PROPUESTA DE MODELO PREDICTIVO EN RETRASO DE DESPEGUE PARA AEROPUERTOS DE MEDIANA COMPLEJIDAD EN COLOMBIA

AUTOR:
SERGIO ANDRES DIAZ CASTRO
ANDRES SANTIAGO SANDOVAL MORALES

PROYECTO DE GRADO PARA OPTAR A TITULO DE INGENIERIA INDUSTRIAL

ASESOR:
INGENIERO LUIS MANUEL PULIDO RICO

UNIVERSIDAD SANTO TOMAS DE AQUINO
BOGOTA COLOMBIA
2020

CONTENIDO

INDICE DE ILUSTRACIONES, GRAFICOS Y TABLAS

Índice de ilustraciones

Índice de Gráficos

Índice de Tablas

1. PLANTEAMIENTO DEL PROBLEMA DE INVESTIGACION

1.1 ANTECEDENTES DEL PROBLEMA

Los retrasos en los vuelos son desde siempre una constante en los viajes aéreos, sin embargo, se intuye que existen desde la popularización de los vuelos comerciales en 1910 con la fundación de las primeras compañías de vuelos comerciales [1]. A partir de este año se popularizaron los vuelos comerciales a través del mundo, pero hasta el año 1919 que los vuelos comerciales llegaron a territorio colombiano de la mano de la alianza colombo Alemana de transporte [1]; este hecho hizo a Colombia un país pionero ya que fue el primer país en implementar el transporte aéreo comercial en territorio americano, con una flota de 25 aeronaves comenzó la historia aeronáutica en Colombia. [1]

Sin embargo, no se tiene certeza de cuándo y cómo fue el primer retraso en un vuelo comercial, pero se podría deducir que su causa fue por fallas mecánicas, condiciones climatológicas adversas o por el estado de la pista, sea cual sea la causa este hecho fue el primero, pero no el ultimo ya que desde entonces los retrasos en los vuelos son un problema que aqueja a los usuarios de las aerolíneas desde su invención. [2]

No fue hasta la sistematización, de la información aeronáutica que se comenzaron a tener registro de los retrasos en los aeropuertos. Desde la aplicación de los ordenadores en el mundo aeroportuario se comenzaron a recopilar datos y la aeronáutica civil colombiana no fue ajena a este avance tecnológico y desde 2004 recopila datos públicos con información relevante sobre los vuelos en Colombia. [2]. La recopilación de datos demuestra que entre 2009 y 2010 Colombia tuvo un alza en 50.000 vuelos realizados en territorio nacional, así mismo los retrasos según el portal américa economía es del 30%. Es decir que en Colombia los vuelos presentan retrasos en más de 129,000 de los 431.664 que se realizaron en 2019. [3].

Esta tasa de retraso en los vuelos en conjunto con el aumento anual de los vuelos a nivel nacional, representa un problema latente, que poco a poco reduce la calidad en los terminales aéreos ya que, según estadísticas, 3´121,373 pasajeros fueron afectados por retrasos en sus vuelos en los últimos 5 años. [2].

En base a lo anterior se puede deducir que los retrasos aeroportuarios son un problema que obstruye el óptimo funcionamiento de las terminales aéreas colombianas y reduce la calidad de los vuelos no solo para turistas sino también para el transporte de mercancía que circula por la red aérea nacional, es por esto que la presente investigación tendrá como objetivo encontrar las causalidades más influyentes en el retraso de despegue en las terminales aéreas de Barranquilla, Cali

y Cartagena; así mismo poder clasificar estas causas según su grado de importancia, con el fin de desarrollar un algoritmo predictivo que permita mediante herramientas estadísticas y en base a su comportamiento, poder prever situaciones de retrasos en despegues; se utilizarán datos de retrasos en despegue del año 2019, en los tres aeropuertos anteriormente mencionados los cuales son considerados aeropuertos de baja complejidad.

1.2 PREGUNTA PROBLEMA

¿Cuál es el modelo más apropiado para pronosticar y entender las causalidades de los retrasos en los despegues presentes en las terminales aéreas de Barranquilla, Cali y Cartagena?

2. JUSTIFICACION

En 2013, más de medio millón de vuelos se vieron afectados por un fallo técnico en el sistema de reserva de vuelos de American Airlines, como consecuencia miles de pasajeros en estados unidos sufrieron retrasos o cancelación en su vuelo, incluso aquellos que ya habían embarcado [4]; este precedente evidencia el delicado equilibrio con el que funcionan las operaciones aéreas de cualquier país.

Este hecho, demuestra que cualquier evento imprevisto afecta el correcto funcionamiento de los aeropuertos, así mismo se demuestra que las actividades involucradas en la operación de un aeropuerto, deben estar en sincronía horaria en todo momento, ya que un error en cualquier parte de la operación implica retrasos o cancelación en los vuelos desde ese momento, lo que conlleva a la reprogramación de los vuelos.

En la historia contemporánea de Colombia, el comercio electrónico y el avance de la industria del turismo han presentado un crecimiento importante, registrando ventas electrónicas por 26 Billones de pesos el año pasado y un crecimiento del 6.2% en las operaciones aéreas [2]. La combinación de medidas de política exterior y lo expuesto anteriormente, permitió a la economía nacional extender sus fronteras comerciales a lugares inexplorados económicamente [5].
Los retrasos en estas cargas representan una pérdida económicamente considerable, las cuales deben ser reparadas por las transportadoras, esto es indicado en la legislación y es mejor conocido como daño emergente y renta cesante, respaldados legalmente por el artículo 1614 del código civil en donde se especifica que se debe reparar el daño causado ya sea por la pérdida de la mercancía o por el dinero que se deja de recibir por la inoperancia. [6]

Este, es bien conocido por la comunidad que frecuenta las terminales aéreas, que cualquier retraso en un vuelo no puede ser cobrado, y en algunos casos las aerolíneas o el aeropuerto deben suplir los gastos adicionales generados por esta demora en el servicio. Es por lo anteriormente mencionado, que las consecuencias económicas que los retrasos han generado son de más de 13 mil millones de dólares en EEUU según lo indica el estudio *“Total Delay Impact Study”* en el cual participaron 10 expertos en el tema y además de la anterior afirmación, también estiman pérdidas por 8 millones de dólares en Latinoamérica y en Colombia, en el 2019 las empresas indemnizaron a más de 100.000 pasajeros con costo total de más de 20 mil millones de pesos. [7]

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Proponer un modelo de predicción para retrasos en despegues aéreos de los aeropuertos de Barranquilla, Cali y Cartagena considerados de mediana complejidad

3.2 OBJETIVO ESPECIFICO

1. Analizar la tendencia de los datos más relevantes a partir de la caracterización de la base de datos de la aeronáutica civil colombiana.
2. Identificar las principales causas de retrasos en despegues, en los aeropuertos de Barranquilla, Cali y Cartagena.
3. Desarrollar un algoritmo en un software que permita prever los posibles retrasos en despegues, en los aeropuertos de Barranquilla, Cali y Cartagena.

4. MARCO REFERENCIAL

4.1 MARCO CONCEPTUAL

4.1.1 Conceptos básicos.

En esta sección el lector encontrará la definición particular de los términos más importantes para la comprensión del documento.

- Terminal aérea: También llamado aeropuerto, es según la RAE *"área destinada al aterrizaje y despegue de aviones dotada de instalaciones para el control del tráfico aéreo y de servicio a los pasajeros"* [8]*;* En resumen, es un espacio adecuado y destinado para el desarrollo de operaciones aéreas.

- Modelo de análisis predictivo: Modelo basado en un área de la estadística, como lo es la minería de datos, que consiste en la extracción de datos de un escenario real, para elaborar un supuesto teórico que permite hacer una estimación numérica de los valores que tomarán las diferentes variables que influyen en un escenario especifico [9].

  El modelo predictivo que más se adapta a los escenarios aéreos es el modelo 'Random forest' el cual es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. [10]

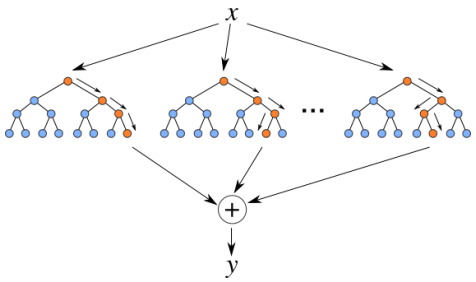

Ilustración 1 Modelo ramdon forest Fuente: Nerea.

- Minería de datos: Es un área de la estadística que tiene como objetivo encontrar anomalías, relaciones y patrones, en un conjunto de datos de gran volumen, esto por diferentes herramientas estadísticas y computacionales [11].

- Modelos de análisis comparativos: Un modelo de análisis comparativo es en esencia un estudio en donde se comparan objetivamente dos o varios escenarios anteriores, que tengan similitudes en sus características, para realizar una hipótesis en base a los resultados de un escenario pasado. [12].

- Estadística Aplicada: Es la aplicación práctica de los conceptos estadísticos teóricos, la cual utiliza un conjunto de probabilidades para hacer la estimación aproximada de las variables estudiadas. [13].

- Recursos informáticos: Los recursos informáticos son las herramientas de apoyo que se utilizarán para el manejo de los datos y la estructuración de un modelo predictivo funcional, entiéndase también como el conjunto de dispositivos computacionales que tendrán uso en el presente documento. [14]

4.1.2 CARACTERIZACION DE LAS TERMINALES AEREAS

En Colombia no existe una clasificación que permita delimitar la información pertinente para este documento, sin embargo, en base al criterio adquirido en este documento se clasifican de la siguiente forma:

- Terminales aéreos de alta complejidad: Estas terminales deberán contar con más de 15 destinos internacionales y un volumen de pasajeros superior a 100.000 pasajeros anuales.

- Aeropuertos de mediana complejidad: Estas terminales deben contar con al menos 1 destino internacional y un máximo de 15 destinos nacionales además deben tener un volumen de pasajeros superior a los 10.000 usuarios anuales y un máximo de 100.000.

- Terminales aéreas de baja complejidad: Estas terminales aéreas se caracterizan por la ausencia de destinos internacionales y un volumen de no más de 10.000 pasajeros anuales, siendo esta última la clasificación de la mayoría de las terminales aéreas en Colombia.

- Las terminales aéreas tengan un flujo de pasajeros menor a 50.000 y mayor a 10.000 pero no posean destinos internacionales serán consideradas como aeropuertos de mediana complejidad. así mismo los aeropuertos que superen los 100.000 pasajeros, pero no posean más de 15 destinos internacionales estarán clasificadas como aeropuertos de mediana complejidad.

4.2 MARCO TEÓRICO

4.2.1 Operación aérea en Colombia

En la ciudad de Barranquilla mientras el mundo celebraba el final de la primera guerra mundial, la ciudad mejor conocida como la arenosa celebraba la construcción de su aeropuerto el cual fue el primer aeropuerto de Colombia bautizado Veranillo, el cual fue construido por la Sociedad Colombo alemana de transporte aéreo, en el año de 1920 como fruto de la relación de varios empresarios colombianos y 3 inversionistas alemanes que conformaban SCADTA.

En el transcurso de las décadas el crecimiento del tráfico aéreo en Colombia ha presentado una creciente demanda y en 1990 mediante una política pública

orientada a liberalizar el espacio aéreo para los mercados internos y externos, con el fin de reenfocar la inversión pública hacia la modernización y actualización de la infraestructura aeroportuaria, externalizando los aeropuertos con mayor tráfico aéreo en el país el cual actualmente cuenta con 18 puertos aéreos en Colombia.

La primera generación de concesiones aeroportuarias se concretó a mediados de la década de 1990, desde entonces se han desarrollado tres generaciones más, contemporáneamente, la geografía fue relevante ya que la concesión del aeropuerto siguió un patrón fuertemente descentralizador y lo hizo geográficamente lo más disperso a lo largo de la geografía cafetera. Todos los aeropuertos en concesión están distribuidos en 13 departamentos administrativos junto con el distrito capital (Bogotá). Y en mayo de 2016 se espera la finalización de la actual cuarta generación de concesiones de concesión a otros dos aeropuertos situados en dos departamentos administrativos diferentes de los anteriores.

Esta amplia cobertura geográfica del sistema aeroportuario colombiano ha mejorado el movimiento de pasajeros aéreos (nacionales), especialmente en la última década, dando una solución de movilidad debido a las circunstancias antes mencionadas.

4.2.2 Retrasos aeroportuarios

4.2.2.1 Causalidad.

Las causas de retrasos aéreos identificadas por la aeronáutica civil superan las 50 causas, estas estar organizadas en 3 clasificaciones, Eventos, Motivos y Causas. [15]

Sin embargo en este documento solo trataremos las causas referenciadas en el documento “Predicción y Análisis de los Retrasos en los Vuelos” presentado como trabajo de grado de Martínez Nerea en el pregrado de gestión aeronáutica para la universidad de Barcelona, estas causas han sido previamente identificadas por la institución educativa virtual Euro aula, esta identificación fue respaldada por un estudio que realizó EasyFly, donde se puede observar que los retrasos aeroportuarios se resumen en 5 causas principales, las cuales son:

1. Acumulación de tráfico aéreo: Entre el aterrizaje y el despegue todas las aerolíneas deben mantener un *tiempo de escala* que debe ser determinados por las mismas. Dónde están incluidas todas las operaciones que deben hacer tanto las empresas, como los usuarios. Entonces, con que en un vuelo no se cumpla con este tiempo, ya generará que las demás se retrasen. Esto puede ser manejable para empresas con pequeñas flotas y finalmente, esta causa compone el 29% de los retrasos. [16]

2. Verificación de los aparatos: Este aspecto, representa el 26% de los retrasos y está muy ligado al *tiempo de escala* ya que, la verificación de aparatos debe estar incluido en el mismo. Puede tratarse de comprobaciones obligatorias,

reparaciones, componentes que aseguran la calidad del vuelo (Tanto técnicos como estéticos). [16]

3. Gestión aeroportuaria: Hace referencia a todos los controles que los aeropuertos tuvieron que crear debido al terrorismo y compone el 14% de los retrasos en este estudio. En estos se incluyen los registros en aduana, registro de equipaje y revisión de equipajes de mano. Sin embargo, este aspecto se presenta con mayor regularidad en la temporada alta. [16]

4. Navegación aérea y pasajeros: Respecto a la navegación aérea, hace referencia a los retrasos que se presentan externos a las compañías prestadoras de servicio, como lo es la gestión de despegues y aterrizajes de cada aeropuerto, donde se da un orden de prioridad en los vuelos; y los pasajeros, ocurre cuando no cumplen con la hora de abordaje establecida y generan incumplimiento en la hora de despegue; estos dos componen el 13% respectivamente de las causas. [16]

5. La meteorología: Este hace referencia a todos los eventos naturales, que son incontrolables por el hombre, como ciclones, tormentas, hielo y tempestades de nieve. Sin embargo, no ocurre con mucha frecuencia por ende solo compone el 5% de las causas. [16]

Las causas mencionadas anteriormente son respaldadas por La organización europea para la seguridad de la navegación aérea en europea mejor conocida como EUROCONTROL, la cual, en su publicación del 3 de agosto de 2018, Identificó 3 causas comunes en los retrasos de la red aérea europea, Llegada tarde del avión precedente, Causas imputables a la aerolínea, Causas imputables al aeropuerto, estas clasificaciones respaldan los motivos de retraso mencionados por Nerea. [17]

4.2.2.2 Consecuencias.

La operación aérea depende de muchos actores, pero en resumen son 3 principales, usuarios, personal del aeropuerto y empresas prestadoras de servicios aéreos, y las consecuencias de un retraso en la operación, afecta directa o indirectamente a todos estos actores.

Para el año de 2018 Colombia tuvo un alza de 6% en cuanto a transporte aéreo de carga, dato mencionado por la revista especializada en logística en el artículo *“El transporte aéreo en Colombia sigue en alza”*; El transporte de carga aérea en Colombia es en su mayoría medicamentos, productos tecnológicos y mariscos exóticos, transportando más de tres millones de toneladas anualmente. [18]

Como ya se mencionó anteriormente las empresas transportadoras tienen la obligación de reparar los daños generados, sin embargo el Código de Comercio cuenta con un reglamento que estipula en caso de cancelación de parte de la aerolínea, se debe hacer devolución del tiquete, si el vuelo es suspendido por causas meteorológicas o de fuerza mayor la aerolínea debe transportar a los

pasajeros a el destino inicial en otro vuelo dispuesto por la aerolínea o descender y reembolsar el dinero equivalente a la distancia restante esto solo en caso que el pasajero lo desee; en caso de quedar en una ciudad o país diferente a la del destino las empresas deben garantizar la alimentación proveer de medios de comunicación y servicio de asesoría, incluso algunas compañías dispondrán viajes en aerolíneas distintas, todos los gastos extras deben ser asumidos por la compañía. lo anterior es respaldado por el Artículo 1870 del código de comercio para empresas extranjeras y el artículo 981 del mismo código, en caso de incumplimiento de estos mandatos puede deducirse en indemnizaciones sanciones restrictivas de operación o en la prohibición de operación en Colombia de una empresa de servicios aéreos. [19]

4.2.3 ESTADISTICA APLICADA

4.2.3.1 Minería de datos.

La minería de datos, consiste en diferentes técnicas que se utilizan para identificar la secuencia o patrones en grandes volúmenes de datos. La finalidad de este consiste en transformar la información a una estructura más comprensible para un uso posterior. [20]

Para utilizar esta herramienta se deben seguir una serie de instrucciones como lo son: obtención, limpieza de datos, comprensión del problema a resolver, y determinación, disponer únicamente de la información necesaria para el proyecto, crear un modelo matemático sujeto a las especificaciones del problema y finalmente la validación de los resultados obtenidos.

La minería de datos, posee diferentes técnicas que provienen básicamente de algoritmos, que se aplican sobre un conjunto de datos para obtener un resultado. Entre ellos se encuentran: Las redes neuronales, regresión lineal, árboles de decisión, modelos estadísticos, agrupamiento y reglas de asociación.

En este proyecto se utilizarán, la regresión lineal, que consiste en la formación de relaciones entre los datos; los árboles de decisión, para la realización del modelo predictivo con el apoyo de la "*inteligencia artificial"* y *"el análisis predictivo"* y los Modelos estadísticos, que son una expresión simbólica en forma de ecuación para indicar las diferentes restricciones que pueden modificar la respuesta.

4.2.3.2 Árboles de decisión

Un árbol de decisión es un diagrama en el cual se relacionan los posibles resultados de un conjunto de decisiones. Su principal atributo es que permite al usuario observar las variables asociadas a cada decisión: costo, beneficios, tiempo etc.

Un uso común es coordinar el intercambio de ideas así mismo, sirve para trazar un algoritmo que comunique la mejor opción del grupo de decisiones

Este diagrama está compuesto comúnmente por un único nodo, del cual se relacionan los resultados posibles, de estos resultados emergen otros, este proceso se repite hasta que todos los resultados posibles son relacionados.

