MODELO PARA IDENTIFICAR LOS VUELOS AFECTADOS POR RETRASOS O CANCELACIONES EN EL AEROPUERTO EL DORADO DE BOGOTÁ, COLOMBIA.

Proyecto de grado

Maestría en Analítica de
Datos para la Inteligencia de Negocios

CAMILO ANDRÉS CHAVARRO CELY;WILLIAM ALFONSO RAMIREZ QUIROGA;CARLOS ALBERTO ARIAS MAURY
caarias@javeriana.edu.co
Camilo Andrés Chavarro Cely
camilo_chavarro@javeriana.edu.co
William Alfonso Ramírez Quiroga
william.ramirezq@javeriana.edu.co

Contenido

1 Información General del Proyecto 6

2 Justificación del Proyecto 6

2.1 Realidad del Sector del Transporte Aéreo de Pasajeros 6

3 Descripción y Contexto del Proyecto 7

3.1 Regulación Vigente para Retrasos y Cancelaciones de Vuelos en Colombia 7

3.1.1 IATA 7

3.1.2 Aerocivil 7

3.2 Descripción General Situación de las Aerolíneas más importantes del mercado colombiano a julio de 2020 8

4 Objetivos 9

4.1 Objetivo General 9

4.2 Objetivos Específicos 10

4.3 Criterios de éxito: 10

5 Participantes 10

6 Inventario de Fuentes Disponibles 10

6.1 Descripción de los datos 11

7 Entregables 13

8 Acercamientos previos que se han tenido al problema 13

9 Disponibilidad de Tecnología para el Proyecto 15

10 Restricciones 16

11 Metodología propuesta para emplear en este trabajo 16

11.1 Entendimiento de los datos 16

11.1.1 Calidad de los datos 16

11.2 Exploración de los datos 19

12 Procesamiento de los datos 24

12.1.1 Filtros aplicados a la base de datos 25

12.1.2 Definición de variables numéricas y categóricas 25

12.1.3 Diagrama de correlación 26

12.1.4 Selección final de variables para el proceso de modelamiento 26

12.1.5 Variables dummy (One – Hot Encoding) 27

12.1.6 Eliminación de variables para evitar multicolinealidad 27

12.1.7 Transformación y normalización de variables numéricas 28
12.1.8 Selección de técnicas y supuestos 28
12.1.9 Conjunto de datos de entrenamiento y prueba 28
13 Modelamiento 28
13.1 Regresión logística 29
13.1.1 Modelo 29
13.2 Redes Neuronales 33
13.2.1 Introducción red neuronal feedforward 33
13.2.2 Librería NNET (R Studio) 34
13.2.3 Modelo 34
13.3 XGBoosting 36
13.3.1 Concepto de Gradient Boosting 36
13.3.2 Variante eXtreme Gradient Boosting 36
13.3.3 Librería xgboost (CRAN) 37
13.3.4 Modelo 37
13.3.5 Cuadro comparativo de los tres modelos creados 39
14 Conclusiones 40
15 Siguientes pasos 41
16 Bibliografía 42

Tablas

Tabla 1 Información general del proyecto. 6
Tabla 2 Participantes del proyecto. 10
Tabla 3 Reporte de cumplimiento de itinerarios 11
Tabla 4 Tráfico por equipo. 12
Tabla 5 Datos del clima. 13
Tabla 6 Entregables 13
Tabla 7 Visualización de impacto por hora del día y por mes para variables climáticas categóricas .. 24
Tabla 8 Variables numéricas y categóricas 26
Tabla 9 One-Hot Encoding 27
Tabla 10 Base de entrenamiento y pruebas por tiempo de retraso 29
Tabla 11 Variables seleccionadas por significancia 31
Tabla 12 Matriz de confusión modelo Regresión logística (60 minutos) 31
Tabla 13 Métricas modelo Regresión Logística (60 minutos) 31
Tabla 14 Resumen Odd ratio y Probabilidades del Modelo sin Balanceo 32
Tabla 15 Selección de neuronas en la capa oculta 34
Tabla 16 Matriz de confusión modelo red neuronal (60 minutos) 35
Tabla 17 Métricas modelo red neuronal (60 minutos) 35
Tabla 18 Matriz de confusión modelo XGBoosting (60 minutos) 37
Tabla 19 Comparativo de Regresión Logística, Redes Neuronales y XGBoosting. 39

Ilustraciones

Ilustración 1 Modelo conceptual 13
Ilustración 2 Etapas de datos del proyecto 15
Ilustración 3 Flujos de trabajo para revisión de Calidad de Datos 16
Ilustración 4 Flujos de trabajo para corrección de formatos Fecha 17
Ilustración 5 Flujos de trabajo para homogenización de valores de diferentes variables 18
Ilustración 6 Instancias de alta diferencia entre fecha de programación y remolque. 18
Ilustración 7 Porcentaje de vuelos por aerolínea. 19
Ilustración 8 Top ciudades origen con mayor tráfico. 19
Ilustración 9 Grado de centralidad de la red 20
Ilustración 10 Porcentaje de estado de los vuelos en el tiempo. 20
Ilustración 11 Distribución de vuelos cancelados por Aerolínea y por día del mes. 20
Ilustración 12 Distribución de vuelos retrasados por Aerolínea y por día del mes. 21
Ilustración 13 Histogramas fuente de datos tráfico por equipo (mensual) 21
Ilustración 14 Porcentaje de afectación de vuelos por año y día del mes. 22
Ilustración 15 Porcentaje de afectación de vuelos por año y de la semana 22
Ilustración 16 Variables climáticas numéricas principales y su variabilidad por hora del día 23
Ilustración 17 Variables climáticas numéricas principales y su variabilidad por mes 23
Ilustración 18 Flujo de la preparación y evaluación de modelos 24
Ilustración 19 Diagrama de correlación 26
Ilustración 20 Curva Roc modelo Regresión Logística a 15, 60 y 90 minutos 33
Ilustración 21 Red Neuronal feedforward 34

Ilustración 22 Rendimiento Red Neuronal..... 35
Ilustración 23 Curva ROC modelo red neuronal 15, 60 y 90 minutos 36
Ilustración 24 Curva ROC modelo XGBoosting 15, 60 y 90 minutos..... 38
Ilustración 25 Métricas de Importancia del modelo basado en XGBoosting 39

1 Información General del Proyecto

Título del proyecto	***Modelo para identificar los vuelos afectados por retrasos o cancelaciones en el aeropuerto El Dorado de Bogotá, Colombia.***
Universidad(es) ejecutora(s)	*Pontificia Universidad Javeriana*
Director general del proyecto	*Luis Manuel Pulido*
Fecha de aprobación del proyecto	*1ro de junio de 2020*

Tabla 1 *Información general del proyecto.*

2 Justificación del Proyecto

2.1 Realidad del Sector del Transporte Aéreo de Pasajeros

Los cambios experimentados por el sector turismo en los últimos años han sido de gran magnitud y con consecuencias de alto impacto para las aerolíneas. La aparición de plataformas de ventas de paquetes turísticos a bajo costo (portales de venta de tiquetes compitiendo entre ellos, Airbnb, etc.) ha llevado a crear un ambiente de competencia tal que los márgenes de negocio son supremamente reducidos y la rentabilidad se ve afectada enormemente.

En adición a lo mencionado, las aerolíneas de bajo costo han surgido en los últimos años como una nueva opción para los consumidores, creando mayores retos para las ya establecidas y haciendo mucho más difícil la viabilidad económica de los diferentes participantes del sector.

Como un alto impacto reciente, la pandemia generada por la propagación del COVID-19 ha puesto a las aerolíneas en una situación económica crítica y sin precedentes. Los gobiernos de diferentes países buscan mecanismos para proteger el sector y evitar escenarios generalizados de quiebra que impactarían el comercio, la competitividad y el desarrollo económico en general.

A nivel mundial, la industria del transporte aéreo maneja estándares de calidad y cumplimiento que cada país adapta por medio de su regulación particular (por ejemplo, un retraso en general se considera una diferencia mayor a 15 minutos entre el tiempo programado de salida de un vuelo y el tiempo real en el que tuvo efecto). En esas regulaciones se especifican umbrales de calidad de servicio esperados en términos de indicadores de cumplimiento y tiempos de respuesta.

Si bien la situación económica de las aerolíneas ya es crítica en el momento de escribir este documento, lograr reducir las penalidades se convierte en un tema de vital importancia para las aerolíneas. La cantidad de vuelos durante la pandemia se redujo y en otros casos hubo cancelación total de la operación, pero a 16 de junio de 2020 ya se tiene noticia de reactivación del sector en diferentes países no sólo de Europa, sino también en América Latina. En síntesis, el estudiar la minimización del impacto económico por penalidades no pierde vigencia. De hecho, se hace más relevante para poder evitar dentro de lo posible mayores impactos económicos para las aerolíneas, evitando más afectaciones a su flujo de caja y, por ende, reduciendo la posibilidad de enfrentar una posible bancarrota y su respectivo impacto, a nivel de empleados, proveedores y del sector en general.

3 Descripción y Contexto del Proyecto

Generalmente cada país tiene una estructura de penalidades económicas al no cumplir niveles mínimos de cancelaciones y retrasos en los vuelos programados. En el caso particular de Colombia, las aerolíneas con operación dentro del territorio nacional están sujetas a cumplir los lineamientos y estándares de calidad dictados por la International Air Transport Association (IATA) e interpretados y vigilados localmente por la Aeronáutica Civil (Aerocivil).

3.1 Regulación Vigente para Retrasos y Cancelaciones de Vuelos en Colombia

En esta sección se introduce un marco general de las regulaciones y políticas relevantes al transporte aéreo de pasajeros en Colombia.

3.1.1 IATA

International Air Transport Association (IATA) es una asociación de carácter comercial que reúne a 290 aerolíneas, equivalente al 82% de tráfico aéreo total a nivel mundial. Esta asociación tiene entre otros propósitos la de emitir estándares y recomendaciones de operación que a su vez son empleadas por los gobiernos para construir el marco regulatorio pertinente en cada país. Como misión, propone el servir, representar y liderar la industria aeronáutica a nivel mundial.

Como prioridades, IATA establece: seguridad, sostenibilidad ambiental, rebalanceo de la cadena de valor (habilitando una reducción continua de costos de operaciones y facilitar un crecimiento sostenible) y facilitar la definición de marcos regulatorios.

3.1.2 Aerocivil

La Aeronáutica Civil es el resultado de la fusión del Departamento Administrativo de Aeronáutica Civil y el Fondo Aeronáutico Nacional, ordenado par el Artículo 67 del Decreto 2171 de 1992 (http://www.aerocivil.gov.co). En la actualidad, la entidad se rige por el Decreto 260 del 28 de enero de 2004 con un nuevo ordenamiento administrativo y con nuevas dependencias.

Con la Ley 105 del 30 de diciembre de 1993, se adscribe la Aeronáutica Civil al Ministerio de Transporte como órgano rector de la política y ejecución de las funciones relativas al transporte aéreo bajo principios de seguridad, oportunidad y eficiencia.

Dentro de los marcos vigentes de regulación relacionados con el reporte de desempeño de las aerolíneas en Colombia a Aerocivil, son relevantes los siguientes documentos:

3.1.2.1 Circular 02

Este documento (*Circular Informativa 02 Version 02b Aerocivil*, n.d.) relaciona el formato y procedimientos que deben seguir las aerolíneas en Colombia para el reporte de gestión y cumplimiento de estándares. Es la guía actual de la cual se derivan diferentes políticas y regulaciones.

3.1.2.2 Regulación RAC 13 (Régimen Sancionatorio)

Esta regulación, mencionada en la circular 02, es la que define las penalidades por incumplimiento de estándares de calidad, aplicables a varias instancias, dentro de las cuales se encuentran las de retraso en vuelos y cancelación por causas internas o propias de la aerolínea y no atribuibles a terceros o condiciones externas, como las meteorológicas.(Aerocivil, 2020)

3.2 Descripción General Situación de las Aerolíneas más importantes del mercado colombiano a julio de 2020

La situación actual originada por la pandemia del coronavirus a nivel mundial ha creado una crisis económica, social e inclusive política de dimensiones alarmantes: recesión, incremento de la pobreza y del desempleo, hambre, miles de muertos, todo lo anterior agravado por las medidas de desconfinamiento ordenado por los diferentes Gobiernos, que aumentará el número de contagiados en el mundo, entre otros factores, lo cual hace prever unas consecuencias de resultados catastróficos. Así lo asegura (France 24, 2020) (1) "quinientos millones de personas podrían ser arrastradas de nuevo a la pobreza". Se escuchan fuertes críticas contra la Globalización. La vida cambiará sustancialmente.

La industria de la aviación nacional e internacional no es ajena a la situación descrita anteriormente. Las aerolíneas han visto incrementadas su crisis, debido a la parálisis general de la aviación, que aún pervive en América Latina y apenas inicia sus operaciones en Europa y Asia. Según (economista.es, junio 2020) (2), cinco son las razones fundamentales que dificultan la recuperación de las aerolíneas en 2021: "La crisis de confianza y de demanda, la incertidumbre de los viajeros para retomar los vuelos, las deudas que han acumulado las diferentes aerolíneas, el aumento de los costos y la conflictividad laboral y las pocas ayudas ofrecidas y recibidas por parte del Estado".

La crisis que vive el sector aeronáutico en América Latina es aún mayor, como es obvio, agravado por la situación económica de los diferentes países. En la actualidad de escuchan voces de protesta por la próxima reactivación del sector sin que se haya llegado el pico de la pandemia, lo cual hará incrementar el número de contagiados. Según (Diana Marcela Tinjacá, La Patria, junio 22 de 2020) (3) plantea cinco claves que amenaza a las aerolíneas en Latinoamérica: "las peores cifras de la historia en número de pasajeros transportados, los tres meses de paralización presentados, las dudas sobre la reapertura ante el incremento de casos de contagio, la crisis financiera de grandes aerolíneas como Latam y Avianca y el escaso apoyo gubernamental, ayudas fiscales y renegociación de deudas".

"Las aerolíneas en América Latina estarán en una situación crítica y necesitan urgentemente el apoyo de los gobiernos para poder sobrevivir. Así lo afirmó la Asociación de Transporte Aéreo Internacional –IATA- en (A21 mx, 15 de junio 2020) (4).

Colombia se mantiene en una etapa de crecimiento acelerado sin llegar aún al pico de la pandemia. Si bien es cierto que se habla de planes piloto para la reactivación de la industria que se iniciarían en el mes de julio para los vuelos nacionales y los internacionales en el mes de septiembre, también es cierto que esto se haría realidad de acuerdo con el análisis concreto de la situación que en ese momento se presenten. La programación del día sin IVA del 19 de junio, durante el cual se eximía del cobro de este impuesto a algunos productos como electrodomésticos, aparatos de computación, hogar, entre otros, hizo volcar a la población, a las diferentes Grandes Superficies para hacerse acreedores de estos descuentos violando las mínimas normas de bioseguridad exigidas. Esto hace predecir incrementos excesivos en los números de enfermos por la Covid-19, que impedirá la posibilidad de la esperada reactivación económica, afectando de igual manera a la industria aeronáutica.

Si particularizamos en algunas aerolíneas nacionales, nos daremos cuenta de lo planteado anteriormente:

AVIANCA surge como SCADTA en 1919 y es la segunda más antigua del mundo. El domingo 10 de mayo se acogió al capítulo 11 del Código de Bancarrotas de Estados Unidos, con lo cual ganaría tiempo para renegociar su deuda, que actualmente asciende a la cifra de 7.000 millones de dólares. Con voces a favor y en contra (estas últimas por considerar que tributa en Panamá y ya no es una aerolínea nacional), el Gobierno busca darle un salvavidas económico a la Empresa.

VIVA AIR es una aerolínea de bajo costo comercial de pasajeros y solicitó un crédito de 50 millones de dólares para enfrentar su crisis. Tal como lo afirmó su presidente en declaraciones a (Portafolio, junio 5 de 2020) (5): "No estamos solicitando ni rescates ni subsidios, pues entendemos las realidades que afronta el país y los esfuerzos que se han puesto en marcha".

AEROREPÚBLICA cambió su nombre por el de "Copa Airlines Colombia" para continuar operando en las mismas rutas que Aerorepublica. La Empresa manifiesta que es difícil resistir la dura situación de la pandemia, a pesar de reducir sus costos y buscar vender 350 millones de dólares en bonos a cinco años convertibles en acciones. Así lo reafirmó la Aerolínea (Reportur, 27 de abril 2020) (6): "que podría consumir casi el 70% de su efectivo disponible para fin de año".

EASYFLY (Empresa aérea de servicios y facilitación logística integral), aerolínea de bajo costo de vuelos regionales, como todas las aerolíneas regionales, ha sufrido las duras secuelas de la pandemia, como así lo hicieron saber en (Reportur, 21 abril 2020) (7): "La emergencia por la pandemia Covid-19 ha puesto en grave riesgo a las aerolíneas como es el caso de la colombiana Easyfly que proyectó pérdidas por más de $30.000 millones a mayo, por lo que va a cancelar rutas y frecuencias y otras medidas como la cancelación de contratos de empleados y la suspensión de compra de aviones".

"Esta es nuestra última oportunidad para sortear la crisis. Estamos contra el tiempo, cada día que pasa, suma a la agonía de una industria que necesita claridad sobre las fechas de regreso a la operación, para poder activarse comercial y operativamente", dijo Peter Cerdá, vicepresidente regional de IATA para las Américas en (A21 mx, 15 de junio 2020) (8).

En este sentido, las diferentes aerolíneas nacionales están preparando las medidas de seguridad requeridas para retornar a las operaciones. De estas medidas la más cuestionada, inclusive por la misma IATA, ha sido la del distanciamiento social durante los vuelos (dejar la silla del medio vacía) y la aplicación de la cuarentena a la llegada.

El ministro del interior autorizó un plan piloto para el retorno de los vuelos nacionales, bajo la verificación de indicadores y comportamiento del virus en el origen y destino y el cumplimiento de los siguientes protocolos: distanciamiento físico, controlar el ingreso al terminal únicamente de las personas que trabajan en los aeropuertos y pasajeros con documento de identidad y el "check in" electrónico, realizar toma de temperatura, tener las sillas intermedias inhabilitadas, entre otras. La primera ruta autorizada será entre el aeropuerto internacional Palonegro, de Bucaramanga y el aeropuerto Internacional Camilo Daza, en Cúcuta.

4 Objetivos

4.1 Objetivo General

Identificar los vuelos programados en el aeropuerto El Dorado de Bogotá (Colombia) que podrían ser afectados por retrasos o cancelaciones.

4.2 Objetivos Específicos

- Proponer un modelo que permita identificar los vuelos con posibles afectaciones por retrasos o cancelaciones en el aeropuerto El Dorado de Bogotá (Colombia) caracterizando las variables operacionales y meteorológicas con mayor relevancia.
- Identificar los factores meteorológicos que inciden en las afectaciones por retrasos o cancelaciones de vuelos que parten o tienen como destino el aeropuerto El Dorado de la ciudad de Bogotá.
- Identificar los factores operacionales de capacidad instalada del aeropuerto El Dorado de Bogotá que tienen mayor relevancia en el retraso o cancelación de vuelos programados en este terminal aéreo.

4.3 Criterios de éxito:

Determinar las variables que permiten clasificar los vuelos programados que podrían ser afectados por retrasos o cancelaciones en el aeropuerto El Dorado que presta servicio al área metropolitana de la ciudad de Bogotá.

