



Treball de Fi de Grau
GRAU D'ENGINYERIA INFORMÀTICA
Facultat de Matemàtiques
Universitat de Barcelona

**Sistema de recomanació per a la compra de
bitllets d'avió**

Jonathan Blasco Barrera

Director: Santi Seguí Mesquida
Realitzat a: Departament de Matemàtica
Aplicada i Anàlisi. UB
Barcelona, 30 de juny de 2016

Abstract

This work is in the frame of data science. Its main aim is to develop a recommender system for purchasing airline tickets. The system will allow the user to purchase airline tickets at the right time when ticket prices are lower. To develop this flight recommender, it has been made an analysis of the evolution of prices on important dates. The analysis is based on the study of the evolution of prices over time to implement a system of price prediction for a specific date and thus give the user a recommendation on the best day to buy a ticket.

Resum

Aquest treball està dins del marc de la ciència de les dades, en el que es desenvolupa un sistema de recomanació per a la compra de bitllets d'avió basat en l'estudi i anàlisi dels preus al llarg del temps. El sistema permetrà a l'usuari realitzar la compra de bitllets d'avió en el moment idoni en què els preus dels bitllets siguin més baixos. Per poder desenvolupar aquest sistema de recomanació s'ha portat a terme un anàlisi sobre la fluctuació dels preus en dates importants i assenyalades. Per una altra banda, dins de l'anàlisi s'examina l'evolució dels preus al llarg del temps per poder implementar un sistema de predicció de preus per a una certa data i així donar-li a l'usuari una recomanació del millor dia per comprar un bitllet.

Resumen

Este trabajo está dentro del marco de la ciencia de los datos, en el que se desarrolla un sistema de recomendación para la compra de billetes de avión basado en el estudio y análisis de los precios a lo largo del tiempo. El sistema permitirá al usuario realizar la compra de billetes de avión en el momento idóneo en que los precios de los billetes sean más bajos. Para poder desarrollar este sistema de recomendación se ha llevado a cabo un análisis sobre la fluctuación de los precios en fechas importantes y señaladas. Dentro del análisis se examina la evolución de los precios a lo largo del tiempo para poder implementar un sistema de predicción de precios para una cierta fecha y así darle al usuario una recomendación del mejor día para comprar un billete.

Continguts

1	Introducció.....	1
2	Descripció del problema	2
2.1	Objectius.....	2
2.2	Ciència de les dades	3
2.3	Etapas del projecte.....	4
2.3.1	Preguntes plantejades	4
2.3.2	Adquisició de dades	4
2.3.3	Neteja de dades	4
2.3.4	Anàlisi de dades	6
2.3.5	Predicció	6
2.3.6	Avaluació	6
2.4	Dades obtingudes	7
2.5	Preguntes Plantejades.....	9
2.5.1	Amb quants dies d'antelació haig de comprar un bitllet d'avió?	9
2.5.2	Vols d'alta demanda vs. Vols de baixa demanda.....	9
2.5.3	Caps de setmana vs. Dies laborals.....	10
2.5.4	Principi de mes vs. Resta del mes.....	11
2.5.5	Dies de la setmana	11
3	Planificació.....	12
3.1	Tasques	12
3.2	Diagrama de Gannt	13
4	Desenvolupament del projecte	15
4.1	Eines	15
4.1.1	Eines de Suport.....	15
4.1.2	Eines de programació.....	16
4.1.2.1	Python	16
4.1.2.2	Biblioteques per a la manipulació de dades.....	16
4.1.2.3	Biblioteques d'estadística	16
4.1.2.4	Biblioteca de visualització	17
4.1.2.5	Biblioteques de Web Scraping.....	17
4.1.3	Eines d'edició	17
4.2	Tècniques utilitzades.....	18
4.2.1	Tècniques d'obtenció de dades	18

4.2.1.1	QPX Express API	18
4.2.1.2	Web Scraping	19
4.2.2	Tècniques d'anàlisi de dades	21
4.2.2.1	Test estadístic T de Student	21
4.2.3	Tècniques de predicció	22
4.2.3.1	Classificador	22
4.2.3.2	Linear Regresion.....	27
4.2.3.3	Logistic Regression	27
4.2.3.4	Random Forest	29
4.2.3.5	Support Vector Machine	30
4.2.3.6	Mètriques classificació.....	30
5	Experiments i resultats	32
5.1	Preparació prèvia als experiments	32
5.2	Amb quants dies d'antelació haig de comprar un bitllet d'avió?	33
5.3	Vols d'alta demanda vs. Vols de baixa demanda	38
5.4	Caps de setmana vs. Dies laborals	40
5.5	Principi de mes vs. Resta del mes	43
5.6	Dies de la setmana.....	46
5.7	Classificadors	49
5.7.1	Estratègia de validació	49
5.7.2	Resultats de la prova 1	50
5.7.3	Resultats de la prova 2	54
5.7.4	Resultats de la prova 3	57
5.8	Connexió amb l'aplicació	60
6	Conclusions	63
6.1	Treball futur	63
7	Bibliografia	64

1 Introducció

Actualment vivim en un món on una de les principals preocupacions és l'economia, tant a escala estatal, industrial o domèstica. La gran varietat de preus que podem trobar per a un únic producte ha augmentat considerablement amb l'evolució tecnològica que estem vivint, fet que ens convida a fer un exercici constant de cerca i contrast de les diferents fonts de distribució per tal de poder adquirir un producte de bona qualitat al millor preu.

Un exemple clar d'aquest fet el trobem en la reserva de vols. La gran quantitat d'aplicacions relacionades amb la cerca i compra de bitllets d'avió junt amb la gran quantitat de companyies aèries que actualment existeixen, fan costós el fet de poder contrastar informació de manera ràpida i senzilla. Si a això li sumes el fet que els preus dels bitllets d'avió evolucionen i canvien constantment al llarg del temps, les tasques de cerca i contrast es tornen en una feina llarga i costosa que acaba provocant que la gent prengui decisions precipitades, comportant una pèrdua de diners innecessària. Davant aquest problema, ha sorgit la idea de crear una eina que permeti als usuaris interessats en la reserva de vols, realitzar la compra de bitllets al millor preu sense la necessitat d'haver de perdre temps cercant i contrastant informació constantment. D'aquí neix el projecte de final de carrera anomenat: Sistema de recomanació per a la compra de bitllets d'avió.

La finalitat d'aquest projecte és crear una eina que permeti als usuaris cercar els vols que siguin del seu gust per a una determinada data i establir una sèrie d'alertes que utilitzarà el sistema per avisar-lo automàticament quan sigui el moment idoni per a realitzar la compra del bitllet, sense la necessitat imperiosa de què hagi de consultar l'eina constantment.

Un dels recursos principals d'aquest projecte són les dades, ja que són les que ens permetran arribar a les conclusions sobre quines dates són les millors per comprar un cert vol i que ens permetran construir l'eina per als usuaris interessats en la reserva de vols. Les nostres dades han sigut obtingudes diàriament a partir d'un aplicatiu anomenat QPX Express API posat en marxa per Google i posat a disposició de tot programador. Per altra banda, també hem codificat un algoritme de web scraping aplicat a Expedia.

El treball es compondrà de dues parts fonamentals: el desenvolupament d'una eina amb la qual l'usuari pugui interaccionar amb el sistema (de la qual no es parlarà en aquest projecte, ja que es competència d'un altre company) i el desenvolupament de l'estudi estadístic que doni suport a l'aplicació. Per tant, en una primera fase, aquest treball se centra en l'estudi estadístic de les dades obtingudes fent consultes diàries sobre els vols d'anada que surten des de l'aeroport de Barcelona fins a les destinacions de Copenhague, Nova York, Madrid, París, Londres, Palma de Mallorca, Moscú-Domodédovo, Dublín, Munich, Brussel·les i Birmingham. També s'ha treballat en un classificador que ens permeti predir quines són les dates més econòmiques per a realitzar la compra d'un cert vol i ens permeti esbrinar si un cert dia X hem de realitzar la compra d'un bitllet d'avió.

2 Descripció del problema

2.1 Objectius

El principal objectiu d'aquest treball de fi de grau, com ja s'ha explicat, és desenvolupar un sistema intel·ligent que permeti a l'usuari cercar vols per a una data determinada, amb la possibilitat d'establir alertes que seran utilitzades pel sistema per avisar-lo quan el preu del bitllet sigui beneficiós per a l'usuari sense la necessitat de què aquest consulti constantment quins són els preus dels bitllets.

Per a dur a terme aquest projecte, hem passat per 4 etapes:

Etapa 1: En aquesta etapa s'han portat a terme les tasques d'adquisició de dades, anàlisi i estructuració de les dades per tal de poder treballar posteriorment amb elles de forma còmoda. Aquesta etapa és la fase inicial on es deixaran les dades preparades per treballar posteriorment amb elles.

Etapa 2: En aquesta etapa analitzarem les dades aplicant tècniques de ciència de les dades. S'ha dut a terme un estudi estadístic de les dades que tenim aplicant la ciència de les dades per tal d'adquirir informació que es troba amagada darrere de les dades i que no es poden intuir a simple vista. La informació que es pretén adquirir sempre es regeix a les preguntes plantejades durant el desenvolupament del projecte. Per tant, en aquesta segona etapa extraurem aquella informació que ens donen les dades referent a les preguntes plantejades per tal de poder desenvolupar el sistema intel·ligent.

Etapa 3: En aquesta etapa es porta a terme l'anàlisi de les dades mitjançant tècniques d'aprenentatge automàtic. Mitjançant la informació extreta a l'etapa anterior, s'apliquen algoritmes de predicció de dades per a poder predir quin serà el millor moment per a realitzar la compra d'un bitllet d'avió, i així poder desenvolupar un classificador que formarà part de les bases del sistema intel·ligent.

Etapa 4: Aquesta quarta etapa consta del desenvolupament de l'eina que permetrà a l'usuari interactuar amb el sistema intel·ligent. En aquesta etapa es desenvolupa una aplicació per Android que permetrà a l'usuari comunicar-se amb el sistema.

Cal esmentar, que aquest projecte de fi de grau només consta de les etapes 1, 2 i 3. L'etapa 4 és desenvolupada per un altre company i no se'n parlarà en aquest treball.

2.2 Ciència de les dades

La ciència de les dades és el conjunt d'etapes i processos que ens permeten extreure coneixement o un millor enteniment de grans volums de dades mitjançant l'aplicació d'un conjunt de tècniques que formen part de diferents àrees, com ara matemàtiques, estadística, teoria de la informació o tecnologia d'extracció d'informació.

En concret, un projecte de ciència de les dades es pot dividir en les següents fases:

- 1) **Plantejament de preguntes:** En aquesta etapa ens preguntem que és el que volem explorar i si té sentit el que ens estem plantejant.
- 2) **Adquisició de dades:** En aquesta etapa ens plantejem quina pot ser la nostra font d'adquisició de dades, i com les podem obtenir.
- 3) **Definició:** Aquesta etapa es pot dividir en processos:
 - a) **Neteja i estructuració de dades:** En aquest procés es planteja com podem netejar, normalitzar i estructurar les nostres dades.
 - b) **Integració:** En aquest procés ens plantejem com recol·lectar, resumir i visualitzar les dades.
 - c) **Enriquiment:** En aquest procés es planteja com podem recol·lectar més informació per a poder augmentar el volum de les nostres dades
- 4) **Segmentació:** En aquesta etapa es planteja una segmentació de les dades que ens permet trobar grups de dades.
- 5) **Anàlisi de dades:** En aquesta etapa es duu a terme un estudi estadístic per tal de mostrear i modelar les nostres dades per poder esbrinar la significança de cada variable utilitzada posteriorment per establir el model de predicció.
- 6) **Predicció:** En aquesta etapa s'intenta predir que succeirà en un futur, tenint com punt de partida les dades de les quals disposem (Regressions, classificadors, recomanadors).
- 7) **Avaluació:** En aquesta etapa avaluem el nostre sistema per tal de respondre a la pregunta de: Com de segurs estem dels nostres resultats?

2.3 Etapes del projecte

A la secció anterior (secció [2.2](#)), hem vist les etapes i processos d'un projecte complet de ciència de les dades. En aquest projecte hem anat seguint les mateixes etapes excepte els processos de segmentació i enriquiment. Ara explicaré que s'ha realitzat en cada etapa dins del context del projecte.

2.3.1 Preguntes plantejades

La primera etapa consta del plantejament de les preguntes que volem resoldre. Entre els integrants del projecte junt amb el tutor del treball vam plantejar preguntes i vam decidir quines eren aquelles preguntes que trobàvem més interessants de resoldre. Aquesta secció s'explica més detalladament a la secció [2.5](#).

2.3.2 Adquisició de dades

L'adquisició de dades s'ha portat a terme mitjançant dues vies de desenvolupament. En una primera fase es va utilitzar la interfície QPX EXPRESS API de Google i finalment, durant les últimes etapes de desenvolupament del projecte, es va decidir enriquir-lo respecte a continguts codificant un algoritme de Web Scraping.

Cal esmentar que les dades utilitzades pel desenvolupament del projecte únicament provenen de l'API posada a disposició per Google, ja que com he explicat, el desenvolupament de l'algoritme de Web Scraping es va decidir a l'etapa final del desenvolupament per tal d'enriquir els continguts del projecte.

Ambdós algoritmes tenen com a output un fitxer csv on s'emmagatzemen les dades recollides.

2.3.3 Neteja de dades

Com s'ha explicat a la secció anterior, la tècnica d'adquisició consisteix a fer peticions a la interfície QPX EXPRESS API de Google. Aquesta interfície rep la petició en format JSON i retorna la resposta en el mateix format. Prèviament a l'execució de qualsevol procés, apliquem una neteja de les dades que consisteix a aplicar els següents passos:

- **Canvi de format:** Per tal de manipular les dades amb més comoditat, s'extreu aquella informació d'interès de la resposta rebuda en format JSON i s'emmagatzema en un fitxer amb format csv. A la secció [2.4](#) es detalla amb més profunditat l'obtenció i estructuració de les dades.

- **Filtració de vols:** A causa del fet que l'API de Google no ens assegura tenir informació de cada vol per cada dia de consulta, s'aplica un filtre de tal manera que només ens quedem amb aquells vols els quals tenen informats com a mínim un 75% dels dies consultats. Amb aquest procés evitem tenir massa soroll en els càlculs que aplicarem posteriorment.
- **Interpolació de nuls estructurals:** Tot i que al procés anterior realitzem el filtratge descrit, seguim tenint vols dels quals no tenim informació certs dies de consulta. Per aquest motiu, portem a terme una interpolació dels preus dels vols els dies dels quals no disposem d'informació. La interpolació s'aplica de la següent manera:
 - **Cas 1:** Un cert dia X no tenim informació del preu del vol, però si disposem del preu del dia anterior i posterior al dia X. Llavors fem una mitja aritmètica del preu dels dies anterior i posterior i establim la mitja com a preu del dia X.
 - **Cas 2:** Pel fet que volem tenir la mateixa quantitat d'informació (mateixos dies consultats) per a cada vol, com procedim si un vol no apareix el primer dia de consulta o l'últim dia de consulta? Si un vol no te informat el primer dia de consulta establim el seu preu amb el valor del primer dia del qual disposem d'informació. Per altra banda, si no disposem del preu del vol de l'últim dia de consulta, establim el seu preu amb l'últim dia del qual disposem d'informació.
- **Normalització de les dades:** Després de dur a terme la interpolació de les dades, apliquem una normalització d'aquestes per tal d'evitar canvis en la mitja i variància en els preus dels bitllets, ocasionats per la varietat de destinacions de viatges amb les que treballem. Per exemple, si volem comparar si viatjar a Xina és més car que viatjar a Madrid, hem de tenir les dades normalitzades per poder comptar amb la mateixa escala de comparació.

S'ha procedit a normalitzar les dades per cada vol aplicant una normalització de mitja 0 i desviació estàndard 1 on s'aplica la fórmula següent:

$$z = \frac{x - \mu}{\sigma}$$

On μ és el promig dels preus de cada dia consultat per vol, σ és la desviació estàndard per vol. Aquest tipus de normalització és anomenada normalització d'unitat tipificada.

2.3.4 Anàlisi de dades

Aquesta etapa és necessària per a fer la posterior predicció i classificació de les dades. S'ha dut a terme un test estadístic t-Student per a cada pregunta plantejada, de tal manera que amb cada test podrem valorar i decidir quines variables utilitzarem com a predictors en el desenvolupament del model de predicció i del classificador.

2.3.5 Predicció

S'implementarà un classificador que ens indicarà si comprar o no un bitllet d'avió un cert dia determinat en referència a la predicció futura del preu del bitllet partint de la base de les dades amb les quals comptem.

2.3.6 Avaluació

Un cop construït el classificador hem d'avaluar quant de bona és la seva predicció i la seva posterior classificació. Aquesta avaluació s'ha dut a terme de forma quantitativa (mitjançant mètriques) i de forma qualitativa (amb la mostra de casos).

2.4 Dades obtingudes

En aquest apartat s'explicarà la informació més rellevant que podem trobar a les nostres dades. Cal recordar que tenim dues vies d'obtenció de dades (API de Google i Web Scraping) tot i que les dades utilitzades per a desenvolupar el projecte només provenen de l'API de Google. Tot i això, les dades necessàries per a desenvolupar el projecte, obtingudes per ambos processos, són les mateixes.

Ambdues tècniques d'obtenció de dades emmagatzemen la informació extreta de les seves respectives fonts en un fitxer csv.

Les dades obtingudes han sigut enriquidores per tal de poder dur a terme un estudi ampli sobre els preus dels bitllets d'avió, on cada vol està identificat al fitxer csv mitjançant un codi de vol que es pot extreure de la resposta retornada per l'API.

La informació més rellevant que obtenim de cada vol és la següent:

- **Identificador de vol:** Mitjançant l'identificador de vol podrem mantenir informació del preu de cada vol en concret, al llarg del temps (dies de consulta).
- **Preu del vol:** Aquesta dada és essencial per al desenvolupament del projecte i ens permetrà dur a terme l'estudi referent a les preguntes plantejades. La dada en concret fa referència al preu d'un vol un cert dia X.
- **Dia de consulta:** Aquesta dada no s'obté directament per ambdues tècniques, però la calculem i l'emmagatzemem estratègicament per a poder respondre a les preguntes plantejades i poder dur a terme diversos processos com els de neteja de dades, segmentació, anàlisi de dades i predicció. Per evitar confusions cal aclarir que aquesta data fa referència a la data en què es realitza la petició dels vols.
- **Dies d'antelació:** Aquesta dada no s'obté directament per ambdues tècniques, però la calculem i l'emmagatzemem estratègicament per a poder respondre a les preguntes plantejades. Els dies d'antelació són el nombre de dies que hi ha des del dia de la consulta fins al dia de la sortida del vol.
- **Data de sortida:** Aquesta dada és d'especial rellevància, ja que juntament amb el dia de consulta, ens serà molt útil per respondre a les preguntes plantejades. Per evitar confusions cal aclarir que aquesta data fa referència a la data en què s'agafa l'avió des de l'aeroport pertinent.
- **Divisa:** Aquesta dada fa referència a la moneda en la qual és donat el preu del vol. No té rellevància al projecte actual, ja que només treballem amb una única moneda, però cal dir que si tinguéssim un altre tipus de moneda, s'hauria d'aplicar el corresponent canvi de moneda per tal de treballar amb dades homogènies.

- **Demanda:** Aquesta dada no s'obté directament per ambdues tècniques, però la calculem i emmagatzemem estratègicament per a poder respondre a una de les preguntes plantejades (Pregunta 2). Aquesta dada té dos possibles valors 0 o 1. Serà 1 si el vol és considerat d'alta demanda, és a dir, si el vol te com a data de sortida un dia d'especial interès per als usuaris, serà 0 en cas contrari. Aquesta dada serà explicada detalladament a la secció [2.5.2](#), on s'explicarà també com es decideix quin valor pren segons el vol tot incloent exemples.
- **Hora de sortida:** Aquesta dada és d'especial interès, ja que ens ajudarà a respondre a la pregunta 2 juntament amb la dada anterior. Aquesta dada fa referència a l'hora en què s'agafa l'avió des de l'aeroport pertinent.