4.2.3.3 Random Forest para algoritmos de predicción

Los algoritmos del árbol de decisión consisten en dividir los datos en clasificaciones para luego tener un algoritmo que prediga dónde aterrizarán los nuevos puntos de datos. Esas clasificaciones se basan en valores de las variables independientes y dependientes. [21]

La técnica de algoritmos Random Forest consiste en tener múltiples árboles, un bosque de árboles. Todos esos árboles pueden ser del mismo tipo o algoritmo o el bosque puede estar formado por una mezcla de tipos de árboles (algoritmos). [21]

En esta técnica la relación causa-efecto es la que prima en la predicción algorítmica, observe el siguiente gráfico en donde se especifica el uso de este método en un algoritmo de predicción:

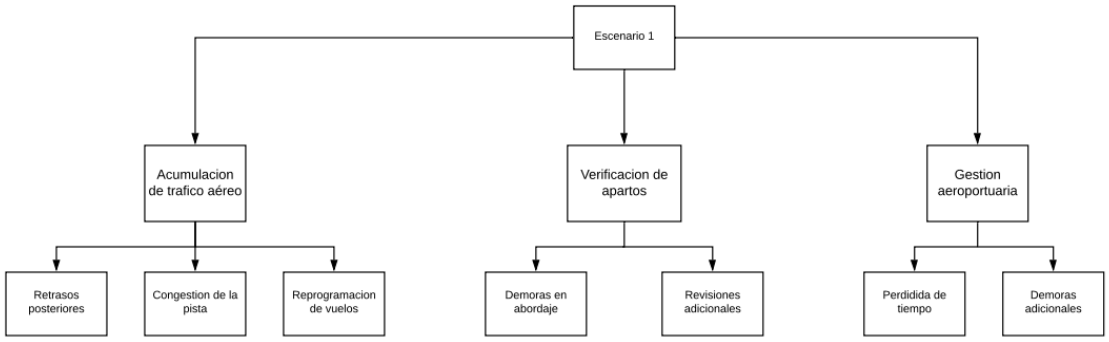

Ilustración 2 Ejemplo de Random Forest fuente: Propia

El algoritmo a partir del escenario insertado comienza a comparar con los arboles los cuales son el conjunto de causa efecto, y de acuerdo a las coincidencias encontradas presenta un escenario de efecto el cual se puede cuantificar con el apoyo de la regresión lineal.

4.2.3.4 Redes neuronales:

Es una técnica de inteligencia Artificial que es posiblemente la familia más popular de algoritmos predictivos, en realidad como modelo computacional existen desde mediados del siglo pasado pero no ha sido hasta hace unos años con la mejora de la técnica y la tecnología que se comenzó a utilizarlas notablemente con capacidades en campos como: reconocimiento de caracteres, imágenes y voz; predicción bursátil, generación de texto, traducción de idiomas, prevención de

fraudes, conducción Autónoma, análisis genético pronóstico de enfermedades entre otras, se trata de una familia de algoritmos muy potentes en la que podemos modelar comportamientos inteligentes, en la mayoría de comportamientos y estructuras avanzadas, la complejidad de estos sistemas emerge de la interacción de muchas partes, más simples trabajando conjuntamente. En el caso de una red neuronal a cada una de estas partes se le denomina neuronas.

4.2.3.4.1 Neurona:

Es la unidad básica de procesamiento dentro de una red neuronal similar a una neurona biológica, estas neuronas tienen conexiones entrada a través de los que reciben estímulos externos, los valores de entrada con estos valores la neurona realiza un cálculo interno y generará un valor de salida, realmente Una neurona no deja de ser una función matemática.

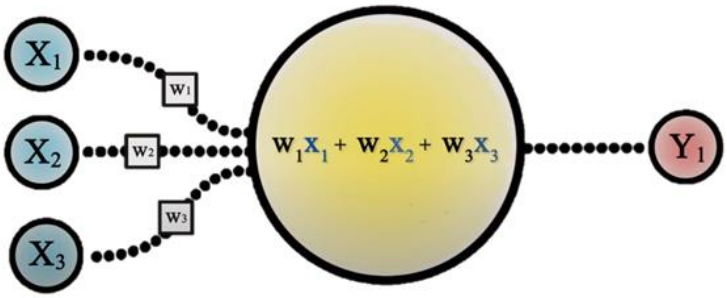

Ilustración 3 estructura de una neurona fuente: CSV

En este cálculo numérico la neurona utiliza todo los valores de entrada para realizar una suma ponderada, en donde cada una de las conexiones entrada se le asigna un valor de acuerdo al peso que tiene esta en el modelo al final se tiene un valor que servirá para definir como cada variable de entrada afecta la neurona internamente esto se suele representar como palancas que podemos subir o bajar para modificar positiva o negativamente el valor de la suma estos pesos son los parámetros del modelo y serán los valores que podremos ajustar para la red neuronal pueda procesar información. [22]

Este proceso matemático es muy parecido a lo que se utiliza en el modelo de regresión lineal, es decir que lo que hace una neurona internamente es una regresión lineal, donde tenemos una variable de entrada que definen una recta a un hiperplano, a la que se puede variar la inclinación utilizando nuestros parámetros, en regresión lineal se tiene el término independiente que sirve para mover verticalmente a la recta, en la neurona también se tiene ese mismo terminó, este valor se le denomina sesgo o Bayas en inglés y se representa básicamente como otra conexión a la neurona pero en el que la variable siempre está asignada a uno y que podemos controlar manipulando el valor del parámetro, sin embargo la diferencia vendrá con un último componente llamado la función de activación. [23]

Este modelo de una única neurona, sirve para problemas de compuertas lógicas AND y OR sin embargo en caso de una compuerta XOR, es imposible resolver es por ello que para solucionar este problema se debe codificar una segunda neurona conformando un sistema más complejo en la medida que esté compuesto por más neuronas, hay dos formas diferentes de organizar a estas neuronas una manera sería colocarlas en la misma columna llamado de forma correcta en la misma capa. Las dos neuronas que se encuentran en la misma capa recibirán la misma información de entrada de la capa anterior y los cálculos que realizan los pasarán a la siguiente capa. La segunda forma es organizarlas de forma secuencial en donde la segunda recibe la información procesada por la primera. [24]

Sin embargo, para construir una red neuronal se crea una combinación de ambas estructuras donde, la primera capa donde está la variable de entrada se le denomina capa de entrada, la última capa de salida y a las capas intermedias ocultas.

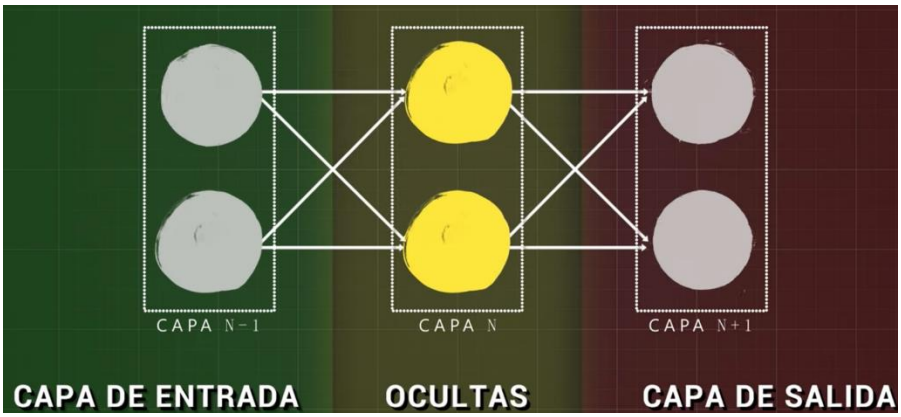

Ilustración 4 Capas de redes neuronales Fuente: CSV

Con esta organización lo que se consigue es que la red tenga algo que se conoce como conocimiento jerarquizado, por medio de este se pueden establecer modelos muchos más complejos en donde la composición de varias decisiones lleva a la red a una conclusión que depende de dos orígenes distintos un ejemplo es la nota que alguien Saca en un examen se puede establecer variables de entrada que son motivación por la asignatura y dificultad del examen esta es la arquitectura de esta red ahora de forma jerarquizada la red neuronal podría aprender conocimientos básicos las primeras capas Como por ejemplo la primera neurona se especialice en si el estudiante está entretenido en la noche y otra neurona que se especializa saber cuál es la motivación para el examen. El conocimiento elaborado en esta capa será procesado nuevamente por las siguientes capas elaborando cada vez conocimiento más complejo abstracto e interesante, un resultado es que la primera neurona podría descubrir si la motivación exámenes baja y la noche del viernes es entretenida posiblemente vaya a estudiar poco y por ende es posible que pierda el examen.

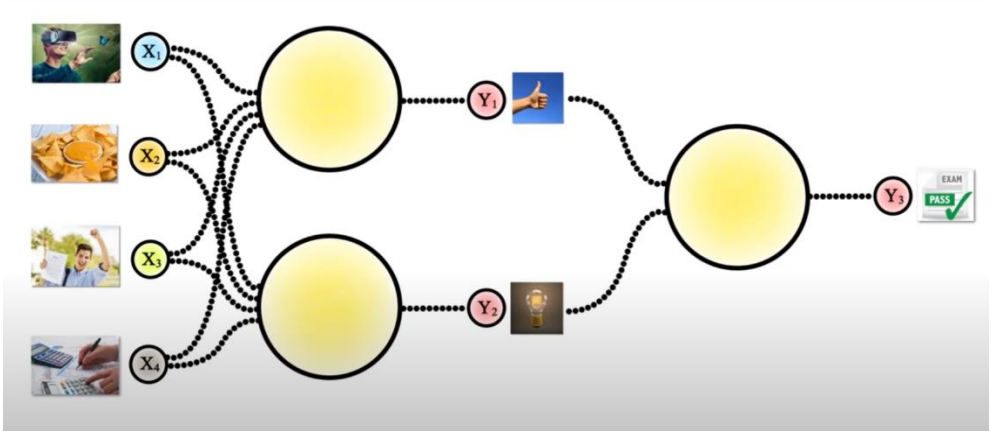

Ilustración 5 Estructura red neuronal Fuente:CSV

Entre más capas se añade más complejo puede ser el conocimiento que se elabora esta profundidad, y depende de la cantidad de capas, es lo que da nombre al aprendizaje profundo el Deep learning, para alcanzar este aprendizaje profundo básicamente lo que se quiere es conectar múltiples neuronas de forma secuencial.

Lo que se hace en cada una de estas neuronas es un problema de regresión lineal es decir planteado matemáticamente es concatenar diferentes operaciones de regresión lineal.

El problema es que matemáticamente se puede comprobar que el efecto de sumar muchas operaciones de regresión lineal equivale solamente una operación y resultado es otra línea recta o visto otra manera tal y como está planteada la red de momento es equivalente a tener una única neurona.

Para conseguir una red útil se necesita un resultado diferente a una línea recta y para eso se necesita que cada una de estas líneas se convierta en una simulación no lineal, es decir que las distorsione y para conseguirlo se debe usar una función de activación.

4.2.3.4.2 Función de activación

La función de activación es la última componente que faltó ver la estructura de la neurona, básicamente si en la neurona lo que se hace es: calcular como valor de salida una suma ponderada de las entradas y posteriormente se debe pasar dicho valor de salida por una función de activación, en donde el valor de salida será distorsionado añadiendo deformaciones no lineales para que así sea posible encadenar de forma efectiva la computación de varias neuronas, existen varias funciones de activación, una es la función sigmoide que sirve para representar probabilidades, también la función tangente hiperbólica que es similar a la sigmoide, por último y la que se utilizara en este proyecto es la función escalonada.

4.3 Recursos informáticos y software.

En la actualidad para encontrar relación entre datos, considerando la magnitud que estos poseen, se convierte en una tarea compleja y casi imposible realizar es por esto que, para realizar esta tarea de una forma más eficiente, existen diferentes recursos informáticos entre ellos se destacan Python y R.

4.3.1 Python y R lenguaje de programación.

Python es un lenguaje y orientación a objetos, lo que clasifica en un lenguaje multi paradigma con programación imperativa pero con menor grado en programación funcional [25]. Que siempre está en constante mejora gracias a sus colaboradores por medio de nuevas funciones.

Enfocado a la minería de datos, Python cuenta con varias librerías capaces de realizar cálculos matemáticos y estadísticos, algoritmos de aprendizaje máquina y visualización de gráficas de datos, por lo cual, es una herramienta que ayuda al análisis de datos gracias a su fácil manejo y lenguaje. [26]

R es un lenguaje de código abierto, con un enfoque orientado al análisis estadístico, presenta múltiples paradigmas y al igual que Python siempre está en constante mejora gracias a su comunidad. [24]

Para la minería de datos, sin embargo, su característica vectorial y multiplataforma, le permite a R poseer opciones especiales para el manejo de elementos estadísticos como lo son, las operaciones con matrices y vectores. R tiene la ventaja de poder manipular los datos de forma rápida, también fue diseñado para la analítica de datos teniendo como pilares la precisión y exactitud. [26].

4.3.2 Precedentes Académicos

Este tipo de investigación a pesar de contar con pocas referencias de estudio en Latinoamérica, cuenta con un amplio desarrollo en países norteamericanos y asiáticos.

La institución hispanohablante que más se ha destacado, en esta área de estudio es la Universidad Autónoma de Barcelona, que a lo largo los años sus estudiantes han aportado casos prácticos e implementado estrategias de optimización aeroportuaria, el aporte más destacado para la presente investigación es *“Predicción y Análisis de los Retrasos en los Vuelos”* por Nerea Martínez Domenech, por el uso de la tecnología de machine learning por medio de Python y el uso de modelos random forest en su algoritmo de base.

Entre otros autores relevantes en el área de la aeronáutica, se destacan a investigadores como Bo Zou, de la universidad de Illinois Chicago y el ingeniero Díaz Olariaga, de la universidad Santo Tomas, que durante la última década han aportado más de 20 investigaciones en el aérea de aeronáutica civil, infraestructura

aeroportuaria, impacto económico en zonas de desarrollo aéreo y creado informes de los constantes cambios de la aeronáutica en EEUU y Colombia.

4.4 Marco Legal

4.4.1 Código Civil, Artículo 1614. Daño emergente y lucro cesante

Según lo establecido en el Código Civil el cual establece el daño emergente como la reparación económica, causada por el incumplimiento de una obligación contraída o el incumplimiento parcial, así mismo el retardo de la obligación.

4.4.2 Codigo de comercio decreto 410 DE 1971.

4.4.2.1 Artículo 1870. Venta de cosa inexistente

Este articulo hace referencia a la transacción de un bien o servicio que no existe. Así mismo se refiriere a la falta de una parte considerable del bien o servicio vendido, y estipula que, en este caso, el comprador podrá desistir del contrato, o darlo por subsistente, y dejando a un lado en precio de justa transacción. El vendedor deberá subsanar los servicios no prestados y resarcirá los perjuicios al comprador de buena fe.

4.4.2.2 Código de Comercio Artículo 981. Contrato de transporte

EL contrato de transporte se pacta entre dos partes, esto a cambio de un precio considerado justo por ambas partes, el cual consiste transportar de un lugar a otro, por medio pactado y en un plazo acordado, este contrato aplica a personas, objetos o animales; El contrato se concreta entre el prestamista y el transportado únicamente y está regulado por las leyes de transporte, en caso que el contrato se viole, y solo por petición de una de las partes un juez podrá intervenir con el fin de establecer las acciones consiguientes.

5. MARCO METODOLÓGICO

5.1 Tipo de Investigación

El presente proyecto responde a un tipo de investigación explicativa, ya que, por medio del análisis del entorno y el rol de las variables en el mismo, se pretende explicar la influencia de las variables en los retrasos aeroportuarios, esta investigación presenta datos cuantitativos y cualitativos, por ende, corresponde a una investigación mixta.

5.1.1 Hipótesis

Se presume que el modelo de Random Forest es el más adecuado para la predicción de retrasos aéreos en despegue dada su efectividad en la clasificación y poca necesidad de suposiciones, así mismo la técnica puede manejar hasta miles de variables de entrada e identificar las más significativas. (Cali Barranquilla y Cartagena)

5.2 Diseño de Investigación

Este proyecto presenta un diseño No experimental, ya que solo se pretende observar las variables en su entorno y analizar su comportamiento en estado natural; así mismo el tipo de diseño es transversal ya que se recolectan los datos en un tiempo específico y el objetivo de estas mediciones es el análisis de los mismos y su influencia en el evento de retraso aéreo.

5.3 Definición de Población y Muestra

Universo

Aeropuertos en Colombia.

Muestra

Tres aeropuertos de mediana complejidad en Colombia, (Barranquilla, Cali y Cartagena).

5.4 Definición de variables

- Tráfico
- Aerolínea
- Origen destino
- Hora salida
- Hora programada de salida
- Hora programada de remolque
- Hora de remolque
- Demora
- Código de demora IATA (Anexo 1)
- Estado de vuelo
- Motivo de demora.

5.5 Herramientas de recolección y análisis de la información

La información, fue recolectada por la AEROCIVIL institución encargada del monitoreo aéreo en Colombia, esta información es captada en cada vuelo y es parametrizada diariamente en una base de datos conjunta con el fin de ser presentada en un reporte anual.

La información fue otorgada con el fin de ser el pilar de esta investigación y fue resultado de la colaboración entre el ingeniero Oscar Eduardo Díaz Olariaga y el docente de la Facultad de ingeniería industrial de la Universidad Santo Tomas el ingeniero Luis Manuel Pulido Rico.

Para el análisis y procesamiento de la información se utilizará una adaptación propia de la metodología CRISP-DM de IBM [27], se trata de un modelo estándar abierto del proceso que describe los enfoques comunes, que utilizan los expertos en minería de datos. Esta metodología se describe en términos de un modelo de procesos jerárquico, que consta de conjuntos de tareas descritas en cuatro niveles de abstracción (de lo general a lo especifico).

Las herramientas para el análisis de la información son pruebas de bondad de ajuste de distribución (Chi Cuadrada, Anderson Darling, Komogorov Sminoff), que serán aplicadas por medio de los Softwares estadísticos como IBM SPS-S®, paquetes de análisis descriptivo de R-Studio y herramientas de análisis de Excel®.

Como herramientas de apoyo, se utilizará Power BI® un software de escritorio y cloud que facilita el análisis de información de diferentes fuentes ya sean de bases de datos con millones de registros o pequeñas hojas de cálculo, así mismo, proporciona visualizaciones interactivas que permiten una mejor comprensión de la información presentada, en un dashboard compuesto por herramientas visuales. [28].

5.6 Pasos de la metodología CRISP-DM

5.6.1 Comprensión de los datos

Esta fase inicia con la recopilación de los datos y a continuación se aplican actividades que permitan la familiarización de los datos, identificación de los problemas de calidad, descubrimiento de los datos superficiales y la detección subconjuntos de información relevante. [27]

5.6.2 Preparación de datos

Esta fase comprende las actividades correspondientes a la construcción de un conjunto útil de datos, los cuales deben cumplir con los requerimientos de las herramientas de modelado. Estas tareas en la mayoría de casos, se deben repetir constantemente hasta obtener un conjunto de datos útil, no tiene ningún orden especifico y el analista aplica este proceso según su criterio, las tareas descritas

anteriormente son la selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para herramientas de modelado. [27]

5.6.3 Modelado

Este proceso comprende la selección y aplicación de técnicas de modelado, y sus parámetros se calibran a valores óptimos. Normalmente, existen varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, a menudo es necesario volver a la fase de preparación de datos. [27]

5.6.4 Evaluación del modelo

En esta etapa del proyecto, se ha creado un modelo (o modelos) que parece tener alta calidad desde la perspectiva del análisis de datos. Antes de proceder a la implementación final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, para asegurarse de que el modelo logre adecuadamente los objetivos. Un objetivo clave es determinar, si existe alguna variable importante que, no se haya tenido en cuenta de manera adecuada. Al final de esta fase, se debe tomar una decisión sobre el uso de los resultados de la minería de datos. [27].