El modelo de clasificación se evaluará mediante la curva ROC y métricas de clasificación como: accuracy, precisión, sensitividad, especificidad, entre otros.

5 Participantes

Equipos de Investigación	**Participante**	**Rol**	**Responsabilidades**
Pontificia Universidad Javeriana	Luis Manuel Pulido	*Director del proyecto*	*Planear y coordinar todas las actividades del proyecto.*
	Camilo Andrés Chavarro Cely	*Estudiante miembro del proyecto.*	*Desarrollar el proyecto.*
	Carlos Alberto Arias Maury	*Estudiante miembro del proyecto.*	*Desarrollar el proyecto.*
	William Alfonso Ramirez	*Estudiante miembro del proyecto.*	*Desarrollar el proyecto.*

Tabla 2 *Participantes del proyecto.*

6 Inventario de Fuentes Disponibles

Los conjuntos de datos adquiridos son:

Reporte de cumplimiento de itinerarios (horario): estado de los vuelos de las principales aerolíneas con operación en los aeropuertos de Colombia para los años 2017 y 2018. La base de datos está agrupada por aerolínea, origen, destino y número de vuelo para un día y hora de programación determinada.

Tráfico por equipo (mensual): operación comercial de todos y cada uno de los trayectos efectuados por las principales aerolíneas con operación en Colombia en los años 2017 y 2018. La base de datos está agrupada por aerolínea, origen, destino y tipo de equipo para un mes determinado.

Datos del clima (horario): archivo histórico de clima del Aeropuerto Internacional El Dorado de la ciudad de Bogotá de los años 2017 y 2018, donde se incluyen datos de temperatura, precipitación, viento y muchos más.

6.1 Descripción de los datos

Se realiza la descripción de los atributos de las tablas Reporte de cumplimiento de itinerarios, tráfico por equipo y clima:

Reporte de cumplimiento de itinerarios (horario):

#	Atributo	Descripción	Tipo
1	Tráfico	Nacional o Internacional	Texto
2	Aerolínea	Nombre comercial de la empresa.	Texto
3	Origen	Corresponde a la sigla IATA del aeropuerto donde se origina el trayecto.	Texto
4	Destino	Corresponde a la sigla IATA del aeropuerto donde termina el trayecto.	Texto
5	# Vuelo	Número del vuelo con el cual se realiza la operación.	Texto
6	Fecha programada de salida referencia Score	Fecha internacional (UTC) en la cual está programada y publicada la operación para su salida.	Fecha
7	Hora programada de salida referencia Score	Hora internacional (UTC) en la cual está programada y publicada la operación para su salida.	Hora
8	Fecha remolque UTC	Fecha internacional (UTC) en la cual se ejecutó la operación atribuible al momento en que la aeronave es remolcada y apartada de la puerta de embarque.	Fecha
9	Hora remolque UTC	Hora internacional (UTC) en el cual la aeronave inicia su remolque desde la plataforma hacía la pista o enciende el primer motor, lo que ocurra primero.	Hora
10	Demora (HH:MM)	Resta entre la Hora de remolque (UTC) y Hora programada de salida de referencia Score.	Hora
11	Estado del vuelo (variable objetivo)	Si se cumplió, canceló, demoró adelantó el vuelo programado en el itinerario aprobado para la Aeronáutica Civil.	Texto
12	Código de demora	Número referente IATA asignado para la identificación de un motivo de demora y/o cancelación de un vuelo.	Texto
13	Motivo de demora	Incontrolables, operacionales, técnicos.	Texto
14	Observaciones	Espacio para comentario de la aerolínea.	Texto

Tabla 3 *Reporte de cumplimiento de itinerarios*

Tráfico por equipo (mensual):

#	Atributo	Descripción	Tipo
1	Sigla Empresa	Sigla OACI con la cual se identifica la empresa ante las Autoridades Aeronáuticas.	Texto
2	Fecha	Año y Mes de operación UTC.	Fecha
3	Origen	Corresponde a la sigla IATA del aeropuerto donde se origina el trayecto.	Texto
4	Destino	Corresponde a la sigla IATA del aeropuerto donde termina el trayecto.	Texto
5	Tipo de Equipo	Código del tipo de equipo utilizado para realizar la correspondiente etapa del itinerario.	Texto

6	***Número de Vuelos***	*Número de vuelos realizados en el mes de referencia, para el correspondiente trayecto.*	*Número*
7	***Horas Bloque***	*Tiempo transcurrido entre calzos.*	*Número*
8	***Sillas***	*Número total de asientos de pasajeros disponibles para la venta.*	*Número*
9	***Carga Ofrecida Kg***	*Capacidad de carga total, encima y debajo de la cubierta, disponible para el transporte de carga y correo.*	*Número*
10	***Pasajeros A Bordo***	*Número de pasajeros de pago (remuneración comercial) que son transportados.*	*Número*
11	***Pasajeros Transito***	*Número de pasajeros en tránsito que son transportados.*	*Número*
12	***Carga Transito***	*El total de carga tránsito transportada.*	*Número*
13	***Distancia***	*Distancia en kilómetros entre los aeropuertos del respectivo trayecto.*	*Número*
14	***Trafico***	***N** Tráfico Doméstico. **I** Tráfico Internacional. **E** Tráfico entre dos aeropuertos fuera de Colombia.*	*Texto*
15	***Tipo de Vuelo***	***R** Operación regular. **A** Vuelos adicionales. **C** Vuelo chárter. **T** Taxi Aéreo (ala fija).*	*Texto*
16	***Carga a Bordo Kg***	*El total carga de pago transportada.*	*Número*
17	***Correo A Bordo Kg***	*El total de correo transportado (correspondencia y otros objetos enviados por una administración postal para ser entregados a otra).*	*Número*
18	***Nombre Empresa***	*Nombre comercial de la empresa.*	*Texto*
19	***Ciudad Origen***	*Ciudad Origen.*	*Texto*
20	***Ciudad Destino***	*Ciudad Destino.*	*Texto*
21	***País Origen***	*País Origen.*	*Texto*
22	***País Destino***	*País Destino.*	*Texto*
23	***Aeropuerto Origen***	*Aeropuerto Origen.*	*Texto*
24	***Aeropuerto Destino***	*Aeropuerto Destino.*	*Texto*

Tabla 4 *Tráfico por equipo.*

Datos del clima (horario):

#	*Atributo*	*Descripción*	*Tipo*
1	***Nombre de la ciudad***	*Nombre de la ciudad.*	*Texto*
2	***Latitud***	*Latitud.*	*Número*
3	***Longitud***	*Longitud.*	*Número*
4	***Temperatura***	*Temperatura (°C).*	*Número*
5	***Percepción***	*Percepción humana del clima.*	*Número*
6	***Presión***	*Presión atmosférica (sobre el nivel del mar), hPa.*	*Número*
7	***Humedad***	*Humedad (%)*	*Número*
8	***Temperatura mínima***	*Desviación de la temperatura que aplica para las grandes ciudades y las megalópolis expandidas geográficamente.*	*Número*
9	***Temperatura máxima***	*Desviación de la temperatura que aplica para las grandes ciudades y las megalópolis expandidas geográficamente.*	*Número*
10	***Velocidad del viento***	*Velocidad del viento. Unidad: metro / segundo*	*Número*
11	***Grado del viento***	*Dirección del viento (Grados)*	*Número*
12	***Nubosidad***	*Nubosidad (%)*	*Número*
13	***Lluvia 1 hora***	*Volumen de lluvia para la última hora (mm)*	*Número*
14	***Lluvia 3 horas***	*Volumen de lluvia para las últimas tres horas (mm)*	*Número*

15	**Nieve 1 hora**	*Volumen de nieve de la última hora en estado líquido (mm)*	*Número*
16	**Nieve 3 horas**	*Volumen de nieve de las últimas tres horas en estado líquido (mm)*	*Número*
17	**Identificación del clima**	*Identificación de la condición del clima*	*Número*
18	**Clima principal**	*Grupo de parámetros climáticos.*	*Texto*
19	**Descripción del clima**	*Condiciones climáticas dentro del grupo.*	*Texto*
20	**Icono del tiempo**	*Icono del tiempo*	*Texto*
21	**Date (ISO)**	*Fecha y hora en formato UTC*	*Fecha*

Tabla 5 *Datos del clima.*

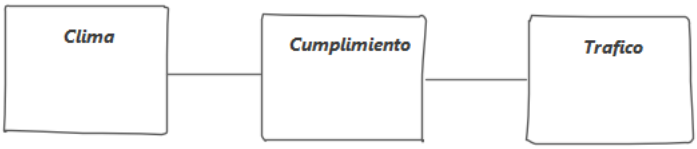


Ilustración 1 *Modelo conceptual*

En el modelo conceptual se evidencia la forma como las diferentes entidades se relacionan, en este caso, las entidades de clima y tráfico tienen información que complementa a la entidad de cumplimiento por medio de atributos comunes como origen, día y hora en el caso del clima y de origen, destino y fecha de los vuelos programados para la información de tráfico.

7 Entregables

Entregable	**Descripción**	**Propiedad**
Documento del contexto del proyecto y entendimiento de los datos	*Documento en formato PDF.*	*Camilo A. Chavarro C.* *Carlos A. Arias M.* *William A. Ramirez Q.*
Presentación de los temas mencionados en el documento de contexto y entendimiento de los datos	*Presentación Power Point*	
Documento del proceso de modelación y resultados.	*Documento en formato PDF*	
Presentación de los temas mencionados en el documento de modelación y resultados.	*Presentación Power Point*	

Tabla 6 *Entregables*

8 Acercamientos previos que se han tenido al problema

Si bien el tema ha sido revisado desde diferentes perspectivas y se ha encontrado material de investigación realizado en los últimos 10 años, las referencias son en su mayoría de actividad aérea en Estados Unidos (E.U.), China, Europa e India. Se tiene referencia de un solo estudio realizado en América Latina (Brasil) y no se tiene referente alguno de este tipo para el caso específico de Colombia. En total se revisaron 47 trabajos de investigación.

De la revisión del material encontrado, uno de los estudios más citados es el realizado por (Rebollo & Balakrishnan, 2014), el cual analiza el problema de retrasos desde una perspectiva de red de todo el sistema de aeropuertos de Estados Unidos y aeropuertos asociados por alguna ruta internacional. Dado el gran volumen y la baja relevancia de algunos aeropuertos con menos de un vuelo diario, los autores se concentran en los 100 peores “links” (parejas de Origen y Destino). Estos autores utilizaron

análisis no supervisado (clusterización por k-means) para definir los núcleos de actividad aérea más representativos de la red de E.U. Por otra parte, realizaron una exploración para determinar potenciales factores de peso, y finalmente aplicaron Random Forest (RF) con el fin de predecir si un vuelo se retrasaría o no en el intervalo de las siguientes 24 horas. Se ejecutaron varios ejercicios con diferentes condiciones de retraso (demora mayor a 45, 60 y 90 minutos), logrando mejorar los resultados a medida que la condición de retraso se hacía mayor (condición de retraso como demora >90 min con mejor desempeño que la de >a 60 min y ésta a su vez mejor que la de > 45 min). Desde esta referencia ya se empieza a considerar Random Forest como una de las opciones de herramientas para predicción, dada la facilidad explicativa asociada a las técnicas derivadas del concepto de árboles de decisión.

Un estudio con buen detalle y extensión es el de (Seelhorst, 2014) que considera tanto retrasos como cancelaciones (la mayoría de la literatura se enfoca en el análisis de los retrasos) a pesar de ser las cancelaciones en este caso de estudio de mucho menor ocurrencia que los retrasos, pero se identifican como un contribuyente de importancia en los retrasos a nivel de red de aeropuertos. Aplicando regresión logística y teoría de colas, el autor llega a proponer un modelo general y presenta un caso de estudio del aeropuerto de San Francisco en el momento en el que tuvo una etapa de remodelación y presentó altos niveles de congestión. Este estudio fue el de mayor detalle y extensión de la literatura encontrada para este ejercicio.

Otro análisis frecuentemente citado es el de (AhmadBeygi et al., 2008) que plantea los impactos de retrasos en cadena y tiene como objetivo principal entender la relación entre la gestión de tripulación y personal de apoyo y los retrasos de vuelo. Los autores proponen el concepto de "árboles de propagación" como herramienta para conceptualizar el problema de los retrasos acumulativos que se van dando a medida que un vuelo retrasado afecta los subsecuentes asociados a la misma aeronave en la agenda inmediata y de obvias consecuencias. Como conclusiones a resaltar, mencionan que en el caso analizado sólo el 40% de los vuelos generaba un efecto acumulativo en los retrasos, y que el interrumpir en lo mínimo la cercanía entre tripulación y naves parece ser un factor determinante en la eficiencia de las operaciones aéreas.

Adicionalmente, se encuentra solamente un estudio en América Latina, aplicando reglas de asociación a observaciones de tráfico aéreo en Brasil en 2014. Como conclusiones se mencionan variables meteorológicas como las de mayor peso, y se menciona también menor probabilidad de retraso en las horas de la mañana, pero que se incrementa fuertemente en caso de niebla.

Otras técnicas utilizadas en los documentos revisados incluyen Gradient Boosting, Support Vector Machines (SVMs), Redes Neuronales y Redes Bayesianas. De estos métodos y los mencionados previamente surge la preferencia de utilizar aquellos relacionados con árboles de decisión, dada su facilidad de implementación y a diferencia de las redes neuronales, el resultado es relativamente sencillo de entender y puede llevar a explicar más fácilmente las variables que resulten aportando mayor peso a la condición de retraso o cancelación.

Como temas no trabajados o que merecieran la pena ser revisados, se podrían mencionar:

- Varios de los casos parten de una capacidad de cada aeropuerto para manejar el tráfico aéreo. Con la información IATA disponible y los registros por cada país sería interesante calcular la capacidad real de un aeropuerto: la que permite la operación con un mínimo de congestión tolerable. Esto puede llevar a proponer prioridades para la expansión de infraestructura a nivel de aeropuertos.
- Un acercamiento al impacto económico que cada aerolínea sufre por concepto de penalidades asociadas a retrasos y cancelaciones.

9 Disponibilidad de Tecnología para el Proyecto

A continuación, se describen los recursos disponibles para el desarrollo del presente proyecto:

Software:

- ***KNIME:*** plataforma de minería de datos que permite la creación de ETL y modelos en un entorno visual. Está construido bajo la plataforma Eclipse.
- ***Microsoft SQL Server:*** sistema de gestión de base de datos relacional, desarrollado por la empresa Microsoft.
- ***Power BI:*** herramienta de visualización interactivas y capacidades de inteligencia empresarial, desarrollado por la empresa Microsoft.
- ***R Studio:*** es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicado a la computación estadística y gráficos**.**

Hardware:

- ***Servidor Contabo:*** 30 GB de RAM, procesador AMD EPYC 7282 16 - Core CPU@ 2.80 GHz, sistema operativo Windows Server 2016 Datacenter de 64 bits.
- ***Equipo personal Camilo Chavarro****:* ASUS, 8 GB de RAM, Intel i7 – 5500 U CPU@ 2.40 GHz 2.39 GHz, sistema operativo Windows 10 de 64 bits.
- ***Equipo personal Carlos Arias:*** Lenovo T480, 16GB de RAM, Intel® Core™ i7-8650U CPU@ 1.90 GHz 2.11GHz, sistema operativo Windows 10 Pro de 64 bits.
- ***Equipo personal William Ramirez:*** ASUS, 8 GB de RAM, intel i7 – UX390U, sistema operativo Windows 10 de 64 bits.

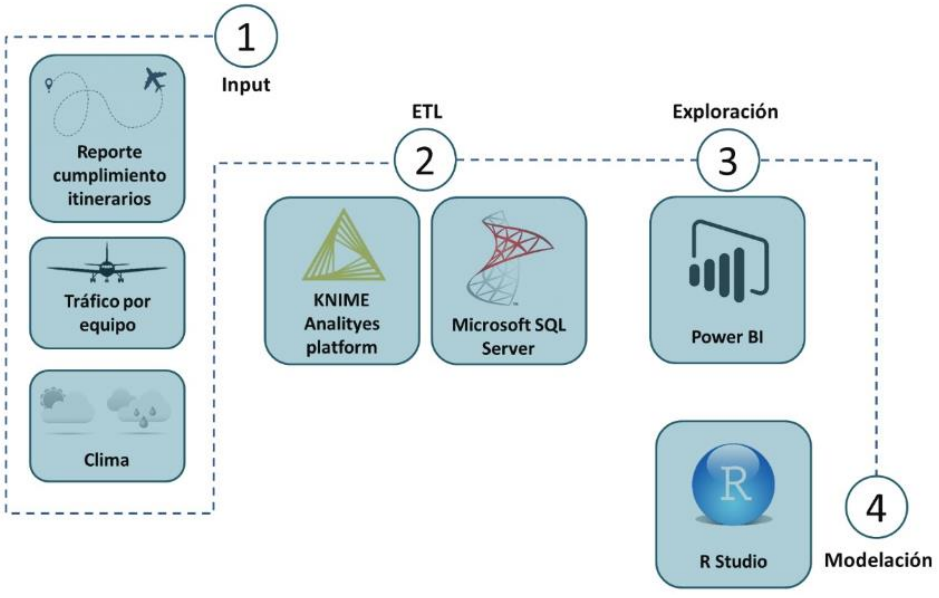


Ilustración 2 *Etapas de datos del proyecto*

10 Restricciones

No fue posible acceder a la cantidad de sillas disponibles y a la cantidad de pasajeros transportados en cada uno de los vuelos que hacen parte de la fuente de datos de cumplimiento, para esta información se tomó el promedio de pasajeros y sillas disponibles obtenidas en la fuente de tráfico aéreo mensual.

No fue posible obtener las características operacionales propias de cada aerolínea, la información de programación de tripulación, y el mantenimiento de aerolíneas podrían aportar información valiosa al objetivo de este documento.

