Departamento de Ciências e Tecnologias da Informação

Predição na aviação não regular escrita por Sónia Afonso

Sónia Cristina Moniz Afonso

Dissertação submetida como requisito parcial para obtenção do grau de:

Mestre em Sistemas Integrados de Apoio à Decisão

Orientadora:
Profª Drª Ana Almeida,
ISCTE-IUL

Outubro, 2019

Agradecimentos

Quero agradecer a compreensão e o acompanhamento de todos os professores que lecionaram no meu percurso académico ao longo destes dois anos.

À minha orientadora, em especial, à Professora Doutora Ana Almeida, agradeço toda a sua orientação e disponibilidade, mesmo por vezes fora do seu horário de expediente.

À familia, quero dedicar em especial esta dissertação às minhas duas avós que entretanto faleceram durante a realização deste mestrado, pela força que demonstraram mesmo perante as suas doenças terminais.

À minha mãe pela força das suas palavras não só durante na realização do mestrado, mas durante todo o percurso académico e a sua ajuda foi deveras fundamental.

Ao Marco, pela sua ajuda e extrema tolerância por nem sempre me encontrar disponível, pelo acompanhamento, flexibilidade e apoio emocional em todos os momentos.

Às amigas, por me darem apoio, e fazerem uso da sua persuasão para foco constante no objetivo com palavras que me deram força e empenho.

Àqueles que disseram que nunca seria capaz!

A todos, o meu sincero agradecimento.

Resumo

A crescente procura na aviação comercial em determinados picos operacionais, as avarias inusitadas e as operações charter *Ad-hoc*, fazem com que as empresas tenham necessidade de procura de aluguer de aeronaves a operadores de aviação. Por vezes deparam-se com dificuldades em encontrar aeronaves disponíveis para realizar o serviço ou simplesmente para que não haja custos acrescidos de aquisição, manutenção e de certo prejuízo em época baixa de operacionalidade, preferem não fazer aquisição de aeronaves de reserva capazes de fazer cobertura para todas as exigências operacionais, ou simplesmente, porque o seu tipo de negócio não abrange a inclusão de compra de aeronaves e preferem recorrência ao aluguer.

As empresas de aviação não regular que conseguem colmatar esta carência necessitam ter uma preparação logística atempada. Assim, com este trabalho pretende-se fazer a predição da próxima tipologia operacional, do modelo de aeronave que será procurado e a consequente tripulação necessária para préstimo de serviço a bordo.

A capacidade de preparação com antecedência na resposta operacional ao cliente, adequar o leque de oferta de aviões à procura e a existência de tripulação adequada às necessidades operacionais adjacentes, permite prestar um serviço de qualidade, melhoria da capacidade de resposta e melhoria de organização interna empresarial.

Com esta dissertação pretende-se encontrar modelos de predição com auxílio a aprendizagem automática, aprendizagem automática com recurso a séries temporais e RNN – LSTM (*Recurrent Neural Network - Long Short Memory Term*), encontrando assim entre estes o modelo mais adequado a permitir fazer predição.

Para a aplicação destas técnicas, foram utilizados os dados de gestão de tripulação e dados de planeamento de aeronaves, onde foi possível encontrar modelação adequada à predição da tipologia operacional, com ANN de classificação, para a modelação para determinação dos modelos de aeronaves, os melhores resultados obtidos foram com Árvores de Decisão de classificação e de tripulação, foi determinado com algumas dificuldades com ANN de regressão, a escolha recaiu na melhor performance.

Palavras-Chave: aviação não regular; predição; aprendizagem automática; Séries temporais; data mining.

Abstract

Growing demand in commercial aviation at certain time of operational peaks, facing maintenance problems as AOG (aircraft on ground) and the procurement for Ad-hoc charter operations, means that companies need to seek aircraft leasing from other aviation operators. Sometimes comercial airlines face some difficulties to find available aircrafts to perform their flights or simply to avoid additional costs of acquisition, maintenance and some losses in low peak operating times, instead they prefer not to purchase but rent aircraft capable of covering all operational requirements, or even simply because their type of business does not include the purchase of aircraft and prefer recurrence to rental.

The non-scheduled aviation companies that can fill this gap need to have in advance a logistic preparation, thus, with this work, is intend to predict the next type of operation, the aircraft model to be searched and the convenient crew required for service on board.

Pre-operational customer adequate response, matching the range of aircraft model supply to demand, and adequate number of crew to the consequent operational requirements, enables quality service, responsiveness improvement and higher internal business organization.

This dissertation aims to find prediction models with the aid of machine learning, machine learning with time series and deep learning RNN - LSTM (Long Term Memory Term), finding amongest them the most suitable model to make predictions.

To apply those techniques, crew management data and aircraft planning data were used, where it was possible to find appropriate modeling to predict the operational typology, with ANN classification, to predict the aircraft models, the best results were obtained with Decision Trees classification, and the necessary crew, it was determined with regression ANN, the choice was done having in mind the best performance of each model.

Keywords: Non-Regular aviation; prediction; machine learning; time series; data mining.

Índice

Índice de Tabelas

Índice de Figuras

Lista de Abreviaturas e Siglas

ACMI - Aircraft, Crew, Maintenance and Insurance

ACP - Análise de Componentes Principais

ANAC - Agência Nacional Aviação Civil

COA - Certificado de Aeronavegabilidade

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

EASA - European Union Aviation Safety Agency

FAA - Federal Aviation Administration

GA - General Aviation

GD - General Declaration

IATA - International Air Transport Association

Capítulo 1 – Introdução

1.2. Enquadramento do tema

O negócio de tráfego aéreo da aviação tem-se revelado em efetiva proliferação. No entanto, as empresas de aviação regular nem sempre dispõem de frota com capacidade de cobertura integral de um período operacional anual, não conseguindo por vezes fazer face a maiores picos de atividade ou que permita dar total resposta a acontecimentos inesperados, tais como avarias, acidentes, ou incidentes que têm como consequência períodos de manutenção inusitados, e sendo assim, sujeitas ao aluguer de aeronaves a outras companhias que permitem fazer face a estes acontecimentos.

Existem diversos fatores para que a dimensão da frota de aeronaves de uma companhia aérea não seja exatamente a adequada a todos os períodos ou que permita continuamente fazer face a imprevistos, em períodos de reduzida procura permanecem as obrigações de manutenção a cumprir definidas pela EASA (*European Union Aviation Safety Agency*), o parqueamento das aeronaves e de tripulação em *standby* (estado de prontidão para eventual operação), que são fatores que têm um elevado impacto no acréscimo de custos, transformando-se, por vezes, num prejuízo em determinados períodos.

Muitas das companhias aéreas existentes e alguns clientes de entidades privadas, consoante o seu perfil e o período em causa, com maior ou menor frequência, recorrem ao aluguer de aeronaves de forma a suprir as suas carências de aeronaves ou alugueres temporários.

Dentro de uma companhia aérea que seja não regular, num fretamento podem existir além dos voos regulares e não regulares, voos privados que se referem a voos de índole privativa de acordo com o perfil do cliente; voos de treino para treino de tripulação de *cockpit* (Pilotos e Co-pilotos); voos de posição que se destinam a posicionar o avião numa determinada localização para poder operar conforme requisitado a partir de determinada localização, portanto o avião é deslocado para o aeroporto conveniente à próxima operação, para manutenção ou posicionamento da aeronave num aeroporto comercialmente estratégico; voos de teste, que se referem a voos de verificação da conformidade de performance pós manutenção ou na aquisição de novo aparelho e

também voos militares que se referem ao destacamento de tropas ou operações específicas militares, toda esta variabilidade operacional de uma forma ou de outra são a preparação ou concretização de determinada tipologia de voo. Assim, a tipologia operacional requer que existam assistências operacionais e preparação logística específicas ao voo, uma organização interna departamental e um planeamento realizado com o intuito de ver cumpridos todos os requisitos do cliente. Pelo que exige uma preparação adequada, se for antecipada, poderá melhorar a capacidade de resposta.

Para fazer face à maior procura de determinado modelo de aeronaves pelo mercado, será necessário encontrar a tendência de procura futura, para que, desta forma, o investimento realizado obtenha a procura pretendida por parte do mercado, bem como a previsão do número de tripulação necessária para préstimo de serviço a bordo e de condução da aeronave adequada à sua formação, para que possa dar assistência e operar os voos que possam surgir. Caso a logística necessária seja proposta com antecipação, a margem de erro reduzirá e é mais provável a obtenção do aprazimento do cliente, não só a nível de cumprimento de horários, como de todos os objetivos que são necessários para que a operação seja realizada com sucesso e a margem de lucro mais elevada.

1.3. Motivação e relevância do tema

A definição de uma boa estratégia interna, é um requisito de extrema significância para fazer a companhia prevalecer perante a concorrência em qualquer indústria. O planeamento estratégico é um pensamento sistémico, devidamente estruturado e, organizado para que a empresa alcance os seus objetivos. O planeamento é realizado com detalhe, como a empresa deverá atuar para alcançar os objetivos gerais e funcionais, cumprindo a sua missão e realizando a visão de futuro, sendo necessário ter em vista o comportamento de atividade, e preparando a empresa para oscilações súbitas de necessidades.

Para uma companhia de aviação, o aumento de requerimentos de viagens aéreas traz, não apenas maiores exigências de pessoal habilitado e treinado, mas, também equipamento técnico, ou seja, o número de aeronaves. A encomenda de produção e a própria produção de um avião são muito onerosos e demorados, assim, este torna-se um

outro motivo para realizar pesquisas estatísticas, mapeamento de mercado e observar as tendências de desenvolvimento do mercado (Anna Torun *et al*, 2018). A frota de uma companhia aérea, regular ou não regular, requer um enorme investimento monetário. O custo de uma aeronave pode ascender facilmente vários milhões, pelo que o retorno desse capital é primordial para justificar um investimento desta dimensão.

O crescimento da procura na aviação está a superar os limites de capacidade das empresas, e com o intuito de conseguir acompanhar esse incremento, as organizações devem fazer recurso a previsões para o planeamento estratégico e melhorar a sua perspetiva do futuro (Kasey Panetta, 2018), tal como poderão observar uma habilidade importante na mudança contínua, e permitindo um vislumbre do futuro à frente da mudança (Daryl Plummer, 2018). Assim, as companhias aéreas procuram cada vez mais, a previsão do comportamento da procura por parte dos seus clientes, com o intuito de propiciar uma melhor alocação dos seus investimentos e minimizar imprevistos indesejáveis no futuro.

A predição com maior precisão pode ser essencial para apoiar o planeamento e o crescimento, minimizando perdas com decisões estratégicas baseadas em procuras superestimadas (K.P.G. Alekseev e J. M. Seixas, 2002). A previsão da procura na GA (*General Aviation*), desempenha um papel importante na gestão da aviação, planeamento e formulação de políticas. Por exemplo, o panorama das atividades da GA (tais como a assistência de *handling* (assistência de terra à descolagem e chegada das aeronaves), custo de fuel, posicionamento de tripulacão, entre outros) é um fator usado pela FAA (*Federal Aviation Administration*) para realizar análises de custo-benefício associadas ao desenvolvimento de aeroportos (GRA, 2011).

As metodologias de previsão auxiliam na adoção de estratégias, investimentos e planeamento, de forma a existir preparação conveniente e poder fazer face ao que o futuro reserva. A necessidade de resposta eficaz e adequada, que permita acolher requisitos em permanente mudança e mais exigentes por parte dos clientes, obriga a que as empresas possuam uma flexibilidade e uma enorme capacidade de adaptação. Por exemplo, sabe-se que os passageiros têm mais probabilidade de procurar voos para certos destinos numa determinada altura do ano envolvendo sazonalidade e periocidade, sabe-se também, que

se a viagem for de negócios a probabilidade de se realizar durante a semana é maior que durante o fim de semana (Lisa Pritscher e Hans Feyen, 2017). Assim, as estratégias definidas deverão ter em conta fatores de comportamento do consumidor que causam variação da procura, e, no caso em específico da aviação não regular a variância da tipologia operacional, tipo de aeronave e consequentemente as necessidades de tripulação, devem adequar-se consoante o tipo de cliente e das suas exigências e prioridades. Tal como no caso da aviação deverão ser tidas em conta as regras IATA (*International Air Transport Association*) e da EASA em vigor, e que restringem a operacionalidade contínua da tipulacão. Esta restrições, podem ser diversas, desde a formação especifica para cada modelo de aeronaves, tempo de descanso obrigatórios depois de operar ou períodos intermédios cíclicos de descanso periódicos, ou outras restrições, e na qual desempenha um papel fundamental na assistência e nas condicionantes da própria operacionalidade de voos na aviação. Desta forma, torna-se de extrema importância, tendo como referência a variação da atividade com probabilidade de aumento na procura, fazer a previsão de tripulação adequada de forma a tomar medidas atempadamente.

Com um modelo de previsão capaz de prever a tipologia operacional, o ajustamento de procura de modelos específicos da frota e a tripulação adequada, poder-se-á reduzir o investimento com pouco retorno, já que o equilíbrio entre oferta e procura será mais adequado. Com este trabalho, pretende-se um aperfeiçoamento na resposta operacional, conhecimento do leque de modelos de aeronaves mais requisitados e de recursos humanos adequados, de forma a possibilitar a preparação atempada dos departamentos de planeamento e consequente logística envolvida para poder ter uma boa capacidade de resposta, mesmo em voos *ad-hoc* (inopinados), que são requisitados com pouco tempo de antecedência, perspetivando a melhoria da qualidade de serviços e melhorar a eficácia da resposta. Procura-se providenciar boas bases que possibilitem adoção de uma estratégia pela empresa de aviação não regular, com diferenciação de resposta mais competitiva de negócio, de forma a firmar-se uma boa exploração do mercado, tendo em vista um maior rendimento e maior lucro económico.

O facto de não existir um comportamento estável na atividade da aviação não regular leva-nos à procura de fundamentos de sirvam de suporte para apoio à decisão.

1.4. Questões e objetivos de investigação

Sabendo-se que cada tipologia operacional ACMI (*Aircraft Crew Maintenance Insurance*), Charter, Treino, Teste e privada, envolve diferentes tipos de logística interna departamental, pretende-se uma previsão da operação que o cliente tem em vista contratualizar, podendo existir maior adequabilidade de logística de planeamento de forma a poder fazer face de forma mais expedita aos requisitos de negócio em tempo útil. A primeira questão de investigação será, assim:

Q1: Será que existe dos modelos lecionados neste mestrado, algum mais adequado para predizer a próxima tipologia operacional?

Pretende-se também, de forma a dar seguimento de resposta à procura por parte dos clientes, encontrar a frota satisfatória e otimizada para o mercado, indo ao encontro dos requisitos do cliente tendo em vista maior procura e preenchimento das necessidades. Ao que se pretende responder à questão:

Q2: Existirá a possibilidade de encontrar a modelação adequada para determinar os modelos de aeronaves mais aconselháveis a adquirir, dada a procura?

Por fim, em sequência dos objetivos anteriores, existirá com grande probabilidade uma necessidade de ajustamento do número de tripulação, que permite não só o cumprimento das normas IATA, mas também a viabilização de prestação de serviço de acordo com os requisitos do cliente. Pretende-se encontrar uma base de predição para contratação eficaz, de forma a permitir a formação antecipada compatível com as necessidades. Assim, tem-se em vista a resposta à questão:

Q3: Existirá um modelo capaz de encontrar a quantidade de tripulação adequada à procura?

De agora em diante, a primeira questão e objectivo será mencionada como Q1, a segunda questão e objectivo como Q2 e a terceira questão e objectivo como Q3.

1.5. Abordagem metodológica

A investigação consiste num estudo de caso de uma companhia aérea não regular, que desenvolve atividade em Portugal desde 2005, em franco crescimento desde o ano de fundação. Esta empresa conta com profissionais de diversas áreas no ramo da aviação que asseguram a boa coordenação operacional e o planeamento futuro das operações entre os quais estão: pilotos, co-pilotos, tripulantes de cabine, *ground coordinators* (coordenação das operações em terra), Engenheiros aeronáuticos, Técnicos de planeamento e Logística Operacional, entre outros.

Com os dados produzidos pela atividade de voos operados pela empresa e dos relatórios internos de voo com o número das tripulações e outros detalhes do voo, denominado *General declaration* (GD), pretende-se atingir os objetivos propostos na [Secção 1.4.], pelo que irá recorrer-se ao uso de diversos métodos que permitirão, de forma gradual e de acordo com as fases principais do CRISP-DM (*Cross- Industry Standard Process for Data Mining*), a preparação dos dados e a construção de modelos para posterior aplicação e investigação pretendidas.

A metodologia que serve de base a esta dissertação é a CRISP-DM [Secção 1.6.]. Trata-se de um modelo de processo aplicável a qualquer tipo de indústria. Esta metodologia permite fazer a abordagem ao género de negócio em estudo e desencadeará o decorrer de um ciclo para resposta às questões de investigação.

Para os objetivos propostos de predição, recorrendo-se à metodologia CRISP-DM, na fase de modelação foram aplicados alguns algoritmos de previsão de aprendizagem automática (*machine learning*), séries temporais e ainda abordagem com um algoritmo de aprendizagem profunda (*deep learning)* – LSTM *(Long Short Term Memory)*. Para poder ser utilizado como um ponto de partida de comparação e observação de desempenho com outros algortimos, será necessário construir um modelo base. Este modelo (*baseline*) representará a forma mais simples de obtenção de uma predição e deverá ser visto como o ponto de partida para obtenção de melhoria de resultados (Howarth, Jaokar & Mutlu, 2016). Por norma são utilizadas essas previsões para comparação das métricas de desempenho do modelo base com os restantes modelos, com

o objectivo final, da obtenção de maior desempenho relativamente a qualquer modelo em que foram utilizados os mesmos dados.

1.6. CRISP-DM

Dado o facto de existência de incerteza do que é reservado no planeamento futuro, tanto a nível de predição de procura, como de modelos de aeronaves que serão mais procurados e qual a tripulacão necessária para operar esses voos, torna-se essencial que se recorra a meios e técnicas que permitam construir os modelos que auxiliem no apoio logístico, de forma a poder dar resposta atempada e com menor desperdício de recursos. A metodologia CRISP-DM suporta e permite que seja desenvolvido um processo gradativo, que se irá desenvolver com base nos dados disponibilizados, referentes ao negócio em questão.

Na década de 1990, à medida que a computação e os dados evoluíram para uma necessidade de todas as empresas, as organizações procuraram a optimização de um processo eficiente e estruturado. Desta procura, nasce o modelo de processo denominado, o CRISP-DM. Esta técnica continua a ser uma metodologia padrão para lidar com projetos que são centrados no uso de dados, porque se mostra robusta e ao mesmo tempo fornece flexibilidade e permite personalização adequada ao tipo de negócio. O modelo CRISP-DM descreve as principais etapas envolvidas na execução de atividades de Ciência de Dados, desde a compreensão do domínio do negócio até ao apoio à decisão, mas, mais importante, define uma estrutura que permite iterações em todas as fases. Quando aplicada ao mundo real, a natureza iterativa permite melhorias constantes através de retrocessos para tarefas anteriores e repetindo certas ações.

O CRISP-DM consiste em seis fases de processos principais que são definidas da seguinte forma: Compreensão de Negócios, Compreensão de Dados, Preparação de Dados, Modelação, Validação de Resultados e Comunicação de Conhecimento para Apoio à Decisão e produção.

- Compreensão do negócio, é uma fase onde deve existir a percepção e apreensão dos objetivos do projeto e dos requisitos numa perspetiva do negócio em causa, convertendo este conhecimento numa definição de um problema de extração de

conhecimento e um plano preliminar destinado a alcançar os objetivos propostos. É também necessária a especificação de critérios que serão usados para avaliação e validação do sucesso do projeto do ponto de vista comercial.

- Compreensão dos dados, tem início com a aquisição de dados inicial e prossegue com diversas explorações com o intuito de familiarização com os dados, identificação dos problemas da qualidade nos dados, descoberta das primeiras tendências e comportamentos ou detetação de subconjuntos interessantes para formar hipóteses a partir de informações dissimuladas.
- Preparação de Dados, em que abrange todas as atividades necessárias para construir o conjunto de dados a ser utilizado na fase de construção de modelos preditivos. Parte de dados não "tratados" ou, se necessário, não estruturados ou semi-estruturados, tendo em vista os objetivos propostos de qualidade, capacidada preditiva e completude. Em geral, há necessidade de proceder a diversas tarefas de limpeza, integração e tratamento de dados, ocasionalmente executadas mais do que uma vez, com o intuito de melhoria da qualidade de dados. Estas tarefas incluem atribuição de índices, registo e seleção de atributos, bem como transformação e limpeza de dados para posterior utilização em ferramentas de modelação.
- Na fase de Modelação, são selecionadas e aplicadas várias técnicas de aprendizagem automática para a construção de modelos preditivos e explicativos. Normalmente, existem várias técnicas para o mesmo tipo de problema de extração de conhecimento. Algumas técnicas têm requisitos específicos sobre o formato dos dados. Portanto, por vezes, torna-se necessário o retorno à fase de preparação de dados.
- Validação, promove a avaliação minuciosa dos modelos construídos e a validação dos resultados obtidos a fim de assegurar que são atingidos adequadamente os objetivos do negócio estabelecidos inicialmente, em especial se envolvem capacidade de previsão. Um objetivo chave, em termos de negócio, é determinar se existe alguma questão comercial importante que não tenha sido considerada. No final desta fase, uma decisão sobre o uso dos resultados de extração de conhecimento deverá ser alcançada, porque caso não tenha sido atingida, dever-se-á regressar à fase de afinação dos modelos.
- Apoio à Decisão, nesta fase o conhecimento que foi extraído deverá ser organizado e apresentado de forma a que o cliente possa compreender e fazer uso.

Tal como referido, qualquer uma destas fases pode ter necessidade de repetição ao longo da aplicação do processo, caso se verifique alguma incongruência ou resultados pouco satisfatórios ou, ainda, uma nova compreensão ou conhecimento (Erskine, 2012).

1.7. Estrutura e organização da dissertação

O presente trabalho está organizado em cinco capítulos que aprofundam as diferentes fases, seguindo a metodologia CRISP-DM adequada ao negócio de aviação não regular em análise.

O primeiro capítulo introduz o enquandramento do tema, motivação e relevância do tema, questões de investigação e como é proposta a resposta de resolução para resposta aos objectivos propostos.

O segundo capítulo, referente à Revisão da literatura, refere-se à investigação dos estudos científicos relacionados com o tema da dissertação.