5.6.5 Flujograma de metodología del proyecto

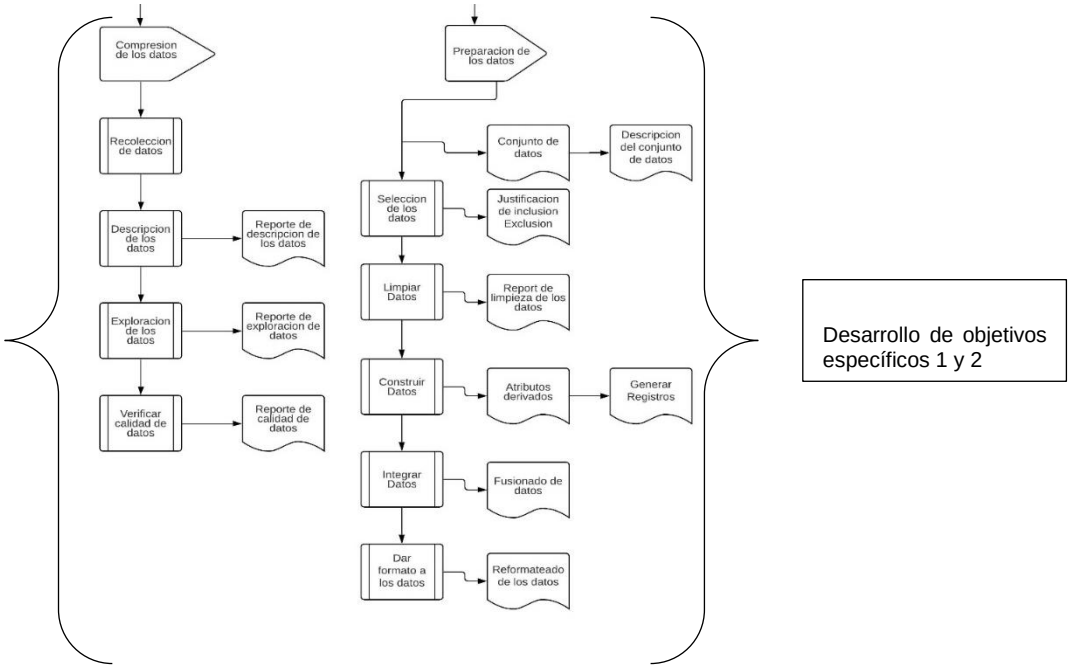

Diagrama 1 diagrama de flujo Metodología Crisp-DM parte 1 Fuente: IBM

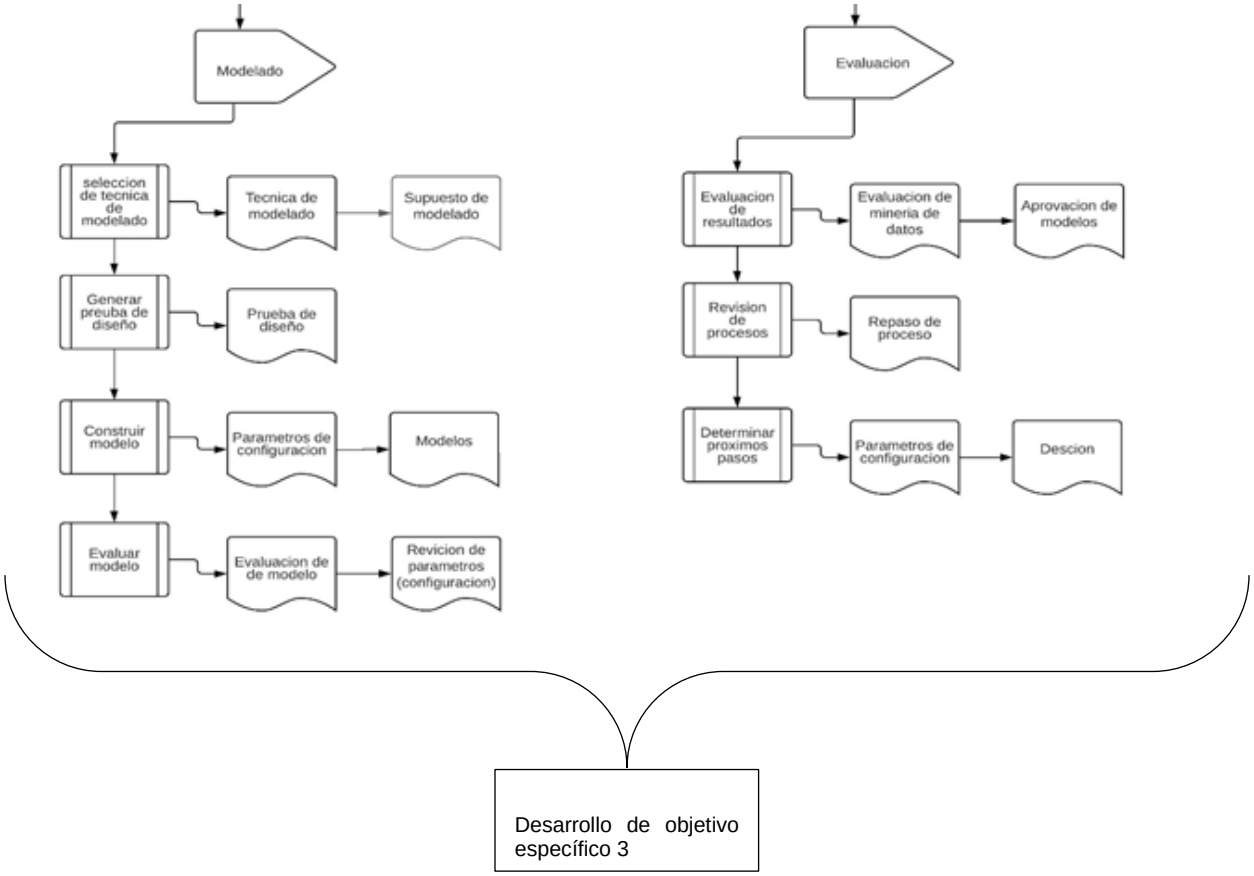

Diagrama 2 diagrama de flujo Metodología Crisp-DM parte 2 Fuente: IBM

6. RESULTADOS Y ANALISIS DE RESULTADOS

6.1 Comprensión de los datos

Como paso inicial, se procederá con la familiarización de los datos y un pre-procesamiento de la información con el fin de identificar datos relevantes para el modelo predictivo, el significado de los registritos y su interpretación, posibles correlaciones e identificar los registros vacíos dentro de las columnas, entre otras incógnitas.

6.1.1 Recolección de datos

La base de datos a trabajar es un registro elaborado por la Aerocivil, día a día del año 2018, en el cuál se registraron todos los vuelos realizados tanto nacionales como internacionales en el año anteriormente mencionado. Se cuenta con un total de 357.596 filas, 16 columnas, y 5´721.536 registros.

6.1.2 Descripción de los datos

Esta base de datos cuenta con la información de vuelos correspondiente a 54 diferentes aeropuertos, entre los cuales se encuentran los 40 aeropuertos

nacionales comerciales (con más de 20.000 pasajeros al año) y los 14 aeropuertos con capacidad internacional.

La data cuenta con 16 diferentes tipos de datos, los cuales estarán descritos a continuación con una breve descripción de cada una:

Motivo	Descripción	Motivo	Descripción
Tráfico	Es el tipo de vuelo y se divide en Nacional e Internacional.	**Hora de remolque UTC**	Hora de remolque programada en el sistema.
Aerolínea	Nombre de todas las aerolíneas que presentaron flujo aéreo en el país.	**Demora AEROCIVIL**	Diferencia entre la hora programada de salida y la hora de remolque.
Origen	Campo que muestra el código IATA del aeropuerto de origen.	**Demora (hh:mm)**	Diferencia entre la hora programada de salida y la hora de remolque.
Destino	Campo que muestra el código IATA del aeropuerto de destino.	**Estado del vuelo**	Indicador de si el vuelo cumplió, tuvo retrasos, se canceló y demás causas.
N° Vuelo	Número registrado de vuelo.	**Código de demora**	Código designado por la Aerocivil a las diferentes causas de retraso.
Fecha programada de salida Referencia SCORE	Fecha de salida registrada en el sistema.	**Motivo de la demora**	Cuando existe demora, este se clasifica en interno o externo dada la causa.
Hora programada de salida Referencia Score	Hora de salida registrada en el sistema.	**Observaciones**	Información adicional de lo que sucedió en el vuelo.
Fecha de remolque UTC	Fecha de remolque registrada en el sistema.	**Estatus AEROCIVIL**	Indicador de si el vuelo cumplió, tuvo retrasos, se canceló y demás causas.

Tabla 1 Variables de la base de datos Fuente: propia

El universo a estudiar son los aeropuertos en Colombia, los cuales en el año 2018 presentaron 231,300 vuelos Cumplidos, 9,150 vuelos adelantados, 20.195 vuelos cancelados, 95.858 demorados y 32 sin estatus. Los aeropuertos en Colombia son:

Aeropuertos Internacionales		
Nombre	**COD**	**Ciudad**
Aeropuerto Internacional El Edén	AXM	Quindío
Aeropuerto Internacional Ernesto Cortissoz	BAQ	Atlántico
Aeropuerto Internacional El Dorado	BOG	Distrito Capital
Aeropuerto Internacional Palonegro	BGA	Santander
Aeropuerto Internacional Alfonso Bonilla Aragón	CLO	Valle del Cauca
Aeropuerto Internacional Camilo Daza	CUC	Norte de Santander
Aeropuerto Internacional Rafael Núñez	CTG	Bolívar
Aeropuerto Internacional Alfredo Vásquez Cobo	LET	Amazonas
Aeropuerto Internacional José María Córdova	MDE	Antioquia
Aeropuerto Internacional Matecaña	PEI	Risaralda
Aeropuerto Internacional Gustavo Rojas Pinilla (3)	ADZ	San Andrés y Providencia
Aeropuerto Internacional Simón Bolívar	SMR	Magdalena
Aeropuerto Internacional Los Garzones	MTR	Córdoba
Aeropuerto Internacional Almirante Padilla (1)	RCH	La Guajira

Tabla 2 aeropuertos internacionales en Colombia *Fuente: propia*

Aeropuertos Nacionales Comerciales					
NOMBRE	**COD**	**CIUDAD**	**NOMBRE**	**COD**	**CIUDAD**
Aeropuerto Antonio Roldán Betancourt	APO	Antioquia	Aeropuerto Benito Salas	NVA	Huila
Aeropuerto Santiago Pérez Quiroz	AUC	Arauca	Aeropuerto Reyes Murillo	NQU	Chocó
Aeropuerto José Celestino Mutis	BSC	Chocó	Aeropuerto Antonio Nariño	PSO	Nariño
Aeropuerto Yariguíes	EJA	Santander	Aeropuerto Contador	PTX	Huila
Aeropuerto Guaymaral	GYM	Distrito Capital	Aeropuerto Guillermo León Valencia	PPN	Cauca
Aeropuerto Juan H. White	CAQ	Antioquia	Aeropuerto El Embrujo	PVA	San Andrés
Aeropuerto Gustavo Artunduaga	FLA	Caquetá	Aeropuerto Tres de Mayo	PUU	Putumayo
Aeropuerto Juan Casiano	GPI	Cauca	Aeropuerto Morelia	PGT	Meta
Aeropuerto Perales	IBE	Tolima	Aeropuerto César Gaviria Trujillo	PDA	Guainía
Aeropuerto San Luis	IPI	Nariño	Aeropuerto Caucaya	LQM	Putumayo
Aeropuerto Javier Noreña Valencia	LMC	Meta	Aeropuerto Germán Olano	PCR	Vichada
Aeropuerto La Nubia	MZL	Caldas	Aeropuerto El Caraño	UIB	Chocó
Aeropuerto Enrique Olaya Herrera	EOH	Antioquia	Aeropuerto Jorge Enrique González	SJE	Guaviare
Aeropuerto de Villagarzón	VGZ	Putumayo	Aeropuerto Los Colonizadores	RVE	Arauca
Aeropuerto Jorge Isaacs	LMN	La Guajira	Aeropuerto Las Brujas	CZU	Sucre
Aeropuerto Fabio Alberto León Bentley	MVP	Vaupés	Aeropuerto La Florida	TCO	Nariño
Aeropuerto Alfonso López Pumarejo	VUP	Cesar	Aeropuerto Puerto Bolívar	URA	La Guajira
Aeropuerto El Alcaraván	EYP	Casanare	Aeropuerto Vanguardia	VVC	Meta

Tabla 3 Aeropuertos nacionales en Colombia Fuente: propia

Las empresas autorizadas para el transporte aéreo en Colombia son:

Aerolíneas en Colombia	
ADA	INTERJET
AEROGAL	JETBLUE
AEROLINEAS ARGENTINAS	KLM
AEROMEXICO	LACSA
AEROREPUBLICA SA	LAN AIRLINES SA
AIR CANADA	LAN Perú SA
AIR EUROPA	LUFTHANSA
AIR FRANCE	ONE
AIR PANAMA	SATENA
AIRES - AEROVIAS DE INTEGRACION REGIONAL	SPIRIT
AMERICAN	TACA INTERNACIONAL
AVIANCA	TACA PERU
AVIOR AIRLINES C.A	TAM LINHAS AEREAS SA
COPA AIRLINES	TAME
CUBANA	TURKISH
DELTA	UNITED
EASYFLY	VIVAAIR COLOMBIA
IBERIA	

Tabla 4 Aerolíneas en Colombia Fuente: propia

6.1.3 Exploración de los datos

Las 35 aerolíneas autorizadas en Colombia realizaron 357.596 vuelos en el año 2018, estos vuelos presentaron un 27% de demora global, así mismo el 65% de los vuelos estuvieron a tiempo solo el 6% fueron cancelados y el 3% adelantados; el 0.01% de la información es obsoleta por errores de transcripción u omisión en la data.

Las aerolíneas que más vuelos presentaron durante el periodo estudiado, son **AVIANCA** con un total del 44.43% de los vuelos, le sigue **Aires** la cual cuenta con un 14.21% del total general, **EASYFLY** con una participación de 11.84%, **SATENA** con un 7.23% y por ultimo **VIVA AIR** con una cuota aérea de 6.69%. A continuación, se observa el comportamiento del status según cada una de las aerolíneas, entre las más cumplidas en relación al número de vuelos, así mismo las aerolíneas más demoradas y las que presentan mayor tasa de cancelación de vuelos durante el año 2018.

Las aerolíneas más cumplidas son American Airlines, Delta y Aerorepublica, con una tasa superior al 85%, de igual forma se observa que las aerolíneas con más retraso son Satena, Aerogal y Easyfly; las aerolíneas con mayor tasa de cancelación son Tame, Air Canadá y Taca Perú.

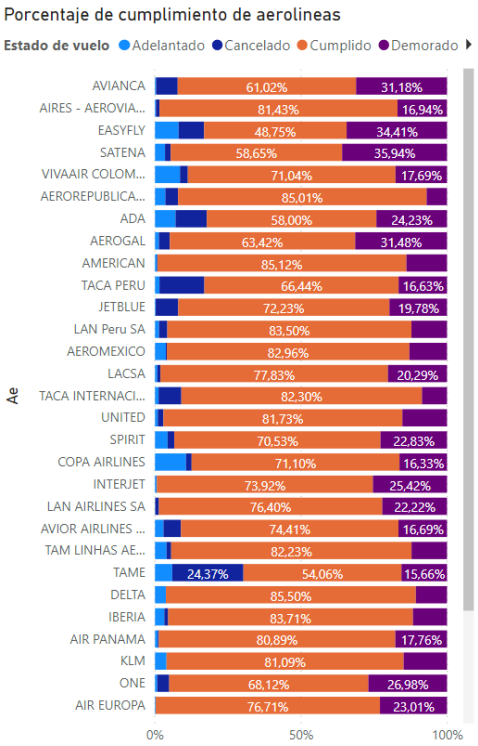

Gráfico 1 Porcentaje de cumplimiento de aerolíneas en Colombia Fuente: propia
Para ver a mayor detalle ir a anexo 2

Posterior al análisis de las aerolíneas se evalúa el comportamiento de los orígenes y destinos más frecuentes en Colombia en materia de transporte aéreo. Como se puede observar en el grafico Bogotá por su naturaleza es el origen del cual despegan mas vuelos tanto a nivel intenacional como nacional, esto es acorde a su reputacion ya que en el actual 2020 ocupa el tercer lugar en el ranking de terminales aéreas más grandes de latino américa. [29] Para ver a mayor detalle ir a anexo 2

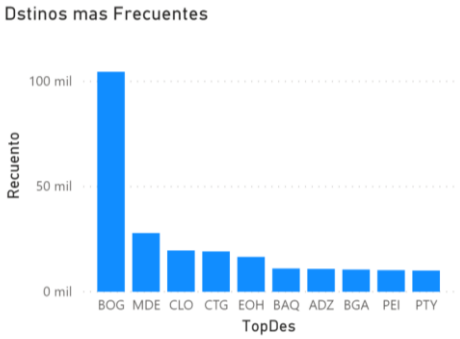

Gráfico 2 Ciudades con más retraso como Destino Fuente: propia

En contraste, los destinos más transitados en 2018 son Bogotá, Medellín, Cali, Cartagena y Barranquilla. Es destacable que el único destino internacional dentro de los 10 con mayor demanda es el aeropuerto de Ciudad de Panamá.

En razón del objetivo del proyecto, en donde el comportamiento de las demoras es el dato más relevante, a continuación, este análisis se enfoca en las demoras presentadas durante el año 2018. Para ver a mayor detalle ir a anexo 2

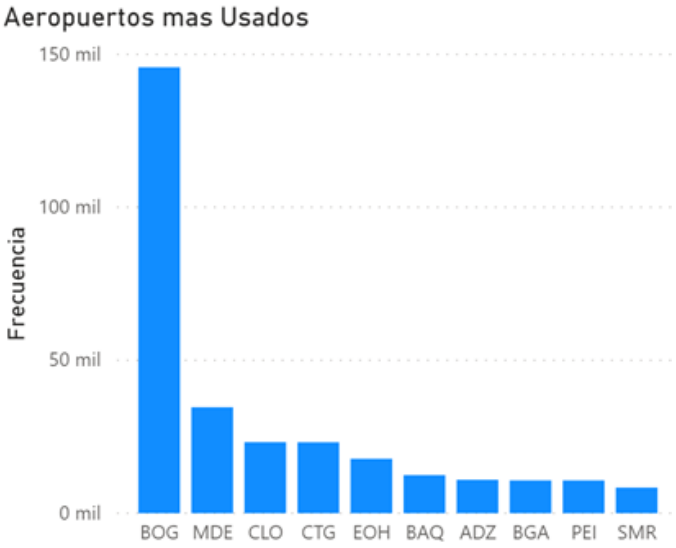

Gráfico 3 Orígenes con más retraso Fuente: propia

En este TOP 10 de aeropuertos se puede observar que destacan 5 ciudades: Bogotá, Medellín, Cali, Cartagena y Barranquilla (Medellín se encuentra dos veces en el TOP). Para elaborar un análisis de cómo se comporta el tráfico aéreo en Colombia, como se había establecido anteriormente en los objetivos, en este proyecto se analizarán los aeropuertos de Cali, Cartagena y Barranquilla. Para ver a mayor detalle ir a anexo 2

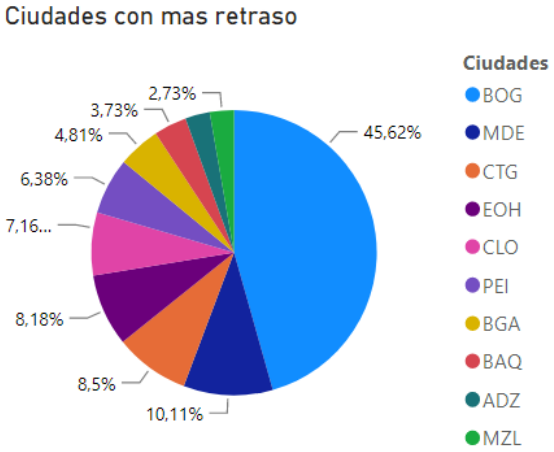

Gráfico 4 Ciudades con más retraso porcentual en Colombia Fuente: propia

Teniendo en cuenta el anterior análisis, se procede con la primera delimitación de la base de datos, que consiste en manejar únicamente vuelos que tengan su origen en Cartagena (CTG), Cali (CLO) y Barranquilla (BAQ) y no aquellos que tengan su destino en las ciudades anteriormente mencionadas, ya que el motivo del proyecto solo toma en cuenta los retrasos por despegue y este por su naturaleza genera un efecto en cadena, se sabe que una vez tenga un retraso en la salida, la llegada también tendrá un retraso de igual proporción,.