11 Metodología propuesta para emplear en este trabajo

11.1 Entendimiento de los datos

11.1.1 Calidad de los datos

Como insumo inicial de datos se tiene la información de cumplimiento de itinerarios de los años 2017 y 2018 en formato XLSX. A continuación, se describe el proceso de transformación, calidad y limpieza de datos:

11.1.1.1 Revisión de Registros Duplicados

Para el proceso de ETL se utiliza la herramienta KNIME que mediante flujos sencillos permite entender, transformar y limpiar los datos. En la primera etapa se eliminan los registros duplicados del 2017 (0.1%) y del 2018 (0.3%):

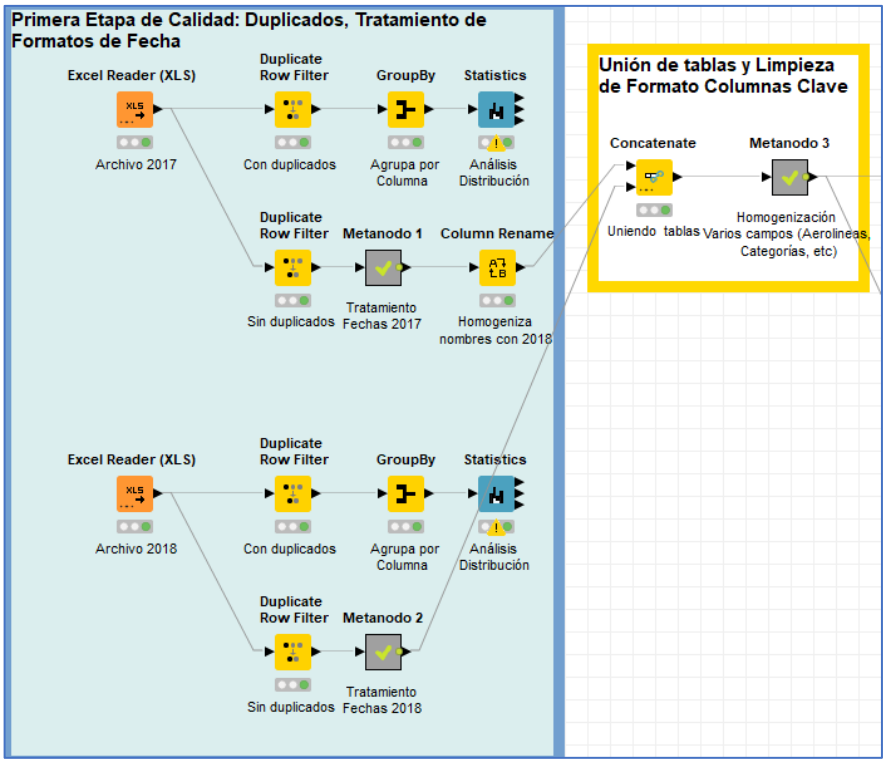


Ilustración 3 *Flujos de trabajo para revisión de Calidad de Datos*

11.1.1.2 Revisión de Integridad de Formatos

Se encontraron algunas instancias de diferencia significativa de formatos de una misma variable. Para el archivo de 2017 fueron observados al menos 4 formatos de fecha (fecha programada y fecha de remolque) diferentes que necesitaron un tratamiento particular:

- *YYYY-MM-DD*
- *DD/MM/YYYY*
- Una variación particular para algunas muestras de septiembre de 2017 expresadas como: *DDSEPYYYY*
- El cuarto formato corresponde al valor numérico de Excel para las fechas. Este formato toma como referencia el 1ro de enero del año 1900, y cada unidad representa 1 día. La exigencia del flujo de KNIME fue mayor para este caso.

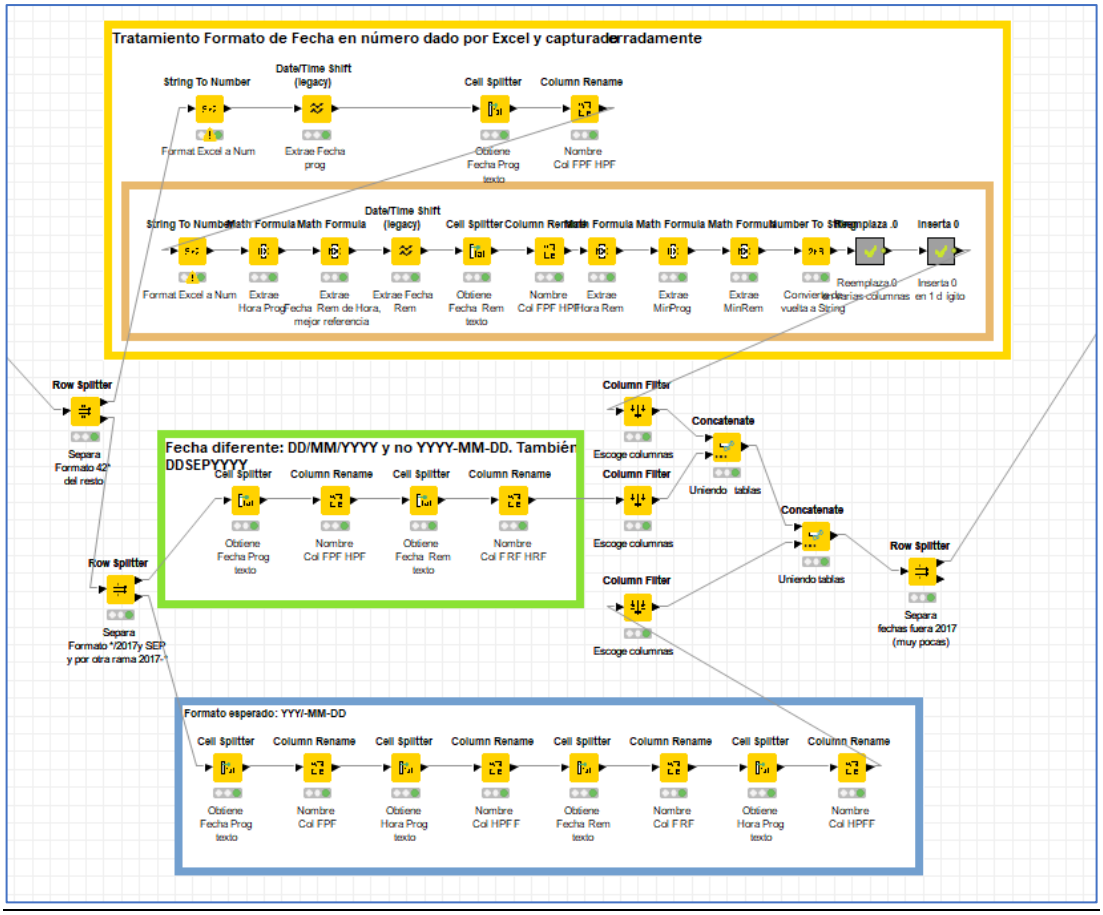


Ilustración 4 *Flujos de trabajo para corrección de formatos Fecha*

Por otra parte, se encontró gran variedad de valores para diferentes variables (por ejemplo, causa de retraso externa tenía valores como Externa, EXTERNO, Externo, externo, etc.). Se procedió a ejecutar una rutina de homogenización de valores para estos casos.

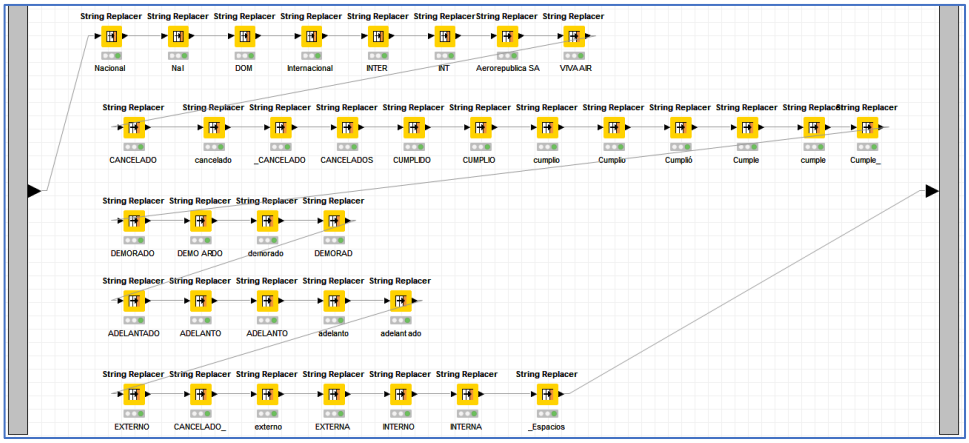


Ilustración 5 *Flujos de trabajo para homogenización de valores de diferentes variables*

11.1.1.3 Algunas transformaciones adicionales

- Se eliminan los registros con estado de Aerocivil "Penalizado" que no están contemplados dentro la regulación RAC 13 (332 casos).
- Se eliminan los registros con estado de Aerocivil vacíos (113 casos).
- Se eliminan los registros con fecha de remolque vacía y estado de Aerocivil diferente a cancelado (1.071 casos).
- Asignación Sigla OACI con la cual se identifica la empresa ante las Autoridades Aeronáuticas.

11.1.1.4 Revisión de Validez

Se encontraron retos de validez en el campo de demora registrado por la Aerocivil. No se identificó distinción de signos negativos en el caso de un "adelanto" (caso en el que la aerolínea empieza a utilizar recursos antes de tiempo programado, lo cual no es necesariamente algo bueno, ya que usa recursos antes de tiempo y en realidad puede ser un método para crear un "slack", o pequeña ventana de tiempo que le permita sobreponerse a imprevistos). En ese caso, se decidió calcular el retraso por cuenta propia derivándolo de los campos de fechas-horas de programación de vuelos y fechas-horas de remolque.

FechaHoraProg	FechaHoraRem	date&time diff
2018-08-17T23:57	2018-08-17T00:10	-1427
2018-09-02T23:57	2018-09-02T00:10	-1427
2018-11-06T23:52	2018-11-06T00:05	-1427
2018-12-04T01:05	2018-12-03T01:18	-1427
2017-06-02T23:55	2017-06-02T00:09	-1426
2017-03-29T23:50	2017-03-29T00:04	-1426
2017-03-18T23:50	2017-03-18T00:04	-1426
2017-05-07T23:50	2017-05-07T00:04	-1426
2017-09-20T23:55	2017-09-20T00:09	-1426
2017-07-25T01:35	2017-07-24T01:49	-1426
2017-08-10T01:35	2017-08-09T01:49	-1426
2017-10-28T01:00	2017-10-27T01:14	-1426
2017-04-11T23:55	2017-04-11T00:09	-1426
2017-06-28T23:55	2017-06-28T00:09	-1426
2017-07-27T00:20	2017-07-26T00:34	-1426
2017-09-05T00:20	2017-09-04T00:34	-1426
2017-10-09T00:20	2017-10-08T00:34	-1426
2017-10-14T00:20	2017-10-13T00:34	-1426
2017-04-26T23:55	2017-04-26T00:09	-1426
2017-06-28T23:55	2017-06-28T00:09	-1426

Ilustración 6 *Instancias de alta diferencia entre fecha de programación y remolque.*

En este proceso, se encontraron unos pocos registros en los que al programarse un vuelo cerca de la media noche, pero darse el remolque en la madrugada, no se corrige la fecha y da la impresión de un adelanto inexplicado. Esos registros deben ser tratados con cuidado, ya que pueden generar la falsa idea de adelantos cuando en realidad representan retrasos.

11.2 Exploración de los datos.

La base de datos tiene 648.384 registros (390 rutas) del reporte de cumplimiento de los itinerarios por parte de las principales aerolíneas con operación en los aeropuertos de Colombia, que indica el estado de los vuelos de los años 2017 y 2018. El estado del vuelo indica si cumplió (61.68%), canceló (9.09%), demoró (26.59%) o adelantó (2.65%) el vuelo programado; los vuelos cumplidos son aquellos que tienen demoras o adelantos inferiores a quince (15) minutos con relación a la hora de salida programada por itinerario.

En esta sección se realiza la exploración de los datos con base en gráficos y datos estadísticos para detección de datos atípicos, concentración de valores, identificación de distribuciones, entre otras. El objetivo es explicar, resumir y detectar patrones de los datos con precisión y confianza.

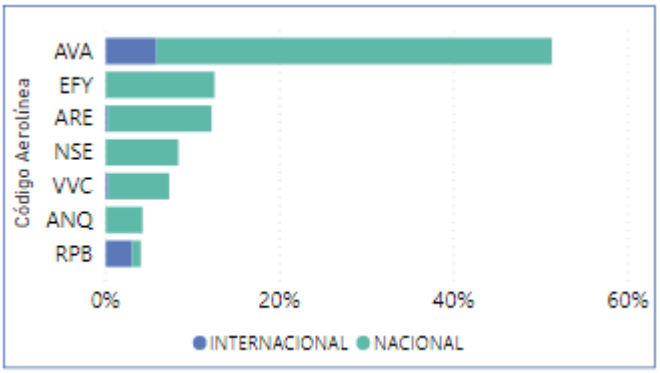


Ilustración 7 *Porcentaje de vuelos por aerolínea.*

La base de datos tiene la información de siete (7) aerolíneas con operación en los aeropuertos de Colombia. En la ilustración 7 se presenta el porcentaje de vuelos programados por aerolínea: Avianca (51.28%), Easyfly (12.52%), Aires (12.19%), Satena (8.37%), Viva Colombia (7.32%), ADA (4.26%) y Aerorepublica (4.06%). De las siete (7) aerolíneas cuatro (4) realizaron rutas internacionales, que corresponden al 9.46% del total de los vuelos programados de los años 2017 y 2018.

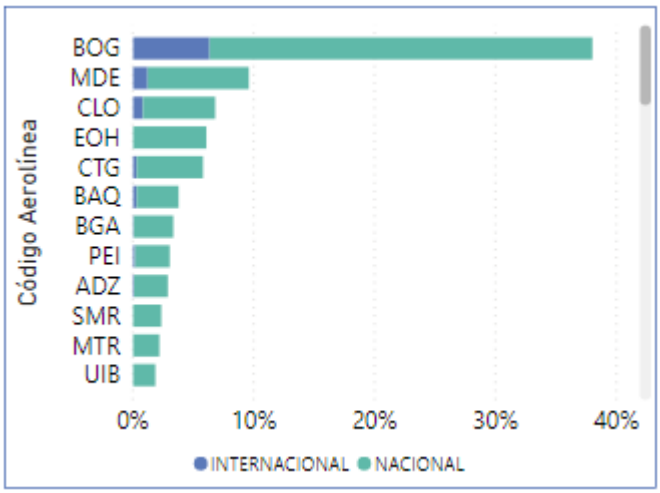


Ilustración 8 *Top ciudades origen con mayor tráfico.*

Para las rutas programadas en los años 2017 y 2018 se identifican 56 aeropuertos de origen (ciudades colombianas) y 113 de destino (tráfico nacional e internacional). De las 56 ciudades de origen, 9 tienen trayectos internacionales: Bogotá (67.14%), Rionegro (12.98%), Cali (8.91%), Cartagena (3.54%), Barranquilla (3.52%), Pereira (2.09%), San Andrés – Isla (1.13%), Bucaramanga (0.68%) y Santa Marta (0.01%).

Bogotá, Rionegro, Cali, Medellín y Cartagena representan el 66.34% de los aeropuertos de origen del itinerario de la Aerocivil Colombiana (Ilustración 8).

La red de tráfico aéreo de los aeropuertos de Colombia de los años 2017 y 2018 se encuentra representada por 113 nodos y 389 conexiones. Los nodos corresponden a los aeropuertos (Sigla IATA) y los arcos son las rutas (origen – destino). El alto nivel de conectividad que se observa permite

establecer que se trata de un *smallworld* (ilustración 9). El color y el tamaño del nodo brinda información sobre el grado de centralidad. Esta métrica sugiere que los aeropuertos con mayor nivel de actividad (origen y destino) son: Bogotá, Rionegro, Medellín, Cali y Lebrija Santander.

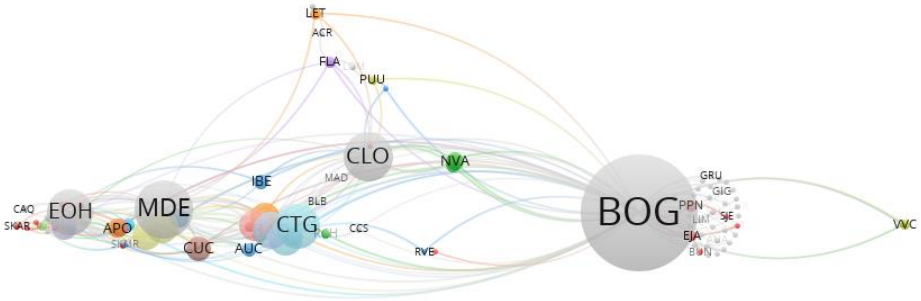


Ilustración 9 *Grado de centralidad de la red*

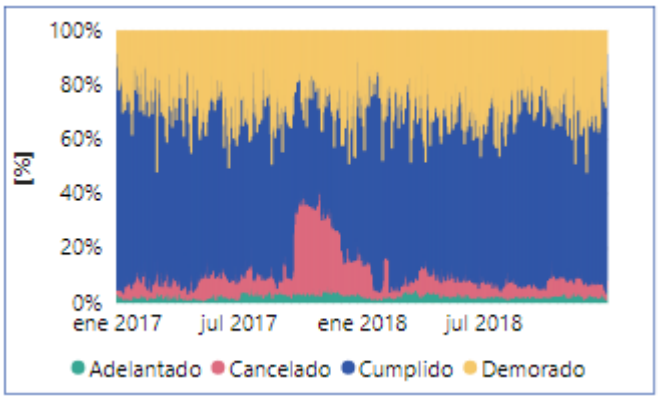


Ilustración 10 *Porcentaje de estado de los vuelos en el tiempo.*

En la ilustración 10 se presenta el estado de los vuelos programados (adelantado, cancelado, cumplido y demorado) en los años 2017 y 2018. Es importante resaltar el incremento de cancelación de vuelos durante las semanas 38 a 52 del 2017, relacionado a la huelga de pilotos de la aerolínea Avianca. Los pilotos solicitaban reducción de la jornada laboral, aumento de días compensatorios, auxilios económicos para desarrollo de actividades de teletrabajo, tiquetes aéreos sin limitaciones para ellos y sus familias, entre otras.

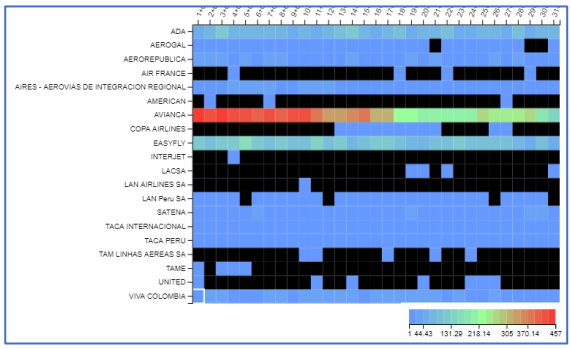


Ilustración 11 *Distribución de vuelos cancelados por Aerolínea y por día del mes.*

En la ilustración 11 se encuentra un análisis de distribución de vuelos cancelados con respecto al día del mes (fuera del período de huelga de Avianca de la semana 38 a la 52 de 2017). En este gráfico se puede observar la particularidad de la concentración de los vuelos en la primera mitad de cada mes con respecto a Avianca que definitivamente tiene el mayor número de vuelos cancelados.

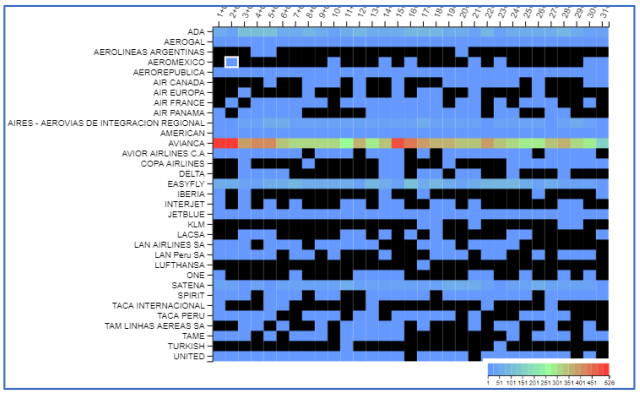


Ilustración 12 *Distribución de vuelos retrasados por Aerolínea y por día del mes.*

De manera similar, en la ilustración 12 se puede ver un patrón de incremento de retrasos mayores a 60 minutos en los 2 primeros días del mes y en los días 15 y 16 de cada mes, de nuevo Avianca concentra la mayor cantidad de retrasos y con un impacto bastante fuerte con respecto a todo el universo de datos.

