrita nesta dissertação por limitação de dimensão e complexidade na explicação. Esse protótipo permite prever a classificação de qualquer novo projeto com base nos modelos construídos na quinta tarefa.

Estes modelos permitirão prever como será o desenrolar de um projeto, ou seja, se tratará complicações para as entidades que atribuem estes fundos. Através destes modelos os técnicos poderão ter uma análise muito mais completa, rápida e automática no momento de decisão sobre apoiar ou não um projeto de uma empresa. Existirá uma melhor monitorização ao longo de todo o projeto e as entidades estarão despertas para situações potencialmente problemáticas antes de elas acontecerem.

1.2. Motivação e relevância do tema

A integração de meios tecnológicos para auxiliar a gestão pública, *E-Government*, é uma mais-valia para a própria organização porque o investimento em tecnologias de informação está relacionado positivamente com o gasto eficiente do dinheiro público (Pang, 2014). Twizeyimana e Andersson (2019) verificaram que as adoções destes sistemas promovem uma maior eficiência administrativa, poupança de recursos, promoção de uma maior transparência, confiança nas instituições e contribuem para melhorias na prestação de serviços.

A implementação de um sistema que visa automatizar a análise de projetos a serem financiados é de extrema importância para o contexto do mundo atual, de modo a aumentar a eficiência neste processo.

Os projetos repartem-se em várias categorias e são destinados, no caso do IAPMEI, a Pequenas e Médias Empresas (PME). Uma PME, é uma empresa que emprega menos de 250 pessoas e que o seu volume de negócios não excede os 50 milhões de euros ou o seu balanço anual não excede os 43 milhões de euros (Decreto-Lei n.º 372/2007).

Segundo dados da PORDATA em 2017, estas representavam 99.9% do número total de empresas em Portugal. Devido a este elevado valor é possível identificar a necessidade económica em promover o crescimento e desenvolvimento das mesmas.

O desenvolvimento de PME está interligado com a obtenção de fundos (Ratnayake, 2014), portanto uma gestão competente por parte das autoridades responsáveis é imprescindível. Através desta gestão criteriosa consegue-se apoiar os projetos que mais necessitam e que serão mais promissores. Essa análise de escolha começa com a análise das candidaturas dos projetos aos diversos Sistemas de Incentivos (SI) e acaba com o fecho do projeto, podendo em qualquer momento haver auditorias aos projetos como forma de validar as informações prestadas.

A atribuição destes incentivos é de extrema importância. De facto, Ratnayake (2014) afirma que as PME desempenham um papel fulcral no desenvolvimento dos países. O desenvolvimento e aumento da competitivade das PME está interligado com a obtenção de financiamento.

1.3. Questões e objetivos de investigação

O objetivo desta investigação é perceber o entendimento da viabilidade de construção de modelos que prevejam o risco de anulação dos projetos no momento da candidatura. Isto vai permitir ajudar os técnicos no seu trabalho de análise de candidaturas a projetos.

As questões de investigação desta dissertação são:

1. Como é possível prever o risco da anulação de uma candidatura no momento da apreciação do projeto?
2. Quais as características que se devem usar para esta previsão?

1.4. Abordagem metodológica

1.4.1. Escolha da metodologia

Como esta investigação procura realizar uma extração de conhecimento a partir de dados (*data mining*) foi necessário perceber qual a metodologia mais indicada para esta problemática.

A metodologia escolhida foi a CRISP-DM (*Cross-Industry Standard Process for Data Mining*). O CRISP-DM foi uma metodologia criada para ser usada em projetos de *data mining*, e é vista como um standard para problemas desta área. O desenvolvimento da CRISP-DM começou em 1996 e foi lançada a sua versão final em 1999, versão essa que perdura até aos dias de hoje quase sem alterações. Os autores descrevem-na como independente de indústria e negócio, sendo por isso utilizável nas mais diversas áreas (Chapman et al., 2000). Alguns autores referem que esta é a metodologia mais usada e relevante na área de análise de dados. Para além disto, não é dependente de software, que faz com que se possa utilizar em qualquer projeto sem custos adicionais (Jaggia et al., 2020)

A framework aqui descrita contempla seis fases ao longo do seu ciclo e que são ligeiramente flexíveis. Cada ciclo pode ser visto como um processo em cascata, ou seja, as fases acontecem posteriormente ao terminus da anterior e o output de uma fase é o input da fase seguinte, dentro do mesmo ciclo. Isto levanta um problema que diz respeito à impossibilidade de paralelização de trabalho dentro do mesmo ciclo, visto que é necessário que a fase anterior termine para que a seguinte seja iniciada (Bošnjak et al., 2009). No entanto poderão estar vários destes ciclos a ocorrer ao mesmo tempo acontecendo uma parelelização de ciclos.

Existe outra metodologia usada para problemas de *data mining*, chamada SEMMA. Foi apresentada pela empresa SAS. O seu nome foi criado usando a inicial de cada nome de fase do seu processo: *Sample, Explore, Modify, Model e Assess.* A *framework* SEMMA, para além de não ser tão completa como a CRISP-DM, foi criada com o objetivo de ser utilizada conjuntamente com o *software* da empresa criadora, neste caso o SAS Enterprise Miner Software, ao contrário da CRISP-DM que é independente de qualquer ferramenta de *data mining* (Azevedo & Santos, 2008).

Na Tabela 1 são apresentadas as fases de ambas as metodologias:

Tabela 1 - Fases previstas das metodologias SEMMA e CRISP-DM

SEMMA	CRISP-DM
---	Compreensão do Negócio
Amostra	Compreensão dos Dados
Exploração	
Modificação	Preparação dos Dados
Modelo	Modelação
Avaliação	Avaliação
---	Instalação/Implementação

Como é possível ver na Tabela 1, a metodologia CRISP-DM é mais completa em termos de fases. Não existe qualquer fase para compreensão do negócio previsto na SEMMA, o que revela ser uma enorme fragilidade.

1.4.2. CRISP-DM – Descrição da metodologia

Esta metodologia inicia-se pelo processo de Compreensão do Negócio, como é possível ver na Figura 1, sendo esta a fase mais importante de todo o ciclo do projeto. É atribuída uma elevada importância a esta fase, visto que para se atingir o sucesso no projeto, um total entendimento do âmbito em que se insere é indispensável. É necessário entender na plenitude tudo o que diz respeito ao "negócio" em questão, para posteriormente existir um risco reduzido do projeto falhar.

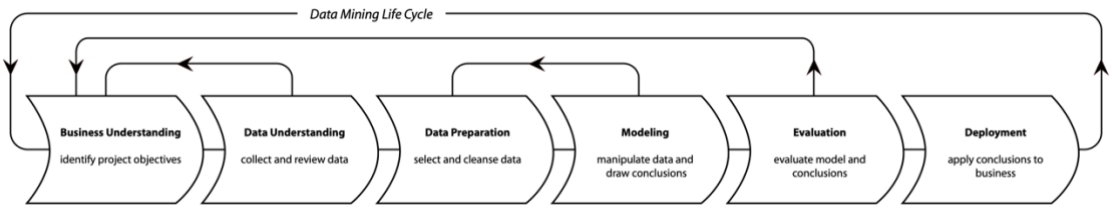


Figura 1 - Fases do CRISP-DM

Como se pode observar na Figura 1, os objetivos de cada fase do ciclo da CRISP-DM são:

1. Compreensão do Negócio – Entendimento total do âmbito e formulação de questões e objetivos do projeto. Para esta fase foram realizadas reuniões com os técnicos do IAPMEI. Para além destas reuniões, foi feita uma pesquisa bibliográfica para o aumento de conhecimento na área e um levantamento dos estudos sobre a temática desta dissertação.
2. Compreensão dos Dados – Recolha, descrição e exploração inicial dos dados a serem utilizados. Esta fase correspondeu à extração de dados partindo dos ficheiros em formato XML para os ficheiros CSV e a produção de estatísticas. Dada a complexidade dos dados esta última tarefa foi imprescindível para o sucesso nas tarefas seguintes.
3. Preparação dos dados – Seleção, tratamento e limpeza dos dados com vista a criar modelos na fase seguinte. Neste caso, foi realizada a codificação das variáveis, o tratamento de nulos, o tratamento de *outliers* e a escolha de *features* a utilizar na fase de modelação.
4. Modelação – Seleção de técnicas e criação dos modelos. Os algoritmos utilizados neste estudo foram Máquinas de Suporte Vectorial, XGBoost, Árvores de Decisão e Regressão Logística.
5. Avaliação – Avaliação dos modelos criados na tarefa anterior e a criação de possíveis otimizações, usando métricas habituais como F-Score e ROC AUC.
6. Instalação/Implementação – Planeamento e instalação do protótipo final

Destas fases, esta dissertação visa cobrir essencialmente as fases 3, 4 e 5: preparação dos dados, modelação e avaliação, respetivamente.

1.5. Estrutura e organização da dissertação

O presente estudo está organizado em cinco capítulos: introdução, revisão de literatura, metodologia, análise e discussão dos resultados e respetivas conclusões.

O primeiro capítulo introduz o tema da investigação e objetivos da mesma, bem como uma breve descrição da estrutura do trabalho.

O segundo capítulo reflete o enquadramento teórico, designado por revisão da literatura, onde são descritos os principais conceitos utilizados para esta investigação.

O terceiro capítulo é dedicado à metodologia utilizada nesta dissertação (CRISP-DM). Ao longo deste capítulo é descrito o trabalho realizado em cada fase bem como a construção de uma *baseline* para servir de base de comparação ao capítulo seguinte.

O quarto capítulo apresenta a análise dos resultados obtidos, de acordo com a metodologia que se entendeu apropriada.

No quinto e último capítulo apresentam-se as conclusões deste estudo, as principais contribuições, bem como as recomendações, limitações e trabalhos futuros.

Capítulo 2 – Revisão da Literatura

2.1. Pequenas e Médias Empresas

As PME são um pilar importantíssimo na economia de qualquer país, e um cuidado dedicado às mesmas é essencial.

Existe sempre um risco associado à candidatura de financiamentos de projetos, no entanto, a maneira como as PME respondem a estes riscos difere das grandes empresas. As empresas grandes têm mais recursos, mais experiência de projetos e maior capacidade de aguentar insucessos em relação a empresas de dimensão mais reduzida (Rosenbusch et al., 2011). No caso das PME, estas estarão sempre mais expostas ao risco e a obterem repercussões mais negativas.

Pissarides (1999) referencia que a aplicação de medidas na atribuição de fundos a PME em países na Europa central e oriental levou a um aumento de 250% no volume de negócios diretamente relacionados com PME, o que se conclui que as medidas referentes a fundos contribuíram para um crescimento destas empresas.

Existem evidências na literatura, que indicam existir uma relação positiva entre inovação e crescimento e/ou produtividade (Love & Roper, 2015). De facto, existe um estudo realizado sobre PME em Itália, que evidenciou uma relação bilateral entre produtividade e inovação. Concluiu-se que as empresas mais produtivas tendem a inovar mais, e as empresas que apostam mais na inovação geralmente são as mais produtivas. Também foi possível concluir que as empresas que optam por investir em investigação e criação de produtos novos ou pela aplicação de métodos inovadores nos seus processos, conseguem obter níveis de produtividade superiores às que não o fazem (Cainelli et al., 2006).

Os autores Lu & Beamish (2001), mostraram que a inovação é uma questão de sobrevivência para PME em mercados competitivos. Estes acreditam que as empresas de menor dimensão conseguiriam competir de forma mais eficiente se apostassem em produtos diferenciados e inovadores ao invés de participarem numa concorrência de preço, pois geralmente não conseguem acompanhar as capacidades das grandes empresas. Estes autores, conseguiram mostrar que a performance tem uma ligação direta com a inovação. As empresas que invistam mais em I&D tendem a obter melhores resultados, que podem ainda ser alavancados se a inovação não for exclusiva de um produto ou

patente, mas também de investimento em inovação nos processos e operações da própria empresa. Este aumento de performance afetará tanto empresas criadas há mais tempo como empresas recentes. No entanto esse aumento é maior em empresas recém-criadas, porque quanto mais recente é a empresa, maior é a sua capacidade de mudança e adaptação às condições exigidas. Assim, a maior flexibilidade leva a uma melhor capacidade de resposta e consequentemente obtenção de melhores resultados.

Tendo em vista o crescimento das empresas, o mercado nacional é reduzido para as ambições de algumas PME. A internacionalização surge como uma possível solução para se atingir esse crescimento, através da chegada a novos potenciais consumidores (Lu & Beamish, 2001). As PME que decidam optar por esta via, estarão menos expostas ao risco de falência do que as não-exportadoras (Esteve-Pérez et al., 2008). Quando uma empresa é exportadora e têm um produto inovador, o aumento de produtividade é ainda maior quando comparado com as não-exportadoras (Love et al., 2010). Porém, a entrada num novo mercado exige ultrapassar várias dificuldades que traz às PME a necessidade de adaptação e geração de conhecimento. Lu e Beamish (2001) sugerem a obtenção de parceiros locais para ultrapassar essa barreira principalmente em mercados tão diferentes dos mercados de origem da PME.

No caso dos investimentos em I&D, existe uma relação direta com a produtividade, e a exportação e inovação de uma empresa dependem do investimento feito nesta área (Love et al., 2010).

2.2. Inteligência Artificial no setor público

O conceito de inteligência artificial (IA), segundo Tecuci (2012), diz respeito ao desenvolvimento de sistemas que tenham um comportamento que pode ser associado a um comportamento tipicamente humano. Outra definição semelhante foi dada como a simulação de comportamento humano, aprendendo, racionalizando e procedendo sobre tarefas previamente definidas (Valle-Cruz & Sandoval-Almazan, 2018). Mais recentemente um autor prôpos a ideia de processos humanos que podem ser automatizados (Doorn, 2021).

O desenvolvimento de sistemas informatizados na gestão governamental é cada vez maior, assumindo assim uma maior importância nos orçamentos dos governos. Os países mais desenvolvidos apresentam-se como líderes na criação de *e-services*, ou seja,

providenciar serviços através de meios eletrónicos (Hassan et al., 2011). Projetos que visam promover o *e-government* têm elevadas taxas de insucesso. Os países em desenvolvimento são os que registam as maiores taxas de insucesso com 35% dos projetos com insucesso total e 50% com insucesso parcial (Mellouli et al., 2020). No caso dos projetos que recorrem a IA, estes seguem a tendência de serem projetos difíceis de atingirem o sucesso sem a existência de complicações. Existem quatro domínios em que é imprescidivel não serem tidos em conta ao implementar projetos com IA, sendo eles as leis e regulamentação, qualidade na implementação da tecnologia, ética e a aceitação da sociedade (Wirtz et al., 2019).

Para uma correta gestão *E-Government*, alguns autores propuseram um conjunto de quatro valores centrais a qualquer serviço criado nesta área:

- Eficiência - gestão eficiente do dinheiro público de modo a haver o menor desperdício possível.
- Utilidade dos serviços - maximização da utilidade dos serviços oferecidos.
- Profissionalismo - providenciar uma gestão independente, robusta e consistente e que esteja em conformidade com as leis em vigor.
- Envolvimento - presença dos cidadãos e melhoria de comunicação com os mesmos (Rose et al., 2015)

Em 2018, na Suécia realizou-se um estudo que visava analisar documentos que continham políticas orientadoras para setor público e privado na temática da IA. Neste âmbito foram encontrados 10 documentos. A análise foi feita com *text mining* com o intuito de retirar informação das frases/*statements* desses documentos. Do total de 522 frases em 10 documentos, 281 referiam benefícios do uso de IA e apenas 50 falavam nos seus riscos, o que indicava boas perspetivas de futuro nesta temática. Posteriormente foram analisadas as frases com base nos valores ideais de uma gestão *e-governement*, com base no modelo proposto por Rose em 2015. Este modelo define como valores ideais a eficiência, o profissionalismo, o envolvimento e a utilidade dos serviços. Assim, o valor profissionalismo foi o que mais foi detetado nos documentos, por outro lado, o envolvimento, que é o valor que diz respeito à presença e comunicação com e para o cidadão, foi o que menos apareceu nas frases, revelando assim ser o valor do modelo que menos foi tido em conta nessas políticas orientadoras. Face a estes resultados é importante

ter em conta a necessidade de envolvimento das populações na administração pública e não descurar este valor (Toll et al., 2019).

A IA pode ser aplicada a diversas áreas da gestão pública, nomeadamente à área da segurança social. Nos Estados Unidos da América, foi implementado um sistema de auxílio na decisão de atribuição de benefícios sociais aos cidadãos (Pang, 2014). Com esta implementação, verificou-se uma maior deteção de casos de declarações fraudulentas para obtenção desses benefícios. Para além desta deteção, houve uma maior rapidez na atribuição de benefícios e uma poupança de recursos, promovendo assim uma gestão eficiente do dinheiro público.

Em 2018, os EUA permitiram pela primeira vez uma campanha que visava publicitar um produto na área da saúde que recorria a algoritmos de IA. Tinha como objetivo a deteção de retinopatia diabética, uma doença do foro ocular que leva a perda parcial ou total da visão. Esta doença é a maior causadora de perda de visão entre a população americana, no entanto, se for detetada em fase precoce poderá ser feito o devido acompanhamento que permitirá atrasar a evolução da doença. O dispositivo chama-se "IDx-DR" e utiliza um algoritmo de inteligência artificial para analisar fotos tiradas pelo médico ao olho do utente. Se o resultado for positivo o doente está a desenvolver a doença (*Food & Drug Administration*, 2018). Na saúde os algoritmos utilizados têm menor erro de diagnóstico do que o ser humano, no entanto existem diversos problemas legais e éticos que impedem a sua maior presença nos serviços de saúde (Ho et al., 2019). Alguns algoritmos tendem a ter comportamentos desviantes que não permitem com que sejam tidos em conta em diversos casos, exemplo disso foi o algoritmo com comportamentos racistas (Angwin et al., 2016) . Este algoritmo foi criado com o objetivo de classificar pessoas com o risco de cometerem crimes futuros, numa escala de 0 a 10. O comportamento deste algoritmo foi melhor do que uma simples previsão aleatória, verificando-se 67% de acerto na identificação de cidadãos que iriam cometer um crime futuro. Porém, houve um grave enviesamento, que atribuía um risco maior de vir a cometer um crime um indivíduo da raça negra do que qualquer outra raça. O sistema atribuía o risco com base em um conjunto de perguntas (137), mas nunca era perguntado a raça da pessoa. A possibilidade da utilização deste algoritmo foi automaticamente boicotada devido a questões éticas e legais.

No âmbito militar, a utilização de sistemas inteligentes é uma realidade desde o século passado. Um bom exemplo disto são as armas que permitem perseguir ou atingir um alvo

facilmente, mesmo que este se mova. Também podemos falar do reconhecimento de padrões. Para além disto, com os avanços mais recentes, já é possível exercer o controlo total dos aviões de guerra através de algoritmos de inteligência artificial (Payne, 2018).

No caso das *smart cities,* através da *Internet of Things* (IoT), grandes volumes de dados são gerados todos os dias, e com o uso de IA é possível extrair informação desses dados (Kankanhalli et al., 2019). O setor dos transportes poderia ser um dos maiores beneficiados com esses dados, no entanto existem poucas aplicações desta área. (Soe & Drechsler, 2018), sugerem uma aplicação que permite a monitorização em tempo real do tráfego rodoviário. Com o uso desta aplicação, numa situação de acidente haverá uma melhor gestão desse incidente, para além disso, gerir o próprio tráfego traria benefícios como a criação de menos congestionamentos.

A IA e os seus recentes avanços levaram países com elevada expressão na economia mundial a dedicarem fatias importantes dos seus orçamentos a projetos nesta área, criando assim um ambiente propício para a criação de mais e melhores projetos. Os EUA gastaram em 2016 a quantia de 1.2 mil milhões de dólares, a China prevê investir até 2030 um total de 147 mil milhões de doláres e a Europa gastou até ao momento 700 mil milhões de euros. Estes investimentos são justificados por existirem previsões que mostram que a IA pode ser responsável pelo aumento de 2% da economia mundial (Wirtz et al., 2019). Para além disto, os países podem conseguir obter uma poupança de 100 mil milhões de euros se investirem em IA bem como em Big Data (Agbozo & Spassov, 2018). O Reino Unido em 2015, decidiu acordar uma parceria público privada em que cedeu dados de 1.6 milhões de utentes do serviço nacional de Saúde, para a criação de uma aplicação de IA que visava detetar a doença renal em fases precoces (Ballantyne & Stewart, 2019).

Estes sistemas no setor público são descritos como eficientes, precisos, de elevada performance e com uma aplicabilidade nos vários domínios (Alexopoulos et al., 2019), e é visto como uma solução para os governos conseguirem manter uma elevada qualidade na tomada de decisão com o minímo de custos (Pang, 2014). Existem ainda outras vantagens identificadas como o combate à corrupção, melhoria na comunicação entre governo e cidadãos, personalização de serviços e prevenção e resposta a desastres (Valle-Cruz et al., 2019). Na temática da tomada de decisão, utilizar Inteligência Artificial evita o “cognitive bias”, ou seja, um enviesamento cognitivo que acontece frequentemente na tomada de decisões em grupo ou individuais (Stone et al., 2020).

2.3. Financiamento de PME

Na Europa, as PME têm um peso enorme na economia europeia. De facto, estas representam 60% do valor criado, 70% de postos de trabalho e 99% dos negócios criados. O acesso a financiamento é de extrema importância para estas conseguirem desenvolverem-se e atingirem os objetivos esperados (Artola & Genre, 2010).