Para el análisis de las ciudades de Barranquilla, Cali y Cartagena. Se detallará el número de vuelos por aeropuerto, compañías que operan en las terminales aéreas de las ciudades anteriormente mencionadas y por último los 5 códigos de demora más frecuentes en cada aeropuerto. En el año 2018 se registraron 14.407 vuelos demorados en Barranquilla, Cali y Cartagena, presentando 3.229, 5.786 y 5.392 respectivamente.

Las aerolíneas presentes en las tres ciudades mencionadas son:

Aerolíneas que operan en CTG CLO y BAQ	
SATENA	SPIRIT
AEROGAL	TACA INTERNACIONAL
AIR CANADA	TACA PERU
AIR PANAMA	TAME
AMERICAN	AEROREPUBLICA SA
COPA AIRLINES	AIRES
DELTA	AVIANCA
JETBLUE	EASYFLY
KLM	VIVAAIR COLOMBIA
LAN AIRLINES SA	ADA
LAN Perú SA	AVIOR AIRLINES C.A

Tabla 5 Aerolíneas que operan CTG CLO y BAQ Fuente: propia

De las cuales, se destacan las aerolíneas nacionales como Avianca, Aires (Actualmente miembro de LATAM), Easyfly, VivaAir y Satena, con mayor presencia en estas ciudades representando el 90 % del flujo aéreo.

En el siguiente diagrama de Pareto, se observa las principales causas de demora en las tres ciudades. Cabe mencionar que el 80% de las causas están conformadas por el 23% de los códigos.

El análisis Pareto es una técnica cualitativa que estima los aspectos más relevantes en un conjunto de datos, así mismo se basa en la premisa 80/20 (el 80% de los retrasos son generados por el 20% de las causas).

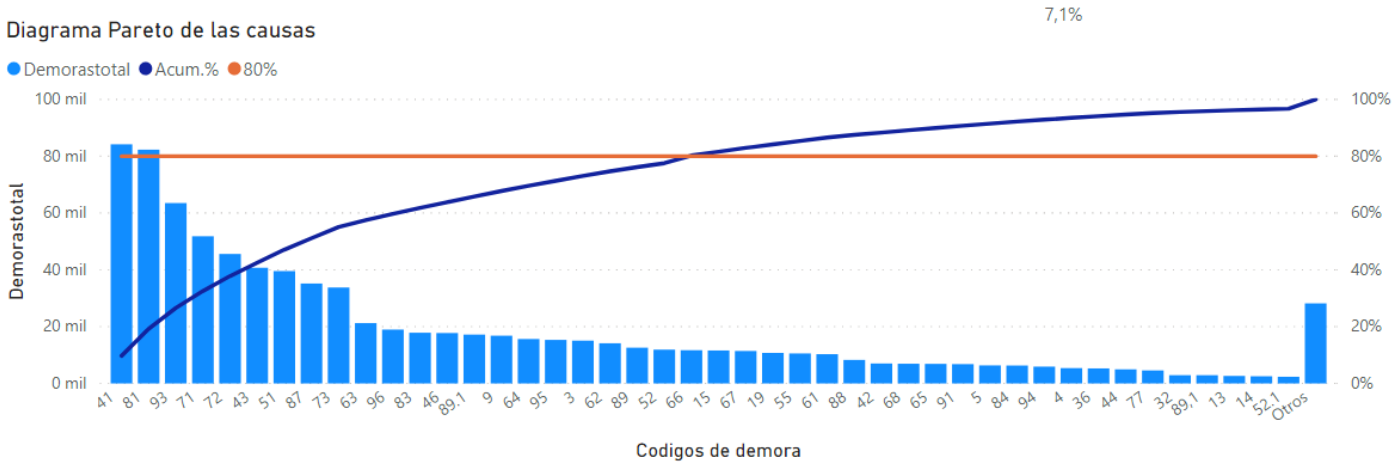

Gráfico 5 Diagrama Pareto de códigos de demora Fuente: propia

Los 5 códigos más frecuentes son por causas externas e internas y con algunas incontrolables como lo son las climáticas. A continuación, una breve descripción de cada una:

Código	Nombre	Descripción
41	Defectos del avión	Falla del avión producida en forma repentina (es decir sin antecedentes en días anteriores) que se presente durante la preparación del mismo o su operación.
81	ATFM (Air Traffic Flow Management)	Mayor tiempo de vuelo que el previsto en el itinerario, durante el trayecto previo, ocasionado por los Servicios de Tránsito Aéreo, lo que causa la llegada tarde del avión.
93	Rotación de naves	Llegada tarde del avión en su vuelo previo. Cambio de avión asignado al vuelo, por necesidades operacionales.
71	Aeropuerto de salida	Condiciones climáticas adversas en el aeropuerto de origen.
72	Aeropuerto de destino	Condiciones climáticas adversas en el aeropuerto de destino.

Tabla 6 Códigos de demora más frecuentes en Colombia Fuente: propia

En las ciudades de Barranquilla, Cali y Cartagena, se identificó que a nivel general los vuelos presentan mayores causas de retraso externas. Las cuales representaron 400,739.2 minutos de retraso.

Las cinco causas de retraso más relevantes en Barranquilla son: atm, aeropuerto alterno de destino, aeropuerto de salida. instalaciones aeroportuarias y posiciones de estacionamiento

Cinco causas más importantes en BAQ			
81	AFTM (air traffic flow management)	RAC	EXTERNO
73	Aeropuerto alterno de destino o en ruta.	INCONTROLABLES	EXTERNO
71	Aeropuerto de salida.	INCONTROLABLES	EXTERNO
87	Instalaciones aeroportuarias,	AGA	EXTERNO
41	Posiciones de estacionamiento de aviones, congestión en rampa	TECNICAS	INTERNO

Tabla 7 Causas más frecuentes en barranquilla Fuente: propia

El 20% de las 5 causas más frecuentes son causas externas, así mismo se identificó que el 26% de los retrasos son producidos por las 5 causas más relevantes. El 40% son causas incontrolables por condiciones climáticas o de visibilidad aérea, las 5 principales causas en barranquilla representan el 5% de los retrasos totales a nivel general

Cinco causas más importantes en CLO			
41	Defectos Del Avión	TÉCNICAS	INTERNO
81	ATFM (air traffic flow management)	RAC	EXTERNO
87	Instalaciones aeroportuarias,	AGA	EXTERNO
93	Rotación de aeronaves,	OPERACIONALES	INTERNO
15	Abordaje,	OPERACIONALES	INTERNO

Tabla 8 causas más frecuentes en Cali Fuente: propia

El análisis Pareto de Cali presenta un comportamiento más disperso sin embargo se repiten 3 causas de Barranquilla en las 5 primeras causas de Cali, el código 81

el 41 y el 87; el 40% de los 5 retrasos más relevantes son causas externas así mismo el 80% corresponden a errores humanos solo el 9% corresponde a problemas climáticos, también se identificó que los 5.738 retrasos generaron 73.719 minutos de retrasos a nivel general.

En la ciudad de Cartagena se observa al igual que en Cali y Barranquilla existe una predominancia de *"Air Traffic Flow Managment*" referente a la administración del espacio aéreo a cargo de la Aerocivil, ente independiente del aeropuerto. Esta causa representa el 13% de los retrasos en el aeropuerto de Cartagena y representa un retraso en tiempo de 71.808 minutos de retraso.

81	ATFM (Air Traffic Flow Management)	RAC	EXTERNO
93	ROTACIÓN DE AERONAVES, (llegada demorada de un avión)	OPERACIONALES	INTERNO
41	Defectos del avión	TECNICAS	INTERNO
87	Instalaciones aeroportuarias	AGA	EXTERNO
15	ABORDAJE, discrepancias y compaginación, pasajeros chequeados faltantes.	OPERACIONALES	INTERNO

Tabla 9 Causas más frecuentes en Cartagena Fuente: propia

Así mismo se encontró que el 60% de las causas más frecuentes son por causas externas y el 20% por causas internas de la aerolínea, referentes a un sobrepaso en el tiempo de vuelo, daños físicos o defectos del avión producidos antes, durante o después del vuelo; los daños físicos a los aviones representaron el 40% de las causas más frecuentes lo cual se traduce a 2.211 minutos de retraso y representa el 9% de los retrasos en la ciudad de Cartagena.

Como conclusión de este análisis por ciudad, se encontró que el 37,5% de las causas más frecuentes se encuentran presentes en todas las ciudades y corresponden a los códigos 81, 41 y 87.

Entre estos tres códigos representaron el 15,62% de los retrasos en general, es decir que estos tres códigos representan un retraso de 2.343 minutos de operación en el país.

Los estadísticos descriptivos de las variables de entrada al modelo son:

Variable	N	Mínimo	Máximo	Media	Desviación estándar
Trafico: Recuento de Trafico por frecuencia total	2	813	7015	3914	4385,74
Aerolínea: Recuento de Aerolínea por frecuencia total	22	5	8752	654.86	1877.79
Origen: Recuento origen por vuelo	3	5229	5786	4802.33	1376.71
Destino: Recuento destino por vuelo	42	1	8145	351,39	95,44
Demora: tiempo de demora en Min	14,407	1	1701	47,484	92,5591
Motivo: Recuento motivo de la demora por frecuencia	5	34	7280	2881,4	3819,69

Tabla 10 Estadística descriptiva de las variables en la base de datos Fuente: propia

6.1.4 Verificar calidad de los datos

Cabe resaltar que los datos de demora presentan una distribución normal, con un comportamiento aleatorio no repetitivo, datos de naturaleza markoviana, con aparente correlación entre origen y demora, 6.571 errores de transcripción con un de 1.83% de perdida de información.

6.2 Preparación de datos

Una vez finalizada la comprensión de los datos, se inicia con la preparación de los datos para el modelo. El proceso se realiza porque algunos de los datos no son necesarios para la elaboración del modelo predictivo, porque algunos datos no se encuentran en el formato más óptimo y porque en algunos casos es necesario crear datos a partir de los ya existentes.

6.2.1 Justificación de inclusión de datos

Entre las variables a trabajar de la base de datos para el modelo predictivo hay una variable subjetiva que no aporta al aprendizaje del modelo, ya que aporta una descripción textual, esta variable es:

- Observaciones.

Las variables de fecha y hora también serán eliminadas debido a que son variables con más de una categoría (1. Fecha, 2. Hora) podrían generar una dificultad a la hora de desarrollar el modelamiento. Las variables a eliminar por esta razón son:

- Hora programada de salida Referencia SCORE.
- Hora de remolque UTC.

Otras variables a eliminar son, Fecha de remolque UTC, ya que existe otra variable con los mismos resultados (Fecha programada de salida Referencia SCORE), Demora AEROCIVIL debido a que es redundante con la variable Demora (hh:mm) y finalmente Estado del vuelo, debido a que también es repetitiva junto con Estatus AEROCIVIL y en algunos casos presentan inconsistencias.

6.2.2 Limpieza de datos

Para mejorar la calidad de los datos, es necesario modificar algunas variables. Las variables a intervenir por esta razón son:

- Tráfico: Al ser una variable con sólo dos posibles respuestas, se opta por volver está a un formato numérico, siendo nacional e internacional, 1 y 2 respectivamente.
- Motivo de la demora: Similar a la variable anterior, esta cuenta con tres posibles respuestas, se opta por volver está a un formato numérico, siendo interno, externo y no específico, 1, 2 y 0 respectivamente.
- Aerolínea: Para no descartar esta variable se decide cambiarla a numérica de 0 y 1, siendo 1 la aerolínea Avianca y 0 las demás aerolíneas.
- Origen y destino: Al igual que con aerolínea para no descartar esta variable se opta por tomar números de 0, 1, 2 y 3 siendo 0 otras ciudades, 1 Barranquilla, 2 Cali y 3 Cartagena.

6.2.3 Construir datos

Esta tarea se hace en caso de crear nuevas variables en base a las ya existentes o de transformar las mismas. En algunos casos, se crea una nueva a partir de una o más o, al contrario, se crean dos o más variables a partir de una. Las variables a intervenir por esta razón son:

- Demora (hh:mm): Para mejorar esta variable, es necesario manejar sólo una unidad de tiempo, por esta razón se crea una nueva variable que posea la demora en minutos únicamente y adicional a esto, con 15 min menos; llamada Demora- 15 min.

- Fecha programada de salida Referencia SCORE: Al ser una variable con datos de fecha para no tener inconvenientes con el modelamiento, se decide crear cuatro nuevas variables con el fin de independizar el año, el mes, el día y el número de día de la semana.

- Demora- 15 min: Al ser una variable de minuto, cuando esta sale de Excel se convierte en números muy pequeños, debido a que, al sólo ser una variable de minutos, Excel la toma con fecha completa del 1/1/1900. Entonces, se debe plantear una ecuación con el número en minutos y el valor arrojado en formato numérico con el fin de conocer cuál es el número que dividido el tiempo en formato numérico da como resultado el tiempo en minutos (Ejemplo: 8min=x/ 0,005555555. X=1440). Una vez hecho este proceso, se debe multiplicar cada tiempo en formato numérico por la “X” con el fin de tener el tiempo real en minutos en formato numérico.

- Descripción: Es una variable que se añade, que hace referencia a qué significa el código de demora de forma cualitativa.

- Diferencia demora: Es una variable numérica de 0 y 1 creada a partir del tiempo de demora de cada registro siendo 0 una demora inferior a 20 min y 1 siendo una demora superior a los 20 min.

  Finalmente, antes de dar por terminado este proceso el nombre de todas las variables debe ser cambiado ya que, en algunos casos, cuando se modela en el programa R se presentan inconsistencias de que las variables “no se encuentran” debido a los diferentes nombres de cada uno, por ende, las variables tendrán una A inicial y quedarán así:

Transito = ATRA	DIA = ADIA
Aerolínea = AALNA	DIASEM = ADSEM
Origen = AORGN	Código = COD
Destino = ADEST	Motivo = AMOT
N Vuelo = ANVLO	DemMin = ADEM
AÑO = AAÑO	DiferenciaDemora = ASUM
MES = AMES	Descripción = ADES

Tabla 11 Nuevas variables de base de datos Fuente: propia

6.2.4 Integración de los datos

Este proceso de la metodología no es necesario ya que solo hay una fuente de información, por lo tanto, no es necesario la fusión de fuentes.

6.3 Reformateado de los datos

Algunas herramientas tienen requisitos en el orden de los atributos, como que el primer campo sea un identificador único para cada registro o el último campo sea el campo de resultado que el modelo debe predecir.

La herramienta R más concretamente los paquetes de predicción (Random Forest, neural net y árboles de decisión) requiere verificar la distribución de los datos, con el fin de hallar desviaciones en la campana de distribución, y en caso de hallar desvió se debe corregir para que la predicción tenga menor porcentaje de error en sus resultados.

Para aplicar esta verificación se utilizan las Pruebas de bondad de ajuste, las tres pruebas a aplicar son:

- Chi Cuadrada (Chi Q)
- Anderson Darling (AD
- Komogorov Sminoff (KS)

Estas pruebas se aplican por medio del paquete estadístico de R estudio 'rriskDistributions', el cual es una colección de funciones para ajustar distribuciones a datos dados o por cuantiles conocidos. La función a utilizar es "fit.cont()", esta función proporciona una interfaz gráfica de usuario para elegir la distribución continua más apropiada ajustada a los datos.

Sin embargo esta interfaz no es lo que interesa del paquete, son los resultados de la prueba que permite saber cuál es la distribución que más se ajusta al comportamiento de los datos [30].

El criterio de selección de mejor distribución es el AIC también llamado criterio del menor Akaike, el cual es un criterio de calidad para modelos probabilísticos, series de tiempo y modelos de regresión lineal; la distribución que tenga el menor Akaike es la que más se ajusta a la distribución de probabilidad.

La distribución de los datos correspondientes a demora en minutos es una distribución exponencial como se observa en los resultados de las pruebas:

Distribución	Akaike	Prueba		
	AIC	Chi Q	AD	KS
Exponencial	469456.5	Inf	Inf	30
Cauchy	476877.8	27829.98	6403.58	26
Student	522266.27	106544.87	31473.94	53
Gompertz	529100.94	Inf	Inf	47
Logística	566555.13	505186269.50	Inf	32
Normal	656152.34	420548.54	Inf	38
Uniforme	NULL	Inf	Inf	41

Tabla 12 Pruebas de bondad de ajuste de variable Dem Min Fuente: propia

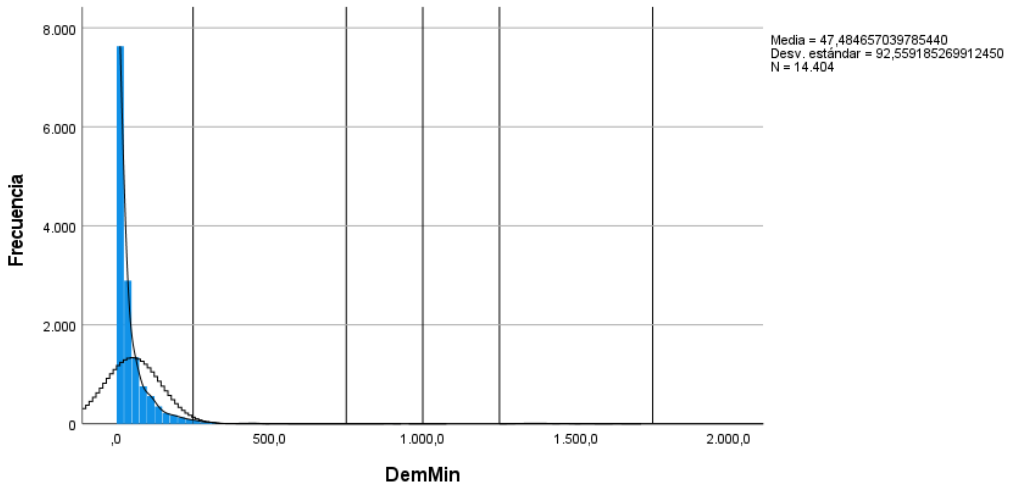

Gráfico 6 Distribución de datos de la variable Dem Min Fuente: propia

Distribución	Akaike	Prueba		
	AIC	Chi Q	AD	KS
Normal	106.42	57084.99	2775.52	0.48
Logística	401.79	624086762.20	2806.92	0.54
Exponencial	-20261.69	Inf	Inf	0.95
Uniforme	NULL	Inf	Inf	0.51
Gompertz	-20258.93	Inf	Inf	0.95