En la ilustración 13 se presentan los histogramas de las variables númericas que se generan a partir de la fuente de datos *Tráfico por equipo mensual* agrupadas por año, mes, sigla empresa, origen y destino:

$$\boldsymbol{Duración\ del\ vuelo} = \sum(horas\ bloque)/\sum(número\ de\ vuelos)$$
$$\boldsymbol{Carga\ del\ vuelo} = \sum(carga\ a\ bordo\ (Kg))/\sum(número\ de\ vuelos)$$
$$\boldsymbol{Pasajeros\ a\ bordo} = \sum(pasajeros\ \ a\ bordo)/\sum(número\ de\ vuelos)$$
$$\boldsymbol{Sillas} = \sum(sillas)/\sum(número\ de\ vuelos)$$
$$\boldsymbol{Distancia} = mean(distancia)$$

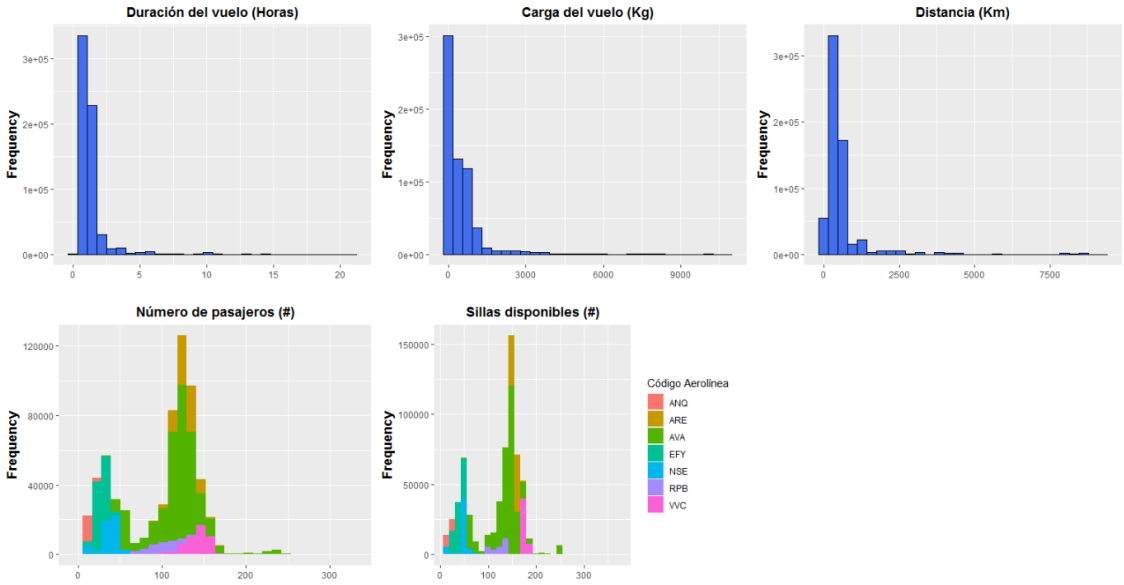


Ilustración 13 *Histogramas fuente de datos tráfico por equipo (mensual)*

La fuente de datos de tráfico por equipo mensual no tiene información de 186 rutas de baja frecuencia que equivale a 18.326 registros en la base de datos principal (cumplimiento de itinerarios (horaria)). En los histogramas de número de pasajeros a bordo y sillas disponibles se identifican tres grupos de aerolíneas: regionales (ADA, Easyfly y Satena), intermedias (Viva Colombia) y de gran escala (Avianca, Aires y Aerorepublica).

A continuación, se realiza el análisis de las variables de temporalidad creadas a partir de la fecha programada del vuelo:

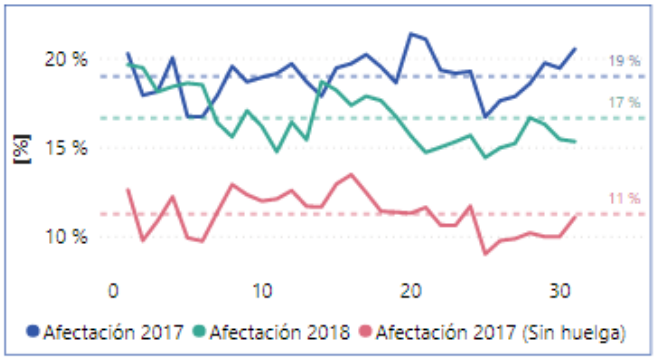


Ilustración 14 *Porcentaje de afectación de vuelos por año y día del mes.*

En la ilustración 15 se presenta el porcentaje de afectación (vuelos demorados y cancelados) en los días del mes. Se identifica que el valor medio de afectación en el año 2017 fue del 19% y sin tener en cuenta el periodo de la huelga de Avianca (de semana 38 a semana 52 del 2017) es del 11% y en el 2018 del 17%.

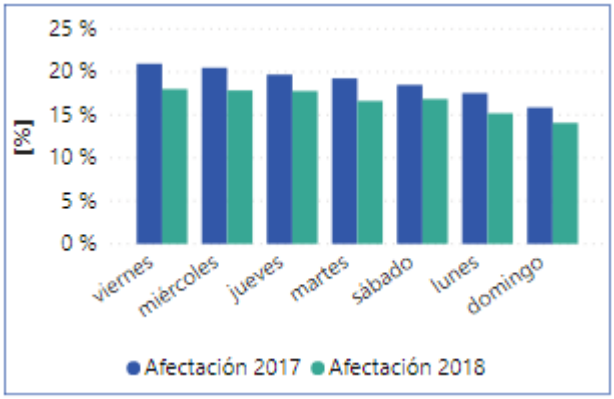


Ilustración 15 *Porcentaje de afectación de vuelos por año y de la semana.*

En la ilustración 15 se presenta el porcentaje de afectación (vuelos demorados y cancelados) en los días de la semana. Se identifica que los miércoles, jueves y viernes es mayor el porcentaje de afectación (entre 18% y 21%) y el domingo es el de menor afectación, entre 14% y 16%.

En cuanto a las variables numéricas asociadas al clima de la ciudad de Bogotá, se observan las siguientes particularidades:

- Al realizar la agrupación por hora del día, se identifican valores más altos de temperatura a partir de las 2:00 de la tarde, coincidiendo con un decremento en los niveles de humedad.
- El viento tiende a aumentar su velocidad alrededor de las 7:00 de la noche, y a tomar un ángulo de incidencia promedio de 180 grados. Se identifica una mayor velocidad del viento para los meses de julio y agosto; el ángulo de incidencia oscila en 150 grados a principio y fin de año, mientras que a mitad de este el registro dominante es de aproximadamente 100 grados.
- Los registros de precipitación (lluvia) muestran la mayor frecuencia entre las 4:00 y las 6:00 am, y en la tarde de las 19:00 a las 23:00 horas. Se identifican mayores niveles de precipitación en febrero, marzo y abril. También se identifican altos niveles de precipitación, pero con baja variabilidad para el mes de agosto.

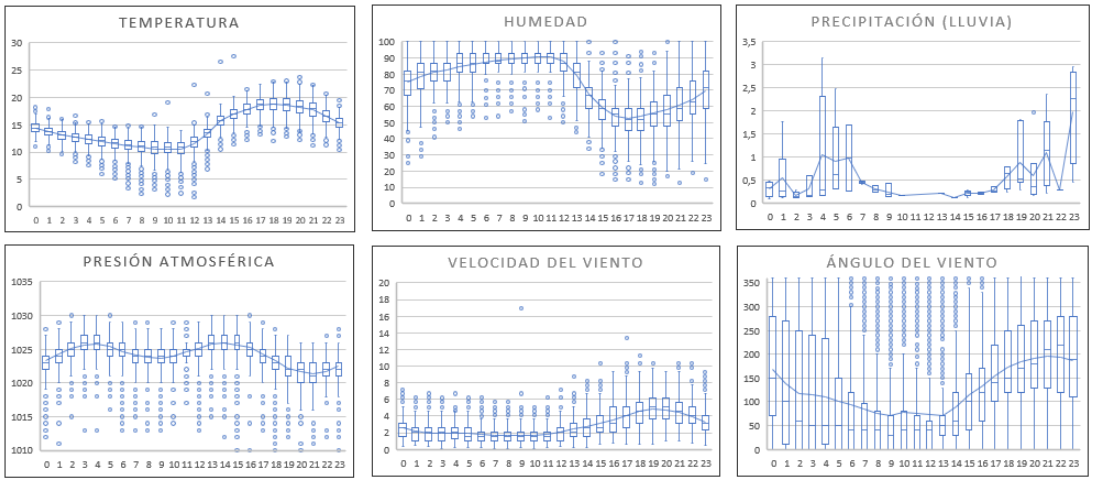


Ilustración 16 *Variables climáticas numéricas principales y su variabilidad por hora del día*

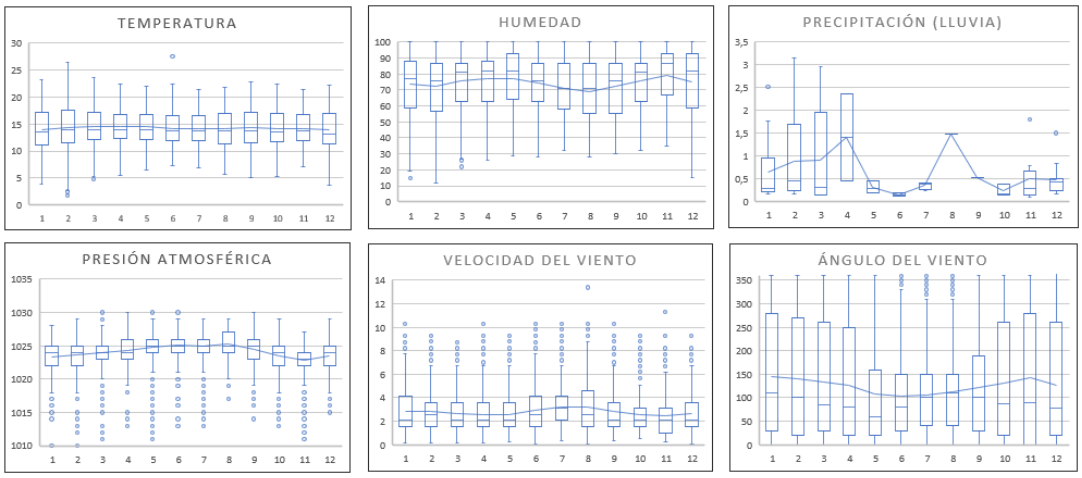


Ilustración 17 *Variables climáticas numéricas principales y su variabilidad por mes*

- La presión atmosférica muestra mayores niveles a las 3:00 am y en el periodo de las 13:00 a las 15:00 horas, y a nivel mensual en el mes de agosto.

Para las variables categóricas climáticas, en la tabla 7 se encuentran las siguientes características:

- Los periodos de nubosidad son mucho más frecuentes que los de cielo despejado. Es más probable encontrar un cielo totalmente despejado en las horas de la mañana; sin embargo, esta ocurrencia no es frecuente. Los meses de principio y fin de año son más propicios para cielos despejados.
- Las horas de la mañana, especialmente 7:00 y 8:00 de la mañana, tienen la mayor concentración de lloviznas, mientras que las horas de tormentas eléctricas se concentran después de las 6:00 pm. Para agregación mensual, las tormentas eléctricas son muy frecuentes en los meses de marzo, abril, octubre y noviembre.
- Los periodos con mayores niveles de lluvia a nivel horario están comprendidos entre las 18:00 y las 22:00 horas, y en el rango de marzo a junio a nivel mensual.

- Después de las 7:00 am se tiene una gran cantidad de eventos de niebla *(fog)*, mientras que la falta de visibilidad se hace más frecuente en las horas de la tarde *(haze).* Estas dos variables tienen mayor impacto en los meses de principio y fin de año.

Hour	Clear	Clouds	Drizzle	Fog	Haze	Mist	Rain	Smoke	Thunder
0	7	1360	89	3	35		113		27
1	23	1371	86	7	27	1	103		16
2	24	1374	101	14	25	3	79	1	12
3	26	1370	81	24	23	6	94		8
4	38	1352	92	34	21	8	83		6
5	49	1320	90	53	15	15	88		3
6	47	1302	86	79	19	17	81		2
7	43	1268	120	100	17	19	67		
8	33	1249	110	149	15	18	59		
9	28	1216	80	212	21	16	62		
10	18	1170	89	274	16	18	51		
11	7	952	81	454	30	46	70		
12	5	1045	71	280	102	53	80		
13	12	1207	48	84	146	42	94	1	
14	4	1301	35	13	170	12	97	1	
15	4	1378	41	1	101	2	105	1	
16	1	1421	53		52	1	103	1	
17		1380	68		47		131	2	4
18	1	1290	75		65		157	2	43
19	1	1242	85		56		172	2	77
20	1	1238	93		50		171	2	86
21		1261	91		50	1	163	2	69
22		1290	81		36		169	1	57
23	1	1314	79		45	1	152		41

Month	Clear	Clouds	Drizzle	Fog	Haze	Mist	Rain	Smoke	Thunder
1	80	2910	51	334	138	54	128	10	16
2	68	2576	98	239	196	37	177	2	19
3	59	2543	204	156	406	60	229	3	72
4	27	2686	259	132	134	22	272	1	75
5	20	2811	297	86	37	7	427		38
6	5	2627	241	73	17	5	364		8
7		2570	160	22	2	3	218		2
8	14	2570	139	43	31	2	171		6
9	26	2424	128	121	37	5	114		30
10	20	2442	128	135	44	16	128		65
11	16	2080	143	234	77	50	198		94
12	38	2432	77	206	65	18	118		26

Tabla 7 *Visualización de impacto por hora del día y por mes para variables climáticas categóricas*

12 Procesamiento de los datos

A continuación, se describe el flujo para la creación y evaluación de los modelos:

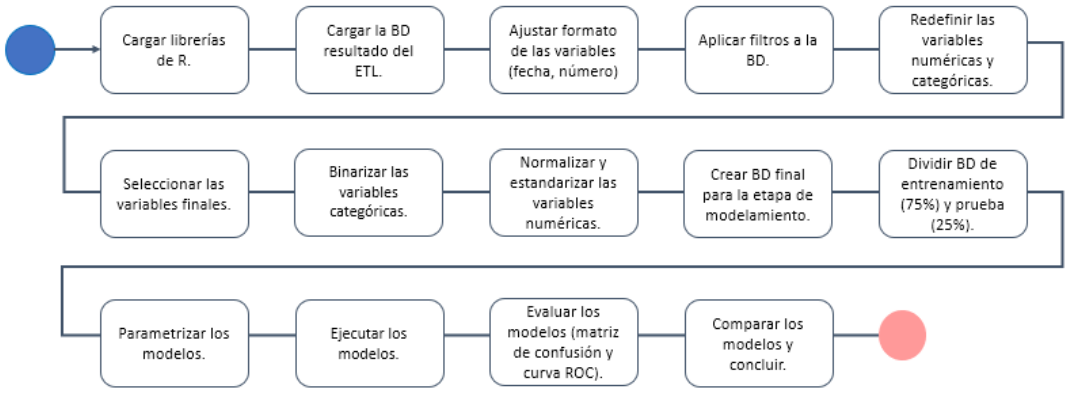


Ilustración 18 *Flujo de la preparación y evaluación de modelos*

12.1.1 Filtros aplicados a la base de datos

Para iniciar con la etapa de modelamiento se eliminan los siguientes registros de la base de datos:

- Registros de semana 38 a 52 del 2017, debido al incremento de cancelaciones de vuelos relacionados a la huelga de los pilotos de Avianca (datos atípicos).
- Registros sin información en la base de datos de *Tráfico aéreo mensual* (rutas de baja frecuencia)
- Registros con adelantos superiores a un día: diferencia entre hora de remolque y hora programada menor a -1.440 minutos.

Igualmente, se filtran los vuelos que despegan o se dirigen al Aeropuerto Internacional El Dorado de la ciudad de Bogotá en los años 2017 y 2018.

12.1.2 Definición de variables numéricas y categóricas

Teniendo en cuenta que la mayoría de las variables del set de datos obtenido son categóricas, se crearon una serie de atributos adicionales con el objetivo que aporten al entendimiento del negocio.

Los nombres de las variables de color verde fueron creados a partir de la información de las bases de datos de reporte de cumplimiento de itinerarios y tráfico aéreo mensual.

#	Atributo	Descripción	Tipo
1	Afectado (variable objetivo)	Variable Objetivo. Describe si el vuelo fue afectado por un retraso mayor a 15 min o por una cancelación.	Categórica
2	Tráfico	Nacional o Internacional	Categórica
3	Aerolínea	Nombre comercial de la empresa.	Categórica
4	Origen	Corresponde a la sigla IATA del aeropuerto donde se origina el trayecto.	Categórica
5	Destino	Corresponde a la sigla IATA del aeropuerto donde termina el trayecto.	Categórica
6	Año	Año en el que está programado el vuelo	Categórica
7	Mes	Mes en el que está programado el vuelo	Categórica
8	Día del mes	Día del mes en el que está programado el vuelo	Categórica
9	Día de la semana	Día de la semana en el que está programado el vuelo.	Categórica
10	Hora (Minutos)	Hora en el que está programado el vuelo. Valor numérico de 0 (00:00:00) a 1439 (23:59:00)	Número
11	Distancia	Distancia en kilómetros entre los aeropuertos del respectivo trayecto.	Número
12	Pasajeros	Cantidad de pasajeros promedio para una ruta definida por mes y aerolínea.	Número
13	Sillas Disponibles	Cantidad de sillas disponibles promedio para una ruta definida por mes y aerolínea.	Número
14	Tiempo Vuelo	Tiempo promedio del vuelo.	Número
15	Días última afectación	Cantidad de días transcurridos desde la última afectación del mismo número de vuelo.	Número
16	Total afectado semana anterior	Cantidad total de veces en las que el vuelo fue afectado durante la semana anterior.	Número
17	Afectado semana anterior	Indica si el número de vuelo de una aerolínea fue afectado la semana anterior.	Número

18	***Afectado día anterior***	*Indica si el número de vuelo de una aerolínea fue afectado el día anterior.*	*Número*
19	***Afectado semana actual***	*Cantidad de veces en las que el vuelo fue afectado durante los siete días previos al día programado de salida.*	*Número*
20	***Cantidad vuelos Hora***	*Cantidad de vuelos programados para despegar y aterrizar en el Aeropuerto El Dorado de Bogotá.*	*Número*
21	***Temperatura***	*Temperatura (°C)*	*Número*
22	***Presión***	*Presión atmosférica (sobre el nivel del mar), hPa.*	*Número*
23	***Humedad***	*Humedad (%)*	*Número*
24	***Velocidad del viento***	*Velocidad del viento. Unidad: metro / segundo*	*Número*
25	***Grado del viento***	*Dirección del viento (Grados)*	*Número*
26	***Nubosidad***	*Nubosidad (%)*	*Número*
27	***Lluvia 1 hora***	*Volumen de lluvia para la última hora (mm)*	*Número*
28	***Clima principal***	*Grupo de parámetros climáticos.*	*Categórica*

Tabla 8 *Variables numéricas y categóricas*

12.1.3 Diagrama de correlación

En el diagrama de correlación se identifican las siguientes características importantes:

- La temperatura tiene un relación inversa con la humedad (r = -0.867).
- Para la nubosidad no se identifican correlaciones significativas con otros atributos.
- Entre las variables días última afectación, total afectado semana anterior, afectado semana anterior, afectado día anterior y afectado semana actual no se identifican correlaciones significativas (< 0.607) que requieren la eliminación de alguna de las variables para evitar multicolinealidad.