Como já foi referido, o acesso ao financiamento é mais difícil nas PME do que nas grandes empresas. Para além dos financiamentos serem menos diversificados a maior fonte de financiamento das PME são as instituições bancárias. Identificou-se que as PME europeias recebem cinco vezes menos apoios financeiros do que as PME nos EUA. É ainda referido que uma maior diversidade de financiamentos leva a uma maior resistência a crises financeiras (Kraemer-Eis, & Lang, 2017)

De acordo com alguns autores, a maioria destas empresas obtêm financiamento internamente ao invés de recorrerem a outras formas de financiamento (Demary et al., 2016). Em países como Portugal, Espanha e Itália, a obtenção de grande parte do seu financiamento é através de empréstimos bancários. Tendo por base que empresas de menor dimensão quando recorrem a esta forma de obter dinheiro acabam por ter condições piores que empresas grandes, percebe-se que a dependência destes financiamentos é prejudicial para as PME dos três países referidos anteriormente. Para estas empresas o problema é ainda maior porque, segundo (Artola & Genre, 2010), existem duas razões para além da dependência bancária que dificultam o crescimento das PME. A primeira razão deve-se ao facto de o acesso à informação ser menor e de pior qualidade e, portanto, os projetos de investimento acabam por ter condições piores. A segunda razão está relacionada com o tempo de vida da empresa, ou seja, quanto mais recente for a empresa menor será a sua exposição/reputação que consequentemente pode ser refletida no pouco interesse por parte dos investidores.

Analisou-se o impacto do acesso a fundos estruturais da União Europeia em PME numa zona sub-desenvolvida da Polónia. Através do acesso a este tipo de fundos, os postos de trabalho criados por PME passaram de 58% do total de emprego em 1995 para 70% em 2008 (após a entrada da Polónia na UE e consequentemente ter acedido a fundos estruturais). Concluíram assim que os fundos têm um impacto direto na criação de emprego e, portanto, no desenvolvimento de zonas menos desenvolvidas (Lewandowska et al., 2015).

Em 2015, num artigo comparativo entre empresas inovadoras e não inovadoras, percebeu-se que empresas inovadoras, ou seja, que criam produtos ou processos novos, tendem a pedir mais financiamento como forma, inicialmente, de suportarem o processo de criação da própria inovação. Porém, as empresas inovadoras são as que encontram maiores dificuldades na obtenção de fundos face às empresas não inovadoras. Para além destas conclusões, concluíram ainda que este entrave se deve a três possíveis razões. A primeira razão é o risco associado ao investimento numa inovação. Para além disso, a análise do valor de uma inovação é complicada e exige especialistas que nem sempre se encontram disponíveis. Por último as inovações poderão ser apenas aplicáveis em contextos específicos, como por exemplo, o contexto em que foram criadas e isso reduzirá bastante a probabilidade de sucesso das mesmas (Lee et al., 2015).

Para além deste estudo de empresas inovadoras, em 2020, existiu um conjunto de autores que se focaram em perceber qual o impacto do financiamento público na inovação tecnológica de empresas em Portugal e outros dois países. No caso de Portugal, o financiamento europeu ou nacional a PME tinha forte influência na criação de tecnologia, indicando também que empresas que conseguiam adquirir bens e serviços de alta tecnologia teria mais condições para criar inovações. Para além disto, levantam um problema que diz respeito à entrega de financiamentos. Segundo esta publicação, o financiamento nem sempre é dado a quem teria as melhores condições para o receber. As empresas que teriam as melhores potencialidades para virem a ter sucesso eram deixadas de parte e os financiamentos eram entregues a empresas que teriam melhores relações com as agências que atribuem esses financiamentos (Anderson et al., 2020)

2.4. Risco de crédito em PME

Para solucionar o problema da previsão de incumprimento de projetos, foi importante perceber o que tinha sido realizado no passado para uma melhor construção da solução. Porém, para o problema em questão, existe uma carência de literatura. Assim, optou-se por alargar a pesquisa e incluir a análise do risco de crédito a empresas deste setor, PME. Na análise destes estudos era importante perceber quais as abordagens utilizadas e os seus resultados.

Em primeiro lugar, importa definir o conceito de risco de crédito que consiste na possibilidade de o financiador não vir a receber o valor acordado nos termos previamente decididos, o financiador pode ser qualquer entidade que forneça crédito (Twala, 2009).

Outra definição dada, define como a possibilidade da não existência futura de dinheiro suficiente para pagar aos credores (Belás et al., 2018). A diferença para o problema desta dissertação é que, ao contrário de um risco de crédito, as empresas recebem financiamento, mas não terão que devolver valor, ao invés disso, devem cumprir com o projeto previamente acordado, provocando o mínimo de atrasos possíveis. Poderá haver situações excecionais que resultam na devolução total ou parcial do dinheiro recebido.

No estudo de risco de crédito em PME, percebeu-se que estas empresas têm a maior taxa de incumprimento de empréstimos (Kalayci & Arslan, 2017). As PME sofrem ainda de um problema acrescido pois encontram-se mais expostas a mudanças económicas do que as grandes empresas (Kim & Sohn, 2010). Neste sentido, existe uma maior preocupação na atribuição de fundos, justificando-se o tratamento individualizado deste tipo de empresas. Empresas com mais crédito, ou seja, maior valor em dívida, tendem a ter um maior risco de incorrerem em incumprimento (Belás et al., 2018).

2.5. IA e análise de risco de crédito

Foi feito um levantamento de estudos na área de análise de risco de crédito em PME usando algoritmos de IA para prever o risco de crédito neste tipo de empresas, como se pode observar na Tabela 2.

Tabela 2 - Artigos acerca de Risco de Crédito

Artigo	Notas importantes	Melhores Resultados accuracy
(Chen & Li, 2009)	O tamanho do conjunto de dados para teste era muito reduzido	SVM – 87.5% BPN – 75%
(Kim & Sohn, 2010)	SVM superou BPN	SVM – 66.16%
(Dereliolu & Gürgen, 2011)		KNN – 80.66% SVM – 75.39% (sem seleção de features)
(Zhu et al., 2016)	ANN superou os modelos híbridos para deteção de casos negativos	ANN superou LR na deteção dos casos negativos: 64.5%(ANN) VS 47.9%(LR)
(Zhu et al., 2017)	DT superou os métodos embedded	DT (C4.5) – 79.58% RandomSubpace-Boosting – 85.41%
(Kalayci & Arslan, 2017)	RF superou MLP	RF – 82.25%
(Gulsoy & Kulluk, 2019)		Multi Objective Evolutionary Fuzzy – 78.85 %

Como é possível reparar existem poucos estudos exclusivos nesta temática, no entanto existem conclusões importantes a retirar, nomeadamente o uso das árvores de decisão, em inglês *Decision Tree* (DT) e *Random Forest*. Com o uso destes algorítmos obteve-se

bons resultados, com 82.25% de acerto num estudo realizado em 2019 (Kalayci & Arslan, 2017). Existiu um estudo em 2009 (Chen & Li, 2009), que dava bons indicativos da performance de *Support Vector Machines*, com 87.5% de acerto. No entanto, é necessário existir algum cuidado ao analisar este estudo devido ao número reduzido de amostras. A amostra apenas contava com 8 empresas. O número reduzido de exemplos nos conjuntos de dados torna difícil a generalização de resultados.

Uma nota importante é que os algoritmos de redes neuronais, nomeadamente BPN, MLP e ANN, não obtiveram melhores resultados, como é observável em alguns dos estudos, portanto o uso destes algoritmos não seria um bom ponto de partida (Chen & Li, 2009; Kalayci & Arslan, 2017; Kim & Sohn, 2010; Zhu et al., 2017)

Existe uma limitação no uso de redes neuronais que diz respeito a não existir interpretabilidade nos modelos criados, e no contexto desta investigação isso seria um problema, dado que a previsão de aceitabilidade dos projetos subentende uma explicação para os projetos que não são aceites.

Capítulo 3 – Metodologia

3.1. Compreensão do problema

Uma das fases previstas para esta investigação passa pela compreensão do negócio. Assim, para esta fase foram realizadas reuniões com os técnicos do IAPMEI. Para além destas reuniões, foi feita uma pesquisa bibliográfica para o aumento do conhecimento na área e um levantamento dos estudos já realizados sobre a temática.

Com o contacto com os técnicos percebeu-se que o objetivo principal na atribuição de incentivos a projetos é dar um impulso às empresas para permitir que se desenvolvam e se tornem mais competitivas. Os incentivos são escassos e é da responsabilidade das instituições competentes, IAPMEI e AICEP, fazerem a melhor gestão possível dos fundos disponíveis. No entanto, os projetos nem sempre decorrem como é suposto, por vezes os projetos são anulados durante o seu tempo de execução. Estas anulações para além de gerarem procedimentos burocráticos morosos e consequentemente custos, criam um problema maior que é a cativação de fundos.

Este problema é particularmente grave porque impede outras empresas de acederem a crédito e concretizarem alguns dos seus projetos, porque o valor disponível é limitado e devido a essa razão existem empresas que não tiveram os seus projetos aceites em detrimento de projetos que vieram a ter um final precoce devido a uma anulação.

As anulações dão lugar maioritariamente a devoluções de valor, mas existem algumas situações em que esse valor não é devolvido, o que é especialmente prejudicial. Existem três tipos de classificações de anulações, segundo o IAPMEI:

- Anulação pós-contrato
- Anulação por caducidade
- Anulação por desistência do promotor

Nesta dissertação foram tratadas apenas as situações de anulações pós-contrato, porque são as situações de anulação que existem em maior número no conjunto de dados. As empresas que tiveram anulação de projetos pós-contrato geraram uma cativação de dinheiro durante o tempo de execução do projeto.

Os dados utilizados neste estudo dizem respeito ao Sistema de incentivos de Inovação e Empreendedorismo, um dos três Sistemas de Incentivos do Portugal2020 geridos pelo

IAPMEI. O conjunto total de dados tem um total de 2792 projetos, desde o ano 2014 até ao ano 2019.

3.2. Compreensão dos dados

Os dados fornecidos dos projetos apresentavam a seguinte organização: cada ficheiro corresponde a um documento de uma fase de projeto, podendo existir vários documentos para uma fase. Estes ficheiros foram fornecidos no formato XML, ou seja, *Extensible Markup Language*.

Para além destes dados foi necessário pedir ao IAPMEI uma listagem de projetos anulados, que veio no formato de Tabela excel. Este documento incluía todos os projetos anulados e continha as seguintes colunas:

- N_projetos – número projeto
- Tipo – indicava o tipo de anulação: pós/pré-contrato
- Motivo – motivo da anulação (descrição)

Dentro das colunas apresentadas, a coluna "Tipo" mostrou-se relevante pois foi a que permitiu selecionar apenas os projetos anulados em situação de pós-contrato, todas as outras situações foram descartadas.

3.3. Preparação dos dados

Os dados fornecidos pelo IAPMEI apresentavam-se num formato pouco adequado à exploração. Assim, nesta terceira fase, foi necessário criar programas automáticos que trataram da estrutura dos ficheiros.

Posteriormente, foi necessário construir um modelo de dados para apoiar a tarefa de extração, este modelo não se encontra nesta dissertação dada a complexidade dos dados e para evitar expor a estrutura de dados interna das agências em demasia. O modelo de dados continha a representação das 73 tabelas da fase de candidatura e das respetivas colunas, bem como as ligações entre elas.

Partindo deste modelo construíram se vários *scripts* em *python* para tratar e extrair os dados. Como se pode observar na Figura 2, cada caixa representa um *script* no ciclo de extração de projetos.

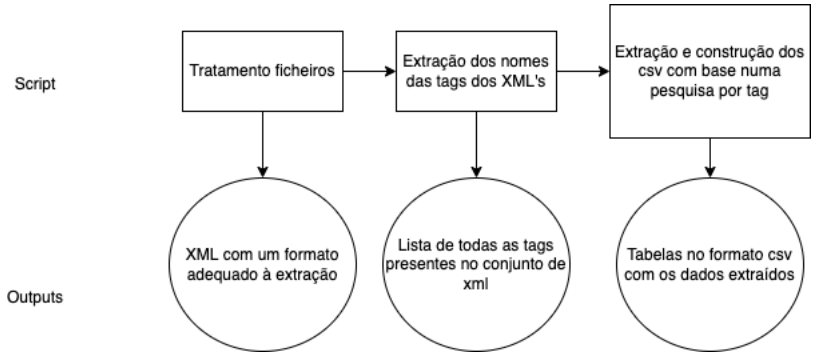


Figura 2 - Descrição do processo de extração de dados dos ficheiros XML

Este ciclo foi repetido para os 5 tipos de ficheiros: candidaturas, análise de candidatura (faci), pedidos de pagamento (ppi), análise pedidos de pagamento (appi) e fecho de projeto (facie). Foi necessário criar um *script* intermédio para extrair apenas os nomes das *tags/features* dos XML por duas razões: eficiência de extração e para uma melhor gestão da falta de informações nos XML. A falta de informação é explicada pela não existência das mesmas *tags* em todos os ficheiros, o que gera um problema na construção da tabela correspondente.

Quando todos os ficheiros se encontraram com uma estrutura semelhante iniciou-se a extração de dados, convertendo-se os XML em ficheiros CSV. Para esta tarefa criou-se um procedimento automático, que permitiu então ler os ficheiros no formato de origem e convertê-lo para linhas em tabelas com um formato aproximado a um modelo relacional. Neste conjunto de tabelas a ligação entre informações do mesmo projeto é feita através de uma chave que é um par com o número do projeto e o número do documento. A única exceção de chave são as informações da candidatura porque a cada projeto apenas corresponde uma e somente uma candidatura, e, portanto, nas tabelas correspondentes à candidatura, a chave é composta apenas pelo número do projeto.

3.4. Análise exploratória dos dados

3.4.1. Exploração inicial

Para um melhor entendimento da dimensão deste problema, é crucial observar-se as estatísticas na Tabela 3:

Tabela 3 - Contagem anulações

Tipo de Anulação	Nº de Projetos	Percentagem de casos
Sem anulação	2297	82.2%
Anulação do pós-contrato	329	11.7%
Anulação por caducidade	101	3.6%
Desistência do Promotor	65	2.3%

É possível concluir que cerca de 12% dos projetos são anulados numa situação de pós-contrato.

Para a comparação da distribuição entre projetos anulados e projetos sem anulação valor elegível foi criada uma tabela como se deve observer na Tabela 4. Assim, utilizou-se o valor correspondente à variável "Dadosprojeto/Elegível" de cada projeto, presente nos documentos de análise de candidaturas (vulgarmente chamados "faci"). Quando este valor não estava disponível, utilizou-se a variável "Elegivel" da Tabela das anulações.

O valor de investimento dos projetos distribuídos por ano observa-se na Tabela 4:

Tabela 4 - Distribuição do total de valor elegível por ano

Ano	Tipo_Anulacao	Valor_Elegivel	Percentagem no ano
2014	Anulação pós-contrato	56161398,17	-
	Sem anulacao	0	-
2015	Anulação pós-contrato	290024473,6	17,04%
	Sem anulacao	1411741627	82,96%
2016	Anulação pós-contrato	151463817,6	19,22%
	Sem anulacao	636445665,9	80,78%
2017	Anulação pós-contrato	129389558,3	12,99%
	Sem anulacao	866360545,9	87,01%
2018	Anulação pós-contrato	25193472,27	2,37%
	Sem anulacao	1036623678	97,63%
2019	Sem anulacao	121784972,8	-

Para os projetos do ano de 2014 não foi possível obter o valor elegível usando o método previamente identificado. Por outro lado, com os dados até ao momento, não existem projetos com ano de candidtura em 2019 anulados, tal como a percentagem de projetos anulados no ano 2018 podem indicar que é cedo para tirar conclusões sobre esse

ano. Os anos de 2015 e 2016 foram particularmente maus no que toca a sucesso de projetos, nestes dois anos 17% e 19% do valor atribuído foi dado a projetos que viriam a ser anulados.

3.4.2. Seleção de projetos a utilizar nos testes

A escolha dos projetos a utilizar no estudo teve que ser feita de forma criteriosa de forma a evitar problemas, tais como a colocação de projetos na amostra como "não anulados" quando na verdade ainda não se sabe o resultado final dos mesmos (não existindo facie).

A amostra será composta por projetos anulados pós contrato e projetos terminados com previsão de finalização até 2019. Com isto, ficaram de fora projetos que não tenham terminado ou tenham terminado após 2019 de forma a evitar possíveis enviesamentos. É possível ver a distribuição de projetos que irão pertencer à amostra com base no ano de previsão de finalização na Tabela 5.

Tabela 5 - Distribuição de projetos por ano de candidatura

Ano Candidatura	**Anulados**	**Sem anulação**	**Total**
2015	2	4	6
2016	14	78	93
2017	119	277	396
2018	120	94	214
Não identificados	73	0	73
Total	328	453	781

Existem 73 projetos anulados que não tinham nas suas candidaturas dados sobre o ano de previsão de finalização, mas foram adicionados à amostra. A amostra contém 328 projetos anulados e 453 projetos que decorreram sem problemas, num total de 781 projetos.

3.4.3. Análise Exploratória

Os 781 projetos utilizados na amostra dizem respeito, em grande parte, a projetos da zona norte e centro do país e essas zonas dizem respeito a projetos com um maior número de anulações.

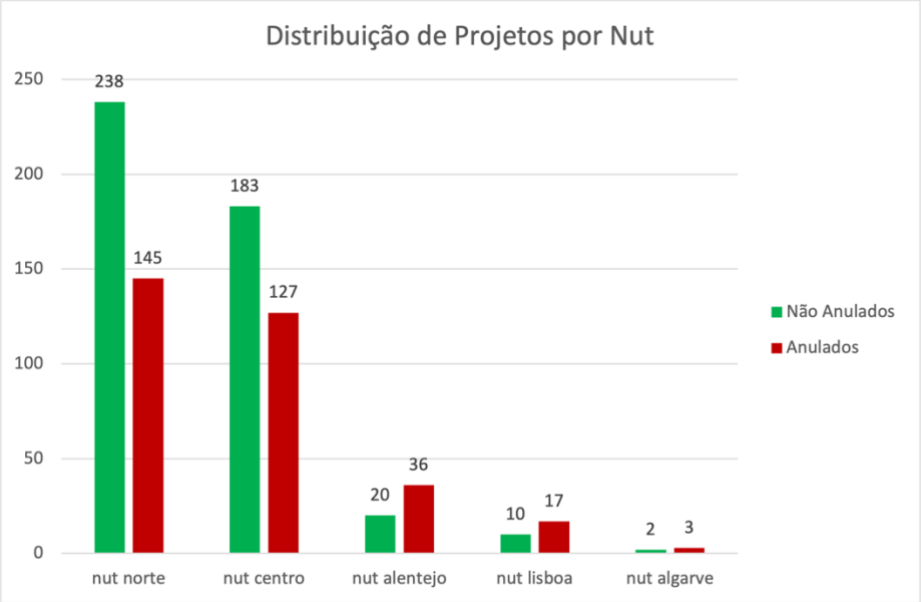


Figura 3 - Distribuição dos projetos por NUT

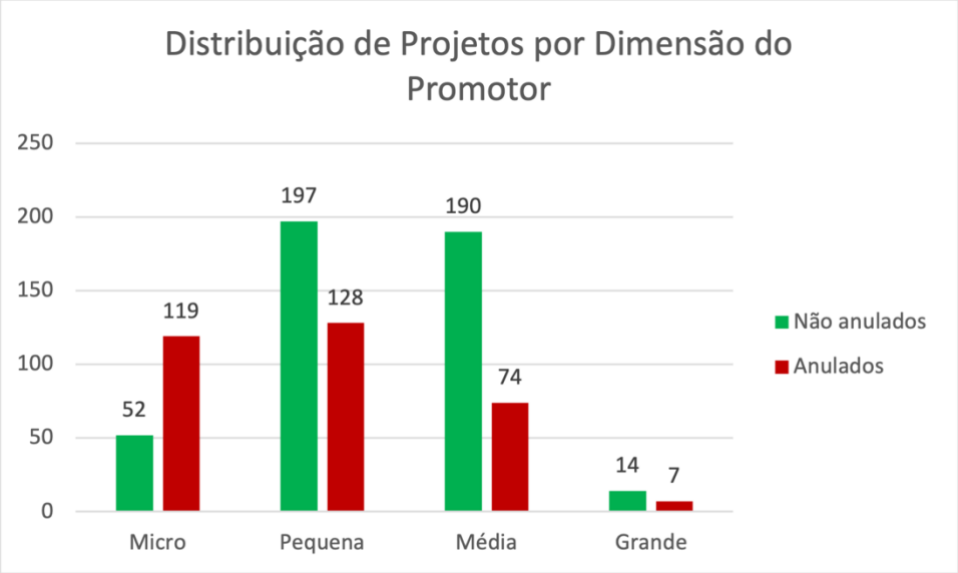


Figura 4 - Distribuição dos projetos por dimensão do promotor

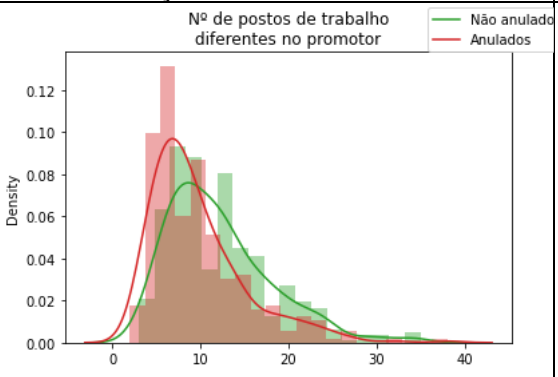
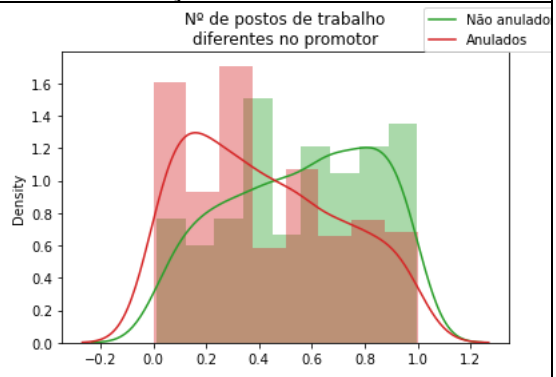
Na distribuição de projetos aceites por dimensão de empresa, existe uma elevada incidência de projetos de PME.