Cal esmentar que aquestes dades no són les úniques dades emmagatzemades que formen el fitxer csv, però la resta de dades no esmentades no són d'interès per aquesta part del projecte sinó que seran utilitzades a la 4a Etapa del projecte desenvolupada per un altre company.

2.5 Preguntes Plantejades

2.5.1 Amb quants dies d'antelació haig de comprar un bitllet d'avió?

La primera pregunta plantejada durant el desenvolupament del projecte és: *Amb quants dies d'antelació haig de comprar un bitllet d'avió?* A partir de la distribució dels preus de cada vol per cada dia de consulta, podem determinar amb quants dies d'antelació haig de realitzar la compra d'un bitllet d'avió perquè aquest em surti més econòmic? Per esbrinar-ho, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i se'ls hi ha aplicat la corresponent neteja de dades esmentada a la secció [2.3.3](#).

L'experiència ens diu que els preus dels bitllets d'avió fluctuen al llarg del temps, de tal manera que és lícit pensar que durant algun moment al llarg del temps, els preus dels avions estaran més baixos de lo comú. Ara que tenim les dades això ho podem saber, convertint les dades en coneixement.

2.5.2 Vols d'alta demanda vs. Vols de baixa demanda

La segona pregunta plantejada durant el desenvolupament del projecte és: *Existeix diferència entre els dies d'antelació amb els quals comprar un vol considerat d'alta demanda respecte als dies d'antelació amb els quals comprar un vol considerat de baixa demanda?* En plantejar aquesta pregunta, la primera observació que ens ve al cap és: Que considerem com a vol d'alta demanda i vol de baixa demanda? Bé doncs, els vols d'alta demanda són tots aquells vols que per certs motius tendeixen a ser més sol·licitats pels consumidors i els vols de baixa demanda són tota la resta de vols. Després d'explicar que considerem com a vols d'alta demanda i vols de baixa demanda, cal explicar quines premisses hem utilitzat per classificar un vol dins del grup d'alta demanda o baixa demanda. La classificació s'ha fet de la següent manera:

Hem considerat com a vols d'alta demanda:

- **Vols que tenen data de sortida un dilluns al matí:** És lògic pensar que aquests tipus de vols són molt sol·licitats pel fet que molta gent viatja els caps de setmana i una de les dates idònies per retornar d'aquests viatges són els dilluns al matí.
- **Vols que tenen data de sortida un diumenge a la tarda-nit:** Igual que a la premissa anterior, els diumenges a la tarda-nit són un dels moments més idonis per a retornar dels viatges realitzats durant el cap de setmana.
- **Vols que tenen data de sortida un divendres a la tarda-nit:** L'experiència ens diu que la gent que viatja els caps de setmana tendeixen a comprar vols amb data de sortida de divendres a la tarda-nit per tal d'aprofitar més el viatge.

- **Vols que tenen data de sortida 3 dies abans o 1 dia després d'algun festiu o pont:**
L'experiència ens diu que els dies festius o ponts, són els dies més aprofitats per la gent per a dur a terme els seus viatges, però és lògic pensar que la gent no viatja el mateix dia festiu, sinó que ho fa uns dies abans de la data en qüestió. Per un altra banda, també és lògic pensar que al retornar d'aquests tipus de viatges, la gent s'espera a l'endemà del dia festiu per tal de tornar a casa.

Per un altra banda, hem considerat vols de baixa de demanda tots els vols que no compleixen alguna de les premisses explicades anteriorment.

Per concloure aquest apartat, tenint en compte les explicacions darreres, l'objectiu principal de la pregunta és esbrinar si existeix diferència entre els dies d'antelació amb els quals comprar un vol d'alta demanda respecte als dies d'antelació amb els quals comprar un vol considerat com a baixa demanda, per tal que tant uns com els altres ens surtin el més econòmic possible. Per tal de desenvolupar aquesta pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i se'ls hi ha aplicat la corresponent neteja de dades esmentada a la secció [2.3.3](#).

2.5.3 Caps de setmana vs. Dies laborals

La tercera pregunta plantejada durant el desenvolupament del projecte és: *És més car comprar un bitllet d'avió un dia de cap de setmana que un dia laboral?*

En plantejar aquesta pregunta, la primera observació que ens ve al cap és quins dies considerem com a laborals i quins dies considerem com a dies de cap de setmana. Bé doncs, en una primera impressió, ens és fàcil pensar en quins dies considerem com a laborals i quins com a cap de setmana:

Dilluns	Dimarts	Dimecres	Dijous	Divendres	Dissabte	Diumenge
Laboral	Laboral	Laboral	Laboral	?	Cap de setmana	Cap de setmana

Però ens trobem amb un dilema: Que fem amb els divendres i els dies festius? El divendres són dies laborals però que per motius de proximitat al cap de setmana es tendeix a considerar com a cap de setmana just quan s'acaba la jornada de treball. El mateix passa amb els dies festius, un dilluns que normalment és un dia laboral, pot ser que sigui festiu, però el considerem com a laboral? Pel fet que no hem sabut a quin grup afegir els divendres i els festius, hem optat per no tenir-los en compte a l'hora de resoldre la pregunta.

Per concloure aquest apartat, tenint en compte les explicacions darreres, l'objectiu principal de la pregunta és esbrinar si és més car comprar un bitllet d'avió un dia de cap de setmana que un dia laboral. El principal motiu que hem trobat per explicar la pregunta,

és el fet de pensar que pot haver-hi una tendència a comprar els caps de setmana, ja que és quan la gent té més temps per cercar i comprar els bitllets, ja que la resta de dies es treballa. Per tal de desenvolupar aquesta pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i se'ls hi ha aplicat la corresponent neteja de dades esmentada a la secció [2.3.3](#).

2.5.4 Principi de mes vs. Resta del mes

La quarta pregunta plantejada durant el desenvolupament del projecte és: *És més car comprar un bitllet a principi de mes (dies d'ingrés de nòmina) que la resta de dies?* Aquesta pregunta té com a objectiu esbrinar si existeix una tendència en el preu dels bitllets que ens digui si és més car comprar un bitllet d'avió a principis de mes que la resta de dies del mes. El principal motiu que hem trobat per explicar la pregunta, és que a principi de mes la gent rep l'ingrés de la seva nòmina, motiu que pot fer que els preus dels bitllets siguin més cars a principis de mes que la resta de dies, ja que la resta de dies la gent ja va més escurada de diners.

Aquesta pregunta ens fa pensar en quins dies considerem com a principi de mes (dies d'ingrés de nòmina) i quins dies considerem com a resta de dies.

- **Dies principi de mes (d'ingrés de nòmina):** Com a dies de cobrament de nòmina s'ha considerat des de l'última setmana del mes fins a la primera setmana del mes següent, ja que hi ha persones que reben la nòmina a l'última setmana del mes.
- **Resta del mes:** Com la seva pròpia descripció indica, prenem la resta de dies com a segon grup de dades.

Per tal de desenvolupar aquesta pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa, s'han eliminat les últimes dues setmanes de consulta, ja que coincideixen amb les dates de sortida dels vols i per tant, si no les eliminéssim, tindríem molt de soroll i se'ls hi ha aplicat la corresponent neteja de dades esmentada a la secció [2.3.3](#).

2.5.5 Dies de la setmana

La cinquena i última pregunta plantejada durant el desenvolupament del projecte és: *Quin dia de la setmana és millor realitzar la compra d'un bitllet d'avió?* Aquesta pregunta té com a objectiu esbrinar si els preus dels bitllets són més alts depenent del dia en què es realitzi la compra del bitllet. L'experiència ens diu que poden haver-hi certs dies on la compra de bitllets augmenti, ja sigui perquè la gent té més temps, va menys estressada, etc. però realment existeix aquesta diferència de preus entre dies de la setmana?

A partir de la distribució dels preus de cada vol per cada dia de consulta, podem determinar si és més econòmic realitzar la compra d'un bitllet d'avió segons el dia de la setmana? Per esbrinar-ho, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i se'ls hi ha aplicat la corresponent neteja de dades esmentada a la secció [2.3.3](#). Posteriorment, han estat agrupats segons el dia de la setmana.

3 Planificació

3.1 Tasques

Les tasques d'aquest projecte estan dividides de la següent manera:

- Formació
- Plantejament de preguntes
- Recollida de dades
- Neteja de dades
- Anàlisi de dades
- Predicció
- Avaluació
- Documentació

A continuació mostrarem com aquestes tasques han estat repartides al llarg del temps, mitjançant un diagrama de Gannt. També mostrarem una avaluació econòmica del projecte.

3.2 Diagrama de Gannt

S'han construït dos diagrames de Gannt, un a partir de la planificació inicial [Figura 1] i l'altre amb la planificació real [Figura 2] per tal de veure les diferències.

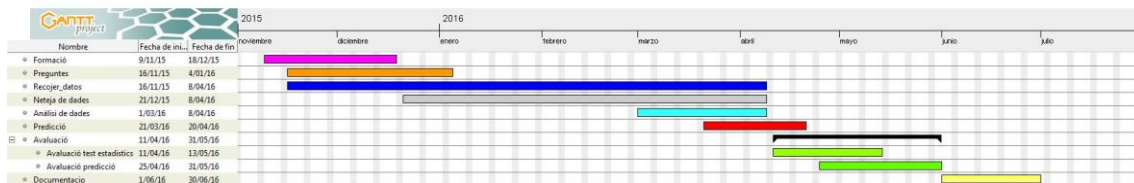


Figura 1: Planificació Inicial

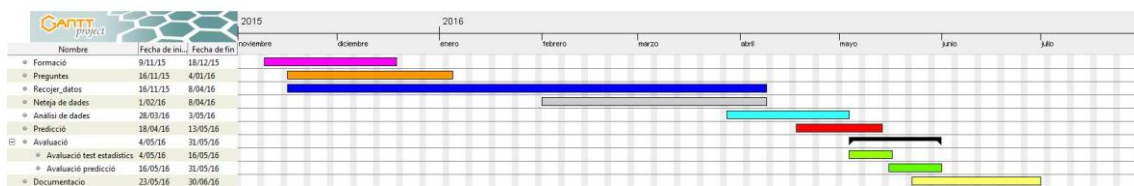


Figura 2: Planificació real

S'observa que en general s'ha pogut seguir correctament la planificació inicial, tot i que en el cas de les tasques de tests estadístics, predicció i avaluació s'ha prolongat el seu començament unes setmanes més tard, pel fet que durant la neteja de dades han sorgit alguns inconvenients que ens han endarrerit.

Cal esmentar, que la neteja de dades s'ha iniciat més tard, pel fet que s'ha cregut oportú realitzar aquesta tasca quan es tingués un volum de dades més substancials.

Finalment, es pot observar que les tasques d'avaluació no han requerit tant temps com estava previst i que la tasca de documentació s'ha iniciat abans de la previsió feta inicialment.

El Treball de Fi de Grau equival a 18 crèdits ECTS. Si cada crèdit equival a 25 hores de treball, llavors tenim que:

$$18 \text{ crèdits} \cdot \frac{25 \text{ hores}}{1 \text{ crèdit}} = 450 \text{ hores}$$

Per tant , totes les tasques s'han de realitzar en 450 hores. Les hores dedicades a cada tasca han sigut les següents:

	Formació	Preguntes	Recollida de Dades	Neteja de Dades	Anàlisi de dades	Predicció	Avaluació	Documentació
Hores	25	25	50	50	100	75	75	50

Taula 1: Hores de dedicació per tasca

4 Desenvolupament del projecte

En aquesta secció explicarem quines són i en què consisteixen totes les eines i tècniques utilitzades per al desenvolupament del projecte.

4.1 Eines

4.1.1 Eines de Suport

Aquestes són les eines de suport que ens han ajudat al llarg del treball per tal de fer més còmode la seva organització tan personal com per equip.

Github

GitHub és una plataforma online per desenvolupar projectes software de forma col·laborativa. Aquesta plataforma utilitza un control de versions anomenat Git. La finalitat de GitHub és l'emmagatzemament massiu de projectes amb la possibilitat d'escollir si el codi font es presenta obert o no. Per això hem optat per la utilització de GitHub, ja que durant la fase de desenvolupament volem que el nostre codi no el pugui veure tothom, fet que ha estat possible fent una petició als administradors de Github per tal que ens obrissin un repositori privat. Amb Github també hem pogut portar al dia un llistat de tasques o issues que hem fet servir per tenir present una planificació.

Microsoft OneNote

Microsoft OneNote és un programa amb llicència gratuïta desenvolupat per Microsoft per facilitar la presa de notes, la recopilació d'informació, i la col·laboració multiusuari. OneNote permet col·locar notes (introduïdes a mà o amb teclat) i ofereix la possibilitat d'agregar dibuixos, diagrames, fotografies, elements multimèdia, àudio, vídeo, i imatges escanejades. Per aquests motius hem optat per la utilització de Microsoft OneNote, ja que mitjançant aquest programa hem pogut establir les tasques a realitzar eventualment i hem pogut compartir de forma ràpida les gràfiques i arxius relacionats amb el desenvolupament del projecte.

Telegram

Per últim, un altre de les eines de suport que hem fet servir ha estat Telegram, un servei de missatgeria instantània que permet una comunicació més clara entre els membres d'un projecte. Amb Telegram hem pogut comunicar-nos de forma fluida aquells dies en els quals se'ns feia impossible comunicar-nos físicament.

4.1.2 Eines de programació

En aquesta secció trobarem el conjunt de llenguatges de programació, el conjunt de biblioteques informàtiques, la base de dades que hem utilitzat i la interfície que ens ha permès realitzar les consultes adients i rebre les respostes sobre els vols d'interès.

4.1.2.1 Python

Python és un llenguatge d'alt nivell interpretat, de codi obert i desenvolupat per Python Software Foundation, una organització sense ànim de lucre. Vam escollir Python per diversos motius: per ser un llenguatge de scripting, per comptar amb diverses biblioteques preprogramades relacionades amb el tractament de dades (com Pandas, NumPy, Scikit-learn) que han permès desenvolupar el projecte de forma més eficaç i productiva i perquè la lectura dels seus codis és molt més senzilla, ja que la seva sintaxi s'aproxima molt al pseudocodi.

4.1.2.2 Biblioteques per a la manipulació de dades

En aquesta secció presentem totes les llibreries utilitzades per a la manipulació, anàlisi i gestió de dades.

Pandas

Pandas és una biblioteca preprogramada i escrita en Python destinada a la manipulació, anàlisi i gestió de grans quantitats de dades. És molt útil per dur a terme el tractament de taules que poden contenir grans quantitats de registres i per a realitzar l'anàlisi i la gestió de la informació continguda en aquestes taules.

Pandas conté tot un conjunt de funcions matemàtiques d'alt nivell per treballar còmodament amb aquestes taules.

Numpy

Numpy és una biblioteca preprogramada i escrita en Python per operar amb vectors i matrius d'una forma més extensa i senzilla de la que et permet el llenguatge Python. Numpy conté tot un conjunt de funcions matemàtiques d'alt nivell per treballar amb aquests vectors i matrius.

4.1.2.3 Biblioteques d'estadística

En aquesta secció presentem totes les llibreries utilitzades per a realitzar la part estadística del treball.

Scikit-learn

Scikit-learn és una biblioteca preprogramada i escrita en Python orientada a l'aprenentatge automàtic. Té suport per algoritmes de classificació, regressió i clusterització i està dissenyat per interoperar amb les dues llibreries esmentades anteriorment. Per aquest projecte hem fet servir regressors.

En la secció [4.2: tècniques utilitzades](#), es detalla cada tècnica que hem utilitzat d'aquesta llibreria.

4.1.2.4 Biblioteca de visualització

En aquesta secció presentem la llibreria utilitzada per a la visualització de dades.

Matplotlib

Matplotlib és una biblioteca preprogramada i escrita en Python per a la generació de gràfics a partir de dades contingudes en llistes, taules o vectors que permet la interoperabilitat amb una de les llibreries esmentades anteriorment com és NumPy. Proporciona una API, pyplot, mitjançant la qual generem les gràfiques necessàries per a dur a terme l'anàlisi i l'estudi estadístic.

4.1.2.5 Biblioteques de Web Scraping

En aquesta secció presentem les llibreries utilitzades en la codificació de l'algoritme de Web Scraping

Selenium

Selenium és un programari que genera scripts en llenguatge Selanese, un llenguatge de scripting especial per a Selenium. Selanese proveeix comandaments que executen accions sobre objectes al navegador, tals com fer clic a un enllaç, seleccionar d'una llista d'opcions, verificar la presència d'un text en particular així com obtenir la totalitat de la pàgina web. Mitjançant Selenium hem pogut portar a terme web scraping sobre la pàgina de Expedia.

Beautiful Soap

Beautiful Soap és una llibreria de Python per a parsejar documents HTML. Aquesta biblioteca crea un arbre amb tots els elements del document HTML retornat per Selenium per així poder extreure la informació d'interès de la pàgina web.

4.1.3 Eines d'edició

IPython Notebook

IPython notebook és un editor per a l'entorn de Python. El fet d'utilitzar notebook resideix en el fet que es pot tenir un codi molt més llegible i a més tenir explicacions d'allò que s'està programant, ja que es pot barrejar codi, la sortida del codi, markdown i HTML.

Hem optat per escollir aquest entorn d'edició, ja que en un projecte de ciència de dades hem de visualitzar feedback constant per poder comentar els resultats obtinguts a cada pas dels algorismes.

JSON

JSON és un format de text lleuger per a l'intercanvi de dades. Un dels suposats avantatges de JSON sobre XML com a format d'intercanvi de dades és que és molt més senzill escriure un analitzador sintàctic (parser) de JSON.

Hem utilitzat el format JSON, ja que l'API de Google esmentada anteriorment rep les peticions en forma d'objecte JSON i retorna la resposta d'igual forma.

4.2 Tècniques utilitzades

A continuació es presenten les tècniques i mètriques d'obtenció de dades, anàlisi de dades i predicció utilitzades en aquest projecte. La finalitat és proporcionar un coneixement per poder entendre la secció d'Experiments i resultats.

4.2.1 Tècniques d'obtenció de dades

A continuació presentem les dues tècniques implementades al projecte per portar a terme l'obtenció de dades: QPX Express API i web scraping. Tot i que ambdues tècniques estan codificades al projecte i són completament funcionals, només hem utilitzat les dades obtingudes a partir de la primera tècnica (QPX Express API) per tal de desenvolupar el projecte, ja que la segona tècnica (web scraping) es va decidir codificar al període final del projecte com a enriquiment d'aquest.

4.2.1.1 QPX Express API

QPX Express és una nova API desenvolupada per Google per a la cerca de tarifes aèries que ens ofereix capacitats bàsiques de compra i facilitat en l'accés a la cerca de viatges aeris mitjançant una sola API. L'API ens permetrà realitzar les peticions i consultes dels vols d'interès que posteriorment utilitzarem per a desenvolupar el nostre projecte.

Aquesta API forma part de la suite d'API's de Google, disponible de cara al públic a través de Google Developers Console, dissenyada per a una fàcil integració, obtenir alta fiabilitat i resposta ràpida. QPX Express fa una cerca a través de les innumerables combinacions de vols, tarifes i normes tributàries per tal de donar una resposta en temps real.

La interfície està dissenyada per a ser utilitzada tant per les empreses intermediàries a la venda de bitllets d'avió, com per a qualsevol desenvolupador que vulgui disposar de la informació proporcionada per l'API per tal de desenvolupar un projecte propi.

Aquesta interfície és un API sense estat, s'executa mitjançant HTTP i tant la petició com la resposta es codifica en format JSON. L'API ens permet realitzar peticions establint una sèrie de paràmetres de cerca, tals com el recompte de passatgers, l'origen del vol, el destí del vol i les dates d'anada i tornada, així com altres paràmetres de configuració de la cerca. Quan la petició és enviada en format JSON, es retorna la resposta en el mateix format, la qual té una estructura que posteriorment serà parsejada per nosaltres per tal d'extreure la informació rellevant que utilitzarem per desenvolupar el projecte.