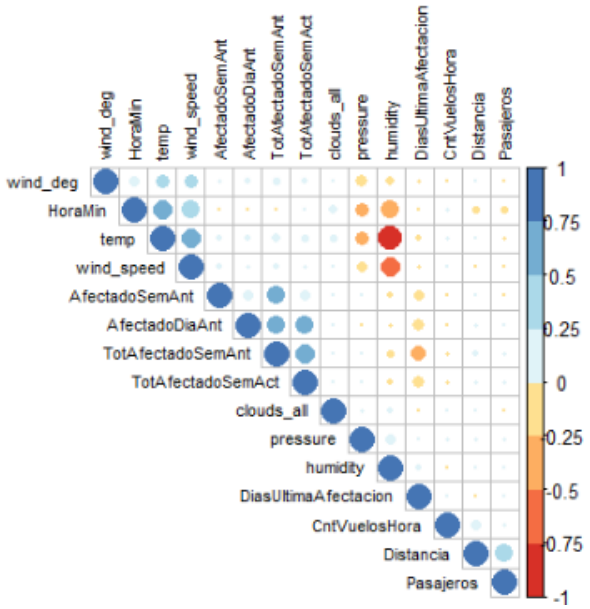


Ilustración 19 *Diagrama de correlación*

12.1.4 Selección final de variables para el proceso de modelamiento

A continuación, se presenta las variables escogidas para la creación de los modelos:




Año	*Código Aerolínea*	*Días última afectación*	*Temperatura*
Mes	*Origen*	*Cuenta vuelos hora*	*Clima Principal*
Día del mes	*Tráfico*	*Total afectado semana anterior*	
Día de la semana	*Distancia*	*Afectado semana anterior*	
Hora (Minutos)	*Pasajeros*	*Afectado día anterior*	
		Afectado semana actual	
		Afectado (Variable Objetivo)	

12.1.5 Variables dummy (One – Hot Encoding)

Una variable categórica es una variable con un número limitado de valores distintos o categorías, que pueden ser nominales u ordinales. Una variable *nominal* consta de dos o más categorías mutuamente excluyentes y en la variable *ordinal* sus categorías están ordenadas por rango y cada clase posee una misma relación posicional con la siguiente.

La estrategia One – Hot Encoding es una estrategia que implementa la creación de columnas para cada valor distinto que exista en las características de la variable categórica que se está codificando y, para cada registro, se marca con un 1 la columna a la que pertenece dicho registro y se deja las demás con 0. La aplicación de esta estrategia se realiza porque los datos categóricos no se pueden representar o no son aplicables en varios algoritmos de aprendizaje automático (regresión logística, redes neuronales, entre otros). Estos deben ser convertidos a valores tipo numérico.

A las siguientes variables se les aplica la estrategia de One – Hot Encoding:

Atributo	***# Categorías***	***Atributo***	***# Categorías***
Año	*2*	*Código aerolínea*	*6*
Mes	*12*	*Origen*	*39*
Día del mes	*31*	*Tráfico*	*2*
Día de la semana	*7*	*Clima Principal*	*9*

Tabla 9 *One-Hot Encoding*

La base de datos pasó de tener 19 variables a 120.

12.1.6 Eliminación de variables para evitar multicolinealidad

Luego de la etapa de One – Hot Encoding se elimina una de las categorías de las variables de la tabla 9 para evitar multicolinealidad. La *multicolinealidad* es la correlación alta entre más de dos variables explicativas en una regresión múltiple que incumple el supuesto de Gauss-Markov cuando la relación es exacta.

Por ejemplo, en código aerolínea se elimina la categoría *RPB (Aerorepublica)* y en clima principal la categoría *Clear (despejado).*

12.1.7 Transformación y normalización de variables numéricas

La transformación de variables se aplica principalmente cuando la distribución de los datos es asimétrica. La simetría es la medida que indica la distribución de una variable respecto a la media aritmética.

Para varios algoritmos de aprendizaje automático es necesario normalizar las variables numéricas de entrada. Por normalización se entiende ajustar los valores medios en diferentes escalas respecto a una escala común. Para el presente análisis se aplica el método de normalización Z Score: se calcula restando al dato, la media de la distribución y dividiendo el resultado por la desviación típica (la distancia que tiene dicho dato respecto de la media).

12.1.8 Selección de técnicas y supuestos

Para la etapa de modelamiento se proponen las siguientes técnicas:

- ***Regresión logística:*** método que permite estimar la probabilidad de una variable cualitativa binaria en función de un conjunto de variables predictoras que pueden ser, tanto continuas como categóricas.
- ***Redes neuronales:*** es un procesador distribuido en paralelo de forma masiva con una propensión natural a almacenar conocimiento experimental y convertirlo en disponible para su uso.
- ***XGBoosting:*** es una variante o aplicación específica del concepto de **"**Gradient boosting", que a su vez es una técnica empleada para problemas de regresión y clasificación, cuyo resultado es un modelo del ensamblaje de modelos (árboles de decisión) más débiles. El aporte específico de XGBoosting (acrónimo derivado de la expresión eXtreme Gradient Boosting) parte de la aplicación de métodos adicionales de regularización para lograr resultados de calidad y de forma muy rápida.

12.1.9 Conjunto de datos de entrenamiento y prueba

La base de datos se divide en dos conjuntos de datos:

- **Entrenamiento:** subconjunto para entrenar el modelo (75% de los datos).
- **Prueba:** subconjunto para probar el modelo entrenado (25% de los datos).

Los subconjuntos de datos son suficientemente grandes para generar resultados significativos desde el punto de vista estadístico.

13 Modelamiento

Los modelos por implementar se evaluarán con métricas como *accuracy, precisión, sensitividad y especificidad*. El *accuracy* indica el porcentaje que el modelo logra predecir correctamente, tanto los vuelos afectados (TP) como los que cumplen con el itinerario programado (TN); la *precisión* indica el porcentaje de vuelos que se logran predecir con afectación (TP) del total de predicciones realizadas; mientras que la *sensitividad* es el porcentaje de vuelos que se predicen con afectación (TP), teniendo en cuenta los valores de referencia de la base de datos de prueba que estaban marcados con afectación. Es decir, la *precisión* tiene en cuenta los falsos positivos (FP) y la *sensitividad* los falsos negativos (FN). Finalmente, la *especificidad* expresa el porcentaje de acierto en los vuelos que cumplen con el itinerario (TN), con relación al total de vuelo no afectados en la base de datos de

referencia. Para este caso se le dará mayor importancia a los modelos que logren mayor *sensitividad,* ya que es más importante identificar aquellos vuelos que posiblemente se van a ver afectado que aquellos que cumplirán con el itinerario.

La curva ROC y el área bajo la curva (AUC) también permiten establecer la efectividad de los modelos. La curva ROC es de gran utilidad dado que compara la tasa positiva verdadera (TPR) frente a la tasa positiva falsa (FPR) y el AUC mide toda el área bidimensional por debajo de la curva ROC, proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles. Un modelo que clasifica perfectamente las dos clases tiene un 100% de sensibilidad y especificidad, por lo que el área bajo la curva es 1 y un modelo que predice por debajo de lo esperado por azar, tiene AUC menor de 0.5.

Es importante recordar que los vuelos cumplidos son aquellos que tienen demoras o adelantos inferiores a quince (15) minutos con relación a la hora de salida programada por itinerario. Para la creación de los modelos se analizan vuelos con retrasos mayores a 15, 60 y 90 minutos. Por la importancia que tienen en la operación y las multas para las aerolíneas, se profundizará en el análisis y predicción de vuelos con retrasos superiores a 60 minutos.

A continuación, se presenta la cantidad de datos de entrenamiento y prueba para los tiempos de retrasos mencionados en el párrafo anterior:

Retraso (minutos)	*Base de Entrenamiento*	*Base de Prueba*
15	*284.608*	*94.856*
60	*228.401*	*76.120*
90	*218.553*	*72.835*

Tabla 10 *Base de entrenamiento y pruebas por tiempo de retraso*

13.1 Regresión logística

Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en donde las observaciones se clasifican en un grupo u otro dependiendo del valor que tomen las variables empleadas como predictoras. Es importante aclarar que, aunque la regresión logística permite clasificar, dicha clasificación se obtiene al modelar el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas.

Para la interpretación de los Odd Ratio se tendrá en cuenta que estos van de cero (0) a infinito. Cuando es igual a uno (1) no hay asociación entre las variables, los valores menores a uno (1) indican una asociación inversa entre las variables y que los que sean mayores a uno (1) señalan una relación positiva entre las mismas. Cuando los valores sean menores que uno (1), la interpretación se realizará calculando la inversa del Odd Ratio.

Por medio de la regresión logística, se crearon y evaluaron tres modelos: **1.** Sin balanceo, **2.** Con sobre muestreo, **3.** Step sin balanceo, para los vuelos afectados con retrasos de 15, 60 y 90 minutos.

13.1.1 Modelo

Teniendo en cuenta la significancia de las variables predictoras en una primera iteración se descartaron un total de 32 variables y el modelo final se ejecuta con los siguientes atributos:

Var	Estimate	Std. Error	z value	Pr(>|z|)	Var	Estimate	Std. Error	z value	Pr(>|z|)
Anio.2018	0.064	0.013	4.844	1.27e-06 ***	VVC	0.664	0.066	10.106	< 2e-16 ***
Mes.10	0.182	0.035	5.189	2.11e-07 ***	Origen.AUC	-0.514	0.117	-4.378	1.20e-05 ***
Mes.11	0.485	0.034	14.189	< 2e-16 ***	Origen.BAQ	-0.219	0.060	-3.631	0.000283 ***
Mes.12	0.140	0.036	3.896	9.80e-05 ***	Origen.BGA	-0.295	0.064	-4.623	3.78e-06 ***
Mes.2	0.078	0.031	2.521	0.011698 *	Origen.BOG	-0.314	0.058	-5.408	6.37e-08 ***
Mes.3	0.296	0.029	10.222	< 2e-16 ***	Origen.CLO	-0.261	0.060	-4.381	1.18e-05 ***
Mes.4	0.295	0.029	10.110	< 2e-16 ***	Origen.CUC	-0.455	0.072	-6.321	2.61e-10 ***
Mes.5	0.340	0.029	11.910	< 2e-16 ***	Origen.EJA	-0.210	0.099	-2.116	0.034311 *
Mes.6	0.500	0.029	17.527	< 2e-16 ***	Origen.EOH	-0.822	0.093	-8.858	< 2e-16 ***
Mes.7	0.402	0.029	14.046	< 2e-16 ***	Origen.EYP	-0.532	0.075	-7.068	1.57e-12 ***
Mes.8	0.220	0.029	7.500	6.36e-14 ***	Origen.IBE	0.187	0.084	2.235	0.025415 *
Mes.9	0.195	0.032	6.143	8.08e-10 ***	Origen.MDE	-0.301	0.059	-5.092	3.54e-07 ***
Dia.1	0.711	0.046	15.614	< 2e-16 ***	Origen.MTR	-0.572	0.083	-6.915	4.68e-12 ***
Dia.11	0.095	0.047	2.007	0.044781 *	Origen.MZL	0.222	0.071	3.149	0.001638 **
Dia.12	0.237	0.046	5.110	3.22e-07 ***	Origen.PEI	-0.294	0.064	-4.602	4.19e-06 ***
Dia.14	0.373	0.046	8.152	3.58e-16 ***	Origen.PUU	-0.336	0.141	-2.379	0.017378 *
Dia.15	0.362	0.046	7.902	2.74e-15 ***	Origen.RCH	-0.516	0.131	-3.952	7.74e-05 ***
Dia.16	0.321	0.046	6.973	3.10e-12 ***	Origen.RVE	0.558	0.247	2.264	0.023589 *
Dia.17	0.357	0.046	7.777	7.44e-15 ***	Origen.SJE	-0.754	0.267	-2.828	0.004688 **
Dia.18	0.214	0.048	4.497	6.88e-06 ***	Origen.SMR	-0.224	0.062	-3.585	0.000337 ***
Dia.19	0.202	0.047	4.305	1.67e-05 ***	Origen.TCO	-0.422	0.153	-2.767	0.005659 **
Dia.2	0.335	0.047	7.179	7.01e-13 ***	Origen.VVC	-0.970	0.134	-7.240	4.47e-13 ***
Dia.20	0.259	0.047	5.526	3.28e-08 ***	NACIONAL	-0.221	0.037	-5.988	2.13e-09 ***
Dia.28	0.099	0.048	2.063	0.039138 *	Drizzle	0.562	0.088	6.346	2.22e-10 ***
Dia.29	0.162	0.049	3.274	0.001059 **	Fog	0.328	0.090	3.626	0.000288 ***
Dia.3	0.186	0.047	3.970	7.19e-05 ***	Haze	0.228	0.098	2.328	0.019936 *
Dia.30	0.128	0.049	2.586	0.009722 **	Mist	0.328	0.121	2.703	0.006875 **
Dia.31	0.178	0.055	3.238	0.001206 **	Rain	0.529	0.095	5.583	2.37e-08 ***
Dia.4	0.385	0.046	8.417	< 2e-16 ***	Thunderstorm	0.962	0.096	10.050	< 2e-16 ***
Dia.5	0.207	0.046	4.489	7.16e-06 ***	HoraMin	-0.045	0.007	-6.046	1.49e-09 ***
DiaSemana.1	-0.098	0.023	-4.158	3.21e-05 ***	Distancia	-0.230	0.012	-18.644	< 2e-16 ***
DiaSemana.2	-0.088	0.023	-3.853	0.000117 ***	Pasajeros	0.037	0.012	2.984	0.002842 **
DiaSemana.3	0.103	0.022	4.699	2.61e-06 ***	temp	-0.033	0.008	-3.984	6.77e-05 ***
DiaSemana.5	0.093	0.022	4.240	2.24e-05 ***	DiasUltimaAfectacion	0.035	0.007	5.065	4.09e-07 ***
DiaSemana.6	-0.081	0.023	-3.548	0.000388 ***	CntVuelosHora	-0.038	0.014	-2.774	0.005545 **
DiaSemana.7	-0.279	0.025	-11.198	< 2e-16 ***	TotAfectadoSemAnt	0.508	0.010	51.845	< 2e-16 ***
AVA	0.630	0.061	10.355	< 2e-16 ***	AfectadoSemAnt	0.069	0.007	10.249	< 2e-16 ***
EFY	1.022	0.066	15.572	< 2e-16 ***	TotAfectadoSemAct	0.092	0.009	9.938	< 2e-16 ***
NSE	1.074	0.069	15.625	< 2e-16 ***	AfectadoDiaAnt	0.106	0.007	15.582	< 2e-16 ***

Tabla 11 *Variables seleccionadas por significancia*

El logaritmo de odds de variables ciudades de origen como Villavicencio (VVC), Medellín (EOH), San Jose del Guaviare (SJE), Montería (MTR), Yopal (EYP), Riohacha (RCH), Arauca (AUC), Cúcuta (CUC), Tumaco (TCO), Puerto Asis (PUU), Bogotá (BOG), Rionegro (MDE), Bucaramanga (BGA), Pereira (PEI), Cali (CLO), Santa Marta (SMR), Barranquilla (BAQ) y Barrancabermeja (EJA); días de la semana lunes (1), martes (2), sábado (6) y domingo (7); distancia, hora a la que está programado el vuelo (HoraMin) y cantidad de vuelos por hora (CntVuelosHora) están negativamente relacionados con la afectación de los vuelos. La mayor parte de estas corresponde a variables que tienen que ver con el origen del vuelo, en donde se destacan los vuelos que parten del aeropuerto La Vanguardia de Villavicencio (-0.97) y del Aeropuerto Olaya Herrera de Medellín (-0.82).

El resto de las variables del conjunto de datos tienen una relación positiva con los vuelos afectados y las aerolíneas como Easy Fly (EFY) y Satena (NSE) tiene una alta importancia en el modelo. En cuanto a los factores climáticos, la tormenta eléctrica (Thunderstorm) es la de mayor peso con un coeficiente de (0.96).

El modelo con mejor balance entre sensitividad y especificidad se obtuvo en el punto de corte 0.16. Con este modelo se logró predecir: 42.256 vuelos no afectados, 8.627 vuelos afectado, 4.128 falsos negativos y 21.109 falsos positivos.

		Referencia	
		No Afectado	*Afectado*
Predicción	*No Afectado*	*42256*	*4128*
	Afectado	*21109*	*8627*

Tabla 12 *Matriz de confusión modelo Regresión logística (60 minutos)*

En la Tabla 13 se presenta las métricas obtenidas con el modelo Logit con los datos sin balancear a 60 minutos. El accuracy de 66.85% indicaría una predicción aceptable para la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

Punto de Corte	*Accuracy*	*Sensitividad*	*Especificidad*
0.16	*0.6685*	*0.6764*	*0.6669*

Tabla 13 *Métricas modelo Regresión Logística (60 minutos)*

En la Tabla 14 se resumen los odd ratios y las probabilidades de cada variable con respecto a la afectación de vuelos.