O terceiro capítulo é dedicado à compreensão do negócio da aviação não regular, exploração dos dados provenientes da empresa relativamente a voos, dos relatórios de voos referente à tripulação e preparação dos dados mencionados para posterior utilização, de acordo com a primeira fase a metodologia CRISP-DM.

O quarto capítulo aborda a modelação dos dados, em consonância com a metodologia CRISP-DM, já anteriormente tratados e respectiva avaliação dos resultados obtidos para extração de conhecimento para apoio à decisão na administração da aviação não regular.

No quinto capítulo apresentam-se as conclusões obtidas deste estudo bem como as recomendações, limitações e proposta de trabalhos futuros.

Capítulo 2 – Revisão de Leitura

Neste capitulo será feita uma introdução sobre os trabalhos desenvolvidos para predição na área da aviação ao longo do tempo. Não se verificou qualquer trabalho da aviação não regular, pelo que, toda a pesquisa se baseou em predição na aviação regular e os seus respectivos serviços também utilizados nas restantes tipologias de aviação.

Previsão na aviação

Só durante a segunda guerra mundial começaram a surgir as primeiras preocupações de planeamento de estratégia na aviação militar. Contudo, para a aviação comercial foi apenas após 1945, com o fim da guerra, que se iniciou um elevado incremento da sua procura. Começou por se aplicar inquéritos para suporte da relevância dos fatores que contribuíam para o crescimento de viagens, em especial nos Estados Unidos, que se concentravam em várias características socio económicas dos passageiros das companhias aéreas (Nawal K. Taneja, 1971).

O aumento da procura na aviação comercial tem sido sempre crescente ao longo das décadas após o seu surgimento. Com o aumento do tráfego de passageiros, o fornecimento significativo de aeronaves terá de ser proporcional ao crescimento de necessidades, uma vez que as aeronaves têm vida limitada de serviço operacional, e eventualmente, a aeronave terá de passar à reforma. O crescimento de aeronaves e a sua aposentação deve ser proporcional em relação ao crescimento de passageiros. (Yash Madhwal et al., 2017).

Na Figura 1, pode-se ver a previsão de tendência de aeronaves na Europa.

Size	2015	2035
Large Widebody	100	60
Medium Widebody	320	460
Small Widebody	750	1150
Single Aisle	4010	6630
Regional jet	1730	1520

Figura 1- Growing plane demand (Europe) in 2035 according to plane type

Fonte:(Yash Madhwal et al., 2017)

Dependendo dos diferentes mercados de rotas, as companhias aéreas usam diferentes tipos de aeronaves devido às suas diferentes capacidades. Partindo desta conjectura, as aeronaves maiores são normalmente usadas para o tráfego regular e *charter*, enquanto as aeronaves menores são mais comuns na aviação executiva irregular, que pode ser fretada, alugada ou de propriedade própria. Associada à necessidade de aeronaves, vem a necessidade de tripulação adequada que faça assistência e opere o voo. Na Figura 2, pode-se observar a previsão da necessidade de aumento de tripulação nas diferentes zonas do globo entre 2018 e 2038.

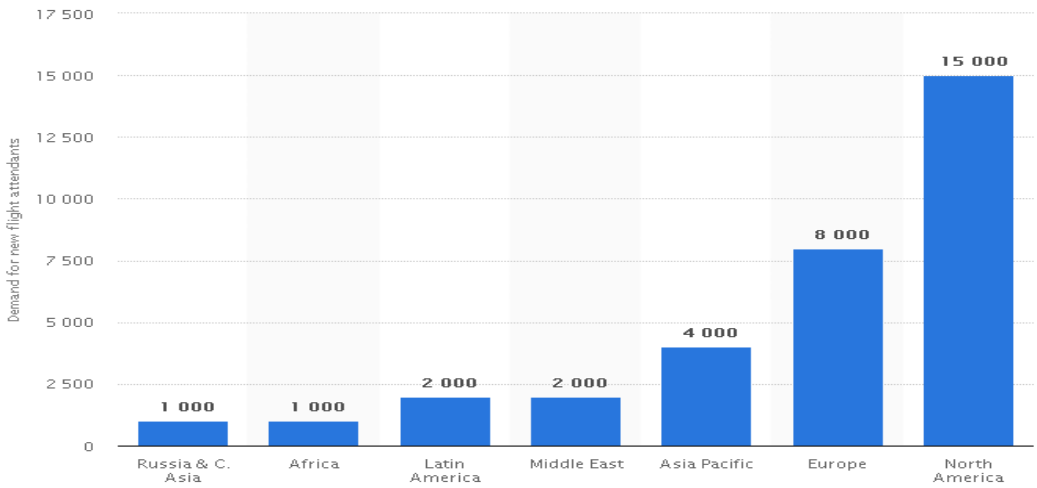

Figura 2- Demand for new cabin crew members in the aviation industry by 2038, by region

Fonte:https://www.statista.com/statistics/617681/new-cabin-crew-demand-globalaviation-industry/

A previsão é a realização de um julgamento sobre o futuro com base no presente conhecimento. Quanto mais precisa for a previsão, melhor será o plano, mas o planeamento pode ser eficaz mesmo que a previsão não seja muito precisa inicialmente (George A. Steiner, 1967). Assim, fazer a previsão da procura das atividades da aviação, tornou-se uma importante incumbência no plano económico (Atef Ghobrial, 1996).

A análise de cenário permite investigar os elementos de um cenário futuro, em particular aqueles que estão em falta no sistema atual, a fim de criar um cenário preferencial. Isto fornecerá algumas orientações sobre como apoiar ou evitar o desenvolvimento em direção a um determinado objetivo. (Johannes Reichmuth e Peter Berster, 2018), portanto, o prognóstico da procura na GA desempenha um importante papel na gestão da aviação, planeamento e políticas de decisão (GRA, 2011). A previsão

da procura de atividades na aviação tornou-se uma tarefa importante no plano económico (Atef Ghobrial, 1996). Por exemplo, a IATA (2018), previu um aumento de transporte aéreo de 3.5 % de taxa de crescimento anual composta nas próximas duas décadas, o que levará ao dobro dos atuais níveis.

A análise dos desenvolvimentos e tendências anteriores relativas ao sistema de transporte aéreo são indicatrizes de um potencial adicional de crescimento a longo prazo, contudo existem alguns indícios em determinados estudos de certos autores que os primeiros efeitos de saturação possam ser vistos nos mercados norte-americano e europeu, ao contrário do asiático. Pelo que a aposta de negócio deverá ter em vista estes estudos e uma maior aposta em mercado com previsões de crescimento mais sólidas.

Para Chen *et al.* (2012) sugerem que a análise de negócio e as tecnologias associadas podem ajudar as organizações a entender melhor o seu negócio, o mercado e as oportunidades presentes perante a abundância de dados e com análise de domínio específico. Para Rajeev Sharma *et al.* (2014), o conhecimento que inclui um entendimento de tendências, operações, clientes e fornecedores, sugere várias opções de exploração e conversão em valor. Algumas opções podem ser óbvias, enquanto outras podem ser o resultado de um processo mais criativo, por exemplo, envolvendo analistas e ignorar uma das restrições atuais e criando novos modelos de negócios.

Os investigadores estão gradualmente a dar início à adoção de técnicas de aprendizagem automática para diversos problemas da aviação, mas esta implementação não é rápida o suficiente devido à indisponibilidade de dados com boa qualidade, assim, é mantida uma elevada confiança na opinião dos especialistas, em vez de modelos complexos orientados para os dados. Adicionalmente é usual ser necessário um grande volume de dados e uma grande capacidade computacional para criação de modelos, que nem sempre se encontram disponíveis (Apoorv Maheshwari et al, 2018).

A disponibilidade de técnicas de aprendizagem automática é elevada e tem sido mais utlizada na aviação nos últimos anos. Foi possível analisar trabalhos de diversos autores, com recurso a várias técnicas de aprendizagem automática para determinação de diferentes objectivos propostos. Dado o facto de não terem sido encontrados trabalhos científicos com o objectivo específico desta dissertação, no restante descrever-se-ão

alguns dos trabalhos na área da aviação que recorreram a algoritmos de aprendizagem automática, que servem de base de investigação e foram também utilizados para investigar as respostas aos objectivos propostos.

O uso de árvores de decisão e de regressão na aviação, em que por norma se obtém bons resultados, é já algo recorrente, como, por exemplo, na determinação de classificação das classes nas falhas do sistema de *de-icing* (quando é efetuada a aplicação de químicos para retirar o gelo ou neve da aeronave) (WANG *et al*., 2013), e de regressão na determinação da procura de viagens aéreas, por DeLaurentis *et al.*, (2018).

Os autores Urkude & Richariya (2016), fizeram um estudo onde utilizaram o algoritmo de *Naive Bayes* com o objetivo de conseguir que as companhias aéreas consigam antecipar e lidar com atrasos antes que aconteçam, de forma a evitar as compensações por atrasos adicionando uma compensação ao sistema, e obtiveram bons resultados na predição de voos tendo em consideração os factores temporais.

Lee *et al*, (2016) usaram dados de voos fornecidos pela companhia aérea *American Airlines* para comparação de várias técnicas de aprendizagem automática para prever o tempo de táxi no aeroporto de Charlotte. Fizeram demonstrações com regressão linear, (SVM), (KNN), (Random Forest) e redes neuronais. Demonstraram que estes algoritmos de aprendizagem automática são capazes de prever o tempo de táxi para partidas numa janela de tempo de 5 minutos. Outros dos trabalhos na área da aviação, que fez recurso ao uso de SVM, terá sido o dos autores Zhang & Mahadevan (2019), com um modelo híbrido, combinando a SVM com um conjunto de redes neuronais de *deep learning*, para quantificar o risco de prever a consequência de eventos perigosos no sistema de transmissão das aeronaves.

Os autores Apoorv Maheshwari et al. (2018) realizam também a comparação entre vários algoritmos de aprendizagem automática, com o propósito de modelar a procura de viagens aéreas. Utilizaram algoritmos de CART (Classification and Regression Trees), de Regressão, SVM, ANN e um conjunto de métodos para estimação e predição da procura de viagens aéreas. Concluíram, de uma forma geral, que os algoritmos de SVM e ANN têm boa precisão para dados com mais elevada complexidade, mas existe a necessidade de maior volume de dados para obtenção de bons resultados. Referem que

existe um consenso geral de que o ANN é muito sensível a características irrelevantes, devido à forma de construção do algoritmo. Isto pode ter como consequência um processo de treino bastante ineficiente e demorado.

Dada a presença de factor temporal nos dados para este estudo, foram também investigados alguns artigos científicos com recurso ao uso de séries temporais.

LALIŠ (2017), no seu artigo resumiu as opções para a aplicação de análises robustas de séries temporais com o objetivo de previsão do índice de desempenho de segurança na aviação, tendo, no entanto, algumas limitações dada a obrigatoriedade de confidencialidade dos dados.

Tang e Deng (2016), usaram predição baseada no modelo ARIMA, e foi possível apurar que o volume de passageiros aéreos manterá uma tendência ascendente constante, bem como, o mercado de passageiros aéreos ficará maior na China, e ainda, a rotatividade de passageiros apresenta uma certa flutuação sazonal, cujas principais razões incluem férias, clima e outros factores sazonais ou feriados. Para Cheng Li (2019), o modelo ARIMA de séries temporais é utilizado para a previsão do volume anual de passageiros da aviação civil na China, e apresenta uma análise matemática precisa. Segundo o autor, o risco de erro de previsão também pode ser reduzido, com o método combinado de previsão baseado em ARIMA-Regression.

Sa (1987), estudou os modelos SARIMA para obter previsões que auxiliassem na gestão de custos de um determinado voo. Contudo, os resultados obtidos não foram promissores. Ainda com recurso a SARIMA, um modelo híbrido de séries temporais SVR (*Support Vector Regression*) foi utilizado, para previsão da procura do setor de aviação por Hing *et al.* (2019), apresentando resultados com bom desempenho.

Ainda usando a abordagem SARIMAX, foi feita uma previsão do número de passageiros, tendo em consideração as oscilações no preço do petróleo, taxa de câmbio SEK / EUR, durante o período da Páscoa. As variáveis explicativas foram usadas por adição recursiva para determinar se seria vantajoso usá-las na totalidade ou não como *input* (Robertson & Wallin, 2014).

Também com recurso ao modelo sazonal-ARIMAX é realizado um estudo empírico combinando o modelo SARIMAX com a intensidade de pesquisa do Google, onde se

mostra que o índice de pesquisa no Google sobre estacionamento, transporte ou outras informações relacionadas com os aeroportos tem um poder explicativo significativo do seu volume aéreo de passageiros. (Ji *et al*. 2015).

Utilizando uma RNN – *recorrent neural network,* foi desenvolvida e implementada uma abordagem para prever o progresso do embarque de passageiros, usando informações de ocorrências na cabine, tais como chegadas tardias, número específico de itens de bagagem de mão ou passageiros prioritários (com embarque privilegiado). Espera-se que no futuro exista uma cabine na aeronave interligada e forneça as condições ambientais através de sensores, que possam ser usados como entrada operacional para uma previsão em tempo real do progresso de embarque e melhorar de forma sustentável a gestão de preparação para uma nova operação (Schultz & Reitmann, 2018).

O modelo de previsão de *hard landing* (aterragem efetuada abruptamente que causa danos na aeronave ou até mesmo nos passageiros a bordo) com recurso a LSTM composto por 5 camadas, com uma camada de entrada, uma camada de saída e três camadas ocultas, obteve o melhor ajuste em relação aos restantes modelos experimentais e baseando-se, na aceleração vertical (VRTG), conseguiram fazer previsão deste valor no próximo momento, a fim de conjecturar a ocorrência de acidentes de *hard landing*. (Tong *et al*., 2018).

Pretende-se com o presente estudo a colmatação da falta de investigação àcerca da previsão sobre tipologias operacionais, modelos de aeronaves adequadas à procura e a respetiva tripulação necessária recorrendo a algoritmos de aprendizagem automática. A intenção é a de apoiar as decisões de optimização de custos, predição de operacionalidade, e de necessidade de pessoal, tendo também em consideração a respectiva sazonalidade. Apesar das pesquisas efetuadas, atrás descritas, não foi possível encontrar este tipo de estudo, em particular no que concerne à aviação não-regular, pelo que fica estabelecida, assim, a necessidade de estudo dos objectivos propostos nesta dissertação.

Capítulo 3 – Compreensão e exploração do negócio da aviação não regular

Tal como anteriormente referido, foi selecionada a metodologia CRISP-DM por se mostrar mais adequada a problemas cuja solução advém da análise de dados de negócio, independentemente do ramo, sendo uma metodologia especificamente desenhada para processos de extração de conhecimento dos dados (*Data Mining*). No que se segue neste capítulo, será associada cada seção às primeiras fases do ciclo do CRISP-DM, portanto a Compreensão do negócio, Compreensão dos dados e Preparação dos dados, estas fases não foram alvo de um desenvolvimento único e directo, pelo que houve necessidade de diversos ciclos de experimentação e reformulação.

3.1. Compreensão do Negócio da aviação não regular

O cliente contacta a empresa solicitando por norma cotação do serviço pretendido, sendo necessária a indicação dos aeroportos, regiões ou países pretendidos ou preferenciais, o intuito do seu voo, tipologia operacional pretendida, com ou sem tripulação incluída (*wet lease e dry lease*), tipo de serviço a bordo e os passageiros previstos.

Em caso de dúvida e a contratação seja para aeroportos que nunca foram operados pela empresa, os principais departamentos com os serviços que têm mais ponderação nos custos associados e os operacionais de suporte a voos, são consultados com o intuito de obtenção de cotações, dada a variação frequente nos destinos pretendidos, tais como o *fuel*, as autorizações de sobrevoo, os *handlings*, as taxas de aeroporto, *catering*, hotéis e posicionamentos da tripulação. As restrições e condições de performance do voo ou voos são também assinaladas pelo o departamento de *Flight Support* (suporte e performance de voos) e é também feito o estudo de adequabilidade do equipamento disponível da aeronave que irá realizar o voo ao aeroporto.

Posteriormente é então calculado o custo dafretagem e feita a orçamentação ao cliente do custo da operação pelo departamento Comercial, tendo em conta não apenas os fatores mencionados, como também a época de maior ou menor procura que se encontra a decorrer de momento e quais as aeronaves que estarão disponíveis. Caso seja um serviço

para um destino regular, a cotação será dada de acordo com a tabela padrão de custos de serviços ACMI ou charter.

Seguidamente a haver aceitação por parte do cliente e fecho de contrato, existe uma divulgação interna geral para a empresa confirmando a data da operação e o cliente ou empresa de contratação (*brooker*) e nessa altura é desencadeado o início de preparação de cada departamento de forma a poder dar resposta e a devida preparação para a operação.

- ***Departamentos***

O departamento de fuel irá solicitar cotações às empresas de fornecimento de combustíveis de aeronaves em cada uma das estações, irá verificar a possibilidade de optimização de abastecimento, com a estação com cotação mais baixa por galão ou tonelada. O departamento de Despacho de Voo, tendo em conta o custo do fuel, o consumo previsto de acordo com o atrito causado pelo tempo, o plano de carga e rota preferencial de acordo, indica, qual a estação em que deverá abastecer mais quantidade.

O departamento de planeamento, actualiza o planeamento e solicita as autorizações de sobrevoo e *slots* (autorização de horário de aterragem ou descolagem associado ao *callsign*). Irá fazer a solicitação do plano de voo e verificar as FIRs (pontos de entrada ou corredores de circulação de um país) sobrevoadas e, desta forma, solicitar a cada uma das autoridades de aviação civil o sobrevoo e aterragem consoante o país, ou *slots*, dependente das normas IATA de cada país.

O departamento de *Travel* irá verificar a informação em sistema das respectivas necessidades de posicionamentos e hotéis de acordo com as escalas determinada pelo departamento de escalonamento de tripulantes (*Crewing Department*) e verificar as cotações de mercado juntos de agências de viagem de companhias aéreas, *transfers* e hotéis tendo também em vista a optimização do custo operacional ou verificação de contratos directos com o fornecedor, consoante a duração da operação.

O departamento de *In-flight*, irá confirmar junto do departamento Comercial os requisitos referentes tanto a refeições que o cliente pretende como o tipo de serviço a

bordo, podendo ser desde o básico a VIP, e no caso de existir algum requisito em especial, existirá consequentemente custo adicional se aplicável.

O departamento de Direção de Operações de Terra (DOT) irá organizar o *handling* e verificar as cotações disponiveis, tendo em vista também taxas aeroportuárias de passageiros, torre de controlo, entre outros serviços necessários de assistência em terra à aeronave, e optar pelo mais vantajoso.

O departamento Centro de Controlo Operacional (CCO) irá dar apoio de serviços de todos os departamentos nas alterações de última hora, inconsistências e falhas, já durante a realização do voo, ou irá substituir na sua ausência, por se tratar de um departamento em funcionamento 24 horas, e requisitar todos os serviços, no caso de uma operação fora do horário de expediente, em que os restantes departamentos não se encontram presentes.

O departamento de Facturação irá confirmação a inclusão de todos os serviços tidos pelos departamentos mencionados na factura final ao cliente e fazer reconciliação de contas sempre que tenha sido necessário custos extra.

O departamento de manutenção irá planear e assegurar que as aeronaves se encontram com todas as inspeções obrigatória pela EASA efetuadas, sempre que possível sem impactar comercial o aluguer das aeronaves.

Os departamentos de *Safety* e *Security* que asseguram o cumprimento das regras de segurança desde os departamentos de planeamento interno até aos posicionamentos que necessitarão de proteção da tripulação em determinados países.

3.2. Recolha de dados

Os dados utilizados provenientes do planeamento de aeronaves e de gestão de tripulações, foram obtidos a partir da ferramenta de gestão de voos da organização. Esta ferramenta permite a extração dos dados num formato que permitirá posterior seguimento ao tratamento dos mesmos e finalmente o carregamento na ferramenta para modelação (ETL).

Embora os dados tenham sido obtidos da mesma aplicação, foram retirados de módulos diferentes. Uma das extrações refere-se à gestão de tripulação e o outro módulo

referente à gestão de voos das aeronaves, onde se obtêm todas as informações referentes aos detalhes de voos previstos (tempo de voo, hora prevista de partida, hora prevista de chegada, tempo entre calços, tempo efetivo de voo, aeroporto de chegada, aeroporto de partida, passageiros e companhia aérea). Referente aos dados das tripulações, é utilizada a informação referente aos relatórios de GD (*General Declaration*), que corresponde à extração de informação decorrida no voo, (tripulação de cabine e de *cockpit* que efetuou o voo, a sua categoria, dados pessoais, número de voo, origem e destino).

A ferramenta de gestão aeronáutica tem suporte de uma empresa externa, onde é possível a realização de gestão e apoio de voos, bem como de tripulação, permitindo ainda efetuar a extração dos dados pretendidos, tanto a nível de voos efetuados, como de tripulação que operou para os voos.

Estes dados foram extraídos num universo de voos ocorridos entre 2010 a 2018 e incluem todos os voos operados pela companhia aérea, voos ACMI, Charter, Treino, Teste e Privado desempenhados ao longo dos anos mencionados, bem como toda a tripulação a nível de cabine e cockpit (tripulação de conduz a nave) que assistiu e operou nestes voos. Na Tabela 1, está presente a descrição das variáveis que compõem os dados, provenientes de origens diferentes, e que serão utilizados nos modelos, com 45929 registos de operacionalidade de voos, e um total de 43047 registos de tripulação, portanto provenientes da General Declaration.