No caso dos projetos de microempresas existe um elevado número de projetos que viriam a ser anulados, com uma incidência maior do que noutras classes. Assim, é percetível que quanto maior a empresa é, menor é a percentagem de projetos anulados.

Tabela 6 - Informações das empresas promotoras

Informações das empresas promotoras	Média	Desvio Padrão	25%	50%	75%
Idade na candidatura	18,08	15,64	-1	5	16
Capital Social	620998	1371082	25000	136669	550955
Nº de postos de trabalho diferentes	45.23	67,13	7	23	56

Tabela 7 - Representação da variável "nº de postos de trabalho diferentes"

Representação gráfica da variável "Nº de postos de trabalho diferentes"	
Sem eliminação de outliers	Com eliminação de outliers
	

Para a representação com eliminação de *outliers* optou-se por utilizar o pré processamento "quantile transformer" que reduz a dimensão da variável para um intervalo entre [0, 1]. O valor do 1º quantil fica o mínimo da variável, ou seja, zero (0), todos os valores inferiores ficarão zero, e acontece o mesmo no 3º quantil ficando este com valor máximo um (1) tal como os valores superiores. Desta forma preserva-se o comportamento da variável dentro do 1º e 3º quantil, rejeitando os outliers da função. Sempre que exista uma representação gráfica ao longo desta dissertação sem *outliers*, recorreu-se a este tratamento.

Esta variável foi construída com a contagem de postos de trabalho diferentes que os promotores colocaram na candidatura. Pode observar-se uma clara tendência para os projetos com maior número de postos de trabalho diferentes terem menos anulações, o que confirma que projetos de empresas maiores têem menos probabilidades de serem anulados.

Com base nas informações recolhidas existem alguns dados que é interessante analisar nomeadamente as informações dos projetos.

Tabela 8 - Informações estatísticas sobre o conjunto de projetos

Informações projetos	Média	Desvio Padrão	25%	50%	75%
Duracao Prevista (meses)	20.32	5.31	17	24	24
Valor Elegível (euros)	1902498	2767886	519243	1011324	2046793

Analisando a Tabela 8, observa-se que a média de duração dos projetos é cerca de 20 meses com 75% dos projetos a terem uma duração prevista acima dos 17 meses. Projetos de curta duração, menos de um ano, são poucos no conjunto de dados. Na variável do valor elegível, o valor médio foi 1 902 498€, mas é importante referir que com um valor tão elevado de desvio padrão (2 767 886€) é difícil retirar conclusões deste valor.

Tabela 9 - Existência de informações acerca de algumas variáveis nos projetos

Variavel	Descrição	Sim (existe)	Não (não existe)
analise_mercados_dir	Direção de crescimento no mercado	561	220
major_plano_acao	Plano de ação para o projeto	237	544
major_sustenta	Sustentabilidade no projeto	483	298
despesas_idt	Despesas IDT previstas projeto	451	330

A variável análise de mercado indica a opção tomada pelo promotor na escolha de direção de mercado, analisando a distribuição desta variável parece haver indícios que a falta de direção de crescimento no mercado na candidatura poderá trazer complicações ao projeto.

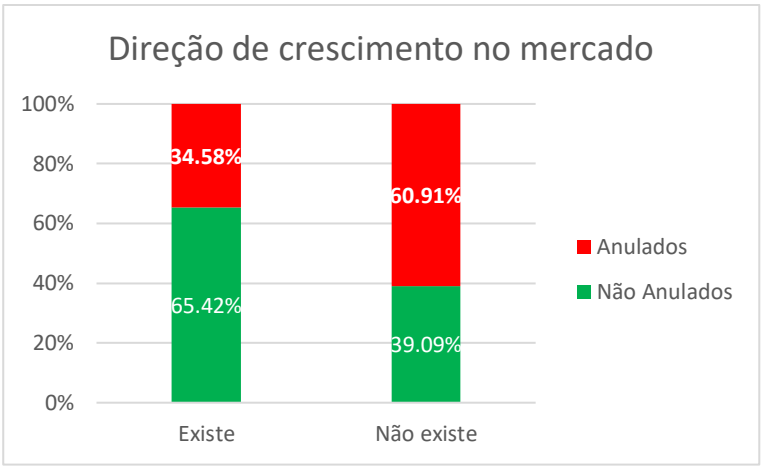


Figura 5 - Distribuição da existência de direção de crescimento no mercado definida no momento da candidatura

Existindo uma direção de crescimento do mercado indicada na candidatura a percentagem de anulações foi 35%. No caso dos promotores que não assinalaram nenhuma direção de crescimento a percentagem de projetos anulados subiu para 60% neste conjunto.

Nas outras variáveis não se verifica essa diferença clara e, portanto, não foram analisadas individualmente.

3.5.Modelação

3.5.1. Descrição da experiência inicial

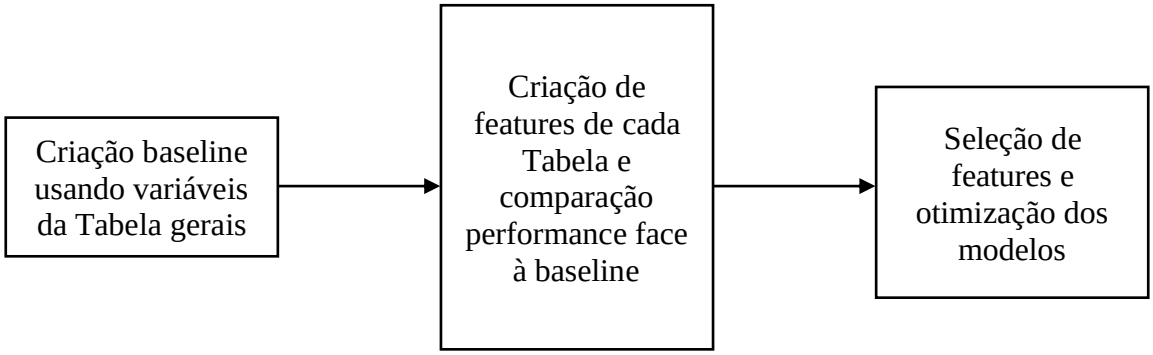


Figura 6 - Esquema representativo da experiência inicial

O trabalho seguiu o esquema representado na Figura 6. Inicialmente, foi criada uma *baseline* usando variáveis da tabela com as informações gerais. Posteriormente essa *baseline* foi usada para comparar com as variáveis das tabelas auxiliares acerca do projeto. Por fim, foram selecionadas as melhores *features*, e procedeu-se à otimização dos modelos.

As tabelas auxiliares dizem respeito a conjuntos de informações associadas a um projeto, todas essas tabelas são geradas em redor de uma tabela que contêm as informações gerais da candidatura.

Analisando as contagens de projetos das tabelas, percebe-se que algumas mostram inviabilidade de serem consideradas por conterem um número extremamente reduzido de projetos presentes. Por isso, optou-se por utilizar todas as tabelas que tenham pelo menos 10% dos projetos da amostra presentes, que corresponde a cerca de 80 projetos. Na Tabela 57 no anexo, encontram-se as contagens de todas as tabelas e a respetiva descrição. Na Tabela 10 apresenta-se a listagem das tabelas que irão ser usadas.

Tabela 10 - Tabelas auxiliares usadas nas experiências

Tabela	Nº de Projetos
AmbitoNov	561
Atividades	781
Balanco_SNC	781
CadeiaValor	561
Cursos	104
DesafiosSocietais	781
DominioPrioritario	765
DR_SNC	781
FormadoresExt	102
Formandos	96
FSE	233
IndicadoresIDT	781
IndicCertf	632
Inv	781
ListaAccoes	551
MarcasOutras	177
MarcasProprias	393
Mercados	781
Mercados2	781
Mercados3	781
ProjCae	781
PromCae	781
PromLocal	781
PromLocalTip	615
Propensao	166
PTrabalho	781
Reforço	166
ResultadosPO	781
Socios	781
TipoInov	561
Tipologia	781
VantagensCompA1	561
VendasExt	185

Como já foi referido, as tabelas foram agrupadas e usando esses grupos de tabelas foram criadas variáveis que foram comparadas com a baseline.

3.5.2. Criação da Baseline

A criação de uma *baseline* para qualquer estudo é de extrema importância pois é esta que permite a possibilidade de comparações durante as diferentes fases do estudo. Para construir esta linha de partida é necessário primeiramente selecionar e construir variáveis da tabela. Assim, criou-se os seguintes critérios:

- Foram escolhidas todas as colunas da Tabela que tinham valores válidos, ou seja, que não contenham um elevado número de nulos e que não fossem do tipo textual.
- Localização: colocou-se a localização expressa em NUT porque as outras colunas são demasiado específicas (distrito, concelhos, ruas, etc).
- As variáveis do grupo Descfisicaemp/ são mais de 90% nulos
- As variáveis do grupo "critselb1" tem um elevado número de nulos (50%) e essas informações encontram-se presentes, de uma forma mais completa, nas Tabelas auxiliares, respetivamente "mercados" e "n_clientes"
- A variável cae apresenta 202 valores diferentes, sendo uma variável categórica torna-se impossível de utilizar nos modelos de forma a não criar um enviesamento de resultados

Todas as variáveis escolhidas na experiência estão numa Tabela em anexo (Tabela 58), tal como os respetivos pré-processamentos utilizados em cada variável (Tabela 59). As condições utilizadas nesta experiência foram:

Utilização de undersampling para contrariar a tendência da amostra

Conjunto de treino (80%) e teste (20%) e random_state = 80

Os resultados poderão ser consultados na íntegra nos anexos (Tabela 60 e 61), tal como os resultados para o teste k-fold com k=10 (Tabela 62).

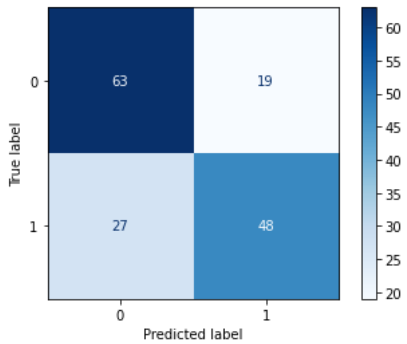


Figura 7 - Matriz de confusão para o modelo LR

Logistic Regression foi o melhor algoritmo a detetar os casos 1, ou seja, os casos dos projetos que viriam a ser anulados. Cobrindo 64% dos casos com classificação 1. Dos 67 projetos identificados como anulados pelo modelo, 48 eram realmente anulados, teve uma precisão elevada destes casos.

3.5.3. Otimização da Baseline

Tendo em conta o conjunto de variáveis da baseline ser demasiado extenso, e potencialmente estar a prejudicar a performance dos modelos, foi necessário selecionar *features*. Essa seleção foi feita com base nas *features* mais importantes para os modelos LR, SVM e Random Forest, visto terem sido os que tiveram uma melhor performance. Com base na importância de *features* destes modelos, presente nos anexos, selecionou-se as 12 variáveis mais importantes para cada um destes três modelos, e estão representadas na Tabela 11.

Tabela 11 - Conjunto de features selecionadas para a otimização

Features Escolhidas	Descrição
idade_candidatura_1	Idade do promotor no momento da candidatura: -1 a 3 anos
n_meses_2	Previsão da duração do projeto(meses): 8.88 a 14.79
n_meses_3	Previsão da duração do projeto(meses): 14.79 a 20.70
n_meses_4	Previsão da duração do projeto(meses): 20.70 a 26.61
n_meses_5	Previsão da duração do projeto(meses): 26.61 a 32.53
nut_norte	Localização promotor: Norte
nut_centro	Localização promotor: Centro
nut_lisboa	Localização promotor: Lisboa
nut_alentejo	Localização promotor: Alentejo
am_dir_0	Não tem direção de crescimento no mercado associada
am_dir_2	Direção de crescimento no mercado: extensão do produto
am_dir_3	Direção de crescimento no mercado: extensão do mercado
am_dir_4	Direção de crescimento no mercado: diversificação
vant_comp_est_2	Natureza das vantagens competitivas: diferenciação global
vant_comp_est_4	Natureza das vantagens competitivas: concentração em diferenciação
capital_social_1	Valor do capital Social do promotor(euros): 100 a 10000
capital_social_2	Valor do capital Social do promotor(euros): 10000 a 75000
capital_social_5	Valor do capital Social do promotor(euros): 756000 a 17500000
major_sustenta	Majorações: plano de ação para demonstração e disseminação
major_plano_acao	Majorações: O investimento enquadra-se na majoração "sustentabilidade"
prom_url_empresa	O promotor tem website
consultora	Existência de consultora associada ao projeto

Tabela 12 - Resultados dos modelos na tarefa de otimização da baseline

	ROC	Accuracy
LR	0.74	0.75
SVM	0.71	0.72
GNB	0.71	0.72
RF	0.67	0.68
DT	0.63	0.63
XGB	0.63	0.64
MLP	0.62	0.62

Os resultados desta experiência estão presentes na Tabela 12, com estes testes, utilizando um número reduzido de *features*, os resultados foram melhores do que os resultados iniciais, para a maioria dos algoritmos, e, portanto, optou-se por utilizar este conjunto reduzido de variáveis como *baseline*. Os resultados k-fold 10 encontram-se nos anexos na Tabela 62, bem como a importância de features nas Tabelas 63.

A variável "n_meses_5" foi retirada do conjunto de variáveis porque o seu coeficiente de importância foi 0 (zero) nos algoritmos SVM, LR, RF e XGBoost.

Esta variável tinha sido colocada devido à importância que tinha no algoritmo RF e no decorrer destes testes com o conjunto reduzido de features esse mesmo algoritmo apresentou um coeficiente zero para esta variável.

3.5.4. Utilização de variáveis das Tabelas auxiliares

Com foi referido anteriormente, as tabelas auxiliares foram agrupadas da forma apresentada na Tabela 13:

Tabela 13 - Conjuntos de Tabelas auxiliares criados para a experiência

Grupo	Tabela	Descrição
1	AmbitoNov Atividades DesafiosSocietais IndicadoresIDT TipoInov Tipologia	Informações Projeto
2	Cursos FormadoresExt Formandos FSE ListaAccoes	Formações
3	MarcasOutras MarcasPropria ProjCae PromCae PromLocal PromLocalTip PTrabalho Socios	Informações Promotor
4	DominiosPrioritarios DominiosPrioritariosAlentejo DominiosPrioritariosAlgarve DominiosPrioritariosCentro DominiosPrioritariosLisboa DominiosPrioritariosNorte	Domínios Prioritários
5	Mercados Mercados2 Mercados3 Propensao VendasExt	Vendas Externas
6	CadeiaValor Reforço ResultadosPO IndicCertf VantagensCompA1	Pré e pós projeto
7	Balanco_SNC DR_SNC	Informação contabilística
8	Inv	Investimento Projeto

O grupo 7 e 8 não foram trabalhados neste estudo devido à complexidade dos dados e por uma opção de limitação do âmbito.

Foram criadas variáveis em cada grupo e testadas separadamente entre elas, mas juntas com as variáveis da experiência 3.5.3. De seguinda, apresentam-se os valores de auc-roc e accuracy para o mesmo conjunto de treino e teste de cada experiência.

Tabela 14 - AUC com a inclusão das variáveis dos diferentes grupos. A cor escura estão os casos em que houve melhoria em relação à baseline.

	Area Under the Curve (AUC ROC)							
	Baseline	Grupo 1	Grupo 1 Otim	Grupo 2	Grupo 3	Grupo 4_5	Grupo 6	Grupo 6 Otim
LR	0.74	0.73	0.75	0.78	0.72	0.72	0.69	0.69
SVM	0.71	0.71	0.75	0.73	0.69	0.73	0.71	0.71
GNB	0.71	0.72	0.73	0.70	0.74	0.70	0.71	0.69
RF	0.67	0.67	0.74	0.71	0.70	0.73	0.71	0.72
MLP	0.63	0.66	0.69	0.68	0.70	0.66	0.63	0.67
XGB	0.63	0.70	0.72	0.68	0.61	0.66	0.73	0.69
DT	0.62	0.66	0.58	0.68	0.63	0.67	0.61	0.64

Tabela 15 - Accuracy com a inclusão das variáveis dos diferentes grupos. A cor escura estão os casos em que houve melhoria em relação à baseline.

	Accuracy							
	Baseline	Grupo 1	Grupo 1 Otim	Grupo 2	Grupo 3	Grupo 4_5	Grupo 6	Grupo 6 Otim
LR	0.75	0.72	0.71	0.79	0.72	0.73	0.69	0.69
SVM	0.72	0.71	0.71	0.74	0.69	0.73	0.71	0.71
GNB	0.72	0.68	0.73	0.71	0.74	0.70	0.69	0.69
RF	0.68	0.70	0.69	0.71	0.70	0.73	0.71	0.73
MLP	0.63	0.60	0.66	0.68	0.70	0.66	0.63	0.69
XGB	0.64	0.68	0.70	0.65	0.61	0.66	0.73	0.69
DT	0.62	0.61	0.66	0.69	0.63	0.68	0.61	0.64

No Grupo 1 e o Grupo 6 repetiu-se a experiência selecionando as 12 variáveis dos modelos com melhor performance e verificou-se que a redução do número de variáveis melhorou o desempenho, são os casos com “Otim”. No caso do grupo 4 e 5 o número de variáveis era bastante reduzido e optou-se por juntar e testar na mesma experiência.

Pode-se observar os resultados dos modelos e perceber que em alguns dos casos houve melhorias nomeadamente no grupo 2, que são as variáveis que dizem respeito a informações ligadas a formações, a adição destas variáveis fez aumentar consideravelmente as prestações dos modelos. Estes resultados dão indícios que poderão ser variáveis bastante importantes para a classificação de uma candidatura de um projeto.

Na generalidade das experiências a adição das novas features permitiu aumentar ligeiramente a performance dos modelos.

3.5.5. Otimização dos modelos

Optou-se por realizar outro tipo de experiência, com a seguinte ordem de trabalhos dividido em 3 fases:

1. Modelos sem otimização de híper parâmetros e com todas as variáveis
2. Modelos com otimização de híper parâmetros e sem seleção de features
3. Modelos com otimização de híper parâmetros e com seleção de features

Primeiramente a intenção foi perceber qual a performance destes modelos sem qualquer otimização e usando todas as variáveis. Utilizando um algoritmo de *Grid Search* executou-se a fase 2 para otimizar híper parâmetros. Posteriormente, selecionou-se as 12 melhores variáveis dos 3 melhores algoritmos e executou-se a fase 3 com os hiperparâmetros otimizados e as melhores variáveis selecionadas.

As variáveis escolhidas para a experiência 3, poderam ser consultadas na Tabela 16 da página 36.

Tabela 16 - Variáveis escolhidas para a tarefa de otimização de hiperparametros

Varíaveis escolhidas	Descrição
n_meses_2	Previsão da duração do projeto(meses): 8.88 a 14.79
n_meses_4	Previsão da duração do projeto(meses): 20.70 a 26.61
am_dir_0	Os projetos não contêm informação de analise de mercadosxs
capital_social_1	Pertencer ou não ao 1º intervalo de valores de capital social
ativ_tipo_2	Existência de tipologia de atividades – aumento da capacidade de um estabelecimento já existente
tipo_inov_organizacional_nan	não existência de classificação de sustentação para o tipo de inovação organizacional
n_socios	nº de sócios de um projeto
n_trabalhadores	nº de trabalhadores do promotor
n_cursos	nº de cursos indicados num projeto
formadores_ext	existência de formadores externos
cadeia_valor_32	classifica a mudança na rubrica 32 (cadeia de valor)
cadeia_valor_41	classifica a mudança na rubrica 41 (cadeia de valor)
vant_comp_32	classificação a mudança na rubrica – relação com fornecedores (vantagens competitivas)
vant_comp_33	classifica a mudança na rubrica – localização (vantagens competitivas)
vant_comp_34	classifica a mudança na rubrica – experiência e qualificação do RH (vantagens competitivas)
resultado_po_2	Existência de contribuição de forma decisiva para o fortalecimento da coesão e inclusão social ao longo do tempo
reforco_2	classifica a mudança na rubrica – Modelo de Gestão orientado para a inovação aberta (reforço)
domínio_3	existência de domínio prioritário - Automóvel, Aeronáutica e Espaço
contagem_mercados	nº total de mercados identificados na Tabela mercados
novo_produtos_mercados	nº de mercados novos (por produtos)
pais_extra_ue	nº de países extra união europeia antes do início do projeto

A performance dos modelos ao longo das três fases está descrita nas Tabelas 16 e 17, página 37.

Tabela 17 - AUC da experiência das experiências de otimização

	AUC-ROC		
Experiência	**3 (10 k-fold)**	**2 (10 cross validation)**	**1 (10 k-fold)**
LR	0.81	0.78	0.80
SVM	0.82	0.73	0.80
GNB	0.81	0.75	0.78
RF	0.81	0.77	0.80
MLP	0.81	0.79	0.74
XGB	0.81	0.76	0.80
DT	0.71	0.59	0.64

Tabela 18 - Accuracy das experiências de otimização

	Accuracy		
Experiência	**3 (10 k-fold)**	**2 (10 cross validation)**	**1 (10 k-fold)**
LR	0.77	0.78	0.75
SVM	0.78	0.73	0.75
GNB	0.75	0.76	0.73
RF	0.76	0.77	0.75
MLP	0.76	0.79	0.69
XGB	0.77	0.76	0.75
DT	0.70	0.59	0.65

Experiência 3 - experiência realizada com feature selection e modelos otimizados, k-fold = 10
Experiência 2 - experiência realizada sem feature selection mas com modelos otimizados, com k-fold= 5
Experiência 1- experiência realizada com as todas as variáveis, k-fold = 10

Os resultados mostram-se melhores quando a otimização de hiperparametros e a seleção de features é realizada. Obtendo-se os melhores resultados neste problema com apenas 21 variáveis.