Var	*Odd*	*Prob*	*Var*	*Odd*	*Prob*	*Var*	*Odd*	*Prob*
(Intercept)	0.08	0.016	*Origen.PSO*	0.97	0.16	*Dia.19*	1.22	0.2
Origen.VVC	0.38	0.07	*temp*	0.97	0.16	*Dia.5*	1.23	0.2
Origen.EOH	0.44	0.08	*Dia.26*	0.97	0.16	*Dia.18*	1.24	0.2
Origen.SJE	0.47	0.09	*ARE*	0.99	0.17	*Mes.8*	1.25	0.2
Origen.MTR	0.56	0.1	*Origen.UIB*	1	0.17	*Origen.MZL*	1.25	0.2
Origen.EYP	0.59	0.11	*Dia.22*	1.02	0.17	*Haze*	1.26	0.2

Origen.RCH	0.6	0.11	*DiasUltimaAfectacion*	1.04	0.17	*Dia.12*	1.27	0.2
Origen.AUC	0.6	0.11	*Pasajeros*	1.04	0.17	*Origen.APO*	1.29	0.21
Origen.CUC	0.63	0.11	*Anio.2018*	1.07	0.18	*Dia.20*	1.3	0.21
Origen.TCO	0.66	0.12	*Dia.6*	1.07	0.18	*Origen.SVI*	1.32	0.21
Origen.BUN	0.67	0.12	*AfectadoSemAnt*	1.07	0.18	*Mes.4*	1.34	0.21
Origen.PCR	0.69	0.12	*Origen.IPI*	1.07	0.18	*Mes.3*	1.34	0.21
Origen.PUU	0.71	0.13	*Dia.9*	1.08	0.18	*Origen.MVP*	1.38	0.22
Origen.BOG	0.73	0.13	*Origen.FLA*	1.08	0.18	*Dia.16*	1.38	0.22
Origen.MDE	0.74	0.13	*Dia.24*	1.08	0.18	*Fog*	1.39	0.22
Origen.BGA	0.74	0.13	*Mes.2*	1.08	0.18	*Mist*	1.39	0.22
Origen.PEI	0.74	0.13	*Dia.13*	1.09	0.18	*Dia.2*	1.4	0.22
DiaSemana.7	0.76	0.13	*Dia.10*	1.09	0.18	*Mes.5*	1.41	0.22
Origen.CLO	0.77	0.13	*Dia.8*	1.09	0.18	*Dia.17*	1.43	0.22
Distancia	0.79	0.14	*Dia.21*	1.09	0.18	*Dia.15*	1.44	0.22
Origen.SMR	0.8	0.14	*Origen.AXM*	1.1	0.18	*Dia.14*	1.45	0.23
NACIONAL	0.8	0.14	*TotAfectadoSemAct*	1.1	0.18	*Dia.4*	1.47	0.23
Origen.BAQ	0.8	0.14	*DiaSemana.5*	1.1	0.18	*Mes.7*	1.49	0.23
Origen.EJA	0.81	0.14	*Dia.11*	1.1	0.18	*Origen.TME*	1.54	0.24
Origen.VUP	0.86	0.15	*Dia.28*	1.1	0.18	*Mes.11*	1.62	0.25
Origen.NVA	0.89	0.15	*DiaSemana.3*	1.11	0.18	*Mes.6*	1.65	0.25
DiaSemana.1	0.91	0.15	*AfectadoDiaAnt*	1.11	0.18	*TotAfectadoSemAnt*	1.66	0.25
Dia.27	0.91	0.15	*Dia.30*	1.14	0.19	*Rain*	1.7	0.25
DiaSemana.2	0.92	0.15	*Origen.LET*	1.14	0.19	*Origen.RVE*	1.75	0.26
DiaSemana.6	0.92	0.16	*Mes.12*	1.15	0.19	*Drizzle*	1.75	0.26
Dia.25	0.92	0.16	*Dia.29*	1.18	0.19	*AVA*	1.88	0.27
Origen.CTG	0.93	0.16	*Clouds*	1.18	0.19	*VVC*	1.94	0.28
Origen.CZU	0.93	0.16	*Dia.31*	1.2	0.19	*Dia.1*	2.04	0.29
Origen.PPN	0.96	0.16	*Mes.10*	1.2	0.19	*Smoke*	2.11	0.3
HoraMin	0.96	0.16	*Dia.3*	1.2	0.19	*Thunderstorm*	2.62	0.34
CntVuelosHora	0.96	0.16	*Origen.IBE*	1.21	0.19	*EFY*	2.78	0.36
Dia.23	0.96	0.16	*Mes.9*	1.21	0.2	*NSE*	2.93	0.37

Tabla 14 *Resumen Odd ratio y Probabilidades del Modelo sin Balanceo*

De la tabla 14 se analizan los siguientes puntos importantes: los vuelos con origen Quibdó no tienen asociación con la afectación. Por cada vuelo se haya registrado con afectación la semana anterior (TotAfectadoSemAnt) y manteniendo el resto de las variables constantes, los odds de afectación aumentan en un 66%. Mientras que por cada vuelo programado en una hora (CntVuelosHora), reduce a

96% el odd base de que el vuelo se vea afectado. El aumento en una unidad de cualquiera de las variables que tienen aumento de la probabilidad de incumplimiento entre el 7% y el 37%.

En la ilustración 20 se muestra el comportamiento entre sensitividad y especificidad para cada una de las predicciones:

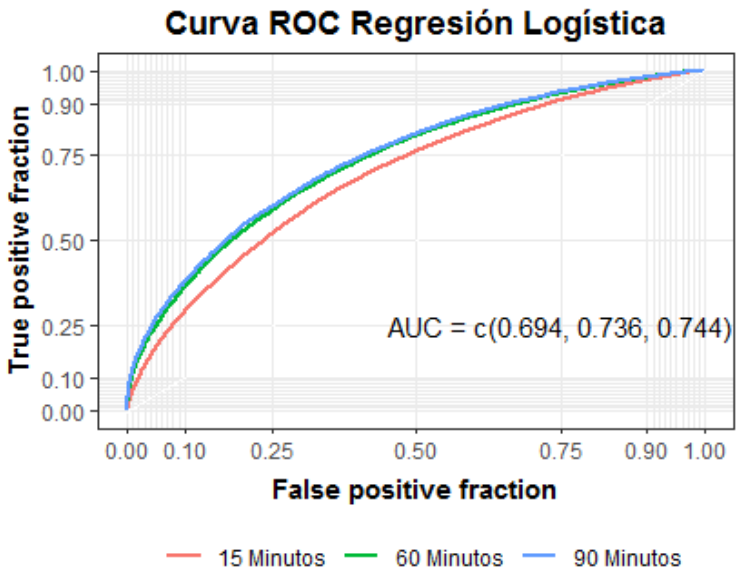


Ilustración 20 *Curva Roc modelo Regresión Logística a 15, 60 y 90 minutos*

13.2 Redes Neuronales

13.2.1 Introducción red neuronal feedforward

Una red neuronal es un procesador distribuido en paralelo de forma masiva con una propensión natural a almacenar conocimiento experimental y convertirlo en disponible para su uso. Se asemeja al cerebro en dos aspectos (1):

- El conocimiento se adquiere por la red mediante un proceso de aprendizaje.
- Las fuerzas de conexión interneuronal, conocidas como ponderaciones sinápticas, se utilizan para almacenar el conocimiento.

Las redes neuronales son un método ideal en muchas aplicaciones de minería de datos predictiva por su potencia, flexibilidad y facilidad de uso. Para el presente análisis se utiliza la arquitectura **feedforward,** donde las conexiones de la red fluyen unidimensionalmente desde la capa de entrada hasta la capa de salida sin ciclos de retroalimentación (ilustración 21).

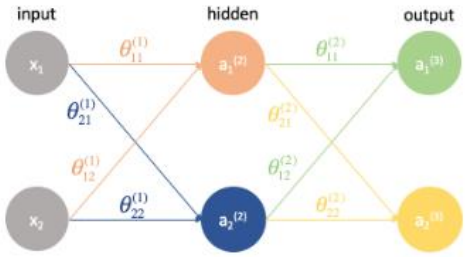


Ilustración 21 *Red Neuronal feedforward*

- La **capa de entrada**, también denominada sensorial, está compuesta por neuronas que reciben datos o señales precedentes del entorno (predictores).
- La **capa oculta** contiene nodos no observables. El valor de cada unidad oculta es una función de los predictores.
- La **capa de salida** se compone de neuronas que proporcionan la respuesta de la red neuronal (2).

13.2.2 Librería NNET (R Studio)

La librería NNET es un software para redes neuronales feedforward con una sola capa oculta y modelos multinomiales logarítmicos lineales. Brian Ripley y William Venables son los autores de la librería NNET; Brian Ripley es un estadístico británico, fue profesor de estadística aplicada en la Universidad de Oxford y profesor titular en el St Peter's Collge.

13.2.3 Modelo

13.2.3.1 Selección de neuronas en la capa oculta

Con la técnica de redes neuronales feedforward se tiene como objetivo clasificar si un vuelo que despega o se dirige al Aeropuerto Internacional El Dorado de la ciudad de Bogotá se afecta (se cancela o se retrasa).

Para el diseño de redes neuronales se recomienda el uso de una sola capa oculta y en algunos casos específicos dos. El aumentar el número de capas ocultas y neuronas hace más lento el entrenamiento, puede aumentar drásticamente el número de mínimos locales y causar sobre-entrenamiento.

Para definir el número de neuronas en la capa oculta se utiliza un método iterativo desde un perceptrón (1 neurona en la capa oculta) hasta 8 neuronas. La selección del mejor modelo se realiza mediante las métricas de *accuracy*, *sensitividad* y *especificidad.* A continuación, se presentan los resultados de cada uno de los 8 modelos creados con la librería *NNET* con 100 iteraciones y decay de 0.0005:

round	*accuracy*	*Accuracy_LL*	*Accuracy_UL*	*sensitivity*	*specificity*	*precision*	*npv*
1	0.6149	0.6114	0.6184	0.7653	0.5846	0.2705	0.9252
2	0.6689	0.6655	0.6722	0.6775	0.6672	0.2907	0.9113
3	0.6893	0.6860	0.6926	0.6506	0.6971	0.3018	0.9083
4	0.6812	0.6779	0.6845	0.7011	0.6772	0.3042	0.9184
5	0.6883	0.6850	0.6916	0.6960	0.6868	0.3091	0.9182
6	0.6727	0.6693	0.6760	0.7039	0.6664	0.2981	0.9179
7	0.6734	0.6701	0.6768	0.6848	0.6712	0.2954	0.9136
8	0.6791	0.6758	0.6825	0.6725	0.6805	0.2976	0.9117

Tabla 15 *Selección de neuronas en la capa oculta*

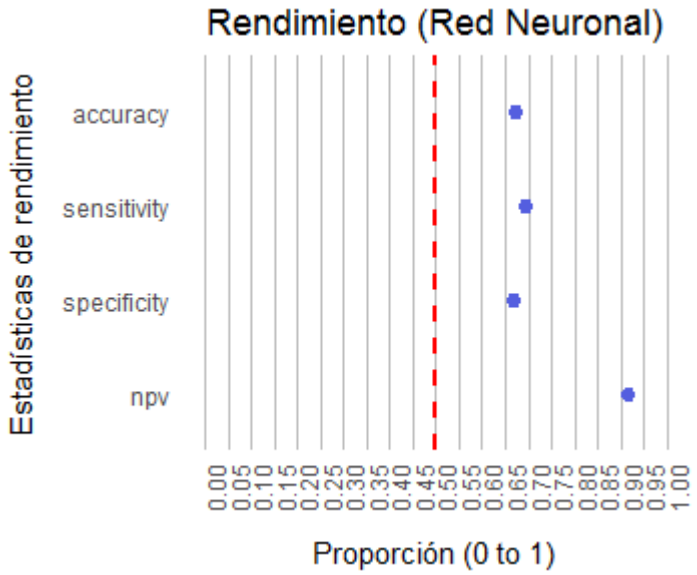


Ilustración 22 *Rendimiento Red Neuronal*

Con cinco (5) neuronas en la capa oculta se logra el mejor resultado en accuracy, sensitividad y especificidad. A partir de 3 neuronas los resultados son comparables y las diferencias en las métricas evaluadas varían entre 1 y 2 puntos porcentuales. En la ilustración 22 se presenta el rendimiento promedio de la red neuronal en las 8 iteraciones.

Para la creación de los modelos finales se tendrá la siguiente arquitectura feedforward:

- **Capa de entrada:** 110 neuronas.
- **Capa oculta:** 5 neuronas.
- **Capa de salida:** 1 neurona.

Los parámetros adicionales para la creación del modelo son: $decay = 0.5, rang = 0.1$ y $maxit = 2.000$. El valor de los parámetros mencionados se ajustó con la iteración y creación de modelos con condición de retraso de 60 minutos.

El modelo con mejor balance entre sensitividad y especificidad se obtuvo en el punto de corte 0.16. Con este modelo se logró predecir: 43.344 vuelos no afectados, 8.767 vuelos afectados, 3.988 falsos negativos y 20.021 falsos positivos.

		Referencia	
		No Afectado	*Afectado*
Predicción	*No Afectado*	*43.344*	*3.988*
	Afectado	*20.021*	*8.767*

Tabla 16 *Matriz de confusión modelo red neuronal (60 minutos)*

En la Tabla 17 se presentan las métricas obtenidas del modelo neuronal con los datos sin balancear a 60 minutos.

Punto de Corte	*Accuracy*	*Sensitividad*	*Especificidad*
0.16	*0.6840*	*0.6873*	*0.6846*

Tabla 17 *Métricas modelo red neuronal (60 minutos)*

En la ilustración 23 se presenta la curva ROC y el AUC para los tres intervalos de tiempo (15, 60 y 90 minutos). Se identifica que el área bajo la curva aumenta al incrementar la condición de retraso, logrando un AUC de 0.771 en 90 minutos.

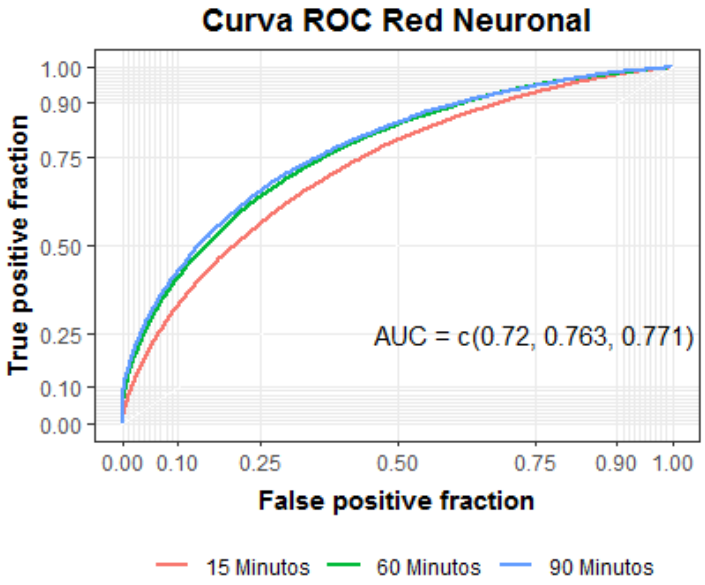


Ilustración 23 *Curva ROC modelo red neuronal 15, 60 y 90 minutos*

13.3 XGBoosting

13.3.1 Concepto de Gradient Boosting

Leo Brieman en sus trabajos de investigación orientados a mejorar métodos de clasificación, observó la ventaja de redefinir los pesos para las variables consideradas en los procesos de entrenamiento de modelos considerando iteraciones previas. Ese primer intento llevó a lo que él llamó "arcing" (o dar forma de arco) y fue tomado por Freund and Schapire generando la primera versión del algoritmo conocido hoy como Adaboost, el cual de manera iterativa redefine los pesos de una base de datos de entrenamiento considerando la historia de los errores de intentos previos, construye un nuevo clasificador y usa el error de clasificación dar un nuevo peso a las muestras clasificadas erróneamente. Como consecuencia, se obtiene un clasificador robusto a partir de la combinación de varios clasificadores débiles.

Posteriormente, Friedman* por propia cuenta y Mason, Baxter, Bartlett and Frean en una iniciativa independiente desarrollaron investigaciones para proponer algoritmos basados en el análisis de los errores residuales (diferencia entre valor de la predicción y valor real) y que de manera iterativa y similar a un algoritmo de optimización, ejecutaban procesos considerando un gradiente descendente, de tal manera que cada iteración llevaba a la reducción de los residuales a un mínimo. De ahí la denominación de "Gradient Boosting" a estas técnicas de clasificación basadas en la optimización de los residuales.

13.3.2 Variante eXtreme Gradient Boosting

Más recientemente, en 2014 Tien propone una mejora a la metodología de Gradient Boosting considerando:

- La modificación de un árbol inicial apuntando a reducir el valor del error residual en cada iteración.
- El uso de un algoritmo "Greedy Aproximado" (un algoritmo "greedy" riguroso implicaría un gran consumo de poder computacional).

- La posibilidad de tener el proceso ejecutado por varios equipos al mismo tiempo (parallel learning).
- *Weighted Quantile Sketch*: generación de "histogramas aproximados" para ubicar las observaciones con bajo nivel de confianza en cuantiles específicos para ello, dejando el resto de las muestras en los cuantiles con la mayor cantidad de observaciones para hacer más eficiente la evaluación de los registros significativos.
- Sparsity-Aware Split Finding: técnicas para procesar adecuadamente "missing values" o datos faltantes en los registros en cuestión.
- *Cache-Aware Access:* XGBoost restringe los cálculos de los *gradientes* (primera derivada) y de los *hessians* (segunda derivada), conceptos clave para los cálculos generales, para la memoria caché del sistema que se esté utilizando.
- Blocks for Out-*of-Core Computation:* manejo particular de bases de datos de gran tamaño, permitiendo comprimirlas en el caso de tener que hacer uso de los recursos en disco duro. Esto partiendo de la base de que es mejor emplear un tiempo descomprimiendo la data de entrenamiento ubicada en el disco duro que esperar a la lectura de toda la base.

13.3.3 Librería xgboost (CRAN)

En el momento de escribir este documento se encuentra disponible la versión 1.1.1.1 de la librería "xgboost" con fecha de generación 6 de junio de 2020. El paquete incluye un modelo lineal "solver" y algoritmos de árboles de decisión que siguen las estrategias ya comentadas para XGBoosting. Se menciona en la documentación de la librería que es "una implementación eficiente del trabajo de Chen & Guestrin" del enfoque de Gradient Boosting y resaltan su rapidez y eficiencia al comparar su desempeño con otras librerías que siguen el mismo enfoque. El paquete soporta diferentes funciones objetivo, incluyendo entre ellas regresión, clasificación y ranking.

13.3.4 Modelo

Para la implementación de este tipo de modelo se dispuso de un código que permitió revisar diferentes valores de la rata de aprendizaje y así poder lograr el mejor resultado.

La librería xgboost ofrece varias opciones de ajuste de manera iterativa. También permite observar según el resultado del modelo las variables de mayor importancia y que ofrecen una guía de los verdaderos factores de peso en este análisis.

De igual manera, como se procedió con los modelos de regresión logística y de red neuronal, se ejecutaron ejercicios de modelación para condiciones de afectación de 15, 60 y 90 minutos.

El modelo con mejor balance entre sensitividad y especificidad se obtuvo en el punto de corte 0,22. Con este modelo se logró predecir: 42.501 vuelos no afectados, 8.846 vuelos afectado, 3.909 falsos negativos y 20.864 falsos positivos.

		Referencia	
		No Afectado	*Afectado*
Predicción	*No Afectado*	*42.501*	*3.909*
	Afectado	*20.864*	*8.846*

Tabla 18 *Matriz de confusión modelo XGBoosting (60 minutos)*

En la Tabla 19 se presentan las métricas obtenidas con el modelo XGBoosting que contiene los datos sin balancear a 60 minutos.

Punto de Corte	*Accuracy*	*Sensitividad*	*Especificidad*
0.22	*0.6746*	*0.6935*	*0.6707*

Tabla 19 *Métricas modelo XGBoosting (60 minutos)*

En la ilustración 24 se presenta la curva ROC y el AUC para los tres intervalos de tiempo (15, 60 y 90 minutos). Se identifica que el área bajo la curva aumenta al clasificar los vuelos con mayor tiempo de retraso.

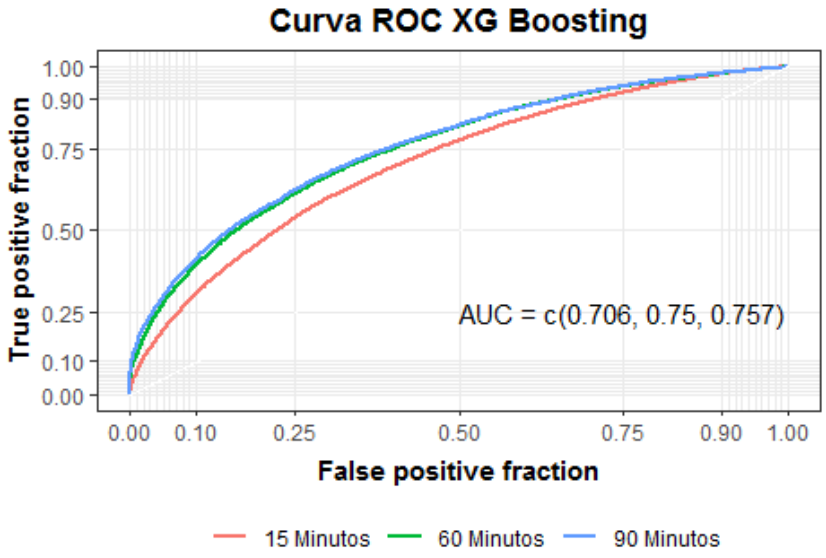


Ilustración 24 *Curva ROC modelo XGBoosting 15, 60 y 90 minutos*

En cuanto a las variables de mayor importancia, es necesario precisar 3 conceptos que la librería xgboost dispone para ilustrar el tema:

- Ganancia *(Gain)*: representa la mejora en precisión aportada por una variable para las ramas en las que se encuentre.
- Cobertura *(Cover)*: mide la cantidad relativa de observaciones asociadas a una variable.
- Frecuencia *(Frequency)*: como una manera más simple de expresar la ganancia, se refiere al número de ocasiones en que una variable ha sido usada en todos los árboles generados.