Tabla 13 Prueba de bondad de ajuste de variable destino Fuente: propia

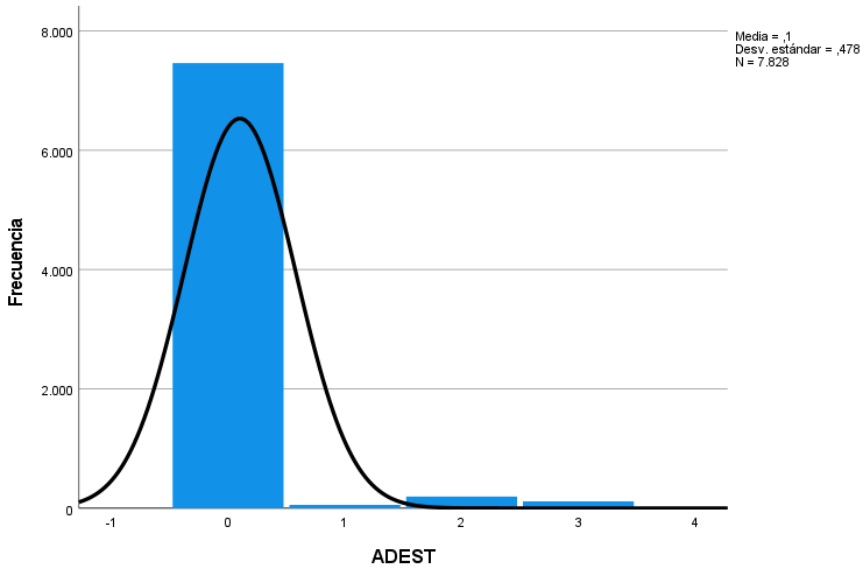

Gráfico 7 Distribución de datos de destinos de vuelo Fuente: propia

Distribución	Akaike	Prueba		
	AIC	Chi Q	AD	KS
Normal	3639.44	53030.16	2559.47	0.53
Logística	-524.72	1015286.10	2616.27	0.45
Exponencial	13945.03	Inf	Inf	0.57
Uniforme	NULL	Inf	Inf	0.49
Student	21418.67	10764.73	6207.70	0.74

Tabla 14 Prueba de Bondad de ajuste de variable trafico Fuente: propia

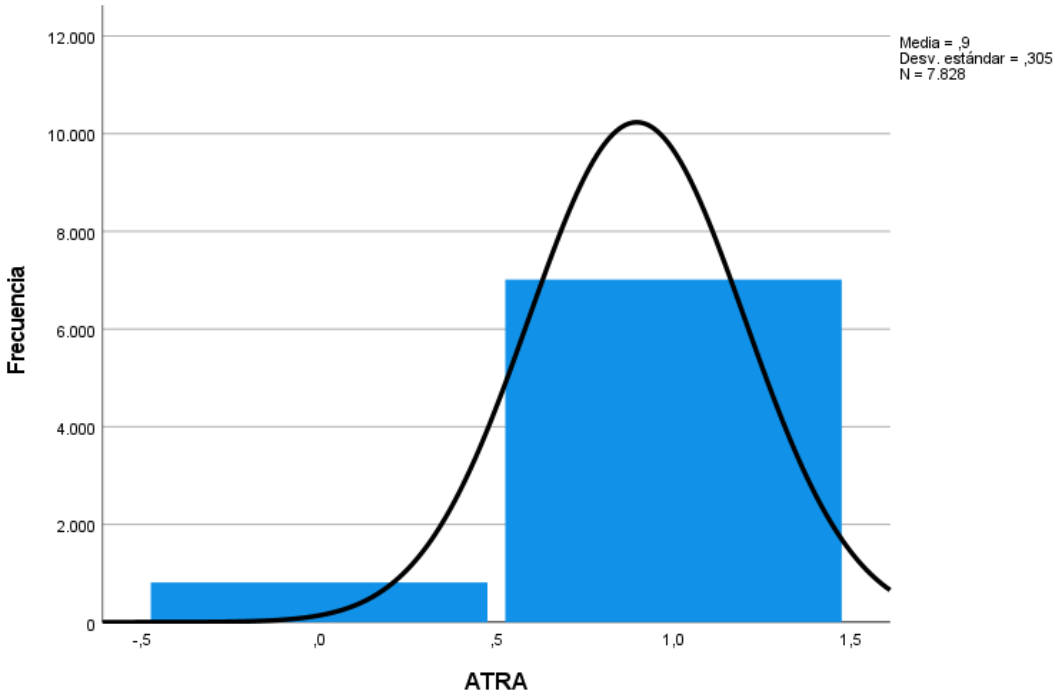

Tabla 15 Distribución de variable trafico

Distribución	Akaike	Prueba		
	AIC	Chis Q	AD	KS
Normal	11359.14	7381.54	1409.70	0.35
Logística	12571.31	6495.31	1248.29	0.33
Exponencial	5414.05	Inf	Inf	0.48
Uniforme	NULL	4875.00	1047.29	0.30
Student	18465.87	1979.09	2317.77	0.50
Gompertz	5379.45	Inf	Inf	0.48

Tabla 16 Prueba de bondad de ajuste de la Variable Aerolínea

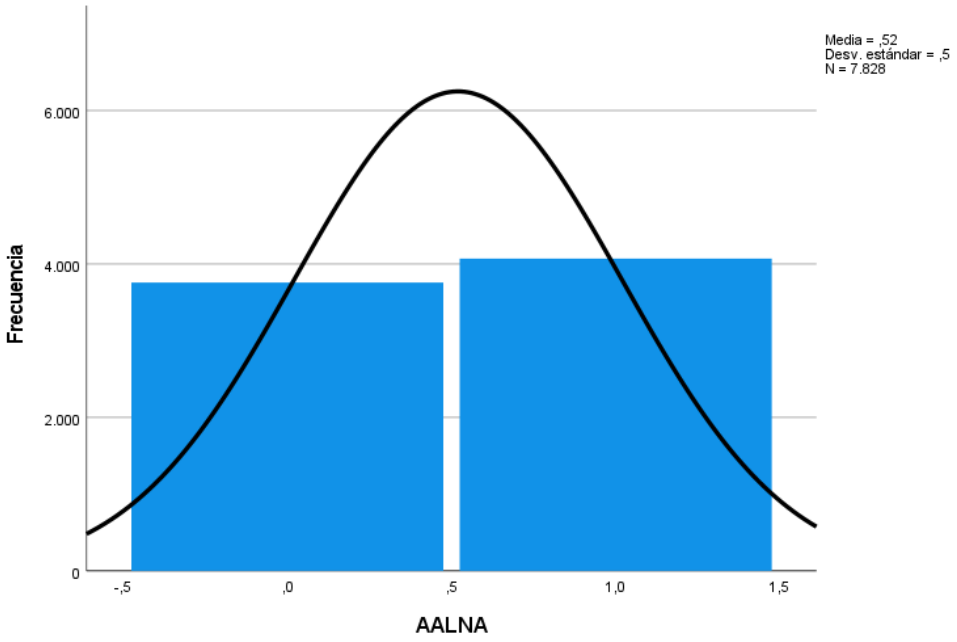

Gráfico 8 Distribución de datos de variable aerolínea Fuente: Propia

Distribución	Akaike	Prueba		
	AIC	Chi Q	AD	KS
Normal	18197.26	4369.97	658.30	0.26
Cauchy	22613.22	2065.80	544.03	0.22
Logística	18921.44	3995.02	612.07	0.24
Exponencial	27805.82	6438.50	1581.55	0.37
Chi-cuadrado	25929.71	5763.60	1116.42	0.38

Tabla 17 Prueba de bondad de ajuste de la variable origen Fuente Propia

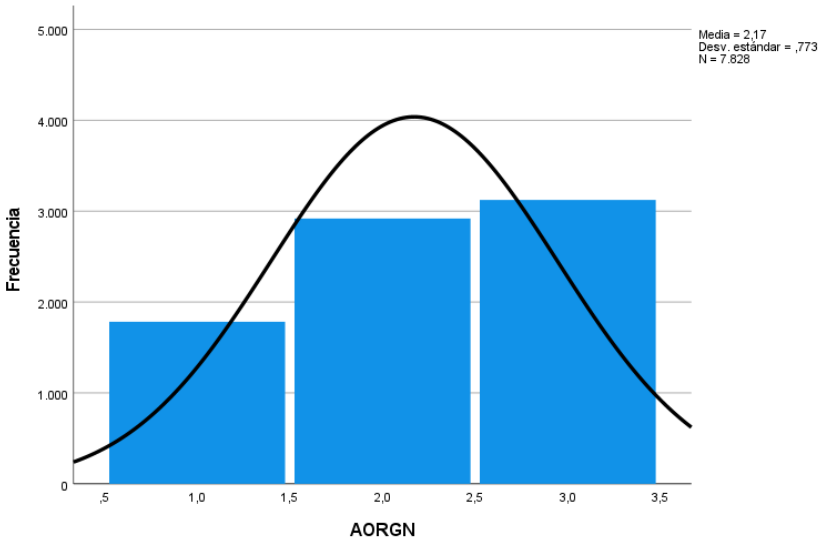

Gráfico 9 Distribución de datos de variable origen Fuente: propia

La distribución que presenta la data en las variables “demora en minutos” y “destino de vuelo”, es una distribución exponencial, es por ello se debe aplicar el método de escalado a la base de datos, con el fin de obtener un rendimiento satisfactorio en el algoritmo.

Este proceso transforma los valores que estén contenidos en un rango especifico, y los transforma en números enteros específicos, por ejemplo:

Datos a escalar: {1-100}

{1-50} ⟶ 0

{51-100} ⟶ 1

El proceso de escalado se refleja en una reducción del error en la predicción del algoritmo.

A continuación, otro reformateo en los datos se da, por el volumen de variables a predecir ya que tener en cuenta todas los códigos, genera conflictos internos en el algoritmo de Random Forest, Arboles de decisión y redes neuronales, por ende, otra modificación de la base de datos, se fundamenta en el análisis Pareto, ya que por medio de la regla 20/80 se puede determinar que las demoras más influentes corresponden solo a 22 causas, en consecuencia, se tendrán en cuenta solo las 10 causas más relevantes en la demora.

6.4 Modelado

6.4.1 Variable a predecir

La variable a predecir es **COD**, a excepción de redes neuronales debido a la alta complejidad de esta, se opta por predecir **ASUM** que es una variable de carácter binaria. En Random Forest predice **COD**, sin embargo, en Árboles de decisión se predice **COD** en conjunto de **ASUM**.

6.4.2 Selección de técnica de modelado

Para el presente proyecto se modelarán 3 técnicas, Random Forest, Arboles de decisión y redes neuronales.

Estas tres técnicas son seleccionadas por su eficiencia en el campo de código predictivo y por su alto nivel de documentación y fuentes. Así mismo el software de modelado es R studio.

Al ser R gratuito, colaborativo y ser un código abierto, permite compartir diferentes algoritmos entre los miembros de la comunidad, es por esta razón, que en este proyecto se hizo uso de los modelos propuestos por el docente Dr. Bharatendra Rai, quien es miembro activo de la Universidad UMass Darthmouth en Massachusetts, Estados Unidos.

Los métodos utilizados fueron los propuestos en Redes Neuronales, Random Forest y Arboles de decisión como se pueden encontrar en su canal de YouTube (Dr. Bharatendra Rai); con adaptaciones a las condiciones que presenta la investigación.

La estructura de los tres modelos es muy similar a las del Dr. Bharatendra, sin embargo, tienen ligeras adiciones hechas por los investigadores, en los tres casos se utilizan funciones para asegurar la calidad de los datos (Lapply, str, prop.table), que variables utilizar en los mismos (hace referencia a las condiciones de la función de cada modelo) y otras de forma específica: Redes neuronales, cuenta con una selección de variables a incluir en la predicción, esto con el fin de lograr dos cosas: reducir el error y evitar que el programa falle al momento de ejecutar dado el número de variables que debe usar; Random Forest, cuenta con un centrado de datos con el fin de reducir el error de la predicción, la selección de número de árboles y la respuesta ante posibles NA (espacios vacíos) con el fin de evitar inconvenientes en la ejecución del modelo, la inclusión de representaciones gráficas del modelo para evidenciar el comportamiento del error vs n de árboles y conocer el peso de cada variable en el modelo; Árboles de decisión, cuenta con la inclusión de gráficas con el fin de observar la respuesta del programa.

6.4.2.1 Técnica de modelado

6.4.2.1.1 Redes Neuronales

La primera técnica de modelado a utilizar son redes neuronales, la cual consiste en la combinación de múltiples neuronas, representadas como nodos.

Estos nodos realizan una regresión lineal interna, con el fin de encontrar los parámetros adecuados y estos parámetros son optimizados hasta llegar al punto de menor error posible.

Posteriormente se proporciona unos datos de entrada, por variable los cuales son transformados por una función de activación, la cual generan una decisión en cada neurona.

La concatenación de estos nodos genera un resultado binario final. Este proceso aplicado a cada registro, entrena la red neuronal con el fin de tomar decisiones en tiempo real.

Los pasos a seguir para la ejecución del modelo están expresados en el siguiente gráfico:

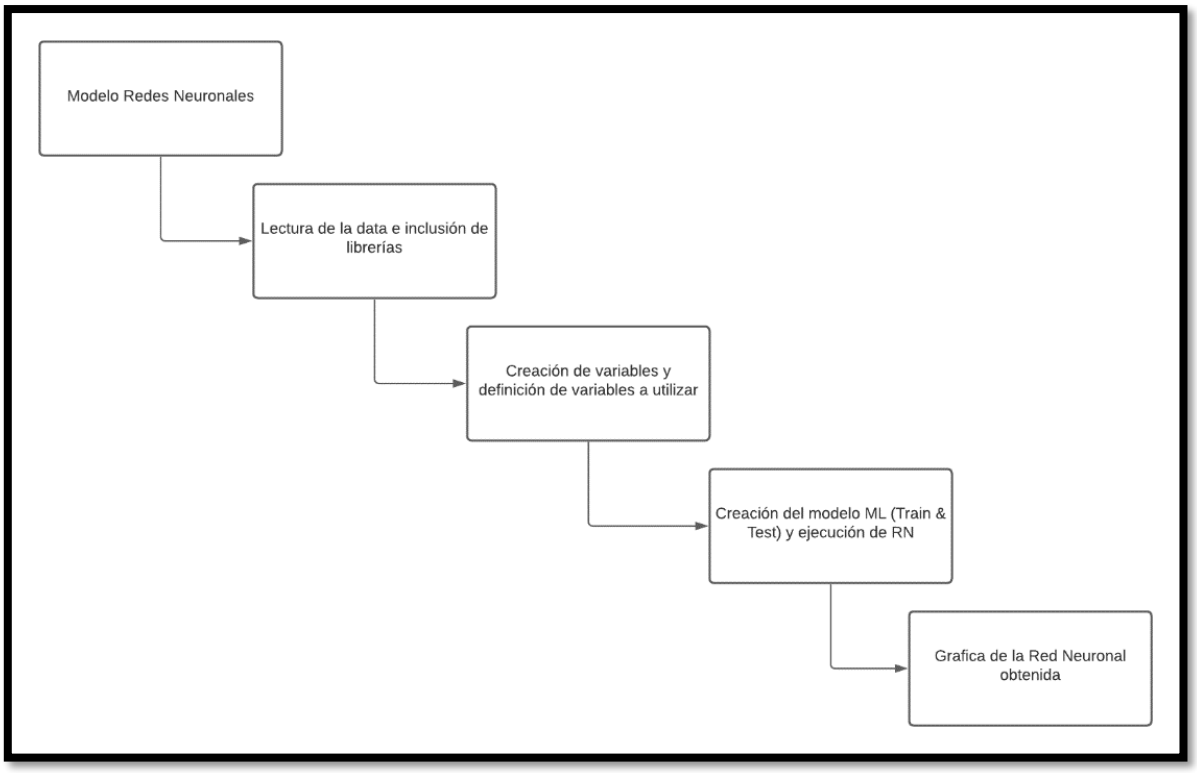

Gráfico 10 Diagrama de construcción para modelo de red neuronal. Fuente: Propia

6.4.2.1.2 Random Forest

La segunda técnica de modelado a utilizar es Random Forest, la cual consiste en una combinación de múltiples árboles de decisión los cuales están compuestos por datos aleatorios tomados de la base de datos por el algoritmo.

Como resultado final se proporciona una categorización de los datos, que permite asignar un valor binario a cada dato de la base real.

Esta categorización se ajusta de acuerdo al peso ponderado de cada variable, como resultado final se obtiene en este caso un número aproximado de los posibles retrasos.

Los pasos a seguir para la ejecución del modelo están expresados en el siguiente gráfico:

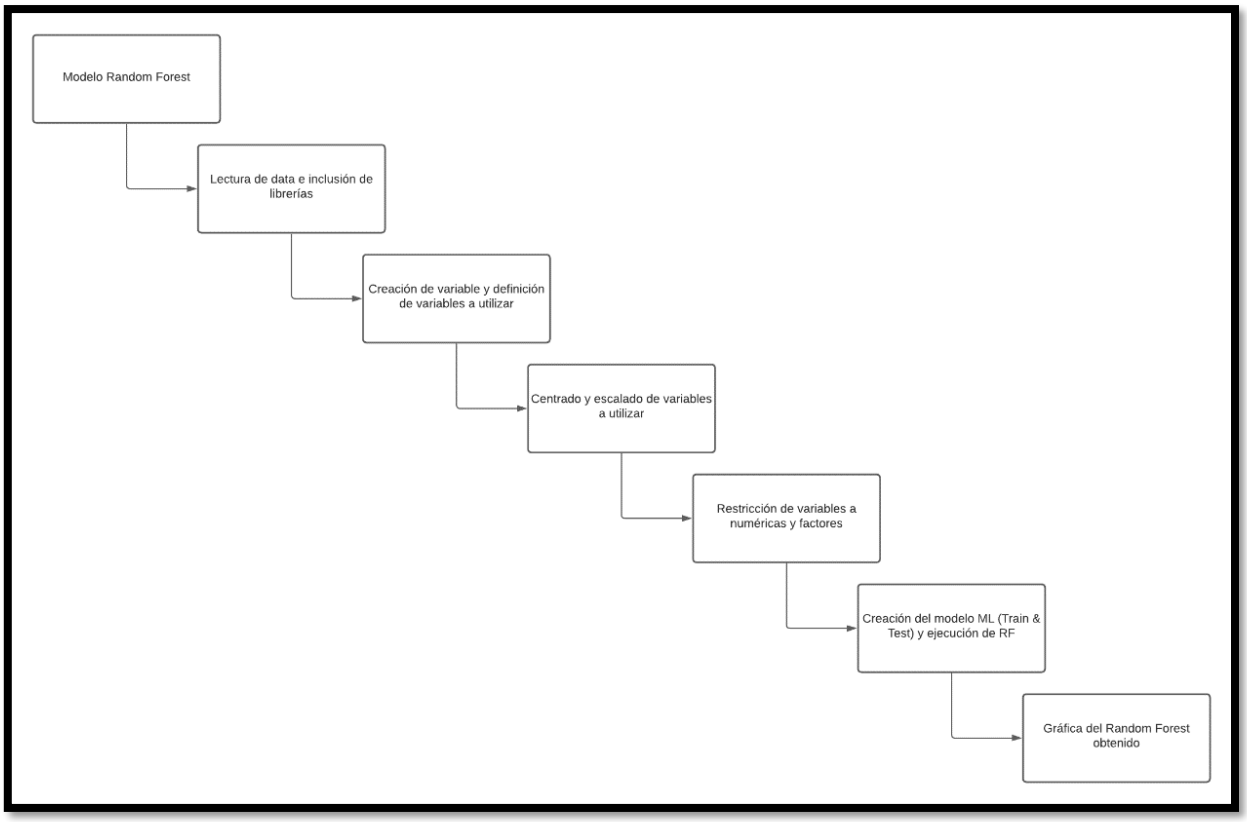

Gráfico 11 Diagrama de flujo para construcción de Random Forest

6.4.2.1.3 Árboles de decisión

La tercera y última técnica a modelar es árboles de decisión. Consiste en un diagrama en el cual, se relacionan los posibles resultados de un conjunto de decisiones. Su principal atributo es que permite al usuario observar las variables asociadas a cada decisión, por ejemplo: costo, beneficio, tiempo etc.

Un uso común es coordinar el intercambio de ideas, así mismo, sirve para trazar un algoritmo que comunique la mejor opción del grupo de decisiones.

Este diagrama está compuesto comúnmente por un único nodo, del cual se relacionan los resultados posibles, de estos resultados emergen otros, este proceso se repite hasta que todos los resultados posibles son relacionados.