3.6. Avaliação

A avaliações dos modelos construídos foi feita no decorrer das experiências, recorrendo a métricas como a AUC score, accuracy e precision. Na última experiência foram adicionadas duas métricas: recall e f-score.

3.7.Experiências seguintes

No capítulo 4 irão ser apresentadas as seguintes experiências, apresentadas na Tabela 19.

Tabela 19 - Breves descrições das experiências do grupo 4

Experiência	Dados utilizados	Breve descrição
4.1	rácios financeiros	Utilização de alguns rácios financeiros das empresas juntamente com os dados da baseline
4.2	dados separados por dimensão de empresa	Nesta experiência foram criados e testados dois conjuntos de dados: - microempresas - pequenas e médias
4.3	classificações dos técnicos do iapmei	Uso das variáveis que dizem respeito à avaliação dos projetos
4.4	informações de um projeto anulado em 2021	Codificação de features de um projeto para posteriormente ser classificado por modelos anteriormente criados
4.5	variáveis dos consultores dos projetos	Utilização de features relacionadas com os consultores ligados aos projetos
4.6	todas as variáveis anteriores	Para esta experiência foram utilizadas todas a variáveis previamente construídas, e utilizou se um algoritmo para selecionar as features de modo a otimizar métricas
4.7	todas as variáveis anteriores	Nesta experiência os projetos foram ordenados para simular uma situação real. Nestes modelos foram otimizados os hiperparametros e a escolha de *features*.

A intenção ao realizar estas experiências foi de testar casos específicos e verificar se trazia melhorias aos resultados. Desta forma, no capítulo seguinte, cada uma destas experiências tem uma secção em que são apresentadas métricas, matrizes de confusão e análises aos resultados.

Capítulo 4 – Análise e discussão dos resultados

Tal como referido no capítulo anterior, neste capítulo estão as descrições e os resultados das experiências que tinham como objetivo a testagem de indícios e problemas sugeridos pelos técnicos.

4.1.Utilização de rácios financeiros como preditores

No momento da candidatura são cedidas várias informações por parte dos promotores, nomeadamente, informações financeiras sobre as empresas. Com base nessas informações, é possível construir rácios financeiros que poderão trazer indicações sobre a saúde financeira da empresa e talvez serem preditores acerca do sucesso ou insucesso dos projetos. Por exemplo, segundo as informações prestadas pelos técnicos, os promotores com melhores resultados financeiros poderiam ter uma influência positiva na forma como o projeto iria decorrer. Optou-se então por construir alguns rácios para t-1(ano anterior ao ano da candidatura) e t-2(ano anterior a t-1) e transformá-los em *features* de forma a testar essa possibilidade. As variáveis introduzidas nesta experiência provenientes da IES dos promotores têm o prefixo "prom" no nome por dizerem respeito aos promotores e os sufixos "t-1" e "t-2" correspondem ao ano em questão.

Esta experiência foi realizada da seguinte forma:

1. Testar todas as variáveis identificadas na experiência 3.5 e perceber o impacto da utilização dos rácios apenas no momento t-1.
2. Selecionar as variáveis mais importantes da opção tomada em 1)
3. Analisar o conjunto de variáveis mais importantes e perceber o impacto dos rácios
4. Repetir a fase de modelação com o conjunto de variáveis selecionadas

Os rácios utilizados nesta experiência e as suas descrições encontram-se em anexo na Tabela 14. Estas variáveis foram juntas a todas as variáveis criadas no capítulo 3.

O primeiro passo visou verificar se a utilização dos rácios com informações do ano anterior (t-1) juntamente a informações de t-2 traria melhorias ao invés de utilizar apenas informações em t-1. Desta forma poder-se-ia reduzir o número de variáveis mesmo não abdicando da performance.

Analisando a Tabela 56 dos anexos podemos concluir que a utilização de rácios em t-2 não aumenta no geral a performance dos modelos e, portanto, a utilização das variáveis dos rácios apenas para t-1 é a opção tomada. O modelo DT aumentou consideravelmente a performance o que pode indiciar uma situação de *overfitting*.

De seguida, selecionaram-se as variáveis mais importantes de cada modelo, para consulta da listagem completa das variáveis selecionadas consultar Tabela 66 dos anexos. A escolha de variáveis mais importantes resulta de uma união de top 12 features de cada modelo considerado importante, SVM, LR, RF e XGBoost. Nesta escolha de variáveis surgiram alguns rácios, como podemos observar na Tabela 20:

Tabela 20 - Rácios escolhidos na seleção de variáveis

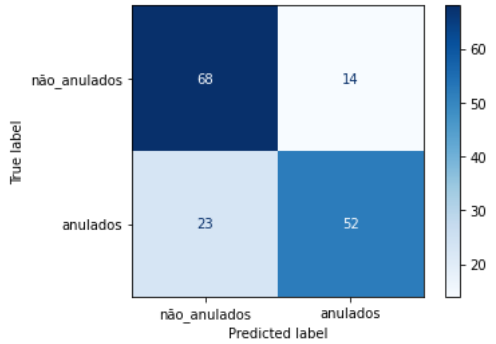
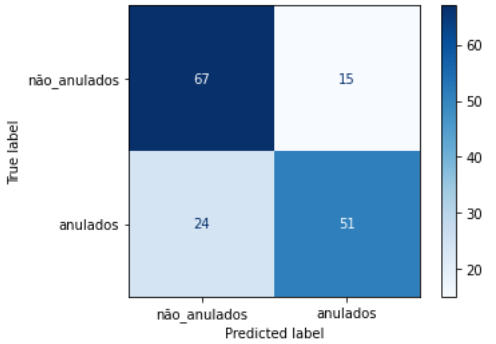
Varíaveis dos rácios escolhidas
prom_cmvmc/inventario_t-1
prom_ebitda/vn_t-1
prom_n_trabalhadores_t-1
prom_emprestimos_obtidos_passivo_ncor_t-1
prom_aqi_t-1
prom_volume_negocios_t-1
prom_cmvmc/fornecedores_t-1
prom_ativo_total_t-1
prom_autonomia_financeira_t-1
prom_resultado_liquido/vn_t-1
prom_tata_t-1
prom_lvgi_t-1
prom_sgai_t-1
prom_emprestimos_obtidos_passivo_cor_t-1

De seguida repetiu-se a experiência, mas utilizando apenas as variáveis escolhidas. Os resultados k-fold 10 poderão ser consultados na Tabela 21:

Tabela 21 - Resultados k-fold 10 incluindo variáveis de rácios

	AUC		Accuracy		Precision	
Rácios	**rácios**	**baseline**	**rácios**	**baseline**	**rácios**	**baseline**
LR	0.82	0.81	0.77	0.77	0.76	-
SVM	0.82	0.82	0.77	0.78	0.77	-
GNB	0.80	0.81	0.75	0.75	0.74	-
RF	0.80	0.81	0.75	0.76	0.76	-
MLP	0.81	0.81	0.76	0.76	0.75	-
XGB	0.80	0.81	0.75	0.77	0.74	-
DT	0.67	0.71	0.64	0.70	0.58	-

Tabela 22 - Matrizes de confusão de dois modelos nos dados do conjunto de teste

Matriz de confusão - LR	Matriz de confusão - SVM
	

Os resultados melhoraram cerca de 1-2% face aos resultados sem o uso de rácios em ambas as métricas comparáveis AUC e Precision. Em termos de interpretabilidade existem outras conclusões importantes retiradas das Tabelas de importância de features (Tabela 67 no Anexo A). No caso dos dois melhores algoritmos: Logistic Regression e SVM, ambos reconhecem a importância dos rácios na classificação de projetos. Os modelos construídos com estes algoritmos têm a particularidade de conseguirem perceber quais variáveis são importantes para classificar cada uma das classes, projetos anulados e projetos bem-sucedidos. Nestes modelos, os rácios importantes para a classificação de projetos indicam-nos, na generalidade dos rácios, que quanto maior o valor do rácio, menor é o risco de anulação. Os rácios que indiciam isto são o ebitda/vn ("Earnings before interest, taxes, depreciation and amortization" a dividir pelo volume de negócios), cmvmc/inventario (diferença entre o valor das vendas e o valor dos custos das mercadorias vendidas e matérias consumidas a dividir pelo valor de inventário) e também a informação acerca do número de trabalhadores. Por outro lado, o modelo SVM utiliza o volume de negócios e o cmvmc/fornecedores (margem bruta a dividir pelo número de fornecedores) para classificar a classe dos anulados, dando a indicação que quando maior o valor que estas variáveis tenham, maior o risco de anulação. Ou seja, as empresas maiores e mais rentáveis tendem a ter os projetos menos anuláveis.

4.2.Modelos baseados na dimensão da empresa

As empresas são classificadas, consoante a sua dimensão, em quatro categorias: micro, pequena, média e grande. A informação sobre a dimensão da empresa é fornecida pelos promotores no momento da candidatura.

Os técnicos, durante as reuniões de compreensão de negócio, alertaram para o facto das empresas de dimensão micro terem características muito diferentes do resto das empresas e que deveriam ser tratadas separadamente.

Tabela 23 - Distribuição de projetos anulados por dimensão de empresa

Dimensão	Nº Total de Projetos	Total de Anulados	% de Anulados na dimensão
Micro	171	119	70%
Pequena	325	128	39%
Média	264	74	28%
Grande	21	14	67%
Total	781	335	43%

A Tabela 23 mostra a distribuição de projetos anulados dentro de cada dimensão de empresas. Dos projetos da amostra pertencentes a microempresas 70% acabam anulados, constrastando com os 39% das pequenas e 28% das médias. As grandes empresas têm 67% de projetos anulados. No entanto, importa referir, que na amostra apenas constavam 21 projetos referentes a grandes empresas. Assim, de forma a evitar enviesamentos, estas empresas não foram consideradas daqui para a frente ao longos das experiências.

Para esta experiência foram criados dois grupos de dados a testar. O grupo 1 é referente a microempresas e o grupo 2 é referente a pequenas e médias empresas. O grupo 2 conta com 589 projetos dos quais 202 vieram a ser anulados (34% de anulações).

O método de realização desta experiência segue a linha das experiências anteriores. Neste caso, para cada grupo, foram criados modelos com todas as variáveis. Posteriormente foram escolhidas as melhores variáveis (selecionando o top 12 de variáveis com maior importância de cada modelo) e criados novos modelos, posteriormente procedeu-se à sua avaliação.

Os resultados dos modelos finais para cada grupo encontram-se na Tabela 24.

Tabela 24 - Resultados k fold dos modelos da experiência da dimensão da empresa

	AUC			Accuracy			Precision		
	Grupo1	**Grupo2**	**baseline**	**Grupo1**	**Grupo2**	**baseline**	**Grupo1**	**Grupo2**	**baseline**
LR	0.79	0.77	0.81	0.74	0.78	0.77	0.70	0.79	-
SVM	0.72	0.76	0.82	0.69	0.78	0.78	0.64	0.77	-
GNB	0.78	0.76	0.81	0.70	0.74	0.75	0.69	0.72	-
RF	0.77	0.75	0.81	0.73	0.76	0.76	0.68	0.75	-
MLP	0.78	0.76	0.81	0.73	0.75	0.76	0.69	0.74	-
XGB	0.76	0.74	0.81	0.74	0.76	0.77	0.69	0.76	-
DT	0.63	0.61	0.71	0.63	0.65	0.70	0.58	0.60	-

Os resultados não foram tão positivos como a classificação de todas as empresas ao mesmo tempo, que por vezes passaram os 0.80 de auc-roc ou 0.75 de taxa de acerto (accuracy). No entanto, ao fazer uma experiência separando em dois grupos a amostra, consegue-se criar modelos que utilizam variáveis mais específicas para cada conjunto.

Numa segunda experiência, para classificar cada grupo de empresas foram utilizadas 28 e 30 melhores variáveis, respetivamente grupo 1 e grupo 2. Na Tabela 25 encontram-se as variáveis que se encontram num conjunto que não estão presentes no outro e como se observa, existem algumas variáveis diferentes. Importa referir que os rácios selecionados pelos modelos não se encontram presentes na Tabela.

Tabela 25 - Váriaveis que diferentem entre o Grupo 1 e Grupo 2

Variáveis exclusivas do Grupo 1	Variáveis exclusivas do Grupo 2
consultora	n_meses_4
idade_candidatura_1	nut_centro
prom_url_empresa	nut_norte
vant_comp_24	n_socios
cadeia_valor_41	contagem_mercados
marcas_propria_N	cadeia_valor_32
vant_comp_est_2.0	am_dir_0
tipo_inov_organizacional_nan	am_dir_4
ativ_inov_marketing	vant_comp_15
novo_produtos_mercados	vant_comp_36
valor_formacoes	desafio_1
	reforco_1
	n_marcas_outras_nan
	n_marcas_outras

Como podemos observar, a existência de consultora é um factor mais importante no sucesso de um projeto para microempresas, enquanto, para empresas de maior dimensão, não é considerada de grande importância. Possivelmente, este facto indica que existe

algum conhecimento (know how) em empresas de maior dimensão e, portanto, o facto de terem o auxílio de uma consultora não é tão decisivo.

Tabela 26 - Distribuição de projetos bem-sucedidos (não anulados) com base na presença de consultora na candidatura

Grupo	Tem consultora	Não tem consultora
1	35.65%	19.64%
2	64.84%	66.12%

A percentagem de projetos bem-sucedidos das microempresas com consultora associada é de 35.65% face aos 19.64% dos projetos sem consultora. Com base nisto, verifica-se um aumento de cerca de 16% de sucesso dos projetos onde existe uma consultora indicada no momento da candidatura. No caso das empresas pertencentes ao Grupo 2, o facto de ter ou não consultora não altera muito a percentagem de projetos bem sucedidos

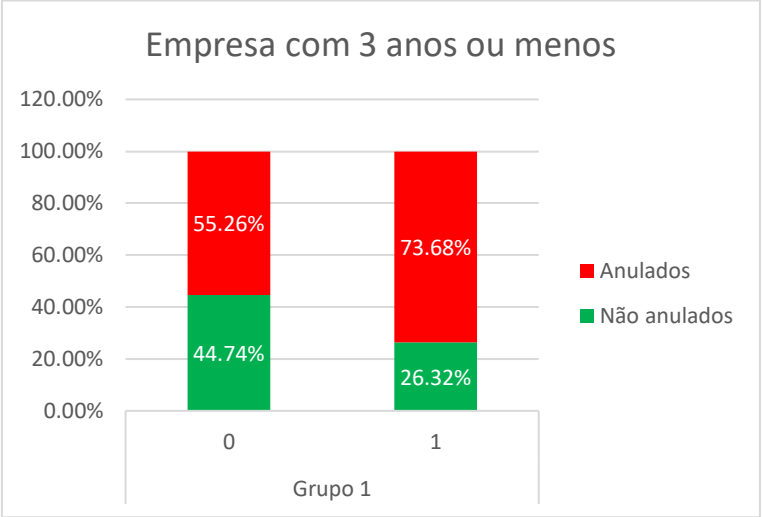


Figura 8 - Distribuição da variável idade_candidatura_1 no Grupo 1

Outro aspecto a salientar nesta análise de variáveis importantes para os modelos, é o facto de a segunda variável mais importante no Grupo 1 ser a da idade da empresa à data da candidatura. Como se observa na Figura 8, as microempresas criadas há 3 ou menos, têm um elevado risco de vir a ter os seus projetos anulados. Neste gráfico, o 1 signfica que a empresa tem 3 anos ou menos. Mais de 70% das microempresas com 3 ou menos anos vieram a ter o seu projeto anulado, o que torna esta característica um factor de risco.

Outra característica importante é a existência de um website, como mostra a variável "prom_url_empresa" na Figura 9.

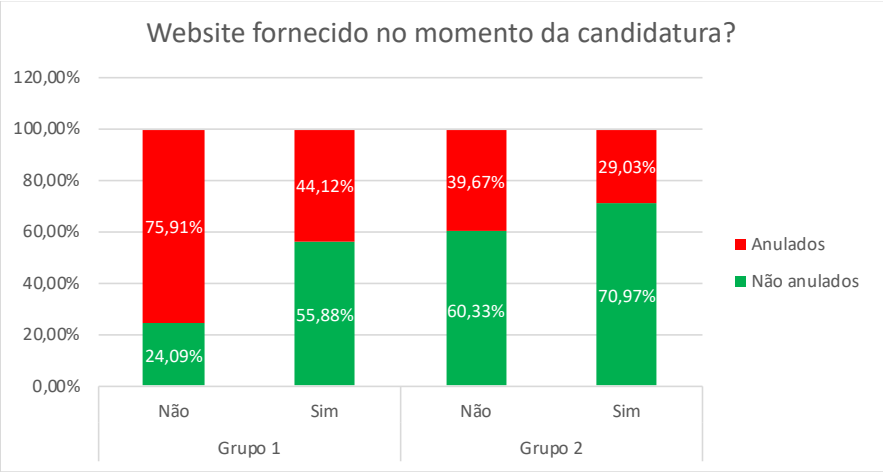


Figura 9 - Distribuição da variável prom_url_empresa

Analisando a distribuição desta variável, existe uma tendência elevada no Grupo 1 (microempresas), observa-se uma relação entre o facto da empresa fornecer website e anulação/sucesso de um projeto. Através da observação do gráfico confirma-se também, a razão pela qual a variável é importante para o Grupo 1 e, mas não para o Grupo 2. Esta variável pode indiciar que o facto de a empresa ser mais sofisticada pode ter relação com o sucesso do projeto.

Existem outras variáveis que caracterizam melhor os projetos microempresas nomeadamente as contagens de novos produtos em novos mercados, ou mesmo o valor de formações, que tal como a contagem de novos mercados, dá indícios que quanto maior o valor da variável, maior será o risco de anulação.

Os projetos enumeram numa parte da candidatura que atividades de inovação irá conter. As possibilidades de atividades de inovação são: produto, processo, marketing e organizacional. Dentro de quatro possibilidades de inovação, cada projeto poderá indicar quantas quiser de cada uma. A presença de cada tipo de atividade foi indicada nas variáveis com o prefixo “ativ_inov”. Analisando esta variável na Figura 10, percebe-se que existem diferenças significativas entre os dois grupos nos projetos com uma atividade de inovação na área de marketing. No caso das empresas do Grupo 2, a existência deste tipo de atividade de inovação não constitui um risco por si só ao projeto. No caso das empresas do Grupo 1, a existência deste tipo de inovação no projeto, pode ser indicativo de um risco elevado de anulação, dado que quase 74% dos projetos com esta características, terminaram anulados.

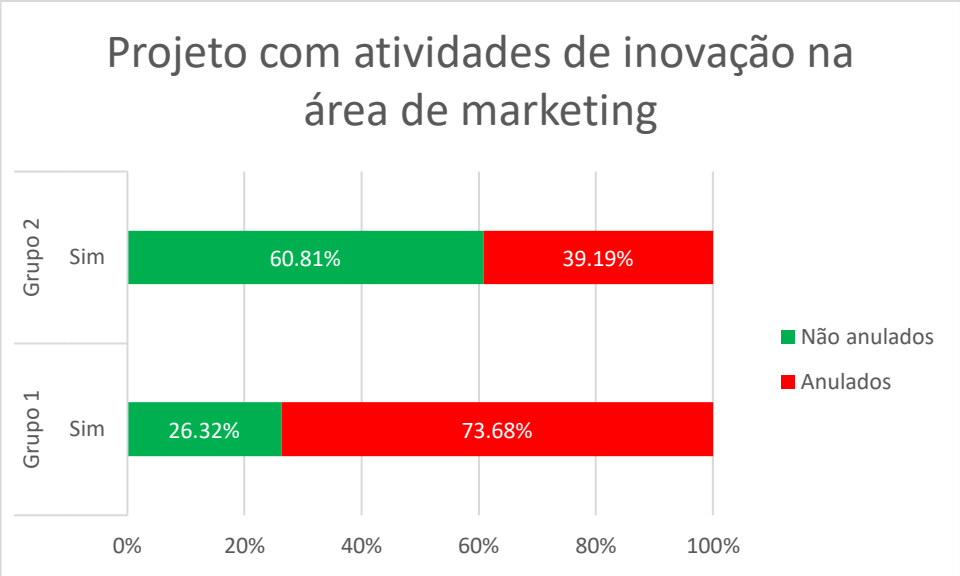
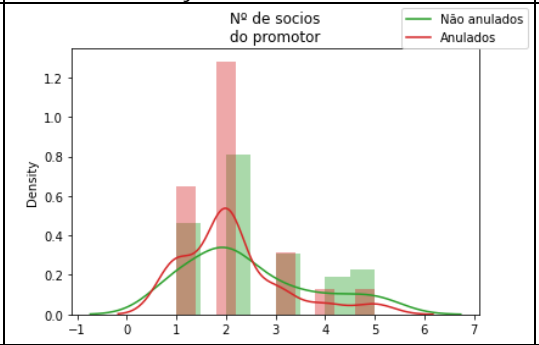
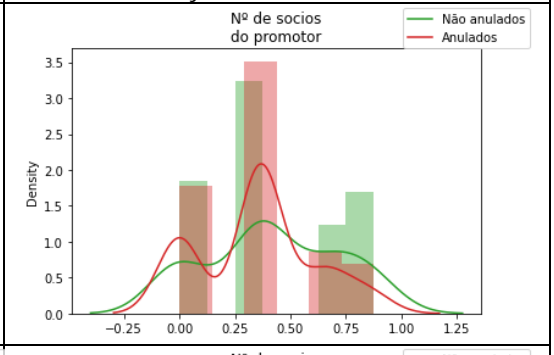
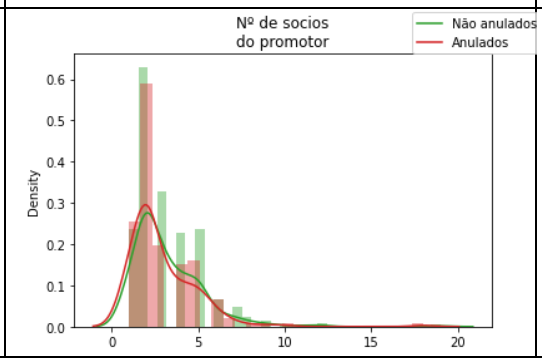
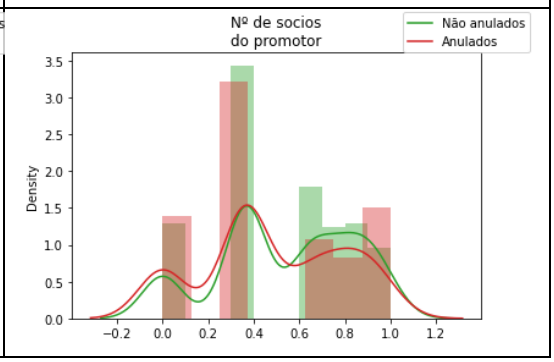


Figura 10 - Distribuição da variável ativ_inov_marketing

O número de sócios é uma variável mais decisiva para as empresas maiores, como seria de esperar, porque nessas empresas talvez exista uma maior dispersão no número de sócios por empresa. As microempresas tendem a ter poucos sócios e com pouca dispersão o que tornaria difícil uma boa predição através dessa informação.