La literatura recomienda tener la ganancia como la métrica para identificar las variables de mayor importancia y en las iteraciones ejecutadas, el total de afectaciones experimentadas por un vuelo la semana anterior *(TotAfectadoSemAnt)*, la hora del día expresada en minutos *(HoraMin)*, la cantidad de pasajeros promedio de un vuelo al mes *(Pasajeros)* y la distancia entre origen y destino *(distancia)* se destacan como las variables de mayor contribución.

Posición	*Variable*	*Gain*	*Cover*	*Frequency*
1	*TotAfectadoSemAnt*	*488288,50*	*241559,20*	*12727,72*
2	*HoraMin*	*75000,71*	*111265,70*	*120708,50*
3	*Pasajeros*	*61787,18*	*89969,16*	*100592,70*

4	Distancia	58335,35	98673,62	67298,50
5	temp	47603,53	58413,20	111394,90
---	---	---	---	---
104	Origen.BUN	9,56	4,21	68,28
105	Origen.CZU	4,52	17,05	54,63
106	Smoke	2,86	15,89	54,63
107	Origen.MVP	2,72	1,47	13,66
108	Origen.IPI	2,32	0,16	13,66

Tabla 20 *Variables de mayor y menor peso según resultado del modelo XGBoosting*

La ilustración 25 permite ver que las primeras 6 variables tienen un peso más significativo que el resto, y que en una segunda instancia y con mucho menor contribución aparecen algunas variables climáticas y temporales.

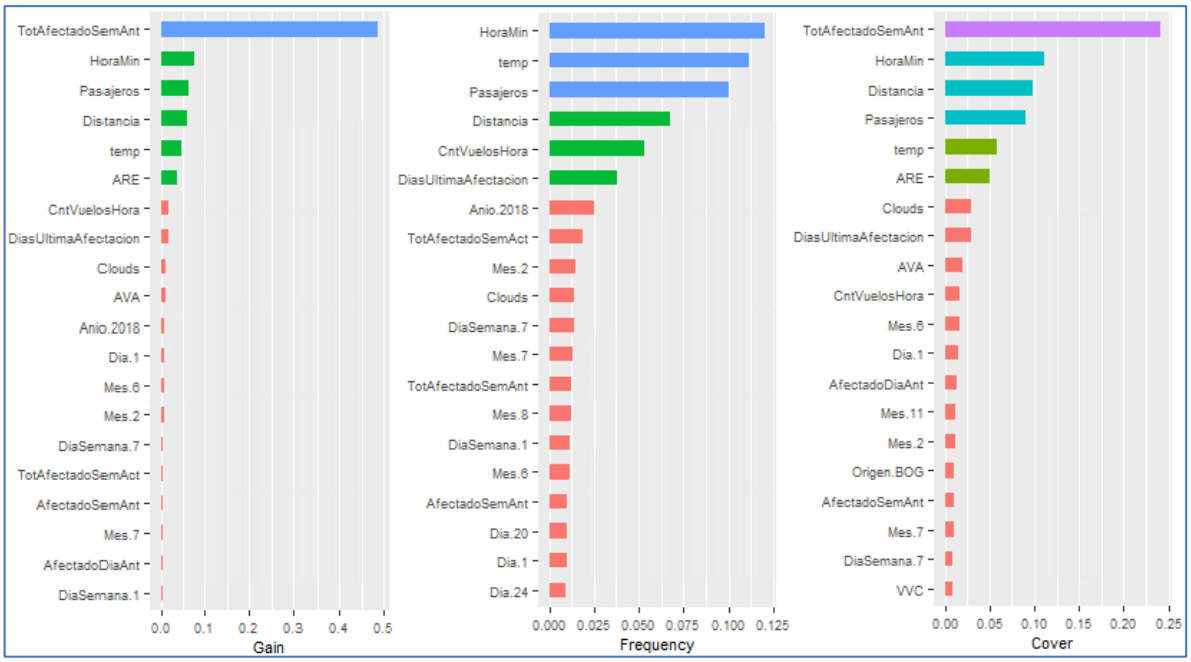


Ilustración 25 *Métricas de Importancia del modelo basado en XGBoosting*

13.3.5 Cuadro comparativo de los tres modelos creados

A continuación, se presenta la tabla comparativa de resultados de los tres modelos creados:

	Regresión Logística	***Redes Neuronales***	***XGBoosting***
Punto de corte	*0.16*	*0.16*	*0.22*
Accuracy	*0.6685*	*0.6840*	*0.6746*
Sensitividad	*0.6764*	*0.6873*	*0.6935*
Especificidad	*0.6669*	*0.6846*	*0.6707*
AUC	*0.736*	*0.763*	*0.75*

Tabla 19 *Comparativo de Regresión Logística, Redes Neuronales y XGBoosting.*

14 Conclusiones

En el presente trabajo se reconocieron factores que influyen en la afectación (cancelación o retraso) de vuelos que despegaron o tenían como destino el Aeropuerto Internacional El Dorado de la ciudad de Bogotá entre los años 2017 y 2018, incluyendo características del vuelo, condiciones operativas, climáticas y de temporalidad.

Los modelos implementados mejoran su desempeño a medida que se aumenta el horizonte temporal para clasificar los vuelos retrasados, entre 15, 60 y 90 minutos. La diferencia de los resultados obtenidos entre los cortes de 60 y 90 minutos es marginal y teniendo en cuenta la cantidad de vuelos con retrasos superiores a 60 minutos, se estableció este último como el mejor escenario de estimación.

La red neuronal es la técnica que presentó mejor desempeño. No obstante, las variables de mayor importancia identificadas entre la regresión logística y el XGBoosting son: total afectado semana anterior (TotAfectadoSemAnt), pasajeros y distancia. Esto permite establecer que, si bien los vuelos con alto número de pasajeros tienen mayor riesgo de experimentar una afectación como sería de esperarse (por la mayor complejidad de su operación), una propensión a la reincidencia en las afectaciones es bastante alta. Es decir, se estaría hablando de condiciones crónicas en las afectaciones estudiadas. En lo que respecta a la distancia, factores como la complejidad de operación entre rutas cortas también deben ser tenidos en cuenta: La pareja BOG-MDE es una ruta más corta que muchas otras, pero su volumen y complejidad la pueden hacer más propensa a afectación. Otro caso particular es la pareja BOG-VVC, que fue encontrada como origen y destino de mayor riesgo especialmente en la regresión logística: esto podría ser un efecto secundario del cierre de la Vía al Llano.

En cuanto a las variables temporales (no climáticas), la hora expresada en minutos fue la que mayor peso tuvo. El día 1 del mes también presentó una importancia mayor que los otros días en los modelos, lo cual es bastante curioso y permite inferir algún inconveniente logístico principalmente por parte de Avianca, el operador de mayor impacto en este renglón y que valida lo observado en la exploración de los datos.

En cuanto a las Aerolíneas: AVA *(Avianca)* fue observada como la más propensa a experimentar afectaciones (relacionado con su mayor número de operaciones); ESY *(EasyFly)* y NSE *(Satena)* destacaron también con un alto riesgo de retraso o cancelación según la regresión logística; ARE *(Aires)* fue otra aerolínea con una contribución considerable en la predicción del XGBoosting. Definitivamente, estas aerolíneas se encuentran más retadas en cumplimiento de itinerarios que el resto de las incluidas en este estudio.

Con las variables climáticas no se obtuvo un peso de tan alto impacto en los modelos (excepto la temperatura en el XGBoosting). No obstante, los factores climáticos son determinantes en la afectación de vuelos con operación en la ciudad de Bogotá y al ser el principal aeropuerto a nivel nacional su afectación no es ajena para el resto de las ciudades. Al ser un factor externo, se deben implementar estrategias operacionales y logísticas que disminuyan el impacto en cadena que se podría presentar.

En resumen, las condiciones observadas con información de los años 2017 y 2018 permiten establecer que los retrasos y cancelaciones en los vuelos entrantes o salientes del Aeropuerto El Dorado de la

ciudad de Bogotá están en mayor grado ligados a problemáticas dependientes de la operación de las aerolíneas que a factores como el clima o logística de los otros aeropuertos involucrados.

15 Siguientes pasos

Las principales causas de retraso de vuelos son: llegada tarde del avión precedente, causas imputables a las aerolíneas (embarque retrasado, demora en la carga y descarga de equipajes, fallas técnicas, problemas en la documentación del vuelo, falta o retraso de algún tripulante) y causas imputables al aeropuerto (problemas con el sistema de gestión de equipajes, tráfico aéreo, clima adverso). Para futuras mejoras de los modelos creados se recomienda el uso y análisis de datos operacionales propios de cada aerolínea, como la información de programación de tripulación, la programación del mantenimiento de los aviones y los datos relacionados con cada uno de los vuelos (tipo de avión, número del vuelo, número de pasajeros, número de sillas disponibles, entre otros). Se considera que al identificar la matricula del vuelo que hace recorridos de ida y vuelta o que hace vuelos consecutivos se podría identificar los casos en los cuales un retraso en uno de ellos pueda afectar el resto de la cadena y se pueden tomar decisiones operativas de reasignación de tripulación o de aerolínea con anterioridad.

Varias de las investigaciones que se han realizado parten de la capacidad de cada aeropuerto para manejar el tráfico aéreo. Con la información IATA disponible y los registros por cada país sería interesante calcular la capacidad real de un aeropuerto: la que permite la operación con un mínimo de congestión tolerable. Esto puede llevar a proponer prioridades para la expansión de infraestructura de los aeropuertos.

Para tener un mayor control sobre los itinerarios y operación de los vuelos se recomienda la creación de un sistema centralizado a nivel Nacional que consigne la información en tiempo de real de las aerolíneas con operación en los aeropuertos de Colombia, con el objetivo de analizar y mejorar la toma de decisiones del tráfico aéreo. La principal razón de esta recomendación es que de acuerdo con la circular informativa número 02 de la Aeronáutica Civil: “las empresas dedicadas a actividades aerocomerciales deberán enviar información que contenga las estadísticas mensuales de vuelos programados, cancelados, demorados y cumplidos dentro de los diez (10) primeros días calendario del mes siguiente al cual corresponde la información”. Es decir, actualmente se consolida la información operativa de las diferentes aerolíneas con 40 días de retraso.

Los incumplimientos (cancelaciones, demoras o sobreventas de vuelos) atribuibles a las aerolíneas genera compensaciones y multas altas. Por ejemplo, cuando un vuelo se cancela la aerolínea debe sufragar gastos de hospedaje, traslados y alimentación de los pasajeros, igualmente deberá compensar con el 30% del valor del trayecto afectado y pago de multas que impone la Aeronáutica civil. El impacto económico real que tienen las aerolíneas anualmente por incumplimiento de itinerarios puede enriquecer la justificación del proyecto.

Hay aerolíneas que tienen mayor incidencia en las afectaciones, por lo que se recomienda implementar estrategias que hagan un seguimiento detallado que permita identificar las causas de las afectaciones y así poner en marcha un plan de acompañamiento que mejore sus indicadores de cumplimiento. Un ejercicio de minería de procesos podría enriquecer el conocimiento existente de las operaciones y contribuir a la calidad de la gestión de éstas para no solo ofrecer un mejor servicio:

el reducir el tiempo en el que se usan los recursos en los diferentes aeropuertos y la reducción de penalidades favorecerán a los pasajeros, a las aerolíneas y a los aeropuertos en general.

16 Bibliografía

IATA - Home. (n.d.). Obtenida en Junio 16, 2020, de https://www.iata.org/

IATA. (2014). Worldwide Slot Guidelines. Wsg, August 2019. https://www.iata.org/policy/slots/pages/slot-guidelines.aspx

Página de inicio Aerocivil. (n.d.). Recopilado Junio 16, 2020, de http://www.aerocivil.gov.co

Aerocivil. (2003). RAC 3 - Regulación Actividades Aéreas Civiles - Aerocivil

Aerocivil. (2020). RAC 13 - Régimen sancionatorio.

COMUNIDAD ANDINA. (n.d.). RESOLUCION 1381 DE LA COMUNIDAD ANDINA.

¿Qué es una red neuronal? (n.d.). Recopilado Julio 14, 2020, de https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/neural_network/nnet_whatis.html

Tablas simples para variables categóricas. (n.d.). Recopilado Julio 14, 2020, de https://www.ibm.com/support/knowledgecenter/es/SSLVMB_sub/statistics_mainhelp_ddita/spss/tables/nt_simple_cat_tables.html

R Tutorials. (n.d.). Recopilado Julio 14, 2020, de http://ww2.coastal.edu/kingw/statistics/R-tutorials/index.html

Larrañaga, P. (n.d.). Redes Neuronales. (n.d.). Recopilado Julio 14, 2020, de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t8neuronales.pdf

Introduction to Extreme Gradient Boosting in Exploratory. (n.d.). Recopilado Julio 4, 2020, de https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). Springer Texts in Statistics An Introduction to Statistical Learning. Recopilado Julio 4, 2020, de http://www.springer.com/series/417

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. In Source: The Annals of Statistics (Vol. 29, Issue 5).

La pandemia y el sistema-mundo - Mundo - La Jornada. (n.d.). Recopilado Junio 23, 2020, de https://www.jornada.com.mx/ultimas/mundo/2020/04/25/ante-lo-desconocido-la-pandemia-y-el-sistema-mundo-7878.html (1)

Los cinco riesgos que amenazan la recuperación de las aerolíneas en 2021 - elEconomista.es. (n.d.). Recopilado Junio 23, 2020, de https://www.eleconomista.es/empresas-finanzas/noticias/10620156/06/20/Los-cinco-riesgos-que-amenazan-la-recuperacion-de-las-aerolineas-en-2021.html (2)

Cinco claves para entender la peor crisis de las aerolíneas en Latinoamérica. (n.d.). Recopilado Junio 23, 2020, de https://www.lapatria.com/internacional/cinco-claves-para-entender-la-peor-crisis-de-las-aerolineas-en-latinoamerica-459627(3)

Aerolíneas latinoamericanas en situación crítica: IATA | Aviación 21. (n.d.). Recopilado Junio 23, 2020, de https://a21.com.mx/index.php/aerolineas/2020/06/15/aerolineas-latinoamericanas-en-situacion-critica-iata (4-8)

Noticias más importantes portafolio 5 junio 2020 | Economía | Portafolio. (n.d.). Recopilado Junio 23, 2020, de https://www.portafolio.co/economia/noticias-mas-importantes-portafolio-5-junio-2020-541487(5)

Easyfly acumula pérdidas de 30.000 millones en su peor crisis | Noticias de turismo REPORTUR. (n.d.). Recopilado Junio 23, 2020, de https://www.reportur.com/aerolineas/2020/04/21/easyfly-acumulara-mayo-perdidas-30-000-millones/(6)

Copa admite que no tendrá liquidez para salir de la crisis | Noticias de turismo REPORTUR. (n.d.). Recopilado Junio 23, 2020, de https://www.reportur.com/aerolineas/2020/04/27/copa-airlines-no-tendria-suficiente-liquidez-resistir-la-crisis/(7)

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. Transportation Research Part C: Emerging Technologies, 44, 231–241. https://doi.org/10.1016/j.trc.2014.04.007

AhmadBeygi, S., Cohn, A., Guan, Y., & Belobaba, P. (2008). Analysis of the potential for delay propagation in passenger airline networks. Journal of Air Transport Management, 14(5), 221–236. https://doi.org/10.1016/j.jairtraman.2008.04.010

Pyrgiotis, N., Malone, K. M., & Odoni, A. (2013). Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, *27*, 60–75. https://doi.org/10.1016/j.trc.2011.05.017

Sternberg, A., Carvalho, D., Murta, L., Soares, J., & Ogasawara, E. (2016). An analysis of Brazilian flight delays based on frequent patterns. Transportation Research Part E: Logistics and Transportation Review, 95, 282–298. https://doi.org/10.1016/j.tre.2016.09.013

Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. (2016). Using scalable data mining for predicting flight delays. *ACM Transactions on Intelligent Systems and Technology*, *8*(1). https://doi.org/10.1145/2888402

Thiagarajan, B., Srinivasan, L., Sharma, A. V., Sreekanthan, D., & Vijayaraghavan, V. (2017). A machine learning approach for prediction of on-time performance of flights. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2017-Septe*. https://doi.org/10.1109/DASC.2017.8102138

Natarajan, V., Meenakshisundaram, S., Balasubramanian, G., & Sinha, S. (2018). A novel approach: Airline delay prediction using machine learning. *Proceedings - 2018 International Conference on Computational Science and Computational Intelligence, CSCI 2018*, 1081–1086. https://doi.org/10.1109/CSCI46756.2018.00210

Dand, A., Saeed, K., & Yildirim, B. (2019). Prediction of airline delays based on machine learning algorithms. *25th Americas Conference on Information Systems, AMCIS 2019*, 1–6.