Descrição das variáveis dos conjuntos de dados referente aos voos operados:

Nome	Tipo	Origem extração	Formatos	Descrição
AircrafType	Categórica	xlsx	String	Modelo de aeronave
AircrafReg	Categórica	xlsx	String	Matricula aeronave
Departure	Categórica	xlsx	Date	Local de partida
Arrival	Categórica	xlsx	Date	Local de chegada
Year	Temporal	xlsx	Date	Ano
Month	Temporal	xlsx	Date	Mês
Dia	Temporal	xlsx	Date	Dia
Weekday_num	Temporal	xlsx	Date	Dia da semana
HEstimTD	Temporal	xlsx	Date	Hora estimada de partida
mtsEstimTD	Temporal	xlsx	Date	Minutos estimados de partida
HActualTD	Temporal	xlsx	Date	Hora efetiva de partida
mtsActualTD	Temporal	xlsx	Date	Minutos efetivos de partida
HEstimTA	Temporal	xlsx	Date	Hora estimada de chegada
mtsEstimTA	Temporal	xlsx	Date	Minutos estimados de chegada
HActualTA	Temporal	xlsx	Date	Hora efetiva de chegada
mtsActualTA	Temporal	xlsx	Date	Minutos efetivos de chegada
stateflightDepartDiff	Temporal	xlsx	String	Estado do voo em relação à partida
HDepartDiff	Temporal	xlsx	Horas	Diferencial entre hora prevista de partida e efetiva
mtsDepartDiff	Temporal	xlsx	Minutos	Diferencial minutos previstos e efetivos de partida
mtstotaisDepartDiff	Temporal	xlsx	Minutos	Diferencial total minutos previstos e efetivos de partida
stateflightArrivalDiff	Temporal	xlsx	Minutos	Estado do voo em relação à chegada
HArrivalDiff	Temporal	xlsx	Horas	Diferencial entre hora prevista de chegada e efetiva
mtsArrivalDiff	Temporal	xlsx	Minutos	Diferencial minutos previstos e efetivos de chegada
mtstotaisArrivalDiff	Temporal	xlsx	Minutos	Diferencial total minutos previstos e efetivos de chegada
HAirborn	Temporal	xlsx	Horas	Número de horas voadas
mtsAirborn	Temporal	xlsx	Minutos	Número de minutos voados
mtstotaisAirborn	Temporal	xlsx	Minutos	Diferencial total minutos voados
HBlockH	Temporal	xlsx	Horas	Horas entre blocos
mtsBlockH	Temporal	xlsx	Minutos	Minutos entre blocos
mtstotaisBlockH	Temporal	xlsx	Minutos	Minutos totais entre blocos
CorrectOperationType	Categórica	xlsx	String	Tipologia Operacional
Airline_anon	Categórica	xlsx	Inteiro	Companhia aérea anonimizada

Tabela 1 – Descrição das variáveis referente aos dados dos voos operados

Descrição das variáveis dos conjuntos de dados referente à gestão de tripulações:

Nome	Tipo	Origem extração	Formatos	Descrição
AircrafReg	Categórica	csv	String	Matricula aeronave
Departure	Categórica	csv	Date	Local de partida
Arrival	Categórica	csv	Date	Local de chegada
Year	Temporal	csv	Date	Ano
Month	Temporal	csv	Date	Mês
Dia	Temporal	csv	Date	Dia
Airline_anon	Categórica	csv	Inteiro	Companhia aérea anonimizada
CrewNum	Numérica	csv	Inteiro	Numero de tripulação

Tabela 2- Descrição das variáveis referente aos dados de gestão de tripulações

3.3. Preparação de Dados

Como referido, as extrações dos sistemas foram realizadas em dois módulos diferentes, referente a voos e tripulação associada. Na extração do primeiro módulo referente a voos foi possível a extração única do sistema em formato *.xlsx*, contudo, para extração do segundo módulo referente ao report de GDs dada dificuldades de extração de grande volume de dados do sistema de informação de tripulação e respectivos detalhes básicos do voo realizado, devido ao enorme volume de dados abrangente nestes ficheiros, foram extraídos os relatórios de *GD* por semestres anuais em formato *.csv*.

Foram realizadas análises de valores nulos e ausentes, de forma a verificar qual a melhor forma de lidar com os mesmos para ambos os conjuntos de dados.

Assim, para a amostra referente aos voos, por existirem poucas observações com estas características, no tratamento de dados omissos ou incorretamente introduzidos, procedeu-se à remoção das 41 observações, de forma a obter uma amostra com dados passíveis de carregamento em *Python*, com eliminação das linhas com presença de *Na*. Este conjunto de dados, com o detalhe dos voos, consiste nos dados utilizados para a obtenção de resposta para as questões objectivo Q1 e Q2 desta dissertação. Procedeu-se, ainda, à remoção das aeronaves que já não fazem parte da frota ou que os modelos foram considerados obsoletos dado que já não são fabricados pela *Airbus*, uma vez que não eram convenientes para aprendizagem com vista a previsões futuras.

Relativamente à amostra de GD das tripulações, que inclui a tripulação que operou os voos, houve necessidade de tratamento individual para cada ano, colocando a informação em *dataframes* independentes, dado que o *Python* fazia fusão e sobreposição de campos com dados dispersos na extração, havendo total desformatação para poderem ser explorados correctamente ou carregados nos algoritmos de modelos. Só desta forma foi possível proceder à limpeza de dados, aplicado especificamente a cada *dataframe.* Houve também necessidade de resolução de valores *Na,* com a sua substituição por média dos valores imediatamente seguinte e anterior da sua vizinhança, quando existiram voos próximo da data da ausência. Caso contrário, procedeu-se à inserção de zeros, já que a ausência de introdução de dados, em especial nos voos de posição, treino e teste, corresponde a lapsos ou omissão de introdução dos dados ou de erro de sistema.

Em ambas as amostras foi necessário recorrer à implementação de técnicas de transformação de dados, de forma a facilitar o manejamento nos algoritmos.

Assim, houve necessidade de implementação de técnicas de para extração de informação a partir de texto. Desta forma, para se determinar qual a companhia aérea que operou o voo, recorreu-se à técnica de extração de *callsign* a partir do número completo de voo (os voos têm a seguinte sequência de informação: os dois caracteres iniciais correspondem ao código IATA da companhia aérea que o opera, a seguinte numeração corresponde à frota deste modelo na companhia, dia da semana e o número de estações que operará continuamente com esse mesmo número de voo). Assim, esta técnica de Linguagem Natural permitiu a eliminação de informação que consta nos dados, eliminando caracteres com informação que não é essencial para posterior utilização do conjunto de dados, (por exemplo do voo “XX1234”, foi extraído apenas “XX” e eliminado o “1234” de forma a identificar a companhia aérea que operou o voo, sem o ruído da informação dos restantes detalhes do número do voo).

Com recurso a técnicas de limpeza de dados, também se procedeu à eliminação de caracteres (tais como “;.*#”), considerados como ruído, de forma a não causar interferência na leitura das observações que constam no conjunto de dados resultante da extração de informação do sistema para o formato *.csv* e *.xlsx* do *Excel*, caso contrário poderia perturbar o desempenho dos algoritmos de aprendizagem automática ou mesmo tornar inviável o seu carregamento nos mesmos.

A primeira transformação foi das variáveis “data” para tipo temporal, em simultâneo com a separação de dias, meses, anos, dias e dias da semana para futuro estudo na exploração de dados. Posteriormente, por forma a simplificar cálculos, foi também necessária a aplicação de transformação de dados referentes à variável horas e minutos, onde anteriormente tinha sido feita a transformação para variável temporal, e aplicação de algumas operações simples aritméticas por forma a conseguir manusear os dados mais simplesmente e extrair conclusões. Assim, procedeu-se à separação das colunas onde constam as horas e os minutos de atrasos de partidas, chegadas, de horas voadas e de horas entre blocos, e efectuou-se a transformação de horas, para minutos, e a sua soma por forma a obter melhor grafismo (na parte de visualização) dos diferenciais de atrasos com a diferença entre horas estimadas e previstas.

Com recurso a técnicas de *Text Mining*, procedeu-se ainda à aplicação de anonimização de dados, portanto desassociando a título definitivo o código IATA do *callsign* das companhias aéreas e transformando cada *callsign* num número. tendo em vista a confidencialidade de informação dos clientes de acordo com nova lei RGPD (Regulamento Geral de Proteção de Dados) em ambos os conjuntos de dados. O *Text Mining* é um campo multidisciplinar, que faz extração de conhecimentos úteis a partir de texto corrente, ou seja, dados não estruturados ou semi-estruturados, e faz uso de um conjunto de métodos para organizar, encontrar e descobrir informação em bases textuais.

A normalização dos dados, foi efectuada usando a técnica z-*scores*. Esta normalização tem em consideração a média e o desvio padrão de cada variável, e produz um valor que permite distribuir de forma central os dados. O valor normalizado indica a distância entre a média e o valor original em termos de desvios-padrão, conforme Equação 1.

$$z = \frac{(x-\mu)}{\sigma} \quad \textbf{(1)}$$

Equação 1 - Fórmula z-score usada

Assim, e para que as escalas das variáveis não tenham impactos diferentes nos algoritmos, procedeu-se à transformação das variáveis categóricas em quantitativas e respectivo mapeamento e das quantitativas em *z-scores*, para homogeneizar as variáveis. O processamento foi feito com recurso ao *package* do *Python* que processa a aplicação desta fórmula em todas as colunas dos conjuntos de dados com a chamada deste método.

Uma vez que os resultados iniciais de análise de avaliação e desempenho do algoritmo modelo de base – construído com a técnica de Regressão Logística Multinomial [Secção 4.1] demonstraram claramente que o desiquilibrio das classes estaria a prejudicar a sua predição, uma vez que determinadas classes se encontram na amostra em número reduzido, foram aplicadas técnicas de tratamento de dados para nova construção de modelos de previsão, com o intuito de melhoria de resultados.

- Foi aplicada a técnica SMOTE, que faz a produção sintética de exemplos de classes minoritárias. Foi feita sempre a comparação da aplicação com a não aplicação deste método nos algoritmos na modelação e respectiva avaliação de desempenho de forma a equacionar a sua aplicação em qualquer uma das questões de investigação.

Foi também feita uma abordagem da ponderação e importância das variáveis dependentes com vários métodos, de forma a analisar se seria viável a redução de dimensão de variáveis do conjunto de dados:

- Pelo método de Lasso, é um tipo de regressão linear, que recorre à redução de dimensão das variáveis, incentivando os modelos a tornarem-se mais simples e esparsos (isto é, modelos com menos parâmetros). Esse tipo específico de regressão é bem adequado para modelos que apresentam altos níveis de muticolinearidade. Com este método irá fazer-se a verificação se existirá mais valia com a seleção de variáveis ou eliminação de parâmetros. Na Figura 3, podemos verificar graficamente a importância atribuída das variáveis por este método, tais como: o aeroporto de chegada, o aeroporto de partida, o diferencial dos minutos totais previstos entre o tempo previsto de partida e os reais, os minutos totais de horas entre blocos, passageiros e a respectiva companhia aérea.

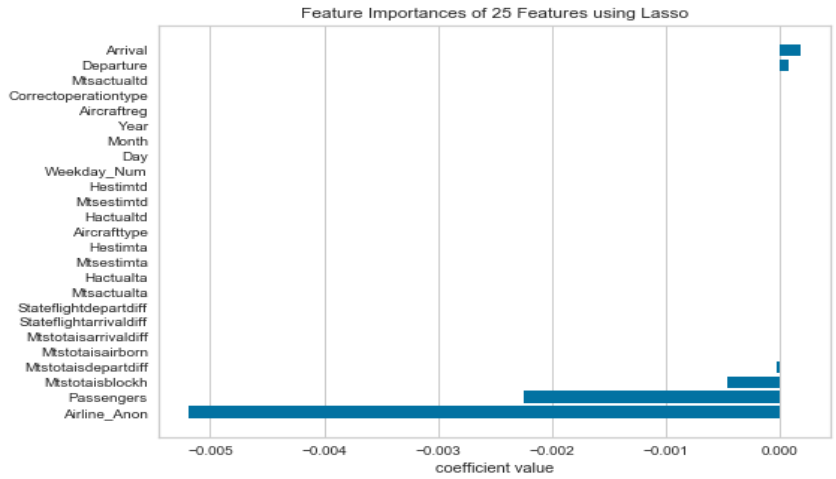

Figura 3 - Método de Coeficiente de Lasso

- Pelo método RFE (Recursive Feature Elimination), efetua-se a seleção de variáveis que se ajusta mais ao modelo e remove a variável (ou variáveis) mais fracas até que o número especificado de variáveis seja atingido. Houve experimentação com 5, 15 e 20 atributos de acordo com a classificação atribuída, mas os resultados no modelo base não demonstraram melhorias com nenhuma destas selecções.

- Pela abordagem com o método *Univariate Selection*, foram obtidos scores pouco significativos em ambas as questões e existe também evidência estatística que demontra a rejeição de relação entre as variáveis categóricas, pelo que corrobora que não existe necessidade de aplicação também de redução de variáveis.

De todas as abordagens mencionadas, para seleção de variáveis, com base nos resultados obtidos, as métricas de avaliação do desempenho revelaram-se baixas, em especial a sensibilidade, pelo que não será efetivamente uma boa opção, a melhor classificação foi obtida com o método SMOTE, que revelou melhor desempenho na aplicação nos algoritmos.

Dada a existência de um factor temporal presente nos dados, embora não sejam observações exactamente sequenciais com o mesmo intervalo de tempo, foi feita também abordagem de predição com séries temporais, na questão Q3, sendo necessária a colocação da “data” como *índex*, para poder fazer recurso destes modelos.

3.4. Análise Exploratória de Dados

Nesta secção será realizada a análise exploratória dos dados extraídos para esta dissertação dos diferentes módulos aplicacionais, os dados referentes às operações e referente às de tripulações (GD).

3.4.1. Dados relativos à operacionalidade de voos

- ***Análise Descritiva***

Na análise exploratória da amostra referente análise de voos, com um total de 45929, verifica-se a existência de desequilíbrio. As operações ACMI são encontradas em grande maioria em relação as restantes operações, distribuindo-se todas as operações da seguinte forma: ACMI - 34819, Charter - 6682, Position - 3769, Private - 390, Train - 245 e Test - 24.

Com recurso a uma função do *Python*, onde se pode verificar a média e o desvio padrão, mínimos e máximos e percentis, pode-se concluir que é mais comum não haver atrasos na operacionalidade dos voos. Na Tabela 3 - Análise médias, desvio padrão, minimos, máximos e percentis de variáveis temporais, observa-se as médias, o desvio padrão, mínimos, máximos, percentis 25%, 50% e 75% referentes aos atrasos e outras estatísticas importantes de voos operados.

	Passageiros	**Minutos atraso partidas**	**Minutos atraso chegadas**	**Minutos totais Voados**	**Minutos totais entre blocos**
count	45929	45929	45929	45929	45929
mean	117	44,80	50,33	305,28	332,79
std	112	82,87	83,53	193,75	195,996
min	0	0	0	0	7
25%	0	6	10	128	154
50%	102	20	27	293	320
75%	217	51	58	441	470
max	467	2610	2595	1025	1039

Tabela 3 - Análise médias, desvio padrão, minimos, máximos e percentis de variáveis temporais

Devendo-se salvaguardar que referente aos registos de passageiros estes registos estão incompletos, dado que nem todas as operadoras fornecem este tipo de informação, ou não

são correctamente actualizados no sistema, estes valores foram excluídos do conjunto de dados.

A aeronave que tem mais voos registados, é a modelo A340 (343) e A332(332), o que pode ser explicado por uma maior procura dadas as suas características, conforme Figura 4.

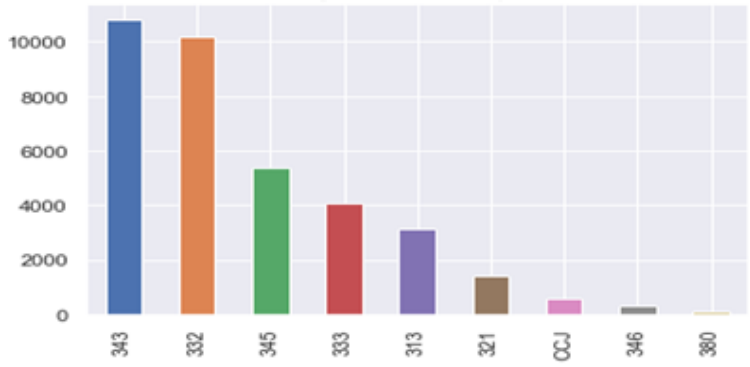

Figura 4- Histograma de voos por tipo de aeronave e modelo

Em relação à operacionalidade, portanto ao número de voos efetuados, do modelo de aeronave e tipologia operacional, observa-se que, à exceção do modelo A321, que opera maioritariamente para um determinado cliente, e o jacto, que opera maioritariamente voos privados, as restantes aeronaves desempenham voos ACMI em maior quantidade, conforme Figura 5. No entanto deverá ser referido que o modelo de aeronave A321 está presente na companhia apenas desde 2014 até ao momento e está agora também alocado a um cliente em especifico. A aeronave A380 também tem menos actividade por se encontrar à pouco tempo na empresa.

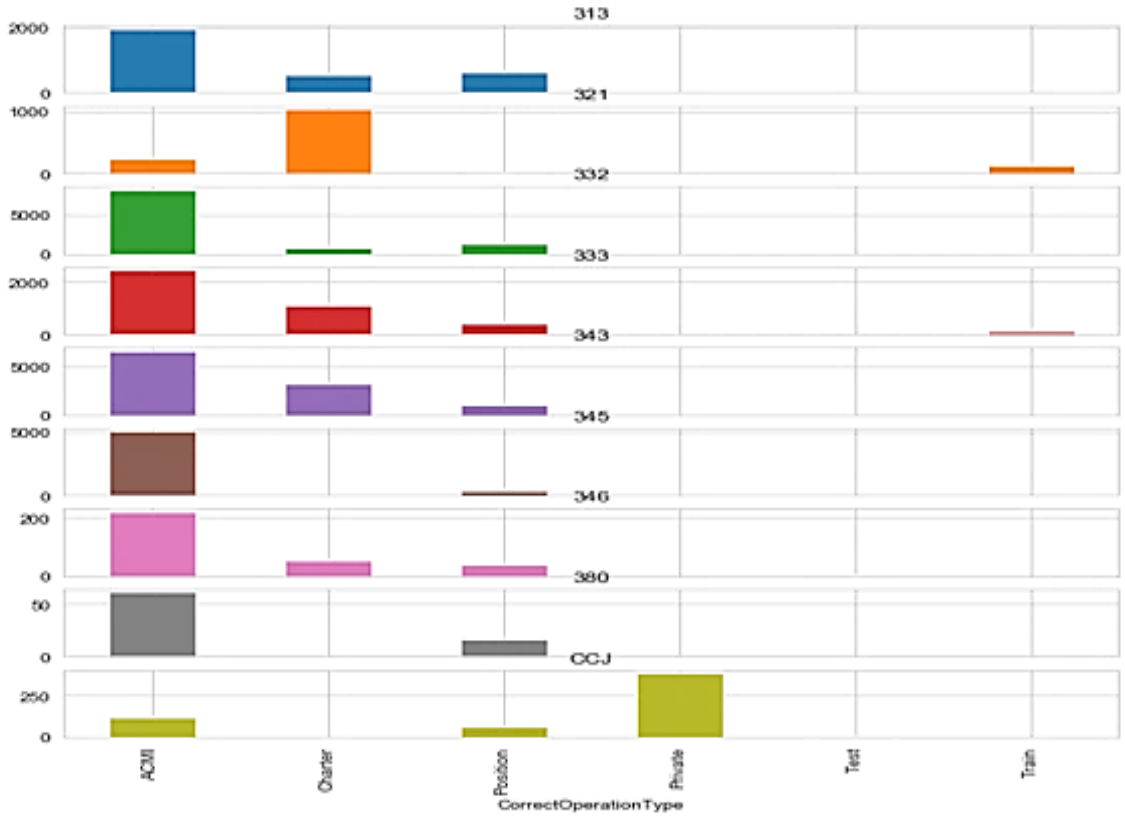

Figura 5 - Modelo de aeronave e Tipologia Operacional

Fazendo-se uma análise anual, observável na Figura 6 verifica-se que, a partir de meados do ano de 2012, houve acréscimo no número de voos ACMI, sendo que no ano e 2015 houve um máximo deste tipo operacional e mantendo-se relativamente constante à posteriori.

No ano de 2011 a 2012, verifica-se maior fluxo de voos charter, havendo posteriormente uma certa diminuição deste tipo operacional. Surge novamente maior actividade deste tipo operacional em finais de 2019.

As tipologias Teste e Treino são relativamente constantes e sem grande expressão. Os voos da tipologia privado só existe registo desde a data em que o jacto entrou na empresa e a sua actividade é mais esparsa e reduzida.

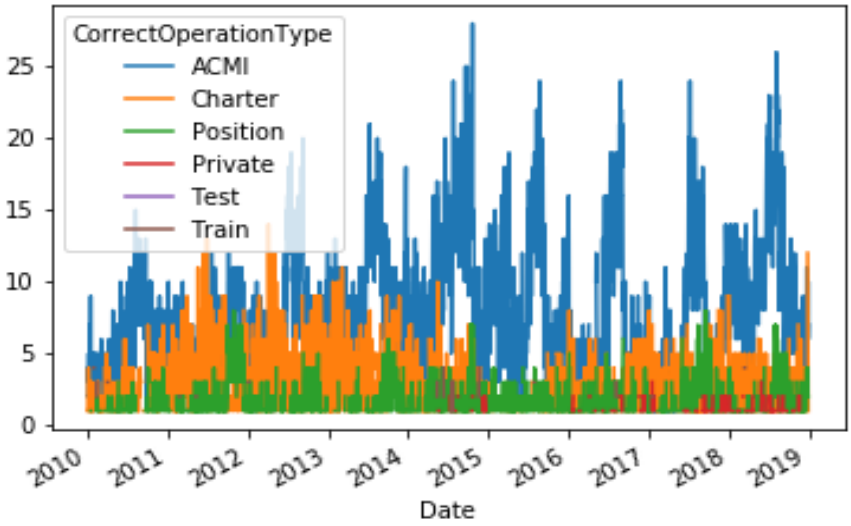

Figura 6 - Ano e Tipologia operacional

Elaborando um gráfico do tipo *boxplot*, Figura 7, entre as aeronaves e os meses do ano, pode-se observar que os aviões de maior porte e o jacto têm a mediana nos meses de Julho e Agosto, com exceção do A346, que esteve relativamente pouco tempo presente na companhia. Já as aeronaves *single aisle* (referente às aeronaves de corredor único) têm mediana localizada nos meses de Agosto e a aeronave A380, a de maior dimensão de todas, ainda não têm voos suficientes para localizar a sua mediana. Na Figura 6, poderá observar-se a caixa-de-bigodes entre o tipo de aeronave no eixo dos *yy* e os meses no eixo dos *xx*.

