Mariana Isabel Ferreira Páscoa

Os desafios da *Machine Learning*: Aplicação ao Mercado Financeiro

Relatório de Estágio Curricular em Economia, na especialidade de Economia Industrial, apresentada à Faculdade de Economia da Universidade de Coimbra para obtenção do grau de Mestre

Julho, 2018

· U C ·

UNIVERSIDADE DE COIMBRA

Mariana Isabel Ferreira Páscoa

Os desafios da *Machine Learning*: Aplicação ao Mercado Financeiro

Relatório de Estágio Curricular em Economia, na especialidade de Economia Industrial, apresentada à Faculdade de Economia da Universidade de Coimbra para obtenção do grau de Mestre

Orientador de estágio na FEUC: Prof. Doutor Pedro Godinho

Orientador na entidade: Dr. Filipe Neves

Coimbra, 2018

Agradecimentos

O meu percurso académico na Faculdade de Economia da Universidade de Coimbra foi marcado por vários desafios. O presente relatório simboliza o fim de uma etapa e não podia terminar esta sem agradecer às pessoas que estiveram desde o primeiro momento ao meu lado.

Quero agradecer ao Professor Doutor Pedro Godinho por todos os conselhos, pela disponibilidade, pelas recomendações e sugestões fundamentais para a elaboração do presente relatório. O meu muito obrigada.

À Feedzai por possibilitar a concretização do meu estágio. Foram meses onde cresci tanto profissionalmente como pessoalmente. Obrigada por todos os ensinamentos e experiência partilhadas.

À minha família, por todo o apoio incondicional durante todo o meu percurso académico, pela força que me transmitiram nos momentos mais difíceis. Doravante espero compensá-los das horas de atenção que lhes devo. Ao meu namorado, Gonçalo, pela compreensão e palavras de conforto ao longo dos cinco anos universitários. A eles ficarei eternamente grata.

Aos meus amigos e colegas de faculdade que me ajudaram a alcançar os meus objetivos. Sem eles tudo teria sido mais difícil.

Índice

Índice de Figuras

Índice de Tabelas

Lista de Siglas

AD- Árvores de Decisão

ANN – *Artificial Neural Network*

CEO - *Chief Executive Officer*

CFO – *Chief Financial Officer*

CSO - *Chief Science Officer*

CTO - *Chief Technology Officer*

GAM – *Generalized Additive Models*

KPI - *Key Performance Indicators*

IT - *Information Technology*

ML- *Machine learning*

SVM -*Support Vector Machine*

PO- *Purchase Order*

PSI - *Portuguese Stock Index*

RF- *Random Forests*

WEKA - *Waikato Environment for Knowledge Analysis*

Resumo

O presente relatório surge na sequência do estágio curricular, decorrido entre 15 de fevereiro de 2018 a 24 de maio de 2018, na Feedzai – Consultoria e Inovação Tecnológica, S.A. com vista à obtenção do grau de Mestre em Economia, pela Faculdade de Economia de Universidade de Coimbra.

O tema central neste trabalho é a *machine learning* (ML). Há várias definições para o que constitui ML, parecendo particularmente apropriada aquela que a define como um processo que permite que os computadores melhorem o seu desempenho numa determinada tarefa através da experiência.

A *machine learning* tem vido a ganhar popularidade. Ao longo das últimas décadas tem contribuído positivamente em vários sectores de atividade. Com a aplicação e conceitos da ML é possível aumentar a produtividade e reduzir custos, detetar fraudes, melhorar o desempenho preditivo e, desta forma, aperfeiçoar o processo de tomada de decisão. Estas são algumas das potencialidades de ML. Nomeadamente a análise preditiva tem vindo a ser considerada como um factor importante para assegurar a vantagem competitiva nas empresas.

Subjacente à escolha do tema, está a curiosidade por saber mais sobre o que é a ML, como ela está a mudar a nossa vida e como ela funciona. Este motivo associado ao facto da Feedzai ser uma empresa de base tecnológica e utilizadora de várias técnicas de ML, o que me suscitou interesse em escrever sobre ML aplicada à minha área de formação base, economia. Ao longo deste trabalho é discutida a importância deste tema no âmbito dos mercados financeiros.

O objetivo deste relatório é descrever as minhas tarefas realizadas no estágio e a respetiva análise crítica e estudar a aplicação das técnicas de ML, nomeadamente as árvores de classificação e *random forests,* através de um caso empírico. O propósito deste caso empírico é testar a capacidade preditiva no mercado financeiro da bitcoin e do PSI 20.

De acordo com os resultados obtidos, não é possível eleger uma técnica de ML melhor pois cada técnica adapta-se melhor consoante as características da base de dados, ou seja, os resultados indicam que a técnica das árvores de classificação oferece melhores

resultados no mercado PSI 20 enquanto o algoritmo *random forests* obtém melhores resultados junto do mercado da bitcoin. Os resultados indicam também que as estratégias com base em ML obtêm, geralmente, melhores resultados do que a estratégia *buy & hold* quando aplicadas ao PSI 20, mas o mesmo não se passa quando são aplicadas ao mercado das bitcoins.

Palavras-chave

Machine learning; Árvores de Classificação; *Random Forests;* Mercados Financeiros

Abstract

The current report is presented as the result of curricular internship, held between 15 February 15, 2018 and May 24, 2018, at Feedzai – Consultoria e Inovação Tecnológica, S.A. with the purpose of obtaining a Master's Degree in Economics from the Faculty of Economics from University of Coimbra.

Machine learning (ML) stands as the main topic of the present work. There are several definitions for what constitutes ML, and it seems particularly appropriate to define it as a process that allows computers to improve their performance in a given task through experience.

Machine learning has been gathering notoriety in the last years. Over the last decades it has contributed positively in several sectors of activity. Using ML's concepts and respective application, it is possible to increase productivity and reduce costs, detect fraud, improve predictive performance and thus improve the decision-making process. These are some of the potentialities of ML. In particular, predictive analysis has been seen as an important factor to ensure competitive advantage between companies.

Underlying theme choice is the curiosity to know more about what ML is, how it is changing our life and how it works. Alongside with this reason is the fact that Feedzai is a technology-based company that uses several ML techniques, which aroused my interest in writing about ML applied to my base training area - economy. Throughout this work the importance of this topic in the financial markets will be discussed.

The goal of this report is to describe tasks performed during the internship followed by respective analysis and also to study the application of ML techniques, namely classification trees and random forests, through an empirical case. This empirical case is used as object to test the predictive capacity in the financial market of bitcoin and PSI 20.

According to test results obtained, no ML technique was found undoubtedly better than the others because each technique is more likely to fit depending on the characteristics of the database, which means, the results indicate that the classification trees technique offers relatively more accurate results in the PSI market 20 while the random forests algorithm fits the most on the bitcoin market.

Additionally according to empirical case, the results tend to indicate that by using methodologies discussed in this report, the forecasting capacity was best performed when applied to PSI 20 market rather than in the bitcoin market.

Key Words

Machine learning; Classification Trees; Random Forests; Financial Market.

1. Introdução

1.1. Enquadramento

"Machine learning is a field of study that gives computers the ability to learn without being explicitly programmed."

Arthur Samuel, 1959

As organizações vivem num mundo cada vez mais globalizado e todos os dias enfrentam novos desafios. As decisões a tomar têm que ser cada vez mais rápidas de forma a garantir a sua sobrevivência num mercado mais competitivo.

Com o uso crescente das ferramentas tecnológicas, as organizações são diariamente responsáveis pela produção de avultados conjuntos de dados. Estes dados conferem informações importantes para a atividade das organizações.

O processamento destes dados em bruto necessita do auxílio de ferramentas da estatística e matemática. De acordo com Vasconcelos (2017) podemos considerar a *machine learning* (ML) como um ramo da estatística que atua sob diversos métodos e compreende dois principais objetivos: a capacidade de aprendizagem e o desempenho preditivo. Implementar num computador a capacidade de aprendizagem tem sido um dos maiores desafios. O processo de aprendizagem computacional constitui assim um importante objeto de estudo da ML (Michalski et al., 1983) assim como a capacidade de previsão da ML.

O desempenho preditivo é uma das tarefas mais reconhecidas da *machine learning*. Existem vários trabalhos na literatura sobre o desempenho preditivo da ML, nomeadamente nos mercados financeiros. A previsão é, nestes mercados, um desafio importante, pois existe muita imprevisibilidade envolvida (Patel et al., 2015). Algumas das técnicas utilizadas na ML para testar a capacidade preditiva nos mercados financeiros são as *Artificial Neural Networks* (ANN), *Support Vector Machine* (SVM) (Patel, et al., 2015), Árvores de Decisão (AD) (Nair et al., 2010) e *Random Forests* (RF) (Kumar et al., 2006).

No presente trabalho darei ênfase às duas últimas técnicas anteriormente citadas. As árvores de decisão são muito utilizadas na ML pois apresentam características como a simplicidade, interpretação fácil, a forte capacidade preditiva e o facto de se ajustarem a problemas de classificação e/ou regressão (Ferreira, 1999). As *random forests* possuem vantagens como a sua natureza aleatória, combinar uma grande quantidade de atributos e um alto nível de precisão (Bastos et al., 2013).

O relatório está organizado da seguinte forma: nesta primeira parte é dado um contexto geral sobre o tema escolhido, da importância da ML nos nossos dias e uma breve descrição da *machine learning* na Feedzai.

Na segunda parte do relatório vou descrever em detalhe as funções executadas na empresa e realizar a respetiva análise critica. Na análise crítica ao estágio é feita uma reflexão sobre as atividade desempenhadas ao longo do mesmo.

A terceira parte do relatório contém a revisão da literatura sobre ML - inicialmente são apresentados alguns conceitos introdutórios. O entendimento prévio destes torna-se crucial para compreensão das técnicas de ML exploradas no relatório (árvores de classificação e *random forests*). É importante compreender a aplicabilidade da *machine learning* no mercado financeiro e, por esta razão, ainda na terceira parte do trabalho, são apresentados vários estudos que observaram a capacidade preditiva nas series financeiras com recurso a diferentes técnicas de ML.

A quarta parte do trabalho descreve a recolha da base de dados, a sua preparação e a definição do modelo a seguir, explica o método adotado e os resultados obtidos, e apresenta a análise destes.

Por fim, na quinta parte, é apresentada uma síntese conclusiva articulando todos os conceitos anteriormente estudados.

1.2. A Importância da *Machine learning*

A *machine learning* tem conquistado o mundo. A razão desta crescente popularidade prende-se com o aumento da produção de dados informáticos e a necessidade de processar um grande volume de dados. Antes da entrada da ML no mundo empresarial as decisões eram tomadas muitas vezes com base nas regras do bom senso e com recurso a modelos matemáticos mais simples. Com a entrada da ML nos negócios, esta tornou-se uma importante ferramenta de apoio às decisões estratégicas (Hall et al., 2016).

A *machine learning* é um importante instrumento de análise na economia e gestão. Por exemplo, no comércio permite monitorizar o comportamento dos consumidores e formular ofertas personalizadas quase instantâneas; nas instituições bancárias proporcionar a comercialização de novos produtos financeiros, examinar o risco de crédito e deteção de fraude; no setor industrial detetar defeitos de fabrico; nos mercados financeiros investir de maneira mais eficiente possível. Os exemplos da aplicabilidade da ML podiam continuar.

À medida que as empresas se apercebem das vantagens da ML, esta adquire maior protagonismo. Estes benefícios podem-se tornar cruciais para o sucesso de uma organização pois a ML é capaz de alavancar informações estratégicas e detetar nichos de mercado. Desta forma as decisões dos empresários serão mais conscientes e informadas.

Não é só nas áreas de economia e gestão que a ML está presente. Exemplo disso é a presença da ML em áreas como a psicologia, neurociências ao estudar o comportamento humano; na biologia ao explorar novos conhecimentos relevantes sobre os vários processos biológicos; na informática com o reconhecimento e classificação de imagens e da fala; na medicina também pode ser uma ajuda preciosa ao recomendar quais os melhores tratamentos perante um determinado quadro clinico e, desta forma, melhorar os procedimentos médicos; neste momento estão a ser desenvolvidos automóveis de condução automática; sugerir aos condutores quais os melhores caminhos a transitar, evitando acidentes, possíveis congestionamentos, condições adversas no clima, etc.; reconhecer padrões relacionados com ocorrências de crimes.

Existem outras mais áreas onde a ML é aplicada, estes são apenas alguns exemplos mencionados. Posto isto é certo que a *machine learning* está a mudar o mundo e nossa maneira de viver.

1.3. A *Machine Learning* na Feedzai

Com o uso intensivo do computador e consequentemente o acesso à internet, as transações eletrónicas têm vindo a ganhar peso ano após ano. Com isto, a possibilidade de fraude *online* também aumenta. Empresas de *eCommerce* e instituições bancárias estão na mira dos *hackers*. O crescimento tanto em quantidade como em qualidade da fraude financeira é uma realidade e, portanto, proteger os utilizadores de ataques informáticos é uma preocupação cada vez maior. A missão da Feedzai passa por combater estes crimes e tornar o mundo informático mais seguro.

A Feedzai oferece um *software* capaz de analisar as transações eletrónicas em tempo real e perceber se estas apresentam sinais de fraude ou não. Cada cliente possui um padrão de normalidade e se houver comportamentos que fujam a esse padrão é dado o alerta. O objetivo é contextualizar cada transação e perceber se está de acordo com o que seria expectável ou não. Através da inteligência artificial e algoritmos de *machine learning* o programa vai aprendendo a identificar o tipo de transação (fraudulenta ou não) a fim de detetar transações de primeira instância, isto é, mesmo sem histórico de dados consegue analisar o tipo de transação com base na aprendizagem dos modelos anteriores.

O *software* disponibilizado pela Feedzai é baseado no programa Feedzai Pulse, que permite o processamento de dados em larga escala produzindo informações sobre o que está a acontecer em cada instante. Este programa constitui uma ferramenta de análise de risco de fraude.

O programa Feedzai Pulse visa fornecer aos decisores ferramentas úteis na sua tomada de decisão. A partir de indicadores de desempenho designados por *Key Performance Indicators* (KPIs) é possível monitorizar os acontecimentos em tempo real. Ao serem constantemente atualizados, é possível comparar esses KPIs com referências históricas e dissecar informações importantes. Esta plataforma aciona alarmes e processa

determinados mecanismos como o bloqueio da transações suspeitas de serem fraudulentas quando ocorrem eventos inesperados.

Além da monitorização das transações bancárias, outro exemplo onde o produto da Feedzai pode ser aplicado é nas empresas de distribuição elétrica. Aqui os operadores necessitam de saber o que está a acontecer a cada momento na sua infraestrutura, ou seja, se existe algum problema em alguma linha e onde. O programa Feedzai Pulse consegue fazer este tratamento de dados e detetar o que está a acontecer.

A figura 1 abaixo mostra um exemplo do painel de controlo da Feedzai Pulse com indicadores de desempenho.