Tabela 27 - Representação gráfica da variável "n_socios"

Representação gráfica da variável "n_socios"		
	Sem eliminação de outliers	Com eliminação de outliers
1		
2		

Projetos com maior número de sócios tendem a ser menos eliminados, em qualquer uma das classes, talvez porque os projetos encarnam um papel importante no planeamento

estratégico das empresas. Empresas com mais sócios tendem a ter um rigor maior dada o número de sócios a exigir resultados e responsabilidades e, portanto, não abandonam projetos com tanta facilidade como as que têm menos.

4.3. Modelos baseados nas classificações de mérito dadas pelos técnicos aos projetos

Após as candidaturas serem apresentadas, os técnicos classificam os projetos em 4 critérios.

- Crit_A – Qualidade do projeto
- Crit_B – Impacto do projeto na competitividade da empresa
- Crit_C – Contributo do projeto para a economia
- Crit_D – Contributo do projeto para a convergência nacional

Com estas classificações, criaram-se 4 variáveis que foram adicionadas ao conjunto de variáveis selecionadas da experiência que visou a utilização dos rácios financeiros (experiência 4.1).

Os resultados k-fold 10 desta experiência podem ser observados na Tabela 28:

Tabela 28 - Resultados k-fold 10 da experiência das classificações de mérito dos projetos

	AUC		**Accuracy**		**Precision**	
	4.3	**baseline**	**4.3**	**baseline**	**4.3**	**baseline**
LR	0.82	0.81	0.77	0.77	0.78	-
SVM	0.82	0.82	0.77	0.78	0.76	-
GNB	0.81	0.81	0.75	0.75	0.74	-
RF	0.81	0.81	0.75	0.76	0.74	-
MLP	0.81	0.81	0.75	0.76	0.74	-
XGB	0.80	0.81	0.74	0.77	0.73	-
DT	0.69	0.71	0.65	0.70	0.59	-

E as variáveis mais importantes para cada modelo podem ser encontradas nas Tabelas 29, 30, 31 e 32:

Tabela 29 - Coeficientes das variáveis mais importantes no modelo do algorítmo LR

LR			
n_meses_2	-0.43	formadores_ext	0.73
prom_ebitda/vn_t-1	-0.42	n_cursos	0.65
ativ_tipo_2	-0.40	resultado_po_2	0.52
reforco_2	-0.40	tipo_inov_organizacional_nan	0.45
dominio_3	-0.36	am_dir_0	0.41
n_trabalhadores	-0.35	n_meses_4	0.40

Tabela 30 - Coeficientes das variáveis mais importantes no modelo do algorítmo SVM

SVM			
reforco_2	-0.54	formadores_ext	0.80
prom_ebitda/vn_t-1	-0.54	n_cursos	0.71
dominio_3	-0.49	resultado_po_2	0.60
n_meses_2	-0.41	am_dir_0	0.54
prom_cmvmc/inventario_t-1	-0.41	tipo_inov_organizacional_nan	0.46
n_trabalhadores	-0.35	prom_tata_t-1	0.46

Tabela 31 - Coeficientes das variáveis mais importantes no modelo do algorítmo RF

Random Forest	
prom_ativo_total_t-1	0.06
prom_n_trabalhadores_t-1	0.05
formadores_ext	0.05
n_cursos	0.05
prom_ebitda/vn_t-1	0.05
prom_volume_negocios_t-1	0.04
prom_emprestimos_obtidos_passivo_ncor_t-1	0.04
resultado_po_2	0.04
prom_sgai_t-1	0.04
prom_lvgi_t-1	0.04
prom_autonomia_financeira_t-1	0.04
prom_tata_t-1	0.03

Tabela 32 - Coeficientes das variáveis mais importantes no modelo do algorítmo XGBoost

XGBoost	
resultado_po_2	0.09
formadores_ext	0.07
prom_ativo_total_t-1	0.06
n_cursos	0.06
prom_n_trabalhadores_t-1	0.04
prom_ebitda/vn_t-1	0.03
n_meses_4	0.03
prom_emprestimos_obtidos_passivo_ncor_t-1	0.03
prom_aqi_t-1	0.03
n_meses_2	0.03
reforco_2	0.03
am_dir_0	0.03

Pode-se observar que nenhuma das variáveis introduzidas, referentes às classificações dos técnicos, consta no top 12 destes modelos, concluindo-se que não foram consideradas como tendo grande poder discriminatório. Para além disso, a performance dos modelos foi sensivelmente a mesma da *baseline*. Seria de esperar que os projetos anulados tivessem avaliações mais baixas, ou seja as variáveis que dizem respeito à classificação dos técnicos dessem indícios de alguma ligação entre anulação e classificação dada, no entanto isso não se verifica. Com isto, percebe-se a dificuldade de classificação de um projeto anulado de um projeto não anulado, visto que até os próprios técnicos têm dificuldade em fazer essa análise.

4.4. Classificação de um projeto anulado

Para se verificar se os modelos construídos por esta investigação são fidedignos, utilizou-se uma notícia publicada em 2021 (Cabrita & Amaral Santos, 2021) acerca de um projeto anulado. O objetivo principal foi verificar se os modelos indicariam este projeto como anulado. O projeto mencionado, para além de ter um valor elegível de 7 milhões de euros, durante os dois anos seguintes à aceitação do projeto, foram apoiados com um total de 2,6 milhões de euros cedidos pelo Estado, onde 1,650 milhões foram obtidos através do IAPMEI. No entanto, este projeto acabou por não avançar, criando-se assim uma cativação do dinheiro que, se o projeto estivesse assinalado atempadamente como projeto de risco, poderia ter se gerido o projeto de modo diferente.

Sendo assim, para a realização da experiência foram utilizados os seguintes modelos:

- Modelo sem rácios (secção 3.5)
- Modelo dimensão (secção 4.2)

No conjunto de dados, o projeto mencionado na notícia foi possível de ser encontrado pelo nome da empresa. No entanto, este não estava incluíndo no conjunto dos dados de trabalho da amostra pois este não foi considerado como terminado nem anulado.

Extraíram-se as informações desse projeto e de seguida construíram-se as features. Após estas tarefas deram-se como input dos modelos as features codificadas. Os modelos posteriormente devolvem uma classificação entre 0 e 1, com zero sendo não anulado e um o anulado.

Modelo sem rácios

Tabela 33 - Classificação dos modelos previamente criados usando modelos sem rácios, as verdes encontram-se os modelos que acertaram na classificação

Algoritmo	Classificação	Classificação por classe (probabilidade)	
		não anulado	anulado
SVM	não anulado	0.59	0.41
Random Forest	anulado	0.46	0.54
GaussianNB	anulado	0.32	0.68
MLP	não anulado	0.55	0.45
XGBoost	não anulado	0.51	0.49
Logistic Regression	anulado	0.44	0.56
Decision Tree	não anulado	0.5	0.5

Dos modelos utilizados dois conseguiram classificar o projeto como anulado, como se observa na Tabela 33. No entanto, as pontuações incorretas foram praticamente todas perto do limiar de classificação de 0.5. A maior falha foi a do modelo SVM, que classificou com mais certeza o projeto e com uma classificação errada. Já no caso do modelo com Decision Tree, classificou mal, mas no geral, este algoritmo não teve uma boa performance ao longo de todo o estudo.

Modelo dimensão

No caso do modelo dimensão, nas informações da candidatura este promotor indicava ser uma microempresa, portanto foi o modelo do grupo 1 a ser utilizado.

Tabela 34 - Classificação dos modelos previamente criados usando modelos com rácios e específico para o grupo 1, a verde encontram-se os modelos que acertaram na classificação

Algoritmo	Classificação	Classificação por classe (probabilidade)	
		não anulado	anulado
SVM	anulado	0.16	0.84
Random Forest	anulado	0.3	0.7
GaussianNB	anulado	0	1
MLP	anulado	0	1
XGBoost	anulado	0.47	0.53
Logistic Regression	anulado	0.25	0.75
Decision Tree	anulado	0	1

Utilizando os modelos criados especificamente para a dimensão do promotor em questão, todos os algoritmos classificaram o projeto como anulado, Tabela 34. Ou seja, no momento da apresentação da candidatura este projeto teria sido claramente detetado como uma potencial futura anulação.

Com esta experiência concluiu-se que, usando os modelos construídos com base na dimensão da empresa, obtiveram-se resultados claramente melhores. Ou seja, com estes modelos, o IAPMEI teria tido um alerta face a este projeto e gerido o incentivo financeiro de um modo mais cauteloso.

4.5.Utilização de variáveis de consultores como preditores

Um dos tópicos abordados pelos técnicos era a existência de consultora associada aos projetos. Na amostra existem 219 projetos anulados com o consultor identificado. Importa ressalvar que o preenchimento do nif de consultora não é um campo obrigatório e, portanto, existem projetos em que os promotores não identificam os consultores, Tabela 35.

Tabela 35 - Distribuição da existência de consultora face à distribuição de anulações

	Anulado	**Não anulados**	**Total**
Preenchido	219	324	543
Não Preenchido	109	129	238
Total	328	453	

Existem 238 projetos em que o promotor não identifica o nif do consultor e, portanto, não se consegue tirar conclusões sobre essa variável nesses projetos.

Para esta experiência criaram-se as variáveis apresentadas na Tabela 36:

Tabela 36 - Descrição das variáveis criadas para a experiência dos consultores

Variáveis	**Descrição**
contagem_consultores	número de projetos aceites do consultor do projeto
consultor_anulacao	o consultor não tem/tem uma anulação noutro projeto
existencia_consultor	existência ou não de consultor

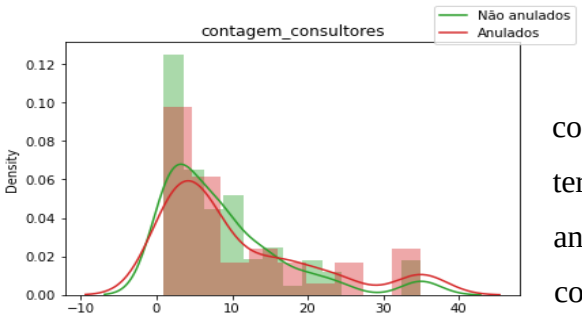


Figura 11 - Distribuição da variável contagem_consultores

Existe uma tendência para as consultoras que aparecem poucas vezes terem um maior número de projetos não anulados, Figura 11. Existem algumas consultoras com mais de 15 projetos que apresentam uma maior quantidade de projetos anulados face aos não anulados.

Para além destas variáveis, foram também adicionadas variáveis que dizem respeito aos rácios dos consultores, essas variáveis poderão ser consultadas na Tabela 69 no anexo.

Os resultados k-fold 10 cross validation desta experiência:

Tabela 37 - Resultados K-fold da experiência 4.5

	AUC		Accuracy		Precision	
Experiência	4.5	Baseline	4.5	Baseline	4.5	Baseline
LR	0.82	0.81	0.77	0.77	0.78	-
SVM	0.81	0.82	0.77	0.78	0.76	-
GNB	0.79	0.81	0.75	0.75	0.74	-
RF	0.81	0.81	0.75	0.76	0.74	-
MLP	0.81	0.81	0.75	0.76	0.74	-
XGB	0.81	0.81	0.74	0.77	0.73	-
DT	0.67	0.71	0.65	0.70	0.59	-

A inclusão das variáveis dos consultores não permitiu aumentar a performance dos modelos, como se observa na Tabela 37. As Tabelas com a importância de features de cada variável são possíveis de serem consultadas nas Tabelas 70 no anexo. Ao contrário da análise feita na secção 4.2, em que se concluiu que nos projetos de microempresas a existência de consultor era importante para a sua classificação, neste caso ao analisar-se os dados de todas as empresas isso não se verificou, e para uma classificação do conjunto de todos os dados a característica de ter ou não consultor não é decisiva.

4.6. Seleção de variáveis utilizando RFE

Esta experiência visa criar os melhores modelos, com base no algoritmo *Recursive Feature Elimination*. Este algoritmo elimina recursivamente as features com base nos seus coeficientes de importância para o modelo, escolhendo um score (*accuracy* ou *recall*, entre outros) objetivo. Realizaram-se duas experiências: uma tentando otimizar a métrica *accuracy* e outra para otimizar a métrica *recall*. E usaram-se os seguintes algoritmos: SVM, LR, RF e XGBoost.

O conjunto de dados utilizado nessa experiência continha: 149 projetos para o Grupo 1 e 546 projetos para o Grupo 2. Cada um desses projetos tinha uma classificação para cada uma das 189 variáveis utilizadas.

Foram criados modelos para os dois grupos de empresas: microempresas; pequenas e médias.

Tabela 38 - Distribuição do target no grupo 1 e grupo 2

	Não anulados	Anulados
Grupo 1	49	100
Grupo 2	363	183

Se o número de variáveis escolhidas for inferior a 30 encontra-se a listagem das mesmas em anexo.

Nesta experiência, o conjunto de dados é divido em conjunto de treino (70% da amostra) e o de teste (30%).

Para além de todas as variáveis conhecidas foram ainda adicionadas algumas variáveis, que foram construídas por análise dos textos proposta dos projetos, extraídas pela equipa de desenvolvimento, usando técnicas de *text mining*. Essas variáveis tinham sido construídas após identificação tópicos, averiguação sobre a semelhança entre textos de candidaturas e medição o comprimento dos textos. Para estas experiências utilizou-se todas as variáveis criadas juntamente com todas as variáveis criadas pela equipa de texto mining. Seguindo a linha de execução das experiências anteriores, escolheu-se o conjunto das variáveis mais importantes e procedeu-se à avaliação dos modelos.

4.6.1. Seleção de features por otimização de modelos tendo em vista a *Accuracy*

Grupo 1 (Microempresas)

Tabela 39 - Número de variáveis selecionadas na experiência 4.6.1. para o Grupo 1

	Nº variáveis
SVM	66
RF	132
LR	185
XGBoost	98

Conjunto de teste

Tabela 40 - Resultados conjunto de teste da experiência 4.6.1 – Grupo 1

	AUC	**Accuracy**	**Precision**	**Recall**	**F-score**
LR	0.58	0.56	0.76	0.52	0.62
SVM	0.56	0.56	0.74	0.55	0.63
RF	0.60	0.58	0.77	0.55	0.64
XGB	0.60	0.58	0.77	0.55	0.64

Um dos melhores algoritmos foi o *Random Forest*, como se observa na matriz de confusão da Figura 12.

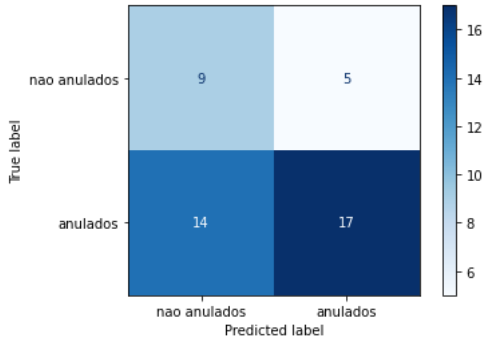


Figura 12 - Matriz de confusão para o modelo RF

K-fold 10 cross validation

Tabela 41 - Resultados k-fold da experiência 4.6.1 – Grupo 1

	AUC	Accuracy	Precision	Recall	F-score
LR	0.75	0.71	0.73	0.89	0.80
SVM	0.71	0.64	0.78	0.72	0.72
RF	0.76	0.71	0.72	0.92	0.80
XGB	0.76	0.71	0.74	0.90	0.80

Grupo 2 (Pequenas e médias empresas)

Tabela 42 - Número de variáveis selecionadas na experiência 4.6.1. para o Grupo 2

	Nº variáveis
SVM	8
RF	57
LR	6
XGBoost	65

Conjunto de teste

Tabela 43 - Resultados conjunto de teste da experiência 4.6.1 – Grupo 2

	AUC	Accuracy	Precision	Recall	F-score
LR	0.67	0.68	0.50	0.64	0.56
SVM	0.65	0.69	0.52	0.53	0.52
RF	0.71	0.70	0.53	0.74	0.61
XGB	0.69	0.68	0.51	0.72	0.59

Neste caso o modelo que teve melhor performance foi o *XGBoost*.

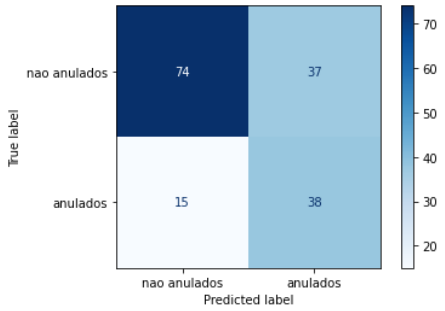


Figura 13 - Matriz de confusão para o modelo XGBoost

K-fold 10 cross validation

Tabela 44 - Resultados k-fold da experiência 4.6.1 – Grupo 2

	AUC	Accuracy	Precision	Recall	F-score
LR	0.73	0.75	0.69	0.46	0.55
SVM	0.74	0.71	0.79	0.28	0.39
RF	0.78	0.75	0.75	0.40	0.51
XGB	0.76	0.76	0.79	0.39	0.52

4.6.2. Seleção de features por otimização de modelos tendo em vista a *Recall*

Grupo 1

Tabela 45 - Número de variáveis selecionadas na experiência 4.6.2. para o Grupo 12

	Nº variáveis
SVM	63
RF	186
LR	3
XGBoost	63

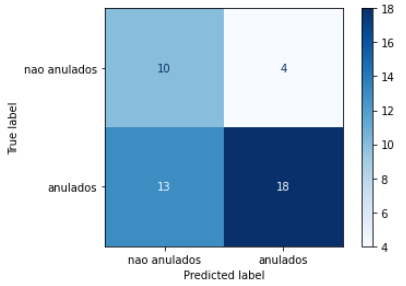
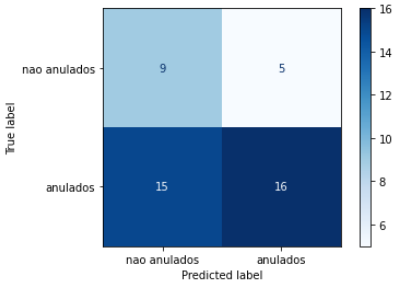
Conjunto teste

Tabela 46 - Resultados conjunto de teste da experiência 4.6.2 – Grupo 1

	AUC	Accuracy	Precision	Recall	F-score
LR	0.65	0.62	0.82	0.58	0.68
SVM	0.56	0.56	0.74	0.55	0.63
RF	0.58	0.56	0.76	0.52	0.62
XGB	0.60	0.58	0.77	0.55	0.64

Tanto o *Random Forest* como o *Logistic Regression* tiveram boas performances, no entanto importa referir que a partição do conjunto de teste deixa poucos dados, não oferencendo uma total confiança nos resultados:

Tabela 47 - Matrizes de confusão para dois dos modelos do Grupo 1 da experiência 4.6.2.