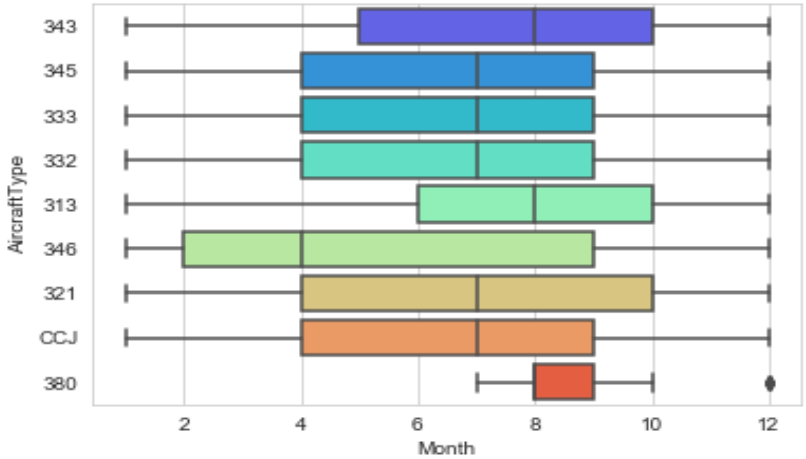

Figura 7 - Caixa-de-bigodes Tipo aeronave e Mês

Realizando-se uma análise de *boxplot*, conforme Figura 8, pode-se verificar que o primeiro e terceiro quartis são mais alargados nas operações charter coincidindo com os meses de Abril e Outubro, enquanto que as operações ACMI têm os seus percentis em Maio e Setembro, que corrobora com os meses de maior procura de operacional. Os restantes tipos operacionais são mais distribuídos por serem realizados consoante as necessidades da empresa e muitos destes posicionamentos são incluídos em contratos ACMI ou charter.

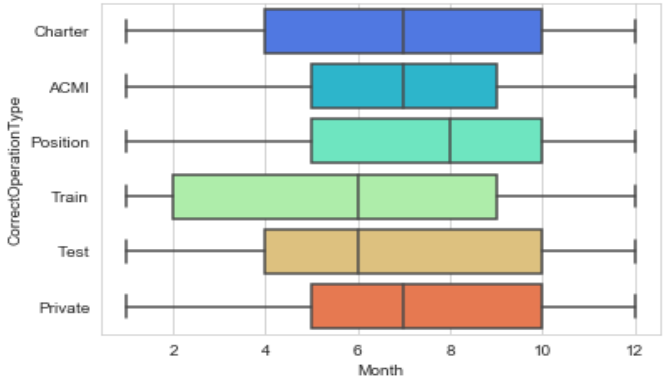

Figura 8 - Caixa-de-bigodes entre Tipologia Operacional e mês

Fazendo o mesmo tipo de análise semanal e a tipologia operacional, existe uma tendência de medianas às terça-feira e quarta-feira, e percentis situarem-se de segunda-feira a sexta-feira, o que é consequência de existência de menos voos ao fim de semana.

A análise de *boxplot* entre a tipologia operacional e os minutos de atraso nas chegadas, conforme Figura 9, verifica-se que a tendência da maioria das operações é de cerca de zero minutos, sendo os seus quartis também próximos desse valor, com exceção das operações de testes, dado muito provavelmente porque esta tipologia operacional está dependente de aval do departamento de manutenção e tem muita volatilidade, verifica-se ainda a existência de prováveis *outliers* (observações discrepantes) por existirem determinadas situações alheias ou inesperadas que provocam atrasos, por vezes bastante elevados.

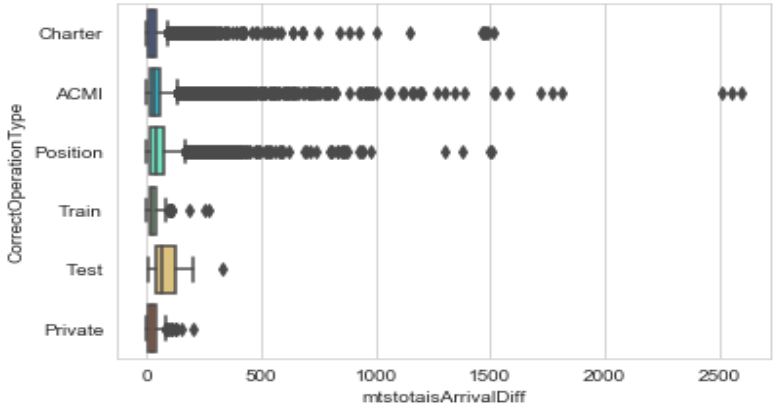

Figura 9 - Caixa-de-bigodes entre números de atraso de chegadas (mts) e Tipologia operacional

Ainda em relação à análise de caixa-de-bigodes, na Figura 10, entre os minutos de *Airborn* e o tipo de operação, pode-se verificar que a mediana é maior em ACMI do que voos charter, derivado ao facto de a empresa operar bastantes mais voos ACMI do que charter, havendo, no entanto, quartis mais aproximados nestas duas operações. Dado que os voos de posicionamento têm propensão a serem curtos, têm mediana inferior, obviamente apresentando alguns *outliers*, dado o facto de certos posicionamentos serem para aeroportos mais distantes, por ser necessário ocasionalmente posicionamentos mais longos. Voos de testes, treino e privado são no geral mais curtos, sendo que os de teste são sempre planeados para duração mais abreviada, enquanto os privados e de treino apresentam por vezes *outliers*, em algumas situações pontuais.

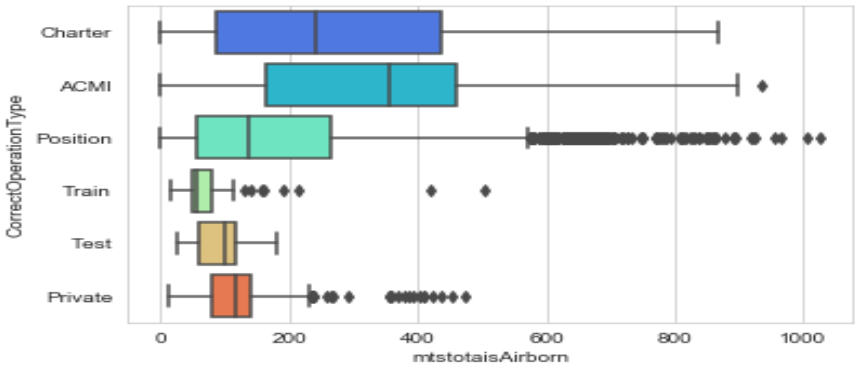

Figura 10 - Caixa-de-bigodes minutos Airborn e Tipologia operacional

Os voos têm mediana entre as 10 e as 15 horas UTC (Tempo Universal Coordenado), e existem menor número de voos antes das 6 horas da manhã e depois das 18 horas.

Fazendo uma análise de gráfico com agrupamento mensal entre os anos de 2010 e 2018 na Figura 11, pode-se verificar um intervalo de pico de actividade entre os meses de Junho a Outubro, havendo descréscimo no mês de Novembro e depois novamente maior procura em Dezembro e ligeiro aumento em Março.

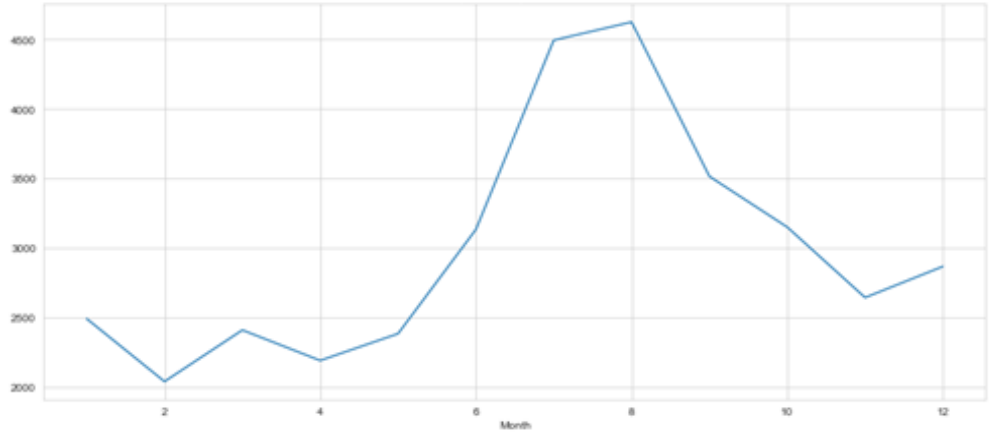

Figura 11 - Gráfico média mensal entre os anos 2010 e 2018

Observando os dados a nível mensal, existe maior actividade no início dos meses e depois um decréscimo, geralmente, em média, uma maior procura a partir do dia 18 até ao final do mês, havendo, no entanto, uma redução de procura a partir dia 21.

Pode-se verificar graficamente na Figura 12, que em média, existem mais operações *Charter* operadas às terças feiras, enquanto que as operações ACMI têm maior ocorrência aos domingos. As restantes operações apresentam uma procura idêntica ao longo da semana.

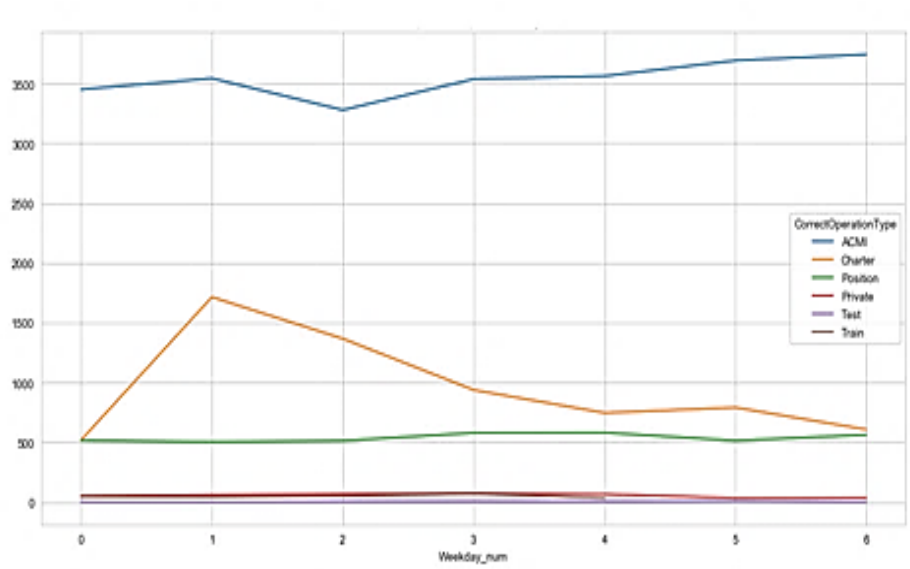

Figura 12 - Gráfico de operações por dia da semana

Investigando a frequência da tipologia de voos por anos, observou-se que os voos ACMI ocorreram maioritariamente em 2014 (meses de Julho e Agosto, conforme análise mensal), os voos charter em 2011 e 2012, (nos meses de Junho e Outubro), de posição também em 2011, (nos meses de Outubro), Privados em 2017, (nos meses de Maio e Junho) e não houve ocorrências deste tipo de voos antes de 2013, de Teste em 2015, (nos meses de Maio e Outubro) e de Treino em 2014, (no mês de Fevereiro), (Figura 13).

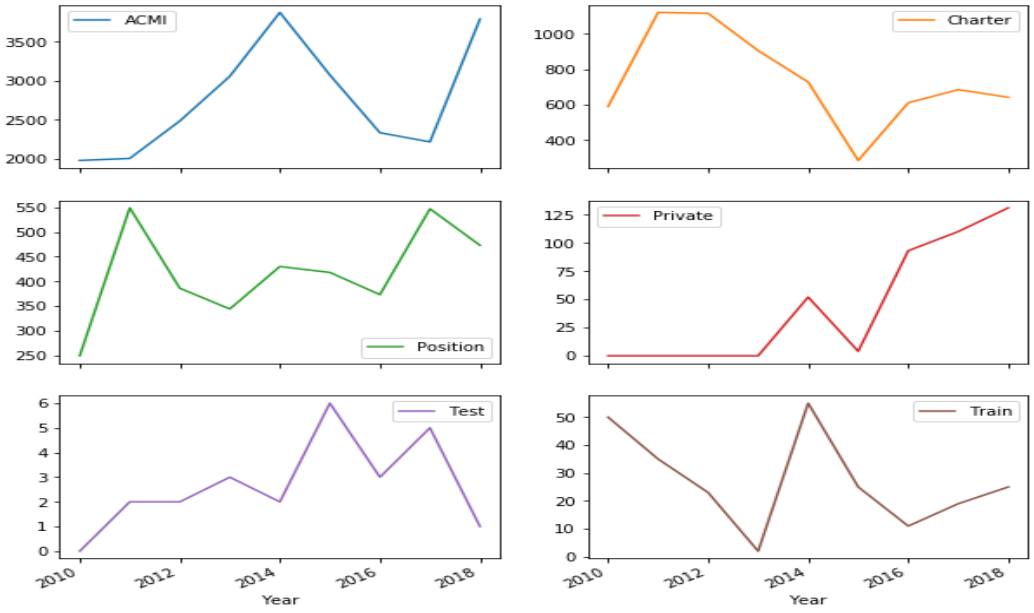

Figura 13 - Ano e tipologia operacional

Referente à frequência de passageiros nas operações charter, verifica-se maior afluência nos meses de Abril, Junho, Outubro, Novembro e Dezembro. Relativamente às operações ACMI, embora uma boa parcialidade do registo de passageiros esteja incompleto, pelo que não poderão ser utilizados para outras quaisquer inferências, existem indícios que os meses de Junho, Julho, Agosto e Setembro são os meses com mais fluxo de passageiros transportados.

Fazendo a análise de atrasos por tipologia operacional, conforme Figura 14, nos voos charter, existe um pico de atrasos para esta operação em 2011 e 2012, coincidindo também com o pico desta tipologia operacional, nos voos de posição denota-se um relativa estabilidade com ligeiras oscilações de atrasos ao longo dos anos, com exceção de 2017 que houve aumento. Para os voos privados, um ligeiro aumento; e uma certa estabilidade nos voos de treino e teste. Para as operações ACMI o máximo de atrasos ocorreu em 2014, tendo o ano de 2018 quase atingido o mesmo nível. Denota-se que as operadoras com mais atrasos são as que têm também mais número de voos realizados.

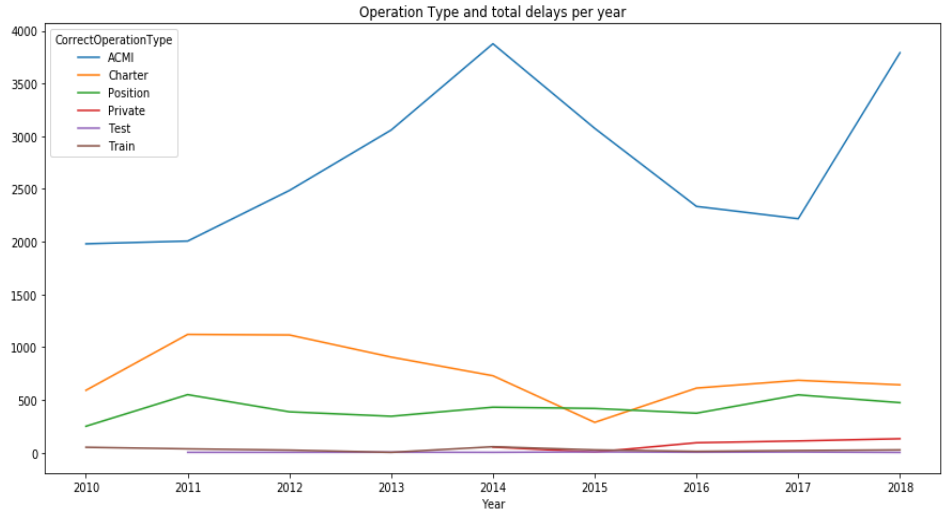

Figura 14 - Gráfico de tipologia de operações por ano

Analisando também as frequências das estações com maior operabilidade, um dos aeroportos que a companhia aérea que maior número de voos operou é BRU (código IATA para Bruxelas), este é um dos pontos de manutenção e centrabilidade da Europa, sabe-se que se trata de um bom ponto estratégico para estacionamento de aeronaves e posicionamento para acessibilidade rápida e possibilidade de custo reduzido de estacionamento, conforme contrato de acordo em vigor.

Houve um total de 117 operadores diferentes registados entre os anos de 2010 e 2018, onde se verifica em primeiro lugar, com maior número de voos, um operador de longo período de contrato e de seguida voos que operaram sob o *callsign* da companhia, portanto os voos charter, de posicionamento e de teste. Os maiores atrasos verificaram-se consequentemente nestes dois *callsigns*, o que poderá estar associado ao facto de serem os mais comuns. Referente ainda aos operadores, denotou-se que têm sempre mais operacionalidade em aeroportos do seu país de registo do COA (*Certificado de Aeronavegabilidade*).

Observando o comportamento das aeronaves A310, na Figura 15, denota-se maior actividade nos meses de maior actividade da companhia, de Junho a Setembro, contudo entre 2010 e 2012 verifica-se uma efetiva maior procura desta aeronave entre os meses de Setembro e Dezembro. Existe ausência de registo de voos entre os meses de Maio e de Julho em 2013, devido a período de manutenção.

Figura 15 - Gráfico para a aeronave A310 por mês/ano

A aeronave A321, estabelecido contrato *leasing* em meados de Março de 2013, tem um comportamento em que o padrão não está de acordo com o pico de actividade da companhia, provavelmente dado o facto de se encontrar dedicado a um cliente específico em que o seu pico de operacionalidade decorre entre os meses de Abril a Agosto e depois no mês de Outubro, conforme Figura 16.

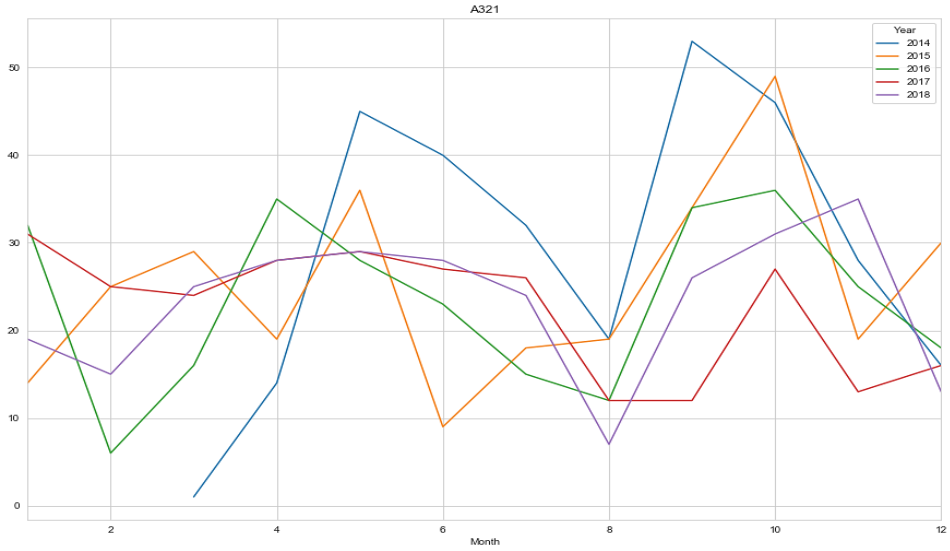

Figura 16 – Gráfico para a aeronave A321 por mês/ano.

As aeronaves A332, A333, A343 têm um comportamento padrão de procura, salvo raras exceções, com o maior pico de actividade da companhia, sendo de Junho a Setembro onde existe maior registo de voos.

As aeronaves A345, observável Figura 17, tiveram uma grande actividade contínua nos primeiros anos de existência na companhia de 2010 a 2013, já que estiveram vinculadas a um cliente específico, à posteriori a sua procura enquadra-se também no padrão tradicional.

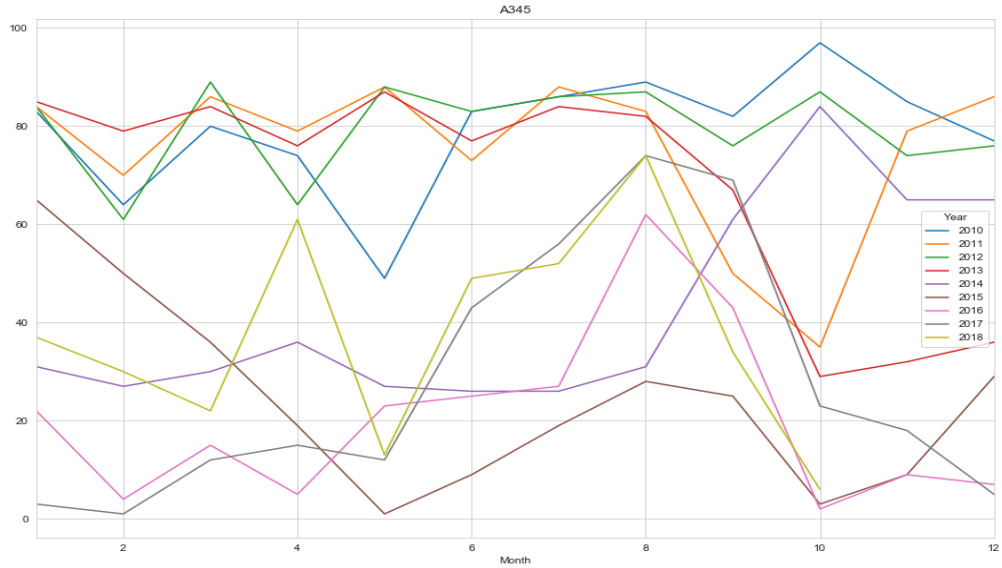

Figura 17 - Gráfico para a aeronave A345 por mês/ano.

A aeronave A346 por se ter encontrado pouco tempo na companhia, não é possível analisar um padrão de comportamento de procura, apresentando apenas actividade no final do ano de 2013 e princípio de 2014.

A aeronave A380 foi adquirida recentemente, e ainda não tem procura mensurável.

O jacto por se encontrar maioritariamente ao serviço da companhia, tem um padrão contrário, a sua actividade é acentuada no início do ano.

- **Análise de Correlações**

Pela matriz de correlações, (Figura 18), denota-se uma correlação positiva fraca entre a hora estimada de partida e a hora atual de partida do voo, podendo ser indício que os atrasos existentes, não são significativos.

Verifica-se também que existe uma relação fraca (0.3) entre as variáveis de tempo entre blocos (tempo que o avião se encontra no chão com calços) e o tempo efetivo de voo (tempo *Airborn*), o que se justifica por norma da preparação necessária, ainda no solo, da aeronave, que é proporcional ao tempo de voo que irá realizar.

O tempo entre blocos, tipo de aeronave, tipologia operacional e a operadora ou cliente também apresentam uma correlação mais forte, onde se pode verificar que o mesmo cliente tem tendência a escolher o mesmo tipo de avião. Isto provavelmente derivado ao facto de a necessidade ser idêntica em períodos anteriores, sempre que efetuam a fretagem de aviões. A relação com o registo será apenas uma consequência, dada a sua associabilidade ao modelo da aeronave.