Una de les principals peculiaritats és que només permet al voltant d'unes 50 peticions per dia de forma gratuïta i que per altra banda, per poder realitzar aquestes consultes necessites una clau d'API per usuari, fet que ens ha limitat en certa manera el desenvolupament del projecte.

4.2.1.2 Web Scraping

Web scraping és una tècnica utilitzada mitjançant programari software per a l'extracció d'informació de pàgines web. Usualment aquests programaris simulen la navegació d'un humà al lloc web ja sigui utilitzant el protocol HTTP manualment o incrustant un navegador a l'aplicació.

El web scraping està molt relacionat amb la indexació de la pàgina web, la qual indexa la informació utilitzant un robot i és una tècnica universal adoptada per la majoria de motors de cerca. De totes maneres, el web scraping es focalitza més en la transformació de dades sense estructura web (com el format HTML). Per un altra banda, el terme web scraping també simula la navegació d'un humà en una pàgina web.

El navegador, browser, és l'eina que habitualment utilitza una persona per accedir als llocs web, moure's entre les seves pàgines, consultar els continguts, intercanviar informació i en definitiva interactuar amb el servidor.

El servidor web rep les peticions del client (navegador), les gestiona i retorna un flux de dades que el navegador és capaç d'entendre, processar, i representar. El protocol HTTP, Hypertext Transfer Protocol, és el que es fa servir en les transaccions web i defineix la sintaxi i la semàntica que s'utilitza en l'entorn web. És a dir, és la manera en què es comuniquen els clients i servidors web. Es tracta d'un protocol orientat a transaccions que segueix el patró petició-resposta.

El codi HTML, HyperText Markup Language, o llenguatge de marques, és la codificació que es fa servir per a la confecció de pàgines web. Fa servir etiquetes o marques per definir les diferents parts de l'estructura.

El XML, eXtensible Mark Language, és un metallenguatge que serveix per definir altres llenguatges i que es pot aplicar a diferents entorns, no només a Internet. És molt utilitzat com estàndard per a intercanviar dades entre programes i sistemes diferents d'una manera senzilla i fiable.

Per a dur a terme Web Scraping, podem utilitzar eines que ens faciliten la feina i que ens permeten extreure i analitzar tot el codi que és intercanviat entre client i servidor. En el context del projecte utilitzarem Selenium, que és un programari que genera scripts en llenguatge Selanese, un llenguatge de scripting especial per a Selenium. Selanese proveeix comandaments que executen accions sobre objectes al navegador, tals com fer clic a un enllaç, seleccionar d'una llista d'opcions, verificar la presència d'un text en particular així com obtenir la totalitat de la pàgina web.

Juntament amb Selenium, podem utilitzar la llibreria de Python anomenada BeautifulSoup, que és una llibreria de Python per a parsejar documents HTML. Aquesta biblioteca crea un arbre amb tots els elements del document HTML retornat per Selenium per així poder extreure la informació d'interès de la pàgina web.

Cal esmentar que no totes les pàgines web permeten ser "raspades", ja que a vegades estan protegides contra els robots de cerca, o també perquè contenen codi o imatges de difícil interpretació. Cada servidor funciona d'una manera particular, és a dir, cada un disposa d'una estructura determinada que utilitza per intercanviar les dades.

Una de les dificultats del web scraping és analitzar aquest flux de dades i ser capaçs d'automatitzar els mecanismes de cerca per aconseguir extreure les dades que ens interessin, en aquest cas les relatives als bitllets d'avió.

La pàgina web que hem escollit per realitzar web scraping és Expedia. Expedia és una pàgina de venda de bitllets d'avió, reserva d'hotels i altres productes. Aquesta pàgina web té com a plana d'inici, una plana on pots escollir el producte d'interès i seguidament introduir els paràmetres de cerca. Un cop s'han introduït els paràmetres de cerca, se'ns presenta una plana de resultats en forma de llista com podem veure a la Figura 3.

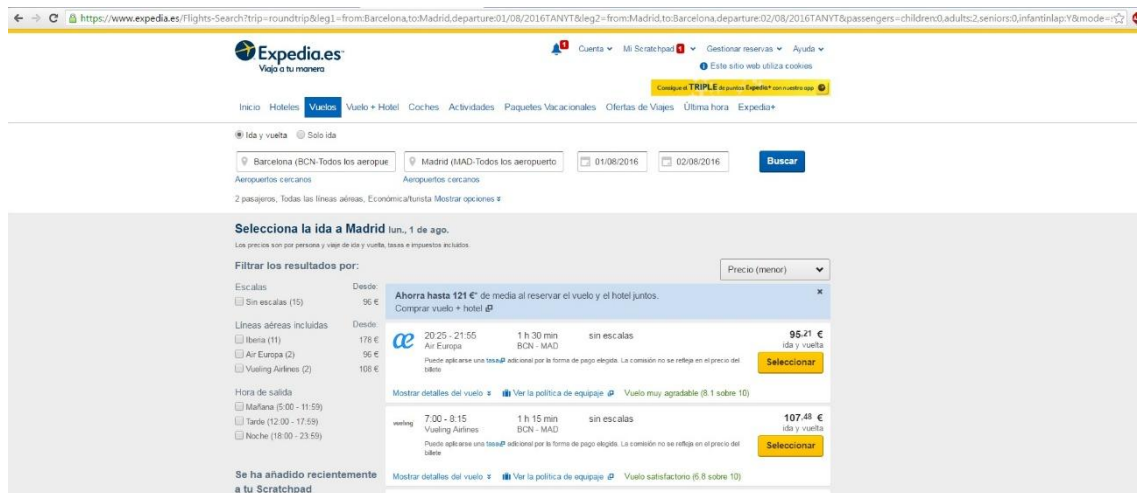


Figura 3: Plana de resultats de Expedia

La implementació de web scraping s'ha codificat directament sobre aquesta plana, de tal manera que hem agafat directament la URL que ens redirigeix a aquesta plana establint els paràmetres de cerca a la pròpia URL i enviant la petició al servidor mitjançant Selenium. Un cop hem obtingut el codi de tota la pàgina web, la parsejem utilitzant BeautifulSoup per extreure tots els vols que formen el llista de resultats que es pot observar a la Figura 3 i que posteriorment són emmagatzemats en un arxiu csv.

4.2.2 Tècniques d'anàlisi de dades

4.2.2.1 Test estadístic T de Student

En estadística, una prova t de Student és qualsevol prova en què l'estadístic utilitzat té una distribució t de Student. La distribució t de Student és una distribució de probabilitat que sorgeix del problema d'estimar la mitja d'una població normalment distribuïda si la mida de la mostra és petita.

Entre els usos més freqüents trobem el de la hipòtesi nul·la. Mitjançant el test d'hipòtesis nul·la, s'intenta demostrar que la diferència entre dues mostres mesurades en les mateixes unitats estadístiques és zero. Per tal de plantejar un test d'hipòtesis nul·la, es determina la hipòtesi que es desitgi contrastar o hipòtesi nul·la H_0 , que és la hipòtesi que indica que no hi ha diferència. Aquesta H_0 representa la hipòtesi que mantindrem a no ser que les dades ens indiquin la seva falsedat. Per un altra banda, s'ha de determinar també la hipòtesi alternativa H_1 , que representa la hipòtesi que desitgem poder concloure com a certa. Un cop establertes tant hipòtesi nul·la com hipòtesi alternativa, s'ha d'establir el nivell de significança (α), que es defineix com l'error màxim adoptat a l'hora de rebutjar la hipòtesi nul·la quan és certa.

Després de realitzar la prova obtenim el p-valor que es defineix com la probabilitat d'obtenir un resultat almenys tan extrem com el que realment s'ha obtingut suposant que la hipòtesi nul·la és certa, en termes de probabilitat condicional.

$$\text{p-valor} = \text{Probabilitat}(\text{dades}|\text{hipòtesi nul·la}) = P(\text{dades}|H_0)$$

Per prendre la decisió de quina hipòtesi s'accepta i quina es rebutja, se segueix la següent regla:

si $\text{p-valor} \leq \alpha$, acceptar H_1 i rebutjar H_0

si $\text{p-valor} > \alpha$, acceptar H_0 i rebutjar H_1

4.2.3 Tècniques de predicció

La predicció és important en un projecte de ciència de les dades, ja que ens permet predir el futur estadístic d'acord amb les observacions que tenim. En aquest apartat explicaré les diverses tècniques de predicció que hem utilitzat en aquest treball per tal de desenvolupar un classificador que ens digui si un cert dia haig de realitzar o no la compra d'un cert bitllet d'avió.

4.2.3.1 Classificador

Un Classificador és un algoritme d'aprenentatge automàtic que assigna un element d'entrada no etiquetat a una categoria concreta coneguda, on les mostres d'entrenament són parelles de dades, la primera representada normalment per un vector i la segona per un valor numèric que identifica la categoria a la qual pertany la primera. A la primera dada se l'anomena vector característic i a la segona etiqueta. El vector característic està format per aquelles propietats o característiques d'un cert objecte, que en el context del projecte, serien les propietats dels vols. Així si disposem d'un conjunt de n mostres i cada vector característic té dimensió p , podem construir la matriu característica i el vector d'etiquetes, que representen tots els vectors característics i totes les etiquetes respectivament.

$$\text{Matriu característica: } X = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

$$\text{Vector d'etiquetes: } Y = [y_1 \quad \cdots \quad y_n]$$

La mesura més immediata de l'eficàcia del nostre classificador és l'*exactitud*, que mesura la proporció de mostres classificades correctament.

$$e = \frac{\text{nombre de classificacions correctes}}{\text{nombre de mostres}}$$

No obstant hi ha casos en què aquesta mesura de classificació no és suficient per validar una bona classificació, com per exemple en els casos en què el nombre de mostres de les diferents classes o etiquetes no és similar.

En el cas d'un problema binari, és a dir, que el vector d'etiquetes Y només estigui format per dues possibles etiquetes $Y = \{0, 1\}$, podem definir una nova mètrica a partir del que anomenem matriu de confusió, que es construeix a partir de les quatre combinacions lògiques que el classificador ens aporta:

- **Vertader positiu (TP):** Quan el classificador prediu una mostra com a positiva i és realment positiva.
- **Fals positiu (FP):** Quan el classificador prediu una mostra com a positiva i és realment negativa.
- **Vertader negatiu (TN):** Quan el classificador prediu una mostra com a negativa i és realment negativa.

- **Fals negatiu (FN):** Quan el classificador prediu una mostra com a negativa i és realment positiva.

La matriu de confusió es defineix de la següent manera:

$$MC = \begin{bmatrix} TP & FP \\ TN & FN \end{bmatrix}$$

Observem com podem recuperar la mètrica d'exactitud descrita anteriorment de la següent manera:

$$e = \frac{TP + TN}{TP + FP + TN + FN}$$

A més a més, la combinació d'aquestes també ens permet definir noves mètriques, entre d'altres trobem les següents:

Sensibilitat: fa referència a l'eficiència de la classificació de tots els elements d'una classe, dit d'una altra manera, mesura la capacitat que té el classificador per detectar petits canvis en les dades.

$$sensibilitat = \frac{TP}{positivos reales} = \frac{TP}{TP + FN}$$

Precisió: fa referència a la qualitat de la resposta del classificador

$$precisió = \frac{TP}{positius predits} = \frac{TP}{TP + FP}$$

Especificitat: representa la probabilitat que un mostra positiva tingui un resultat negatiu.

$$especificitat = \frac{TN}{negatius reals} = \frac{TN}{TN + FP}$$

L'eficàcia i fiabilitat del nostre classificador dependrà de la quantitat de falsos positius i negatius que obtinguem a l'etapa d'entrenament i del nombre de mostres utilitzades. El que fem és subdividir les mostres en dos conjunts, el d'entrenament i el de test. Mitjançant el conjunt d'entrenament podem ajustar el nostre model de predicció i educar al nostre classificador. Seguidament testem aquest model amb les mostres del conjunt de test i l'avaluem seguint un conjunt de mètriques de predicció.

Un cop definit el concepte de classificador, es passaran a explicar els classificadors construïts en el context del projecte, quines són les característiques que formen la matriu de característiques, com estan definides i quins valors poden prendre. Per una altra banda, també s'explicarà quines etiquetes formen part del vector d'etiquetes i com estan definides.

Els classificadors definits en el context del projecte estan formats per una matriu de característiques on cada vector característic que la forma, estaria representat per les característiques definides de cada vol. Primer, hem definit un classificador on cada vol estaria definit per 4 característiques, i després hem definit un segon classificador

introduint una cinquena característica per tal de comparar els resultats de classificació i veure si obtenim una millor predicció. Les característiques definides són les següents:

- **Dies d'antelació:** Aquesta característica es defineix com el nombre de dies que hi ha des del dia de la consulta fins al dia de la sortida del vol, dit d'una altra manera, defineix el nombre de dies entre la suposada compra del bitllet fins al dia de sortida del vol. El valor d'aquesta característica oscil·la dels 95 als 0 dies d'antelació.
- **Demanda:** Com ja s'ha definit a la secció [2.5.2](#), els vols d'alta demanda són tots aquells vols que per certs motius tendeixen a ser més sol·licitats pels consumidors i els vols de baixa demanda són tota la resta de vols. Per tant, aquesta característica podrà prendre dos possibles valors: 1 si el vol és considerat d'alta demanda i 0 si el vol és considerat de baixa demanda. En concret, aquesta característica prendrà valor 1 en els següents casos:
 - a) **Vols que tenen data de sortida un dilluns al matí:** És lògic pensar que aquests tipus de vols són molt sol·licitats pel fet que molta gent viatja els caps de setmana i una de les dates idònies per retornar d'aquests viatges són els dilluns al matí.
 - b) **Vols que tenen data de sortida un diumenge a la tarda-nit:** Igual que a la premissa anterior, els diumenges a la tarda-nit són un dels moments més idonis per a retornar dels viatges realitzats durant el cap de setmana.
 - c) **Vols que tenen data de sortida un divendres a la tarda-nit:** L'experiència ens diu que la gent que viatja els caps de setmana tendeixen a comprar vols amb data de sortida de divendres a la tarda-nit per tal d'aprofitar més el viatge.
 - d) **Vols que tenen data de sortida 3 dies abans o 1 dia després d'algun festiu o pont:** L'experiència ens diu que els dies festius o ponts, són els dies més aprofitats per la gent per a dur a terme els seus viatges, però és lògic pensar que la gent no viatge el mateix dia festiu, sinó que ho fa uns dies abans de la data en qüestió. Per una altra banda, també és lògic pensar que al retornar d'aquests tipus de viatges, la gent s'espera a l'endemà del dia festiu per tal de tornar a casa.

Per una altra banda, prendrà valor 0 en aquells vols que no compleixen algunes de les premisses exposades.

- **Dia de la setmana:** Aquesta característica es defineix com el número de dia de la setmana en què es realitza la consulta, és a dir, el dia de la setmana en què suposadament es realitza la compra del bitllet d'avió. Per tant, aquesta característica prendria valors entre 1 i 7. Un exemple seria el següent: Suposem que es realitza la compra d'un bitllet d'avió un dilluns, llavors aquesta característica prendria valor 1. Per una altra banda, si es realitzés un dimarts, prendria valor 2 i de la mateixa manera per a la resta de dies.
- **Cap de setmana:** Aquesta característica prendrà valor 0 o 1 si el dia en què es realitza la consulta, és a dir, el dia que es compra el bitllet d'avió, és un dia de cap de setmana o un dia laboral.

- **Preu:** Aquesta característica només serà utilitzada pel segon classificador i prendrà el valor normalitzat del preu del bitllet comprat un cert dia X. La normalització aplicada en aquest cas, es realitza agafant el preu de tots els vols consultats des de l'origen de les dades fins al dia de compra del bitllet (dia X).

Un cop definides les 4 característiques que formaran la matriu de característiques del primer classificador i la cinquena característica introduïda al segon classificador, passarem a explicar quines etiquetes formen el vector d'etiquetes.

El vector d'etiquetes està format per dos possibles etiquetes:

- **Comprar:** Aquesta etiqueta defineix que hem de comprar el bitllet d'un cert vol i prendrà valor 1. Per tal d'entendre com es decideix si un cert bitllet d'avió es compra, hem d'explicar quin és el raonament fet per prendre la decisió. Cal recordar que l'objectiu del projecte és recomanar-li a l'usuari la compra d'un bitllet d'avió en el moment en què preu del bitllet és més beneficiós per ell. Per explicar-ho ens recolzarem sobre la figura 4 presentada a continuació:

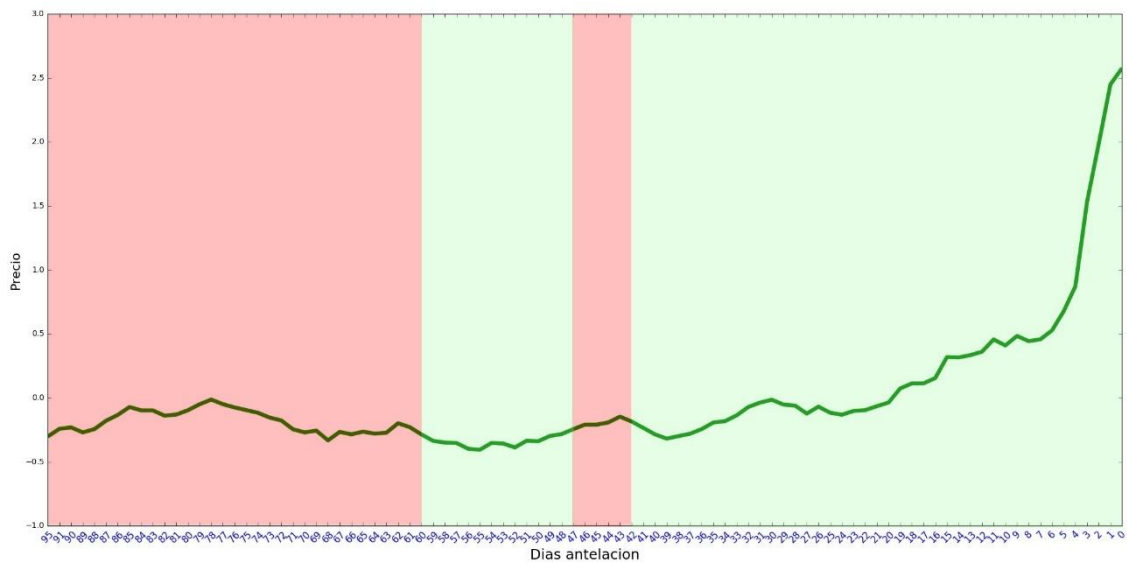


Figura 4: En aquesta figura es mostra l'evolució del preu d'un vol amb data de sortida durant els dies de setmana santa respecte als dies que queden fins a la sortida del vol. Les zones vermelles indiquen els dies que no s'ha de comprar el bitllet d'avió i les zones verdes mostren els dies que s'ha de comprar el bitllet d'avió.

Com es pot observar a la figura 4, el preu del bitllet no és constant a mesura que ens apropem al dia de sortida del vol. Per tant, es considerarà que hem de comprar un bitllet d'avió un cert dia X, si no existeix cap dia entre el dia X i el dia de sortida del vol (dia d'antelació 0) en què el preu del bitllet hagi baixat com a mínim un 5% del preu del bitllet del dia X. Per tal d'aclarir-ho més, posem un exemple: basant-nos en la figura 4, suposem que volem realitzar la compra del bitllet d'avió els 7 dies (dia d'antelació 7) abans que surti el vol (dia d'antelació 0). Pel fet que el preu del bitllet d'avió durant els dies d'antelació 6, 5, 4, 3, 2, 1 i 0 no ha baixat d'un 5% del preu del bitllet del dia d'antelació 7, s'establirà que el dia 7 s'ha de realitzar la compra del bitllet d'avió, ja que sinó ens sortirà més car.