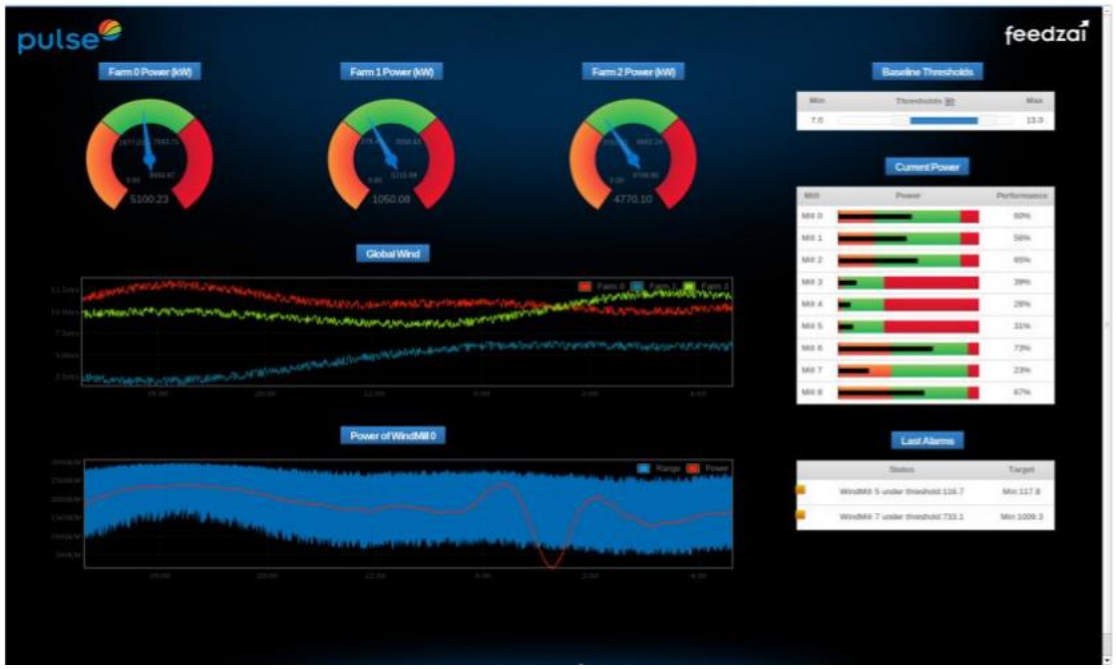

Figura 1 - Exemplo painel de controlo Feedzai Pulse.

O programa da Feedzai consegue dar assistência contínua aos negócios ao analisar na totalidade as transações e classificá-las como suspeitas ou não. Esta aptidão no mercado da deteção da fraude permite aos decisores das grandes empresas, como instituições financeiras, empresas em telecomunicações, empresas do setor energético, etc., organizar uma grande quantidade de dados, fornecer a alarmística e prever o que está a acontecer no negócio. Através deste produto desenvolvido pela Feedzai, os utilizadores conseguem diagnosticar as ocorrências e encontrar os motivos para tal.

2. O Estágio

2.1. Apresentação da entidade de acolhimento – Feedzai

Com sede em Coimbra, originária de um *spin-off* da Universidade de Coimbra, a Feedzai é uma empresa de base tecnológica especializada em deteção de fraude e utiliza técnicas de ML no processamento de dados.

Fundada em 2008 por Nuno Sebastião (CEO), Pedro Bizarro (CSO) e Paulo Marques (CTO), a *startup* mantém uma ligação umbilical à Universidade de Coimbra e à cidade de Coimbra, com um forte envolvimento em palestras e formações junto dos estudantes, oferecendo estágios em vários domínios aos alunos. Esta cooperação tem possibilitado uma inserção a nível local, contribuindo para o desenvolvimento regional. O seu principal centro de decisão é na cidade de Coimbra e pretende continuar no futuro pois é onde a equipa nuclear de desenvolvimento de produto se localiza. A empresa tem presentemente a seguinte estrutura:

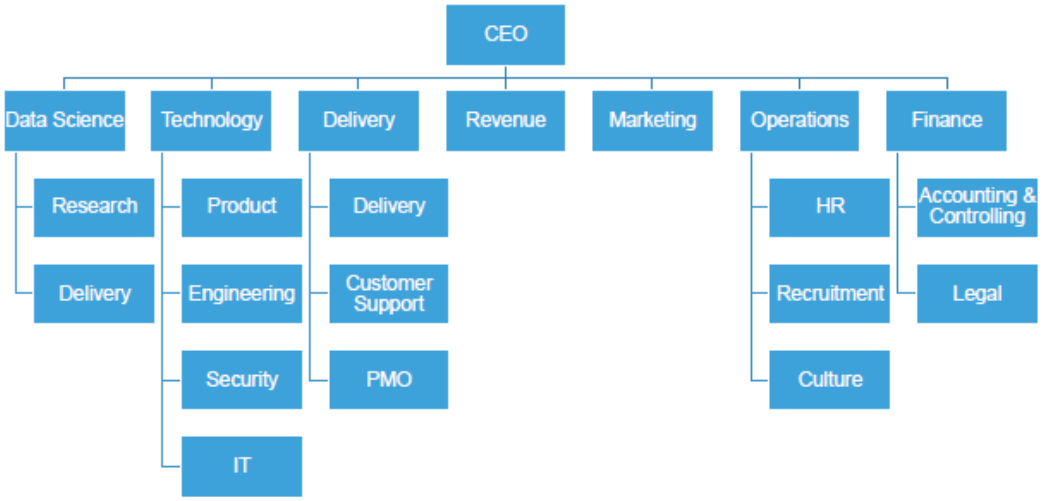

Figura 2 - Organograma Feedzai.

O processo de internacionalização acompanhou intrinsecamente a evolução da Feedzai, levando a um crescimento exponencial - em 2011 abria escritórios em Lisboa e contava com 12 funcionários ao serviço. Nesse ano, a Feedzai foi premiada pela Gartner pelo seu *software,* Feedzai Pulse, como uma das melhores ferramentas de análise em *Business Intelligence* a nível mundial. Assim "A Gartner destaca em particular a facilidade

de desenvolvimento e capacidade de expansão da Feedzai Pulse (…) a ferramenta estabelece facilmente referências de base de monitorização de indicadores de negócio, ao analisar o histórico de dados antigos".[1]

Decidida a ultrapassar barreiras geográficas, no ano 2013 a Feedzai instalou-se em Silicon Valley, localizado no Estado da Califórnia, nos Estados Unidos da América. Até esse momento, a *startup* esteve essencialmente focada na área "*Analytics and Business Intelligence*". A partir daí, foram anos de grande transformação devido à própria estratégia "*Born Gobal*[2]": a abertura dos escritórios nos anos de 2014 em Londres e em 2015 em Nova Iorque e no Porto, contabilizando 75 funcionários no total. No ano de 2016 foi a vez de abrir escritórios em Atlanta, no Estado Norte-Americano da Geórgia, com um total de 117 funcionários. A empresa Feedzai foi estruturada em duas empresas: a Feedzai S.A. a empresa mãe em Portugal e a Feedzai Inc. a subsidiária nos Estados Unidos da América.

No ano seguinte, a Feedzai captou uma ronda de investimentos de "*series C*[3]" no valor de 50 milhões dólares e o número de funcionários subiu em flecha para 237 funcionários em diferentes setores: *data science*, engenharia, departamento financeiro, recursos humanos e *marketing*. As novidades não param de chegar e já no início do ano de 2018 anunciou a sua estreia por terras asiáticas - a abertura dos novos escritórios em Hong Kong.

A Feedzai é reconhecida: está integrada na lista *Tech Tour Growth* 50 pelo terceiro anos consecutivo[4] - um ranking que engloba as 50 empresas europeias com crescimento mais promissor, é destacada na lista Forbes das fintech para 2018[5] e está perto de ser

[1] https://www.computerworld.com.pt/2011/05/19/gartner-distingue-ferramenta-da-feedzai/ . Consultado a 23 de julho de 2018.
[2] As *Born Gobal* são empresas que nos primeiros anos de vida procuram entrar no mercado externo (Oviatt, 2005).
[3] As rondas de investimentos em *startup* são classificadas por series de A a D+. Investimentos de Série C e D+ são para empresas com um volume de negócios significativo. Fonte: http://saldopositivo.cgd.pt/empresas/ronda-de-investimento-o-que-e/ . Consultado a 17 de junho de 2018.
[4] https://www.jornaldenegocios.pt/empresas/tecnologias/detalhe/feedzai-esta-na-lista-da-tech-tour-pelo-terceiro-ano. Consultado a 29 de março de 2018.
[5] https://www.jornaldenegocios.pt/empresas/pme/start-ups/detalhe/portuguesa-feedzai-entra-na-lista-da-forbes-das-fintech-para-2018. Consultado a 29 de março de 2018.

considerada numa *startup* "unicórnio"[6] – empresas com uma avaliação de valor igual ou superior a mil milhões de dólares.

A Feedzai conta com clientes de grande relevo pelos quatro cantos do mundo. Entre as instituições financeiras estão First Data, Goldman Sach, Lloyds Bank, etc. Numa vertente de eCommerce, marcas como a Nike e Starbucks estão na lista de utilizadores da tecnologia da Feedzai.

2.2. Objetivos do Estágio

No final do percurso académico o estágio funciona como elo de ligação entre o mundo académico e o mundo laboral. Este é um estímulo importante para um aluno em fase final do mestrado, prestes a entrar no mercado de trabalho.

O objetivo geral do estágio foi oferecer um primeiro contacto com a área financeira e uma experiência *hands-on* nos diversos campos de atuação do departamento financeiro da Feedzai, entre as quais:

- Controlo e faturação dos projetos da empresa;
- Controlo e acompanhamento do cumprimento orçamental;
- Participação ativa em projetos internos da empresa;
- Participação na preparação e gestão de candidaturas/projetos no âmbito do Portugal 2020;
- Execução de *reports* financeiros mensais;
- Apoio à função contabilística do grupo Feedzai;
- Cooperação com o CFO, *accounting* e *reporting manager* e com os *controllers* em tarefas de *compliance* fiscal.

[6] http://expresso.sapo.pt/economia/2016-03-17-Feedzai.-Vai-nascer-um-unicornio-em-Coimbra-#gs.XEo8a1k. Consultado a 29 de março de 2018.

2.3. Tarefas Desenvolvidas

Umas das primeiras tarefas consistiu na leitura de informação interna da Feedzai, nomeadamente as políticas de funcionamento da empresa, a história da empresa, o tipo de produto que a Feedzai desenvolve e como desenvolve. Perceber qual a conduta da empresa foi importante num primeiro contacto. De seguida irei detalhar as atividades que desenvolvi no estágio curricular.

2.3.1. Gestão de Projetos

As tarefas por mim realizadas incidiram essencialmente sobre a gestão de projetos. No entanto obtive uma visão geral de todas as tarefas executadas no departamento financeiro, o que foi bastante gratificante para a minha formação enquanto profissional.

É importante perceber previamente o que se entende por projetos na Feedzai. Estes projetos que vão ser descritos de seguida são considerados os principais projetos da Feedzai. Cada projeto compreende a relação entre um cliente e a empresa, descrevendo todas as interações numa ótica financeira. Tendo em conta que os recursos (humanos, financeiros, etc.) são limitados, um projeto estabelece os objetivos a cumprir dentro de um determinado prazo com vista à otimização dos resultados. Os contactos frequentes com os gestores de projetos – estes são do departamento de engenharia – são cruciais para o acompanhamento da evolução dos projetos em termos técnicos.

O montante monetário estabelecido entre o cliente e a empresa serve para o consumo do produto que a Feedzai oferece. Este produto é fornecido sob forma de licença e/ou sob forma de serviços. Com base no contrato acordado, há clientes que possuem só licença e outros que possuem licença e serviços.

Cada cliente tem as suas particularidades, desde a forma como as licenças e/ou serviços são faturados, o tipo de trabalhadores, por qualificação, alocados ao projeto, o número de horas trabalhadas, a necessidade de deslocação dos trabalhadores, a calendarização das tarefas a serem cumpridas no projeto por diversas fases, etc. De acordo com os diferentes objetivos e necessidades, cada cliente detém o seu próprio projeto.

No decorrer no meu estágio participei ativamente em nove projetos e, com o decorrer destes, supervisionei-os nas várias vertentes que assim exigem:

- **Controlo do orçamento:** Requer a atenção constante sobre os fluxos financeiros que vão ocorrendo no projeto. Tendo em conta o montante monetário acordado, a função de controlo do orçamento passa por notificar todos os que estão envolvidos no projeto - desde a equipa técnica de engenharia, o departamento comercial e financeiro até ao contacto direto com o cliente - do ponto da situação orçamental do projeto e avisar prontamente quando este se está quase a esgotar. A comunicação clara entre os diferentes departamentos é crucial.

- **Controlo das despesas com viagens e/ou alojamento e pedido de aprovação:** Tendo em conta a internacionalização da Feedzai, é muito comum a necessidade de viagens. Quando há necessidade de trabalho *onsite* alocado aos projetos de clientes, é necessário certificar a situação orçamental destinada a este tipo de despesas (condições definidas no contrato) e, ainda, garantir previamente a aprovação junto do cliente destas deslocações. Só depois da confirmação do cliente, as viagens podem ir avante. Esta função exige uma eficiente articulação na comunicação entre o cliente, o trabalhador que vai viajar, e quem está encarregue de marcar a viagem.

- **Executar previsões:** Esta é uma tarefa importante pois possibilita antever as necessidades futuras do projeto e comunicá-las o mais breve possível. Ter presente o estado atual do projeto é essencial para levar a cabo uma boa gestão.

- **Reuniões com os gestores de projetos**: Estes encontros têm essencialmente o objetivo de transmitir qual o ponto da situação numa perspetiva financeira e obter junto dos gestores de projetos as informações técnicas necessárias. Estas reuniões são uma mais-valia para a execução das previsões e para conhecer melhor alguns aspetos técnicos do projeto. As reuniões eram mensais; no entanto, caso houvesse necessidade, eram agendadas mais reuniões.

- **Coletar todas as faturas:** Todas as despesas incorridas nos projetos necessitam, para efeito de prova, de um documento onde se descrevem detalhadamente os custos, designado como fatura. Esta tarefa consiste no devido armazenamento oportuno de forma a evitar atrasos no momento da faturação e por uma questão de organização diária.

- **Faturação aos clientes:** O envio da documentação necessária para poder faturar o cliente da licença e/ou serviços prestados pela Feedzai é um passo importante. A emissão destas faturas é executada com recurso à plataforma Netsuite. No momento do envio, geralmente via *e-mail*, tinha que garantir que os destinatários selecionados estavam todos presentes e anexar corretamente os documentos. Também são enviadas as faturas com despesas de deslocações e/ou alojamento posteriormente recolhidas.

- **Controlo das contas a receber por parte dos clientes:** Uma vez emitidas e enviadas as respetivas faturas via *e-mail*, supervisionei as contas a receber e registei todos estes acontecimentos à medida que decorriam. Quando havia alguma anomalia nos pagamentos era minha tarefa também entrar diretamente em contacto com o cliente.