Entre a companhia e o tipo de operação, também existe alguma relação, negativa, (0.3), o que seria expectável, já que a tipologia charter opera sempre com o *callsign* da própria companhia e para a tipologia ACMI opera sempre com o seu próprio *callsign*.

No entanto, no geral, as correlações não são mensuráveis, positivas ou negativas, estando situadas entre correlações fracas a desprezíveis.

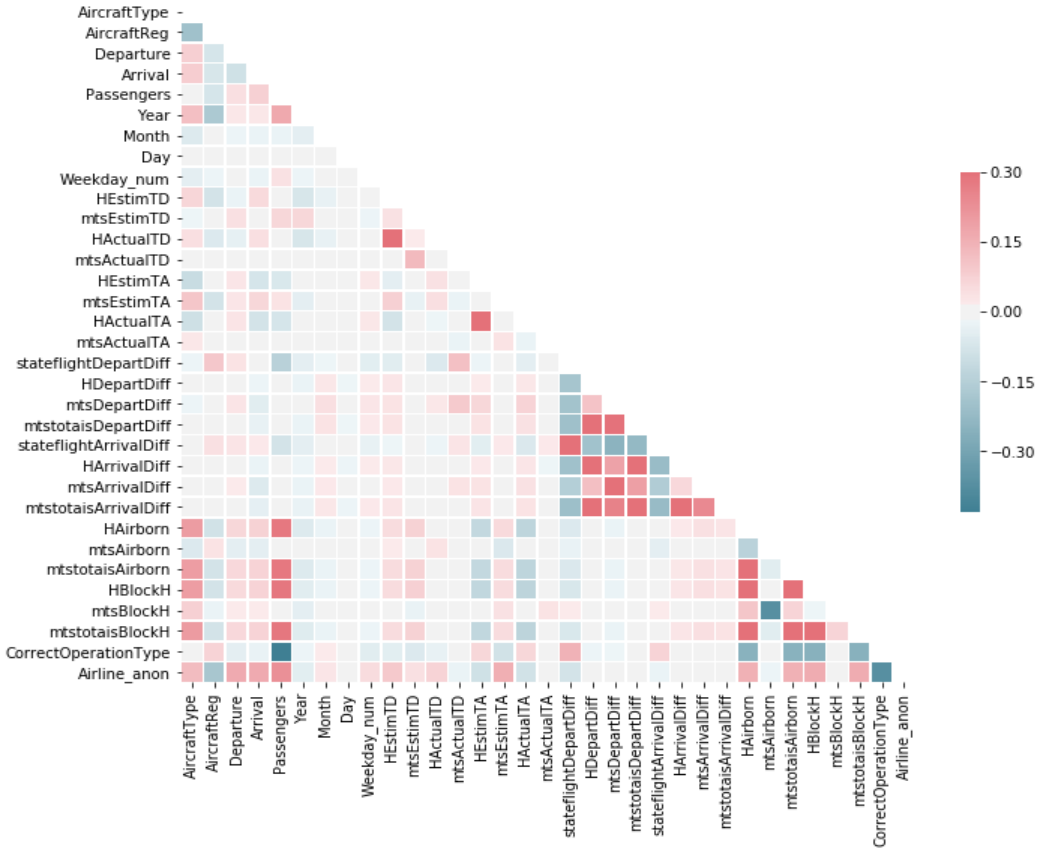

Figura 18 - Matriz correlações

- **Análise de Componentes Principais**

Nesta análise exploratória, foi também realizada uma Análise de Componentes Principais (ACP) na amostra de voos, a fim de confirmar se traria vantagens a aplicação de redução de dimensionalidade da amostra. A ACP é um método exploratório que auxilia na elaboração de hipóteses gerais a partir dos dados extraídos, tem a capacidade de separar a informação importante da informação redundante e aleatória. Na análise exploratória de dados estatísticos multivariados, a análise de componentes principais é um método também tradicionalmente utilizado com o objectivo de projetar dados com n-dimensões num espaço de menor dimensão. Após a aplicação da técnica, 17 variáveis explicavam cerca de 95.5% da variância, denotado na Figura 19, sendo que serão as primeiras 17 componentes.

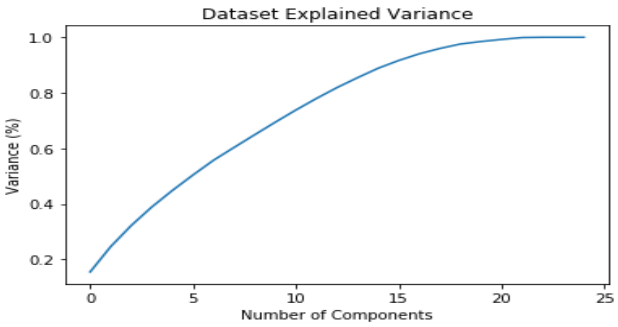

Figura 19 - Número de Componente Principais e a variância explicada

A ACP nem sempre consegue ser aplicada, nomeadamente quando estamos perante variáveis pouco correlacionadas ou categóricas, as componentes principais são as próprias variáveis originais (Regazzi, 2000). Assim dado o facto de a correlação ser bastante fraca ou quase inexistente entre as variáveis e estarmos na presença de alguns *outliers*, a ACP não deve ser aplicada.

3.4.2. Dados relativos à gestão de tripulações

- **Análise descritiva**

Relativamente aos dados provenientes para a terceira questão Q3, num total de 43047 observações, para predição da tripulação ideal para o bom funcionamento da empresa, advêm, como mencionado das General Declaration de actividade da companhia, e as variáveis que constam nos dados têm o comportamento consoante a Figura 20 pode-se observar o comportamento de cada uma das variáveis que compõem o conjunto de dados relativos às tripulações.

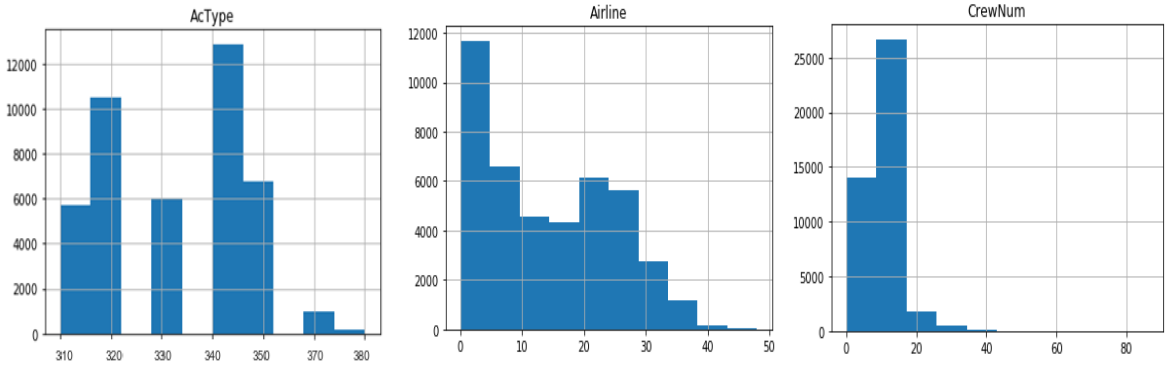

Figura 20 – Histogramas relativos às variáveis

A quantidade de tripulação mais comum refere-se à tripulação mínima para operar um voo correspondente ao modelo de aeronave em maior número na empresa. Um maior número de tripulação a bordo num voo, poderá corresponder a tripulação reforçada para operar outro voo da operação, ou, posicionada para ir operar outros voos, ou ainda, de regresso à sua base. Na Figura 21 pode-se observar a variabilidade do número de tripulantes para cada avião em cada ano, sendo, ainda, possível verificar a evolução de contratação. Notam-se picos, com comportamento *outlier*, em 2012, 2016 e 2017. Nota-se ainda que os anos de 2016 e 2017 houve maior variabilidade nos quartis, o que provavelmente se deverá à implementação da nova directiva de optimização de custos e a tipologia operacional, posicionando mais vezes a tripulação nos próprios voos operados. A mediana da quantidade de tripulação para a maioria dos anos é 10.

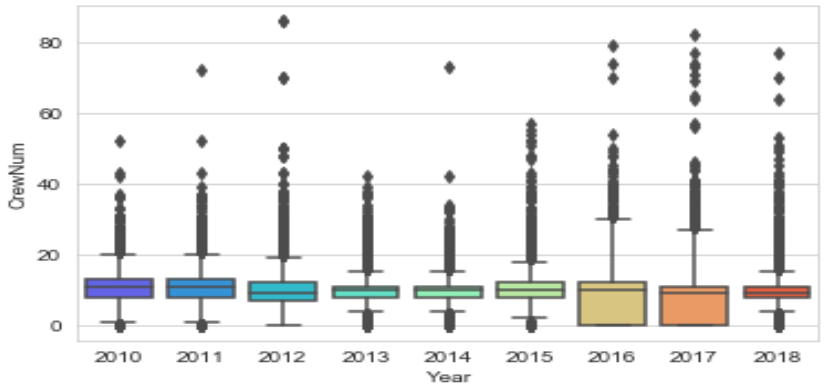

Figura 21 - Caixa-de-bigodes Número de tripulação e ano

Na Figura 22, verifica-se a existência de sazonalidade, entre os meses de Maio e Outubro, e novo pico, embora com menos quantidade de operações, para o mês de Dezembro para determinados modelos de aeronaves. Nos meses iniciais do ano, observa-se uma certa estabilização do número de tripulações posicionadas em voos e um aumento no período de pico operacional. A aeronave A380 ainda não surge no gráfico por não existirem registos suficientes de voos.

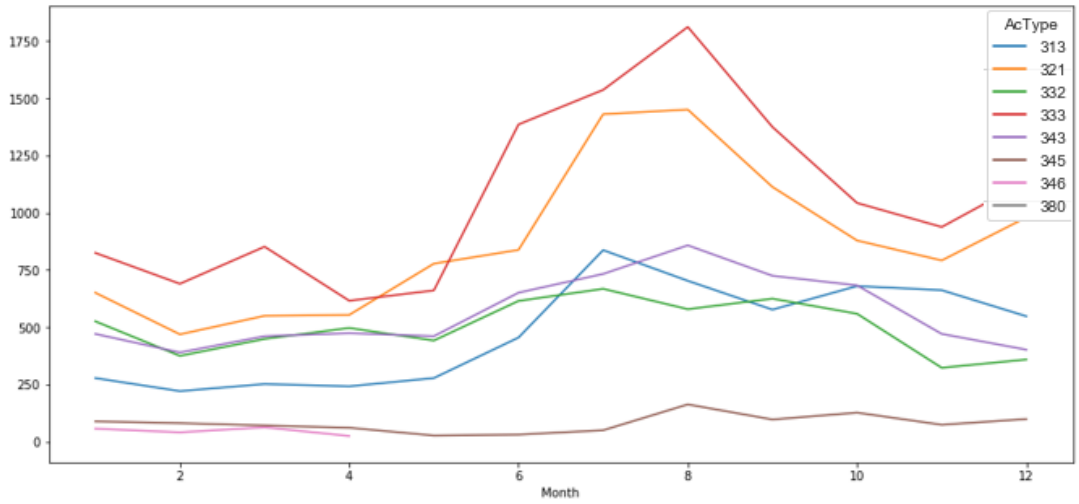

Figura 22 - Gráfico de Tipo de aeronave/mês e o número de tripulação a bordo das aeronaves

A Figura 23 mostra que os meses de Junho e de Outubro detêm mais variabilidade de número de tripulantes posicionados, o que corresponde ao início e ao fim do pico operacional anual. Deverá salientar-se a existência de menos variabilidade do número de tripulantes posicionados em pico de actividade, já que a grande maioria da tripulação estará operacionalmente atribuída a uma operação em específico, havendo ocasionalmente mobilidade entre operações.

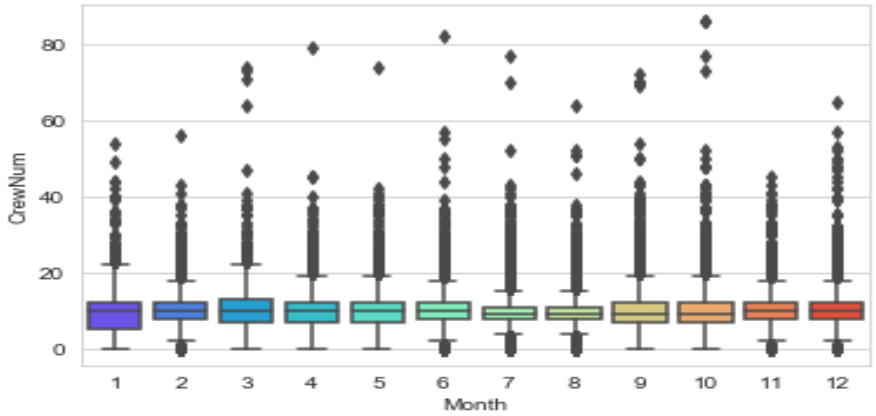

Figura 23 - Caixa-de-bigodes por tripulação e mês

- **Análise de correlações**

Relativamente à correlação dos dados GD, faz-se notar a correlação entre número de tripulantes e o tipo de aeronave, e entre a companhia aérea (Figura 24). Provavelmente devido à especificidade de formação dos tripulantes para cada aeronave, bem como dado o acordo com, ou sem, inclusão de tripulação ser uma opção contratual da companhia aérea, e por norma, na repetição de fretagem da aeronave, irá estabelecer os contratos em que as claúsulas contêm as mesmas condições. O ano e companhia aérea, também apresentam alguma relação, provavelmente dado o facto de algumas companhias aéreas regulares efectuam contratações de aeronaves exporadicamente e excecionalmente, tendo algumas companhias aéreas, a tendência a culmatar as suas necessidades no ano seguinte optimizando os seus recursos. Contudo, todas estas relações são valores baixos, encontram-se entre valores de 0.24, positivas, e 0.08, negativas, pelo que são consideradas bastante fracas.

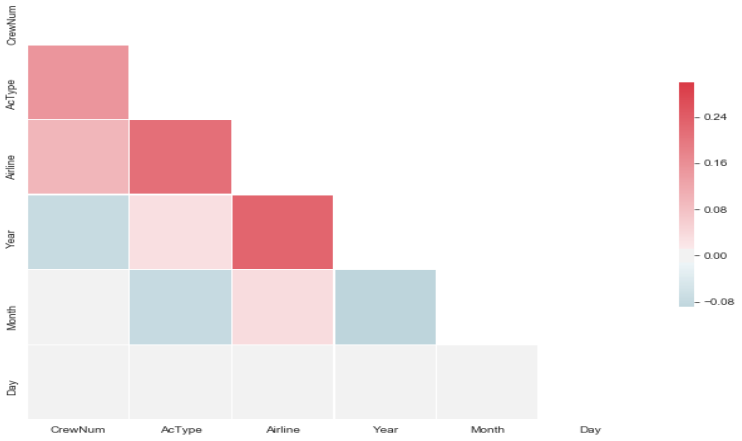

Figura 24 - Matriz de Correlações (GD)

Na Figura 25, pode observa-se os dados referente aos voos desempenhados pelas tripulações ao longo dos anos de 2010 a 2019 (incluíndo a média e o desvio padrão) de acordo com factor temporal. Observa-se um pico de operações com maior número de tripulações em finais de 2012 e em meados de 2017. A partir de meados de 2015 nota-se um aumento do número de operações, havendo mais períodos de picos de actividade operacional.

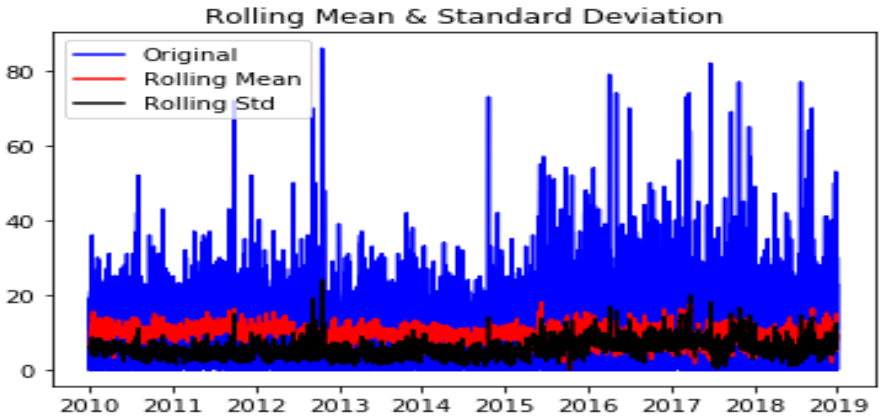

Figura 25 - Séries temporais, (com média e desvio padrão)

Capítulo 4 – Modelação e avaliação dos resultados

4.1. Modelação

A modelação irá produzir um modelo a partir dos dados preparados e estudados nos capítulos anteriores. Esta fase, tem como objectivo a predição de um atributo ou variável alvo (Y) e tendo como base outras variáveis ou atributos (X), previamente conhecidas. Se a variável de resposta do modelo for categórica, então estamos perante um problema de classificação, em que cada categoria é designada como valor de classe (Fonseca, 2006). Se a variável for contínua, estamos perante um problema de regressão. O processo de modelação tem diversos passos, tanto para os problemas de regressão como de classificação, de forma a poder encontrar o modelo mais adequado, isto é, aquele que apresenta melhor desempenho.

Para poder ser utilizado como um ponto de partida de comparação das métricas e observação do desempenho com outros algortimos, será necessário construir um modelo base. A técnica do modelo base representará a forma mais simples de obtenção de uma predição e deverá ser visto como o ponto de partida para obtenção de melhoria de resultados (Howarth, Jaokar & Mutlu, 2016). Por norma são utilizadas essas predições para comparação de base nas métricas de desempenho, com o objectivo final, da obtenção de maior desempenho relativamente a qualquer modelo base em que foram utilizados os mesmos dados. O modelo base é um modelo simples e consegue providenciar resultados de referência para avaliar o desempenho dos modelos mais complexos, no entanto, muitas vezes os modelos de base correspondem ou superam os modelos mais complexos. Estes modelos são mais rápidos de treinar, mais acessíveis de estudar e auxiliam na compreensão dos dados.

De seguida, encontra-se uma descrição dos modelos e as métricas de avaliação utilizados para construção dos modelos e avaliar o desempenho dos dados explorados no capítulo anterior, com o intuito de dar resposta aos objectivos propostos para as questões Q1, Q2 e Q3 de forma a perceber qual o modelo mais adequado. Sendo que, para as questões Q1 e Q2 foram aplicadas técnicas de Classificação e para responder à questão Q3, ténicas de regressão.

- ***Avaliação nos problemas de Classificação***

Nos problemas de classificação a avaliação do modelo, passa não apenas pela análise das métricas obtidas, como pode, e muitas vezes, deve ser complementada a análise de performance com a verificação das respectivas contagens correctamente predictas. A matriz de confusão simplifica a visualização da proporção de classificações corretas e do número de classificações preditas para cada classe de um determinado conjunto de exemplos, segundo o classificador em análise, sendo que, na diagonal principal da matriz pode-se observar a proporção de predições correctas. Assim, torna-se uma ferramenta útil para verificar a qualidade do classificador na identificação de exemplos das diferentes categorias (Han e Kamber, 2006). Na Tabela 4, observa-se a matriz de confusão para duas classes, contudo poderá ser extrapolado para várias classes.

		Predição de Classes	
		Positivo	Negativo
Classe verdadeira	Positivo	**VP**	**FN**
	Negativo	**FP**	**VN**

Tabela 4 - Tabela de predição de duas classes

- VP é o número de previsões correctas para uma predição positiva;

- FN é o número de previsões incorrectas para uma predição positiva;

- FP é o número de previsões incorrectas para uma predição negativa;

- VN é o número de previsões correctas para uma predição negativa;

Com auxílio da matriz de confusão obtêm-se as seguintes métricas que apuram a qualidade do modelo predito, que se traduzem com as seguintes fórmulas, sendo que a escolha da métrica que se dará mais realce dependerá do caso em estudo:

- **Sensibilidade** (*Recall*) – VP/ (VP+FN)
- **Precisão** ***(****Precision*) – VP/ (VP + FP)
- **Especificidade** – VN / (VN + FP)
- **Acurácia/Acerto** (*Accuracy*) – (TP + TN) / (TP + TN + FP + FN)

- ***Avaliação nos problemas de Regressão***

Tal como nos problemas de classificação, existe necessidade de existência de uma métrica de avaliação de desempenho que compare os valores previstos com os valores reais e que quantifique as diferenças entre eles. Existem diversas métricas disponíveis, mas neste projecto recorreu-se ao RMSE (*root mean square error*). As unidades obtidas do RMSE serão as mesmas que as da variável objetivo (*target*), dependendo do tipo de dados, o que torna mais acessível a compreensão do resultado e compreender se o tamanho do erro é significativo ou não (Kantardzic, 2003). O RMSE deverá ser o menor possível, para que determine se os valores estimados estão próximos dos reais. O RMSE é o erro quadrático médio e é calculado consoante a Equação 2:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Valor\ Predito_i - Valor\ Actual_i)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \quad \textbf{(2)}$$

Equação 2- RMSE

4.1.1. Preparação dos dados para modelação aos objetivos propostos

Na construção dos modelos de base para cada questão, recorreu-se às funções encontradas nas bibliotecas *Scikit Learn* do *Python*. Estes modelos (Apêndice A), serão comparados, de forma a avaliar o melhor desempenho obtido.

Dado que ambas as amostras apresentam sazonalidade, não será viável a aplicação do método de avaliação com *cross validation,* já que a extração com aleatoriedade poderia originar o fenómeno 'fuga' ou *leakage.* Assim, a avaliação foi feita com dados de treino para aprendizagem dos algoritmos e teste, para testar o desempenho dos algoritmos e validar a sua generalização. Tendo sido efetuada uma partição da amostra, entre dados de

treino e teste, em que os dados de 2010 a 2017 foram utilizados para treino e os dados de 2018, para teste.

Relativamente para às questões Q1 e Q2 em causa, os dados usados para treino constituem uma amostra desequilibrada, uma vez que se encontra um número de observações muito mais elevado para a classe ACMI do que para as restantes. Assim, foi realizada a experiência de modelação com e sem a aplicação da técnica de replicação sintética da amostra, SMOTE. O algoritmo (SMOTE) cria dados artificiais com base nas similaridades de espaço entre a classe minoritária existente, introduzindo minorias não replicadas. Espera-se que a introdução dos novos exemplos sirva para mudar o viés da aprendizagem relativamente à classe maioritária. As novas instâncias minoritárias são extrapoladas e criadas a partir dos desequilíbrios existentes da classe minoritária usando o algoritmo KNN. Os vizinhos são escolhidos aleatoriamente de acordo com a quantidade de sobre amostragem que é requerida. (Bowyer *et al*, 2002).