- **No comprar:** Aquesta etiqueta defineix que no hem de comprar el bitllet d'avió per a un cert vol i prendrà valor 0. Igual que amb l'etiqueta anterior, per tal d'entendre com es decideix si un cert bitllet d'avió no es compra, hem d'explicar quin és el raonament fet per prendre la decisió. Per explicar-ho ens tornarem a recolzar sobre la figura 4.

Es considerarà que no hem de comprar un bitllet d'avió un cert dia X, si existeix algun dia entre el dia X i el dia de sortida del vol (dia d'antelació 0) en què el preu del bitllet hagi baixat com a mínim un 5% del preu del bitllet del dia X. Per tal d'aclarir-ho més, posem un exemple: basant-nos en la figura 4, suposem que volem realitzar la compra del bitllet d'avió els 62 dies (dia d'antelació 62) abans de què surti el vol (dia d'antelació 0). Pel fet que el preu del bitllet d'avió baixa d'un 5% per algun dia entre els dies 62 i 0, s'establirà que el dia 62 no s'ha de realitzar la compra del bitllet d'avió, ja que si ens esperem, ens sortirà més barat.

Un cop explicades tant la matriu de característiques com el vector d'etiquetes, passarem a explicar els diferents models de predicció implementats per ambdós classificadors. En concret, els dos classificadors definits estan implementats sobre quatre models de predicció diferents, que són els següents:

- Classificador amb *Linear Regresion*
- Classificador amb *Logistic Regresion*
- Classificador amb *Random Forest*
- Classificador amb *Suport Vector Machine*

Els 4 models de predicció s'explicaran en els apartats que vénen a continuació, juntament amb les mètriques de predicció i classificació que s'han utilitzat per avaluar cadascun d'aquests classificadors.

4.2.3.2 Linear Regresion

Un regressor lineal modelitza una recta de regressió a partir d'un núvol de punts. La recta modelada, és la recta més propera que passa per tots els punts. Aquesta recta, defineix la relació entra una variable dependent Y , un conjunt de variables independents X_i i un terme aleatori ε , és a dir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

on:

Y : variable dependent

X_i : variables independents

B_i : paràmetres que mesuren la influència que tenen les variables independents sobre la variable dependent.

En el cas de tenir una regressió lineal amb una variable dependent i una variable independent, tindríem un resultat semblant al de la Figura 4.

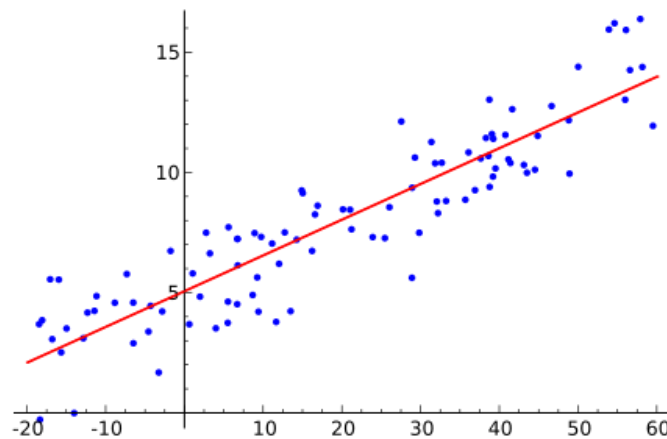


Figura 5: Regressió lineal

4.2.3.3 Logistic Regression

El model de regressió logística permet estimar la probabilitat d'un succés que depèn del valor de certes variables, permeten modelar la probabilitat de què una certa variable Y pertanyi a una certa categoria.

Per exemple, es suposa que un succés d'interès A pot presentar-se o no en cadascun dels individus d'una certa població. Llavors, considerem una variable binària y que pren els valors:

$$y = 1 \text{ si } A \text{ es presenta, } y = 0 \text{ si } A \text{ no es presenta}$$

Si la probabilitat d'A no depèn d'altres variables, sent $P(A) = p$, tenim que la versemblança d'una única observació y és:

$$L = p^y(1 - p)^{1-y}$$

on $L = p$ si $y = 1$, $L = 1 - p$ si $y=0$.

Per una altra banda, si la probabilitat p depèn dels valors de certes variables $X_1, \dots, X_n$ on $x = (X_1, \dots, X_n)$ son les observacions d'un cert individu ω , llavors la probabilitat d'esdevenir A donat x és $p(y = 1|x) = p(x)$ i la probabilitat de què no succeeixi A donat x serà $p(y = 0|x) = 1 - p(x)$.

Per tal d'establir la transformació logística de la probabilitat, se suposa un model lineal, d'on obtenim que:

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}}$$

La regla de decisió logística ens diu si l'individu ω posseeix la característica A o no. Aquesta regla és la següent:

$$\text{si } \ln \left[\frac{p(X)}{(1 - p(x))} \right] > 0 \text{ aleshores } y = 1, \text{ si } \ln \left[\frac{p(X)}{(1 - p(x))} \right] \leq 0 \text{ aleshores } y = 0$$

A la Figura 5 tenim un possible exemple d'un model de regressió logística.

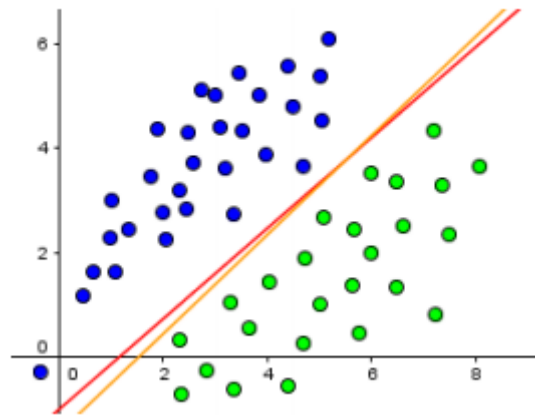


Figura 6: Regressió logística que mostra les dades de dues classes (verd i blau) juntament amb la frontera de decisió (línia taronja) i el model de regressió (línia blava).

4.2.3.4 Random Forest

Abans d'explicar la tècnica de *Random Forest*, hem d'explicar el concepte d'arbre de regressió. Un arbre de regressió és una tècnica aplicada en aprenentatge automàtic, que es defineix com un model predictiu que mapeja observacions sobre una característica a conclusions sobre el valor objectiu d'aquesta característica. En aquestes estructures d'arbre, les fulles representen un valor real d'aquella característica i les branques les conjuncions de característiques que han portat fins a la fulla. A la Figura 6 podem veure un exemple d'arbre de regressió.

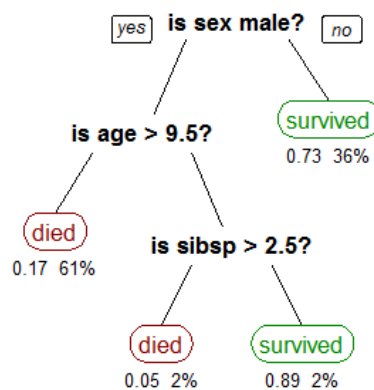


Figura 7: Arbre de Regressió

Random forest són un conjunt d'arbres de regressió on el seu resultat és la mitja de cada arbre. A més a més, per cada arbre s'aplica un soroll aleatori a les dades sense variar la seva mitja i variància. Aquest fet permet que aplicant el promig de la sortida de cada arbre de regressió, aquesta tècnica obtingui beneficis.

4.2.3.5 Support Vector Machine

Support Vector Machine (SVC) són un conjunt d'algoritmes d'aprenentatge automàtic que donat un conjunt de mostres etiquetades en una certa categoria, s'encarrega de predir a quina categoria pertany una nova mostra.

El funcionament d'aquest model de regressió és el següent: Es representen els punts de mostra a l'espai, separant les categories o classes per un espai que sigui el més ampli possible, de tal manera que quan les noves mostres s'introdueixin al model, en funció de la seva proximitat a alguna de les categories, podrà ser classificada dins de la categoria més pròxima a la mostra. Per tant, una bona separació entre classes permetrà una classificació millor. A la Figura 7 podem veure un exemple del model de regressió SVC.

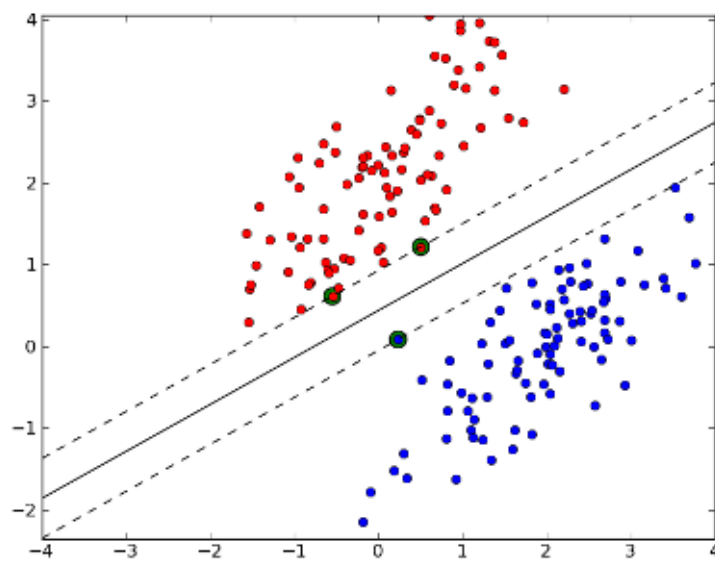


Figura 8: Màquina de vectors de suport

4.2.3.6 Mètriques classificació

En aquesta secció mostrarem les diferents tècniques que hem utilitzat per validar les nostres prediccions i classificacions. Anomenem y_{pred} a l'etiqueta que s'ha predit per un cert bitllet d'avió i y_{test} a l'etiqueta real del bitllet d'avió. Recordem que les dues possibles etiquetes assignades a un bitllet d'avió són comprar o no comprar.

Les mètriques utilitzades per avaluar els nostres classificadors, són algunes de les explicades a la secció [4.2.3.1](#). En concret, les que hem utilitzat en el context del nostre projecte, tenint en compte les bases ja explicades, són les presentades a continuació:

- **Exactitud:** mesura la proporció de mostres classificades correctament.

$$e = \frac{TP + TN}{TP + FP + TN + FN}$$

En concret, hem utilitzat la tècnica de *cross-validation* per a mesurar l'exactitud del nostre model de predicció. *Cross-validation* és una tècnica que prediu l'ajust d'un model de predicció a un conjunt de test sense disposar del conjunt explícit de dades de test. En el context del projecte, hem utilitzat la tècnica de *cross-validation* anomenada *k-fold cross-validation*, que és una tècnica basada en dividir el conjunt d'observacions en k subconjunts d'aproximadament la mateixa grandària, de tal manera que un dels subconjunts és utilitzat com a conjunt de test i els altres k-1 subconjunts són utilitzats com a conjunts d'entrenament. Aquest procés de *cross-validation* és repetit durant k iteracions, utilitzant diferents conjunts de test per a cada iteració. Finalment realitza la mitja aritmètica dels errors obtinguts a cada iteració per obtenir un únic resultat.

A la Figura 8 podem observar un exemple de *cross-validation*.

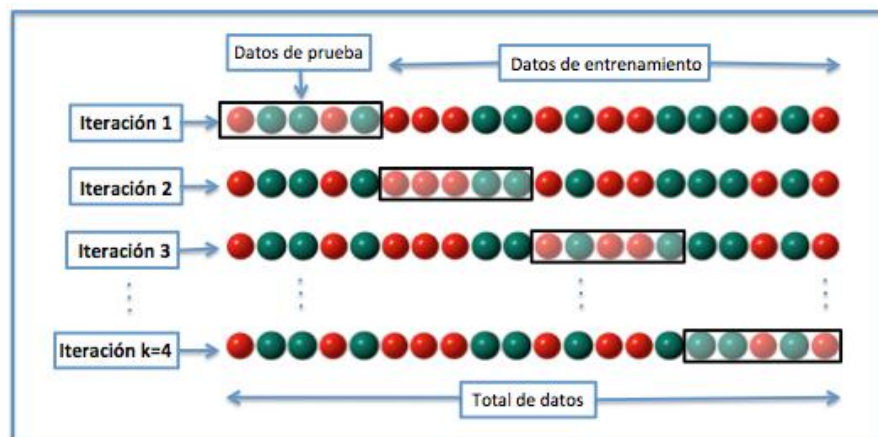


Figura 9: Cross-Validation de k iteracions on k = 4

- **Sensibilitat** : fa referència a l'eficiència de la classificació de tots els elements d'una classe, dit d'una altra manera, mesura la capacitat que té el classificador per detectar petits canvis en les dades.

$$sensibilitat = \frac{TP}{positivos reales} = \frac{TP}{TP + FN}$$

- **Precisió**: fa referència a la qualitat de la resposta del classificador, es a dir, donada una predicció positiva per part del classificador , quant probable és que sigui correcte.

$$precisió = \frac{TP}{positius predits} = \frac{TP}{TP + FP}$$

5 Experiments i resultats

En aquest apartat s'explicarà pas per pas els experiments realitzats i els resultats obtinguts per cada pregunta plantejada a la secció [2.5](#). Fins ara s'han presentat tots els conceptes necessaris per poder entendre aquesta secció.

Primer es mostraran els resultats a les preguntes de la secció [2.5](#) i posteriorment es presentaran els resultats obtinguts dels classificadors explicats a la secció [4.2.3](#).

5.1 Preparació prèvia als experiments

Abans de començar a comentar els resultats, s'explicarà quines són les dades per respondre a cada pregunta. Inicialment tenim una taula on cada fila representa la informació que tenim de cada vol, per tant, a cada fila tenim informació com *l'identificador del vol*, *preu del vol*, *hora de sortida*, *hora d'arribada*, *data de sortida*, *data d'arribada*, etc. A partir d'aquesta taula fem una conversió de tal manera que a cada fila ens queda un vol i cada columna representa el preu del bitllet d'avió el dia en què hipotèticament es realitza la compra del bitllet per aquest vol, construint una matriu que presenta la següent estructura:

$$M = \begin{matrix} & d_1 & \dots & d_n \\ \begin{matrix} v_1 \\ \vdots \\ v_n \end{matrix} & \begin{pmatrix} p_{11} & \dots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nm} \end{pmatrix} \end{matrix}$$

on:

v_i és un vol

d_j és el dia en què es realitza la compra del bitllet per al vol i

p_{ij} és el preu del bitllet per al vol i , el dia j

Hi ha dies en què el preu del bitllet per a un vol és desconegut, el que fem és interpolar el valor del preu com ja s'ha explicat a la secció [2.3.3](#).

Segons la pregunta que volem respondre, la matriu variarà, ja que introduïrem alguna característica que sigui rellevant per a la resolució de la pregunta. A continuació, resoldrem les preguntes plantejades, explicant tot el desenvolupament d'aquesta, incloent-hi l'estructura particular de la matriu a cada pregunta.

5.2 Amb quants dies d'antelació haig de comprar un bitllet d'avió?

La resposta a la pregunta: *Amb quants dies d'antelació haig de comprar un bitllet d'avió?*, consisteix en trobar quants dies abans de la sortida del vol haig de comprar un bitllet d'avió perquè aquest em surti el més econòmic possible.

Per esbrinar-ho, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i s'ha construït una matriu que té la següent forma:

$$M = \begin{matrix} & da_1 & \dots & da_n \\ \begin{matrix} ds_1 v_1 \\ \vdots \\ ds_n v_n \end{matrix} & \begin{pmatrix} p_{11} & \dots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nm} \end{pmatrix} \end{matrix}$$

on:

ds_x és la data de sortida del vol.

v_i és el vol.

da_j és el nombre de dies entre el dia de compra i el dia de sortida del vol.

p_{ij} és el preu del bitllet per al vol v_i , amb data de sortida ds_x , el dia da_j .

Un cop tenim construïda la matriu, apliquem la corresponent neteja de dades esmentada a la secció [2.3.3](#). Arribats a aquest punt, tenim els preus normalitzats de cada vol amb data de sortida durant els dies de setmana santa, per a cada dia hipotètic de compra (dies consultats).

Com podem veure a la Figura 10, si visualitzem cada vol per separat, no podem extreure resultats immediats, per tant hem de buscar una alternativa que ens permeti visualitzar i entendre millor com evolucionen els preus dels vols al llarg dels dies.

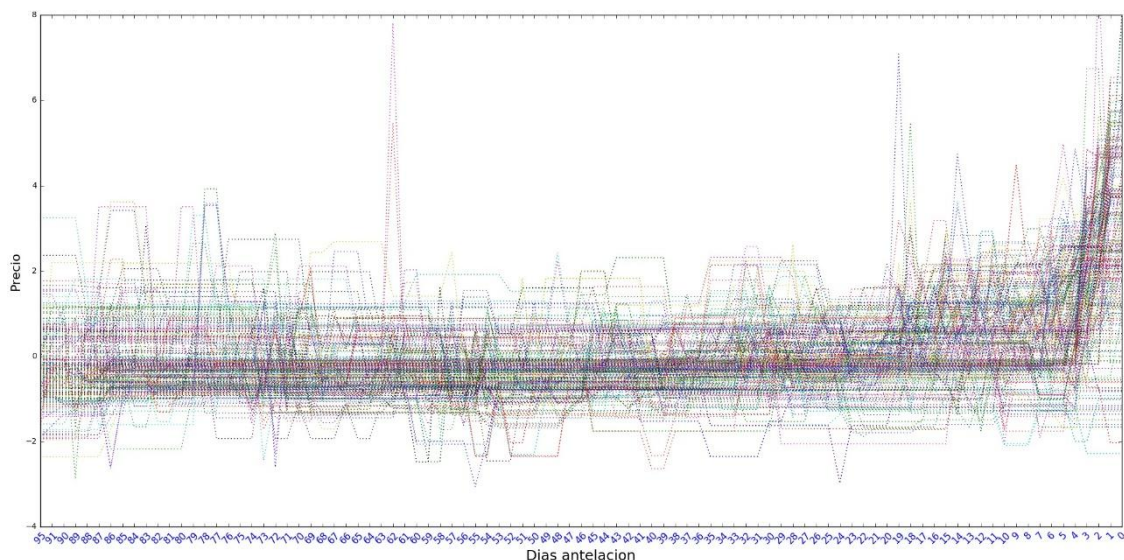


Figura 10: Evolució dels preus de cada vol a mesura que ens apropem al dia de sortida del vol.

La millor manera de veure amb quants dies d'antelació realitzar la compra del bitllet, és analitzant la tendència de tots els vols a mesura que ens apropem al dia de la seva sortida. La tendència ens dirà com fluctuen els preus al llarg dels dies, i la calculem fent la mitja aritmètica dels preus de cada vol per a cada dia de consulta.

En la següent figura 11, podem observar la tendència dels preus dels vols a mesura que ens apropem al dia de sortida.

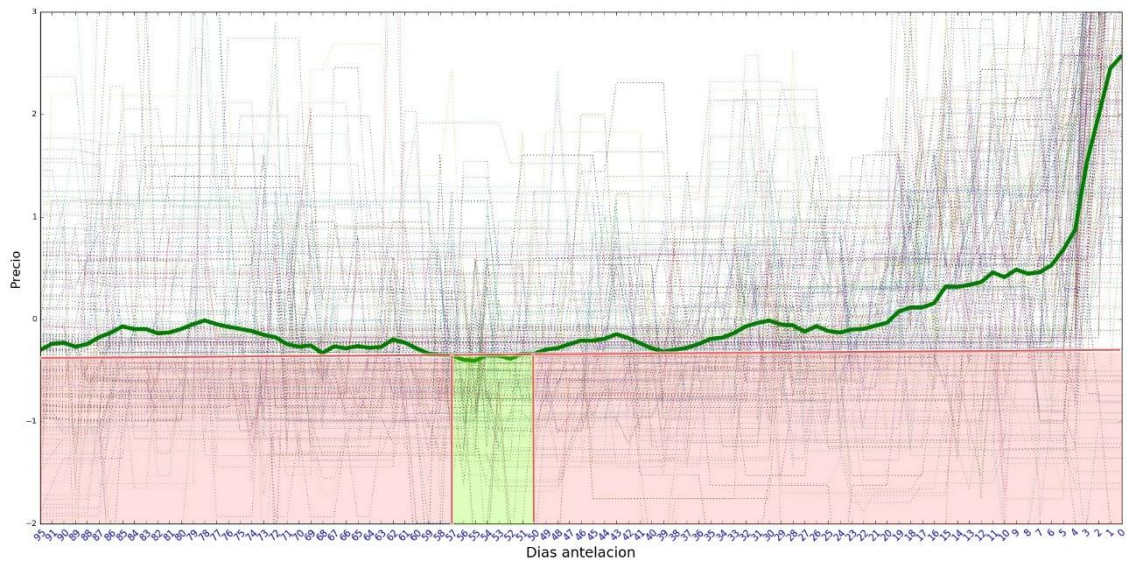


Figura 11: Tendència dels preus dels vols a mesura que ens apropem al dia de sortida dels vols (línia verda). Las franges vermelles mostren els dies d'antelació en que els preus aparentment estan més baixos. En verd es mostren els dies d'antelació amb que és més barat comprar el bitllet d'avió, i en vermell les zones on és més car comprar-lo.