  Por vezes era-me solicitado por parte dos clientes alguns pedidos de esclarecimento numa ótica financeira do projeto.

- **Utilização da plataforma Amazon Web Services[7]*:*** Alguns dos projetos que a Feedzai possui, utilizam a plataforma da Amazon Web Services e, portanto, também tive oportunidade de explorar esta ferramenta. A utilização desta plataforma também envolve o processo de faturação.

[7] "A Amazon Web Services (AWS) é uma plataforma de serviços em nuvem segura, oferecendo poder computacional, armazenamento de banco de dados, distribuição de conteúdo e outras funcionalidades para ajudar as empresas no dimensionamento e crescimento." Fonte: https://aws.amazon.com/pt/what-is-aws/ . Consultado a 17 de junho de 2018.

2.3.2. Portugal 2020

A Feedzai é candidata a vários projetos de incentivo ao investimento por parte do Governo de Portugal, através de um programa designado por Portugal 2020 - "trata-se do acordo de parceria adotado entre Portugal e a Comissão Europeia (…) consagram a política de desenvolvimento económico, social e territorial para promover, em Portugal, entre 2014 e 2020."[8]. As candidaturas são aprovadas pelo Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico em Copromoção[9].

Nos vários projetos pertencentes ao Portugal 2020 que a Feedzai tem a decorrer, colaborei em três projetos, nomeadamente ExpansionZai, InsightZai e MomentumZai.

Apoiei a elaboração e gestão de candidaturas/projetos do Portugal 2020, submeti a formalização de pedidos de pagamento no Balcão 2020, agreguei a documentação necessária (extratos do banco dos vencimentos dos recursos humanos; faturas das viagens e eventos realizados, etc.) para a preparação dos *dossiers* financeiros com as despesas elegíveis de cada projeto, assegurando a sua constante atualização.

2.3.3. *Report* Financeiro

Numa dimensão financeira, realizei o *reporting* de pagamentos e *reporting* de recebimentos referente a toda a empresa, isto é, da Feedzai S.A. e da Feedzai Inc. Esta atividade refere-se à preparação da informação dos *reports* financeiros mensais - classificar e descrever corretamente todos os movimentos e, no final, agrupar todos os *cash-flows* conciliando todas as rubricas existentes num só documento.

2.3.4. *Procurement*

A Feedzai participa em muito eventos, feiras de divulgação, conferências e formações, etc. Estes acontecimentos envolvem quase sempre a deslocação dos

[8] https://www.portugal2020.pt/Portal2020/o-que-e-o-portugal2020. Consultado a 28 de Março de 2018.
[9] http://ani.pt/incentivos/idt-em-co-promocao/ Consultado a 28 de Março de 2018.

participantes. A tarefa de *procurement* é a procura por serviços externos em nome da empresa, como, maioritariamente, viagens e alojamento.

É comum a Feedzai contratar trabalhadores estrangeiros que vêm para Portugal trabalhar. Nestes casos também é necessário garantir alojamento local. Neste sentido, ao longo do meu estágio realizei e acompanhei de perto algumas funções de *procurement.*

À medida que as viagens iam sendo reservadas era necessário o registo num documento em *Excel* que descreve detalhadamente informações como: o nome do passageiro, a cidade de origem/destino, data, nome do hotel, número da fatura relacionada com despesa, caso seja uma despesa alocada a um projeto, indicar qual o cliente, a respetiva referência da ordem de compra, etc. Em parte do meu estágio fiquei responsável pelo registo e constante atualização deste documento.

2.3.5. *Purchase Order*

Todas as faturas pagas pela Feedzai, têm que ser registadas na contabilidade. Cada fatura, antes de ser lançada contabilisticamente, precisa estar associada a uma ordem de compra, usualmente designada por *purchase order* (PO). As PO são emitidas a partir do programa Netsuite e consistem numa breve descrição da fatura e sua classificação por centro de custo - se a despesa diz respeito ao departamento de engenharia, *marketing*, serviços externos, despesas gerais, o tipo de IVA taxado, etc.

Sempre que havia esta necessidade, foi uma das tarefas que assumi ao longo do estágio. Para associar as faturas ao correto centro de custo, este exercício exige, de maneira geral, o conhecimento prévio do tipo de despesas efetuadas em cada departamento, os habituais fornecedores, o tipo de serviço prestado, etc. À medida que fui ganhando mais experiência, esta tarefa foi-se tornando mais intuitiva.

2.3.6. Outras Atividades

A organização das tabelas salariais também foi uma das tarefas executadas no estágio. Esta tarefa consiste no tratamento estatístico de dados dos colaboradores como a caracterização profissional de todos os empregados, a respetiva localização do escritório a

que estão alocados, etc. O *Microsoft Excel* é o programa mais utilizado na concretização destas tarefas.

Ao longo do estágio a Feedzai proporcionou à equipa do departamento financeiro uma formação durante uma semana sobre o programa Netsuite. Esta formação foi útil para domínio desta ferramenta utilizada no dia-a-dia da Feedzai.

Outra atividade praticada foi a seguinte: depois da equipa de contabilidade apurar quais as despesas por pagar, é partilhado um documento interno onde se descreve qual é a despesa e o respetivo montante. Por uma questão de confirmação, antes do pagamento ser efetivamente concluído, é necessária a validação por um dos elementos do departamento financeiro. Esta validação compreende confirmar se a descrição e os valores batem certo com as faturas. Nesta tarefa, de maneira a não haver erros, é necessária a análise individual de cada despesa e atenção aos pormenores.

2.4. Integração no Estágio

Numa primeira abordagem, fui integrada com recetividade por toda a equipa e prontamente mostraram-se disponíveis para esclarecer todas as minhas questões. Rapidamente me contextualizaram na empresa e explicaram a conduta relativamente ao departamento financeiro.

Semanalmente a equipa do departamento financeiro da Feedzai S.A. e Feedzai Inc. convoca uma reunião com a finalidade de alinharmos, todos juntos, os objetivos individuais e/ou coletivos, e manter-nos a par dos últimos acontecimentos. Esta constitui numa discussão saudavelmente aberta em que todos podemos intervir. A participação nas reuniões foi essencial para perceber como tudo funciona internamente e o qual o papel de cada um de nós lá dentro.

A Feedzai S.A. e Feedzai Inc. une também, oportunamente, todos os departamentos, harmonizando momentos de convívio num ambiente de aprendizagem mútua. Estas circunstâncias foram cruciais para a minha integração, tanto a nível profissional como pessoal pois permitiu compreender todas as outras funções fora da minha formação académica e obter conhecimentos noutras áreas.

Na primeira semana, tive uma sessão de acolhimento dirigida pelos recursos humanos. Essa sessão foi repartida em 2 dias perfazendo um total de 4 horas. Proporcionaram-nos um primeiro enquadramento sobre a empresa, a sua missão e valores, as políticas de trabalho, os benefícios adicionais a que temos direito, forneceram-nos contactos úteis e ilustraram como a empresa esta dividida, esquematizando tudo de maneira a facilitar o nosso entendimento sobre "o que fazem", "quem são" e "como fazem".

2.5. Análise Crítica do Estágio

Como forma de terminar este capítulo, a análise crítica é fundamental na medida em que permite uma reflexão sobre a forma como as atividades por mim realizadas se foram desenvolvendo ao longo do estágio, analisar a minha evolução nestes meses, o cumprimento dos objetivos iniciais e da eventual necessidade de sugerir algumas alterações que, do meu ponto de vista, me pareçam válidas.

Participar nos projetos internos da empresa foi uma atividade desafiante. Durante esta tarefa tive sempre supervisão superior porém concederam-me autonomia e atribuíram-me responsabilidades. Dadas as caraterísticas desta atividade, permitiu-me compreender a dinâmica da gestão de projetos internos da Feedzai, conhecer de perto cada cliente, desenvolver capacidade analítica e sentido estratégico, e desenvolver competências de comunicação por ser o elo de ligação entre o cliente e a Feedzai em termos financeiros do projeto.

Os projetos Portugal 2020 permitiram-me conhecer os procedimentos teóricos nas várias etapas do projeto, entender as respetivas diferenças entre os projetos e desenvolver competências administrativas.

Com o exercício de tarefas como o *reporting* financeiro e realização de PO comecei a entender a dinâmica de funcionamento no âmbito financeiro e contabilístico. A capacidade de identificar e interpretar também foram bastante estimuladas.

Para além de consolidar os meus conhecimentos académicos através de um primeiro contacto com o mundo laboral na minha área de formação, o estágio alargou os

meus conhecimentos no domínio da língua inglesa devido ao contínuo contacto com esta língua.

Com o rápido crescimento da empresa, sente-se alguma fragmentação geográfica nos vários departamentos da empresa. Ainda que o uso da Internet minimize imenso esta questão, é necessário um esforço para integrar os vários departamentos o máximo possível. Esta integração é exequível através de uma boa comunicação interna. A Feedzai enfrenta este desafio pois nos últimos anos sofreu uma grande expansão. A minha sugestão vai neste sentido, trabalhar cada vez mais para uma comunicação interna mais clara.

Num balanço geral, a minha passagem pela Feedzai possibilitou um crescimento pessoal e profissional constituindo uma experiência indubitavelmente enriquecedora ao atribuírem-me responsabilidade em várias tarefas estimulando o meu sentido de análise crítica e autonomia.

3. Revisão da Literatura

3.1. O que é *Machine learning*

De acordo com a definição no *site* da SAS[10], *machine learning* pode ser definida como "um método de análise de dados que automatiza a construção de modelos analíticos. É um ramo da inteligência artificial baseado na ideia que os sistemas podem aprender com dados, identificar padrões e tomar decisões com intervenção humana mínima." Quanto mais dados forem inseridos, mais experiência o computador obtém, o que torna o seu desempenho melhor (Cielen et al., 2016).

Machine learning compreende diversos métodos que partilham os mesmos objetivos: a capacidade preditiva do modelo e automatizar o processo de treino com a base de dados inserida. A ML vem maximizar o desempenho preditivo dos modelos através de ferramentas estatísticas (Alpaydin, 2004). A ML trabalha com modelos direcionados para a previsão *out of sample*, isto é, embora os modelos sejam construídos com uma determinada amostra de dados pretende-se que tenha um bom desempenho quando efetuam previsões em novos dados (De Souza, 2016).

A *machine learning* incorpora várias etapas até construir um modelo suficientemente robusto. A construção deste modelo compreende quatro fases (Cielen et al., 2016): seleção dos recursos – este é um passo extremamente importante pois o modelo selecionado vai combinar os recursos para efetuar previsões; treinar o modelo – tendo os preditores corretos o modelo pode começar a ser treinado; validação – esta etapa averigua se o modelo realmente funciona; e por fim aplicação do modelo treinado a novos dados – se todas a etapas anteriores forem bem aplicadas então podemos adotar este modelo a um novo conjunto de dados e executar previsões.

[10] https://www.sas.com/pt_pt/insights/analytics/machine-learning.html. Consultado a 18 de junho de 2018

3.2. Tipos de Aprendizagem

De um modo geral, podemos dividir as diferentes abordagens da *machine learning* através da quantidade de esforço humano necessário para serem implementadas. A classificação destas abordagens difere um pouco consoante os autores. Segundo Cielen et al. (2016) dividem-se em:

Aprendizagem supervisionada – Esta abordagem tenta distinguir resultados e aprender ao extrair padrões dos dados rotulados[11] - de maneira a perceber melhor o que são dados rotulados é dado um exemplo[12] bastante popular: imaginemos que pretendemos ensinar o computador a identificar o que é um gato e o que é um cão. A partir de imagens destes, é colocado um rótulo de "gato" em imagens com gatos e um rótulo de "cão" em imagens com cães. A ideia é que o treino do modelo use dados que identificam corretamente o resultado pretendido. De seguida podemos começar a usar esses dados nos algoritmos de aprendizagem, para estes aprenderem a classificar imagens corretamente.

A interação humana é necessária para rotular os dados. O algoritmo recebe um conjunto de entradas com as saídas corretas correspondentes e aprende ao comparar a saída real com as saídas corretas e encontrar os erros.

Aprendizagem não-supervisionada – Esta abordagem não depende de dados rotulados e tenta encontrar padrões num conjunto de dados sem interação humana. O algoritmo pretende descobrir o que está a visualizar. O objetivo é explorar os dados e encontrar uma estrutura no seu interior. Esta abordagem é usada com dados transacionais; por exemplo, podem ser usados para identificar segmentos de clientes com atributos similares em campanhas de marketing e recomendar itens.

Aprendizagem semi-supervisionada – Nesta abordagem são utilizados tanto dados rotulados como não rotulados. Na maior parte dos casos apenas uma pequena parte dos

[11] De acordo com o *site* SAS, rótulo é o alvo em *machine learning*. https://www.sas.com/pt_br/insights/analytics/machine-learning.html. Consultado 11 de Maio de 2018.
[12] Exemplo retirado de: http://nkonst.com/machine-learning-explained-simple-words/ . Consultado a 18 de junho de 2018.

dados possuem rótulos e, por esta razão, esta forma de aprendizagem torna-se bastante frequente. Além disso o custo de rotular dados é geralmente elevado (Sanches, 2003).

No *site* SAS, fazem a distinção de mais uma abordagem de aprendizagem:

Aprendizagem de esforço – O algoritmo com base em testes "tentativa erro" deteta quais as ações que conduzem a um melhor resultado. Desta forma, esta abordagem atinge o objetivo mais rápido ao descobrir a qual a melhor política a seguir. Este método é frequentemente utilizado na robótica, jogos e navegação. Por exemplo[13]: num computador que esteja a aprender jogar xadrez, o algoritmo apenas recebe a informação final se ganhou ou perdeu o jogo pois nem todos os movimentos do jogo são rotulados como bem-sucedidos ou não. O algoritmo de ML, quanto mais jogar, vai atribuir maior peso aos movimentos que antes resultaram numa combinação vencedora.

3.3. Conceitos da Aprendizagem Supervisionada

No tópico anterior definimos os diferentes tipos de aprendizagem. Agora importa fornecer alguns conceitos intrínsecos à ML. Monard et al. (2003) descreve alguns conceitos relativos a aprendizagem supervisionada. Alguns destes conceitos estão definidos brevemente na tabela 1[14] abaixo.

Tabela 1 - Definição de alguns conceitos da ML.

Conceito	Definição
Atributo	Descreve certas características dos exemplos. A escolha dos atributos é crucial para um bom desempenho preditivo do problema a considerar. Por exemplo, para determinar uma gripe a escolha de atributos como a cor de cabelo, dos olhos, a altura, etc., em vez de atributos com os sintomas da doença tal como a temperatura corporal, o mal-estar geral, etc., vai comprometer os resultados da previsão. Os atributos podem ser de dois tipos:

[13] Exemplo retirado de: http://nkonst.com/machine-learning-explained-simple-words/ . Consultado a 18 de junho de 2018.
[14] Esta tabela baseia-se parcialmente em Monard et al. (2003).