Los pasos a seguir para la ejecución del modelo están expresados en el siguiente gráfico:

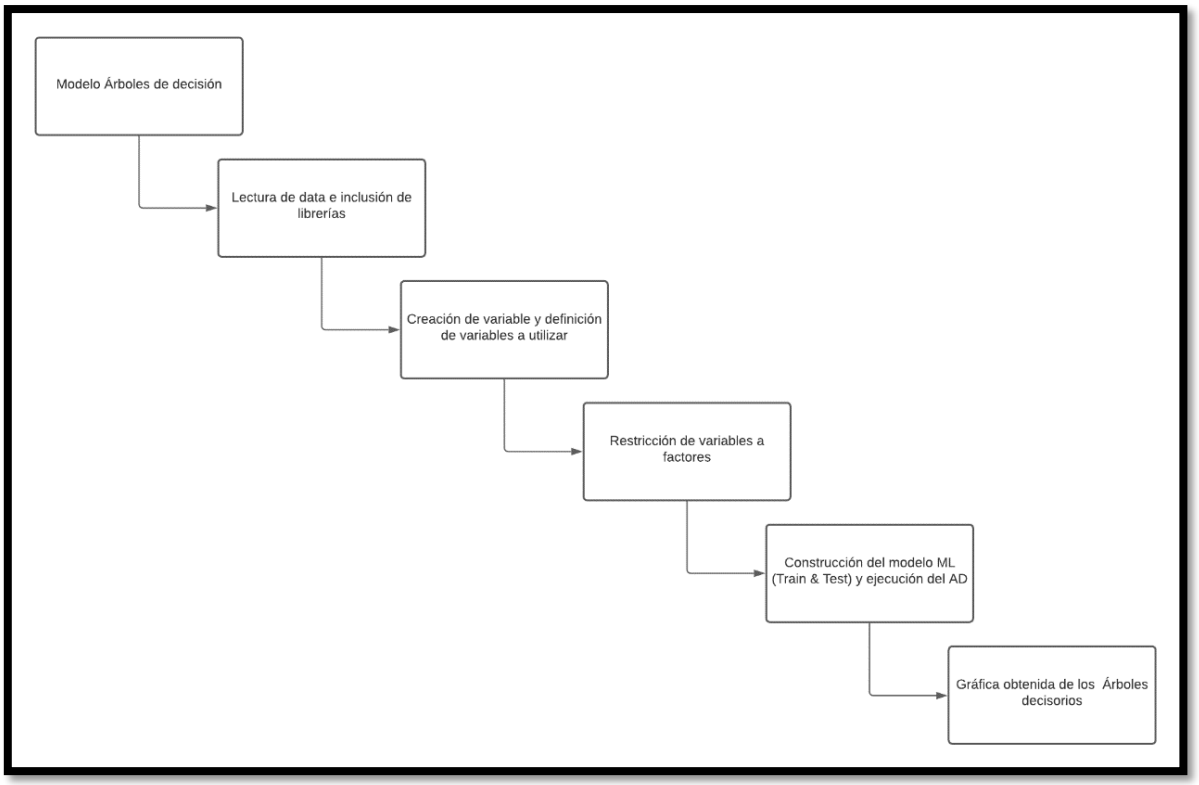

Gráfico 12 Diagrama de flujo de Construcción de árboles de decisión. Fuente: propia

6.4.2.2 Supuesto de modelado

Este aspecto se refiere a los requerimientos de modelado en las variables de la base de datos, en cada técnica (Random Forest, Arboles de decisión y Redes neuronales), así como la transformación de los datos.

Las tres técnicas de modelado requieren una distribución normal en todas sus variables, con el fin de un resultado en la predicción satisfactoria, por ello, se realizó un escalado y centrado de los datos como se explicó en el punto 6.3.1.

6.4.2.2.1 Arboles Decisión

La data set de orígenes que deben ser variables discreteas no continuas, ya que como se expresó en el marco teórico la técnica de árboles de decisión, expresa afirmaciones de si o no (0-1), por ende, requiere una variable que represente el estado de retraso o a tiempo, la cual correlaciona las demás variables. Las variables continuas en esta técnica, se traducen como probabilidad en base a los registros y las relaciona con la variable categórica binomial asociada al estado categórico.

Para representar otras variables de estado por ejemplo ciudades, o destinos se deben utilizar variables categóricas representadas como números enteros a partir

de 1, otro inconveniente de este algoritmo es la impureza de los datos que se reduce con una técnica llamada *"Gini impurity"* sin embargo el data set utilizado no presenta este tipo de problemas ya que a la base de datos se le realizo un proceso de limpieza anterior al modelado, por lo tanto esta técnica no se utilizó para el presente proyecto.

6.4.2.2.1 Random Forest

Esta técnica selecciona 100 datos aleatorios para el data set nuevo, conocido como *"Bootstrapped Dataset"* de los cuales se les crea un árbol de decisión, este proceso se repite n veces hasta optimizar el algoritmo y tenga el mínimo error posible, la decisión que genera, funciona analógicamente como una democracia, por ejemplo, si un dato es clasificado como 80 por la mayoría de los árboles se dice que ese dato será siempre clasificado como 80, sin embargo, a partir del data set original se puede observar los valores con más peso ponderado y ajustar esta. Al ser compuesto, por múltiples árboles de decisión, los requerimientos son iguales. En R la técnica de ramdon Forest solo trabaja con factores y variables numéricas.

6.4.2.2.2 Redes neuronales

Las variables de entrada convergen en una única variable de salida que comunica la decisión de la neurona, es por eso que las variables deben transformarse en variables categóricas discretas, ya que la función de activación escalonada, posee una estructura binomial, y las variables continuas no permiten que el algoritmo obtenga aprendizaje profundo.

6.4.3 Generar prueba de diseño

El plan previsto para capacitar, probar y evaluar los modelos consta de tres etapas, primero se procederá con la creación del modelo ML (Train & Test), a continuación, se procede a evaluar por medio de **prop.table** y **unique** y finalmente se procederá con la ejecución del modelo.

6.4.4 Construcción de modelo: Red neuronal

Antes de explicar el paso a paso, se deben mencionar y definir todas las librerías y funciones a utilizar en el modelo:

6.4.4.1 Parámetros de configuración: Red Neuronal

Librerías: Las librerías son funciones utilizadas para cargar paquetes. Los paquetes son colecciones de funciones y conjuntos de datos realizados por la comunidad. [31]

Readxl	Librería que permite la importación de datos que se encuentran en Excel.

Scales	Librería que permite la modificación de todos los datos de una variable con el fin de centrar y escalar los datos.
Neuralnet	Librería que permite la ejecución del modelo de redes neuronales.

Tabla 18 Librerías de la red neuronal Fuente: Propia

6.4.4.2 Parámetros de configuración

Función	Descripción	Función	Descripción
Library	Función que permite la inclusión de un paquete en el código. Se debe escribir el nombre de la librería correctamente, a medida que se escribe, R muestra las posibles búsquedas.	Set.seed	Función que permite establecer una semilla aleatoria
Read_Excel	Función que permite importar archivos que sean de tipo Excel (.xls y .xlsx). Se debe escribir entre comillas (“”) la ubicación exacta del documento en el equipo, para hacer la búsqueda más fácil, se debe buscar el documento en archivos y en la parte superior dar clic derecho en la búsqueda y copiar url de archivo.	Sample	Función que permite tomar una muestra del tamaño especifico de los elementos de una data. Se debe escribir: nrow, que hace referencia a la data a trabajar; replace, que hace referencia a si el muestreo debe ser con reemplazo; prob, que hace referencia a un vector con probabilidades para obtener elementos de la data a trabajar.
View	Función que permite visualizar objetos que se hayan guardado bajo un nombre determinado. Se debe escribir el nombre del objeto.	Prop.table	Función que permite crear tablas de frecuencia relativa a partir de tablas absolutas. Se debe escribir la tabla a evaluar junto con la tabla de la variable específica.
Str	Función que muestra los detalles de los objetos guardados en memoria, bajo un nombre determinado. Se debe escribir el nombre del objeto.	Unique	Función que permite observar una matriz con la cantidad de cada elemento, sin duplicar elementos. Se debe escribir la data y entre [] la variable a evaluar.

Lapply	Función que permite cohesionar los elementos de una lista a un objeto que se pida. Se debe escribir nombre del objeto y la característica	Neuralnet	Función que permite la aplicación, visualización e implementación de redes neuronales. Se debe escribir: formula, la variable a predecir y variables a participar; data, que hace referencia a la data a trabajar; hidden, el número de neuronas ocultas en cada capa; err.fct, se utiliza para el cálculo del error; act.fct, función para suavizar el resultado del producto cruzado de las neuronas y los pesos; linear.output, es de carácter lógica.

Tabla 19 Funciones a utilizar para redes neuronales Fuente: propia

6.4.4.3 Modelo: Red neuronal

Los pasos de construcción de la red neuronal son:

1. Lectura de la data e inclusión de librerías: Como primer paso en el software R se deben ingresar todas las librerías a utilizar en el modelo, entre las cuales se encuentran:

- Readxl.
- Scales.
- Neuralnet.

Una vez instaladas las librerías se puede continuar con el modelado, importando la base de datos en el programa y comprobando que esta haya pasado sin errores, realizando una lectura de los primeros 50 datos por medio de **view**, asignándole un nuevo nombre a la data y observando la clase que posee cada una de las variables (Entiéndase por clase, como el tipo de variable, numérica, de carácter, de fecha, entre otras) por medio del comando **str**.

```
> library(readxl)
> library(scales)
> library(neuralnet)
Warning message:
package 'neuralnet' was built under R version 4.0.3
>
> dataset_avianca <- read_excel("C:/Users/ASUS/Downloads/Base final con fecha.xls
x")
> View(dataset_avianca)
> str(dataset_avianca)
```

Ilustración 6 código red neuronal parte 1 Fuente: Propia

2. Creación de variable y definición de variables a utilizar:

Una vez contando con la data en R, se procede a la creación de una variable de clase fecha concatenando las variables de **AAÑO, AMES, ADIA**; separadas mediante un guion (-), almacenadas en **date** y en forma (año-mes-día).

```
> # Se concatenan 3 variables para crear la columna date
> dataset_avianca$date <- as.Date(with(dataset_avianca, paste(AAÑO, AMES, ADIA,
 sep="-")), "%Y-%m-%d")
> lapply(dataset_avianca, class)
```

Ilustración 7 código red neuronal parte 2 Fuente: Propia

Se verifica nuevamente la clase de cada una de las variables por medio de **lapply** y se seleccionan las variables a intervenir en la red neuronal entre las cuales están:

- **ATRA**
- **AORGN**
- **ADEST**
- **COD**
- **ADEM**
- **ASUM**

Se rectifica nuevamente, la clase de estas variables y que se hayan descartado las demás por medio de **view y lapply**.

```
> dataset_avianca = dataset_avianca[, c("ATRA", "AORGN", "ADEST", "COD", "ADEM",
 "ASUM")]
> View(dataset_avianca)
> lapply(dataset_avianca, class)
```

Ilustración 8 código red neuronal parte 3 Fuente: Propia

3. Creación del modelo ML (Train & Test) y ejecución de la Red Neuronal

Para crear el modelo de machine learning lo primero que se debe establecer la semilla aleatoria **set.seed**, en este caso el número elegido es “1234”. Seguido de esto, se continúa utilizando el comando **sample** para dividir los datos en entrenamiento y evaluación bajo el nombre de **indDT**, con peso de 80% y 20% y almacenándose como **trainDT** y **TestDT** respectivamente.

```
> set.seed(1234)
> indDT <- sample(2, nrow(dataset_avianca), replace = TRUE, prob = c(0.8, 0.2))
> trainDT <- dataset_avianca[indDT==1,]
> testDT <- dataset_avianca[indDT==2,]
```

Ilustración 9 código red neuronal parte 4 Fuente: Propia

Para evaluar que el proceso de separación quedó correctamente, se evalúa por medio de **prop.table** y **unique** el cual permite verificar los porcentajes del total que posee la variable a predecir en los correspondientes grupos y observar los

```
> prop.table(table(trainDT$ASUM))

        0         1
0.4540092 0.5459908
> prop.table(table(testDT$ASUM))

        0         1
0.4186495 0.5813505
> unique(trainDT[["ASUM"]]) # ver valores diferentes de la columna COD
[1] 0 1
> un
[1]
>
```

Ilustración 10 código red neuronal parte 5 Fuente: Propia

diferentes valores que la variable puede tomar. En este caso la variable a evaluar es ASUM, la cual responde de manera satisfactoria a estas pruebas.

Una vez comprobado esto, se puede proceder a realizar el modelo de redes neuronales almacenado en el objeto **neural_model.** Utilizando el comando **neuralnet**, prediciendo la variable **ASUM**, con todas las variables anteriormente seleccionadas "aportando", con la data de **trainDT**, tres neuronas por capa, utilizando **sse** para calcular el error, suavizando el resultado con **logistic** y con carácter lógico de falso. Al ser un proceso tan lento, dado los procesos que se deben realizar, se opta por sólo evaluar la data de entrenamiento.

```
> neuralnet_model = neuralnet(ASUM~., data = trainDT, hidden = 3, err.fct = "ss
e",
+                             act.fct = "logistic", linear.output = FALSE)
```

Ilustración 11 código red neuronal parte 6 Fuente: Propia

6.4.5 Evaluación de modelo: Red neuronal

Finalmente, por medio del comando **plot** se imprime la gráfica de la red neuronal realizada.

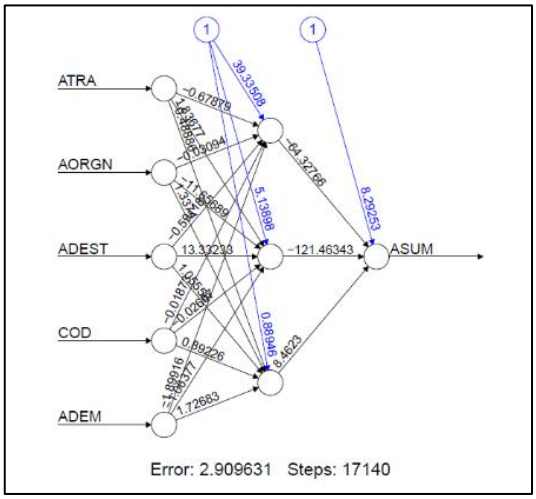

Ilustración 12 resultados redes neuronales Fuente: Propia

Como se observa la red entrenada asigna los valores ponderados a cada nodo de decisión, por motivos sanitarios no es posible evaluar la red neuronal, ya que este modelo funciona como criterio de decisión en tiempo real de la operación, es por ello que para este modelo el único criterio de evaluación es el porcentaje de error, el cual es del 2.9%.

Para observar a detalle ver en anexo 3.

6.4.6 Construcción de modelo: Random Forest

Antes de explicar el paso a paso, se deben mencionar y definir todas las librerías y funciones a utilizar en el modelo:

6.4.6.1 Parámetros de configuración

Librerías

Las librerías son funciones utilizadas para cargar paquetes. Los paquetes son colecciones de funciones y conjuntos de datos realizados por la comunidad. [31]

Readxl	Librería que permite la importación de datos que se encuentran en Excel.
Scales	Librería que permite la modificación de todos los datos de una variable con el fin de centrar y escalar los datos.
RandomForest	Librería que permite la ejecución del modelo de random forest.

Tabla 20 Librerías de Random Forest Fuente: propia Fuente: Propia

Funciones

Library:	Función que permite la inclusión de un paquete en el código. Se debe escribir el nombre de la librería correctamente, a medida que se escribe, R muestra las posibles búsquedas.
Read_Excel:	Función que permite importar archivos que sean de tipo Excel (.xls y .xlsx). Se debe escribir entre comillas (“”) la ubicación exacta del documento en el equipo, para hacer la búsqueda más fácil, se debe buscar el documento en archivos y en la parte superior dar clic derecho en la búsqueda y copiar url de archivo.
View	Función que permite visualizar objetos que se hayan guardado bajo un nombre determinado. Se debe escribir el nombre del objeto.
Str	Función que muestra los detalles de los objetos guardados en memoria, bajo un nombre determinado. Se debe escribir el nombre del objeto.
Lapply	Función que permite cohesionar los elementos de una lista a un objeto que se pida. Se debe escribir nombre del objeto y la característica.
Set.seed	Función que permite establecer una semilla aleatoria.
Sample	Función que permite tomar una muestra del tamaño especifico de los elementos de una data. Se debe escribir: nrow, que hace referencia a la data a trabajar; replace, que hace referencia a si el muestreo debe ser con reemplazo; prob, que hace referencia a un vector con probabilidades para obtener elementos de la data a trabajar.
Prop.table	Función que permite crear tablas de frecuencia relativa a partir de tablas absolutas. Se debe escribir la tabla a evaluar junto con la tabla de la variable específica.
Unique	Función que permite observar una matriz con la cantidad de cada elemento, sin duplicar elementos. Se debe escribir la data y entre [] la variable a evaluar.
Rescale	Cuando las variables de una data, tienen mucha diferencia numérica, se escala las variables en un rango de [0,1]. Se debe escribir la variable a escalar y en qué objeto se guardará, por lo general, se reescribe la variable.
RandomForest	Función que permite la aplicación, visualización e implementación de random forest. Se debe escribir: la variable a predecir, junto con la selección de las variables a influir en el modelo; data, la data con la que se va a trabajar; ntree, el número de árboles a realizar; na.action, qué hacer en caso de encontrar un espacio vacío en la data.
Predict	Función que se puede aplicar a un modelo para obtener los valores de “y”. Se debe escribir el modelo a predecir y la data a trabajar.
Round	
	Función que devuelve un valor numérico redondeado a lo que se digite. Se debe escribir el objeto y el número de decimales.
Importance	Función que imprime la importancia de las variables en un objeto guardado.Se debe escribir el nombre del objeto.

Print	Función que permite la visualización de objetos guardados en un objeto. Se debe escribir el nombre del objeto.
VarImpPlot	Función que grafica los puntos de importancia-variable según lo establecido en el modelo RF. Se debe escribir el nombre del modelo en que fue guardado; col, el color para imprimir los puntos; pch, la forma de los puntos (Triángulos, puntos, cuadrados, entre otros).
Plot	Comando que permite crear gráficos de un objeto. Se debe escribir: El nombre del objeto guardado en memoria.

Tabla 21 fuentes a utilizar para algoritmo de Random Forest Fuente: propia

6.4.6.2 Modelo: Random Forest

1. Lectura de data e inclusión de librerías

Como primer paso en el software R se deben ingresar todas las librerías a utilizar en el modelo, entre las cuales se encuentran:

a. Readxl.
b. Scales.
c. Neuralnet.