Com podem observar a la Figura 11, sembla que el millor moment per realitzar la compra d'un bitllet d'avió és amb 50 – 57 dies d'antelació. Si volem corroborar-ho, no ens hem de fixar només en la mitja de les dades, sinó que també hem de tenir en compte la desviació estàndard d'aquestes, que ens dirà en quina quantitat es poden separar les dades de la seva mitja.

A la Figura 12, podem veure la desviació superior e inferior de les dades ($\mu \pm std$, on, μ : mitja dels preus dels bitllets d'avió per dia d'antelació, std : desviació estandard)

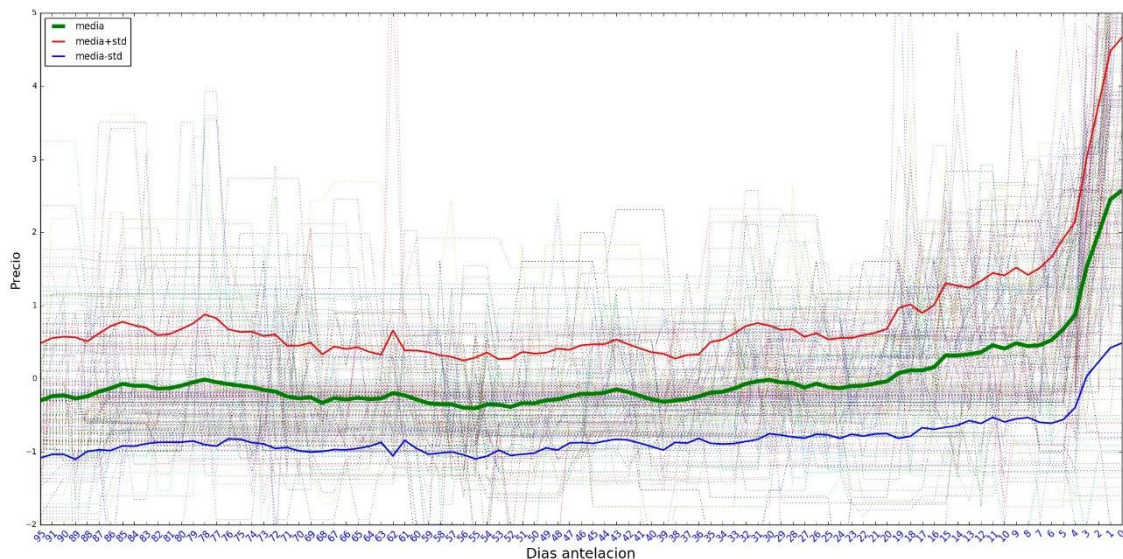


Figura 12: La línia verda mostra la mitja dels preus dels bitllets d'avió per dia d'antelació, la línia vermella mostra la desviació superior de les dades ($\mu + \text{std}$) y la línia blava mostra la desviació inferior de les dades ($\mu - \text{std}$)

Com podem observar a la Figura 12, els preus dels vols mantenen una tendència creixent durant bona part del temps. Tenint en compte la desviació de les dades i la seva mitja, podem dir que els últims 5 dies abans de la sortida del vol els preus augmenten exponencialment el seu valor, possiblement, a causa de la seva proximitat amb la data de sortida del vol, per tant, el pitjor moment per realitzar la compra del bitllet és durant els últims 5 dies. Per una altra banda, tot i que la diferència no és tan marcada, durant els últims 20 dies d'antelació el preu també manté una tendència creixent, tot i que no tan marcada com els últims 5 dies, però que també ens permet dir que tampoc és el moment idoni per realitzar la compra del bitllet d'avió.

Per una altra banda, si tenim en compte les desviacions superiors e inferiors de les dades, no és fàcil veure el moment en què el preu es troba en el seu moment més baix. Si a la Figura 11 podíem distingir sense gaire dificultats quina era el moment en què els preus eren més baixos (entre els 50 – 57 dies d'antelació), a la Figura 12, tenint en compte la desviació de les dades, no està tan clar quin és el moment idoni per comprar el bitllet d'avió. Per aquest motiu, comprovarem si realment les diferències entre comprar amb 50-57 dies d'antelació respecte a realitzar la compra del bitllet la resta de dies són significatives. Per fer-ho realitzarem un test d'hipòtesi nul·la.

A continuació, segmentem les dades en els dos grups d'interès: bitllets comprats amb 50-57 dies d'antelació i bitllets comprats amb la resta de dies. Visualitzarem ambdós grups d'interès per tal de comentar les diferències entre ells i poder resoldre la pregunta plantejada durant el desenvolupament del projecte.

A la figura 13 podem veure un diagrama de caixes dels dos grups d'interès: vols comprats amb 50-57 dies d'antelació i vols comprats amb la resta de dies d'antelació. Aquest tipus de diagrama representa visualment una distribució. Cada caixa té les mateixes característiques: la línia del centre representa la mediana de la distribució, la caixa visualitza un 50% de la mostra (des del primer quartil fins al tercer), les línies verticals determinen el límit de les distribucions i per últim, tota mostra que estigui fora de la normalitat (mostra atípica) es visualitza per fora dels límits amb una creu.

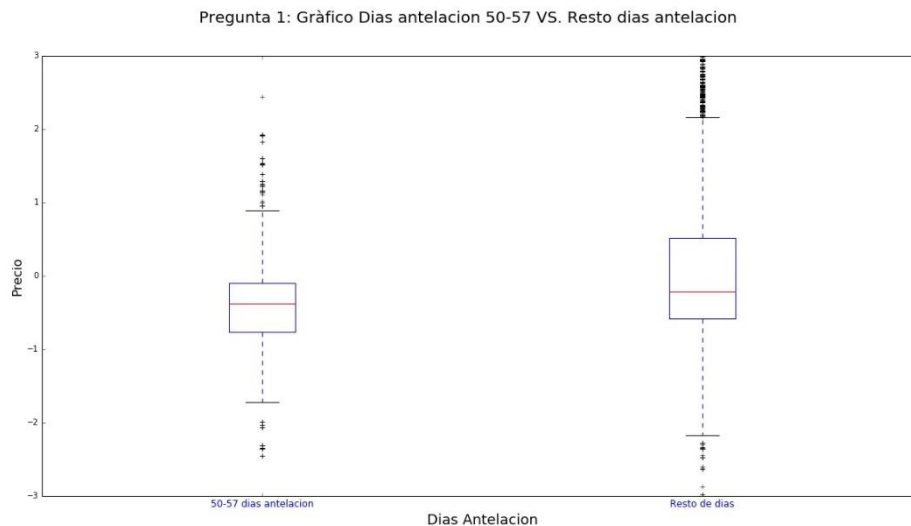


Figura 13: Visualització de la distribució dels vols comprats amb 50-57 dies d'antelació i del vols comprats amb la resta de dies.

Com podem observar a la figura 13, visualment els preus dels bitllets comprats amb 50-57 dies d'antelació són més barats que els bitllets comprats en la resta de dies, però la diferència que observem és prou significativa per dir que comprar amb X dies d'antelació és més barat?

Comprovem si la diferència visual és significativa mitjançant el test d'hipòtesis nul·la.

Com s'explica a la secció [4.2.2.1](#), per tal de plantejar el test d'hipòtesis nul·la, s'ha de determinar la hipòtesi que es vol contrastar o hipòtesi nul·la H_0 i la hipòtesi alternativa H_1 , que representa la hipòtesi que desitgem poder concloure com a certa.

En el context de la nostra pregunta, H_0 i H_1 són les següents:

H_0 : *Comprar un bitllet d'avió amb 50-57 dies d'antelació és igual que comprar-lo amb la resta de dies d'antelació.*

H_1 : *Comprar un bitllet d'avió amb 50-57 dies d'antelació és diferent que comprar-lo amb la resta de dies d'antelació.*

Transformem aquestes dues hipòtesis a notació matemàtica. Tenint en compte que:

M_1 : *Preu dels bitllets comprats amb 50 – 57 dies d'antelació.*

M_2 : *Preu dels bitllets comprats amb la resta de dies d'antelació.*

Llavors tenim:

$$H_0: M_1 = M_2$$

$$H_1: M_1 \neq M_2$$

Per tal de dur a terme el test d'hipòtesi nul·la hem pres com a nivell de significança:

$$\alpha = 0.05.$$

Els resultats del test d'hipòtesi nul·la ha estat el següent:

$$p\text{-valor} = 1.12757111548e-117$$

Com que el p-valor obtingut és més petit que el nivell de significança establert, podem acceptar la hipòtesi alternativa i rebutjar la hipòtesi nul·la. Per tant, podem concloure que el millor moment per realitzar la compra dels bitllets d'avió és amb 50-57 dies d'antelació.

5.3 Vols d'alta demanda vs. Vols de baixa demanda

Recordem que aquesta segona pregunta pretén esbrinar si existeix diferència entre els dies d'antelació amb els quals comprar un vol considerat d'alta demanda respecte als dies d'antelació amb els quals comprar un vol considerat de baixa demanda. Fem un petit recordatori de què consideràvem vols d'alta demanda i que consideràvem vols de baixa demanda abans de resoldre la pregunta. Els vols d'alta demanda són tots aquells vols que per certs motius tendeixen a ser més sol·licitats pels consumidors i els vols de baixa demanda són tota la resta de vols. Partint de la classificació dels vols com a alta o baixa demanda explicada a la secció [2.5.2](#) procedim a explicar com s'ha resolt la pregunta.

Per resoldre la pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i s'ha construït una matriu que té la següent forma:

$$M = \begin{matrix} & da_1 & \dots & da_n \\ \begin{matrix} ds_1 v_1 d_1 \\ \vdots \\ ds_n v_n d_n \end{matrix} & \begin{pmatrix} p_{11} & \dots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nm} \end{pmatrix} \end{matrix}$$

on:

ds_x és la data de sortida del vol.

v_i és el vol.

d_y és la demanda del vol.

da_j és el nombre de dies entre el dia de compra i el dia de sortida del vol.

p_{ij} és el preu del bitllet per al vol v_i , amb data de sortida ds_x , el dia da_j .

Un cop tenim construïda la matriu, apliquem la corresponent neteja de dades esmentada a la secció [2.3.3](#). Arribats a aquest punt, tenim els preus normalitzats de cada vol d'alta o baixa demanda, amb data de sortida durant els dies de setmana santa, per a cada dia hipotètic de compra (dies consultats). Posteriorment, segmentem les dades en els dos grups d'interès: vols de baixa demanda i vols d'alta demanda.

La intuïció ens fa pensar que els preus dels bitllets d'avió considerats d'alta demanda tendiran a mantenir-se per sobre dels preus dels bitllets d'avió considerats com a baixa demanda durant la major part del temps, ja que seran més sol·licitats pels consumidors i això serà aprofitat per encarir el seu preu. A continuació, visualitzarem ambdós grups d'interès per tal de comentar les diferències entre ells i poder resoldre la pregunta plantejada durant el desenvolupament del projecte.

A la figura 14, podem observar com varien els preus dels vols considerats d'alta demanda i dels vols considerats de baixa demanda, a mesura que ens apropem a la data de sortida dels vols.

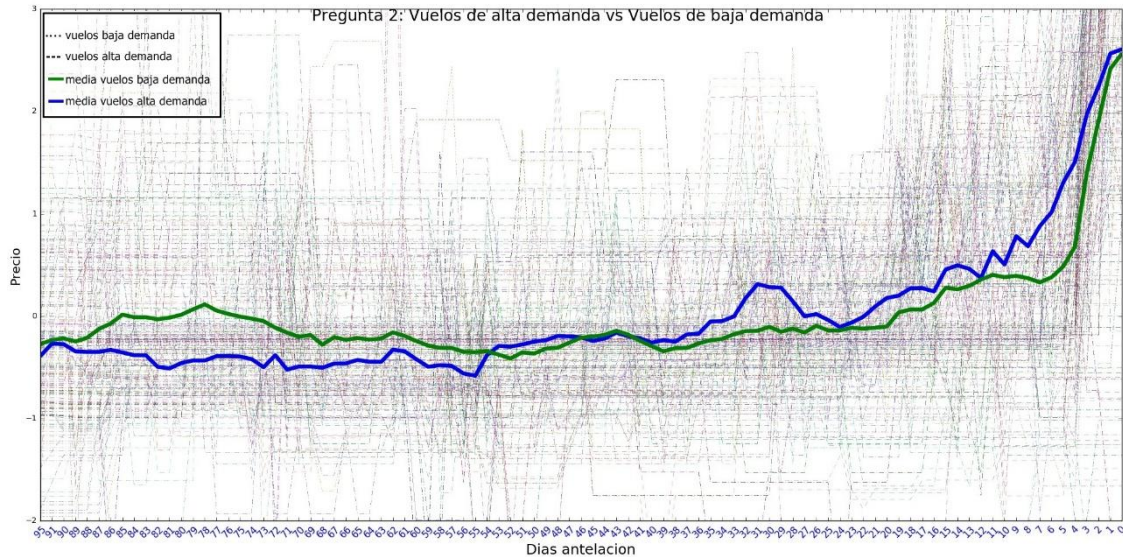


Figura 14: Figura que mostra la mitja dels preus dels vols d'alta demanda (línia blava) i dels vols de baixa demanda (línia verda) a mesura que ens apropem a la data de sortida dels vols.

Com podem observar a la figura 14, tal com havíem previst, els preus dels vols d'alta demanda tendeixen a estar per sobre dels preus dels vols de baixa demanda gran part del temps. Durant els dies més allunyats de la data de sortida dels vols (des del dia d'antelació 95 fins al dia d'antelació 54), observem que la mitja dels preus dels vols de baixa demanda, està bastant per sobre de la mitja dels vols considerats d'alta demanda. Aquest fet pot ser en certa manera il·lusori, ja que el més normal és pensar que com a mínim els vols d'alta demanda es mantindran a la mateixa mitja que els vols de baixa demanda, mai per sota. L'origen d'aquest fet pot residir en la segmentació que hem definit dels dos grups d'interès, de tal manera que es possible que no s'hagi detectat alguna característica amagada darrere de les dades que ens ha fet classificar alguns vols d'alta demanda com a baixa demanda. Per tal d'acabar de resoldre la pregunta, no tindrem en compte aquest fet i prendrem com a correcte la segmentació plantejada inicialment.

Com podem observar a la figura 14, el millor moment per a realitzar la compra d'un vol considerat d'alta demanda és sempre com més aviat millor, ja que els preus d'aquests tipus de vols tendeixen a créixer a mesura que ens apropem a la data de sortida dels vols (dia d'antelació 0). Per una altra banda, podem observar que els preus dels vols considerats com a baixa demanda són alts des dels 95 dies d'antelació fins als 60 dies d'antelació aproximadament, després disminueixen des dels 60 fins als 40 dies d'antelació i posteriorment no paren de créixer fins al dia de sortida del vol (dia d'antelació 0). Per tant, podem concloure que si es tracta d'un vol d'alta demanda, hem de realitzar la compra com més aviat millor, però que si es tracta d'un vol de baixa demanda, hem de realitzar la compra entre els 60-40 dies abans de la sortida del vol.

5.4 Caps de setmana vs. Dies laborals

La resposta a la pregunta: *És més car comprar un bitllet d'avió un dia de cap de setmana que un dia laboral?*, consisteix en esbrinar i demostrar mitjançant un test d'hipòtesis nul·la, si comprar un bitllet d'avió un dia de cap de setmana és més car que comprar un bitllet d'avió un dia laboral.

Abans de començar a explicar com hem resolt aquesta pregunta, fem un petit recordatori per explicar quins dies hem considerat com a caps de setmana i quins dies hem considerat com a laborals. Com ja s'ha explicat a la secció [2.5.3](#), hem considerat com a laborals els dilluns, dimarts, dimecres i dijous no festius, i hem considerat com a caps de setmana els dissabtes i diumenges no festius.

Per resoldre la pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i s'ha construït una matriu que té la següent forma:

$$M = \begin{matrix} & d_1 & \dots & d_n \\ \begin{matrix} ds_1 v_1 \\ \vdots \\ ds_n v_n \end{matrix} & \begin{pmatrix} p_{11} & \dots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nm} \end{pmatrix} \end{matrix}$$

on:

ds_x és la data de sortida del vol.

v_i és el vol.

d_y es la data de compra del bitllet d'avió (dia de consulta).

p_{ij} és el preu del bitllet per al vol v_i , amb data de sortida ds_x , el dia d_y .

Un cop tenim construïda la matriu, apliquem la corresponent neteja de dades esmentada a la secció [2.3.3](#), tenint en compte que abans de realitzar la normalització dels preus, eliminem els divendres i festius per tal de no introduir errors en la normalització, ja que si els eliminéssim després de normalitzar les dades, els estariem tenint en compte. Arribats a aquest punt, tenim els preus normalitzats de cada vol, amb data de sortida durant els dies de setmana santa, per a cada dia hipotètic de compra (dies consultats).

Podem arribar a pensar que els preus dels bitllets d'avió són més alts els caps de setmana que els dies laborals, ja que són els dies en què la gent té més temps lliure per realitzar la compra d'aquests. A continuació, visualitzarem ambdós grups d'interès per tal de comentar les diferències entre ells i poder resoldre la pregunta plantejada durant el desenvolupament del projecte.

A la figura 15 podem veure un diagrama de caixes dels dos grups d'interès: vols comprats en cap de setmana i vols comprats en dies laborals. Com ja s'ha explicat, aquest tipus de diagrama representa visualment una distribució. Cada caixa té les mateixes característiques: la línia del centre representa la mediana de la distribució, la caixa visualitza un 50% de la mostra (des del primer quartil fins al tercer), les línies verticals determinen el límit de les distribucions i per últim, tota mostra que estigui fora de la normalitat (mostra atípica) es visualitza per fora dels límits amb una creu.

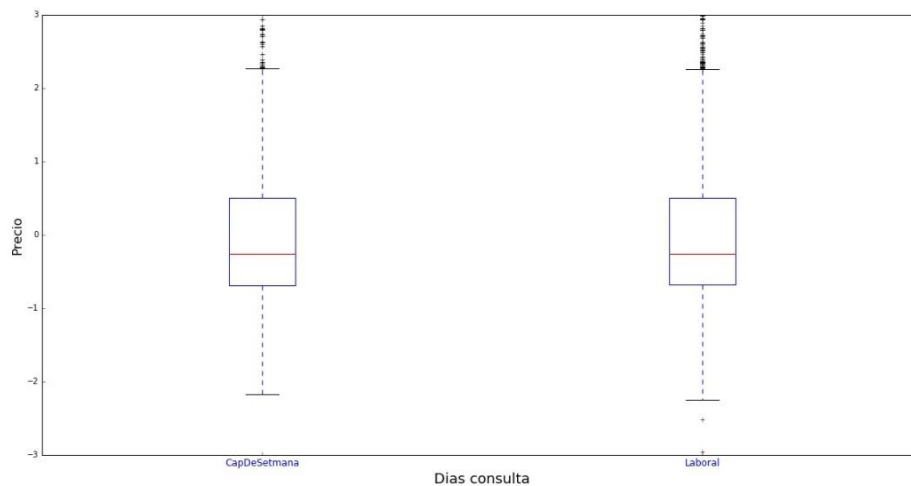


Figura 15: Visualització de la distribució dels vols comprats en cap de setmana i del vols comprats en un dia laboral

Com podem observar a la figura 15, les dues distribucions tenen la mateixa mediana, cosa que ens indica que la normalització de les dades realitzada abans de visualitzar les dades s'ha dut a terme correctament. Per una altra banda, les mostres atípiques presentades en ambdues distribucions són causades pel fet que hi ha bitllets d'avió en què el seu preu està molt per sobre de la mediana de la distribució, per aquest motiu surten fora dels límits. Si ens hi fixem atentament, podem observar que la distribució de bitllets d'avió comprats en cap de setmana, està una mica per sobre de la distribució de bitllets comprats en dies laborals, però no és fet suficient per afirmar que els bitllets comprats en cap de setmana són més cars que els bitllets comprats en dies laborals. Per tal de poder afirmar-ho, hem d'esbrinar si hi ha diferències significatives entre ambdues distribucions. Per poder resoldre-ho, hem aplicat un test d'hipòtesis nul·la com l'explicat a la secció [4.2.2.1](#).