	Atributo Nominal Assume valores num conjunto finito e pré-definido de possibilidades, sendo estas possibilidades designadas também por categorias. Muitas vezes, este tipo de atributo não segue uma sequência ordinal. Ex.: sol, nublado, chuva, etc.
	Atributo Contínuo ou Numérico Mede um o valor real ou inteiro (Witten et. al., 2016). Ex.: altura de uma pessoa, o peso, etc.
Algoritmo de indução	Este conceito funciona com base na dedução lógica para obter conclusões; caracteriza-se por ser um dos principais métodos utilizados para gerar conhecimento novo e prever eventos futuros. Consoante os exemplos disponíveis – tendo em conta que cada exemplo possui atributos próprios e está normalmente associado a um rótulo da classe – vai tentar determinar corretamente a classe dos novos exemplos que ainda não possuam um rótulo.
	Indutor na Aprendizagem supervisionada É fornecido um conjunto de exemplos já rotulados de forma a treinar o modelo. Os exemplos subdividem-se em problemas de classificação quando os rótulos são discretos ou de regressão para valores contínuos.
	Indutor na Aprendizagem Não Supervisionada Com base nos exemplos anteriormente fornecidos, vai tentar encaixá-los em *clusters* (Hanson et al., 1991). Após a determinação dos *clusters*, requer uma análise para averiguar o seu significado no contexto do problema.
Classe	Como já mencionado anteriormente os atributos e rótulos pertencem a conjuntos nominais de classes no caso de classificação.
Classificador	A partir de um conjunto de exemplos, o algoritmo de indução vai criar como saída um classificador, também pode ser denominado como hipótese. O objetivo é prever corretamente a classe dos novos exemplos inseridos.
Ruído	Existe ruído nos dados quando trabalhamos com dados imperfeitos. Estas imperfeições podem ser oriundas da aquisição dos dados, de classes mal rotuladas, etc.
Overfitting* / *Underfitting	***Overfitting*** No processo de indução do problema podem ser criados classificadores demasiado específicos. Neste caso, os classificadores ajustam-se em excesso no conjunto de dados de treino, podendo por isso ter um mau desempenho preditivo quando aplicados a novos dados.
	Underfitting Designa-se como sendo o caso contrário do *overfitting,* ou seja, ao induzir com base nos exemplos de treino inseridos, existe uma carência de ajustamento no classificador.

Dados *In of sample*	Dados usados no treino inicial do modelo. [15]
Dados *Out of sample*	Dados usados para testar o desempenho preditivo do modelo. [16]

3.4. Métodos de *Machine learning*

Os métodos da *machine learning* podem ser identificados como Modelos Lineares e Modelos Não Lineares. Apesar de neste relatório se estudarem os modelos não lineares segue-se uma breve descrição de cada.

3.4.1. Modelo Linear

O modelo linear, tal como o próprio nome indica é linear nos seus preditores preservando a relação da causalidade direta e simplista do problema a observar. Uma das técnicas mais aplicadas nesta abordagem é a dos MQO (Método dos Quadrados Ordinários) Este é um método simples com resultados satisfatórios em muitos casos (Vasconcelos, 2017). Este modelo tem, tipicamente, uma interpretação simples (James et al., 2013).

Como vimos anteriormente, o principal objetivo da *machine learning* é a capacidade para prever o que vai acontecer a partir de modelos e a automatização do processo de aprendizagem. A utilização de modelos lineares mostra-se, geralmente, insuficiente para responder satisfatoriamente a situações mais complexas, existindo a necessidade de buscar técnicas mais elaboradas.

[15] Definição retirada do site: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:In-sample_vs._out-of-sample_forecasts. Consultado 11 de Junho de 2018.
[16] Definição retirada do site: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:In-sample_vs._out-of-sample_forecasts. Consultado 11 de Junho de 2018.

3.4.2. Modelo Não Linear

O modelo não linear possui estruturas não lineares relativamente aos parâmetros a estimar. O pressuposto de não linearidade nos parâmetros do modelo traz uma complexidade adicional na interpretação dos valores estimados. Alguns dos métodos não lineares supervisionados utilizados em *machine learning* são: regressões polinomiais, métodos de *kernel*, *smoothing*, modelos aditivos generalizados (GAMs), árvores de regressão e classificação, redes neuronais, etc. (Vasconcelos, 2017). No presente relatório vou dar destaque às árvores de regressão e classificação e às *random forests.*

3.4.2.1. Árvores de Decisão

As árvores de decisão (AD) são uma ferramenta de ML aplicável na resolução de problemas de regressão e classificação, usando estatística não-paramétrica - dados provenientes de uma população que não segue necessariamente uma distribuição normal, assumindo poucas hipóteses sobre a distribuição de probabilidade da população. Na parte empírica deste relatório é aplicada a técnica das árvores de classificação e, assim, opta-se por descrever os procedimentos relativos às árvores de classificação.

As árvores de decisão seguem uma abordagem de "dividir-para-conquistar" perante o problema de decisão. A lógica desta abordagem é tornar um problema complexo em sub-problemas mais simples. Este modo de representar o conhecimento é constituído por quatro elementos (Gama, 2002): a "raiz", os "nós", os "ramos" e as "folhas". A "raiz" é o ponto de partida da árvore de decisão. A partir daqui é feita uma divisão, esta divisão do espaço é definida pelos atributos em sub-espaços denominados como os "nós" da árvore.

Seguindo a estratégia inicial "dividir-para-conquistar" vão descender dos nós para os "ramos". Estes ramos, no final da árvore de decisão, conduzem às "folhas". Cada folha está associada uma classe. O trajeto desde a "raiz" até à "folha" corresponde a uma regra de classificação. Segundo Costa (2011) a construção de uma árvore de classificação envolve os seguintes aspetos:

1. O número de subconjuntos gerados a partir de uma questão;
2. Definir qual a característica a ser testada em cada nó;

3. Em que momento é que deve considerar um nó final, isto é, uma folha;
4. Associar a folha a uma classe.

Em geral, AD representam uma separação de combinações dos atributos. Cada caminho da raiz da árvore até à folha corresponde a uma combinação de testes de atributos e a própria árvore a uma separação dessas conjunções (Mitchell, 1997). A figura 3 ilustra um exemplo simples de uma árvore de classificação.

Baseado do exemplo de Mitchell (1997), suponhamos que se pretende jogar ténis mas antes é necessário averiguar se as condições meteorológicas vão estar favoráveis. Para tal são tidos em conta certos atributos sobre o estado do tempo (solarengo, nublado ou chuvoso), a humidade e o vento. A resposta final será, sim no caso de haver condições para jogar, ou não no caso contrário. Uma das possíveis trajetórias seria: tempo solarengo, humidade normal, então sim, existem condições favoráveis para jogar ténis.

Ao aplicar a estratégia "dividir-para conquistar" a árvore vai sofrendo contínuas divisões conforme as respostas às sucessivas questões em cada nó até chegar à folha. É na folha que se encontra a resposta final ao problema de decisão. É com base nos dados anteriormente treinados que a árvore se baseia para prever a classe (sim ou não) dos novos dados.

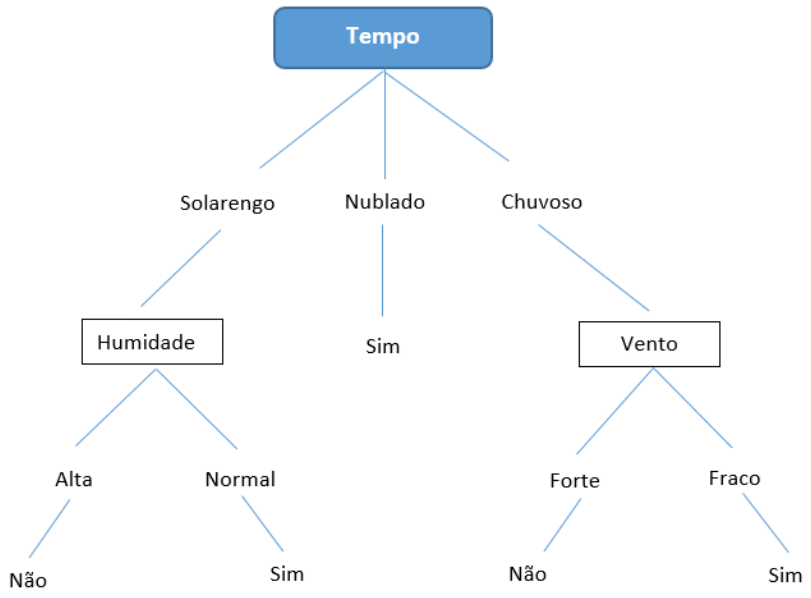

Figura 3 - Exemplo Árvore de Decisão para classificação (Mitchell, 1997).

À medida que os problemas se tornam mais complexos, vão crescendo mais ramos nas árvores de decisão. No entanto a base de raciocínio não sofre alterações, ou seja, os procedimentos na resolução do problema em si, por mais complicado que pareça, são os mesmos.

A utilização das árvores de decisão possui naturalmente algumas vantagens e desvantagens. Segundo Costa (2011) a utilização das AD apresentam benefícios como:

- Fácil interpretação - Decisões mais complexas são transformadas numa série de decisões mais simples e locais.
- Flexibilidade na construção da AD - As árvores de decisão não assumem nenhuma distribuição para os dados pois, como já foi referido anteriormente, utilizam estatística não-paramétrica.
- Robustez – Bom desempenho mesmo se as hipóteses iniciais forem alteradas.

Apesar das vantagens previamente mencionadas, as AD apresentam alguns problemas, como (Costa, 2011):

- Instabilidade – Pequenas perturbações no conjunto de treino podem provocar grandes alterações na árvore final;
- Risco de replicação – Duplicação de uma sequência de testes em diferentes ramos de uma árvore de decisão, levando a uma representação não concisa.
- *Overfitting (*Mitchell, 1997) - É uma das principais questões que se coloca relativamente à inclusão de novos ramos nas árvores de decisão, pois é necessário controlar o seu crescimento. Determinar a melhor dimensão para AD pode ser complicado quando há ruído na amostra ou quando o número de exemplos na amostra de treino é demasiado pequena para generalizar.

Um dos procedimentos para ultrapassar esta questão é o *pruning* ou poda. Existem duas fases neste processo: Pré-poda - Processo efetuado durante a construção da árvore de decisão. A segmentação dos nós termina e transforma o nó corrente numa folha; Pós-poda - Após a árvore estar construída, este processo vai remover ramos inteiros da árvore a partir de um determinado nó. Este nó, por ventura, vai ser transformado numa folha.

3.4.2.2. *Random Forests*

Random forests (RF) são um método de *machine learning* criado por Breiman (2001). As RF definem-se como uma combinação de árvores de decisão de tal forma que cada árvore é construída usando um subconjunto dos atributos selecionados aleatoriamente a partir do conjunto de dados originais (Lorenzett, 2016).

De acordo com o *site*[17] Statsoft, RF consistem num conjunto de várias árvores de decisão para determinar o resultado final. Em cada nó formado na árvore é escolhido um subconjunto aleatoriamente, tanto das variáveis independentes (atributos) como das observações contidas na amostra de treino. Este subconjunto selecionado é usado para definir o teste que conduzirá à ramificação da árvore.

A aplicação do algoritmo *random forests* compreende os seguintes conceitos (Costa, 2012):

1. A amostra escolhida para o conjunto de treino é feita de forma aleatória;
2. A escolha dos subconjuntos de variáveis em cada nó também é feita de forma aleatória.

Na figura 4 é possível visualizar a lógica subjacente às RF. Supondo que se quer identificar imagens de cães ou de gatos. São inseridas as bases dados, neste caso as imagens e a partir daí são criadas várias árvores de decisão. Através do algoritmo das *random forests*, cada nó é dividido de acordo com o melhor subconjunto de indicadores escolhidos de forma aleatória (Liaw et al., 2002).

Para problemas de classificação, primeiro, ocorre a seleção aleatória de atributos depois a classe mais popular é selecionada – isto é, podendo diferentes árvores dar indicações diferentes quanto à classe a escolher, seleciona-se a classe indicada pelo maior número de árvores. No final deste processo é encontrado um padrão predominante, Y, que leva à resposta ao problema inicial, identificar se é um cão ou se é um gato.

[17] http://www.statsoft.com/Textbook/Random-Forest. Consultado a 19 de junho de 2018.

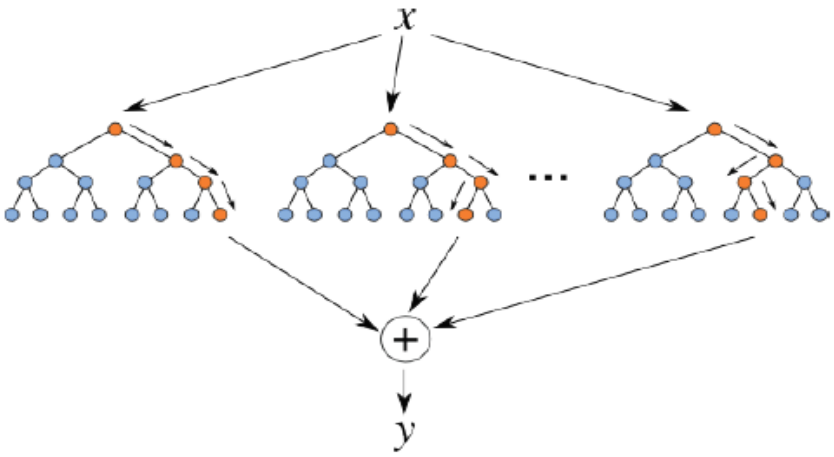

Figura 4 - Ilustração do algoritmo r*andom forests.*

Atendendo às características que as *random forests* possuem, as vantagens da sua utilização são (Lorenzett et al., 2016):

- Constituem um algoritmo mais poderoso comparativamente à utilização de apenas uma árvore de decisão. Com efeito, o uso de várias árvores de decisão pode maximizar o desempenho preditivo (Breiman, 2001). Esta combinação de classificadores é denominada por *ensemble learning* (Freund et al., 1996). O objetivo do *ensemble learning* é combinar múltiplos classificadores treinados individualmente, retirar a contribuição de cada um e no final obter melhores resultados de classificação (Marmanis et al., 2009). Portanto uma *random forest* atua como um algoritmo do tipo *ensemble learning* ao combinar várias árvores de decisão, utilizando um subconjunto de atributos selecionados aleatoriamente;
- Evitam o problema *overfitting* – as *random forests* conseguem generalizar melhor o resultado através do método de escolha aleatória dos atributos;
- São menos sensíveis a dados com ruído;
- Bom desempenho preditivo;
- Permitem processar um grande volume de dados de forma aleatória sem intervenção humana.

Apesar de constituírem um algoritmo com bastantes benefícios, as *random forests* também possuem desvantagens (Costa, 2011):

- Difícil interpretação;
- Podem enviesar os resultados quando existe a presença de um elevado número de atributos distintos.

3.5. *Machine Learning* no Mercado Financeiro

As técnicas de *machine learning* possuem uma vasta aplicabilidade a vários sectores de atividade. Neste trabalho aparece em destaque a aplicação das técnicas de ML ao mercado financeiro. O principal objetivo é a capacidade de previsão em séries financeiras.

