Flights Delay Prediction

Integrantes:
Marcos Antonio Ochil Trujillo C-412
Davier Sánchez Bello C-412
Maykol Luis Martínez Rodríguez C-412

July 7, 2024

1 Introduccion

La aviación ha experimentado un crecimiento exponencial en las últimas décadas, convirtiéndose en uno de los medios más eficientes y rápidos de transporte global. Sin embargo, este avance no está exento de desafíos, siendo uno de los más significativos el problema de los retrasos de vuelo. Los retrasos pueden tener consecuencias directas e indirectas, afectando tanto a los pasajeros como a las aerolíneas y, por ende, al sistema económico global.

1.1 Motivaciones

A pesar de los numerosos estudios y modelos desarrollados para predecir el retraso de los vuelos, la mayoría de estos se basan en datos complejos y detallados, los cuales son difíciles de obtener en la práctica. La literatura existente sugiere que la predicción precisa de los retrasos de vuelo requiere modelos altamente sofisticados y datos exhaustivos, lo cual rara vez se logra en situaciones reales. Esta brecha entre la teoría y la aplicación plantea una oportunidad para mejorar la precisión y accesibilidad de las predicciones de retraso de vuelo.

Además, hemos decidido continuar el trabajo presentado en "An Intelligent Approach for Flight Delays Prediction" de Mahmout y Ezzas, adoptando su enfoque basado en Time Flight Delay (TFD) y Previous Flight Delay (PFD), TFD es la demora entre las operaciones y PFD es la demora del vuelo anterior. Este enfoque nos permite explorar la posibilidad de utilizar la menor cantidad de datos posible, enfocándonos en características medias que representen el comportamiento típico de aeropuertos y aerolíneas.

1.2 Problematica

Las principales problemáticas identificadas en el ámbito de la predicción de retrasos de vuelo incluyen: - Complejidad de los Modelos: Muchos modelos existentes requieren grandes cantidades de datos detallados, lo cual puede ser impracticable en entornos operacionales. - Acceso a Datos: La disponibilidad de datos de alta calidad y relevancia para entrenar modelos de machine learning es limitada. - Saturación de Aeropuertos: El fenómeno de la saturación de aeropuertos, caracterizado por el aumento de tráfico y la falta de capacidad, no se considera adecuadamente en la mayoría de los modelos actuales.

1.3 Objetivos generales y especificos o hipotesis o preguntas cientificas

2 Estado del arte/Preliminares

El objetivo básico de nuestro proyecto es crear un sistema capaz de predecir los retrasos de los vuelos en un sistema de aviación. Siguiendo la ideología newtoneana, este trabajo se ha apoyado en los hombros de los gigantes con el fin de ver más lejos. Para este fin, se han recurrido a numerosos artículos relacionados con este tema con el propósito de apoyarse en sus aportes, corregir sus errores y desarrollar sus ideas hasta lograr desarrollar nuestra propia solución que, si bien es absurdo y presuntuoso llamar concluyente, sí podemos decir con confianza que constituye una nueva y sólida basa sobre la cual otros, como nosotros, puedan apoyarse para ver aún más lejos en este largo y potencialmente infinito camino de estudio e investigación.

En los artículos estudiados se observan una multitud de algoritmos y enfoques para enfrentar el problema,muchos de ellos incluso opuestos entre sí. Un ejemplo de esto es el hecho de centrarse en clasificar los vuelos en función de si tienen retraso o no retraso basándose en si superan o no cierto rango de tiempo como hacen artículos como ¨A CNN-LSTM framework for flight delay prediction¨ publicado por Jingyi Qu, Shixing Wu y Jinjie Zhang el 17 de enero del 2023, o si se enfocan más en simplemente medir el tiempo que se alejó del horario planificado independientemente de si esa diferencia se considera retraso o no, como ¨Flight Delay Regression Prediction Model Based on Att-Conv-LSTM¨ publicado por Jingyi Qu, Min Xiao, Liu Yang and Wenkai Xie el 8 de mayo del 2023. Estas pueden considerarse las dos categorías en que se divide el problema de predicción de vuelos, naturalmente algunos artículos analizan ambas en profundidad, unos pueden considerarse modelos de clasificación en niveles de retraso para las cuales de puede variar el umbral hasta obtener el rango que presente la mayor precisión, mientras que los modelos de predicción de regresión de tiempos de retraso puede predecir tiempos de retraso específicos proporcionando una orientación más granulada para la aplicación práctica en los sectores relevantes.

Otro aspecto a tener en cuenta es la viabilidad de implementación de las soluciones planteadas, teniendo en cuenta la alta complejidad computacional de los algoritmos de predicción de retrasos de vuelos existentes, que no son propicios para su implementación en dispositivos móviles y otros dispositivos sin el poder de cómputo necesario. Este es un problema abordado en el trabajo ¨Flight Delay Prediction Model Based on Lightweight Network ECA-MobileNetV3¨ publicado por Jingyi Qu, Bo Chen, Chang Liu y Jinfeng Wang el 17 de marzo del 2023 que propone un algoritmo ligero mejorado

ECA-MobilenetV3, que reemplaza el modelo SE con un módulo liviano ECA (Efficient Channel Attention), reduciendo efectivamente la complejidad computacional del modelo sin perder precisión; sienta las bases para la aplicación del modelo en dispositivos móviles.