Para dar resposta às questões Q1 e Q2, dado que se trata de problemas de classificação, recorreu-se à modelação por algoritmos de classificação. Relativamente à questão Q3, para além da utilização dos algoritmos de regressão, foi também desenvolvida uma abordagem com séries temporais e, por fim, recorreu-se ao uso de uma técnica de *deep learning* recorrendo a uma LSTM para a regressão. Enquanto, para modelar as questões Q1 e Q2, foi efetuada análise de resultados com (e sem) a aplicação da técnica SMOTE. Na questão Q3, pela característica da amostra e como se tem por objectivo uma previsão de valor numérico contínuo, não se recorre à aplicação desta técnica.

4.1.2. Estudo de modelação para responder à questão Q1

Ao criar a curva ROC (*Receiver Operating Characteristic*), para o modelo base, em que se pode verificar o rácio entre a sensibilidade e a especificidade, observa-se o poder discriminante de um teste de diagnóstico. Trata-se da curva representativa da taxa de verdadeiros positivos (Taxa de VP) em função da taxa de falsos positivos (Taxa de FP). A área ROC é habitualmente escolhida em detrimento do acerto, ou para complementar a avaliação fornecida por outra métrica, aquando da utilização de conjuntos de dados que

apresentam classes equilibradas, dado que têm tendência a capturar eficazmente o equilíbrio entre verdadeiros positivos e verdadeiros negativos (Raeder, 2008).

Para todas as classes existem valores bastante razoáveis que determinam a qualidade deste teste. Tal como a AUC (*Area under the curve*), onde se constata resultados razoáveis de equilíbrio entre os VP e FP. Fazendo a curva ROC com o modelo de Regressão Logística e com a aplicação da metodologia SMOTE, obtém-se resultados em que todas as classes de tipologia operacional, têm uma AUC acima de 70% o que significa que este modelo está relativamente bem ajustado, conforme Figura 26.

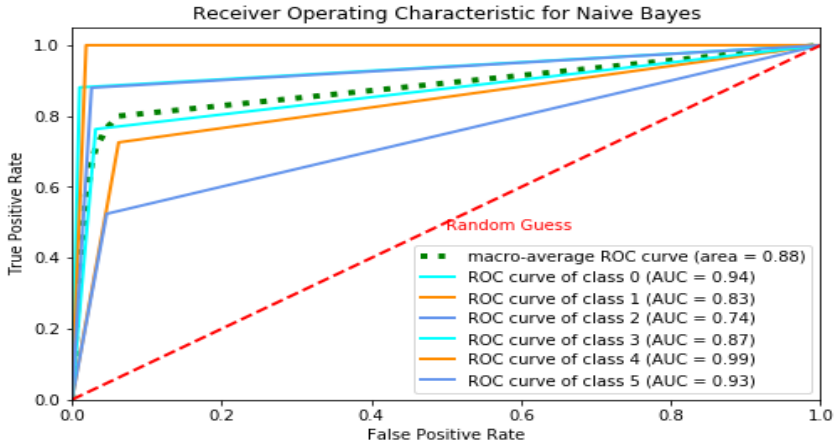

Figura 26 – ROC

Para responder à questão Q1, o modelo base selecionado de comparação, foi construído usando a técnica de Regressão Logística. Esta técnica é adequada para a classificação binária, embora possa ser estendida à classificação multiclasses. Baseia-se na determinação de um valor de probabilidade de pertença a (cada) uma das classes a estudar. Na verdade, as distribuições de probabilidade, que podem estar na origem dos dados que são utilizados para treinar o modelo, são estimadas através de uma transformação logarítmica – logit – que estima a probabilidade do evento, isto é, da observação, pertencer a uma dada classe. Assim, é feita a comparação do modelo base com os modelos de classificação de: Naive Bayes, Árvores de Decisão, KNN (K – Vizinhos mais próximo), SVM (Máquinas de Vectores de Suporte), ANN (Rede Neuronal Artificial) e LSTM. Na Tabela 5 observam-se os parâmetros do *Scikit Learn - Python* que optimizaram os resultados de desempenho. Para todos os algoritmos foi também aplicada a técnica SMOTE para avaliar se os resultados melhoravam ou pioravam.

Regressão Logistica Multinomial (base)	**Multiclasses**	**Extração aleatória**	**Algoritmo Optimizaçã o**		
	sim	não	saga		
Gaussian Naive Bayes Classificação	**Prioridade de classes**	**Estabilidade do cálculo**			
	sem	1,00E-04			
Árvores Decisão Classificação	**Critério separação dados**	**Estratégia de separação**	**Peso classes**		
	Gini	best	multi output		
KNN Classificação (ciclo k=1,…,k=10)	**Algortimo**	**Distância**	**Pesos**		
	brute	*Manhattan*	distribuidos pela distância		
SVM Classificação	**Kernel**	**Função de retorno**			
	scale	one-vs-one			
ANN classificação	**Função activação**	**Solver**	**Termo penalizador**	**Número neurónios – camadas escondidas**	
	logistica	adam	1,00E-11	50	
LSTM classificação (ciclo 10 repetições)	**Épocas**	**Argumento de ativação**	**Dropout**	**Camadas**	**Batchsiz e**
	80	relu	0,5	100	64

Tabela 5- Parâmetros Scikit Learn Python Q1

Os resultados de treino para a questão Q1 obtidos foram obtidos e descritos na Tabela 6:

	Acerto (accuracy)	**Precisão (precision)**	**Sensibilidade (recall)**	**SMOTE**		**Matriz confusão - classes c/ mais previsões incorrectas**		
Regressão Logistica Mult. (base)	0,84	0,84	0,84	**Sem**	-	3	5	
	0,84	0,84	0.84	**Com**	-	3	5	
Naive Bayes	0,84	0,87	0,84	**Sem**	+	3	5	
	0,83	0,83	0,83	**Com**	-	3	5	
KNN k=4	0,85	0,85	0,86	**Sem**	-	4	5	6
k=8	0,87	0,86	0,82	**Com**	+	4	5	6
SVM	0,86	0,86	0,86	**Sem**	-	4	5	6
	0,88	0,89	0,83	**Com**	+	4	5	6
LSTM	0,87	-	-	-	-			
	0,85	-	-	-	-			

Tabela 6 - Resultados de avaliação de treino para a questão Q1

A precisão para alguns dos algoritmos é bastante razoável, mas verifica-se, de acordo com a matriz confusão, maior dificuldade dos modelos em predizer correctamente as classes de tipologia operacionais 3, 4, 5 e 6. Portanto estas classes apresentaram poucas predições correctas ou nenhumas, provavelmente consequência de se encontrarem em número mais reduzido. No entanto, com a técnica SMOTE, observou-se algumas melhorias no desempenho dos modelos SVM (Figura 27), não só em termos de métricas de avaliação como também na matriz de confusão, as classes têm menos lacunas de predição nas classes 3, 4, 5 e 6, embora tenha baixado ligeiramente a capacidade de predição das classes 1 e 2, com o modelo KNN o comportamento é idêntico.

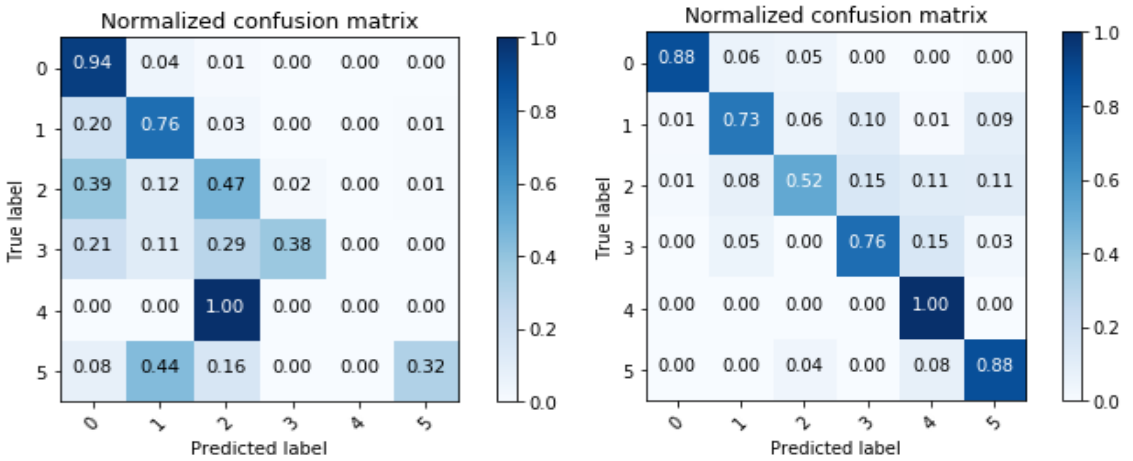

Figura 27 - Matriz de confusão para SVM sem e com técnica SMOTE (respectivamente)

4.1.3. Estudo de modelação para responder à questão Q2

Para a questão Q2 foi tomada a mesma abordagem como na questão Q1. Neste caso, os melhores resultados de desempenho obtidos, com os parâmetros *Scikit Learn* do *Python* descritos na Tabela 7. Para todos os algoritmos foi também aplicada a técnica SMOTE e analisado se os resultados apresentavam melhorias ou não.

Regressão Logística Multinomial (base)	**Multiclasses**	**Extração aleatória**	**Algoritmo Optimização**		
	sim	não	saga		
Gaussian Naive Bayes Classificação	**Prioridade de classes**	**Estabilidade do cálculo**			
	sem	1,00E-20			
Árvores Decisão Classificação	**Critério separação dados**	**Estratégia de separação**	**Peso classes**		
	Gini	best	multi output		
KNN Classificação (ciclo k=1,…,k=10)	**Algortimo**	**Distância**	**Pesos**		
	brute	*Manhattan*	distribuidos pela distância		
SVM Classificação	**Kernel**	**Função de retorno**			
	-	-			
ANN classificação	**Função activação**	**Solver**	**Termo penalizador**	**Número neurónios – camadas escondidas**	
	relu	adam	1,00E-05	50	
LSTM classificação (ciclo 10 repetições)	**Épocas**	**Argumento de ativação**	**Dropout**	**Camadas**	**Batchsize**
	60	relu	0,5	100	70

Tabela 7 - Parâmetros Scikit Learn Python Q2

Com o algoritmo SVM a máquina não consegue efetuar o processamento do modelo em causa para esta questão, fica pendente o resultado por tempo indeterminado, sem qualquer demonstração de resultados. Numa tentativa de obter o resultado deste algoritmo, foi efetuada a variação dos parâmetros e nova redução da dimensionalidade da amostra de treino progressivamente, excluíndo os anos de 2010, de 2010 e 2011 e assim sucessivamente até 2010 a 2015, por várias interações, no entanto, sem sucesso.

Os resultados de treino, para a questão Q2, foram obtidos e descritos conforme Tabela 8:

	Acerto (accuracy)	Precisão (precision)	Sensibilidade (recall)	SMOTE		Matriz confusão – classes c/ mais previsões incorrectas	
Regressão Logistica Mult. (base)	0,32	0,47	0,32	**Sem**	+	3	6
	0,41	0,46	0,41	**Com**	-	3	6
Naive Bayes	0,77	0,75	0,77	**Sem**	+	3	6
	0,77	0,75	0,77	**Com**	+	3	6
KNN K = 6	0,49	0,57	0,50	**Sem**	-	3	6
	0,46	0,58	0,49	**Com**	+	3	6
SVM	-	-	-	**Sem**	-	-	-
	-	-	-	**Com**	-	-	-
ANN - Classificação	0,55	0,58	0,55	**Sem**	-	3	6
	0,83	0,83	0,83	**Com**	+	3	6
LSTM	0,76	-	-	-	-	-	-
	0,70	-	-	-	-	-	-

Tabela 8 - Resultados de avaliação de treino para a questão Q2

Com exceção do algoritmo de ANN que demonstrou razoável desempenho, os restantes algoritmos têm fracos resultados, em especial observando a matriz de confusão. Analisando as matrizes de confusão, revelam que as classes das aeronaves 3 e 6 têm poucas predições correctas, ou, não são de todo predictas. Houve algumas situações com melhorias na aplicação da técnica SMOTE relativamente à predição destas classes, mas não significativas.

4.1.4. Estudo de modelação para responder à questão Q3

A abordagem para esta questão foi efetuada usando um modelo de regressão. Os melhores resultados obtidos de desempenho foram com os parâmetros descritos na Tabela 9.

Regressão Múltipla	**fit_intercept**	**Normalização**		
	none	false		
Árvore de Decisão de Regressão	**Critério**	**Critério de separação**	**Estado aleatório**	
	mse	best	0	
KNN Regressão (ciclo k=1,…,k=20)	**weights**	**Algoritmo**	**p**	**Distância**
	distance	auto	1	minkowski
ANN de regressão	**Activação**	**Solver**	**Número neurónios – camadas escondidas**	**max_iter**
	relu	lbfgs	50	50
LSTM regressão	**Optimizador**	**Função Objectivo**	**Épocas**	**batch_size**
	sgd	'mean_squared_error'	200	50

Tabela 9 - Parâmetros Scikit Learn Python Q3

Os dados obtidos de treino para resposta à questão Q3, descritos na Tabela 10:

	RMSE
Regressão Múltipla	66,74
LSTM	21,093
Séries Temporais - ARIMA	6,64
Séries Temporais - SARIMA	3.35e+32
Séries Temporais - SARIMAX	10,5

Tabela 10 - Resultados de avaliação de treino para a questão Q3

Para a terceira questão, Q3, os dados, pelas suas características, têm comportamento idêntico a uma série temporal, dada a existência de alguma periocidade. Contudo, existem dias onde não se verificam observações por não terem existido voos realizados.

De forma a poder efetuar a abordagem com séries temporais, existe necessidade de verificar a sua estacionaridade, para verificar a necessidade de aplicação ou não de transformação dos dados. Assim, com o teste de hipóteses de *Dickey-Fuller*, em que se rejeita a hipótese nula (H0) e pelo teste KPSS, mostram evidência estatística que existe estacionaridade na série temporal e assim, descarta-se a necessidade de transformação dos dados. Os gráficos da PACF e ACF, que demonstram a autocorrelação entre observações e entre observações e a sua repectiva *lag* e auxiliam na extração dos parâmetros *p* e *d*, são indicadores de baixa correlações e auto-corelações. Para determinação através de função do *Python*, e comparação de resultados, foi também utilizada a função 'auto_arima' para auxílio na determinação dos melhores parâmetros.

Com os modelos com séries temporais univariado de ARIMA, AR e MA e misto, foi obtido um RMSE bastante razoável para os resíduos, mas verifica-se incapacidade na captação da componente sazonal, fazendo uma predição que graficamente se apresenta constante, sem qualquer indício de variabilidade ou tendência.

Para o método SARIMA, O *QQ-Plot* (Figura 28) mostra que a distribuição ordenada dos resíduos tem desvios em relação distribuição normal. O gráfico correlograma, mostra que os resíduos da série temporal têm baixa correlação com a suas próprias *lags*. Contudo o gráfico *Normal Q-Q* surgem dúvidas da normalidade, dado que a distribuição não se encontra adequada à recta.

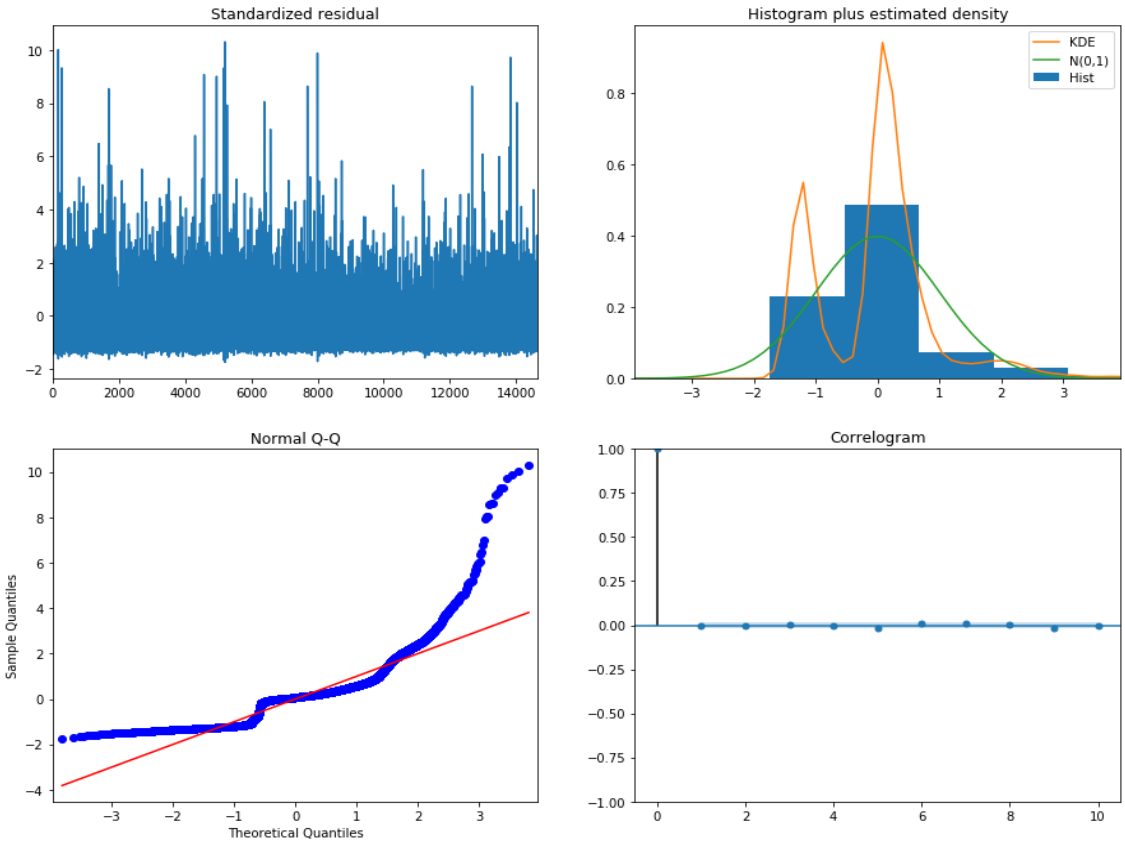

Figura 28 – Gráficos de diagnóstico

Os resultados do método SARIMA são indicadores que o modelo produz um ajuste relativamente satisfatório, contudo na representação gráfica da predição não existe uma nítida capacidade de predição com este modelo, denota-se uma estabilidade de previsão ao prever valores de tripulação a contratar no futuro. O resultado obtido de RMSE não foi o mais fraco, mas não houve a inclusão do factor tempo por limitação dos dados ao *software* deste modelo.

A abordagem de séries temporais com o método SARIMAX apresenta relativamente bom RMSE, contudo dado o facto de não estarmos perante uma periodicidade assídua de registos, o resultado também não foi determinado com inclusão do factor tempo.

4.2. **Avaliação**

As avaliações de desempenho dos algoritmos garantem a fiabilidade dos resultados. As medidas são numéricas e fazem a quantificação da performance de um determinado classificador (Troy, 2008). As métricas de avaliação, conforme mencionado na secção anterior, permitem determinar se se está perante um bom algoritmo de modelo para os dados ou não, verificando o desempenho de aprendizagem do modelo com os dados de teste.

No processo de modelação na aprendizagem automática realiza-se a divisão entre um conjunto de treino e um conjunto de validação de previsão relativamente à variável objectivo, tendo também como missão evitar o ajuste exagerado (*overfitting*). A avaliação do desempenho é efetuada com um conjunto de teste, onde se compara a predição realizada pelo modelo treinado com os valores reais que deviam ter sido atingidos, sendo verificado o desempenho com as métricas de avaliação (Apoorv Maheshwari *et al*.,2018).

- ***Valiação resultados obtidos relativamente à questão Q1***

	Acerto (accuracy)	**Precisão (precision)**	**Sensibilidade (recall)**	**SMOTE**		**Matriz confusão - classe c/ mais previsões incorrectas**	
Árvores Decisão Classificação	0,91	0,91	0,91	**Sem**	-	5	6
	0,87	0,88	0,87	**Com**	+	5	6
ANN - Classificação	0,89	0,89	0,89	**Sem**	-	5	6
	0,90	**0,92**	**0,90**	**Com**	+	-	-

Tabela 11 -Resultados de avaliação de teste para a questão Q1

Portanto, conforme Tabela 11, verifica-se que houve um bom desempenho para os algoritmos de ANN e de Árvores de Decisão para Classificação, dado a obtenção das métricas com resultados acima dos 90%. Contudo, para as Árvores de Decisão de classificação, os resultados demonstram a incapacidade de predição para as classes de tipologia operacional 5 e 6. Houve uma melhoria com a técnica SMOTE mas, ainda assim, existem bastante lacunas. O modelo construído com a ANN também se mostrou incapaz na predição correcta das classes 5 e 6, contudo houve melhorias substanciais devido à

aplicação da metodologia SMOTE e todas as classes foram predictas com sucesso, conforme matriz de confusão (Figura 29). Para esta questão, o mais importante será a métrica de precisão, para que a classe seja predita com sucesso, portanto com precisão.

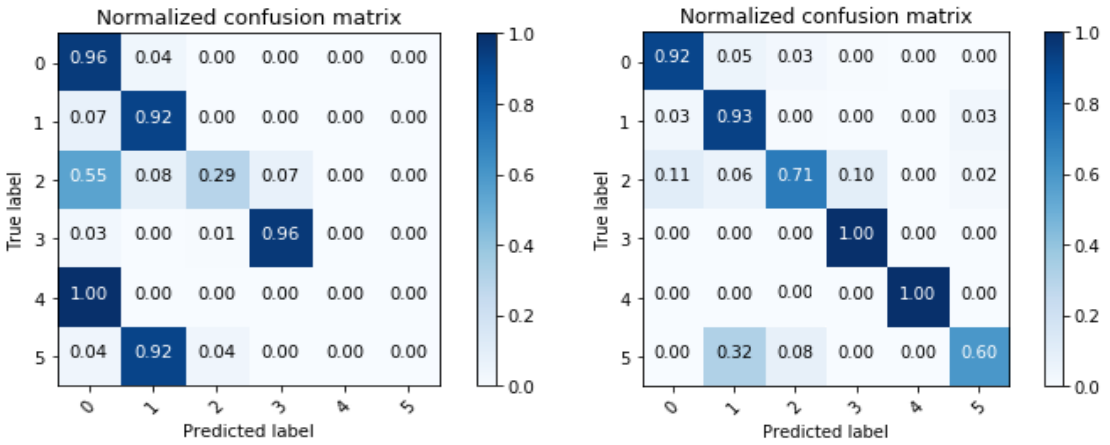

Figura 29 - Matriz de confusão para ANN sem e com técnica SMOTE (respectivamente)

- ***Validação resultados obtidos relativamente à questão Q2***

	Acerto (accuracy)	Precisão (precision)	Sensibilidade (recall)	SMOTE		Matriz confusão – classes c/ mais previsões incorrectas	
Árvores Decisão Classificação	**0,93**	**0,87**	**0,93**	**Sem**	+	5	6
	0,87	0,86	0,89	**Com**	-	5	6

Tabela 12 - Resultados de avaliação de teste para a questão Q2

De acordo com a Tabela 12, a melhor avaliação obtida de desempenho foi para a Árvore de Decisão de Classificação, embora se note que, de acordo com a matriz de confusão (Figura 30) , que a classe 5, não obteve predições correctas, e na classe 6, da aeronave teve 88% de predições incorrectas, logo, o modelo não tem uma sólida capacidade de predição. Sem a aplicação do SMOTE as classes das aeronaves, 1, 2 e 4 têm maior taxa de predições correctas, com a aplicação da metodologia SMOTE a classificação das classes 3 e 6 obtêm melhores resultados comparativamente, mas a classe 5 continua a não se obter uma previsão correcta, e as restantes classes pioram o seu desempenho de capacidade de predição, conforme matriz de confusão.