Logistic Regression	Random Forest
	

K-fold 10 cross validation

Tabela 48 - Resultados k-fold da experiência 4.6.2 – Grupo 1

	AUC	Accuracy	Precision	Recall	F-score
LR	0.75	0.74	0.77	0.89	0.82
SVM	0.70	0.69	0.81	0.75	0.76
RF	0.75	0.73	0.74	0.91	0.81
XGB	0.75	0.68	0.71	0.88	0.78

Grupo 2

Tabela 49 - Número de variáveis selecionadas na experiência 4.6.2. para o Grupo 2

	Nº variáveis escolhidas
SVM	23
RF	74
LR	22
XGBoost	65

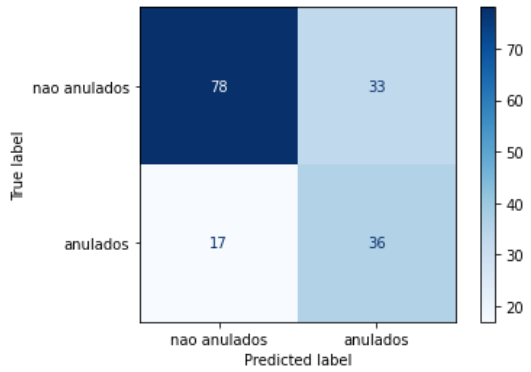
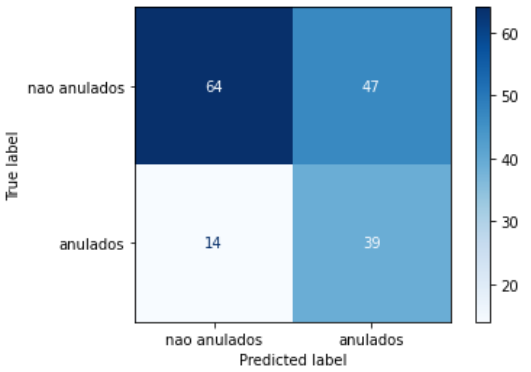
Conjunto teste

Tabela 50 - Resultados conjunto de teste da experiência 4.6.2 – Grupo 2

	AUC	Accuracy	Precision	Recall	F-score
LR	0.69	0.70	0.52	0.68	0.59
SVM	0.67	0.67	0.49	0.66	0.56
RF	0.66	0.63	0.45	0.74	0.56
XGB	0.69	0.68	0.51	0.72	0.59

As matrizes dos melhores modelos RF e LR encontram-se na Tabela 51.

Tabela 51 - Matrizes de confusão para dois dos modelos do Grupo 2 da experiência 4.6.2.

Logistic Regression	Random Forest
	

K-fold 10 cross validation

Tabela 52 - Resultados k-fold da experiência 4.6.2 – Grupo 2

	AUC	Accuracy	Precision	Recall	F-score
LR	0.79	0.77	0.77	0.49	0.59
SVM	0.77	0.75	0.71	0.43	0.53
RF	0.77	0.77	0.79	0.43	0.55
XGB	0.76	0.76	0.79	0.39	0.52

4.6.3. Conclusões da experiência

Com esta experiência, existiram várias conclusões importantes a serem mencionadas. Em primeiro lugar, podemos observar que é muito mais fácil classificar o grupo 1. Isto é, a distribuição de valores nas variáveis do grupo das microempresas é mais discriminatória que no grupo das pequenas e médias empresas. Os modelos criados para os projetos deste grupo de empresas atingiram resultados de 90% de cobertura (*recall*) em k-fold 10 e com taxas de precisão acima dos 80%. Por outro lado, os projetos das empresas do grupo 2 foram pior classificados, mesmo com a precisão dos modelos estar por vezes perto dos 80% em k-fold 10. Observando as melhores matrizes dos melhores modelos do grupo 2, percebemos que a única classe que está a ser classificada com precisão é a classe dos projetos não anulados.

Um ponto interessante desta experiência é que podemos perceber que o modelo usando *Logistic Regression*, otimizado para *recall* no grupo 1, apenas escolheu 3 variáveis, am_3.0, despesas_idt e emp_menos_4_anos_cand. Estas variáveis dizem respeito, respetivamente a:

- am_dir_3 - Direção de crescimento no mercado: Extensão do mercado (sim/não)
- despesas_idt – O projeto contempla despesas em investigação e desenvolvimento (sim/não)
- emp_menos_4_anos_cand – A empresa candidatura tem menos de 4 anos de vida (sim/não)

Com estas 3 variáveis o modelo conseguiu uma excelente prestação, apesar de, quando da análise exploratória, estas variáveis não se mostrarem particularmente cacterizadoras, surgem como criadoras de padrões na análise de modelação, justificado a necessidade de ser necessário ir além de uma análise meramente descritiva neste tratamento de dados.

4.7. Criação de modelos para uso real

A última experiência deste estudo tem como principal objetivo a avaliação da capacidade de uso real destes modelos. Para isso, ordenou-se os projetos por data de candidatura e utilizou-se os primeiros 75% para treinar os modelos e os últimos 25% de teste.

Tabela 53 - Distribuição dos projetos no conjunto treino e teste

	Não anulados	Anulados
treino	356	182
teste	70	108

Utilizaram-se todas as variáveis das experiências que foram surgindo ao longo de todo o estudo, utilizou-se um algoritmo para escolha de híper parâmetros, o *GridSearch*, e usou-se o algoritmo *Recursive Feature Elimination* para escolher as melhores *features* tendo como objetivo o *recall* da classe dos projetos anulados. A escolha da métrica *recall* recai sobre a necessidade maior de detetar os projetos anulados em detrimento da taxa de acerto. Os projetos utilizados dizem respeito a projetos de todo o tipo empresas.

Os hiperparâmetros escolhidos para cada algoritmo poderão ser consultados nas Tabelas 75, 76, 77 e 78, no anexo.

Tabela 54 - Nº de variáveis escolhidas para cada modelo

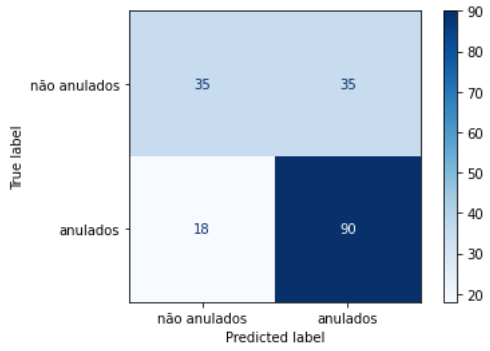
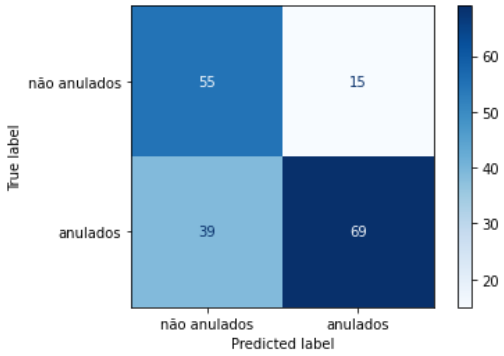
	Nº variáveis escolhidas
SVM	185
RF	10
LR	8
XGBoost	142

Resultados conjunto de teste

Tabela 55 - Resultados dos modelos na experiência dos dados simulando situação real

	AUC	Accuracy	Precision	Recall	F-score
LR	0.62	0.58	0.73	0.48	0.58
SVM	0.67	0.70	0.72	0.83	0.77
RF	0.66	0.64	0.77	0.58	0.66
XGB	0.71	0.70	0.82	0.64	0.72

Tabela 56 - Matrizes dos modelos SVM e XGBoost

SVM	Xgboost
	

Como se deve observar na Tabela 55, o modelo SVM criado nesta experiência conseguiu uma cobertura de 83% dos projetos que viriam a ser anulados, e pode-se afirmar que teve uma precisão aceitável de 72%. No caso do modelo XGBoost, este teve uma melhor precisão com 82%, no entanto a cobertura dos casos anulados foi de apenas 64%. Uma nota importante é que ambos os modelos referidos anteriormente usaram bastantes variáveis: 185 no caso do SVM e 142 no caso do XGBoost, ou seja, evidenciam a extrema dificuldade e complexidade da análise de projetos para previsão de sucesso/insucesso.

Capítulo 5 – Conclusões e recomendações

5.1. Principais conclusões

Esta dissertação tinha como principal objetivo verificar a precisão com que se consegue modelar o trabalho de avaliação e suporte de projetos por parte de entidades que fazem a distribuição de fundos comunitários.

Como ponto de partida, foi importante perceber como funcionava o ciclo de um projeto bem como entender a real importância da integração de sistemas de inteligência artificial no âmbito da administração pública. Durante a pesquisa bibliográfica não se encontraram trabalhos semelhantes, realçando-se assim a extrema importância desta dissertação. Ao longo da tarefa de revisão de literatura percebeu-se o impacto que a atribuição de fundos tem no desenvolvimento das PME e consequentemente no desenvolvimento do país.

As conclusões aqui presentes dizem respeito à análise exploratória e às tarefas de modelação realizadas durante este estudo. Na fase exploratória de dados percebeu-se que existe uma elevada diversidade de empresas no conjunto de dados, e que todas elas têm características muito diferentes.

No caso das microempresas detetou-se facilmente que estas têm uma elevada taxa de anulação. De facto, estas empresas têm 70% dos seus projetos anulados. Em contrapartida, do conjunto dos projetos de pequenas e médias empresas apenas 34% terminaram em anulação. Outra conclusão deste tipo de empresas diz respeito ao número de diferentes postos de trabalho, sendo que se identificou que quanto maior o número de postos de trabalho diferentes menor o risco de anulação. Para além disto, a não existência de uma direção de mercado identificada no momento da candidatura indicia uma provável anulação.

Na fase de modelação, os rácios mostraram ser uma importante fonte de informação para a classificação de projetos. É importante salientar que o uso de mais que um ano de informações financeiras não trouxe melhorias aos modelos de previsão. Também foi possível observar que quanto melhor os rácios financeiros de uma empresa menos anulável serão os seus projetos. Para além disso, como seria esperado, um maior número de trabalhores também indicia menos probabilidade de anulação.

Microempresas recentes, com menos de 3 anos, apresentam um risco acrescido. Se a empresa for demasiado ambiciosa também é um indício que existirá maior probabilidade

de os projetos serem anulados, neste caso os modelos identificaram a variável que contabiliza o número de novos mercados externos que os promotores tencionam entrar durante ou após o projeto, como um fator de risco. Projetos com inovações na área de *marketing* tendem a ser mais anulados. Um pouco de sofisticação nas empresas, sejam micro ou de outro tipo, indicam um menor risco de anulação como se verifica no caso de a empresa ter ou não *website.*

No caso das pequenas e médias empresas não foi tão claro retirar explicações que permitam caracterizar se os projetos destas empresas são mais ou menos anuláveis. No entanto, foi possível identificar que as empresas com maior número de sócios tendem a ter os seus projetos menos anulados.

As classificações obtidas no mérito do projeto no momento da análise da candidatura não permitiram separar os projetos de sucessos dos anulados, pois essas classificações dos projetos não permitiram a diferenciação entre estes. Os técnicos no momento da avaliação das candidaturas têm imensas dificuldades em diferenciar os melhores dos piores, usando a amostra de projetos aceites.

No caso particular do projeto identificado à posterior como anulado, verificou-se que os modelos relevantes que tinham sido criados com base na dimensão da empresa conseguiram prever a sua anulação no momento da candidatura, o que teria poupado tempo e fundos à instituição em questão.

A existência de consultor e o uso de informações dos consultores presente na candidatura não permite aumentar a performance da previsão dos projetos.

Os modelos criados usando a técnica de RFE são melhores para o grupo 1, mostrando que usando as variáveis que se teve acesso os projetos anulados são mais facilmente separados dos não anulados do que os projetos no grupo 2. No entanto, os modelos criados para o grupo 2 têm uma elevada precisão a identificar os projetos não anulados. A sugestão seria usar como classificador para ambas as classes (anulados e não anulados) para projetos de microempresas. Nos projetos de pequenas e médias empresas a utilidade dos modelos passaria por servir como primeira validação aos projetos que não viriam a ser anulados, isto é, se os modelos classificassem como "não anulado" seria muito provável ele não ser realmente anulado. Porém, se fosse classificado como "anulado" a classificação já não seria tão interessante porque o modelo não conseguiu captar bem esses projetos. Como já foi referido, é importante ter em mente que as microempresas têm

elevadas taxas de anulação, o que se torna bastante prioritário despistar o maior número de situações em que isso aconteça, preferencialmente no momento antes da aceitação do projeto. No grupo 2, a percentagem de anulações desce para 34%. Embora não seja considerado um valor ótimo não é tão preocupante.

Quando os projetos foram ordenados por data da candidatura de forma a simular qual seria o comportamento dos modelos criados numa situação real, a performance baixou ligeiramente e mostrou a dificuldade e complexidade em transpor estes modelos para uma situação real. Por outro lado, um dos modelos conseguiu boas performances e o uso deste modelo tinha permitido detetar 83% das anulações que viriam a aparecer posteriormente à data do último projeto no conjunto de treino, o que tinha significado uma enorme poupança para as instituições.

Em suma, a atribuição de fundos a microempresas exige maior atenção e mais especificamente empresas recentes de dimensão reduzida com rácios baixos e pouco sofistificadas. Esta necessidade de atenção deve-se à inexperiência destas empresas e não a uma intenção clara de abandono de projetos ou à incapacidade de cumprir com o acordado. Recomendar-se-ia um maior acompanhamento dos projetos destas empresas, porque estas empresas têm uma elevada importância no tecido empresarial português e o seu desenvolvimento é crucial para a economia nacional.

Conclui-se assim que o uso destes modelos poderá ajudar os técnicos na avaliação dos projetos e por consequência criar uma melhor eficiência na atribuição de fundos. As empresas que não tiveram os seus projetos aceites por limitação de fundos poderiam ter uma oportunidade. Para além disso, as empresas que conseguem a aprovação nos seus projetos, mas que mais tarde vêm a ser anulados, seriam mais facilmente detetadas no momento da candidatura.

Para além do trabalho que esta dissertação contempla, foi criado um protótipo. Esse protótipo permite que um projeto seja classificado automaticamente por vários modelos, sendo apenas necessário a introdução do mesmo e o output contempla um ficheiro de resultados de predições.

5.2. Contributos para a comunidade científica e empresarial

O trabalho desenvolvido ao longo desta dissertação, permitirá uma melhor avaliação dos projetos por parte das entidades responsáveis. Para além de conseguir evitar situações de anulação antes de elas acontecerem, permitirá perceber quais os projetos com maior risco. Conseguindo identificar estas situações, poderá mover um maior acompanhamento e evitar umas possíveis situações futuras de anulações. As empresas consideradas boas e com baixa probabilidade de problemas, são libertas antecipadamente de demorosos e custosos processos burocráticos.

5.3.Limitações do estudo

Este estudo apresenta algumas limitações começando pela falta de literatura na temática aqui apresentada. Sendo a área da inteligência artificial, mais concretamente aplicada à administração pública, um tema recente é difícil arranjar estudo comparativos. Para além disso, existem muito poucos estudos sobre o tema adaptados à realidade portuguesa, o que dificulda ainda mais esta investigação.

Outro grande desafio desta dissertação foi a limitação da dimensão permitida no currente documento. Este problema agregado à falta de bibliografia, mencionada anteriormente, criou alguns problemas na seleção da informação apresentada ao longo do trabalho, podendo ter se gerado algumas lacunas de contexto. Estas falhas foram evitadas ao máximo, mas é importante realçar que limitação de dimensão foi de facto um problema.

Esta dissertação como foi referido anteriormente, esteve inserida num contexto de um projeto de investigação, do ISCTE, com o nome "IA-SI - Inteligência Artificial nos Sistemas de Incentivos". Existiram outros trabalhos que não estão aqui descritos, nomeadamente criação de modelos de aprendizagem automática para deteção de anulações de projeto para a AICEP, a criação de modelos de aprendizagem automática para deteção de faturas inelegíveis para o IAPMEI e criação de modelos de aprendizagem automática para deteção de faturas inelegíveis para a AICEP. A colocação destes estudos poderia ter contribuído para a melhor preceção da dimensão e complexidade do problema.

Nas limitações técnicas desta dissertação surge o problema do reduzido conjunto de dados. Este problema limitou a utilização de algumas técnicas e requer alguma atenção na interpretação dos dados, dado que os dados são poucos para uma generalização a todo e qualquer exemplo de financiamento de projetos. Para além disto as experiências realizadas ao longo da dissertação foram feitas seguindo práticas comuns de *data mining*, nomeadamente: criação de variáveis, pre processamento de variáveis, escolha de hiperparametros. No entanto, com outras opções tomadas os resultados poderiam ser ligeiramente diferentes.

5.4.Propostas de investigação futura

Para futuros trabalhos as sugestões passam pela aquisição de mais dados provenientes de agências de outros países, de forma a perceber se a existe alguma semelhança entre os problemas na atribuição de fundos nos diversos países.

Outra sugestão seria a criação de modelos usando todos os projetos que concorrem a fundos, ao invés da utilização apenas de projetos que foram aceites. Estes modelos poderiam prever quer a sua aceitação, quer o potencial risco de anulação.

Para além destas sugestões, um trabalho que seria importante realizar seria usando técnicas de *text-mining*, criar variáveis com base na descrição de projetos e através disso perceber o seu risco de anulação ou a possibilidade de ter faturas inelegíveis ou irregulares.

Referências Bibliográficas

Agbozo, E., & Spassov, K. (2018). Establishing efficient governance through data-driven e-government. *ACM International Conference Proceeding Series*, 662–664. https://doi.org/10.1145/3209415.3209419

Alexopoulos, C., Diamantopoulou, V., Lachana, Z., Charalabidis, Y., Androutsopoulou, A., & Loutsaris, M. A. (2019). How machine learning is changing e-government. *ACM International Conference Proceeding Series*, *Part F1481*, 354–363. https://doi.org/10.1145/3326365.3326412

Anderson, H. J., Gyamfi, S., & Stejskal, J. (2020). Public Funding as a Catalyst for Firms' Technological Innovation: Case of Cyprus, Croatia and Portugal. *17th International Conference on Intellectual Capital, Knowledge Management & Organisational Learning ICICKM 2020*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Artola, C., & Genre, V. (2010). *Euro area SMEs under financial constraints: Belief or reality?*

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, 182–185. http://recipp.ipp.pt/handle/10400.22/136%0Ahttp://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf

Ballantyne, A., & Stewart, C. (2019). Big Data and Public-Private Partnerships in Healthcare and Research: The Application of an Ethics Framework for Big Data in Health and Research. *Asian Bioethics Review*, *11*(3), 315–326. https://doi.org/10.1007/s41649-019-00100-7

Belás, J., Dvorský, J., Kubálek, J., & Smrčka, L. (2018). Important factors of financial risk in the SME segment. *Journal of International Studies*, *11*(1), 80–92. https://doi.org/10.14254/2071-8330.2018/11-1/6

Bošnjak, Z., Grljević, O., & Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. *Proceedings - 2009 5th International Symposium on Applied Computational Intelligence and Informatics, SACI 2009*, *114*, 509–514. https://doi.org/10.1109/SACI.2009.5136302

Cabrita, F., & Amaral Santos, J. (2021, May 17). IAPMEI anula apoios a 'fábrica fantasma.' *SOL*. https://sol.sapo.pt/artigo/734907/iapmei-anula-apoios-a-fabrica-fantasma

Cainelli, G., Evangelista, R., & Savona, M. (2006). Innovation and economic performance in services: A firm-level analysis. *Cambridge Journal of Economics*, *30*(3), 435–458. https://doi.org/10.1093/cje/bei067

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rudiger, W. (2000). Crisp-Dm 1.0. In *CRISP-DM Consortium*.

Chen, W., & Li, J. (2009). Machine for Credit Risk Identification in Small-and-Medium. *Management Science, July*, 12–15.

Demary, M., Hornik, J., & Watfe, G. (2016). SME Financing in the EU: Moving Beyond One-Size-Fits-All. In *Institute for Economic Research*.