Os mercados financeiros possuem alguma complexidade e dinamismo pois existem muitas variáveis capazes de influenciar o valor no mercado financeiro. Variáveis como as condições socioeconómicas, as expectativas dos investidores, eventos políticos, etc., fazem com que os mercados financeiros sejam a suscetíveis a rápidas mudanças, causando flutuações aleatórias no preço dos produtos financeiros (Khaidem et al., 2016).

A ML aplicada ao mercado financeiro tem como principal função a previsão da evolução das séries financeiras. Este é um objetivo muito ambicionado na literatura existente. A fórmula perfeita para a previsão do comportamento das séries financeiras não existe, no entanto a ML entra como uma preciosa ajuda nesta tarefa.

Entre as técnicas de *machine learning* exploradas na literatura estão as *Artificial Neural Networks* (ANN), *Support Vector Machines* (SVM), *Random Forests* (RF) e *Naive-Bayes* (NB). Os autores dos estudos apresentados usam os valores ou os sinais dos indicadores técnicos para prever a direção e/ou a variação dos preços num determinado mercado financeiro.

Patel et al. (2015) estudaram o desempenho preditivo na bolsa Indiana usando todas as técnicas supramencionadas, aplicadas com base em sinais obtidos a partir de indicadores técnicos. Por ordem decrescente, as técnicas que mostraram melhor capacidade de previsão foram as NB, as RF, de seguida as SVM e por último ANN. Apesar desta ordem de preferência todas as técnicas tiveram resultados são bastante satisfatórios. Em particular, as três primeiras técnicas obtêm resultados muito semelhantes.

As SVM e as RF têm vindo a ganhar destaque na literatura. Kumar et al. (2006) observou no seu estudo um bom desempenho preditivo com estas técnicas, especialmente com as SVM. No entanto, cada técnica possui os seus pontos fortes e fracos. Kumar et al. (2006) apontou para no futuro combinar modelos de SVM com outros de modelos classificação, assim a fraqueza de um método pode ser colmatada pelos pontos fortes de outro.

Outros estudos indicam a técnica das ANN como uma das mais populares neste tipo de aplicação. No entanto, a sua aplicação a séries financeiras possui algumas limitações pois os dados provenientes destas séries podem conter ruído. Kim (2003) aplicou as técnicas ANN e SVM no mercado financeiro da Coreia fazendo as previsões com base em indicadores de análise técnica. Os resultados indicam que a técnica SVM conduz a melhores resultados. O autor concluiu ainda que a técnica SVM é uma alternativa promissora na previsão de séries temporais financeiras.

A técnica ANN tem vindo a mostrar uma boa capacidade de resolução de problemas. Porém, uma das fraquezas é o facto de necessitar de um grande número de parâmetros o que vai dificultar uma solução estável, havendo perigo para também desenvolver *overfitting* (Tay et al., 2001). Na literatura foi testado o desempenho preditivo nas series financeiras a técnica ANN comparativamente com a técnica SVM. A conclusão, segundo o autor Tay et al. (2001), é que a técnica SVM mostrou significativamente melhores resultados.

Huang et al. (2005) avaliaram a capacidade de previsão no mercado financeiro das seguintes técnicas: Análise Discriminante Linear, Análise Discriminante Quadrática, ANN e SVM. Uma vez mais, com base em indicadores de análise técnica, o método SVM apresentou os melhores resultados. Os autores sugerem uma combinação de vários métodos para obter ainda melhores resultados.

No presente relatório as técnicas exploradas foram as árvores de classificação e *random forests*. Khaidem et al. (2016) explica que árvores de classificação são uma das técnicas de *machine learning* com possível aplicação nas séries financeiras mas, tal como já foi dito, os dados das séries financeiras possuem normalmente ruído e uma complexidade intrínseca ao mercado financeiro e, com a aplicação das árvores de decisão,

pode haver tendência de *overfitting*. A técnica RF supera este problema ao treinar múltiplas árvores de decisão.

Khaidem et al. (2016) usou o algoritmo RF para testar o desempenho preditivo. Através dos sinais dos indicadores técnicos construiu um modelo com o objetivo de prever a tendência no futuro do movimento do preço das ações no mercado financeiro. Os resultados com o algoritmo RF foram bastante positivos. Os autores apontam que esta técnica de *machine learning* merece ser mais explorada na literatura.

4. Aplicação Empírica

4.1. Método

Numa fase final do trabalho foi implementado um exercício empírico aplicando as técnicas árvores de classificação e *random forests.* Os dados utilizados são as cotações da bitcoin em dólares desde 18 de agosto de 2010 até dia 18 de maio de 2018 e as cotações do PSI 20 desde 1 de setembro de 2010 até 18 de maio de 2018. Estes dados foram retirados do *site* Yahoo Finance[18]. A utilização de técnicas de ML é feita através do programa WEKA[19], que permite a utilização de várias técnicas de ML através de uma interface gráfica. Para as árvores de classificação o programa permite escolher vários algoritmos de construção, tendo sido utilizado um algoritmo denominado J48.

Inicialmente foi feita uma preparação das bases de dados a partir do *Microsoft Excel*. A partir das cotações, foram determinados os valores de vários indicadores de análise técnica, calculados da forma sugerida por Kara et al. (2011), tendo ainda sido considerado o volume de transações. Na tabela 2 é possível ver os indicadores usados.

Após a aplicação destas fórmulas cada indicador é convertido para um sinal que pode ser 0 ou 1. Isto é, para cada indicador, tendo em conta a comparação dos resultados do dia anterior com os do dia corrente, quando o valor diminui o sinal representado, em princípio, é zero, dando a indicação de venda; caso os valores tenham uma variação positiva, ou seja, os valores do dia atual sejam maiores que os registados no dia anterior, é convertido para 1, indicando ordem de compra. Este critério anteriormente mencionado é válido para todos os indicadores exceto para os sinais de RSI e CCI. Para o primeiro a regra é a seguinte: se o valor calculado for maior que 70 então o sinal é 0; se o valor for menor que 30 então o sinal é 1; se o valor estiver entre 30 e 70, o sinal é 1 se o valor do dia estiver acima do do dia anterior e 0 no caso contrário. Para o indicador CCI a regras usada é a seguinte: para valores menores de 200, o sinal é 1; para o caso de o valor ser maior do que 200, então o critério é o seguinte: comparar o valor do dia corrente com o do dia anterior, se for maior então o sinal é 1, senão o sinal é 0.

[18] https://finance.yahoo.com/?guccounter=1. Consultado a 18 de maio de 2018.
[19] https://www.cs.waikato.ac.nz/ml/weka/ . Consultado a 23 de julho de 2018.

Finalizando este exercício resulta uma matriz com os valores dos indicadores convertidos em sinais 0 ou 1. Cada linha desta matriz corresponde a um dia, e a essa linha é adicionado o sinal da rentabilidade do ativo no dia seguinte, que é o que se pretende prever. Esta matriz é dividida numa amostra de treino e numa amostra de teste como forma de validar o desempenho preditivo do modelo treinado. Tendo em conta que o horizonte temporal total são 8 anos foram efetuadas várias divisões: a amostra de treino de 3 anos com os restantes 5 anos na amostra de teste, perfazendo os 8 anos; depois tanto a amostra treino como a de teste com um horizonte temporal de 4 anos e assim sucessivamente até atingir 7 anos para a amostra de treino e 1 ano para a amostra de teste. O objetivo destas sucessivas divisões é obter uma indicação sobre a dimensão dos horizontes de treino e teste que conduz a melhor capacidade preditiva.

Tabela 2 - Descrição dos indicadores (Kara et al., 2011).

Nome do indicador	Fórmula	Descrição
Média móvel simples (10 dias)	$\frac{C_t + C_{t-1} + \cdots + C_{t-9}}{n}$	No cálculo desta média todos os valores possuem o mesmo peso.
Média móvel ponderada (10 dias)	$\frac{(10)C_t + (9)C_{t-1} + \cdots + C_{t-9}}{n + (n-1) + \cdots + 1}$	No cálculo desta média os valores mais recentes possuem o peso maior.
Momento	$\mathrm{C_t} - \mathrm{C_{t-9}}$	Quantifica o montante proveniente da variação do preço num determinado período de tempo.
Estocástico (K%)	$\frac{C_t - LL_{t-(n-1)}}{HH_{t-(n-1)} - LL_{t-(n-1)}} \times 100$	Compara o preço de fecho com as variações do preço num período de tempo anterior.
Estocástico (D%)	$\frac{\sum_{i=0}^{n-1} \mathrm{K_{t-i}}}{10}\%$	Média móvel do indicador anterior.
Índice de Força Relativa (RSI)	$100 - \frac{100}{1 + (\sum_{i=0}^{n-1} UP_{t-i/n})/(\sum_{i=0}^{n-1} DW_{t-i/n})}$	Mede a oscilação dos preços, variando de 0 a 100.
Média Móvel Convergente e divergente (MACD)	$MACD(n)_{t-1} + \frac{2}{n+1} \times (DIFF_t - MACD(n)_{t-1})$	Mostra a tendência entre duas médias móveis.

***Larry William's* R% (WR)**	$\frac{H_n - C_t}{H_n - L_n} \times (-100)$	Mede os níveis de sobrecompra ou sobrevenda.
Oscilador A / D (Acumulação / Distribuição)	$\frac{H_t - C_{t-1}}{H_t - L_t}$	É um indicador que está associado às mudanças nos preços.
Índice *Commodity Channel Index* (CCI)	$\frac{M_t - SM_t}{0{,}015 D_t}$	Mede a variação do preço relativamente à média estatística. Utilizado para identificar os ciclos nos preços.
Volume (10 dias)	$\frac{V_t}{(V_t + V_{t-1} + \cdots + V_{t-9})/n}$	Calcula a relação entre o volume do dia e a média do volume de títulos que foram transacionados nos últimos 10 dias.

C_t é o preço de fecho no dia t, L_t é o preço mais baixo, V_t é o volume de títulos transacionados e H_t é o preço mais alto no dia t. $DIFF_t = EMA\,(12)_t - EMA(26)_t$, EMA é a média móvel exponencial, $EMA\,(k)_t = EMA\,(k)_{t-1} + \alpha \times (C_t - EMA(k)_{t-1})$,α é o fator de suavização que é igual a $\frac{2}{k+1}$, k é o período de tempo da média móvel exponencial de k dias. LL_t e HH_t referem-se ao menor e a maior preços nos últimos t dias, respetivamente. $M_t = \frac{H_t + L_t + C_t}{3}$, $SM_t = \frac{(\sum_{i=1}^{n} M_{t-i+1})}{n}$. $D_t = \frac{(\sum_{i=1}^{n} |M_{t-i+1} - SM_t|)}{n}$, UP_t significa variação de preço no sentido ascendente no dia t, enquanto DW_t é variação de preço no sentido descendente no dia t .

De seguida, as matrizes são inseridas no programa WEKA. O método de amostragem é *supplied test set.* Neste método é inserido previamente o documento da amostra de treino com o objetivo de treinar o modelo e depois o documento da amostra de teste para validar os resultados. O resultado consiste numa previsão do sinal da rentabilidade de cada dia, com base nos sinais dos indicadores técnicos no dia anterior, para cada dia da amostra de teste, e o programa apresenta ainda um conjunto de indicadores sobre a qualidade das previsões. Uma vez concluído este processo, as previsões efetuadas para o conjunto de teste foram exportadas para um documento *Excel*. Com base nas previsões e nas cotações dos ativos, foram calculadas, para cada horizonte temporal, as rentabilidades acumuladas das estratégias baseadas em ML e da estratégia *buy & hold.* Em todos os casos foram usadas rentabilidades logarítmicas, tendo também sido calculado o desvio-padrão das rentabilidades de cada estratégia.

4.2. Resultados

O resultado final retirado do programa WEKA (na figura 9 em anexo é possível observar o painel de controlo da programa WEKA) foi exportado para um documento *Excel* e foram calculadas as rentabilidades acumuladas das estratégias baseadas em ML e da estratégia *buy & hold*. As estratégias baseadas em ML consistem em transacionar os ativos de acordo com as previsões produzidas, ou seja, comprar quando se prevê subida do preço e vender quando se prevê descida. A estratégia *buy & hold* passa pela compra dos títulos numa determinada data e mantê-los durante todo o período de tempo em análise.

De forma a simplificar a interpretação dos resultados obtidos com este exercício prático foi construída uma tabela onde é possível visualizar a soma dos valores da rentabilidade com as estratégias baseadas em ML e com a estratégia *buy & hold*, apresentando valores para as diferentes divisões temporais. Estas divisões temporais coincidem com o horizonte temporal da amostra de teste. No caso do PSI 20 o mercado funciona só nos dias úteis e, portanto, algumas datas tiveram que ser ajustadas. É importante realçar que não foram considerados custos de transação no presente exercício.

4.2.1. Bitcoin

As percentagens de acertos para cada horizonte temporal não variam muito. No algoritmo J48 os resultados foram os seguintes: no horizonte temporal mais reduzido da amostra de teste, ou seja, 1 ano, a percentagem que o modelo conseguiu prever corretamente foi de 50,4%; para 2 anos de amostra de teste obteve o melhor resultado, com 52,5%; para 3 anos de amostra de teste a percentagem desce para 50,7%; de seguida volta subir, 52,2% para 4 anos e, por último, 51,6% para o horizonte temporal maior na amostra de teste, 5 anos.

No algoritmo *random forests* as percentagens de acertos são, em geral, ligeiramente melhores. Apresentando os resultados do menor para o maior horizonte temporal da amostra de teste, temos: 52,3%; 54,3%; 51,3%; 52,4% e 49,9%, respetivamente. Na figura 12 em anexo é possível observar o resultado no programa WEKA para o horizonte temporal da amostra de teste 1 ano.

Quanto às rentabilidades acumuladas, estas tendem a subir à medida que o horizonte temporal é alargado. No algoritmo J48 a rentabilidade acumulada para o horizonte temporal mais pequeno, de 1 ano, é de 103%, para 2 anos é de 143%, para 3 anos é registada uma ligeira quebra para 129% mas é logo recuperada no horizonte temporal de 4 anos para 198%, subindo até 298% em 5 anos. O mesmo padrão se regista quando se usam *random forests*. A rentabilidade acumulada para 1 ano é de 111%, para 2 anos sobe para 159% e para 3 anos para 188%, atingindo o seu máximo quando a amostra de teste é de 4 anos, 264 %. Para o horizonte temporal da amostra de teste de 5 anos, a rentabilidade desce para valores bastante abaixo, 33%.

A estratégia *buy & hold* também segue esta tendência crescente, começando, no horizonte temporal de 1 ano com uma rentabilidade acumulada 144%, para 2 anos 292%, para 3 anos confirma-se a tendência crescente com 357%, para 4 anos observa-se uma queda para 292%. No entanto para 5 anos consegue-se a melhor rentabilidade acumulada, com 421%.