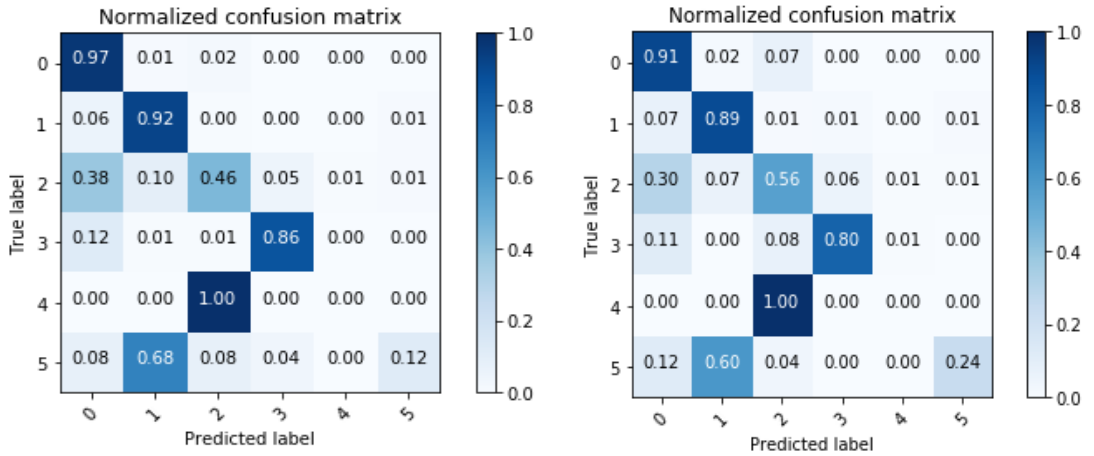

Figura 30 - Matriz de confusão para Árvores de Decisão sem e com técnica SMOTE (respectivamente)

- ***Validação resultados obtidos relativamente à questão Q3***

Para a terceira questão Q3, foi demonstrado que o facto de o comportamento temporal não ser totalmente assíduo e periódico, existindo dias onde não se verificam observações por não terem sido realizados voos, teve como consequência a não inclusão deste factor temporal no carregamento dos dados nos algoritmos dos modelos, ou, no caso de ter sido considerado, os modelos têm fraco ou péssimo desempenho.

	RMSE
Árvore de Decisão Regressão	5,789
KNN (ciclo k=1,…,k=20)	5,92
ANN	**5,0E-05**
Prophet	[5.64, 6.44]

Tabela 13 - Resultados de avaliação de teste para a questão Q3

Com o método *Prophet* foi obtida a predição conforme Figura 31, em que demonstra uma predição com intervalo alargado, esta metodologia ajusta-se à sazonalidade anual, semanal e diária, além de efeitos de férias ou feriados. Este método é robusto perante lacunas de dados, alterações nas tendências e *outliers*.

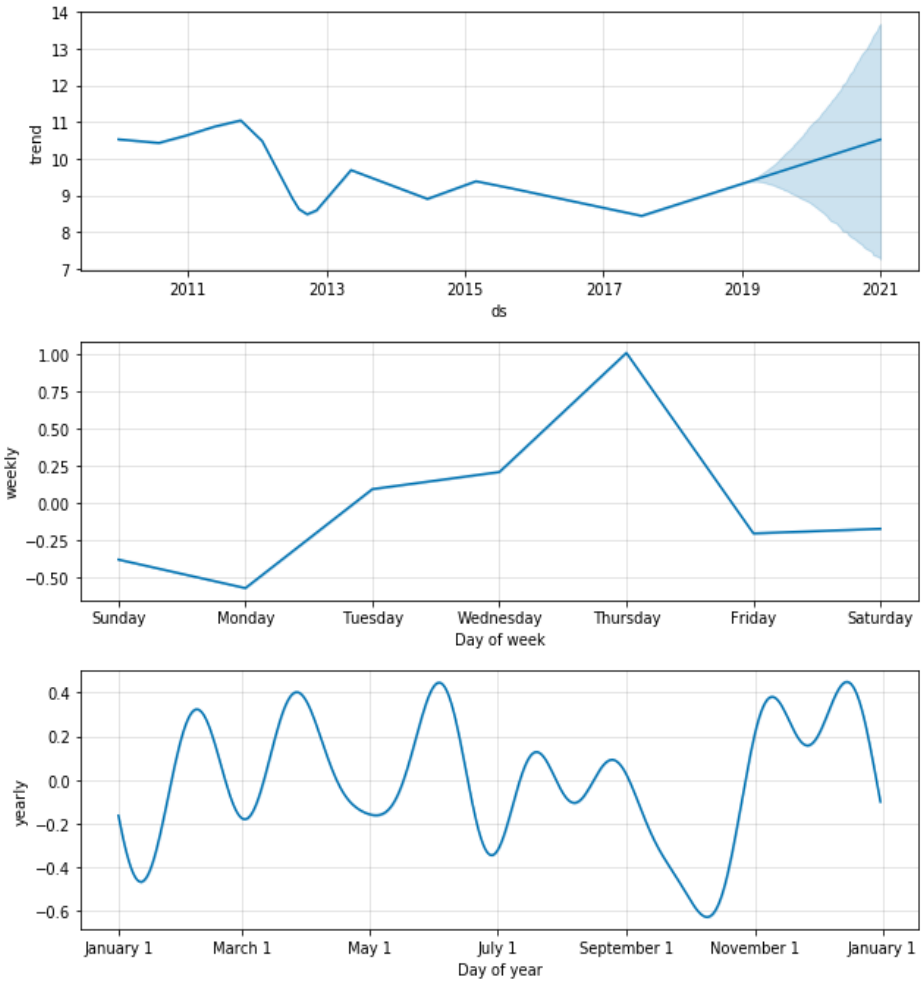

Figura 31 - Predição Prophet

Pela análise de tendência existe um intervalo de necessidade de tripulantes que poderá diminuir ou aumentar bastante, existe uma tendência semanal de maiores necessidades tendencial à quinta-feira, que corrobora com a análise exploratória [Secção 3.4.1.], com maior número de tripulantes e/ou um maior número de voos com tripulantes posicionados, que coincide com maior actividade operacional. Denota-se ainda uma disposição para o aumento de actividade noutros meses. Com a predição intervalar, obtém-se também um RMSE em intervalo.

4.3. Conhecimento para Apoio à Decisão

A ferramenta escolhida para este projeto foi o *Python*, versão 3.7. O software *Python* pode ser instalado gratuitamente usando o site *Anaconda*, em www.anaconda.com. Apesar de gratuita é uma ferramenta com bastantes potenciais, com boas metodologias a nível de programação e com um conjunto de *packages* em constante desenvolvimento. O *Python* é uma linguagem de programação orientada a objetos e permite criação de funções e rotinas pelo utilizador para a análise e manipulação de dados, desde a fase inicial de análise exploratória, para carregamento nos algoritmos e avaliação dos modelos, até à última fase, de comunicação. A fase de conhecimento de Apoio à Decisão é a demonstração dos resultados de forma a obter fácil compreensão e servir de suporte no apoio à decisão. Os modelos finais implementados não representam a finalização do projecto, sendo necessária a sua contínua monitorização, os resultados obtidos terão de ser organizados e apresentados de forma facilmente acessível, de forma a serem compreendidos e correctamente utilizados. Dependendo do que foi definido no plano de Conhecimento para Apoio à Decisão, o relatório poderá apenas incluir um resumo de todo o projecto e algumas notas, ou então um relatório final detalhado, com uma apresentação exaustiva dos resultados obtidos (Shearer, 2000).

- ***Resultados obtidos - Q1***

A resposta à questão Q1, é positiva, existe alguma evidência estatística que o modelo terá a capacidade de predizer a próxima tipologia operacional, com bom desempenho, esse modelo será a Rede Neuronal Artificial de classificação. Os resultados obtidos, tanto a nível de acerto, como de precisão e sensibilidade, foram favoráveis. Observando em especial a precisão que será a métrica mais importante, e a matriz de confusão, este modelo consegue predizer todas as próximas tipologias de classes operacionais. Assim, será possível a preparação logística departamental atempada para a próxima tipologia operacional, fazendo a sua determinação com recurso à modelação que apresentou melhor desempenho para esta questão.

- ***Resultados obtidos - Q2***

A resposta à questão Q2, é de que é parcialmente possível, dado que, embora exista alguma evidência estatística que o modelo de Árvores de Decisão de Classificação consiga predizer com sucesso qual o modelo de aeronave com maior procura e tenha uma boa precisão, existem classes de aeronaves em que o modelo com melhor desempenho não consegue determinar (classe 6, portanto o jacto e classe 5, portanto A380), e tem algumas dificuldades de classificação na classe 3, portanto o A333, conforme a matriz de confusão. No entanto esta dificuldade de classificação, pode ser justificada pelo tempo reduzido que as aeronaves se encontram na companhia ou porque se encontra recentemente na companhia e logo não existe histórico que sustente devidamente o treino do modelo. Portanto, existem algumas evidências do modelo de aeronave que será alvo de procura em determinadas circunstâncias, sustentado pelo desempenho da modelação para esta questão, e deverá ser adquirido ou realizado contrato de *leasing* (arrendamento de aeronaves) por forma a dar resposta à procura de acordo.

- ***Resultados obtidos - Q3***

A resposta à questão Q3, é parcialmente, sem recurso à informação temporal o melhor resultado obtido foi com a Rede Neuronal Artificial de regressão com um RMSE de 0.00005 para determinar a tripulação necessária para operar voos da empresa, portanto o valor mais baixo de erro e o melhor de todos os resultados obtidos, salvaguardando que não faz inclusão do factor tempo, pelo que se torna mais complexa a resposta exacta, já que, está-se perante dados que apresentam sazonalidade acentuada. Desta forma, existem algumas evidências do modelo ser capaz de predizer o número de tripulação, com razoável aproximação, a ser necessária a contratação, ou a ter disponibilidade para operar, e dar resposta às necessidades em determinadas circunstâncias.

Capítulo 5 – Conclusões e recomendações

Apesar do facto de que o negócio de tráfego aéreo da aviação tem revelado crescimento efetivo, as empresas de aviação regular nem sempre dispõem de frota com capacidade de cobertura integral de um período operacional anual, que permita dar total resposta a picos de atividade ou acontecimentos inesperados, estando sujeitas ao recurso ao aluguer de aeronaves a outras companhias para fazer face a estes acontecimentos. Por consequência, as companhias aéreas procuram cada vez mais fazer a previsão do comportamento da procura, com o intuito de propiciar uma melhor alocação dos seus investimentos e minimizar imprevistos indesejáveis no futuro.

Este estudo pretende explorar, a partir dos dados recolhidos de uma empresa particular de aviação civil, dar resposta à capacidade destes dados para fazer previsões de apoio à decisão e gestão em termos de frota e tripulação. Em especial, pretende explorar modelos de resposta a três questões específicas: (Q1) Será que existe dos modelos estudados, algum mais adequado para predizer a próxima tipologia operacional? (Q2) Existirá a possibilidade de encontrar a modelação adequada para determinar os modelos de aeronaves mais aconselháveis a adquirir, dada a procura? e (Q3) Existirá um modelo capaz de encontrar a quantidade de tripulação adequada à procura?

Neste sentido, foram exploradas várias técnicas em diferentes fases importantes no estudo: preparação e análise de dados e treino de modelos, que resultam nas conclusões que se apresentam em seguida.

5.1. Principais conclusões

A utilização de determinadas metodologias de exclusão de variáveis na amostragem, ou de replicação da amostra, podem influenciar o desempenho dos algoritmos dos modelos de diferentes formas, optimizando-os ou não. Nas questões Q1 e Q2 em análise existiu, em alguns casos, um bom desempenho de aplicação da metodologia SMOTE de replicação sintética da amostra. A exclusão de variáveis pelas metodologias: *Lasso, RFE* ou *Univariate Selection* não trouxeram melhorias nos resultados de desempenho dos modelos para os dados em estudo. Em particular pode-se concluir que as variáveis não se

encontram correlacionadas, ou apresentam correlações bastante fracas ou desprezíveis e dada a existência de alguns *outliers*, a abordagem para a redução de dimensionalidade ACP, não se demonstra adequada para aplicação nos algoritmos de aprendizagem automática.

A standardização das variáveis denota uma melhoria nas métricas de avaliação e também na rapidez do desempenho de resolução dos algoritmos, em especial dado o facto de se estar na presença de variáveis nas amostras de dados com diferentes escalas.

O facto da complexidade de determinados algoritmos ser maior, nem sempre implica que melhorem a capacidade de predição, podendo mesmo por vezes levar a uma degradação do desempenho.

Cada questão requer uma forma singular de abordagem dos dados, tal como a afinação do modelo do algoritmo, sendo variável o tempo de processamento com a inclusão do total das observações no algoritmo ou parciais.

Na investigação de resposta à questão Q1 em que se pretende obter a modelação adequada para predizer qual a próxima tipologia operacional, conclui-se que o melhor desempenho obtido foi com o modelo ANN, o que corrobora com alguns artigos científicos, é comum no ramo da aviação o recurso a este algortimo.

Para a questão Q2, o melhor desempenho obtido foi com a Árvores de Decisão de Classificação, existem também diversos artigos científicos com a implementação desta técnica. Este modelo demonstrou dificuldades na classificação da classe da aeronave A380 e o jacto, no entanto, um dos modelos de avião está há pouco tempo presente na companhia e outro tem voos mais exporádicos, pelo que corrobora com a necessidade de algum histórico de dados para poder realizar inferências robustas.

Para a questão Q3, referente a predição do número de tripulação, o modelo que se demonstrou mais adequado foi o de Rede Neuronal Artificial de Regressão, o que também se revelou comum neste tipo de aplicação de modelo na aviação. O factor limitativo dos dados possuírem alguma periocidade, mas por não se tratarem de uma série temporal, foi restritivo para a aplicação dos modelos de séries temporais, exceto com a nova

metodologia *Prophet*, que aceita as lacunas temporais, mas que, no entanto, se obteve uma predição com um intervalo demasiado alargado.

5.2. Contributos para a comunidade científica e empresarial

5.2.1. Implicações ao nível académico

Ao nível académico não foi encontrado qualquer estudo idêntico a esta dissertação, tendo sido encontrada a grande maioria dos trabalhos e pesquisas ao nível da aviação regular para serviços, previsão de procura, equipamentos e prevenção a nível de segurança.

Este estudo demonstra que dependendo das questões objectivo a aplicação de metodologias de tratamento de dados, tal como os algoritmos para as predições pretendidas é diferente, pelo que se torna necessária a experimentação das diversas técnicas e observar o desempenho de cada modelo para decisão da melhor sequência metodológica a seguir.

5.2.2. Implicações ao nível empresarial

A nível empresarial este estudo torna-se uma mais valia, já que auxilia no apoio à decisão.

Foi possível concretizar a análise do comportamento dos dados, e com estes, poder inferir as necessidades da empresa tanto a nível de preparação logística, como de equipamentos necessários, e de recursos humanos.

A decisão planificada traz mais valias para a empresa, e ir ao encontro da optimização das necessidades e ter oferta adequada à procura, desta forma pode-se aumentar a margem de lucros, dado que a qualidade de resposta que o cliente procura, consegue ser atingida se for aplicada a preparação logística departamental de acordo com as soluções de respostas encontradas às questões e objectivos colocados.

5.3. Limitações do estudo

Dever-se-á começar por fazer notar que a confidencialidade exigida pela empresa relativamente aos dados não permitiu inclusão de dados mais detalhados sobre as companhias aéreas de forma a poder inferir-se outras conclusões.

As variáveis disponibilizadas foram apenas referentes à base de dados de voos e tripulação, pelo que, se fosse possível aceder a outras variáveis disponíveis de outros departamentos, seria uma mais valia para auxiliar na análise e aplicação de metodologias de extração de conhecimento. Em especial, departamentos que envolvem custos operacionais permitiriam obter também tendências dos factores económicos da empresa.

Alguns dos algoritmos por exigirem maior capacidade de processamento, do que a máquina usada, (DELL modelo Optiplex 900 com velocidade processador Intel ® Core ™ i5-2400 CPU @3.10 Ghz 3.10 GHz, RAM 8 GB), consegue suportar e acaba por se tornar necessário a redução da dimensão da amostra de treino, ou mesmo, após esta (re)dimensionalidade, inviabilizou a resolução. A redução dos dados de treino poderá levar a uma pior aprendizagem dos modelos.

O facto de não haver registos regulares diários limita a aplicação de modelos de séries temporais que exigem essa regularidade periódica nos dados para carregamento nos modelos.

O tempo foi um factor determinante de limitação para alargar a pesquisa e para a possibilidade de maior amplitude de uso de outros algoritmos e aprofundar mais a afinação dos respectivos modelos.

5.4. Propostas de investigação futura

Para futuros trabalhos apresentam-se as seguintes propostas:

A possibilidade de extração de mais dados para inclusão de mais observações nos dados de forma a permitir maior capacidade de treino dos algoritmos e redução da desiquilibrio entre classes, já que se revelou bastante presente nos dados.

Experimentação de outras técnicas de replicação de amostras de forma a poder obter melhor representação de classes.

Aplicação de outras técnicas tais como *Ensemble*, *XGbost*, *Random Forest*, entre outros, dado que poderão demonstrar melhores resultados de desempenho.

A possibilidade de inclusão de dados referentes à informação de custos para que seja possível a inclusão de fatores económicos na modelação, de forma a fundamentar de forma mais robusta a optimização de margens de lucro no apoio à decisão.

Bibliografia

AbdEl Rahman ElSaida Fatima El Jamiy, James Higgins, BrandonWild, Travis Desella, "Optimizing long short-term memory recurrent neural networks using ant colony optimization to predict turbine engine vibration" [Journal]. - [s.l.] : Applied Soft Computing, 2018. - Vol. Volume 73.

Adeniyi D. A., Wei, Z., & Yongquan, Y. " Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method" [Article]. - [s.l.] : Applied Computing and Informatics, 2016. - Vols. 12(1), 90–108..

Alpaydın Ethem Introduction to Machine Learning [Book]. - London, England : Massachusetts Institute of Technology, 2010.

André A.P. Santos ,Luciano N. Junkes e Floriano C.M.Pires Jr " Forecasting period charter rates of VLCC tankers through neural networks: A comparison of alternative approaches" [Journal]. - [s.l.] : Santos, Maritime Economics & Logistics, 2013. - Vols. 16(1), 72–91. doi:10.1057/mel.20.

Anna Torun Czelaw Burniak, Jerzy Bialy, Justyna Tomaszewska, Norbert Grzesik, Sarka H., Marta W., Marius Z e Adam R. "Challenges for Air Transport Providers in Czech Republic and Poland" [Journal]. - [s.l.] : Journal of Advance Transportation., 2018.

Apoorv Maheshwari Navindran Davendralingam e Daniel A. DeLaurentis "A comparative study of Machine Learning Techniques for Aviation Applications" [Conference]. - [s.l.] : Aviation Technology, Integration, and Operations Conference, 2018.

BaFail A.O. "Applying data mining techniques to fore-cast number of airline passengers in Saudi Arabia (domestic and international travels)" [Journal]. - [s.l.] : Journal of Air Transport-ation , 2004. - Vols. 9(1): 100–115.

Berster Johannes Reichmuth and Peter "Past and Future Developments of the Global Air Traffic" [Conference]. - Springer-Verlag GmbH Germany : [s.n.], 2018.

Billy M. Williams M.ASCE and Lester A. Hoel, F.ASCE " Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results" [Journal]. - [s.l.] : Journal of Transportation Engineering - ASCE, 2003.

Bishop Cristopher M. "Pattern Recognition and Machine Learning" [Book]. - Cambridge : Springer, 2006.

Boser B. E., Guyon, I. M., & Vapnik, V. N. "A Training Algorithm for Optimal Margin Classfiers" [Conference] // Proceedings of the Fifth Annual Workshop on Computational Learning Theory. - 1992.

Bramer Max Data for Data Mining [Book Section] // Principles of Data Mining. - [s.l.] : Springer-Verlag London, 2016.

Burnett R. A., and Si, D “Prediction of Injuries and Fatalities in Aviation Accidents through Machine Learning" [Conference] // Proceedings of the International Conference on Compute and Data Analysis. - [s.l.] : ACM, 2017. - Vols. pp. 60–68.

C Shearer "The CRISP-DM model: the new blueprint for data mining" [Journal]. - [s.l.] : Journal of Data Warehousing, 2000. - Vols. vol. 5, no. 4, pp. 13-22.

ChaoTong XiangYin, Jun Li, Tongyu Zhu, Renli Lv, Liang Sun, Joel P. C. Rodrigues "An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data" [Journal]. - [s.l.] : Applied Soft Computing – Elsevier, 2018. - Vol. Volume 73.

Cheng Tao-ran "Research status of Artificial Neural Network and Its Application Assumption in Aviation" [Conference] // 12th International Conference on Computational Intelligence and Security. - 2016.

Chin-Shan Lu Petrus Choy, Kee-hung Lai, Y.H. Venus Lun, Tsz Leung Yip " Empowering Excellence in Maritime and Air Logistics: Innovation Management and Technology" [Conference]. - [s.l.] : Proceedings of the International Forum on Shipping, Ports and Airports (IFSPA) , 2015.