Dereliolu, G., & Gürgen, F. (2011). Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey. *Expert Systems with Applications*, *38*(8), 9313–9318. https://doi.org/10.1016/j.eswa.2011.01.012

Doorn, N. (2021). Artificial intelligence in the water domain: Opportunities for responsible use. *Science of the Total Environment*, *755*, 142561. https://doi.org/10.1016/j.scitotenv.2020.142561

Esteve-Pérez, S., Máñez-Castillejo, J. A., & Sanchis-Llopis, J. A. (2008). Does a "survival-by-exporting" effect for SMEs exist? *Empirica*, *35*(1), 81–104. https://doi.org/10.1007/s10663-007-9052-1

Food & Drug Administration. (2018). FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems. https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye

Gulsoy, N., & Kulluk, S. (2019). A data mining application in credit scoring processes of small and medium enterprises commercial corporate customers. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), 1–12. https://doi.org/10.1002/widm.1299

Hassan, H. S., Shehab, E., & Peppard, J. (2011). Recent advances in e-service in the public sector: State-of-the-art and future trends. *Business Process Management Journal*, *17*(3), 526–545. https://doi.org/10.1108/14637151111136405

Ho, C. W. L., Soon, D., Caals, K., & Kapur, J. (2019). Governance of automated image analysis and artificial intelligence analytics in healthcare. *Clinical Radiology*, *74*(5), 329–337. https://doi.org/10.1016/j.crad.2019.02.005

IAPMEI. (2020). *Incentivos Portugal 2020*. https://www.iapmei.pt/PRODUTOS-E-SERVICOS/Incentivos-Financiamento/Sistemas-de-Incentivos/Incentivos-Portugal-2020.aspx

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2020). Applying the CRISP-DM Framework for Teaching Business Analytics. *Decision Sciences Journal of Innovative Education*, *18*(4), 612–634. https://doi.org/10.1111/dsji.12222

Kalayci, S., & Arslan, S. (2017). Early NPLWarning for SME credit risk: An experimental study. *IC3K 2017 - Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *1*(Kdir), 190–197. https://doi.org/10.5220/0006496601900197

Kankanhalli, A., Charalabidis, Y., & Mellouli, S. (2019). IoT and AI for Smart Government: A Research Agenda. *Government Information Quarterly*, *36*(2), 304–309. https://doi.org/10.1016/j.giq.2019.02.003

Kim, H. S., & Sohn, S. Y. (2010). Support vector machines for default prediction of SMEs based on technology credit. *European Journal of Operational Research*, *201*(3), 838–846. https://doi.org/10.1016/j.ejor.2009.03.036

Kraemer-Eis, H., & Lang, F. (2017). Access to funds: how could CMU support SME financing? *Vierteljahrshefte Zur Wirtschaftsforschung*, *86*(1), 95–110. https://doi.org/10.3790/vjh.86.1.95

Lee, N., Sameen, H., & Cowling, M. (2015). Access to finance for innovative SMEs since the financial crisis. *Research Policy*, *44*(2), 370–380. https://doi.org/10.1016/j.respol.2014.09.008

Lewandowska, A., Stopa, M., & Humenny, G. (2015). The European Union Structural Funds and Regional Development. The Perspective of Small and Medium Enterprises in Eastern Poland. *European Planning Studies*, *23*(4), 785–797. https://doi.org/10.1080/09654313.2014.970132

Love, J. H., & Roper, S. (2015). SME innovation, exporting and growth: A review of existing evidence. *International Small Business Journal: Researching Entrepreneurship*, *33*(1), 28–48. https://doi.org/10.1177/0266242614550190

Love, J. H., Roper, S., & Hewitt-Dundas, N. (2010). Service innovation, embeddedness and business performance: Evidence from Northern Ireland. *Regional Studies*, *44*(8), 983–1004. https://doi.org/10.1080/00343400903401568

Lu, J. W., & Beamish, P. W. (2001). The internationalization and performance of SMEs. *Strategic Management Journal*, *22*(6–7), 565–586. https://doi.org/10.1002/smj.184

Mellouli, M., Bouaziz, F., & Bentahar, O. (2020). E-government success assessment from a public value perspective. *International Review of Public Administration*, *25*(3), 153–174. https://doi.org/10.1080/12294659.2020.1799517

Pang, M. S. (2014). IT governance and business value in the public sector organizations - The role of elected representatives in IT governance and its impact on IT value in U.S. state governments. *Decision Support Systems*, *59*(1), 274–285. https://doi.org/10.1016/j.dss.2013.12.006

Payne, K. (2018). Artificial Intelligence: A Revolution in Strategic Affairs? *Survival*, *60*(5), 7–32. https://doi.org/10.1080/00396338.2018.1518374

Pissarides, F. (1999). Is lack of funds the main obstacles to growth? L. *Journal of Business Venturing*, *9026*(14), 519–539. http://ac.els-cdn.com/S0883902698000275/1-s2.0-S0883902698000275-main.pdf?_tid=4fda2c14-1fba-11e7-ab7d-00000aab0f6c&acdnat=1492027252_9abe81c3d618fe02896bc8676fe0b930

Ratnayake, R. M. C. (2014). Small and medium enterprises project finance:Identifying optimum settings of controllable factor. *International Journal of Applied Decision Sciences*, *7*(2), 136–150. https://doi.org/10.1504/IJADS.2014.060327

Rose, J., Persson, J. S., Heeager, L. T., & Irani, Z. (2015). Managing e-Government: Value positions and relationships. *Information Systems Journal*, *25*(5), 531–571. https://doi.org/10.1111/isj.12052

Rosenbusch, N., Brinckmann, J., & Bausch, A. (2011). Is innovation always beneficial? A meta-analysis of the relationship between innovation and performance in SMEs. *Journal of Business Venturing*, *26*(4), 441–457. https://doi.org/10.1016/j.jbusvent.2009.12.002

Soe, R. M., & Drechsler, W. (2018). Agile local governments: Experimentation before implementation. *Government Information Quarterly*, *35*(2), 323–335. https://doi.org/10.1016/j.giq.2017.11.010

Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., Laughlin, P., Machtynger, J., & Machtynger, L. (2020). Artificial intelligence (AI) in strategic marketing decision-making: a research agenda. *Bottom Line*, *33*(2), 183–200. https://doi.org/10.1108/BL-03-2020-0022

Tecuci, G. (2012). Artificial intelligence. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 168–180. https://doi.org/10.1002/wics.200

Toll, D., Lindgren, I., Melin, U., & Madsen, C. Ø. (2019). Artificial Intelligence in Swedish Policies: Values, Benefits, Considerations and Risks. In *Electronic Government* (pp. 301–310). https://doi.org/10.1007/978-3-030-27325-5

Twala, B. (2009). Combining classifiers for credit risk prediction. *Journal of Systems Science and Systems Engineering*, *18*(3), 292–311. https://doi.org/10.1007/s11518-009-5109-y

Twizeyimana, J. D., & Andersson, A. (2019). The public value of E-Government – A literature review. *Government Information Quarterly*, *36*(2), 167–178. https://doi.org/10.1016/j.giq.2019.01.001

Valle-Cruz, D., & Sandoval-Almazan, R. (2018). Towards an understanding of Artificial Intelligence in government. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3209281.3209397

Valle-Cruz, D., Sandoval-Almazan, R., Ruvalcaba-Gomez, E. A., & Ignacio Criado, J. (2019). A review of artificial intelligence in government and its potential from a public policy perspective. *ACM International Conference Proceeding Series*, 91–99. https://doi.org/10.1145/3325112.3325242

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration*, *42*(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Zhu, Y., Xie, C., Sun, B., Wang, G. J., & Yan, X. G. (2016). Predicting China's SME credit risk in supply chain financing by logistic regression, artificial neural network and hybrid models. *Sustainability (Switzerland)*, *8*(5), 433. https://doi.org/10.3390/su8050433

Zhu, Y., Xie, C., Wang, G. J., & Yan, X. G. (2017). Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, *28*(s1), 41–50. https://doi.org/10.1007/s00521-016-2304-x

Anexos e Apêndices

Anexo A

Tabela 57 - Tabelas - contagens e descrições

Ficheiro/Tabela	**Nº de Projetos**	**Descrição**
Activos	0	
AmbitoNov	561	caracteriza a inovação do projeto em três categorias - abertura a um novo mercado/industria (sim/não) - inovação ao nível do produto ou ao nível de mercado - inovação de mercado/região/mundo
Ameacas	10	caracteriza as ameaças ao projeto
Areas	5	contem as áreas de construção existentes ou a construir
Atividades	781	caracteriza a inovação: - tipo (4 tipos, aumento capacidade, alteração processo, etc) - marketing/processo/produto/organizacional -grau de novidade (empresa, nacional, internacional)
Balanco_SNC	781	valores do balanço desde t-3 até previsão t+7
CadeiaValor	561	Apresenta previsões de melhoria pós projetos de rubricas da cadeia de valor
Capacidade	0	
Concorrentes	10	lista de concorrentes com o tipo de empresa (comércio/industrial etc), localização, etc
CondFin	10	valores de conta s31 ou superiores para t-1
ConstrEdificios	7	discrimina os custos das construções de edifícios dos projetos (preço por m2, área, descrição)
Cursos	104	caracteriza os cursos associados à candidatura: - Area de formação - Horas
DesafiosSocietais	781	Desafios societais que os projetos iram responder contém um conjunto de categorias definidas (proteção climática, transportes, saúde, etc)
Dividas	9	caracteriza as dividas (valor, entidades, etc)
DividasFinanc	3	layout igual ao csv “dividas”, acrescenta algumas dívidas
DominiosPrioritario	765	caracteriza os domínios da candidatura: - domínio (TI, energia, automóvel, etc)

		- área (área dentro do domínio)
DominiosPrioritariosAlentejo	20	caracteriza os domínios da candidatura no Alentejo
DominiosPrioritariosAlgarve	1	caracteriza os domínios da candidatura no Algarve
DominiosPrioritariosCentro	90	caracteriza os domínios da candidatura no Centro
DominiosPrioritariosLisboa	8	caracteriza os domínios da candidatura no Lisboa
DominiosPrioritariosNorte	101	caracteriza os domínios da candidatura no Norte
DR_SNC	781	valores da demonstração de resultados desde t-3 até previsão t+7
EconomiaCircular	10	caracteriza os projetos em termos economia circualr
EstudosDiagnosticos	2	caracteriza os estudos no projeto (tipo, custo, fornecedor)
FormadoresExt	102	caracteriza formadores externos envolvidos no projeto - origem (Nacional/internacional) ; - custo hora ;- entidade
FormadoresInt	6	caracteriza formadores internos envolvidos no projeto - nº horas ; - custo hora
Formandos	96	caracteriza os formandos envolvidos no projeto
FSE	233	custos gerais com formadores/formandos -custos totais -custos t+1 até t+7 se aplicável
Geral	781	informações gerais do projeto
IndicadoresIDT	781	identifica a existência de despesas do projeto com I&D e quais esses custos em t+1 até t+7
IndicCertif	632	Indica quais os certificados (existentes/ que se pretende adquirir/ etc) - tipo de certificado (iso/segurança alimentar, etc) - pre proj. vs pos proj (sim/não)
IndicPress	1	valores de contas
Industria40_i	10	identifica o tipo de tecnologia envolvida no projeto
Industria40_ii	10	identifica o tipo de tecnologia envolvida no projeto
Industria40_iii	10	identifica o tipo de tecnologia envolvida no projeto
Inv	781	identifica os investimentos previstos: - valor elegível - classe/tipo de investimento - localidade e classe

ListaAccoes	551	informação sobre os formandos (horas, nível, curso,etc)
ListaUploads	220	url para ficheiros que foram carregados na candidatura
MajorSocios	0	percentagem de sócios maioritarios
MapaOrigem	10	tem alguns valores sobre compras e fse
MarcasOutras	177	percentagem detida noutras marcas
MarcasProprias	393	identifica marcas detidas e comercializadas e percentagens
MaterialCirculante	0	não identificável
Mercados	781	informação das vendas/compras por países: - tipo (compra/venda) - bem servido - país - valores
Mercados2	781	Informação de mercados - nacional/internacional - valor pré e pós projeto (percentagem)
Mercados3	781	Informação de mercados - país - valor pré e pós projeto (percentagem)
Oportunidades	10	Identifica o tópico de oportunidade
OutrosFinanc	1	identifica financiamentos
Part	10	identifica participações na empresa (nif, percentagem, país e se detêm controlo)
PontosFortes	10	identifica pontos fortes do projeto
PontosFracos	10	identifica pontos fracos do projeto
ProjCae	781	identifica o/os CAES's do projeto e as percentagens
PromCae	781	identifica o/os CAE's do promotor e as percentagens
PromLocal	781	identifica a localização das instalações (velhas e novas)
PromLocalTip	615	identifica a localização das instalações (nuts)
Propensao	166	-vendas internacionais sustentáveis em mercados internacionais sustentáveis (sim/não) - elevada vocação de vendas internacional (sim/não) - vendas internacionais (diretas/intermedias/estruturantes) - vendas internacionais serão marca própria/terceiros
PTrabalho	781	informação sobre os trabalhadores da empresa: sexo, qualificações, numero
PTrabalhoResumo	10	outras informações sobre os trabalhadores da empresa: sexo, qualificações, numero
QuadrosTecn	0	identifica os quadros técnicos da empresa (remuneração/nível/segurança_socia

RefBancarias	1	identifica a data e origem das contas
Reforço	166	Classificação de 1-5 no pré e pós projetos: - Utilização ferramentas de marketing - Modelo de gestão orientado para a inovação - Qualidade recursos humanos - Parcerias I&D (entidades não empresariais) - Sofisticação dos Processos Produtivos
ResultadosPO	781	Classificação de sim/não no pré e pós projetos: - Contribui para empregabilidade sustentada - Fortalecimento de coesão e inclusão social - Uso sustentável de recursos - Contributo do projeto para um Portugal dinâmico, exportador, competitivo, internacional
Socios	781	Identifica os sócios (tipo/nif/percentagem)
SubstImp	0	valores de importações e compras (pré e pós projeto)
TabProj	2	informações sobre o projeto (datas e designações)
TipoInov	561	caracteriza o tipo de inovação do projeto para cada tipo (Produto/Processo/Marketing/Organizacional) - sustentação (interna/externa) - perceção (0-5) e impacto (0-5)
Tipologia	781	Identificação do tipo de projeto (criação/aumento de estabelecimento ; melhoria de produto; etc)
TransfTecnologia	1	informações sobre transferências de tecnologia
TransicaoEnergetica	10	informação da mudança de energia utilizada
VantagensCompA1	561	pre e pós projeto em diversas rubricas (gama produtos/preços/etc)
VendasExt	185	define o valor das vendas externas pré e pós projetos

Tabela 58 - Variáveis escolhidas 3.5.2

Variável	**Localização**
nut_norte	Resumo/Nute_Norte
nut_centro	Resumo/Nute_Centro
nut_lisboa	Resumo/Nute_Lisboa
nut_alentejo	Resumo/Nute_Alentejo
nut_algarve	Resumo/Nute_Algarve
inicio_atividade	Resumo/Dt_Inicio_Act
ano_candidatura	Parametros/Ano_Cand
prom_nat_jur	Promotor/Nat_Jur
prom_capital_social	Promotor/Cap_Social
prom_url_empresa	Promotor/Url
consultora	Consultora/Nif
vantagens_comp_est	Vantagenscomp/Estrategia
analise_mercados_dir	Analisemercados/Direcao
proj_n_meses	Dadosprojecto/N_Meses
proj_elegivel	Dadosprojecto/Elegivel
proj_investimento	Dadosprojecto/Investimento
major_plano_acao	Majoracoes/Plano_Acao
major_sustenta	Majoracoes/Sustentab

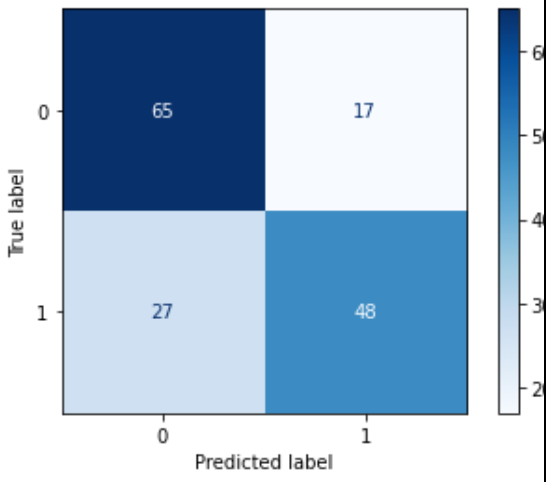
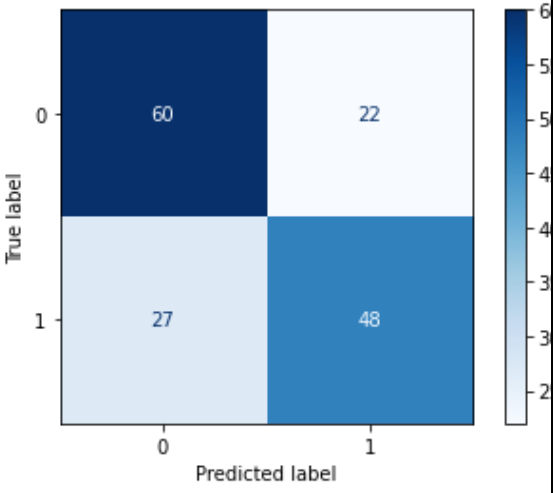
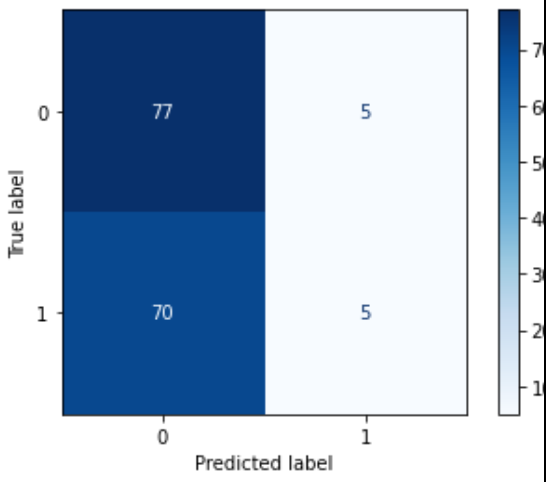
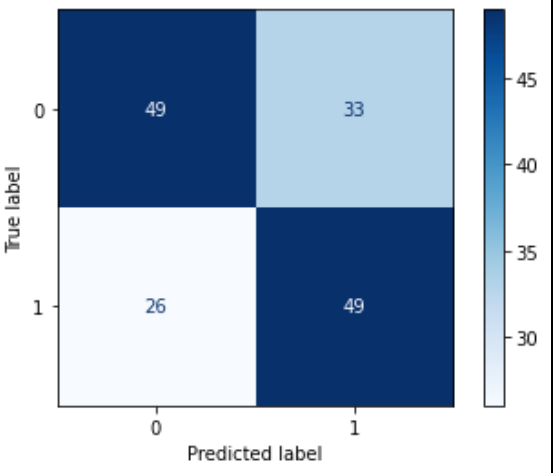
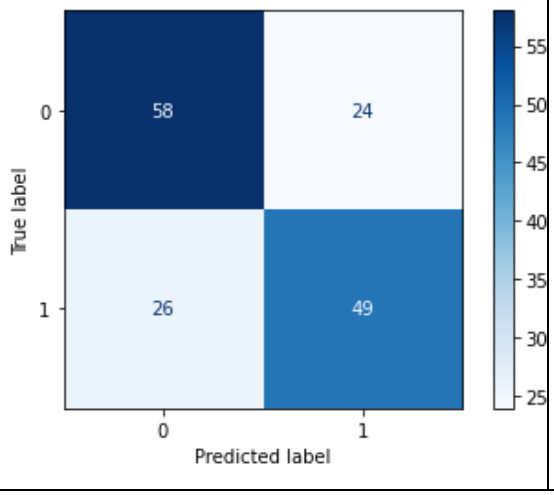
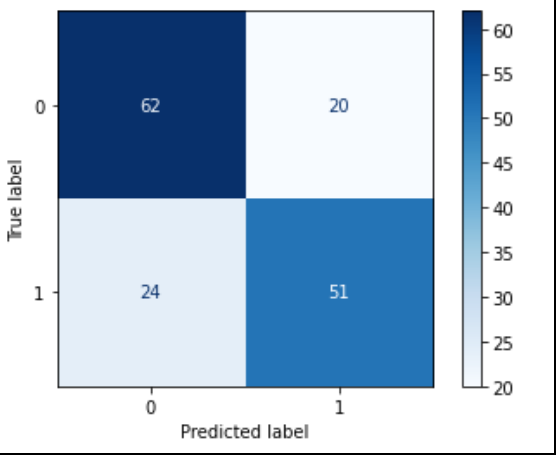
Tabela 59 - Pré-processamento variáveis 3.5.2.

variável	**get_dummies**	**qcut**	**cut**	**fill_na**	**exist**	**Tratamento especial**
nut_norte						
nut_centro						
nut_lisboa						
nut_alentejo						
nut_algarve						
inicio_atividade ano_candidatura	x	x				Criou-se a variável- "idade na candidatura" -> "ano_candidatura" - "inicio de atividade"
prom_nat_jur	x					
prom_capital_social		x		x		
prom_url_empresa				x	x	
consultora				x	x	
vantagens_comp_est	x					
analise_mercados_dir	x			x		
proj_n_meses			x			
proj_elegivel		x				
proj_investimento	x		x			Criou-se a variável "percent_eleg_invest" -> proj_elegivel / proj_investimento
major_plano_acao				x	x	
major_sustenta				x	x	

Tabela 60 - Resultados modelação das variáveis gerais 3.52

	ROC_AUC	Accuracy
SVM	0.72	0.72
RF	0.69	0.69
GNB	0.50	0.52
MLP	0.63	0.62
XGBoost	0.68	0.68
LR	0.72	0.72
DT	0.63	0.63

Tabela 61 - Matrizes confusão 3.5.2

SVM	Random Forest
	
Gaussian NB	**MLP**
	
XGboost	**Logistic Regression**
	

Decision Tree

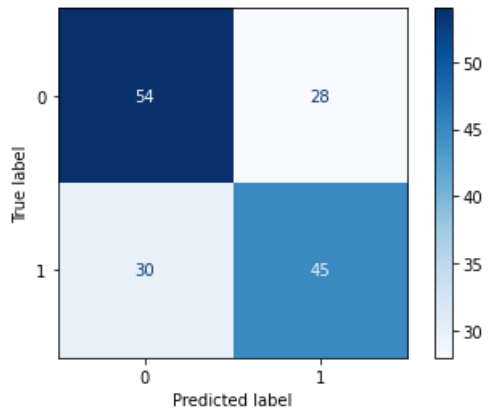


Tabela 62 - Resultados k-fold modelação das variáveis gerais 3.5.2

	ROC_AUC	Accuracy
SVM	0.73	0.69
RF	0.72	0.70
GNB	0.70	0.58
MLP	0.67	0.64
XGBoost	0.70	0.65
LR	0.73	0.70
DT	0.59	0.59