Cao, Y., Zhu, C., Wang, Y., & Li, Q. (2019). A Method of Reducing Flight Delay by Exploring Internal Mechanism of Flight Delays. *Journal of Advanced Transportation*, *2019*, 1–8. https://doi.org/10.1155/2019/7069380

Chakrabarty, N. (2019). A data mining approach to flight arrival delay prediction for American airlines. IEMECON 2019 - 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference, 102–107. https://doi.org/10.1109/IEMECONX.2019.8876970

Chen, X., Yu, H., Cao, K., Zhou, J., Wei, T., & Hu, S. (2020). Uncertainty-Aware Flight Scheduling for Airport Throughput and Flight Delay Optimization. IEEE Transactions on Aerospace and Electronic Systems, 56(2), 853–862. https://doi.org/10.1109/TAES.2019.2921193

Baluch, M., Bergstra, T., & El-Hajj, M. (2017). Complex analysis of united states flight data using a data mining approach. 2017 IEEE 7th Annual Computing and Communication Workshop and Conference, CCWC 2017, 1–6. https://doi.org/10.1109/CCWC.2017.7868414

Alonso, H., & Loureiro, A. (2015). Predicting flight departure delay at Porto airport: A preliminary study. IJCCI 2015 - Proceedings of the 7th International Joint Conference on Computational Intelligence, 3, 93–98. https://doi.org/10.5220/0005587700930098

Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. Transportation Research Part E: Logistics and Transportation Review, 125, 203–221. https://doi.org/10.1016/j.tre.2019.03.013

Wu, W., Cai, K., Yan, Y., & Li, Y. (2019). An Improved SVM Model for Flight Delay Prediction. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2019-Septe*, 1–6.

https://doi.org/10.1109/DASC43569.2019.9081611

Abdel-Aty, M., Lee, C., Bai, Y., Li, X., & Michalak, M. (2007). Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, *13*(6), 355–361. https://doi.org/10.1016/j.jairtraman.2007.06.002

Ariyawansa, C. M., & Aponso, A. C. (2016). Review on state of art data mining and machine learning techniques for intelligent Airport systems. *Proceedings of 2016 International Conference on Information Management, ICIM 2016*, 134–138. https://doi.org/10.1109/INFOMAN.2016.7477547

Proenca, H. M., Klijn, R., Bäck, T., & Van Leeuwen, M. (2019). Identifying flight delay patterns using diverse subgroup discovery. Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018, 60–67. https://doi.org/10.1109/SSCI.2018.8628933

Meel, P., Singhal, M., Tanwar, M., & Saini, N. (2020). Predicting flight delays with error calculation using machine learned classifiers. 2020 7th International Conference on Signal Processing and Integrated Networks, SPIN 2020, 71–76. https://doi.org/10.1109/SPIN48934.2020.9071159

Lambelho, M., Mitici, M., Pickup, S., & Marsden, A. (2020). Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management*, *82*(May 2019), 101737. https://doi.org/10.1016/j.jairtraman.2019.101737

Chung, S. H., Ma, H. L., Hansen, M., & Choi, T. M. (2020). Data science and analytics in aviation. *Transportation Research Part E: Logistics and Transportation Review*, *134*(January). https://doi.org/10.1016/j.tre.2020.101837

Li, Q., Lei, W., Rong, F., Bin, W., & Hei, X. (2015). An analysis method for flight delays based on Bayesian network. Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015, 2561–2565. https://doi.org/10.1109/CCDC.2015.7162353

Prakash, J., & Bharathi, A. (2016). Predicting Flight Delay using ANN with Multi-core Map Reduce Framework. *Communication and Power Engineering*. https://doi.org/10.1515/9783110469608-028

Khanmohammadi, S., Tutun, S., & Kucuk, Y. (2016). A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. *Procedia Computer Science*, *95*, 237–244. https://doi.org/10.1016/j.procs.2016.09.321

Qin, Q. L., & Yu, H. (2014). A statistical analysis on the periodicity of flight delay rate of the airports in the US. *Advances in Transportation Studies*, *3*, 93–104. https://doi.org/10.4399/978885487831010

Xu, X., Yuan, H., & Qian, Y. (2015). Analyzing the system features of the flight delays: A network perspective. 2015 12th International Conference on Service Systems and Service Management, ICSSSM 2015. https://doi.org/10.1109/ICSSSM.2015.7170279

Dhanawade, R., Deo, M., Khanna, N., & Deolekar, R. V. (2019). Analyzing factors influencing flight delay prediction. *Proceedings of the 2019 6th International Conference on Computing for Sustainable Global Development, INDIACom 2019*, 1003–1007.

Rong, Y., Wang, J., & Xu, T. (2009). Prediction model and algorithm of flight delay propagation based on integrated consideration of critical flight resources. *2009 Second ISECS International Colloquium on Computing, Communication, Control, and Management, CCCM 2009*, *2*, 98–102. https://doi.org/10.1109/CCCM.2009.5267970

Du, W. B., Zhang, M. Y., Zhang, Y., Cao, X. Bin, & Zhang, J. (2018). Delay causality network in air transport systems. Transportation Research Part E: Logistics and Transportation Review, 118(February), 466–476. https://doi.org/10.1016/j.tre.2018.08.014

Ding, Y. (2017). Predicting flight delay based on multiple linear regression. IOP Conference Series: Earth and Environmental Science, 81(1). https://doi.org/10.1088/1755-1315/81/1/012198

Tu, Y., Ball, M. O., & Jank, W. S. (2008). Estimating flight departure delay distributions - A statistical approach with long-term trend and short-term pattern. Journal of the American Statistical Association, 103(481), 112–125. https://doi.org/10.1198/016214507000000257

Pamplona, D. A., Weigang, L., De Barros, A. G., Shiguemori, E. H., & Alves, C. J. P. (2018). Supervised Neural Network with multilevel input layers for predicting of air traffic delays. *Proceedings of the International Joint Conference on Neural Networks*, *2018-Julio*. https://doi.org/10.1109/IJCNN.2018.8489511

Wang, Y., Cao, Y., Zhu, C., Wu, F., Hu, M., Duong, V., Watkins, M., Barzel, B., & Stanley, H. E. (2020). Universal patterns in passenger flight departure delays. *Scientific Reports*, *10*(1), 6890. https://doi.org/10.1038/s41598-020-62871-6

Tang, C.-H. (2011). A Gate Reassignment Model for the Taiwan Taoyuan Airport Under Temporary Gate Shortages and Stochastic Flight Delays. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 41(4), 637–650. https://doi.org/10.1109/TSMCA.2010.2089512

Novianingsih, K., & Hadianti, R. (2014). Modeling flight departure delay distributions. *Proceeding - 2014 International Conference on Computer, Control, Informatics and Its Applications: "New Challenges and Opportunities in Big Data", IC3INA 2014*, 30–34. https://doi.org/10.1109/IC3INA.2014.7042596

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P., & Barman, S. (2018). A statistical approach to predict flight delay using gradient boosted decision tree. *ICCIDS 2017 - International Conference on Computational Intelligence in Data Science, Proceedings*, *2018-Janua*, 1–5. https://doi.org/10.1109/ICCIDS.2017.8272656

Cheng, J. (2015). Estimation of flight delay using weighted Spline combined with ARIMA model. Proceedings of 2014 IEEE 7th International Conference on Advanced Infocomm Technology, IEEE/ICAIT 2014, 8–20. https://doi.org/10.1109/ICAIT.2014.7019523

Moreira, L., Dantas, C., Oliveira, L., Soares, J., & Ogasawara, E. (2018). On Evaluating Data Preprocessing Methods for Machine Learning Models for Flight Delays. *Proceedings of the International Joint Conference on Neural Networks*, *2018-Julio*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489294

Cheng, S., Zhang, Y., Hao, S., Liu, R., Luo, X., & Luo, Q. (2019). Study of Flight Departure Delay and Causal Factor Using Spatial Analysis. *Journal of Advanced Transportation*, *2019*, 1–11. https://doi.org/10.1155/2019/3525912

Hopane, J., & Gatsheni, B. (2019). A computational intelligence-based prediction model for flight departure delays. Proceedings - 6th Annual Conference on Computational Science and Computational Intelligence, CSCI 2019, 564–571. https://doi.org/10.1109/CSCI49370.2019.00107

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight delay prediction based on aviation big data and machine learning. IEEE Transactions on Vehicular Technology, 69(1), 140–150.

https://doi.org/10.1109/TVT.2019.2954094

Zhang, H., Wu, W., Zhang, S., & Witlox, F. (2020). Simulation Analysis on Flight Delay Propagation Under Different Network Configurations. *IEEE Access*, *8*, 103236–103244. https://doi.org/10.1109/ACCESS.2020.2999098

Dou, X. (2020). Flight Arrival Delay Prediction And Analysis Using Ensemble Learning. 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 836–840. https://doi.org/10.1109/itnec48623.2020.9084929

ANEXO 1
PONTIFICIA UNIVERSIDAD JAVERIANA
BIBLIOTECA ALFONSO BORRERO CABAL, S.J.
ENTREGA DE TESIS Y TRABAJOS DE GRADO

FACULTAD: Ingeniería Industrial
PROGRAMA: Maestría en Analítica para la Inteligencia de Negocios
FECHA DE ENTREGA: 31-07-2020

APELLIDOS COMPLETOS	NOMBRES COMPLETOS	TITULO DE LA TESIS O DEL TRABAJO DE GRADO	NOMBRE DEL DIRECTOR	AÑO	Documentos adjuntos (Marque con x)		
					Anexo 2	Anexo 3	Carta de confidencia-lidad
Ramírez Quiroga	**William Alfonso**	**MODELO PARA IDENTIFICAR LOS VUELOS AFECTADOS POR RETRASOS O CANCELACIONES EN EL AEROPUERTO EL DORADO DE BOGOTÁ, COLOMBIA**	**Luis Manuel Pulido Moreno**	**2020**	**X**	**X**	**N/A**
Chavarro Cely	**Camilo Andrés**						
Arias Maury	**Carlos Alberto**						

DILIGENCIADO POR (Nombres y Apellidos): _____

CARGO: _____

FIRMA: _____

ANEXO 2

CARTA DE AUTORIZACIÓN DE LOS AUTORES

(Licencia de uso)

Bogotá, D.C., Julio 30 de 2020

Señores
Biblioteca Alfonso Borrero Cabal S.J.
Pontificia Universidad Javeriana
Cuidad

Los suscritos:

William Alfonso Ramírez Quiroga, con C.C. No 80.497.013

Camilo Andrés Chavarro Cely, con C.C. No 1.019.051.959

Carlos Alberto Arias Maury, con C.C. No 80.842.615

En mi (nuestra) calidad de autor (es) exclusivo (s) de la obra titulada:

MODELO PARA IDENTIFICAR LOS VUELOS AFECTADOS POR RETRASOS O CANCELACIONES EN EL AEROPUERTO EL DORADO DE BOGOTÁ, COLOMBIA (por favor señale con una "x" las opciones que apliquen)

Tesis doctoral [ ] Trabajo de grado [X] Premio o distinción: **Si** [ ] **No** [X]

cual: N/A

presentado y aprobado en el año 2020 , por medio del presente escrito autorizo (autorizamos) a la Pontificia Universidad Javeriana para que, en desarrollo de la presente licencia de uso parcial, pueda ejercer sobre mi (nuestra) obra las atribuciones que se indican a continuación, teniendo en cuenta que en cualquier caso, la finalidad perseguida será facilitar, difundir y promover el aprendizaje, la enseñanza y la investigación.

En consecuencia, las atribuciones de usos temporales y parciales que por virtud de la presente licencia se autorizan a la Pontificia Universidad Javeriana, a los usuarios de la Biblioteca Alfonso Borrero Cabal S.J., así como a los usuarios de las redes, bases de datos y demás sitios web con los que la Universidad tenga perfeccionado un convenio, son:

AUTORIZO (AUTORIZAMOS)	SI	NO
1. La conservación de los ejemplares necesarios en la sala de tesis y trabajos de grado de la Biblioteca.	X	
2. La consulta física (sólo en las instalaciones de la Biblioteca)	X	
3. La consulta electrónica – on line (a través del catálogo Biblos y el Repositorio Institucional)	X	
4. La reproducción por cualquier formato conocido o por conocer	X	

AUTORIZO (AUTORIZAMOS)	SI	NO
5. La comunicación pública por cualquier procedimiento o medio físico o electrónico, así como su puesta a disposición en Internet	X	
6. La inclusión en bases de datos y en sitios web sean éstos onerosos o gratuitos, existiendo con ellos previo convenio perfeccionado con la Pontificia Universidad Javeriana para efectos de satisfacer los fines previstos. En este evento, tales sitios y sus usuarios tendrán las mismas facultades que las aquí concedidas con las mismas limitaciones y condiciones	X	

De acuerdo con la naturaleza del uso concedido, la presente licencia parcial se otorga a título gratuito por el máximo tiempo legal colombiano, con el propósito de que en dicho lapso mi (nuestra) obra sea explotada en las condiciones aquí estipuladas y para los fines indicados, respetando siempre la titularidad de los derechos patrimoniales y morales correspondientes, de acuerdo con los usos honrados, de manera proporcional y justificada a la finalidad perseguida, sin ánimo de lucro ni de comercialización.

De manera complementaria, garantizo (garantizamos) en mi (nuestra) calidad de estudiante (s) y por ende autor (es) exclusivo (s), que la Tesis o Trabajo de Grado en cuestión, es producto de mi (nuestra) plena autoría, de mi (nuestro) esfuerzo personal intelectual, como consecuencia de mi (nuestra) creación original particular y, por tanto, soy (somos) el (los) único (s) titular (es) de la misma. Además, aseguro (aseguramos) que no contiene citas, ni transcripciones de otras obras protegidas, por fuera de los límites autorizados por la ley, según los usos honrados, y en proporción a los fines previstos; ni tampoco contempla declaraciones difamatorias contra terceros; respetando el derecho a la imagen, intimidad, buen nombre y demás derechos constitucionales. Adicionalmente, manifiesto (manifestamos) que no se incluyeron expresiones contrarias al orden público ni a las buenas costumbres. En consecuencia, la responsabilidad directa en la elaboración, presentación, investigación y, en general, contenidos de la Tesis o Trabajo de Grado es de mí (nuestro) competencia exclusiva, eximiendo de toda responsabilidad a la Pontifica Universidad Javeriana por tales aspectos.

Sin perjuicio de los usos y atribuciones otorgadas en virtud de este documento, continuaré (continuaremos) conservando los correspondientes derechos patrimoniales sin modificación o restricción alguna, puesto que de acuerdo con la legislación colombiana aplicable, el presente es un acuerdo jurídico que en ningún caso conlleva la enajenación de los derechos patrimoniales derivados del régimen del Derecho de Autor.

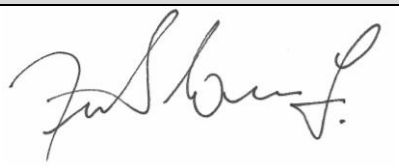
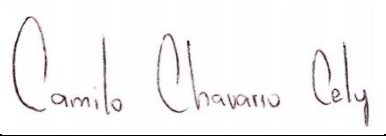
De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, *"Los derechos morales sobre el trabajo son propiedad de los autores"*, los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. En consecuencia, la Pontificia Universidad Javeriana está en la obligación de RESPETARLOS Y HACERLOS RESPETAR, para lo cual tomará las medidas correspondientes para garantizar su observancia.

NOTA: <u>Información Confidencial:</u>

Esta Tesis o Trabajo de Grado contiene información privilegiada, estratégica, secreta, confidencial y demás similar, o hace parte de una investigación que se adelanta y cuyos

resultados finales no se han publicado. Si [ ] No [X]

En caso afirmativo expresamente indicaré (indicaremos), en carta adjunta, tal situación con el fin de que se mantenga la restricción de acceso.

NOMBRE COMPLETO	No. del documento de identidad	FIRMA
WILLIAM ALFONSO RAMIREZ QUIROGA	80.497.013	
CAMILO ANDRÉS CHAVARRO CELY	1.019.051.959	
CARLOS ALBERTO ARIAS MAURY	80.842.615	

FACULTAD: Ingeniería Industrial

PROGRAMA ACADÉMICO: Maestría en Analítica para la Inteligencia de Negocios.

ANEXO 3

BIBLIOTECA ALFONSO BORRERO CABAL, S.J.
DESCRIPCIÓN DE LA TESIS O DEL TRABAJO DE GRADO

FORMULARIO

TÍTULO COMPLETO DE LA TESIS DOCTORAL O TRABAJO DE GRADO			
MODELO PARA IDENTIFICAR LOS VUELOS AFECTADOS POR RETRASOS O CANCELACIONES EN EL AEROPUERTO EL DORADO DE BOGOTÁ, COLOMBIA			
SUBTÍTULO, SI LO TIENE			
N/A			
AUTOR O AUTORES			
Apellidos Completos	**Nombres Completos**		
Ramírez Quiroga	William Alfonso		
Chavarro Cely	Camilo Andrés		
Arias Maury	Carlos Alberto		
DIRECTOR (ES) TESIS O DEL TRABAJO DE GRADO			
Apellidos Completos	**Nombres Completos**		
Pulido Moreno	Luis Manuel		
FACULTAD			
Ingeniería Industrial			
PROGRAMA ACADÉMICO			
Tipo de programa (seleccione con "x")			
Pregrado	Especialización	Maestría	Doctorado
		X	
Nombre del programa académico			
Maestría en Analítica para la Inteligencia de Negocios			
Nombres y apellidos del director del programa académico			
Jorge Andrés Alvarado Valencia			
TRABAJO PARA OPTAR AL TÍTULO DE:			
Magister en analítica para la Inteligencia de Negocios			
PREMIO O DISTINCIÓN *(En caso de ser LAUREADAS o tener una mención especial)*:			
N/A			

CIUDAD	AÑO DE PRESENTACIÓN DE LA TESIS O DEL TRABAJO DE GRADO	NÚMERO DE PÁGINAS
Bogotá D.C	**2020**	**45**

TIPO DE ILUSTRACIONES (seleccione con "x")

Dibujos	Pinturas	Tablas, gráficos y diagramas	Planos	Mapas	Fotografías	Partituras
		x				

SOFTWARE REQUERIDO O ESPECIALIZADO PARA LA LECTURA DEL DOCUMENTO

Nota: En caso de que el software (programa especializado requerido) no se encuentre licenciado por la Universidad a través de la Biblioteca (previa consulta al estudiante), el texto de la Tesis o Trabajo de Grado quedará solamente en formato PDF.

N/A

MATERIAL ACOMPAÑANTE

TIPO	DURACIÓN (minutos)	CANTIDAD	FORMATO		
			CD	DVD	Otro ¿Cuál?
Vídeo	N/A				
Audio	N/A				
Multimedia	N/A				
Producción electrónica	N/A				
Otro Cuál?	N/A				

DESCRIPTORES O PALABRAS CLAVE EN ESPAÑOL E INGLÉS

Son los términos que definen los temas que identifican el contenido. (*En caso de duda para designar estos descriptores, se recomienda consultar con la Sección de Desarrollo de Colecciones de la Biblioteca* Alfonso Borrero Cabal S.J *en el correo biblioteca@javeriana.edu.co, donde se les orientará).*

ESPAÑOL	INGLÉS
Aerocivil, IATA, Aeropuerto Internacional El Dorado, Bogotá, retrasos, cancelaciones, Clima, itinerario, exactitud, precisión, Sensitividad, especificidad, pasajeros, Vuelos.	**Aerocivil, IATA, El Dorado International Airport, Bogotá, Delays, Cancelation, weather, itinerary, accuracy, precision, sensitivity, specificity, passengers, Flights.**

RESUMEN DEL CONTENIDO EN ESPAÑOL E INGLÉS

(Máximo 250 palabras - 1530 caracteres)

Este trabajo está basado en el análisis de factores climáticos y operacionales de las aerolíneas con operación en Colombia. El factor operacional contiene el detalle de los vuelos que tienen lugar en los aeropuertos del país con variables como origen, destino, número de vuelo, aerolínea, fecha y hora programada, fecha y hora de remolque, estado del vuelo (adelantado, cumplido, retrasado y cancelado), cantidad de pasajeros, cantidad de carga, distancia y tiempo de vuelo entre otras. Por el gran peso e importancia que tiene el Aeropuerto El Dorado de Bogotá, el análisis y modelo resultado de este trabajo se centró en la operación y factores climáticos que tienen incidencia en este terminal aéreo.

Por medio de técnicas como regresión logística, redes neuronales y XGboosting se logró predecir en la base de datos de pruebas cerca del 70% de los vuelos afectados por cancelaciones o retrasos en el aeropuerto de la capital colombiana.

This work is based on the analysis of weather and operational factors of the airlines operating in Colombia. The operational factor contains the detail of the flights that take place in the country's airports with variables such as origin, destination, flight number, airline, scheduled date and time, towing date and time, flight status (early, on time, delayed and canceled), number of passengers, amount of cargo, distance and flight time among others. Due to the great weight and importance of the El Dorado Airport of Bogotá, the analysis and model resulting from this work focused on the operation and weather factors that have an impact on this Airport.

Using techniques such as logistic regression, neural networks and XGboosting, it was possible to predict in the test database about 70% of flights affected by cancellations or delays at the Colombian capital airport.