Com s'explica a la secció [4.2.2.1](#), per tal de plantejar el test d'hipòtesis nul·la, s'ha de determinar la hipòtesi que es vol contrastar o hipòtesi nul·la H_0 i la hipòtesi alternativa H_1 , que representa la hipòtesi que desitgem poder concloure com a certa.

En el context de la nostra pregunta, H_0 i H_1 són les següents:

H_0 : *Comprar un bitllet d'avió en cap de setmana és igual que comprar-lo un dia laboral.*

H_1 : *Comprar un bitllet d'avió en cap de setmana és diferent que comprar-lo un dia laboral.*

Transformem aquestes dues hipòtesis a notació matemàtica. Tenint en compte que:

M_1 : *Preu dels bitllets comprats en cap de setmana.*

M_2 : *Preu dels bitllets comprats en dies laborals.*

Llavors tenim:

$$H_0: M_1 = M_2$$

$$H_1: M_1 \neq M_2$$

Per tal de dur a terme el test d'hipòtesi nul·la hem pres com a nivell de significança:

$$\alpha = 0.05.$$

Els resultats del test d'hipòtesi nul·la ha estat el següent:

$$p\text{-valor} = 0.0432378501721$$

Com que el p-valor obtingut és més petit que el nivell de significança establert, podem acceptar la hipòtesi alternativa i rebutjar la hipòtesi nul·la. Per tant, podem concloure que els bitllets d'avió comprats en cap de setmana són significativament més cars que els bitllets d'avió comprats en dies laborals.

5.5 Principi de mes vs. Resta del mes

La resposta a la pregunta: *És més car comprar un bitllet a principi de mes (dies d'ingrés de nòmina) que la resta de dies?*, consisteix en esbrinar i demostrar mitjançant un test d'hipòtesis nul·la, si comprar un bitllet d'avió a principi de mes (dies d'ingrés de nòmina) és més car que comprar un bitllet d'avió la resta del mes.

Abans de començar a explicar com hem resolt aquesta pregunta, fem un petit recordatori per explicar quins dies hem considerat com a principi de mes i quins dies hem considerat com a resta del mes. Com ja s'ha explicat a la secció 2.5.4, hem considerat com dies de principi de mes (dies d'ingrés de nòmina) des de l'última setmana del mes fins a la primera setmana del mes següent, ja que hi ha persones que reben la nòmina a l'última setmana del mes. Per un altre banda, hem considerat com a resta del mes, la resta de dies que no són considerats com a dies de principi de mes.

Per resoldre la pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i s'ha construït una matriu que té la següent forma:

$$M = \begin{matrix} & d_1 & \dots & d_n \\ \begin{matrix} ds_1 v_1 \\ \vdots \\ ds_n v_n \end{matrix} & \begin{pmatrix} p_{11} & \dots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nm} \end{pmatrix} \end{matrix}$$

on:

ds_x és la data de sortida del vol.

v_i és el vol.

d_y es la data de compra del bitllet d'avió (dia de consulta).

p_{ij} és el preu del bitllet per al vol v_i , amb data de sortida ds_x , el dia d_y .

Un cop tenim construïda la matriu, apliquem la corresponent neteja de dades esmentada a la secció 2.3.3, tenint en compte que abans de realitzar la normalització dels preus, eliminem les últimes dues setmanes abans de la sortida del vol, per tal de no introduir soroll a les dades, ja que com hem vist a la pregunta 1 durant les últimes dues setmanes els preus creixen de forma més pronunciada. Arribats a aquest punt, tenim els preus normalitzats de cada vol, amb data de sortida durant els dies de setmana santa, per a cada dia hipotètic de compra (dies consultats).

Podem arribar a pensar que els preus dels bitllets d'avió són més alts a principi de mes que a la resta de dies, ja que són els dies en què la gent rep l'ingrés de la seva nòmina i per tant, té més diners per fer la compra dels bitllets. A continuació, visualitzarem ambdós grups d'interès per tal de comentar les diferències entre ells i poder resoldre la pregunta plantejada durant el desenvolupament del projecte.

Igual que a l'apartat anterior, a la figura 16 podem veure un diagrama de caixes dels dos grups d'interès: bitllets comprats a principi de mes i bitllets comprats en la resta de dies.

Partint de l'explicació ja feta sobre que és un diagrama de caixes i quines característiques presenta, passem a visualitzar-lo en el context de la pregunta.

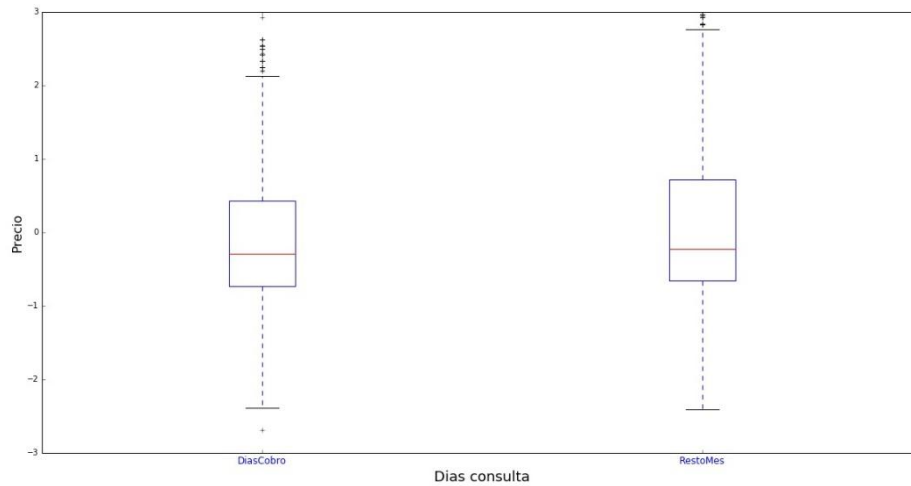


Figura 16: Visualització de la distribució dels bitllets comprats a principi de mes i del vols comprats en un dia considerat com a resta del mes.

Com podem observar a la figura 16, les dues distribucions tenen la mateixa mediana, cosa que ens indica que la normalització de les dades realitzada abans de visualitzar les dades s'ha dut a terme correctament. Per una altra banda, les mostres atípiques presentades en ambdues distribucions són causades pel fet que hi ha bitllets d'avió en què el seu preu està molt per sobre de la mediana de la distribució, per aquest motiu surten fora dels límits. Si ens hi fixem atentament, podem observar que la distribució de bitllets d'avió comprats a principi de mes, està bastant per sota de la distribució de bitllets comprats en la resta de dies del mes, per tant, en una primera impressió sembla que els bitllets d'avió comprats a principi de mes, no són més cars que els bitllets comprats la resta de dies del mes. Però, hem de realitzar un test d'hipòtesis nul·la, ja que no és fet suficient per afirmar que els bitllets comprats a principi de mes no són més cars que els bitllets comprats la resta de dies del mes. Per tal de poder afirmar-ho, hem d'esbrinar si hi ha diferències significatives entre ambdues distribucions. Apliquem, doncs, el test d'hipòtesis nul·la per resoldre la pregunta.

Com ja s'ha explicat en diverses ocasions al llarg del projecte, per tal de plantejar el test d'hipòtesis nul·la, s'ha de determinar la hipòtesi que es vol contrastar o hipòtesi nul·la H_0 i la hipòtesi alternativa H_1 , que representa la hipòtesi que desitgem poder concloure com a certa.

En el context de la nostra pregunta, H_0 i H_1 són les següents:

H_0 : *Comprar un bitllet d'avió a principi de mes és igual que comprar-lo la resta de dies del mes.*

H_1 : *Comprar un bitllet d'avió a principi de mes és diferent que comprar-lo la resta de dies del mes.*

Transformem aquestes dues hipòtesis a notació matemàtica. Tenint en compte que:

M_1 : *Preu dels bitllets comprats a principi de mes.*

M_2 : *Preu dels bitllets comprats un dia de la resta del mes.*

Llavors tenim:

$$H_0: M_1 = M_2$$

$$H_1: M_1 \neq M_2$$

Per tal de dur a terme el test d'hipòtesi nul·la hem pres com a nivell de significança:

$$\alpha = 0.05.$$

El resultat del test d'hipòtesi nul·la ha estat el següent:

$$p\text{-valor} = 9.53652718017e-24$$

Com que el p-valor obtingut és molt més petit que el nivell de significança establert, podem acceptar la hipòtesi alternativa i rebutjar la hipòtesi nul·la. Per tant, podem concloure que els bitllets d'avió comprats en cap de setmana no són significativament més cars que els bitllets d'avió comprats en dies laborals.

Els resultats obtinguts, poden trastocar la nostra lògica, ja que pot ser difícil de creure que el preu dels bitllets d'avió no siguin més cars a principi de mes que a la resta del mes. Hi ha dues possibles opcions per les quals s'ha pogut produir aquest fet: el primer és pensar que hi ha hagut un error a la classificació dels dies de compra i que, per tant, hi ha dies classificats com a resta del mes que haurien d'haver estat classificats com a principi de mes, fet que és factible de què s'hagi produït, ja que el criteri de classificació ha estat establert per nosaltres. El segon motiu pot ser que simplement els resultats siguin correctes i no ens concordin amb la lògica i que, per tant, no hi hagi cap estratègia per part de les companyies aèries a l'hora d'establir els preus dels bitllets en referència al context de la pregunta. De qualsevol manera, prendrem com a certs els resultats obtinguts per tal de continuar desenvolupant el projecte.

5.6 Dies de la setmana

La resposta a l'última pregunta que es va plantejar: *Quin dia de la setmana és millor realitzar la compra d'un bitllet d'avió?*, consisteix en esbrinar i demostrar mitjançant un test d'hipòtesi nul·la, quin dia de la setmana és més barat i quin és més car per comprar un bitllet d'avió.

Per resoldre la pregunta, s'han agafat tot els vols que tenien data de sortida durant tot el període de setmana santa i s'ha construït una matriu que té la següent forma:

$$M = \begin{matrix} & d_1 & \dots & d_n \\ \begin{matrix} ds_1 v_1 \\ \vdots \\ ds_n v_n \end{matrix} & \begin{pmatrix} p_{11} & \dots & p_{1m} \\ \vdots & \ddots & \vdots \\ p_{n1} & \dots & p_{nm} \end{pmatrix} \end{matrix}$$

on:

ds_x és la data de sortida del vol.

v_i és el vol.

d_y es la data de compra del bitllet d'avió (dia de consulta).

p_{ij} és el preu del bitllet per al vol v_i , amb data de sortida ds_x , el dia d_y .

Un cop tenim construïda la matriu, apliquem la corresponent neteja de dades esmentada a la secció [2.3.3](#). Arribats a aquest punt, tenim els preus normalitzats de cada vol amb data de sortida durant els dies de setmana santa, per a cada dia hipotètic de compra (dies consultats).

És lògic pensar que els preus varien segons els dies en què realitzes la compra del bitllet. Per exemple, no ens sorprendria saber que el preu dels bitllets d'avió són més cars els dissabtes que els dimarts. Però, hi ha un dia que és més barat que tots els altres o en cas contrari, hi ha un dia que és més car que tots els altres? Procedim a esbrinar-ho.

A continuació, visualitzarem els set grups d'interès (un per cada dia de la setmana) per tal de comentar les diferències entre ells i poder resoldre la pregunta plantejada durant el desenvolupament del projecte.

A la figura 17 podem veure un diagrama de caixes dels set grups d'interès: bitllets comprats en dilluns, dimarts, dimecres, dijous, divendres, dissabte i diumenge.

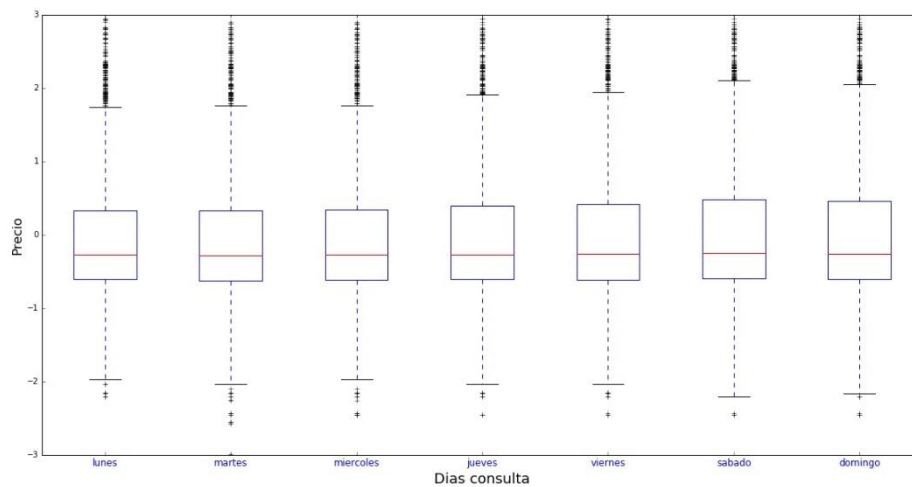


Figura 17: Visualització de la distribució dels bitllets comprats cada dia de la setmana.

Com podem observar a la figura 17, totes les distribucions tenen la mateixa mediana, cosa que ens indica que la normalització de les dades realitzada abans de visualitzar les dades s'ha dut a terme correctament. Per una altra banda, les mostres atípiques presentades en ambdues distribucions són causades pel fet que hi ha bitllets d'avió en què el seu preu està molt per sobre de la mediana de la distribució, per aquest motiu surten fora dels límits.

Si ens hi fixem atentament, podem observar que la distribució de bitllets d'avió comprats en dissabte, està bastant per sobre de la distribució de bitllets comprats en la resta de dies de la setmana, per tant, en una primera impressió sembla que el dia més car per a comprar un bitllet d'avió és el dissabte. També podem observar que els dilluns, dimarts i dimecres són els dies més econòmics per a comprar un bitllet d'avió. Però, hem de realitzar un test d'hipòtesi nul·la, per tal de demostrar que les diferències observades a priori a la figura 16 són significatives. Cal esmentar, que s'ha dut a terme un test d'hipòtesi nul·la comparant cada dia amb la resta de dies, tenint un total de 21 proves.

L'estructura que han seguit els 21 tests, és la següent:

H_0 : Comprar un bitllet d'avió el dia X és igual que comprar – lo el dia Y.

H_1 : Comprar un bitllet d'avió el dia X és diferent que comprar – lo el dia Y.

Transformem aquestes dues hipòtesis a notació matemàtica. Tenint en compte que:

M_1 : Preu dels bitllets comprats el dia X .

M_2 : Preu dels bitllets comprats el dia Y .

Llavors tenim:

$$H_0: M_1 = M_2$$

$$H_1: M_1 \neq M_2$$

Per tal de dur a terme el test d'hipòtesis nul·la hem pres com a nivell de significança:

$$\alpha = 0.05.$$

A la taula 2, podem observar els resultats de cada test d'hipòtesis nul·la, on el valor de cada cel·la representa el p-valor resultant de cada test. Les cel·les pintades de color verd indiquen els tests on la hipòtesi alternativa es pren com a certa i es rebutja la hipòtesi nul·la, en cas contrari, estan pintades en vermell aquelles cel·les on s'accepta la hipòtesi nul·la i es rebutja la hipòtesi alternativa.

	Dilluns	Dimarts	Dimecres	Dijous	Divendres	Dissabte	Diumenge
Dilluns	-	0.5878	0.7957	0.0077	0.0022	2.30e-06	1.59e-05
Dimarts	0.5878	-	0.4350	0.0018	0.0004	2.99e-07	2.99e-07
Dimecres	0.7957	0.4350	-	0.0185	0.0060	1.16e-05	7.35e-05
Dijous	0.0077	0.0018	0.0185	-	0.6960	0.0392	0.1232
Divendres	0.0022	0.0004	0.0060	0.6960	-	0.0942	0.2549
Dissabte	2.30e-06	2.99e-07	1.15e-05	0.0392	0.0942	-	0.5548
Diumenge	1.58e-05	2.99e-07	7.35e-05	0.1232	0.2549	0.5548	-

Taula 2: Resultats dels tests d'hipòtesis nul·la en termes de p-valor.

Les conclusions que podem extreure a partir dels resultats dels tests mostrats a través de la taula 2 són els següents: només podem arribar a afirmar que els dies més econòmics per comprar un bitllet d'avió són dilluns, dimarts i dimecres, i que per altra banda, els dies més cars per a realitzar la compra d'un bitllet d'avió són divendres, dissabte i diumenge. Com es pot veure a la taula 2, no podem afirmar que hi hagi un únic dia en què els preus siguin més baixos que la resta de dies, ni tampoc podem afirmar que hi hagi un únic dia en què els preus siguin més cars que la resta de dies.

5.7 Classificadors

En aquest apartat, s'expliquen les tècniques de predicció utilitzades per a la construcció dels dos classificadors esmentats a la secció [4.2.3.1](#) i quins resultats s'han obtingut. Recordem que l'objectiu principal del projecte, és construir un classificador que ens permeti recomanar la compra d'un cert bitllet d'avió.

Abans de donar una recomanació a l'usuari, necessitem esbrinar quina és la tècnica de predicció que s'ajusta més a les nostres dades. Les tècniques de predicció utilitzades per aquest experiment han sigut les següents, anteriorment explicades secció [4.2.3](#):

- Classificador amb *Linear Regresion*
- Classificador amb *Logistic Regresion*
- Classificador amb *Random Forest*
- Classificador amb *Suport Vector Machine*

Anem a mesurar quins dels predictors s'adequa més a les nostres dades. Les mètriques que s'utilitzaran per avaluar els predictors són les que ja s'han explicat en la secció [4.2.3.6](#): *Mètriques de classificació*.

5.7.1 Estratègia de validació

Totes les proves realitzades s'han fet a partir dels vols que tenen data de sortida durant tot el període de setmana santa. Per realitzar aquestes proves, hem hagut de fer una partició del conjunt de dades en un subconjunt de *training* i un subconjunt de *test*, de tal manera que s'entrena el classificador amb el subconjunt de *training* i s'avalua amb el subconjunt de *test*. Per tal de dur a terme aquesta partició, hem utilitzat la tècnica de *cross-validation*. Fent aquestes proves, podrem saber quina de les tècniques de predicció és la més adequada pel nostre conjunt de dades. Les diferents proves realitzades es diferencien entre elles segons la partició que s'ha fet del conjunt de dades. En concret hem realitzat les següents 3 proves:

- **Prova 1:** el conjunt total de dades inicials (100%) es divideix en un 90% de dades destinades a *training* i un 10% de dades destinades a *test*. És a dir, el nostre conjunt de *training* està format pel 90% de les dades i el nostre conjunt de *test* està format pel 10% restant de les dades.
- **Prova 2:** el conjunt total de dades inicials (100%) es divideix en un 80% de dades destinades a *training* i un 20% de dades destinades a *test*. És a dir, el nostre conjunt de *training* està format pel 80% de les dades i el nostre conjunt de *test* està format pel 20% restant de les dades.
- **Prova 3:** el conjunt total de dades inicials (100%) es divideix en un 70% de dades destinades a *training* i un 30% de dades destinades a *test*. És a dir, el nostre conjunt de *training* està format pel 70% de les dades i el nostre conjunt de *test* està format pel 30% restant de les dades.