Como forma de complementar a análise anterior é apresentada a evolução das cotações no mercado bitcoin. Na figura 5 é evidente o forte crescimento a partir do ano de 2017. Foi especialmente neste ano que as cotações da bitcoin dispararam. Esta evolução positiva durou até ao final do ano de 2017 (atingindo o seu auge dia 18/12/2017) e, desde aí, a tendência, apesar de algumas flutuações, foi decrescente.

Ainda nesta secção é apresentada uma parte da árvore de classificação. Nessa árvore foi escolhido o horizonte temporal da amostra teste de 3 anos e, devido à sua grande dimensão, foi cortada. Na figura 6 é possível observar uma parte desta.

Neste excerto da árvore de classificação está implícito o seguinte raciocínio: no caso do sinal da média móvel simples (SMA1) ser 1 então a prosseguimos para a análise do indicador RSI, se o sinal deste for 0 então é observado o sinal do indicador CCI. Se o sinal do CCI for 1, passamos para o indicador volume (VA1), se este apresentar sinal de 1 então chegamos a uma das folhas da árvore de classificação concluindo, neste caso, que se prevê uma subida do preço no dia correspondente (isto é, prevê-se um sinal de 1). Esta é a lógica subjacente a toda árvore de classificação. Em anexo nas figuras 10, 11 e 12 é possível observar outros exemplos de árvores de classificação completas.

Tabela 3 - Resultados do algoritmo J48 no mercado bitcoin.

Horizonte Temporal Teste			1 Ano	2 Anos	3 Anos	4 Anos	5 Anos
Estratégia ML	J48	**Instâncias Corretamente Classificadas**	**50,4%**	**52,5%**	**50,7%**	**52,2%**	**51,6%**
		Rentabilidade Acumulada	103%	143%	129%	198%	298%
		Desvio-Padrão Diário	4,66%	3,47%	2,93%	3,18%	5,29%

Tabela 4 - Resultados do algoritmo RF no mercado bitcoin.

Horizonte Temporal Teste			1 Ano	2 Anos	3 Anos	4 Anos	5 Anos
Estratégia ML	***Random Forests***	**Instâncias Corretamente Classificadas**	**52,3%**	**54,3%**	**51,3%**	**52,4%**	**49,9%**
		Rentabilidade Acumulada	111%	159%	188%	264%	33%
		Desvio-Padrão Diário	4,41%	3,53%	3,02%	3,07%	3,81%

Tabela 5 - Resultados da estratégia *Buy & Hold* no mercado bitcoin.

Horizonte Temporal Teste		1 Ano	2 Anos	3 Anos	4 Anos	5 Anos
Estratégia *Buy & Hold*	Rentabilidade Acumulada	143%	292%	357%	292%	421%
	Desvio-Padrão Diário	5,54%	4,51%	3,98%	3,97%	6,60%

Horizonte temporal da amostra de teste: 1 Ano - de 18/05/2017 até 18/05/2018; 2 Anos - de 18/05/2016 até 18/05/2018; 3 - Anos de 18/05/2015 até 18/05/2018; 4 - Anos de 18/05/2014 até 18/05/2018; 5 - Anos de 18/05/2013 até 18/05/2018.

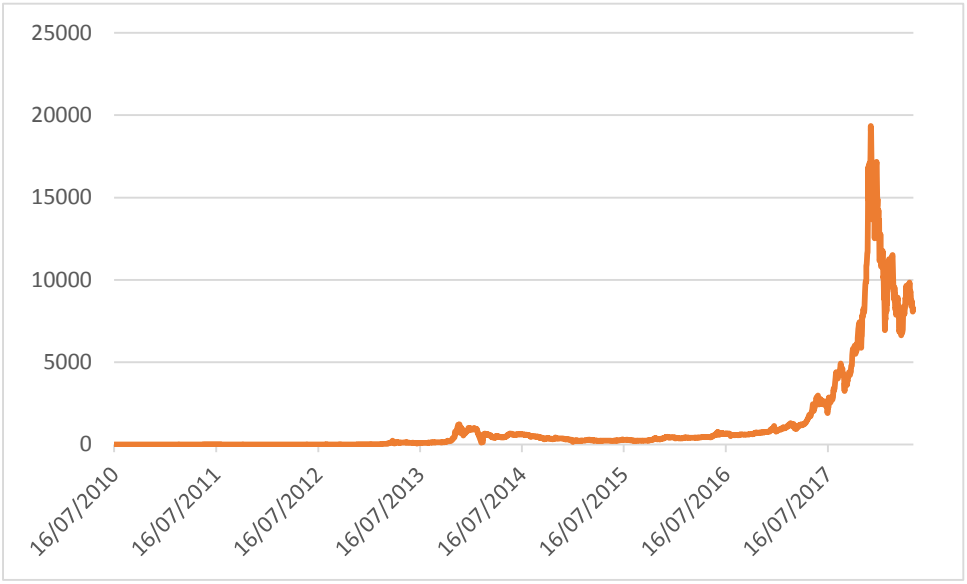

Figura 5 - Evolução das cotações das bitcoin.

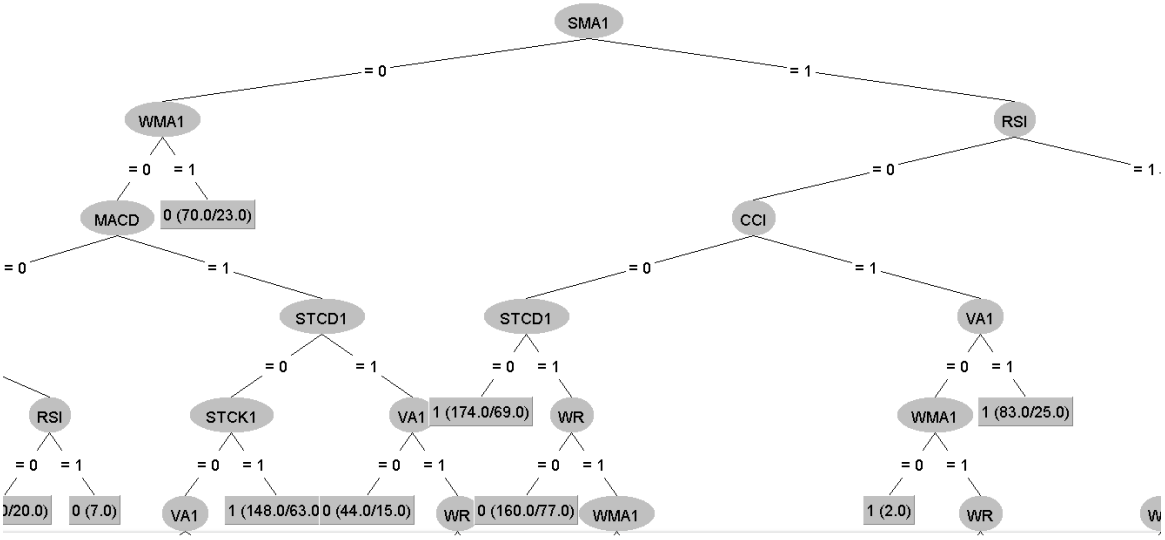

Figura 6 - Árvores de classificação bitcoin - Amostra teste 3 Anos.

4.2.2. PSI 20

As percentagens de acertos no mercado do PSI 20 são, de maneira geral melhores com o algoritmo J48. Para 1 ano de amostra de teste, o resultado é de 50,6%, ou seja, o modelo consegue prever um pouco mais de metade dos valores. Para 2 anos, o resultado sobe para 55%; para 3 anos, a percentagem de acertos é 54,4%; para 4 anos consegue atingir 54,6% descendo depois para 53,6% para 5 anos.

Com o algoritmo *random forests* os resultados são ligeiramente mais baixos à exceção de um dos horizontes temporal. Por ordem crescente de número de anos na amostra de teste, as percentagens de acerto são: 45,9%; 51,8%; 53,6%; 53,1% e 53,3%, respetivamente. No horizonte temporal de 5 anos para a amostra de teste obtém a melhor percentagem de todas. Na figura 15 em anexo é possível observar o resultado no programa WEKA para o horizonte temporal da amostra de teste 5 anos.

No PSI 20, de acordo com os resultados da tabela 6, olhado para os diferentes horizontes temporais existem as seguintes observações sobre a estratégia baseada em ML na rentabilidade acumulada: para o algoritmo J48 começa com uma percentagem de 7%, subindo para 25% para 2 anos, depois com uma ligeira descida para 19% para 3 anos, para 4 a tendência decrescente é contrariada, atingindo 45% e finalizando com 70%.

Para a mesma estratégia, o algoritmo *random forests,* a rentabilidade acumulada possui uma tendência crescente, estando sempre a subir à exceção no horizonte temporal de amostra de teste maior – para 1 ano começa para um valor de 4%, subindo nos próximos horizontes temporais, isto é, para 2 anos 12%, 19% para 3 anos, para 4 anos atinge 21 % e 69% para 5 anos.

Na estratégia *buy & hold* os resultados são alternados pois para o horizonte temporal de 1 ano de teste é de 12%, para 2 anos sobre para 16%, para 3 e 4 anos desce para -7% e -19% respetivamente.

O gráfico 3 mostra a evolução das cotações na bolsa do PSI 20. Apesar das oscilações ao longo do período considerado mantém uma tendência bastante mais estável comparativamente com o mercado das bitcoin. Na figura 6 está ilustrada uma parte da árvore de classificação para o PSI 20. O raciocínio subjacente é o mesmo: neste caso árvore de classificação inicia com o indicador WR (Larry William's R %), se este possuir sinal de 0 desce um ramo em direção ao indicador STCK1 (Estocástico K%) e este, por sua vez, se for igual a 1 desce para uma das folhas da árvore dando fim a este percurso. Em anexo nas figuras 13, 14 e 15 é possível observar outros exemplos de árvores de classificação completas.

Tabela 6 - Resultados do algoritmo J48 no mercado PSI 20.

Horizonte Temporal Teste			1 Ano	2 Anos	3 Anos	4 Anos	5 Anos
Estratégia ML	J48	**Instâncias Corretamente Classificadas**	**50,6%**	**55%**	**54,4%**	**54,6%**	**53,6%**
		Rentabilidade Acumulada	7%	25%	19%	45%	70%
		Desvio-Padrão Diário	0,5%	0,62%	0,88%	0,91%	0,85%

Tabela 7 - Resultados do algoritmo RF no mercado PSI 20.

Horizonte Temporal Teste			1 Ano	2 Anos	3 Anos	4 Anos	5 Anos
Estratégia ML	***Random Forests***	**Instâncias Corretamente Classificadas**	**45,9%**	**51,8%**	**53,6%**	**53,1%**	**53,3%**
		Rentabilidade Acumulada	4%	12%	19%	21%	69%
		Desvio-Padrão Diário	0,52%	0,67%	0,81%	0,88%	0,81%

Tabela 8 - Resultados da estratégia *Buy & Hold* no mercado PSI 20.

Horizonte Temporal Teste		1 Ano	2 Anos	3 Anos	4 Anos	5 Anos
Estratégia *Buy & Hold*	Rentabilidade Acumulada	12%	16%	-7%	-19%	-7%
	Desvio-Padrão Diário	0,68%	0,88%	1,14%	1,21%	1,21%

Horizonte temporal da amostra de teste: 1 Ano - de 18/05/2017 até 18/05/2018; 2 Anos - de 18/05/2016 até 18/05/2018; 3 - Anos de 18/05/2015 até 18/05/2018; 4 - Anos de 19/05/2014 até 18/05/2018; 5 - Anos de 17/05/2013 até 18/05/2018.

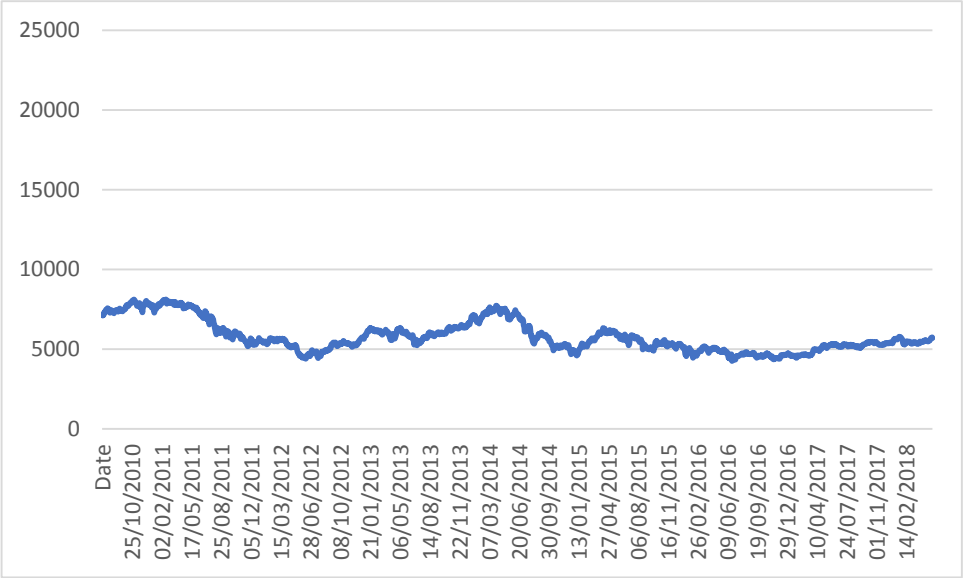

Figura 7 - Evolução das cotações PSI 20.

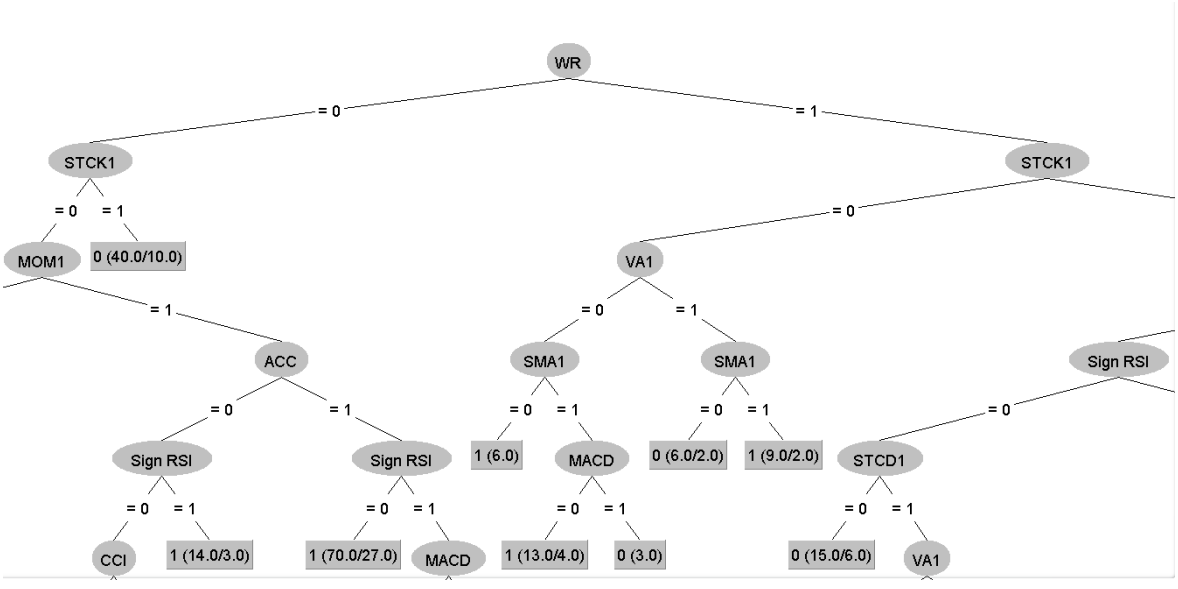

Figura 8 - Árvores de Classificação PSI 20 - Amostra teste 3 anos.