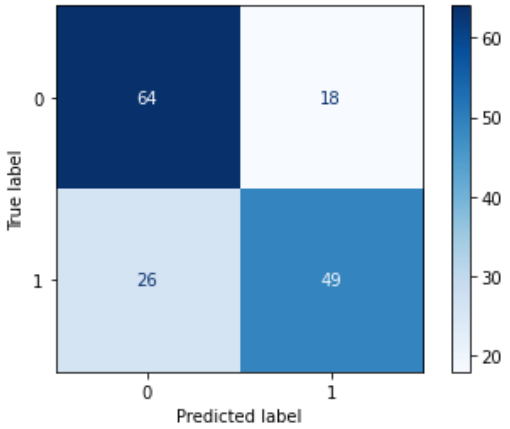
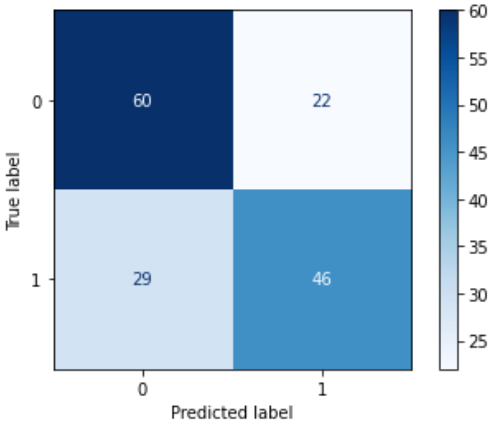
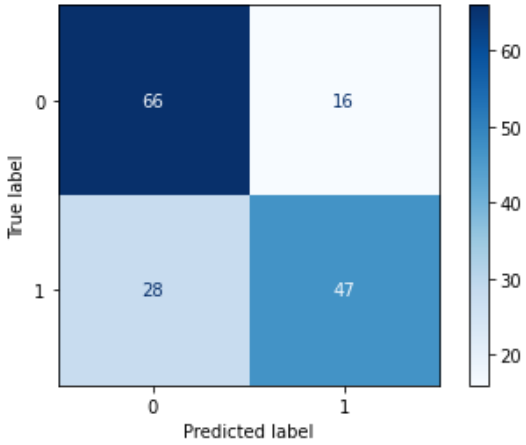
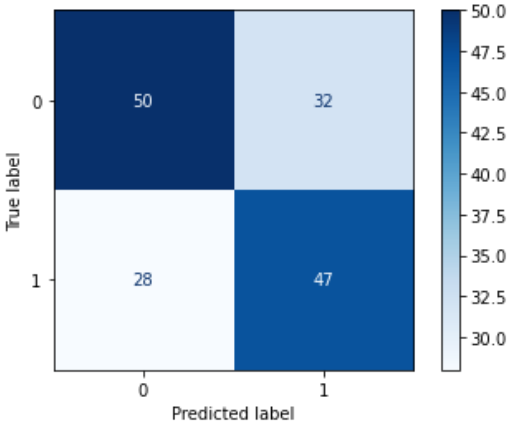
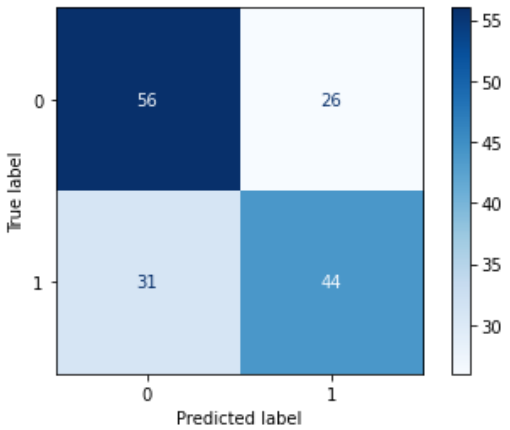
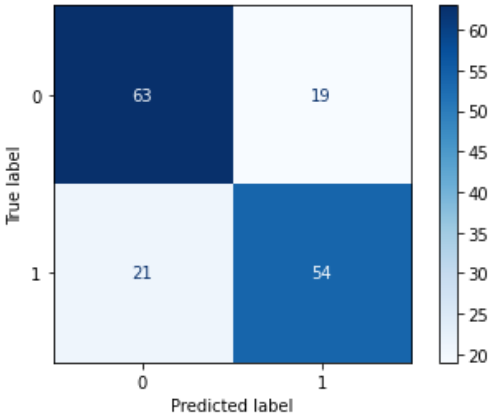
Tabela 63 - Importância de features da experiência 3.5.2

SVM			
n_meses_2	-1.32	am_dir_0	1.46
nut_centro	-1.20	capital_social_1	0.91
nut_norte	-1.20	nut_lisboa	0.71
am_dir_2	-0.41	idade_candidatura_1	0.59
am_dir_4	-0.38	n_meses_4	0.50
nut_alentejo	-0.37	n_meses_3	0.44

Random Forest	
am_dir_0	0.05
major_sustenta	0.04
prom_url_empresa	0.04
n_meses_4	0.04
capital_social_1	0.04
consultora	0.03
vant_comp_est_2.0	0.03
nut_norte	0.03
idade_candidatura_1	0.03
major_plano_acao	0.03
vant_comp_est_4	0.03
nut_centro	0.02

LR			
nut_centro	-0.85	am_dir_0	1.05
nut_norte	-0.81	nut_lisboa	0.74
n_meses_2	-0.73	n_meses_4	0.73
am_dir_3	-0.50	capital_social_1	0.62
am_dir_2	-0.44	capital_social_2	0.32
capital_social_5	-0.35	idade_candidatura_1	0.03

Tabela 64 - Matrizes confusão variáveis otimizadas 3.5.3

SVM	Random Forest
	
Gaussian NB	**MLP**
	
XGboost	**Logistic Regression**
	

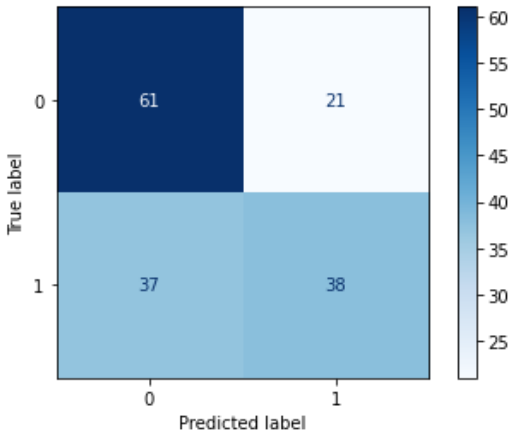
Decision Tree


Tabela 65 - Variáveis dos rácios promotor e as suas descrições

Rácios - variáveis criadas	**Descrição**
prom_n_trabalhadores_t-1	Número de Trabalhadores do Promotor em t-1
prom_volume_negocios_t-1	Volume de Negócios do Promotor em t-1
prom_ativo_total_t-1	Ativo Total do Promotor em t-1
prom_emprestimos_obtidos_passivo_cor_t-1	Empréstimos Obtidos Passivo Corrente do Promotor em t-1
prom_emprestimos_obtidos_passivo_ncor_t-1	Empréstimos Obtidos Passivo Não Corrente do Promotor em t-1
prom_autonomia_financeira_t-1	Autonomia Financeira do Promotor em t-1
prom_ebitda/vn_t-1	EBITDA do Promotor em t-1
prom_resultado_liquido/vn_t-1	Resultado Líquido do Volume de Negócios do Promotor em t-1
prom_alavancagem_financeira_t-1	Alavancagem Financeira do Promotor em t-1
prom_vab/trabalhador_t-1	VAB Trabalhador do Promotor em t-1
prom_liquidez_geral_t-1	Liquidez Geral do Promotor em t-1
prom_rentabilidade_capitais_proprios_t-1	Rentabilidade dos Capitais Próprios do Promotor em t-1
prom_cmvmc/fornecedores_t-1	CMVMC Fornecedores do Promotor em t-1
prom_cmvmc/inventario_t-1	CMVMC Inventário do Promotor em t-1
prom_resultadoliquido/ativo_t-1	Resultado Líquido Ativo do Promotor em t-1
prom_vn/clientes_t-1	Volume de Negócios Clientes do Promotor em t-1
prom_crescimento_vn_t-1	Crescimento do Volume de Negócios do Promotor em t-1
prom_dsri_t-1	DSRI do Promotor em t-1
prom_aqi_t-1	AQI do Promotor em t-1

prom_depi_t-1	DEPI do Promotor em t-1
prom_sgai_t-1	SGAI do Promotor em t-1
prom_lvgi_t-1	LVGI do Promotor em t-1
prom_tata_t-1	TATA do Promotor em t-1
prom_mscore_t-1	MSCORE do Promotor em t-1
prom_n_trabalhadores_t-2	Número de Trabalhadores do Promotor em t-2
prom_volume_negocios_t-2	Volume de Negócios do Promotor em t-2
prom_ativo_total_t-2	Ativo Total do Promotor em t-2
prom_emprestimos_obtidos_passivo_cor_t-2	Empréstimos Obtidos Passivo Corrente do Promotor em t-2
prom_emprestimos_obtidos_passivo_ncor_t-2	Empréstimos Obtidos Passivo Não Corrente do Promotor em t-2
prom_autonomia_financeira_t-2	Autonomia Financeira do Promotor em t-2
prom_ebitda/vn_t-2	EBITDA do Promotor em t-2
prom_resultado_liquido/vn_t-2	Resultado Líquido do Volume de Negócios do Promotor em t-2
prom_alavancagem_financeira_t-2	Alavancagem Financeira do Promotor em t-2
prom_vab/trabalhador_t-2	VAB Trabalhador do Promotor em t-2
prom_liquidez_geral_t-2	Liquidez Geral do Promotor em t-2
prom_rentabilidade_capitais_proprios_t-2	Rentabilidade dos Capitais Próprios do Promotor em t-2
prom_cmvmc/fornecedores_t-2	CMVMC Fornecedores do Promotor em t-2
prom_cmvmc/inventario_t-2	CMVMC Inventário do Promotor em t-2
prom_resultadoliquido/ativo_t-2	Resultado Líquido Ativo do Promotor em t-2
prom_vn/clientes_t-2	Volume de Negócios Clientes do Promotor em t-2
prom_crescimento_vn_t-2	Crescimento do Volume de Negócios do Promotor em t-2
prom_dsri_t-2	DSRI do Promotor em t-2
prom_aqi_t-2	AQI do Promotor em t-2
prom_depi_t-2	DEPI do Promotor em t-2
prom_sgai_t-2	SGAI do Promotor em t-2
prom_lvgi_t-2	LVGI do Promotor em t-2
prom_tata_t-2	TATA do Promotor em t-2
prom_mscore_t-2	MSCORE do Promotor em t-2

Tabela 66 - Resultados usando variáveis de rácios em t-1 e em t-1 + t-2

	AUC		Accuracy		Precision	
Rácios	**t-1**	**t-1 e t-2**	**t-1**	**t-1 e t-2**	**t-1**	**t-1 e t-2**
LR	0.80	0.80	0.75	0.75	0.75	0.75
SVM	0.79	0.77	0.72	0.71	0.69	0.66
GNB	0.79	0.78	0.73	0.73	0.71	0.70
RF	0.80	0.79	0.77	0.75	0.76	0.74
MLP	0.79	0.79	0.74	0.72	0.71	0.68
XGB	0.80	0.79	0.75	0.74	0.75	0.73
DT	0.67	0.80	0.65	0.75	0.59	0.75

Tabela 67 - Variáveis escolhidas no exercício dos rácios

Varíaveis escolhidas
n_meses_2
n_meses_4
am_dir_0
resultado_po_2
reforco_2
dominio_3
formadores_ext
n_cursos
nut_norte
n_marcas_outras_nan
n_marcas_outras
n_trabalhadores
tipo_inov_organizacional_nan
ativ_tipo_2
contagem_mercados
mais_nacional_pos
vant_comp_15
vant_comp_32
vant_comp_33
vant_comp_34
vant_comp_37
prom_cmvmc/inventario_t-1
prom_ebitda/vn_t-1
prom_n_trabalhadores_t-1
prom_emprestimos_obtidos_passivo_ncor_t-1
prom_aqi_t-1
prom_volume_negocios_t-1
prom_cmvmc/fornecedores_t-1
prom_ativo_total_t-1
prom_autonomia_financeira_t-1
prom_resultado_liquido/vn_t-1
prom_tata_t-1
prom_lvgi_t-1

prom_sgai_t-1
prom_emprestimos_obtidos_passivo_cor_t-1

Tabela 68 - Importância de features do exercicio dos rácios, para LR, RF, XGBoost e SVM

LR			
n_meses_2	-0.45	formadores_ext	0.68
prom_ebitda/vn_t-1	-0.44	n_cursos	0.61
ativ_tipo_2	-0.43	resultado_po_2	0.49
vant_comp_34	-0.31	tipo_inov_organizacional_nan	0.47
prom_cmvmc/inventario_t-1	-0.31	n_meses_4	0.43
dominio_3	-0.30	vant_comp_37	0.29

Random Forest	
prom_ativo_total_t-1	0.05
prom_ebitda/vn_t-1	0.04
n_cursos	0.04
prom_n_trabalhadores_t-1	0.03
formadores_ext	0.03
prom_volume_negocios_t-1	0.03
resultado_po_2	0.03
prom_autonomia_financeira_t-1	0.02
prom_resultado_liquido/vn_t-1	0.02
n_trabalhadores	0.02
vant_comp_32	0.02
prom_tata_t-1	0.02

XGBoost	
resultado_po_2	0.04
prom_n_trabalhadores_t-1	0.03
formadores_ext	0.03
prom_ativo_total_t-1	0.03
n_cursos	0.03
reforco_2	0.02
am_dir_0	0.02
tipo_inov_organizacional_nan	0.02
prom_emprestimos_obtidos_passivo_ncor_t-1	0.02
prom_aqi_t-1	0.02
vant_comp_15	0.01
mais_nacional_pre	0.01

SVM			
prom_n_trabalhadores_t-1	-3.32	n_marcas_outras	5.33
prom_cmvmc/inventario_t-1	-2.24	n_marcas_outras_nan	5.00
n_meses_2	-2.22	prom_volume_negocios_t-1	2.81
prom_ebitda/vn_t-1	-2.12	prom_cmvmc/fornecedores_t-1	2.39
nut_norte	-2.07	n_cursos	2.37
contagem_mercados	-1.93	vant_comp_32	2.35

Tabela 69 - Variáveis escolhidas para os modelos de cada grupo

Grupo 1		Grupo 2
consultora		n_meses_2
n_meses_2		n_meses_4
idade_candidatura_1		nut_centro
prom_url_empresa		nut_alentejo
vant_comp_24		nut_norte
nut_alentejo		n_socios
cadeia_valor_41		contagem_mercados
marcas_propria_N		formadores_ext
vant_comp_est_2.0		n_cursos
tipo_inov_organizacional_nan		cadeia_valor_32
ativ_inov_marketing		am_dir_0
novo_produtos_mercados		am_dir_4
resultado_po_2		vant_comp_15
formadores_ext		vant_comp_36
n_cursos		desafio_1
valor_formacoes		reforco_1
prom_vn/clientes_t-1		resultado_po_2
prom_ebitda/vn_t-1		n_marcas_outras_nan
prom_resultado_liquido/vn_t-1		n_marcas_outras
prom_rentabilidade_capitais_proprios_t-1		prom_emprestimos_obtidos_passivo_ncor_t-1
prom_resultadoliquido/ativo_t-1		prom_volume_negocios_t-1
prom_n_trabalhadores_t-1		prom_ebitda/vn_t-1
prom_dsri_t-1		prom_lvgi_t-1
prom_vab/trabalhador_t-1		prom_dsri_t-1
prom_ativo_total_t-1		prom_ativo_total_t-1
prom_volume_negocios_t-1		prom_rentabilidade_capitais_proprios_t-1
prom_depi_t-1		prom_autonomia_financeira_t-1
prom_crescimento_vn_t-1		prom_resultado_liquido/vn_t-1
		prom_resultadoliquido/ativo_t-1
		prom_n_trabalhadores_t-1

Tabela 70 - Variáveis dos rácios consultor e as suas descrições

Rácios - variáveis criadas	Descrição
consul_alavancagem_financeira_t-1	Alavancagem Financeira da Consultora em t-1
consul_ano_t-1	T-1 Relativo ao Ano do Último Ppi (inelegibilidades) ou da Candidatura do Projeto Associado à Consultora (anulações)
consul_aqi_t-1	AQI da Consultora em t-1
consul_ativo_total_t-1	Ativo Total da Consultora em t-1
consul_autonomia_financeira_t-1	Autonomia Financeira da Consultora em t-1
consul_CAE_t-1	CAE da Consultora em t-1
consul_CAE_t-2	CAE da Consultora em t-2
consul_cmvmc/fornecedores_t-1	CMVMC Fornecedores da Consultora em t-1
consul_cmvmc/inventario_t-1	CMVMC Inventário da Consultora em t-1
consul_crescimento_vn_t-1	Crescimento do Volume de Negócios da Consultora em t-1
consul_depi_t-1	DEPI da Consultora em t-1
consul_dsri_t-1	DSRI da Consultora em t-1
consul_ebitda/vn_t-1	EBITDA da Consultora em t-1
consul_emprestimos_obtidos_passivo_cor_t-1	Empréstimos Obtidos Passivo Corrente da Consultora em t-1
consul_emprestimos_obtidos_passivo_ncor_t-1	Empréstimos Obtidos Passivo Não Corrente da Consultora em t-1
consul_falta_IES_t-1	Temos IES disponível para este consultor para t-1?
consul_liquidez_geral_t-1	Liquidez Geral da Consultora em t-1
consul_lvgi_t-1	LVGI da Consultora em t-1
consul_mscore_t-1	MSCORE da Consultora em t-1
consul_n_trabalhadores_t-1	Número de Trabalhadores da Consultora em t-1
consul_rentabilidade_capitais_proprios_t-1	Rentabilidade dos Capitais Próprios da Consultora em t-1
consul_resultado_liquido/vn_t-1	Resultado Líquido do Volume de Negócios da Consultora em t-1
consul_resultadoliquido/ativo_t-1	Resultado Líquido Ativo da Consultora em t-1
consul_sgai_t-1	SGAI da Consultora em t-1
consul_tata_t-1	TATA da Consultora em t-1
consul_vab/trabalhador_t-1	VAB Trabalhador da Consultora em t-1
consul_clientes/vn_t-1	Volume de Negócios Clientes da Consultora em t-1
consul_volume_negocios_t-1	Volume de Negócios da Consultora em t-1

Tabela 71 - Feature Importance experiência consultor 4.5

LR			
n_meses_2	-1.10	vant_comp_32	1.24
consul_cmvmc/inventario_t-1	-1.09	formadores_ext	1.19
reforco_2	-1.08	n_cursos	1.14
prom_ebitda/vn_t-1	-1.07	consul_vab/trabalhador_t-1	0.93
vant_comp_34	-0.95	prom_lvgi_t-1	0.91
prom_cmvmc/inventario_t-1	-0.81	tipo_inov_organizacional_nan	0.87

SVM			
consul_ativo_total_t-1	-2.03	n_marcas_outras	4.62
consul_cmvmc/inventario_t-1	-2.01	n_marcas_outras_nan	4.38
consul_rentabilidade_capitais_proprios_t-1	-1.99	n_cursos	3.09
prom_ebitda/vn_t-1	-1.62	vant_comp_32	1.99
prom_n_trabalhadores_t-1	-1.60	consul_vab/trabalhador_t-1	1.65
reforco_2	-1.29	consul_ebitda/vn_t-1	1.54

Random Forest	
prom_ativo_total_t-1	0.05
prom_ebitda/vn_t-1	0.04
prom_n_trabalhadores_t-1	0.04
formadores_ext	0.04
n_cursos	0.04
resultado_po_2	0.03
prom_lvgi_t-1	0.03
prom_sgai_t-1	0.03
prom_autonomia_financeira_t-1	0.03
prom_emprestimos_obtidos_passivo_ncor_t-1	0.03
prom_volume_negocios_t-1	0.03
consul_vn/clientes_t-1	0.02

XGBoost	
resultado_po_2	0.06
prom_ativo_total_t-1	0.05
formadores_ext	0.05
prom_n_trabalhadores_t-1	0.04
n_cursos	0.04
prom_aqi_t-1	0.03
am_dir_0	0.03

n_meses_2	0.02
prom_volume_negocios_t-1	0.02
prom_emprestimos_obtidos_passivo_ncor_t-1	0.02
prom_ebitda/vn_t-1	0.02
consul_vn/clientes_t-1	0.02

Tabela 72 - RFE variáveis escolhidas - Accuracy Grupo 2

LR
n_meses_2
am_dir_0
n_cursos
formadores_ext
vant_comp_24
sim_max

SVM
n_meses_2
am_dir_3
am_dir_4
nut_alentejo
n_cursos
prom_emprestimos_obtidos_passivo_cor_t-1
prom_falta_IES_t-1
direto

Tabela 73 - RFE variáveis escolhidas - Recall Grupo 1

LR
am_dir_3
despesas_idt
emp_menos_4_anos_cand

Tabela 74 - RFE variáveis escolhidas - Recall Grupo 2

LR
n_meses_2
n_meses_4
am_dir_0
nut_norte
nut_centro
nut_alentejo
ativ_tipo_1
ativ_inov_marketing

tipo_inov_organizacional_nan
n_cursos
formadores_ext
vant_comp_15
vant_comp_24
vant_comp_32
prom_emprestimos_obtidos_passivo_cor_t-1
prom_ebitda/vn_t-1
prom_sgai_t-1
consul_emprestimos_obtidos_passivo_ncor_t-1
prom_falta_IES_t-1
sim_max
topic1
topic10

SVM
n_meses_2
am_dir_3
am_dir_4
nut_alentejo
nut_lisboa
n_cursos
novo_produtos_mercados
contagem_mercados
pais_extra_ue
vant_comp_15
vant_comp_32
vant_comp_36
prom_emprestimos_obtidos_passivo_cor_t-1
prom_ebitda/vn_t-1
prom_tata_t-1
consul_ebitda/vn_t-1
consul_rentabilidade_capitais_proprios_t-1
prom_falta_IES_t-1
emp_recente
sim_max
direto
topic5
topic10

Tabela 75 - Hiperparâmetros escolhidos 4.7 - SVM

SVM	parametros
C	1
gamma	1
kernel	linear

Tabela 76 - Hiperparâmetros escolhidos 4.7 - LR

LR	parametros
C	0.1
max_iter	100
penalty	l1
solver	liblinear

Tabela 77 - Hiperparâmetros escolhidos 4.7 - RF

RF	parametros
max_depth	10
max_features	auto
min_samples_leaf	1
min_samples_split	10
n_estimators	100
criterion	entropy

Tabela 78 - Hiperparâmetros escolhidos 4.7 - XGBoost

XGBoost	parametros
learning_rate	0.01
max_depth	10
min_child_weight	1
subsample	0.5
colsample_bytree	0.7
n_estimators	200