Cal esmentar, que dins d'una mateixa prova, els conjunts de *training* i *test* utilitzats pels quatre models de predicció són els mateixos, ja que sinó, la comparació que estariem duent a terme seria incorrecte.

A partir d'aquest moment, anomenarem Classificador 1 al classificador que utilitza només 4 característiques per fer la predicció i Classificador 2 al classificador que introdueix una cinquena característica per fer la predicció.

5.7.2 Resultats de la prova 1

En aquesta primera prova, s'utilitza el 90% de les dades com a dades de *training* i el 10% de les dades com a dades de *test*. Primer mostrarem i analitzarem els resultats obtinguts de cada classificador (Classificador 1 i Classificador 2) per separat, i posteriorment conclourem l'apartat amb una petita comparació de resultats entre ambos classificadors.

Classificador 1

Els resultats obtinguts del Classificador 1, testejat amb els 4 models de predicció ja explicats anteriorment, es poden observar a la següent taula:

	Exactitud	Precisió			Sensibilitat		
		No Comprar (0)	Comprar (1)	Mitja	No Comprar (0)	Comprar (1)	Mitja
Linear Regression	0,5497	0,54	0,57	0,56	0,81	0,28	0,55
Logistic Regression	0,5472	0,54	0,57	0,56	0,81	0,28	0,55
Random forest	0,6172	0,58	0,54	0,56	0,56	0,56	0,56
Support vector machine	0,5480	0,54	0,57	0,56	0,80	0,29	0,55

Taula 3: Resultats de la prova 1 per al Classificador 1

Per tal de visualitzar més clarament quines són les diferències entre els diferents models, mostrarem les dades en una sèrie de gràfics que ens ajudaran a entendre quin dels models s'ajusta més a les nostres dades.

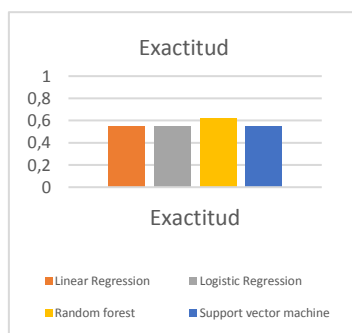


Figura 18: Exactitud del Classificador 1 amb els diferents models de predicció per a la prova 1.

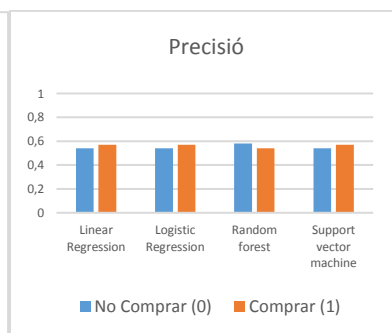


Figura 19: Precisió del Classificador 1 amb els diferents models de predicció per a la prova 1.

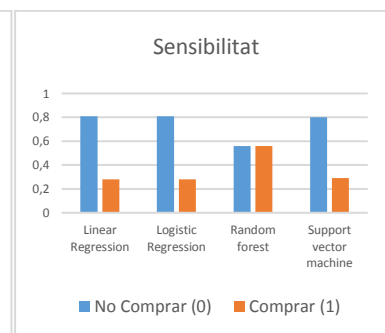


Figura 20: Sensibilitat del Classificador 1 amb els diferents models de predicció per a la prova 1.

Com podem observar a la figura 18, el Classificador 1 amb model de predicció *random forest* classifica correctament les mostres amb una exactitud superior a la resta de models de predicció. Per una altra banda, també es pot veure que la resta de models tenen un percentatge d'encert molt similar, al voltant del 55%. Aquest és un dels principals motius pel qual el Classificador 1 amb model de predicció *random forest*, és el model que millor s'ajusta a les nostres dades en aquesta primera prova.

A la figura 19, podem observar que la precisió de tots els models de predicció és bastant similar, és a dir, tots són capaços de classificar les dades dins d'una etiqueta amb la mateixa probabilitat d'encert.

Com podem veure a la figura 20, un altre dels motius pel qual el model de predicció *random forest* és el model que millor s'adequa a les nostres dades és el fet que la seva sensibilitat és més equilibrada, motiu que fa que pugui detectar millor els petits canvis en les dades per tal d'etiquetar-les en una categoria concreta (Comprar o No Comprar). Tot i que els altres predictors tenen una sensibilitat molt alta per detectar els canvis que indueixen a donar una resposta negativa (No comprar), la sensibilitat per detectar els canvis que indueixen a donar una resposta positiva (Comprar) és tan baixa que penalitza la seva efectivitat. Podem observar aquest fet, si ens hi fixem en la taula 3, la sensibilitat mitja per al *random forest* és una mica superior a resta.

Classificador 2

Els resultats obtinguts del Classificador 2 testejat amb els 4 models de predicció ja explicats anteriorment, es poden observar a la següent taula:

	Exactitud	Precisió			Sensibilitat		
		No Comprar (0)	Comprar (1)	Mitja	No Comprar (0)	Comprar (1)	Mitja
Linear Regression	0,7137	0,69	0,75	0,72	0,81	0,62	0,71
Logistic Regression	0,7159	0,70	0,74	0,72	0,79	0,64	0,72
Random forest	0,8460	0,75	0,76	0,75	0,79	0,72	0,75
Support vector machine	0,7164	0,70	0,75	0,72	0,80	0,63	0,72

Taula 4: Resultats de la prova 1 per al Classificador 2

Per tal de visualitzar més clarament quines són les diferències entre els diferents models, mostrarem les dades en una sèrie de gràfics que ens ajudaran a entendre quin dels models s'ajusta més a les nostres dades.

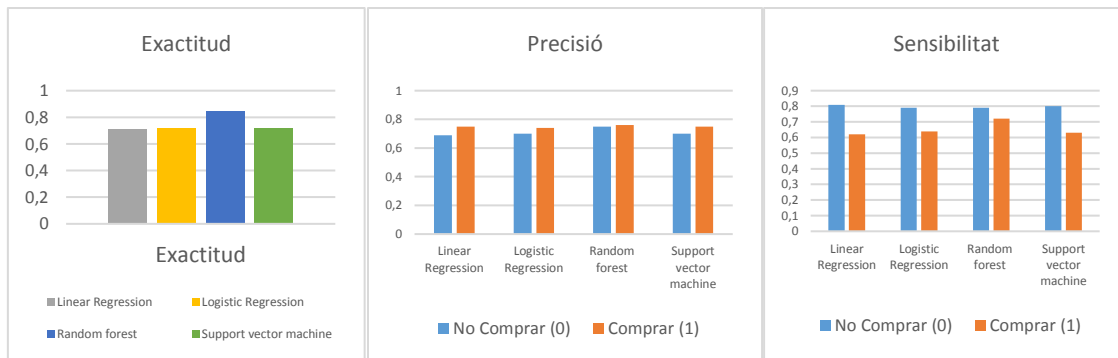


Figura 21: Exactitud del Classificador 2 amb els diferents models de predicció per a la prova 1.

Figura 22: Precisió del Classificador 2 amb els diferents models de predicció per a la prova 1.

Figura 23: Sensibilitat del Classificador 2 amb els diferents models de predicció per a la prova 1.

Com podem observar a la figura 21, el Classificador 2 amb model de predicció *random forest* classifica correctament les mostres amb una exactitud superior a la resta de models de predicció. Per una altra banda, també es pot veure que la resta de models tenen un percentatge d'exactitud molt similar, al voltant del 72%. Aquest és un dels principals motius pel qual el Classificador 2 amb model de predicció *random forest*, és el model que millor s'ajusta a les nostres dades en aquesta primera prova.

A la figura 22, podem observar que la precisió del Classificador 2 amb *random forest* és més alta i equilibrada que la precisió de la resta de models, motiu que li permet classificar les dades dins d'una etiqueta coneguda amb un percentatge d'encert superior a la resta de models.

Com podem veure a la figura 23, visualment la sensibilitat dels 4 models és molt similar, tot i que si observem numèricament els resultats a la taula 4, observem una petita diferència en la mitja a favor del model de *random forest*, ja que té una sensibilitat més equilibrada respecte a la detecció de petits canvis per tal d'etiquetar les dades com a positives (Comprar). Tot i que els altres predictors tenen una sensibilitat molt alta per detectar els canvis que indueixen a donar una resposta negativa (No comprar), la sensibilitat per detectar els canvis que indueixen a donar una resposta positiva (Comprar) és més baixa que la del model *random forest*. Podem observar aquest fet, si ens hi fixem en la taula 4, la sensibilitat mitja per al *random forest* és una mica superior a resta.

Conclusions

Com hem pogut comprovar, està clar que el model de predicció que més s'adequa a les nostres dades és el *random forest*, tant per al Classificador 1 com per al Classificador 2.

Un cop tenim clar que el model de predicció que s'ajusta millor a les dades és el *random forest*, anem a comprovar si existeix alguna diferència entre el Classificador 1 i el Classificador 2 amb la introducció de la cinquena característica.

Com es pot observar a les figures 24, 25 i 26, amb la introducció d'una cinquena característica al Classificador 2, hem pogut obtenir millors resultats de classificació que amb les 4 característiques utilitzades pel Classificador 1, ja que tant l'exactitud, la precisió i la sensibilitat són molt més altes en el Classificador 2. Per tant, el millor classificador que podem utilitzar és el Classificador 2 amb model de predicció *random forest*. Cal esmentar, que la introducció de noves característiques no sempre millorarà la qualitat de les nostres prediccions, és possible que hi hagi característiques dels vols que no tinguin valor significatiu en la predicció i que per tant, baixin la nostra efectivitat de classificació. Per això, s'ha de tenir cura a l'hora d'escollir les característiques que poden aportar significança a la predicció.

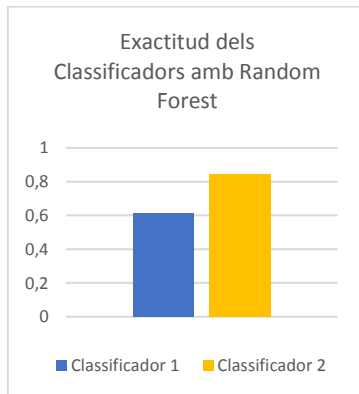


Figura 24: Exactitud dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 1.

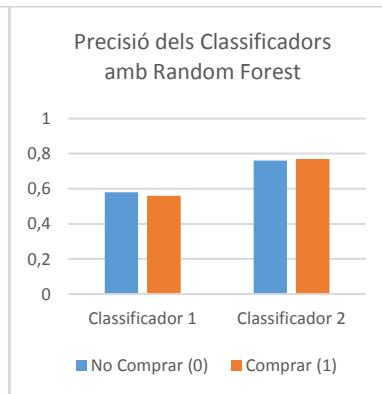


Figura 25: Precisió dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 1.

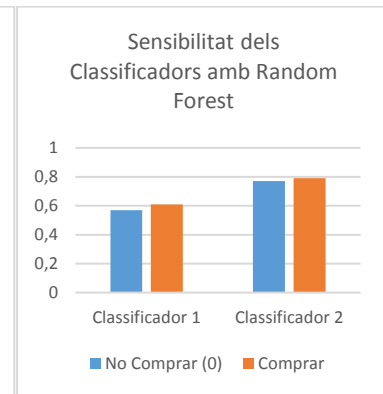


Figura 26: Sensibilitat dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 1.

5.7.3 Resultats de la prova 2

Aquesta segona prova utilitza com a dades de training el 80% de les dades inicials i com a dades de test el 20% de les dades inicials. Igual que a la prova 1, primer mostrarem i analitzarem els resultats obtinguts de cada classificador (Classificador 1 i Classificador 2) per separat, i posteriorment conclourem l'apartat amb una petita comparació de resultats entre ambdós classificadors.

Classificador 1

Els resultats obtinguts del Classificador 1 testejat amb els 4 models de predicció ja explicats anteriorment, es poden observar a la següent taula:

	Exactitud	Precisió			Sensibilitat		
		No Comprar (0)	Comprar (1)	Mitja	No Comprar (0)	Comprar (1)	Mitja
Linear Regression	0,5502	0,54	0,56	0,56	0,81	0,28	0,55
Logistic Regression	0,5454	0,54	0,56	0,56	0,81	0,28	0,55
Random forest	0,6154	0,58	0,56	0,57	0,61	0,54	0,57
Support vector machine	0,5491	0,54	0,57	0,56	0,81	0,28	0,55

Taula 5: Resultats de la prova 2 per al Classificador 1

Per tal de visualitzar més clarament quines són les diferències entre els diferents models, igual que a la prova 1, mostrarem les dades en una sèrie de gràfics que ens ajudaran a entendre quin dels models s'ajusta més a les nostres dades.

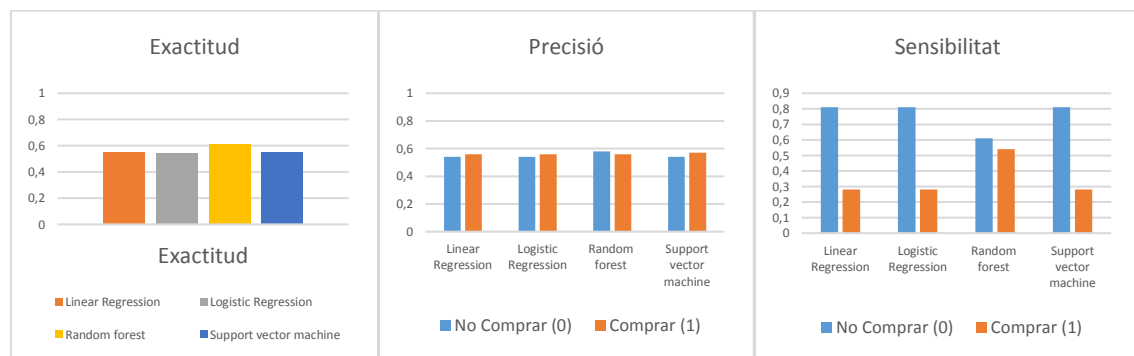


Figura 27: Exactitud del Classificador 1 amb els diferents models de predicció per a la prova 2.

Figura 28: Precisió del Classificador 1 amb els diferents models de predicció per a la prova 2.

Figura 29: Sensibilitat del Classificador 1 amb els diferents models de predicció per a la prova 2.

Com podem observar a la figura 27, el Classificador 1 amb model de predicció *random forest* classifica correctament les mostres amb una exactitud superior a la resta de models de predicció, igual que a la prova 1. Per una altra banda, també es pot veure que el model de predicció *linear regression* obté millors resultats que els models *logistic regression* i *support vector machine*. Aquest és un dels principals motius pel qual el Classificador 1 amb model de predicció *random forest*, és el model que millor s'ajusta a les nostres dades en aquesta segona prova.

A la figura 28, podem observar que el model *random forest* etiqueta les dades com a negatives (No comprar) amb una precisió més elevada que la resta de models, però que per altre banda, la seva precisió al etiquetar les dades com a positives (Comprar), és una mica inferior a la dels altres models. De totes maneres, si ens hi fixem en la taula 5, la mitja aritmètica entre la precisió d'etiquetatges positius i negatius del *random forest* és una mica més alta que la resta de models, per tant, *random forest* té una precisió més alta.

Com podem veure a la figura 29, igual que a la prova 1, un altre dels motius pel qual el model de predicció *random forest* és el millor model, és el fet que la seva sensibilitat és més equilibrada, motiu que fa que pugui detectar millor els petits canvis en les dades per tal d'etiquetar-les en una categoria concreta (Comprar o No Comprar). Tot i que els altres predictors tenen una sensibilitat molt alta per detectar els canvis que indueixen a donar una resposta negativa (No comprar), la sensibilitat per detectar els canvis que indueixen a donar una resposta positiva (Comprar) és tan baixa que penalitza la seva efectivitat. Podem observar que l'equilibri en la sensibilitat és un punt a favor, si ens hi fixem en la mitja aritmètica de les dades a la taula 5.

Classificador 2

Els resultats obtinguts del Classificador 2 testejat amb els 4 models de predicció ja explicats anteriorment, es poden observar a la següent taula:

	Exactitud	Precisió			Sensibilitat		
		No Comprar (0)	Comprar (1)	Mitja	No Comprar (0)	Comprar (1)	Mitja
<i>Linear Regression</i>	0,7137	0,70	0,74	0,72	0,79	0,63	0,71
<i>Logistic Regression</i>	0,7146	0,70	0,73	0,72	0,78	0,65	0,72
<i>Random forest</i>	0,8442	0,76	0,77	0,77	0,79	0,74	0,77
<i>Support vector machine</i>	0,7138	0,70	0,73	0,72	0,79	0,64	0,71

Taula 6: Resultats de la prova 2 per al Classificador 2

Per tal de visualitzar més clarament quines són les diferències entre els diferents models, igual que a la prova 1, mostrarem les dades en una sèrie de gràfics que ens ajudaran a entendre quin dels models s'ajusta més a les nostres dades.

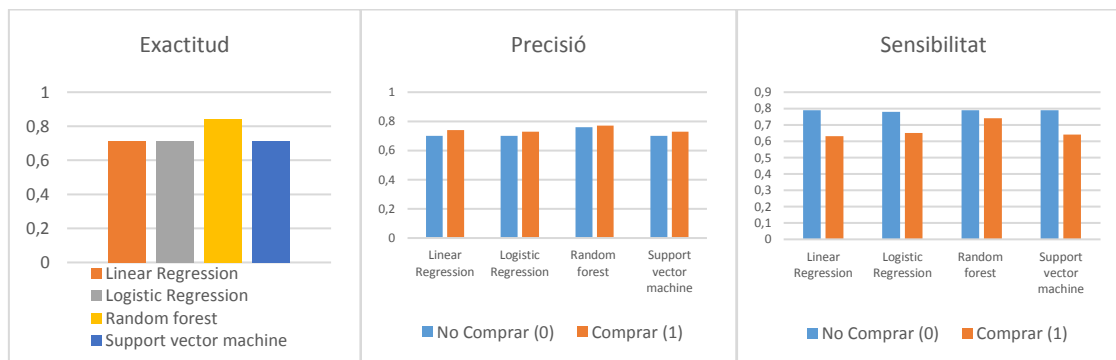


Figura 30: Exactitud del Classificador 2 amb els diferents models de predicció per a la prova 2.

Figura 31: Precisió del Classificador 2 amb els diferents models de predicció per a la prova 2.

Figura 32: Sensibilitat del Classificador 2 amb els diferents models de predicció per a la prova 2.

Com podem observar a la figura 30, el Classificador 2 amb model de predicció *random forest* té un percentatge d'exactitud més elevat que la resta de models, igual que a la prova 1. Per un altre banda, també es pot veure que la resta de models tenen un percentatge d'exactitud molt similar, al voltant del 71%. Aquest és un dels principals motius pel qual el Classificador 2 amb model de predicció *random forest*, és el model que millor s'ajusta a les nostres dades en aquesta segona prova.

A la figura 31, podem observar que la precisió del Classificador 2 amb *random forest* és més alta i equilibrada que la precisió de la resta de models, motiu que li permet classificar les dades dins d'una etiqueta coneguda amb un percentatge d'encert superior a la resta de models.

Com podem veure a la figura 32, la sensibilitat dels models *random forest* i *logistic regression* són molt similars, sent les més equilibrades respecte a la resta de models. Podem observar aquest fet si ens hi fixem en la taula 6, la sensibilitat mitja per al *random forest* és una mica superior a la resta.

Conclusions

Com hem pogut observar en aquesta segona prova, està clar que el model de predicció que més s'adequa a les nostres dades torna a ser el *random forest*, tant per al Classificador 1 com per al Classificador 2.

Un cop tenim clar que el model de predicció que s'ajusta millor a les dades és el *random forest*, anem a comprovar si la diferència entre el Classificador 1 i el Classificador 2 observada a la prova 1, es torna a donar.