4.3. Análise dos Resultados

Numa primeira análise a partir dos gráficos 1 e 2, da evolução das cotações de fecho, para cada mercado é possível reconhecer padrões de comportamento bastante distintos. Enquanto o mercado das bitcoin obteve um verdadeiro *boom* durante o ano de 2017, o mercado PSI 20, funcionando como um índice de referência no mercado financeiro português, possui uma tendência mais estável. Apesar das flutuações ao longo do tempo, as amplitudes nas variações dos valores de fecho é muito menor. É possível visualizar esta análise comparando a figura 6 com a figura 7.

Olhando para os resultados da tabela 5, no mercado das bitcoin, de uma maneira geral, a estratégia *buy & hold* apresenta resultados melhores. Este facto pode estar correlacionado com o forte crescimento nos últimos anos deste produto financeiro a partir do ano 2017. O valor mínimo registado no período em análise foi no dia 4 de julho de 2013 a fechar com 68,5 e o valor mais elevado registado foi no dia 16 de dezembro de 2017 com 19.345,49. É um aumento cerca de 282 vezes superior, aproximadamente, em quatro anos. Posto isto, a estratégia *buy & hold* é claramente mais vantajosa pois a moeda sofreu uma incrível valorização ao longo do tempo, permitindo uma elevada rentabilidade aos investidores que optaram por a manter. Uma análise da empresa Bloomberg[20] considerou o mercado das bitcoin como um dos mais voláteis em 2017.

A análise dos resultados no mercado PSI 20 mostra-se completamente diferente. Este é um mercado com valores mais estáveis atingindo o mínimo no dia 27 de junho de 2016 (4.260,13) e o máximo dia 1 de abril de 2014 (7.734,95) dentro do período considerado para análise. Dado o padrão de comportamento bastante distinto das bitcoin, as regras obtidas a partir de ML conduzem por vezes a resultados interessantes, melhores do que o *buy & hold*. Observando a figura 7 da evolução das cotações de fecho é possível ver uma leve tendência descendente. Logo a estratégia *buy & hold* não é a melhor neste cenário. Ainda no mercado PSI 20 e seguindo as estratégias dadas pelas técnicas de ML, o algoritmo J48 dá-nos melhores resultados comparativamente com as *random forests*.

Se compararmos a rentabilidade acumulada das estratégias baseadas em ML com a rentabilidade acumulada da estratégia *buy & hold*, no mercado da bitcoin estas

[20] https://www.bloomberg.com/graphics/2017-bitcoin-volume/ . Consultado a 2 de julho de 2018.

estratégias apresentam uma rentabilidade significativamente mais baixa enquanto no mercado PSI 20 obtêm um rentabilidade mais alta. Desta forma, é possível concluir que para o caso do mercado da bitcoin os algoritmos têm um desempenho, em geral, inferior ao da estratégia *buy & hold,* e no caso do mercado PSI 20 possuem um desempenho superior ao da estratégia *buy & hold*.

O desvio-padrão é um importante indicador na análise dos mercados financeiros. O desvio-padrão calculado neste exercício indica uma medida de dispersão da rentabilidade e, desta forma, serve para comparar o risco das estratégias. Fazendo a comparação entre o mercado das bitcoin e o PSI 20, o desvio-padrão para as moedas virtuais é sempre maior do que no PSI 20, facto que não surpreende dadas as caraterísticas deste mercado. Como vimos anteriormente, este é um mercado com determinadas particularidades e a grande amplitude nos valores das cotações de fecho provoca um desvio-padro mais alto. No PSI 20, surge um desvio-padrão menor explicado pela maior estabilidade do mercado.

Analisando as percentagens do desvio-padrão por estratégia, verifica-se que no mercado das bitcoin o desvio-padrão é, de forma geral, mais elevado na estratégia *buy & hold*. No mercado do PSI 20, os valores registam próximos de 1% se arredondarmos à primeira casa decimal. É possível concluir que a estratégia *buy & hold* acarreta mais riscos comparativamente com a estratégia baseado em ML.

Nas tabelas 3 e 4 estão também os valores das instâncias classificadas corretamente para o mercado das bitcoin. Através da sua leitura, é possível verificar o que no mercado das bitcoin o algoritmo *random forests* obteve ligeiramente melhores resultados. Com uma percentagem mais elevada de acertos, o que podemos concluir que neste mercado a *random forests* desempenha melhor a função preditiva. Já no mercado PSI 20, nas tabelas 6 e 7, o algoritmo J48 obtém, de forma geral, melhores resultados.

5. Conclusão

As organizações vivem numa era em que estão conscientes da importância da informação que produzem diariamente. Olhar para esta informação e saber interpretá-la é fundamental para criar uma vantagem competitiva. A exportação de conhecimento a partir de grandes bases de dados não é possível com técnicas de estatística simples. As técnicas da *machine learning* permitem o processamento de avultadas quantidades de informação. Esta informação é uma mais-valia no processo de tomada de decisão, constituindo a principal matéria-prima na construção de modelos de *machine learning.*

A *machine learning* é um método de análise de dados com a particularidade de identificar padrões, aprender com a própria experiência e, assim, gerar conhecimento e recomendar decisões com a mínima intervenção humana. Tal como vimos ao longo deste trabalho a aprendizagem pode ser supervisionada, não-supervisionada, semi-supervisionada ou aprendizagem de esforço.

A aplicabilidade das técnicas de *machine learning* atinge vários sectores de atividade. Porém, no presente relatório o objetivo foi estudar a capacidade preditiva em séries financeiras. Na literatura existente são várias as técnicas de *machine learning* exploradas para testar o desempenho preditivo nos mercados financeiros. Neste trabalho as técnicas estudadas foram as árvores de classificação e *random forests*.

As árvores de decisão apresentam uma boa capacidade preditiva e permitem uma simples ilustração gráfica, facilitando a interpretação do modelo. No entanto, a principal limitação das árvores de decisão é o problema de *overfitting*. Recorrendo à poda este problema pode ser minimizado (Lucas, 2011). Devido à seleção de atributos, as árvores de decisão apresentam robustez face a atributos irrelevantes (Vasconcelos, 2017).

As *random forests* surgem de várias combinações das árvores de classificação para, no final, gerar apenas um modelo (Vasconcelos, 2017). Possuem caraterísticas como boa capacidade de precisão e de generalização das amostras. Os modelos preditivos com *random forests* em geral são tão bons ou melhores do que as árvores de decisão (Lucas, 2011).

No caso empírico foram observadas as séries financeiras bitcoin e PSI 20. São mercados com comportamentos bastante distintos ao longo de tempo, como foi possível verificar. Ao analisar os resultados dos diferentes algoritmos, nomeadamente a percentagem das instâncias corretamente classificadas do modelo, para cada série financeira, foi possível verificar que houve um melhor desempenho preditivo do algoritmo J48 no mercado PSI 20 e do algoritmo de *random forests* no mercado bitcoin.

Após esta experiência empírica foi possível perceber a importância na escolha do algoritmo a explorar dadas as diferentes características da base de dados, neste caso, os diferentes mercados financeiros.

Outra conclusão importante retirada nos resultados do caso empírico neste trabalho é que, se compararmos a rentabilidade acumuladas das estratégias baseadas em *machine learning* com a rentabilidade acumulada da estratégia *buy & hold,* as metodologias da ML parecem ser incapazes de suplantar uma estratégia *buy & hold* no mercado das bitcoin, mas parecem ser superiores a esta estratégia no caso do PSI 20, pelo menos se os custos de transação forem ignorados.

Por fim, a realização do estágio curricular na empresa Feedzai deve ser relevado. Este foi como um processo de vivência prático-pedagógica bastante rico e permitiu-me solidificar os conhecimentos académicos num contexto laboral. Ao longo do estágio fui integrada na conduta da empresa e tive um papel bastante ativo nas minhas funções. O estágio foi uma experiência sem dúvida enriquecedora.

6. Referência Bibliográficas

1. Alpaydin, E. (2004). Introduction to machine learning. Cambridge, Massachusetts: MIT Press.
2. Bastos, D. G., Nascimento, P. S., & Lauretto, M. S. (2013). Proposta e análise de desempenho de dois métodos de seleçao de caracterısticas para random forests s. IX Simpósio Brasileiro de Sistemas de Informação, 49-60.
3. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
4. Hanson, R., Stutz, J., & Cheeseman, P. (1991). "Bayesian Classification Theory." NASA Ames Research Center, Artificial Intelligence Research Branch.
5. Cielen, D., Meysman, A., & Ali, M. (2016). Introducing data science: big data, machine learning, and more, using Python tools. Manning Publications Co..
6. Costa, J. F. A. (2011). Um ambiente gráfico para facilitar tarefas de data mining via ferramenta R (Doctoral dissertation). Universidade do Minho. Disponível em: https://repositorium.sdum.uminho.pt/handle/1822/19829 (25 Julho, 2018).
7. De Souza, R. G. (2016). Previsões Dentro E Fora Da Amostra Da Regra De Taylor Utilizando Fatores Comuns Para O Período De 2002: 02 À 2015: 04. In Anais do XLIII Encontro Nacional de Economia [Proceedings of the 43rd Brazilian Economics Meeting] (No. 063). ANPEC-Associação Nacional dos Centros de Pós-graduação em Economia [Brazilian Association of Graduate Programs in Economics].
8. Ferreira, M. D. F. M. (1999). Árvores de regressão e generalizações: Aplicações. Tese de Mestrado. Universidade do Porto.
9. Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. International Conference on Machine Learning. (Vol. 96, pp. 148-156).
10. Gama, J. (2002). Árvores de decisão. Palestra ministrada no Núcleo da Ciência de Computação da Universidade do Porto, Porto. Disponível em: www.dcc.fc.up.pt/~ines/aulas/MIM/arvores_de_decisao.pdf (25 Julho, 2018).
11. Hall, P., Phan, W., & Whitson, K. (2016). Opportunities and Challenges for Machine Learning in Business. Disponivel em: https://pdfs.semanticscholar.org/cc62/c04074334d1d39b1c9f6a47b1ada99858529.pdf (25 Julho, 2018).

12. Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. Computers & Operations Research, 32(10), 2513-2522.

13. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: Springer.

14. Kara, Y., Boyacioglu, M. A., & Baykan, Ö. K. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. Expert systems with Applications, 38(5), 5311-5319.

15. Kim, K. J. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1-2), 307-319.

16. Khaidem, L., Saha, S., & Dey, S. R. (2016). Predicting the direction of stock market prices using random forest. Applied Mathematical Finance Month. 00(20), 1–20.

17. Kumar, M., & Thenmozhi, M. (2006). Forecasting stock index movement: A comparison of support vector machines and random forest. Indian Institute of Capital Markets 9th Capital Markets Conference Paper.

18. Liaw, A., & Wiener, M. (2002). Classification and regression by random Forest. R news, 2(3), 18-22.

19. Lorenzett, C. D. C., & Telöcken, A. V. (2016). Estudo Comparativo entre os algoritmos de Mineração de Dados Random Forests e J48 na tomada de Decisão. Simpósio de Pesquisa e Desenvolvimento em Computação (SPDC), 2(1).

20. Lucas, L. D. S. (2011). Árvores, Florestas e Sua Função Como Preditores: Uma Aplicação na Avaliação do Grau de Maturidade de Empresas. Revista Pmkt 6 (1): 6– 11. Disponível em: http://www.revistapmkt.com.br/Portals/9/Edicoes/Revista_PMKT_006_01.pdf (26 de Julho, 2018).

21. Marmanis, H., & Babenko, D. (2009). Algorithms of the intelligent web (pp. 69-120). Greenwich: Manning.

22. Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1983). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

23. Mitchell, T. M. (1997). Machine learning. McGraw-Hill.

24. Monard, M. C., & Baranauskas, J. A. (2003). Indução de regras e árvores de decisão. Sistemas Inteligentes Inteligentes Fundamentos e Aplicações, Cap. IV. Rezende, SO Editora Manole Ltda, 115-140.

25. Nair, B. B., Mohandas, V. P., & Sakthivel, N. R. (2010). A decision tree—rough set hybrid system for stock market trend prediction. International Journal of Computer Applications, 6(9), 1-6.

26. Oviatt, B., McDougall P. (2005) Defining International Entrepreneurship and Modeling the Speed of Internationalization. Entrepreneurship: Theory and Practice 29, (5), 537- 553.

27. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.

28. Sanches, M. K. (2003). Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados. Doctoral dissertation. Universidade de São Paulo.

29. Tay, F. E., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. Omega, 29(4), 309-317.

30. Vasconcelos, B. F. B. D. (2017). Poder preditivo de métodos de Machine learning com processos de seleção de variáveis: uma aplicação às projeções de produto de países. Dissertação (mestrado). Universidade de Brasília.

31. Witten, I. H., Frank, E., Hall, M. A. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

7. Anexos

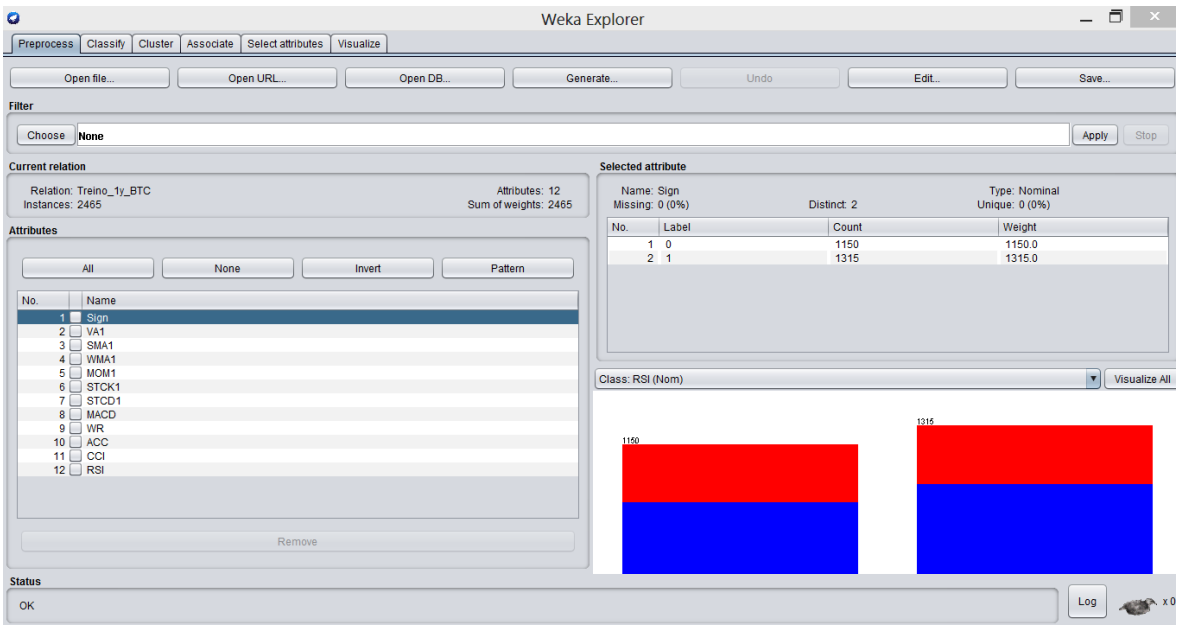

Figura 9 - Painel de controlo do programa WEKA.