Christopher A. A., Vivekanandam, V. S., Anderson, A. A., Markkandeyan, S., and Sivakumar, V. “Large-scale data analysis on aviation accident database using different data mining techniques" [Journal]. - [s.l.] : The Aeronautical Journal, 2016. - Vol. 120, No. 1234, pp. 1849.

David Rios Insua CesarAlfaro, JavierGomez, PabloHernandez-Coronado e FranciscoBernal "Forecasting and assessing consequences of Aviation safety

ocurrences", [Article]. - [s.l.] : Safety Science Volume 111, January 2019, 2019. - Pages 243-252.

Dean J. "Big Data, Data Mining and Machine Learning. Value Creation for Business Leaders and Practitioners" [Article]. - New Jersey : Wiley, 2014.

Deng Xixin Tang e Guangming "Prediction of Civil Aviation Passenger Transportation Based on ARIMA Model" [Article]. - [s.l.] : Scientific Research Publishing – College of Science, Gullin University of Technology, Guilin, China and Institute of Applied Statistics, Guil, 2016.

Eibe Frank Leonard Trigg, Geoffrey Holmes, Ian H. Witten "Naive Bayes for Regression" [Article]. - University of Waikato, Hamilton, New Zealand : Department of Computer Science, 1999.

Erskine Joseph R. "Developing Cyberspace Data Understanding: Using CRISP-DM for Host-Based IDS Feature Mining Paperback" [Article]. - Ohio : Air Force Institute of Technology, 2012.

Fawcett Foster Provost and Tom "Data Science for Business " [Book]. - 1005 Gravenstein Highway North, Sebastopol, CA 95472 : Published by O'Reilly Media, Inc., 2013. - Vols. isbn: 978-1-449-36132-7.

Fayyad U., Piatetsky-Shapiro, G., & Smyth, P "From data mining to knowledge discovery in databases" [Article]. - 1996. - AI Magazine. - 17(3), 37–53.

Fonseca. Nuno A. " Parallelism in Inductive Logic Programming Systems, PhD thesis" [Report]. - Porto : Faculdade de Ciências da Universidade do Porto, 2006.

Garcia S Dissertação: "O uso de árvores de decisão na descoberta de conhecimento na área da saúde" [Report]. - [s.l.] : Universidade Federal do Rio Grande do Sul. Instituto de Informática, 2003.

Garrido C., De Oña R. and De Oña J "Neural networks for analyzing service quality in public transportation" [Article]. - [s.l.] : Expert Systems with Applications, 2014. - Vols. 41(15): 6830–6838.

Ghobrial Atef "A Model to Forecast Aircraft Operations at General Aviation Airports" [Journal]. - [s.l.] : Jornal of Advanced Transportation, 1994. - Vols. Vol 31, Nº 3. pp. 311-323.

Gregory R. Herman Russ S. Schumacher "Using Reforecasts to Improve Forecasting of Fog and Visibility for Aviaton" [Journal]. - Lima : Universidade Federal de Santa Catarina, 2015.

Halford Alan J. Stolzer & Carl "Data Mining Methods Applied to Flight Operations Quality Assurance Data: A Comparison to Standard Statistical Methods" [Journal]. - [s.l.] : Journal of Air Transportation, 2007.

Han J. e Kamber, M. "The Data Mining: Concepts and Techniques" [Book]. - San Francisco : Second Edition, Morgan Kaufmann Publishers, 2006.

Han Jiawei [Article] // Data Mining Techniques. - Montreal, Canada : School of Computing Science, Simon Fraser University, British Columbia, Canada V5A 1S6, 1996.

Hanbong Lee Waqar Malik e Yoon C. Jung "Taxi-Out Time Prediction for Departures at Charlotte Airport Using Machine Learning Techniques" [Conference] // 16th AIAA Aviation Technology, Integration, and Operations Conference. - 2016.

Jain N., & Srivastava, V. Data Mining Techniques [Article]. - [s.l.] : International Journal of Research in Engineering and Technology,, 2013. - Vols. IJRET: 2(11) , 116–119.

Janakiraman V. M., and Nielsen, D. “Anomaly detection in aviation data using extreme learning machines” [Conference] // Neural Networks (IJCNN), 2016 International Joint Conference on , IEEE. - 2016. - Vols. pp. 1993–2000.

Jordán Gladys Castillo "Theses PhD: Adaptive Learning Algorithms for Bayesian Network Classifiers" [Report]. - [s.l.] : Departamento de Matemática, Universidade de Aveiro, 2006.

K.P.G. Alekseev J.M.Seixas "Forecasting the Air Transport Demand for Passengers with Neural Modeling, Signal Processing Laboratory" [Journal]. - [s.l.] : COPPE/EE – Federal University of Rio de Janeiro, 2002.

Kantardzic M "Data mining: concepts, models, methods, and algorithms" [Article]. - [s.l.] : WileyInterscience, 2003.

Koh H. C., & Tan, G. " Data Mining Applications in Healthcare" [Journal]. - [s.l.] : Journal of Healthcare Information Management, 2005. - No. 2, pp. 64-72. : Vol. Vol. 19.

Kotegawa T., DeLaurentis, D. A., and Sengstacken, A. “Development of network restructuring models for improved air traffic forecasts” [Conference] // Transportation Research Part C: Emerging Technologies Conference. - 2010. - Vols. Vol. 18, No. 6, pp. 937–949..

Kumba Sennaar "How the 4 Largest Airlines Use Artificial Intelligence" [Online]. - Emer, May 17, 2018. - Jan 19, 2019. - https://emerj.com/ai-sector-overviews/airlines-use-artificial-intelligence/.

LALIŠ Andrej "Time-Series Analysis and modelling to predict aviation safety performance index - Czech Technical University in Prague" [Article]. - Prague : Czech Technical University in Prague, Faculty of Transportation Sciences, 2017.

Lee H., Malik, W., and Jung, Y. C., “Taxi-out time prediction for departures at Charlotte airport using machine learning techniques” [Conference] // 16th AIAA Aviation Technology, Integration, and Operations Conference. - 2016. - Vol. p. 3910..

Li Cheng "Combined forecasting of civil aviation passenger volume based on ARIMA-REGRESSION" [Article]. - [s.l.] : Springer - Li, C. Int J Syst Assur Eng Manag, 2019.

Liang Zhang Na Lu "Structural Behavioral Study on the General Aviation Network Based on Complex Network" [Conference] // IOP Conference Series: Materials Science and Engineering. - 2017.

Lipton Z. C., Kale, D. C., Elkan, C., & Wetzel, R. "Learning to diagnose with LSTM Recurrrent Neural Networks" [Conference] // International Conference on Learning Representations. - 2016.

Lyasoff Rodin "How Artificial Intelligence Will Impact The Aviation Industry" [Journal]. - [s.l.] : Forbes, 2018.

Machado Marta Alexandra Lourenço "Modelos de previsão aplicados à optimização da gestão das actividades de um Call Center" [Article]. - [s.l.] : Mestrado em Estatística - Universidade de Lisboa Faculdade de Ciências Departamento de Estatística e Investigação Operacion, 2012.

Matthews B., Das, S., Bhaduri, K., Das, K., Martin, R., and Oza, N. “Discovering anomalous aviation safety events using scalable data mining algorithms” [Journal]. - [s.l.] : Journal of Aerospace Information Systems, 2013.

Mengchen Ji Gang Xie and Shouyang Wang "Can Google Search Data Help Nowcasting Air Passenger Volume? A Study of Hong Kong International Airport" [Journal]. - Hong Kong : Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 2015.

Merve Şahin Recep Kızılaslan and Ömer F. Demirel "Forecasting Aviation Spare Parts Demand Using Croston Based Methods and Artificial Neural Networks" [Journal]. - [s.l.] : Journal of Economic and Social Research, 2013. - Vols. Vol 15(2) 2013, 1-21.

MichaelSchultz and StefanReitmann "Machine learning approach to predict aircraft boarding" [Conference]. - [s.l.] : Transportation Research Part C: Emerging Technologies , 2018. - Vols. Volume 98, January 2019, Pages 391-408.

Mitchell T. M. Machine Learning [Book Section]. - [s.l.] : McGraw-Hill Companies, Inc., 1997.

Mutlu. Ajit Jaokar and Dan Howarth. With contributions from Ayse "Classification and Regression In a Weekend" [Book]. - 2016.

Nitesh V. Chawla Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique [Journal]. - [s.l.] : Journal of Artificial Intelligence Research 16, 2002. - 321–357.

Osman Erman Gungor Imad L. Al-Qadi "Developing Machine-Learning Models to Predict Airfield Pavement Responses" [Journal]. - [s.l.] : University of Illinois at Urbana-Champaign, Illinois Center for Transportation, Rantoul, IL, 2018. - Vols. Volume: 2672 issue: 29, page(s): 23-34.

Panarat Srisaeng Gleen Baxter "Modelling Australia's Outbound Passenger Air Travel Demand Using An Artificial Neural Network Approach [Article]. - Huahin : School of Tourism and Hospitality Management, Suan Dusit University, Huahin Campus, 2017.

Panetta Kasey 5 Trends Emerge in the Gartner Hype Cycle for Emerging Technologies [Online] // gartner.com. - Smarter with Gartner, 08 16, 2018. - 04 2019, 15. - https://www.gartner.com/smarterwithgartner/5-trends-emerge-in-gartner-hype-cycle-for-emerging-technologies-2018/.

Pestana Sílvio Filipe Velosa e Dinis Duarte "Introdução à Probabilidade e à Estatística" [Book]. - [s.l.] : Fundação Calouste Gulbenkian, 2006. - Vols. 4ª edição, isbn: 9789723111507.

Plummer Daryl Sieben Tech-Trends Fur CEOS und Cios 2019 [Article]. - AUSBLICK : DIGITAL BUSINESS CLOUD, 2019.

Pritscher Lisa and Feyen Hans "Data Mining And Strategic Marketing In The Airline Industry" [Article]. - CH-8058 Zurich-Airport, Switzerland : Atraxis AG, Swissair Group, Data Mining and Analysis, CKCB, 2017.

Priyanka "Prediction of Airline delays using K Nearest Neighbor Algorithm" [Journal]. - College of Technology : International Journal of Emerging Technology and Innovative Engineering, 2018. - Issue 12 - 125-126 : Vol. Volume 4.

Raeder Troy "Model Monitor User's Guide version 1.0" [Article]. - [s.l.] : Department of Computer Science and Engineering, University of Notre Dame,, 2008.

Raghavendra Totamane Amit Dasgupta, and Shrisha Rao " Air Cargo Demand Modeling and Prediction" [Journal]. - [s.l.] : IEE Systems Journal , 2014.

Rajeev Sharma Sunil Mithas e Atreyi Kankanhalli "Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations" [Journal]. - [s.l.] : European Journal of Information Systems, 2017.

Regazzi "R Development Core Team. R: a language and environment for statistical computing." Vienna: R Foundation for StatisticalComputing, Vienna, 2014. REGAZZI, A.J. Análise multivariada, notas de aula INF 766, Departamento de Informática da Unive

[Journal]. - Vienna : Departamento de Informática da Universidade Federal de Viçosa, v.2, 2014. - A.J. Análise multivariada, notas de aula INF 766.

Robertson Fredrik and Wallin Max "Forecasting monthly air passenger flows from Sweden - Evaluating forecast performance using the Airline model as benchmark" [Article]. - [s.l.] : Bachelor thesis - Department of Statistics - Uppsala University, 2014.

Rodrigo Marcos Oliva García-Cantú, Ricardo Herranz "A Machine Learning Approach to Air Traffic Route Choice Modelling" [Journal]. - [s.l.] : Nommon Solutions and Technologies, Madrid, 28006, Spain, 2017.

Rodrigues João Gabriel das Neves Dissertação:" Aprendizagem Automática Aplicada à Condução de um Veículo com Direção Ackermann" [Report]. - Universidade de Aveiro : [s.n.], 2018.

Ryan C. Boyer William T. Scherer William & T. Scherer "Trends Over Two Decades of Transportation Research: A Machine Learning Approach" [Journal]. - [s.l.] : Journal of the Transportation Research Board, 2017.

Sa J. "Reservations forecasting in airline yield management" [Article]. - [s.l.] : Flight Transportation Lab Report R87-1. Massachusetts Institute of Technology, p. 116-117. , 1987. - Vols. p. 116-117. .

Santos F. M., & Azevedo, C. S. Data Mining - Descoberta de conhecimento em bases de dados [Book Section]. - Lisboa, Portugal : Editora de Informática, 2005.

Sebastien Maire and Chris Spafford "The Data Science Revolution That's Transforming Aviation" [Journal]. - [s.l.] : Forbes, 2017.

Shmueli Deborah "Application of neural networks in transport planning, Progress in Planning" [Article] // Concept analysis for business aviation decision-making. - Charlottesville, VA 22904-4747, 2017. : Department of Systems and Information Engineering, School of Engineering and Applied Science, University of Virginia, 1998. - 151 Engineers Way. - Issue 3, Pages 141-204 : Vol. Volume 50.

Shuojiang Xu Hing Kai Chan, TiantianZhang "Forecasting the demand of the aviation industry using hybrid time series SARIMA - SVR approach" [Article]. - [s.l.] : Transportation Research Part E: Logistics and Transportation Review - Elsevier, 2019.

Smith A. Kyle Collins e Dimitri Mavris "Survey of Technology Forecasting Techniques for Complex Systems" [Conference] // 58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. - 2017.

Smith A., Collins, K., and Mavris, D., “Survey of Technology Forecasting Techniques for Complex Systems” [Conference] // 58th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. - 2017.

Song Min Wang e Haiyan "Air Travel Demand Studies: A Review" [Journal]. - [s.l.] : Journal of China Tourism Research, 2010.

Steiner George A. "Approaches to Log-Range Planning for Small Business" [Journal]. - 1967. - Vols. Volume: 10 issue: 1, page(s): 3-16.

Takaya Ukai Hsun Chao, Purdue and Daniel A. DeLaurentis "An Aircraft Deployment Prediction Model Using Machine Learning Techniques" [Conference]. - [s.l.] : 17th AIAA Aviation Technology, Integration, and Operations Conference, 2017.

Takeichi Noboru " Adaptive prediction of flight time uncertainty for ground-based 4D Trajectory management" [Journal]. - Tokyo : Department of Aeronautics and Astronautics, Tokyo Metropolitan University, 2018.

Taneja N. K “A model for forecasting future air travel demand on the North Atlantic” [Article]. - [s.l.] : Tech. rep., Cambridge, Mass. Massachusetts Institute of Technology, 1971. - Flight Transportation Laboratory.

Taneja Nawal K. Airline Survival Book [Book]. - New Yor USA : Routledge, 2003. - Vol. I.

Tao Li Antonio A. Trani "A Model to forecast airport-level General Aviation Demand" [Journal]. - [s.l.] : Journal of Air Transport Management, 2005.

Ujjwala Urkude Pratibha Richariya "Naive baye’s classification algorithm in prediction of Flight delays using MR" [Journal]. - Maxim institute of technology Affiliated to RGPV Bhopal : International Journal of Innovative Research in Technology, 2016.

Ukai T., Chao, H., and DeLaurentis, D. A. “An Aircraft Deployment Prediction Model Using Machine Learning Techniques” [Conference] // 17th AIAA Aviation Technology, Integration, and Operations Conference. - 2017. - Vol. p. 3081.

William T. Scherer Michael C. Smith " Artificial Intelligence and Machine Learning in Aviation" [Conference]. - [s.l.] : IATA, 2017.

Y. & Ramos, I. "Business Intelligence - Tecnologias da Informação na Gestão de Conhecimento" [Book Section]. - Lisboa, Portugal : Editora de Informática, 2009. - Vols. (2ª edição ed., Vol. 1).

Yash Madhwal Zinaida Avdeeva "Planning in Aircraft Industry based on prediction of Air Traffic" [Article]. - [s.l.] : Procedia Computer Science, 2017. - Vols. Volume 122, Pages 1047-1054.

YunBaia Bo Zeng, Chuan Li, Jin Zhang "An ensemble long short-term memory neural network for hourly PM2.5 concentration forecasting" [Journal]. - 2019 : Chemosphere – Elsevier. - Vols. Volume 222, Pages 286-294.

ZeFeng WANG Jean-Luc ZARADER, Sylvain ARGENTIERI, Karim YOUSSEF "A Decision System for Aircraft Faults Diagnosis Based on Classification Trees and PCA" [Article]. - Institut des Systèmes Intelligents et de Robotique : Université Pierre et Marie Curie, 2013.

Zhang X., & Mahadevan, S. Ensemble machine learning models for aviation incident risk [Article] // Ensemble machine learning models for aviation incident risk. - Nashville, USA : Elsevier, 2019. - Decision Support Systems. - 46-63 : Vol. 116.

Apêndices

Apêndice A - Descrições sumárias de algoritmos de Aprendizagem Automática

Nesta seção encontra-se a descrição sumária dos modelos de aprendizagem automática supervisionada de classificação e regressão, a informação foi extraída do livro de Cristopher M. Bishop - «Pattern Recognition and Machine Learning».

- **Árvore de Decisão (*Decision Trees*)**

Um algoritmo de árvore de decisão faz a aprendizagem, construindo iterativamente várias decisões usando os atributos do conjunto de dados no sentido de conseguir isolar um valor de resposta (ou objectivo) o mais rapidamente possível, quer se trate de uma regressão ou de uma classificação. No último caso, pretende-se medir o contributo de cada variável independente, de modo a escolher a que mais rapidamente consegue uma classificação, segmentando assim, o conjunto de dados de modo a tentar obter conjuntos de dados todos da mesma classe. Quanto menos mistura, maior o grau de pureza da segmentação obtida. O objetivo é criar um modelo que tenha a capacidade de prever um valor objectivo com base no conjunto de treino, ou seja, nos dados anotados (ou rotulados) que descrevem, portanto, exemplos da associação que deve ser aprendida. Esta técnica é, portanto, supervisionada.

- **K - Vizinhos mais próximos (KNN – *K-nearest neighboors*)**

A técnica de KNN, K-vizinhos mais próximos, é, também, um algoritmo para aprendizagem supervisionada. O seu processamento é simples: em face de todos os casos disponíveis para exemplo, prediz o valor objectivo (numérico ou uma classe) com base numa medida de similaridade, (ou seja, uma função de distância, como: Euclidiana, Manhattan, Minkowski, entre outras). O algoritmo KNN é usado principalmente para problemas de classificação. A técnica para o KNN consiste na análise dos *k* exemplos (de treino) mais próximos no espaço de variáveis. O(s) resultado(s) será (serão) igual (iguais) à classe maioritária nesses *k* exemplos, no caso de se tratar de um problema de classificação (Priyanka, 2018).

- **Naive Bayes**

Em termos simples, o classificador de Naive Bayes assume que o valor de uma variável específica não está relacionado com a presença ou ausência de qualquer outra característica, dada a classe da variável (Urkude & Richariya, 2016), o que se denomina de independência de acontecimentos (eventos) aleatórios. O núcleo desta técnica encontra-se no teorema de Bayes, em que se assume independência entre eventos preditores (Equação 3):

$$P(y|x1,\dots,x_n) = \frac{P(x1|y)P(x2|y)\dots P(x_n|y)P(y)}{P(x1)P(x2)\dots P(x_n)} \quad \textbf{(3)}$$

Equação 3 - Teorema de Bayes

Em problemas categóricos, a previsão ideal deverá ser compreendida entre zero e um, ou seja, de acordo com a distribuição subjacente. Contudo, em problemas de regressão, a previsão ideal é a média ou a mediana. Assim, quando usado numa previsão numérica, o algoritmo de Naive Bayes é mais sensível a estimativas de probabilidade imprecisas do que quando é usado para classificação (Eibe, 1999).

- **SVM (Support Vector Machine)**

Os modelos SVM implicam a determinação de um hiperplano óptimo (com a margem máxima) para separação das diferentes classes. A introdução da função do processo de minimização estrutural (termo de regularização) permite que o SVM tenha uma boa característica de generalização, garantindo assim o menor erro de classificação numa ou mais variáveis (Zhang & Mahadevan, 2019). As máquinas de suporte de vetores são utilizadas como técnica de classificação para o reconhecimento de padrões. As SVM adaptadas a multiclasses visam a classificação de classes fazendo uso de máquinas de suporte de vetores, onde a classe é desenhada a partir de um conjunto finito de vários elementos.

- **ANN – Artificial Neural Network**

As redes neuronais usam métodos de aprendizagem supervisionada, como a maioria das abordagens de aprendizagem automática, e comparam o resultado previsto com um valor de referência (Schultz & Reitmann, 2018). Os modelos de redes neuronais são úteis na modelação de relações não-lineares entre as variáveis. Têm robustez na presença de *outliers* e ruído nos dados. As ANN capturam as informações inerentes de um conjunto considerável de variáveis e aprendem com os dados existentes (Garrido *et al*., 2014). Estes modelos têm a capacidade de diversas variáveis de *input* (correspondentes aos atributos ou características do problema) e apenas uma variável de saída que corresponde ao que se pretende predizer, podendo esta ser discreta, em problemas de classificação, ou contínua em problemas de regressão. Para ser realizada a aprendizagem da rede neuronal vai existindo ajuste dos pesos da rede de forma a minimizar o erro cometido entre a saída da rede e o objetivo de saída pretendido (*backpropagation*). A função de ativação (geralmente não-linear), com passagem das somas dos pesos, é aplicada a cada neurónio (ou unidade) no *input* para determinar o seu *output* pretendido. Na Figura 32 pode-se observar o comportamento de um neurónio.

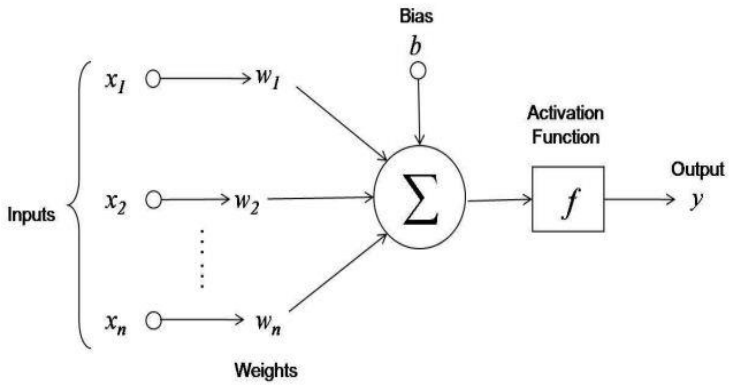

Figura 32- Neural Network – Neuron representation.

Fonte: https://naadispeaks.wordpress.com/2017/11/08/artificial-neural-networks-with-net-in-azure-ml-studio/