Una vez instaladas las librerías se puede continuar con el modelado, importando la base de datos en el programa y comprobando que esta haya pasado sin errores, realizando una lectura de los primeros 50 datos por medio de **view**, asignándole un nuevo nombre a la data y observando la clase que posee cada una de las variables (Entiéndase por clase, como el tipo de variable, numérica, de carácter, de fecha, entre otras) por medio del comando **str.**

```
> library(readxl)
> library(scales)
> library(randomForest)
> dataset_avianca <- read_excel("C:/Users/ASUS/Downloads/Base final con fecha.xlsx")
> View(dataset_avianca)
> str(dataset_avianca)
```

Ilustración 13 Código Random Forest parte 1 Fuente: Propia

2. Creación de variable y definición de variables a utilizar

Una vez contando con la data en R, se procede a la creación de una variable de clase fecha concatenando las variables de **AAÑO, AMES, ADIA**; separadas mediante un guion (-), almacenadas en date y en forma (año-mes-día).

```
> dataset_avianca$date <- as.Date(with(dataset_avianca, paste(AAÑO, AMES, ADIA, sep="-
")), "%Y-%m-%d")
> lapply(dataset_avianca, class)
```

Ilustración 14 Código Random Forest parte 2 Fuente: propia

Se verifica nuevamente la clase de cada una de las variables por medio de **lapply** y se seleccionan las variables a intervenir en el Random Forest, entre las cuales están:

1. ATRA.
2. AALNA
3. AORGN.
4. ADEST.
5. ANVLO.
6. ADSEM.
7. COD.
8. ADEM.
9. ASUM.
10. ADES.
11. Date.

Se rectifica nuevamente, que se hayan descartado las demás por medio de **view**.

```
> dataset_avianca = dataset_avianca[, c("ATRA", "AALNA", "AORGN", "ADEST", "ANVLO", "A
DSEM", "COD","AMOT", "ADEM", "ASUM", "ADES", "date")]
> View(dataset_avianca)
```

Ilustración 15 Código Random Forest parte 3 Fuente: Propia

Centrado y escalado de variables a utilizar: Una vez seleccionadas las variables a utilizar, dado que no todas las variables manejan la misma escala numérica (Mientras que ATRA maneja datos de [0,1], ANVLO maneja datos de cuatro cifras) se opta por escalar todas las variables por medio de la función **rescale**. Cabe aclarar que esto no es necesario para todas las variables, sin embargo, se opta por hacerlo de manera equitativa para todas, con el fin de que el algoritmo no presente errores a posteriori. Seguido de esto se observa nuevamente la clase y las variables de la data por medio de **lapply**.

```
> dataset_avianca$ATRA <- rescale(dataset_avianca$ATRA)
> dataset_avianca$AALNA <- rescale(dataset_avianca$AALNA)
> dataset_avianca$AORGN <- rescale(dataset_avianca$AORGN)
> dataset_avianca$ADEST <- rescale(dataset_avianca$ADEST)
> dataset_avianca$ANVLO <- rescale(dataset_avianca$ANVLO)
> dataset_avianca$ADSEM <- rescale(dataset_avianca$ADSEM)
> dataset_avianca$AMOT <- rescale(dataset_avianca$AMOT)
> dataset_avianca$ADEM <- rescale(dataset_avianca$ADEM)
> lapply(dataset_avianca, class)
```

Ilustración 16 Código Random Forest parte 4 Fuente: Propia

Restricción de variables numéricas y factores: En el modelo de Random Forest, sólo se pueden utilizar variables de carácter numérica y de factor, por ende, a las variables que no cuentan con dicha restricción se modifican a alguna de las alternativas. Luego, se observan los datos que contiene **COD** en la base de datos por medio de **table** y nuevamente se rectifica la clase de las variables por medio de **lapply**.

```
> dataset_avianca$date <- as.numeric(dataset_avianca$date)
> dataset_avianca$COD <- as.factor(dataset_avianca$COD)
> table(dataset_avianca$COD)

  41   43   51   63   71   72   73   81   87   93
1132  516  534  497  677  506  367 1862  844  893
> dim(dataset_avianca)
[1] 7828   12
> lapply(dataset_avianca, class)
```

Ilustración 17 Código Random Forest parte 5 Fuente: Propia

3. Creación del modelo ML (Train & Test) y ejecución de RF

Para crear el modelo de machine learning lo primero que se debe establecer la semilla aleatoria **set.seed**, en este caso el número elegido es “300”. Seguido de esto, se continúa utilizando la función **sample** para dividir los datos en entrenamiento y evaluación bajo el nombre de **ind**, con peso de 70% y 30% y almacenándose como **train** y **test** respectivamente.

```
> set.seed(300)
> ind <- sample(2, nrow(dataset_avianca), replace = TRUE, prob= c(0.7,0.3))
> train <- dataset_avianca[ind==1,]
> test <- dataset_avianca[ind==2,]
```

Ilustración 18 Código Random Forest parte 6 Fuente: Propia

Para evaluar que el proceso de separación quedó correctamente, se evalúa por medio de **prop.table** y unique el cual permite verificar los porcentajes del total que poseen la variable a predecir en los correspondientes grupos y observar los diferentes valores que la variable puede tomar. En este caso la variable a evaluar es **COD**, la cual responde de manera satisfactoria a estas pruebas.

```
> prop.table(table(train$COD))

        41         43         51         63         71         72         73
0.14309796 0.06873946 0.06705376 0.06386964 0.08597116 0.06312043 0.04701255
        81         87         93
0.23674845 0.10975838 0.11462821
> prop.table(table(test$COD))

        41         43         51         63         71         72         73
0.14785054 0.05986340 0.07071113 0.06267577 0.08758538 0.06789875 0.04660506
        81         87         93
0.24025713 0.10365609 0.11289675
> unique(train[["COD"]]) # ver valores diferentes de la columna COD
 [1] 81 72 71 41 87 93 63 51 43 73
> unique(test[["COD"]]) # ver valores diferentes de la columna COD
 [1] 73 81 43 51 71 72 41 87 63 93
```

Ilustración 19 Código Random Forest parte 7 Fuente: Propia

Una vez comprobado esto, se puede proceder a realizar la selección de la semilla “12345”, el modelo de random forest almacenado en el objeto **“rf”**, utilizando la función **randomForest**, prediciendo la variable **COD**, con todas las variables anteriormente seleccionadas “aportando”, con la data **train**, 1000 árboles a realizar. Una vez finalizado el proceso se imprime el objeto rf para observar la predicción.

```
> set.seed(12345)
> rf <- randomForest(COD~., data = train, ntree = 1000, na.action = na.roughfix)
```

Ilustración 20 Código Random Forest parte 8 Fuente: Propia

```
Call:
 randomForest(formula = COD ~ ., data = train, ntree = 1000, na.action = na.roughf
ix)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.22%
Confusion matrix:
    41  43  51  63  71  72  73   81  87  93 class.error
41 764   0   0   0   0   0   0    0   0   0 0.000000000
43   0 367   0   0   0   0   0    0   0   0 0.000000000
51   0   0 354   0   2   0   0    0   2   0 0.011173184
63   0   0   0 341   0   0   0    0   0   0 0.000000000
71   0   0   0   0 459   0   0    0   0   0 0.000000000
72   0   0   1   0   1 335   0    0   0   0 0.005934718
73   0   0   0   0   0   0 246    0   5   0 0.019920319
81   0   0   0   0   0   0   0 1264   0   0 0.000000000
87   0   0   0   0   0   0   0    0 586   0 0.000000000
93   0   0   0   0   0   0   0    0   1 611 0.001633987
```

Ilustración 21 Código Random Forest parte 9 Fuente: Propia

Se obtiene una matriz en la que se puede observar la cantidad de retrasos que el modelo las atribuye a las 10 causas principales de forma individual, con un error inferior al 1% lo cual demuestra que el modelo responde de manera satisfactoria a las exigencias.

Dado que los resultados obtenidos con la data de entrenamiento fueron positivos, se examina la data de evaluación para corroborar los resultados y confirmar que el modelo es óptimo.

```
Call:
 randomForest(formula = COD ~ ., data = test, ntree = 1000, na.action = na.roughfi
x)
               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 3

        OOB estimate of  error rate: 0.72%
Confusion matrix:
    41  43  51  63  71  72  73  81  87  93 class.error
41 368   0   0   0   0   0   0   0   0   0 0.000000000
43   0 148   0   0   0   0   0   0   0   1 0.006711409
51   0   0 167   0   2   3   1   0   3   0 0.051136364
63   0   0   0 154   0   0   0   0   1   1 0.012820513
71   0   0   0   0 218   0   0   0   0   0 0.000000000
72   0   0   0   0   0 168   0   0   1   0 0.005917160
73   0   0   0   0   0   0 112   0   4   0 0.034482759
81   0   0   0   0   0   0   0 598   0   0 0.000000000
87   0   0   0   0   0   0   0   0 258   0 0.000000000
93   0   0   0   0   0   0   0   0   1 280 0.003558719
```

Ilustración 22 Código Random Forest parte 10 Fuente: Propia

La data de evaluación presenta un error mayor que la de entrenamiento, sin embargo, no es significativo ya que tampoco supera el 1% del error y al igual que en entrenamiento, se mantienen las mismas predominaciones en los respectivos códigos como era de esperarse.

6.4.7 Evaluación de modelo: Random Forest

Finalmente, por medio de la función **plot** se imprime la gráfica del random Forest, para observar el error obtenido; por medio de **VarImpPlot**, se evidencia gráficamente el peso de cada variable en el modelo, con color azul y forma de triángulos.

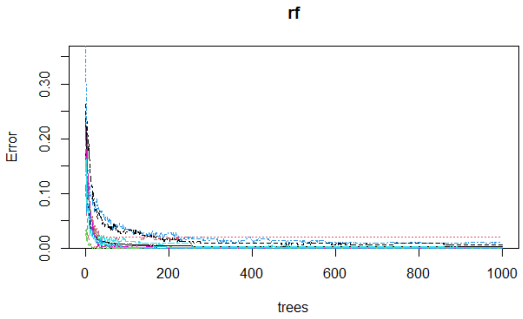

Ilustración 23 Código Random Forest parte 11 Fuente: Propia

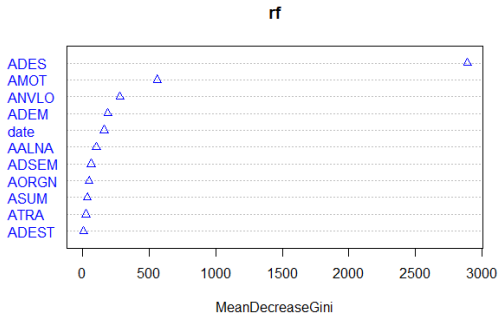

Ilustración 24 Código Random Forest parte 12 Fuente: Propia

Para observar a detalle ver en anexo 4.

A medida que se aumentan la cantidad de árboles, se disminuye el error, cabe resaltar, que después de 200 árboles se observa que el error es inferior al 5%

También se hace uso de las funciones **Round** concatenada con **importance**, para conocer la importancia de cada variable en forma numérica.

	MeanDecreaseGini
ATRA	28.83
AALNA	105.84
AORGN	51.77
ADEST	11.54
ANVLO	285.00
ADSEM	69.14
AMOT	564.14
ADEM	193.48
ASUM	38.95
ADES	2889.39
date	163.48

Ilustración 25 Código Random Forest parte 13 Fuente: Propia

Las variables que más peso tienen son los números inferiores entre las cuales se encuentran **ADEST, ATRA, ASUM y AORGN**.

El modelo Random Forest, da una respuesta muy positiva ante las exigencias del proyecto, entrega un error inferior al 1%, con una confiabilidad superior al 95% y el uso de todas las variables en la predicción, lo que hace al modelo el mejor hasta el momento a falta de ejecutar Árboles de decisión.

6.4.1 Construcción de modelo: Árbol de decisión

6.4.1.1 Parámetros de árboles de decisión

Antes de explicar el paso a paso, se deben mencionar y definir todas las librerías y funciones a utilizar en el modelo:

Librerías

Las librerías son funciones utilizadas para cargar paquetes. Los paquetes son colecciones de funciones y conjuntos de datos realizados por la comunidad.

Readxl	Librería que permite la importación de datos que se encuentran en Excel.
Scales	Librería que permite la modificación de todos los datos de una variable con el fin de centrar y escalar los datos.
Party	Librería que permite la ejecución del modelo de árboles de decisión.
Rpart	Librería que permite la ejecución del modelo de árboles de decisión.

Tabla 22 Librerías de árboles de decisión Fuente: propia Fuente: Propia

Funciones

Library	Función que permite la inclusión de un paquete en el código. Se debe escribir el nombre de la librería correctamente, a medida que se escribe, R muestra las posibles búsquedas.
Read_Excel	Función que permite importar archivos que sean de tipo Excel (.xls y .xlsx). Se debe escribir entre comillas (“”) la ubicación exacta del documento en el equipo, para hacer la búsqueda más fácil, se debe buscar el documento en archivos y en la parte superior dar clic derecho en la búsqueda y copiar url de archivo.
View	Función que permite visualizar objetos que se hayan guardado bajo un nombre determinado. Se debe escribir el nombre del objeto.
Str	Función que muestra los detalles de los objetos guardados en memoria, bajo un nombre determinado. Se debe escribir el nombre del objeto.
Lapply	Función que permite cohesionar los elementos de una lista a un objeto que se pida. Se debe escribir nombre del objeto y la característica.
Set.seed	Función que permite establecer una semilla aleatoria.
Sample	Función que permite tomar una muestra del tamaño especifico de los elementos de una data. Se debe escribir: nrow, que hace referencia a la data a trabajar; replace, que hace referencia a si el muestreo debe ser con reemplazo; prob, que hace referencia a un vector con probabilidades para obtener elementos de la data a trabajar.
Prop.table	Función que permite crear tablas de frecuencia relativa a partir de tablas absolutas. Se debe escribir la tabla a evaluar junto con la tabla de la variable específica.
Unique	Función que permite observar una matriz con la cantidad de cada elemento, sin duplicar elementos. Se debe escribir la data y entre la variable a evaluar.
Predict	Función que se puede aplicar a un modelo para obtener los valores de “y”. Se debe escribir el modelo a predecir y la data a trabajar.
Ctree	Función que permite la aplicación, visualización e implementación de los árboles de decisión. Se debe escribir la variable a predecir y las variables que contribuyen en el modelo; data, la data a trabajar en el modelo; controls, …
Rpart	Función que permite la aplicación, visualización e implementación de los árboles de decisión. Se debe escribir la variable a predecir, junto a las variables que contribuyen en el modelo; data, la data a trabajar en el modelo y el método que se utilizará.
Rpart.plot	Función que muestra gráficamente el árbol de decisión obtenido en la función rpart. Se debe escribir el nombre del objeto realizado en la función Rpart y extra, que selecciona un valor basado en el modelo.
Print	Función que permite la visualización de objetos guardados en un objeto. Se debe escribir el nombre del objeto.

Tabla 23 Funciones de árboles de decisión Fuente: propia

6.4.1.2 Modelo: árbol de decisión

1. Lectura de data e inclusión de librerías

Como primer paso en el software R se deben ingresar todas las librerías a utilizar en el modelo, entre las cuales se encuentran:

- Readxl.
- Scales.
- Rpart.
- Party.

Una vez instaladas las librerías se puede continuar con el modelado, importando la base de datos en el programa y comprobando que esta haya pasado sin errores, realizando una lectura de los primeros 50 datos por medio de view, asignándole un nuevo nombre a la data y observando la clase que posee cada una de las variables (Entiéndase por clase, como el tipo de variable, numérica, de carácter, de fecha, entre otras) por medio del comando **str**.

```
> library(readxl)
> library(scales)
> library(party)
> library(rpart)
> dataset_avianca <- read_excel("C:/Users/ASUS/Downloads/Base final con fecha.xls
x")
> View(dataset_avianca)
> str(dataset_avianca)
```

Ilustración 26 Código árboles de decisión parte 1 Fuente: Propia

2. Creación de variable y definición de variables a utilizar

Una vez contando con la data en R, se procede a la creación de una variable de clase fecha concatenando las variables de **AAÑO, AMES, ADIA**; separadas mediante un guion (-), almacenadas en **date** y en forma (año-mes-día).

```
> dataset_avianca$date <- as.Date(with(dataset_avianca, paste(AAÑO, AMES, ADIA, sep="-
")), "%Y-%m-%d")
> lapply(dataset_avianca, class)
```

Ilustración 27 Código árboles de decisión parte 2 Fuente: Propia

Se verifica nuevamente la clase de cada una de las variables por medio de **lapply** y se seleccionan las variables a intervenir en los árboles de decisión, entre las cuales están:

- ATRA.
- AALNA
- AORGN.
- ADEST.
- ANVLO.
- ADSEM.
- COD.
- ADEM.
- ASUM.
- ADES.

- Date.

Se rectifica nuevamente, que se hayan descartado las demás por medio de **view**.

```
> dataset_avianca = dataset_avianca[, c("ATRA", "AALNA", "AORGN", "ADEST", "ANVLO", "A
DSEM", "COD","AMOT", "ADEM", "ASUM", "ADES", "date")]
> View(dataset_avianca)
```

Ilustración 28 Código árboles de decisión parte 3 Fuente: Propia

4. Restricción de variables a factores

En el modelo de Arboles de decisión, sólo se pueden utilizar variables de carácter factor, por ende, a las variables que no cuentan con dicha restricción se modifican. Luego, se observan los datos que contiene **COD** en la base de datos por medio de **prop.table** concatenada con **table** y nuevamente se rectifica la clase de las variables por medio de **lapply**.

```
> dataset_avianca = dataset_avianca[, c("ATRA", "AALNA", "AORGN", "ADEST", "ANVL
O", "ADSEM", "COD","AMOT", "ADEM", "ASUM", "ADES", "date")]
> View(dataset_avianca)
> dataset_avianca$COD = as.factor(dataset_avianca$COD)
> dataset_avianca$ADEM = as.factor(dataset_avianca$ADEM)
> dataset_avianca$AORGN = as.factor(dataset_avianca$AORGN)
> dataset_avianca$ADEST = as.factor(dataset_avianca$ADEST)
```

Ilustración 29 Código árboles de decisión parte 4 Fuente: Propia

5. Creación del modelo ML (Train & Test) y ejecución de AD

Para crear el modelo de machine learning lo primero que se debe establecer la semilla aleatoria **set.seed**, en este caso el número elegido es “1234”. Seguido de esto, se continúa utilizando el comando **sample** para dividir los datos en entrenamiento y evaluación bajo el nombre de **indDT**, con peso de 80% y 20% y almacenándose como **trainDT** y **testDT** respectivamente.

```
> set.seed(1234)
> indDT <- sample(2, nrow(dataset_avianca), replace = TRUE, prob = c(0.8, 0.2))
> trainDT <- dataset_avianca[indDT==1,]
> testDT <- dataset_avianca[indDT==2,]
```

Ilustración 30 Código árboles de decisión parte 5 Fuente: Propia

Para evaluar que el proceso de separación quedó correctamente, se evalúa por medio de **prop.table** y unique el cual permite verificar los porcentajes del total que poseen la variable a predecir en los correspondientes grupos y observar los diferentes valores que la variable puede tomar. En este caso la variable a evaluar es **COD**, la cual responde de manera satisfactoria a estas pruebas.

```
> prop.table(table(trainDT$COD))

        41         43         51         63         71         72         73
0.14124024 0.06791009 0.06727244 0.06456241 0.08767735 0.06233062 0.04734577
        81         87         93
0.24071417 0.10808226 0.11286466
> prop.table(table(testDT$COD))

        41         43         51         63         71         72         73
0.15819936 0.05787781 0.07202572 0.05916399 0.08167203 0.07395498 0.04501608
        81         87         93
0.22636656 0.10675241 0.11897106
> unique(trainDT[["COD"]]) # ver valores diferentes de la columna COD
 [1] 73 81 43 72 51 71 41 87 93 63
Levels: 41 43 51 63 71 72 73 81 87 93
> unique(testDT[["COD"]]) # ver valores diferentes de la columna COD
 [1] 81 72 51 63 41 43 93 87 71 73
Levels: 41 43 51 63 71 72 73 81 87 93
```

Ilustración 31 Código árboles de decisión parte 6 Fuente: Propia

Una vez comprobado esto, se puede proceder a realizar el modelo de árboles de decisión almacenado en el objeto "**tree**", utilizando la función **ctree**, prediciendo la variable **COD**, con las variables **ASUM, AORGN y ADEST** "aportando", con la data **trainDT** y método **class**. Seguido de esto, se evalúa el segundo modelo de árboles de decisión, guardando el objeto en "**fit**", utilizando la función **rpart**, prediciendo la variable **COD**, con las variables **ASUM, AORGN y ADEST** "aportando", con la data **trainDT** y método **class**.

```
> tree <- ctree(COD~ASUM+AORGN+ADEST+ADEM, data = trainDT,
+               controls = ctree_control(mincriterion = 0.8, minsplit = 200))
```

Ilustración 32 Código árboles de decisión parte 7 Fuente: Propia

```
> library(rpart)
> fit <- rpart(COD~ASUM+AORGN+ADEST, data = trainDT, method = 'class')
```

Ilustración 33 Código árboles de decisión parte 8 Fuente: Propia

El modelo de árboles de decisión cuenta con dos métodos distintos, sin embargo, el modelo con la función **ctree** se debe descartar debido a su lento tiempo de respuesta en ejecución y su gráfica poco explicativa. A continuación, se muestran los resultados obtenidos mediante la función **rpart**:

```
predict_unseen  41  43  51  63  71  72  73  81  87  93
            41 600 295 311 176 390 286 219 515 225 408
            43   0   0   0   0   0   0   0   0   0   0
            51   0   0   0   0   0   0   0   0   0   0
            63   0   0   0   0   0   0   0   0   0   0
            71   0   0   0   0   0   0   0   0   0   0
            72   0   0   0   0   0   0   0   0   0   0
            73   0   0   0   0   0   0   0   0   0   0
            81 286 131 111 229 160 105  78 995 453 300
            87   0   0   0   0   0   0   0   0   0   0
            93   0   0   0   0   0   0   0   0   0   0
```

Ilustración 34 Código árboles de decisión parte 9 Fuente: Propia

Sin embargo, la predicción no es clara sobre cómo se ejecutaron los árboles de decisión y el por qué sólo existen valores asignados a 41 y a 81.