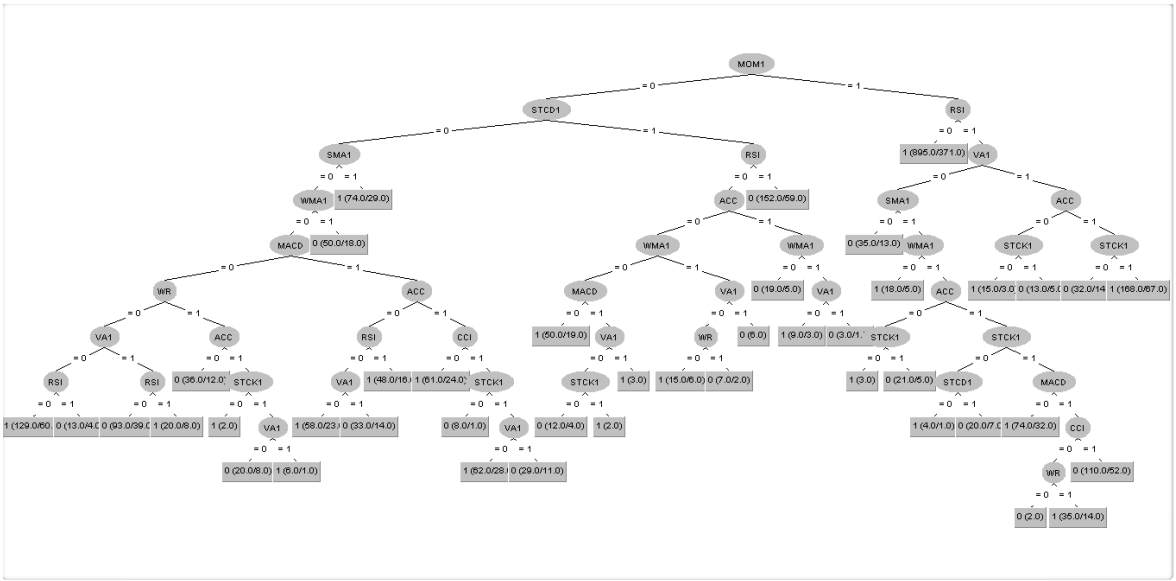

Figura 10 - Árvore de decisão relativa à Bitcoin (teste com 1 ano).

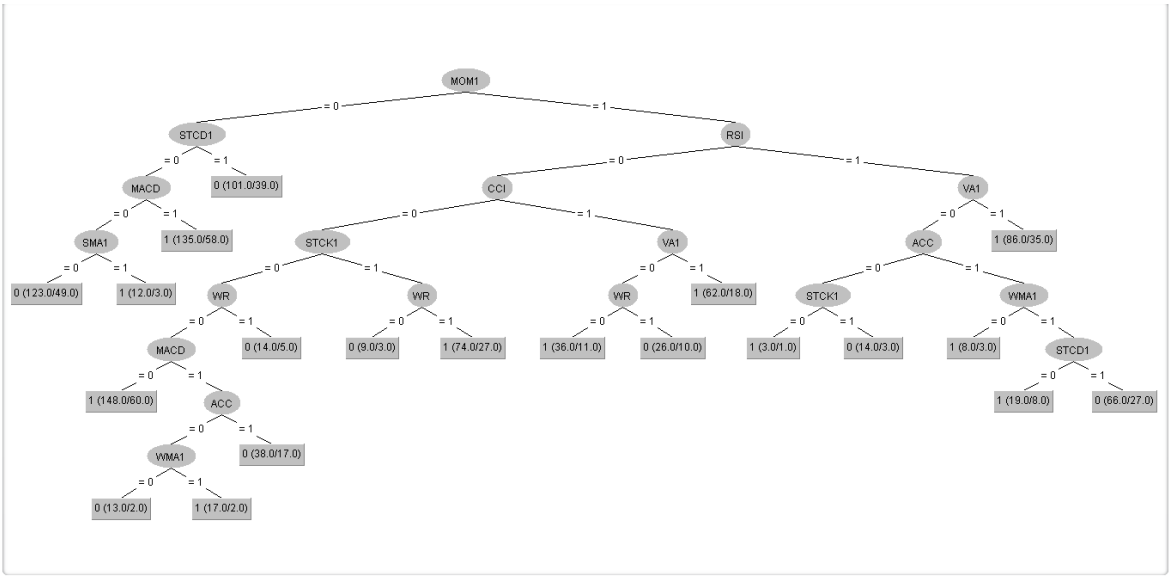

Figura 11 - Árvore de decisão relativa à Bitcoin (teste com 5 anos).

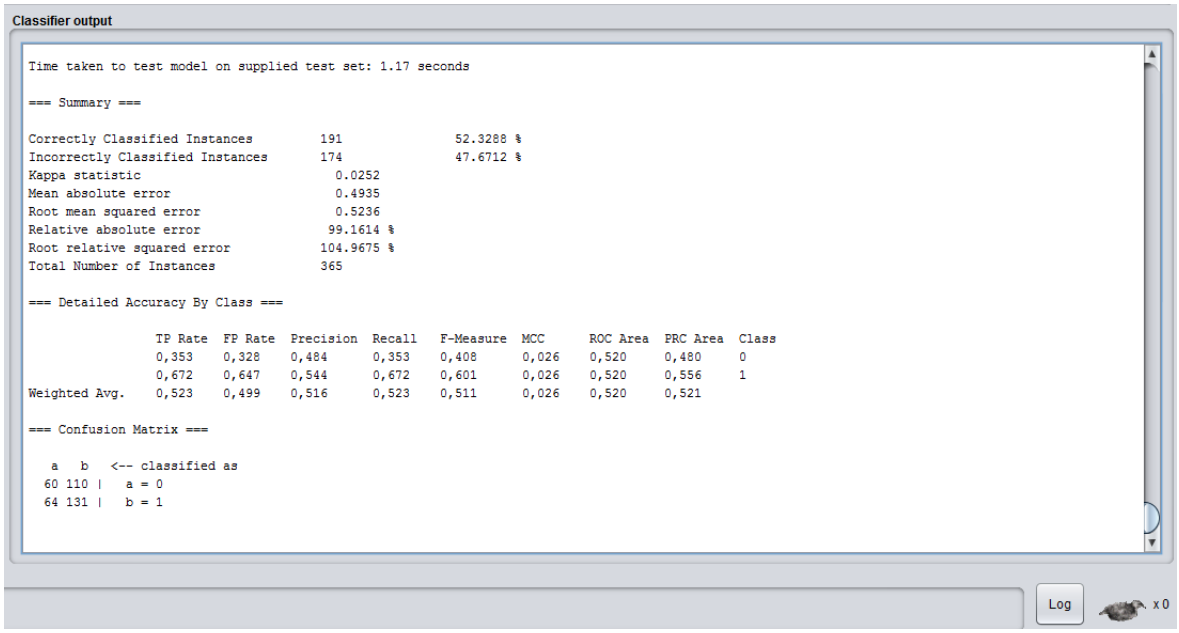

Figura 12 – *Random Forests* relativas à Bitcoin (teste com 1 ano).

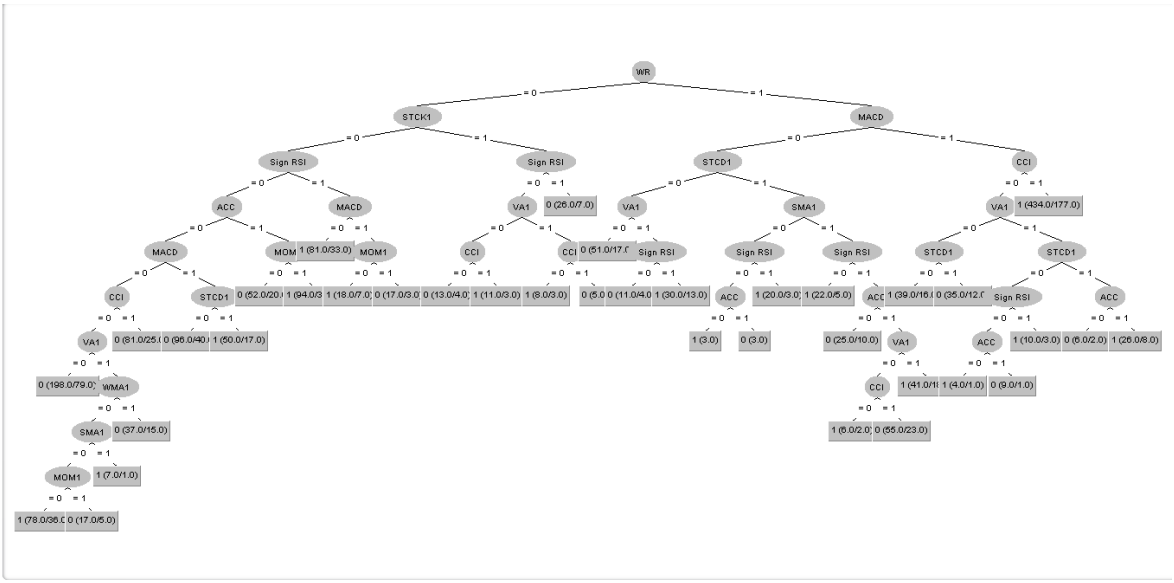

Figura 13 - Árvore de decisão relativa ao PSI 20 (teste com 1 ano).

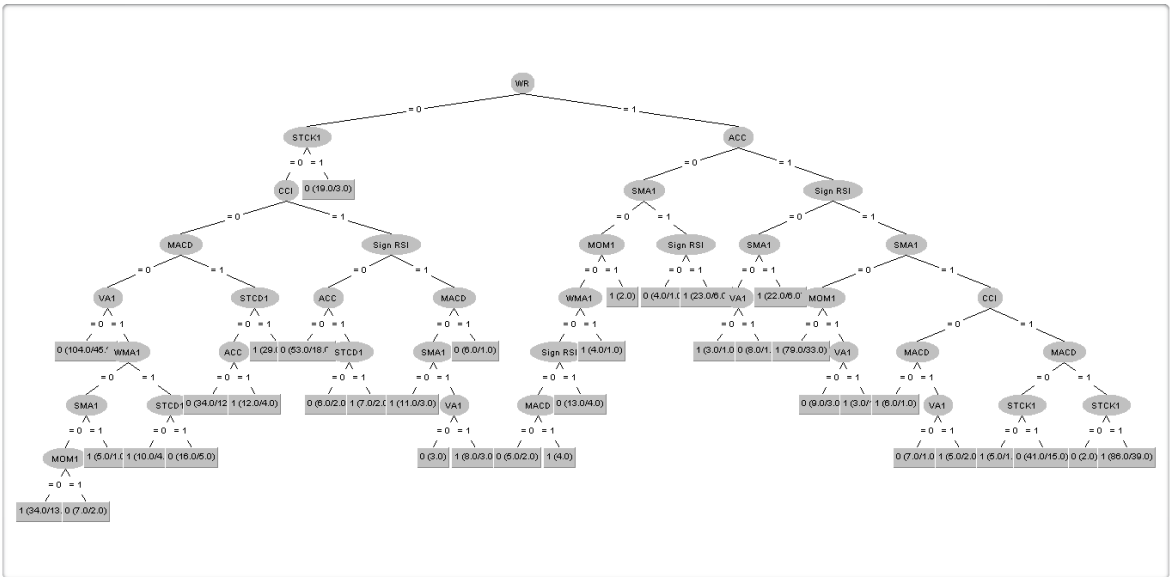

Figura 14 – Árvore de decisão relativa ao PSI 20 (teste com 5 anos).

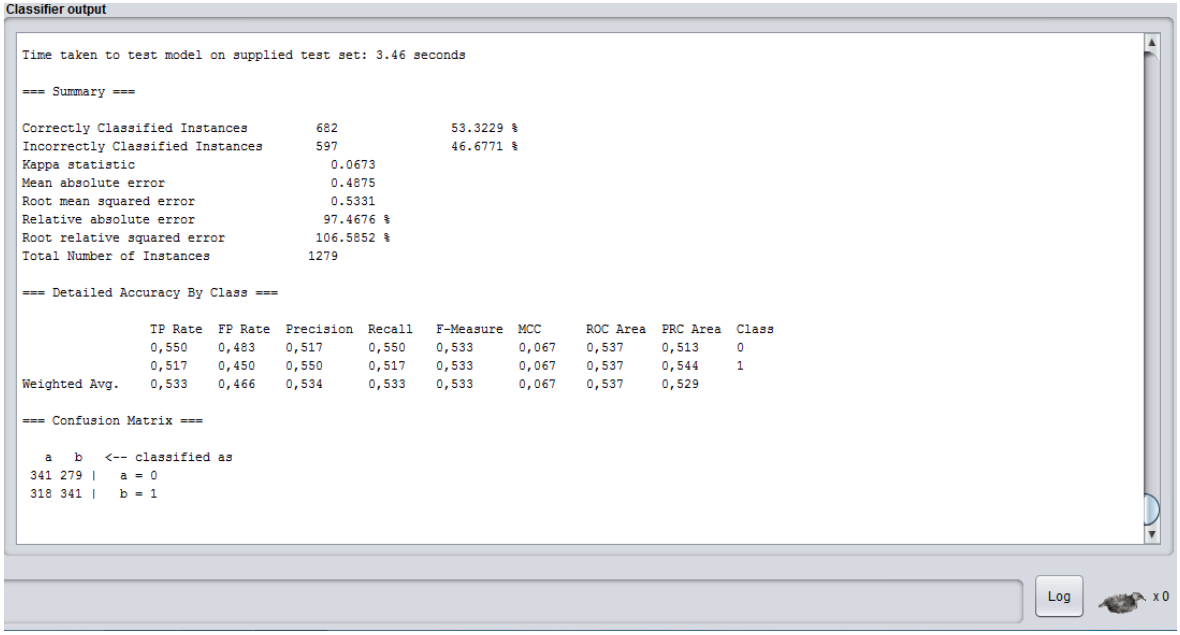

Figura 15 - *Random Forests* relativas ao PSI 20 (teste com 5 anos).