Igual que a la prova 1, a les figures 33, 34 i 35 es pot observar que amb la introducció d'una cinquena característica al Classificador 2, hem pogut obtenir millors resultats de classificació que amb les 4 característiques utilitzades pel Classificador 1, ja que tant l'exactitud, la precisió i la sensibilitat són molt més altes en el Classificador 2. Per tant, el millor classificador que podem utilitzar és el Classificador 2 amb model de predicció *random forest*.

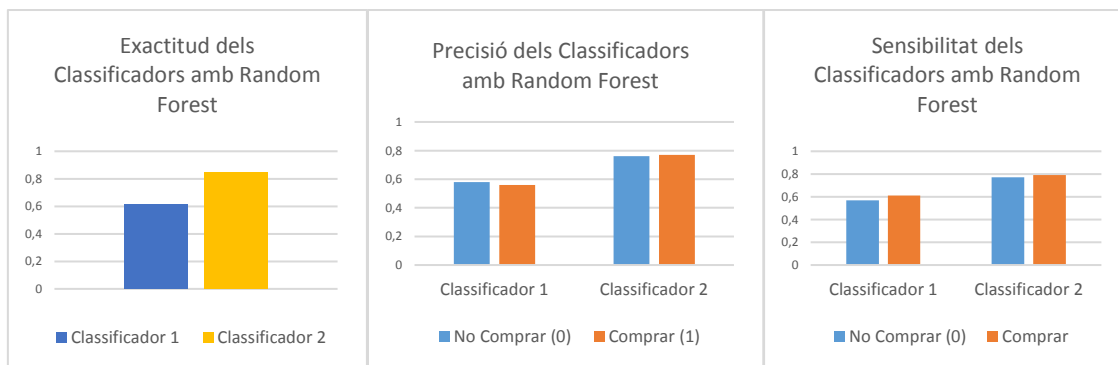


Figura 33: Exactitud dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 2.

Figura 34: Precisió dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 2.

Figura 35: Sensibilitat dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 2.

5.7.4 Resultats de la prova 3

Per últim, realitzarem aquesta tercera prova per acabar de confirmar els resultats obtinguts a les dues proves anteriors. Aquesta tercera prova utilitza com a dades de training el 70% de les dades inicials i com a dades de test el 30% de les dades inicials. Igual que a les proves anteriors, primer mostrarem i analitzarem els resultats obtinguts de cada classificador (Classificador 1 i Classificador 2) per separat, i posteriorment conclourem l'apartat amb una petita comparació de resultats entre ambdós classificadors.

Classificador 1

Els resultats obtinguts del Classificador 1 testejat amb els 4 models de predicció ja explicats anteriorment, es poden observar a la següent taula:

	Exactitud	Precisió			Sensibilitat		
		No Comprar (0)	Comprar (1)	Mitja	No Comprar (0)	Comprar (1)	Mitja
Linear Regression	0,5562	0,55	0,58	0,56	0,81	0,29	0,56
Logistic Regression	0,5572	0,55	0,58	0,56	0,80	0,29	0,56
Random forest	0,6205	0,59	0,57	0,58	0,62	0,54	0,58
Support vector machine	0,5518	0,55	0,58	0,56	0,81	0,29	0,56

Taula 7: Resultats de la prova 3 per al Classificador 1

Per tal de visualitzar més clarament quines són les diferències entre els diferents models, igual que a les altres proves, mostrarem les dades en una sèrie de gràfics que ens ajudaran a entendre quin dels models s'ajusta més a les nostres dades.

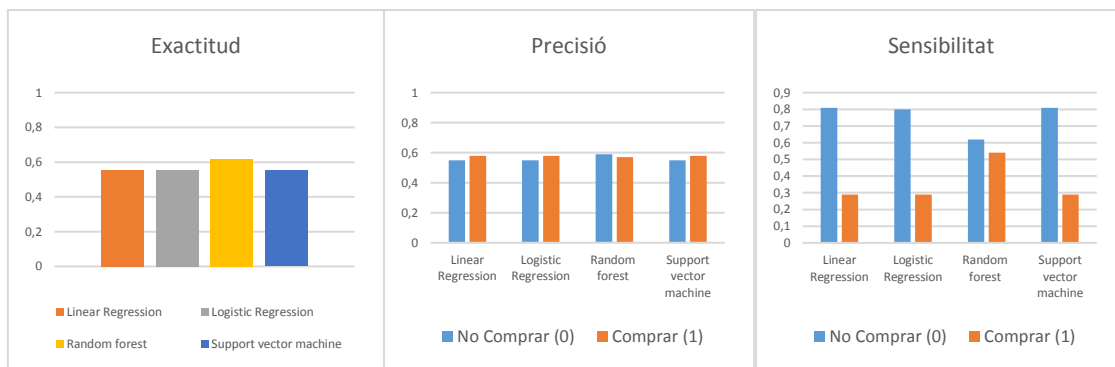


Figura 36: Exactitud del Classificador 1 amb els diferents models de predicció per a la prova 3.

Figura 37: Precisió del Classificador 1 amb els diferents models de predicció per a la prova 3.

Figura 38: Sensibilitat del Classificador 1 amb els diferents models de predicció per a la prova 3.

Com podem observar a les figures 36, 37 i 38, a la tercera prova obtenim resultats similars a les proves anteriors. El Classificador 1 amb *random forest*, té major exactitud, té una sensibilitat més equilibrada i per tant superior si parlem en termes de mitja aritmètica i una precisió superior a la dels altres predictors. Per tant, el model de *random forest* per al Classificador 1 torna a ser el que dona millors resultats.

Classificador 2

Els resultats obtinguts del Classificador 2 testejat amb els 4 models de predicció ja explicats anteriorment, es poden observar a la següent taula:

	Exactitud	Precisió			Sensibilitat		
		No Comprar (0)	Comprar (1)	Mitja	No Comprar (0)	Comprar (1)	Mitja
Linear Regression	0,7096	0,69	0,73	0,71	0,79	0,63	0,71
Logistic Regression	0,7197	0,70	0,73	0,72	0,78	0,65	0,71
Random forest	0,8465	0,78	0,79	0,78	0,81	0,76	0,78
Support vector machine	0,7166	0,70	0,73	0,71	0,78	0,64	0,71

Taula 8: Resultats de la prova 3 per al Classificador 2

Per tal de visualitzar més clarament quines són les diferències entre els diferents models, igual que a les altres proves, mostrarem les dades en una sèrie de gràfics que ens ajudaran a entendre quin dels models s'ajusta més a les nostres dades.

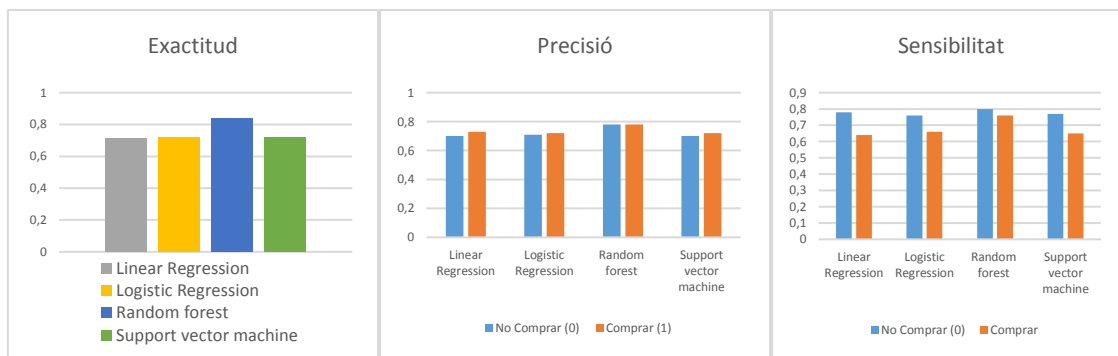


Figura 39: Exactitud del Classificador 2 amb els diferents models de predicció per a la prova 3.

Figura 40: Precisió del Classificador 2 amb els diferents models de predicció per a la prova 3.

Figura 41: Sensibilitat del Classificador 2 amb els diferents models de predicció per a la prova 3.

Com podem observar a les figures 39, 40 i 41, podem reafirmar que el Classificador 2 amb *random forest* s'adapta millor al model de les nostres dades que els altres 3 models de predicció pel Classificador 2. L'exactitud segueix sent superior, juntament amb la sensibilitat i la precisió, motiu pel qual, la qualitat de la predicció feta per el Classificador 2 amb *random forest* és superior a la dels altres tres models.

Conclusions

Com hem pogut observar en aquesta tercera i última prova, està clar que el model de predicció que més s'adequa a les nostres dades torna a ser el *random forest*, tant per al Classificador 1 com per al Classificador 2.

Comprovem per últim cop, que la diferència entre el Classificador 1 i el Classificador 2 observada a les altres proves, es torna a donar.

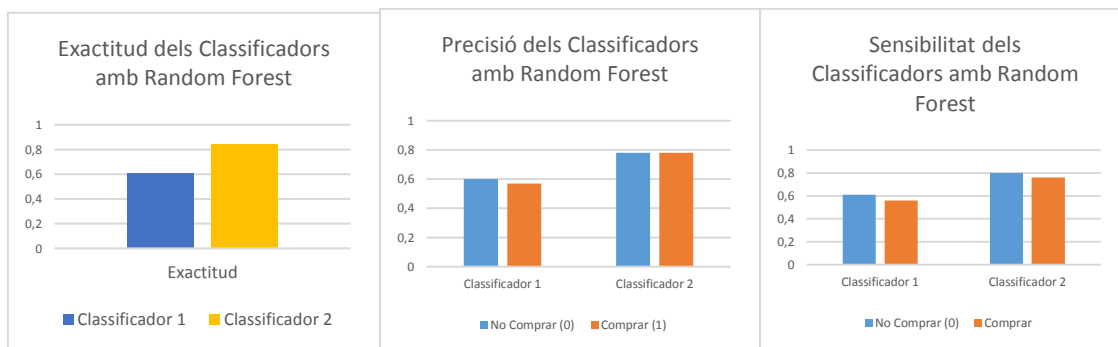


Figura 42: Exactitud dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 3.

Figura 43: Precisió dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 3.

Figura 44: Sensibilitat dels Classificadors 1 i 2 amb model de predicció random forest per a la prova 3.

Efectivament, a les figures 42, 43 i 44 es pot observar que amb la introducció d'una cinquena característica al Classificador 2, obtenim millors resultats de classificació que amb les 4 característiques utilitzades pel Classificador 1, ja que tant l'exactitud, la precisió i la sensibilitat són molt més altes en el Classificador 2. Per tant, i concloent l'apartat d'experiments i resultats, podem afirmar que el Classificador 2 amb model de predicció *random forest* és el que s'adapta millor a les nostres dades.

5.8 Connexió amb l'aplicació

En aquest apartat explicarem com s'ha dut a terme la connexió del nostre classificador amb l'eina que manipularà l'usuari per utilitzar el sistema de classificació. Aquesta eina, és competència d'un altre company i no se'n parlarà en aquest projecte. Proveïrem a l'eina d'ambdós classificadors (Classificador 1 i Classificador 2), però només amb els models de predicció *Linear Regression* i *Logistic Regression* per motius de temps, ja que són els més senzills de codificar.

A cada prova realitzada, a part de calcular les mètriques de classificació que ens han permès avaluar cada classificador i veure quin d'ells s'adequa més a les nostres dades, també hem calculat els coeficients que permeten modelar els models de predicció.

Els coeficients obtinguts en ambdós models de predicció per a cada classificador (Classificador 1 i Classificador 2), utilitzats a l'eina, els podem veure a les taules següents (taula 9 i 10):

	Terme independent	β_1	β_2	β_3	β_4
Linear Regression	0.48749	0.00098	0.04474	-0.01403	0.01117
Logistic Regression	-0.04973	-0.00397	0.17959	-0.05618	0.04466

Taula 9: Coeficients utilitzats a cada model de predicció al Classificador 1.

	Terme independent	β_1	β_2	β_3	β_4	β_5
Linear Regression	0.47934	-0.05064	0.04424	-0.01804	0.00270	0.21173
Logistic Regression	-0.04088	-0.26026	0.20075	-0.11300	0.03616	1.24236

Taula 10: Coeficients utilitzats a cada model de predicció al Classificador 2.

Recordem que un regressor lineal modelitza una recta de regressió a partir d'un núvol de punts i que la recta modelada, és la recta més propera que passa per tots els punts. Aquesta recta, la definim de la següent manera:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

on:

Y : variable dependent

X_i : variables independents

β_i : paràmetres que mesuren la influència que tenen les variables independents sobre la variable dependent.

Llavors, en el cas del Classificador 1, a partir dels coeficients obtinguts mitjançant les proves realitzades, tenim que:

$$Y = 0.48749 + 0.00098 X_1 + 0.04474 X_2 + (-0.01403) X_3 + 0.01117 X_4$$

on X_1 , X_2 , X_3 , X_4 representen el valor de les característiques que defineixen un vol, explicades a la secció [4.2.3.1](#): dies d'antelació, demanda, cap de setmana i dia de la setmana, respectivament.

En el cas del Classificador 2, a partir dels coeficients obtinguts mitjançant les proves realitzades, tenim que:

$$Y = 0.47934 + (-0.05064) X_1 + 0.04424 X_2 + (-0.01804) X_3 + 0.00270 X_4 + 0.21173 X_5$$

on X_1 , X_2 , X_3 , X_4 y X_5 representen el valor de les característiques que defineixen un vol, explicades a la secció [4.2.3.1](#): dies d'antelació, demanda, cap de setmana, dia de la setmana i preu, respectivament.

Recordem que en el cas de la regressió logística, la predicció es pot reduir al modelatge d'una regressió lineal del següent estil:

$$Y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}}$$

on:

Y: variable dependent

X_i : variables independents

B_i : paràmetres que mesuren la influència que tenen les variables independents sobre la variable dependent.

Llavors, en el cas del Classificador 1, a partir dels coeficients obtinguts mitjançant les proves realitzades, tenim que:

$$Y = \frac{1}{1 + e^{-((-0.04973) + (-0.00397) X_1 + 0.17959 X_2 + (-0.05618) X_3 + 0.04466 X_4)}}$$

on X_1 , X_2 , X_3 y X_4 representen el valor de les característiques que defineixen un vol, explicades a la secció [4.2.3.1](#): dies d'antelació, demanda, cap de setmana i dia de la setmana, respectivament.

En el cas del Classificador 2, a partir dels coeficients obtinguts mitjançant les proves realitzades, tenim que:

$$Y = \frac{1}{1 + e^{-((-0.04088) + (-0.26026) X_1 + 0.20075 X_2 + (-0.11300) X_3 + 0.3616 X_4 + 1.24236 X_5)}}$$

on X_1 , X_2 , X_3 , X_4 y X_5 representen el valor de les característiques que defineixen un vol, explicades a la secció [4.2.3.1](#): dies d'antelació, demanda, cap de setmana, dia de la setmana y preu, respectivament

A partir d'aquests models de predicció definits, només es tractarà de donar valor a les característiques d'un vol (X_1 , X_2 , X_3 , X_4 en el cas del Classificador 1 y X_1 , X_2 , X_3 , X_4 , X_5 en el cas del Classificador 2) per tal d'obtenir una predicció.

6 Conclusions

Per a dur a terme aquest treball, hem hagut de seguir les etapes d'un projecte de ciència de les dades. Hem pogut transformar les dades en coneixement, que era un dels objectius d'aquest Treball de Fi de Grau. Fer un estudi estadístic dels vols amb la posterior construcció d'un classificador ens serveix com a base del nostre projecte.

Com a resultats de la investigació, és possible concloure resultats com que els dilluns, els dimarts i els dimecres són els millors dies per comprar un bitllet d'avió, que els dissabtes i els diumenges són els pitjors dies per comprar bitllets, que els bitllets comprats en cap de setmana surten més cars que els comprats en dies laborals o que els bitllets d'avió s'han de comprar amb uns 50 a 57 dies d'antelació per tal de què ens surtin més econòmics. Tot i que en alguns resultats no s'ha obtingut els resultats que s'esperaven, s'han explorat nous resultats i conclusions que no s'havien arribat a explorar.

Amb els resultats obtinguts, podem arribar a proveir a l'usuari d'una eina que li permeti comprar els bitllets d'avió en el moment en què el seu preu és el més econòmic possible, evitant que aquest hagi d'estar contrastant contínuament múltiples fonts d'informació.

6.1 Treball futur

Com s'ha explicat al llarg del projecte, l'objectiu principal del projecte era desenvolupar un sistema intel·ligent que permetés recomanar a l'usuari la compra d'un cert bitllet d'avió. El treball futur a implementar en aquest projecte són els següents punts:

- **Augment de la base de dades:** és evident, que si augmentem les nostres dades, la fiabilitat dels resultats del nostre sistema intel·ligent augmentarà, ja que l'anàlisi estadístic dut a terme, tindria una base molt més solida sobre la que treballar.
- **Exploració de noves característiques:** una de les bases del nostre sistema intel·ligent (classificador) són les característiques definides per a cada vol, que són utilitzades per entrenar-lo i que ens doni una predicció el més fiable possible. Per tant, una millora destacable seria l'exploració de noves característiques que ens permetessin millorar la fiabilitat de predicció del nostre sistema.
- **Connexió amb l'aplicació:** actualment el nostre sistema només proveeix a l'eina que l'usuari utilitzarà per interactuar amb ell, dos models de predicció: *linear regression* i *logistic regression*. Un altre de les grans millores que es podrien dur a terme és proveir a l'eina del model de predicció que millor rendiment ha demostrat tenir en els experiments i proves realitzades al treball, el *random forest*.

7 Bibliografia

- [1]. An Introduction to Statistical Learning [Llibre]
Disponible: <http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf>
- [2]. Normalització d'unitat tipificada – Wikipedia [Online]
Disponible: https://en.wikipedia.org/wiki/Standard_score
- [3]. Plana web de Github [Online]
Disponible: <https://github.com/>
- [4]. Plana web Expedia [Online]
Disponible: <https://www.expedia.es>
- [5]. Plana web QPX Express API [Online]
Disponible: <https://developers.google.com/qpx-express/>
- [6]. Llenguatge de programació Python [Online]
Disponible: <https://www.python.org/>
- [7]. Entorn de programació python: Ipython Notebook [Online]
Disponible: <http://ipython.org/notebook.html>
- [8]. Biblioteca informàtica de Python: Pandas [Online]
Disponible: <http://pandas.pydata.org/>
- [9]. Biblioteca informàtica de Python: Numpy [Online]
Disponible: <http://www.numpy.org/>
- [10]. Biblioteca informàtica de Python: Matplotlib [Online]
Disponible: <http://matplotlib.org/>
- [11]. Biblioteca informàtica de Python: Selenium [Online]
Disponible: <http://selenium-python.readthedocs.io/>
- [12]. Biblioteca informàtica de Python: BeautifulSoup [Online]
Disponible: <https://pypi.python.org/pypi/beautifulsoup4>
- [13]. Preprocessing data d'Scikit-learn [Online]
Disponible: <http://scikit-learn.org/stable/modules/preprocessing.html>
- [14]. Cross-Validation d'Scikit-learn [Online]
Disponible: http://scikit-learn.org/stable/modules/cross_validation.html

- [15]. T-Test d'SciPy [Online]
Disponible: http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ttest_ind.html
- [16]. Random Forest Classifier d'Scikit-learn [Online]
Disponible: <http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>
- [17]. Linear Regression d'Scikit-learn [Online]
Disponible: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- [18]. Logistic Regression d'Scikit-learn [Online]
Disponible: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- [19]. Support Vector Machine d'Scikit-learn [Online]
Disponible: <http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>
- [20]. Mesura d'Exactitud d'Scikit-learn [Online]
Disponible: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
- [21]. Classification Report d'Scikit-learn [Online]
Disponible: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
- [22]. Matriu de confusió – Wikipedia [Online]
Disponible: https://en.wikipedia.org/wiki/Confusion_matrix
- [23]. Diagrama de Caixes – Wikipedia [Online]
Disponible: https://en.wikipedia.org/wiki/Box_plot