6.4.1.3 Evaluar modelo: Árbol de decisión

Finalmente, por medio de la función **plot** se imprime la gráfica de los árboles de decisión para entender cómo se ejecutó el modelo.

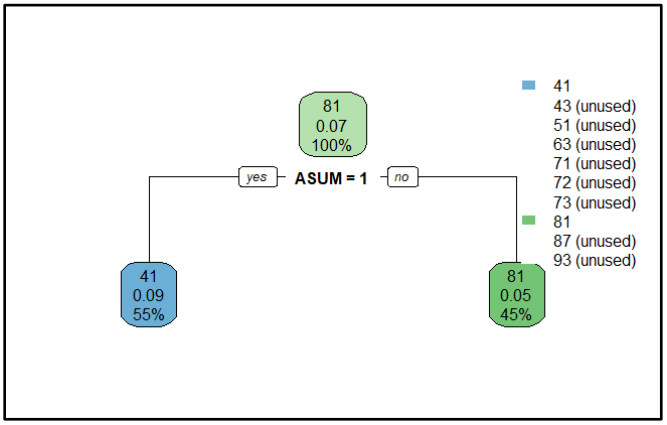

Ilustración 35 Código arboles de decisión parte 10 Fuente: Propia

Se puede observar de una forma más clara cuáles fueron las decisiones tomadas por el modelo. Se creó una correlación entre la variable **COD** y **ASUM** para desarrollar la función, tomando una decisión si **ASUM** era o no igual a uno (que la demora sea más de 20 minutos). Dando como respuesta que existe un 55% de posibilidades de que sea igual a uno y la asocie a las variables azules (41, 43, 51, 63, 71, 72 y 73) y un 45% de posibilidades de que no sea igual a uno y la asocie a las variables verdes (81, 87 y 93). Pero para los respectivos colores, sólo tuvo en cuenta un código de demora por ende todas las demoras solo fueron atribuidas a 41 y 81.

Finalmente, aunque los porcentajes de error en el modelo son inferiores al 10% (como se puede apreciar en el número medio de cada cuadro) que es un buen resultado, no puede ser tenido en cuenta, ya que sólo hizo una predicción enfocada a dos códigos, por ende, no es un modelo óptimo.

6.4.2 Revisión de paramentos de configuración

Como primer acercamiento el primer modelado que se programó, se concibió con las distribuciones sin escalado, es por ello que el primer modelo genero un error del 40% de desacierto en el algoritmo de Random Forest.

En el algoritmo de árboles de decisión los valores numéricos se convirtieron en factores para obtener una reducción del 20% en el error como resultado el error final es menos del 10%.

Así mismo la red neuronal inicial no fue entrenada correctamente y sus errores ascendían a 2000 decisiones incorrectas, el resultado era un error global de 94.5%, posterior a un correcto entrenamiento el error se redujo al 2.4%.

6.5 EVALUACION

6.5.1 Evaluación de resultados

Los resultados del proyecto, son satisfactorios en términos de objetivos, ya que se obtuvo un análisis en la tendencia de los datos más relevantes, se logró Identificar las principales causas de retrasos en los aeropuertos de estudio y como resultado se logró desarrollar un algoritmo en el software R que permite prever los posibles retrasos en despegues, en los aeropuertos de Barranquilla, Cali y Cartagena, así mismo obtuvo un criterio para definir cuál es el algoritmo más adecuado en términos de predicción de retrasos.

6.5.2 Evaluación de minería de datos

Los datos recabados por la Aerocivil cumplen con los requerimientos del proyecto, como resultado los modelos predictivos generaron resultados satisfactorios, y son fruto de la recaudación de datos, como es una fuente directa de observación se presume que los datos en cuestión de calidad, son confiables y representan la situación a estudiar.

Los resultados de la minería de datos, en el numeral 6.1.3, dan respuesta al objetivo específico número uno y parcialmente al objetivo número dos, el cual es complementado como resultado de desarrollo del objetivo número 3. Sin embargo, en el proceso de recaudación de los datos se recomienda una sistematización, ya que durante el proceso de limpieza de la base de datos se identificaron varios conflictos, a raíz de errores humanos como espacios vacíos, inconsistencia en la información, formato inadecuado y diferencias entre registros.

6.5.3 Evaluación de modelos

Las comparaciones de los tres modelos permiten, determinar el modelo más preciso y acorde para la predicción de los retrasos aeroportuarios en las ciudades de estudio.

El modelo más acorde es Random Forest, ya que es el que en cuestión de resultados genera la respuesta más satisfactoria, el modelo de redes neuronales es descartado ya que este modelo actúa en tiempo real, en base a las entradas que se proporciona, genera una decisión. Como en el actual 2020 el mundo experimenta una pandemia, las operaciones aéreas tienen una anormalidad en su funcionamiento, esta situación vuelve obsoleto el algoritmo ya que el entrenamiento se hizo con datos anteriores y totalmente distintos.

El algoritmo de árboles de decisión a pesar de ser un algoritmo muy sólido y eficiente, por sí solo no genera una respuesta satisfactoria, es cuando se combinan los resultados de Random Forest en donde se observa un resultado que da respuesta al objetivo central del proyecto.

6.6 Revisión de Procesos

6.6.1 Repaso de proceso

El proceso de formulación del proyecto, se tuvo que corregir durante el desarrollo de la segunda etapa, a fines de delimitar mejor el alcance del proyecto.

En cuanto al proceso de minería de datos, se presume que los resultados obtenidos son satisfactorios y no hay información relevante que se haya pasado por alto, así mismo, los modelos resultantes, son satisfactorios y responden a los objetivos del proyecto.

Los tres modelos construidos son satisfactorios cuando se utilizan en conjunto, sin este enfoque el único que presenta un comportamiento satisfactorio es Random Forest.

6.7 Determinación de próximos pasos

En términos de alcance, las próximas etapas del proyecto en condiciones normales, son la implementación y evaluación del desempeño de los modelos, sin embargo, por la situación sanitaria de Colombia, no es posible implementar con éxito los modelos ni evaluarlos en un escenario real, por ello el proyecto se da por finalizado y se dispone los resultados teóricos en pro de próximas investigaciones.

7. CONCLUSIONES

Las causas de retraso más influyentes en términos de tiempo en Colombia son:

- Defectos del avión asociadas a fallas producidas en la preparación del mismo o durante la operación.
- El mal manejo del espacio aéreo relacionado con los cambios de itinerario durante el trayecto que se le atribuye al manejo del espacio aéreo.
- Rotación de aeronaves el cual se debe por llegada tarde del avión en su vuelo previo. Cambio de avión asignado al vuelo, por necesidades operacionales.
- Demora en el aeropuerto de salida Condiciones climáticas adversas en el aeropuerto de origen.
- Demora en el aeropuerto de destino Condiciones climáticas adversas en el aeropuerto de destino.

Las causas más influyentes en términos de tiempo para la ciudad de Barranquilla Son:

- El mal manejo del espacio aéreo relacionado con los cambios de itinerario durante el trayecto que se le atribuye al manejo del espacio aéreo.
- Aeropuerto alterno de destino o en ruta condiciones climáticas adversas en el aeropuerto de alternativa.
- Demora en el aeropuerto de salida Condiciones climáticas adversas en el aeropuerto de origen.

- Instalaciones aeroportuarias traducidos como la no disponibilidad de posición de parqueo o puente de embarque
- Mal Posiciones de estacionamiento de aviones, y alta utilización de aeropuerto traducido como congestión en rampa.

Las causas más influyentes en términos de tiempo para la ciudad de Cali son:

- Defectos del avión asociadas a fallas producidas en la preparación del mismo o durante la operación.
- El mal manejo del espacio aéreo relacionado con los cambios de itinerario durante el trayecto que se le atribuye al manejo del espacio aéreo.
- Instalaciones aeroportuarias traducidos como la no disponibilidad de posición de parqueo o puente de embarque.
- Rotación de aeronaves que se refleja en la llegada demorada de un avión.
- Demora en el abordaje por imprevistos con los pasajeros discrepancias y compaginación, pasajeros chequeados faltantes.

Las causas más influyentes en términos de tiempo para la ciudad de Cartagena son:

- El mal manejo del espacio aéreo relacionado con los cambios de itinerario.
- Defectos del avión asociadas a fallas producidas en la preparación del mismo o durante la operación.
- Instalaciones aeroportuarias traducidos como la no disponibilidad de posición de parqueo o puente de embarque.
- Demora en el abordaje por imprevistos con los pasajeros discrepancias y compaginación, pasajeros chequeados faltantes.
- Rotación de aeronaves que se refleja en la llegada demorada de un avión.

Las aerolíneas más cumplidas son American Airlines, Delta y Aerorepublica así mismo las aerolíneas con más retraso son Satena, Aerogal y Easyfly por ultimo las aerolíneas con mayor tasa de cancelación son Tame, Air Canadá y Taca Perú.

En las ciudades de Barranquilla, Cali y Cartagena, los vuelos presentan mayores causas de retraso externas y representaron 400,739.2 minutos de retraso en las ciudades de Barranquilla Cali y Cartagena.

El algoritmo de árboles de decisión determino, que es probable que los códigos 41, 43 ,51 ,63 ,71 ,72 y 73, tendrán un retraso superior a 20 minutos, de igual forma los códigos 81, 87 y 93 tendrán un retraso estimado menor a 20 minutos.

El modelo de Random Forest estimo, que los códigos que más demoras presentarán son 81, 41 y 93. Por lo tanto, es correcto afirmar que según las estimaciones del modelo la tendencia de los datos en el año 2018 se mantiene.

El modelo de redes neuronales debe ser reentrenado con datos actuales, ya que este método actúa tomando decisiones en tiempo real y como único resultado un criterio de selección que solo se puede observar en la etapa de despliegue del

algoritmo y debe ser con data nueva ya que la operación aérea no tiene el mismo comportamiento de la base de datos de referencia.

El modelo más óptimo en términos de resultados es Random Forest, basándose en 3 criterios fundamentales, el primer criterio es el error ya que es un parámetro común entre los tres modelos y es un método cuantificable de comparación, este método obtuvo el menor porcentaje de error entre los 3 modelos inferior al 1%.

El segundo argumento es la confiabilidad del modelo en el cual los 3 modelos obtuvieron una confiabilidad superior al 90% es por ello que los modelos son indiferentes en este argumento.

El tercer punto de comparación es la inclusión de las variables, en donde Random Forest es el único modelo en donde todas las variables tienen un nivel de influencia que aporta al aprovechamiento de la información.

Sin embargo, en términos de utilización los tres modelos construidos, tienen enfoques distintos en su predicción, es por ello que el uso mixto de los mismos es el que permite dar un resultado óptimo.

Los modelos presentados, dan como resultado una predicción en la frecuencia de las demoras de acuerdo a las 10 causas más influyentes, para obtener una predicción en términos de en qué momento del año se presentarán, se debe construir un nuevo modelo especializado en la predicción de series de tiempo, utilizando diferentes técnicas a las planteadas ya que los algoritmos de inteligencia artificial contemplados en el proyecto, no son acordes para predecir series de tiempo.

8. RECOMENDACIONES

Se recomienda a próximos investigadores, desarrollar los modelos e implementarlos en condiciones normales, con el fin de evaluar los modelos y comparar los resultados de la predicción, con los resultados prácticos para decidir cambios sistémicos en caso que el modelo no tenga un desempeño optimo o decidir si se descartan los modelos.

Se recomienda la integrar la sistematización de la toma de datos, con el fin de garantizar la calidad de la información en la investigación futura, ya que se encontró perdida de información por errores de transcripción y desarticulación en los parámetros de formato, así mismo gran cantidad de sinónimos que tergiversan la información de la data.

Se recomienda en base al presente proyecto de investigación, continuar con una predicción en base a series de tiempo, con el fin de dar resultados más contundentes e información más sustanciosa.

9. Bibliografía

[1] Avianca, «Avianca,» 16 02 2020. [En línea]. Available: https://www.avianca.com/co/es/descubre-y-compra/vuela-con-nosotros/.

[2] Aeronautica Civil, «Aeronautica Civil Colombiana,» 16 02 2020. [En línea]. Available: http://www.aerocivil.gov.co.

[3] America economia, «America economia,» 16 02 2020. [En línea]. Available: www.americaeconomia.com.

[4] El Mundo, «El mundo,» 16 04 2013. [En línea]. Available: https://www.elmundo.es/america/2013/04/16/estados_unidos/1366141051.html.

[5] Eter Colombia, «eter,» 2020. [En línea]. Available: https://www.enter.co/especiales/empresas/crecen-transacciones-linea-col/.

[6] Código Civil, Artículo 1614. Daño emergente y lucro cesante, 2020.

[7] M. Ball, C. Barnhart, M. Dresner, M. H. . K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani y B. Zou, Total Delay Impact Study, 2019.

[8] RAE, «https://dle.rae.es/aeropuerto,» 2020.

[9] Microstrategy, «Modelado predictivo: la única guía que necesitará,» 19 04 2020. [En línea]. Available: https://www.microstrategy.com/es/resources/introductory-guides/predictive-modeling-the-only-guide-you-need.

[10] A. S. Rico, «Universidad Autonoma de Barcelona,» 09 Febrero 2017. [En línea]. Available: https://ddd.uab.cat/pub/tfg/2017/181125/SerranoRicoAitor-TFGAa2016-17.pdf.

[11] SAS, «Minería de datos,¿Que es y porque es importante?,» 19 04 2020. [En línea]. Available: https://www.sas.com/es_co/insights/analytics/data-mining.html#dmtechnical.

[12] C. C. Regin, The Comparative Method: Moving beyond Qualitative and Quantative Strategies, 1987.

[13] ASOCIACIÓN ESPAÑOLA PARA LA CALIDAD, «ESTADÍSTICA APLICADA,» 19 04 2020. [En línea]. Available: https://www.aec.es/web/guest/centro-conocimiento/estadistica-aplicada.

[14] Weebly, «¿QUÉ SON RECURSOS INFORMÁTICOS?,» 04 2020. [En línea]. Available: https://recursoinformatico.weebly.com/inicio/que-son-recursos-informaticos.

[15] Hosteur, «Identifican las cinco causas fundamentales de los retrasos aéreos,» 2020. [En línea]. Available: https://www.hosteltur.com/57156_identifican-cinco-causas-fundamentales-retrasos-aereos.html.

[16] M. D. Nerea, «Predicción y Análisis de los Retrasos en los vuelos,» Barcelona, 2016.

[17] EUROCONTROL, «Retrasos: tres preguntas y muchas respuestas,» 3 Agosto 2018. [En línea]. Available: https://www.eurocontrol.int/news/delays-three-questions-and-many-answers.

[18] DIAN, «Bases de datos abiertas,» 2020. [En línea]. Available: https://www.dian.gov.co/atencionciudadano/Paginas/Datos-Abiertos.aspx.

[19] CODIGO DE COMERCIO, DECRETO 410 DE 1971, 2020.

[20] Sinnexus, «Minería de datos,» 2020. [En línea]. Available: https://www.sinnexus.com/business_intelligence/datamining.aspx.

[21] G. Ligdi, «ligdigonzalez,» 2019. [En línea]. Available: https://ligdigonzalez.com/aprendizaje-supervisado-random-forest-classification/.

[22] atria innovation, «https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/#:~:text=Como%20se%20ha%20mencionado%20el,a%20un%20nodo%20llamado%20neurona.&text=Cada%20una%20de%20las%20neuronas,que%20modifica%20la%20entrada%20recibida.,» 17 11 2020. [En línea]. Available: https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones/#:~:text=Como%20se%20ha%20mencionado%20el,a%20un%20nodo%20llamado%20neurona.&text=Cada%20una%20de%20las%20neuronas,que%20modifica%20la%20entrada%20recibida..

[23] R ORG, «ntroducción a las Redes Neuronales mediante el,» 10 11 2020. [En línea]. Available: http://madrid.r-es.org/wp-content/uploads/2016/01/Redes_Neuronales_01.pdf.

[24] R-project, 2020. [En línea]. Available: https://www.r-project.org/.

[25] kdnuggets, «Best Python modules for data mining,» 2020. [En línea]. Available: https://www.kdnuggets.com/2012/11/best-python-modules-for-data-mining.html.

[26] P. Rochina, «Inesem,» 2020. [En línea]. Available: https://revistadigital.inesem.es/informatica-y-tics/python-r-analisis-datos/.

[27] International Business Machines Corporation , Manual CRISP-DM de IBM SPSS Modeler, GSA ADP, 2012.

[28] Deloitte, «¿Qué es Power BI?,» [En línea]. Available: https://www2.deloitte.com/es/es/pages/technology/articles/que-es-power-bi.html.

[29] Dinero, «Los mejores aeropuertos de Latinoamérica: ¡ya no somos los primeros!,» 2020. [En línea]. Available: https://www.dinero.com/pais/articulo/mejores-aeropuertos-de-america-latina-y-el-mundo/269100.

[30] N. Belgorodski, G. Matthias, T. Kristin, S. Katharina, F. Matthias y L. Göhring, «Fitting Distributions to Given Data or Known Quantiles,» R Projetc, 2017.

[31] R Proyect, «Git Boocks,» 10 11 2020. [En línea]. Available: https://rsanchezs.gitbooks.io/ciencia-de-datos-con-r/content/paquetes/paquetes.html.

[32] Logística, «El transporte aéreo en Colombia sigue en alza,» *Revista Logistica: Supply Chain Industri,* 2018.

[33] O. Díaz Olariaga y J. F. Zea, «Influence of the liberalization of the air transport industry on configuration of the traffic in the airport network,» *Science Direct,* 2018.

[34] USC, «USC.gal/gl,» 2020. [En línea]. Available: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140128_RegresionMultiple.pdf.

[35] SAP, «Potenciación del gradiente de árboles de decisiones,» 2019. [En línea]. Available: https://help.sap.com/viewer/feae3cea3cc549aaa9d9de7d363a83e6/2002/es-ES/a2a3eada1d954aeaaaa8259ac2720767.html.

10. ANEXOS

Anexo 1: Código de demora IATA

Anexo 2: Dashboard de gráficos

Anexo 3: Código Modelo Red neuronal

Anexo 4: Código Modelo Random Forest

Anexo 5: Código Modelo Arboles de decisión

Para acceder a estos anexos dirigirse al siguiente link:

https://drive.google.com/drive/folders/1_n3zmMaCjToR-C_P5aKXgS6JjAInbmwG?usp=sharing