Hay algunos que se centran en los algoritmos de aprendizaje utilizados, buscando con cual se obtiene el mejor resultado como es el caso del artículo ¨Flight Delay Classification Prediction Based on Stacking Algorithm" publicado por Jia Yi, Honghai Zhang, Hao Liu, Gang Zhong, y Guiyi Li el 18 de agosto del 2021 en el que se utiliza el algoritmo de apilamiento, se disponen para ello alumnos que utilizan varios algoritmos de ml sobre el mismo conjunto de datos a fin de explorar la estabilidad del algoritmo de apilamiento. Por el contrario hay trabajos como ¨ON THE PERFORMANCE OF MACHINE LEARNING BASED FLIGHT DELAY PREDICTION – INVESTIGATING THE IMPACT OF SHORT-TERM FEATURES¨ publicado por Schösser D, y Schönberger J. el 8 de julio del 2022 que solo utilizan algoritmos estándar y sin mejorar como Random Forest (RF) y XGBoost basados en árboles, así como el algoritmo Neural Network(NN) basado en aprendizaje profundo, y en su lugar están centrados en la investigación de la influencia de las características a corto plazo en la calidad de los resultados, para lo cual se crean diferentes escenarios que se caracterizan por diferentes conjuntos de características de entrada.

También se destacan algunos estudios que, aunque innovadores, no son factibles dada la naturaleza de los datos que se requieren para implementarse, un ejemplo de esto es el artículo ¨Predicting Flight Delay with Spatio-Temporal Trajectory Convolutional Network and Airport Situational Awareness Map" en el cual se hace énfasis en la predicción del retraso de la salida pero se demanda para ello información del estado de la pista para generar un mapa de conciencia situacional, lo cual obviamente no está a nuestro alcance.

Entre los conocimientos adquiridos tenemos:

Técnicas de Aprendizaje Automático y Profundo

• Aprendizaje Espaciotemporal: Los modelos basados en aprendizaje automático han sido adaptados para capturar la naturaleza espaciotemporal de los datos de vuelos, incluyendo la ubicación geográfica de los aeropuertos y los patrones temporales de operaciones y retrasos [1].

• Propagación de Retrasos: Algunos estudios se han centrado específicamente en predecir cómo los retrasos en un vuelo pueden afectar a otros vuelos conectados, utilizando redes neuronales profundas para modelar estas dinámicas complejas [3].

Integración con Tecnologías de Voz Domésticas

• Interfaz de Usuario Amigable: La integración de sistemas de predicción de retrasos con tecnologías de voz domésticas, como Amazon Alexa, ha demostrado ser una forma efectiva de hacer que la información sea accesible para los usuarios de manera rápida y eficiente [2].

Mejora Continua y Perspectivas Futuras

• Evaluación y Comparación: Los modelos propuestos han sido evaluados comparativamente con métodos tradicionales, mostrando una mejora significativa en la precisión de las predicciones de retrasos de vuelos [1, 2, 3].

• Expansión de Aplicaciones: Existe un potencial considerable para expandir estos enfoques a otras áreas donde la predicción precisa y la adaptabilidad sean críticas, como el transporte público, el tráfico urbano, y la logística [2].

[1] Spatiotemporal Propagation Learning for Network-Wide Flight Delay Prediction

[2] Alexa, Predict My Flight Delay

[3] Flight Delay Propagation Prediction Based on Deep Learning

Para ahondar más en los artículos mencionados, se hizo un breve resumen de estos en la carpeta de State of Art. Nuestra conclusión fue que dado el gran número de posibles enfoques y consideraciones, intentar abarcarlo todo dará como resultado un trabajo superficial incapaz de profundizar en nada en concreto, por lo que en su lugar optamos por un enfoque específico entre los disponibles que se adaptó más a nuestros intereses.

3 Propuestas de solucion

Nuestra propuesta de solucion, sigue la tonica del paper 'A novel intelligent approach for flight delay prediction'. En ese paper se proponen predecir tardanzas con un uso de datos minimal, emulando la realidad de los planificadores de vuelos, que conocen no mucho mas alla respecto a los vuelos futuros que sus fechas de partida y de llegada. Su enfoque, al igual que el nuestro, va hacia un preprocesado especial de la data que permite enfatizar ciertas verdades importantes del Conocimiento del Negocio y parte, al igual que nosotros de la asuncion de que los retrasos de vuelos pueden predecirse mejor si seguimos la ruta de cada avion de manera indivual. Pero los autores del paper dejan sin cubrir campos importantes, respecto a la manera en que la data categorica debe ser codificada. Ademas, creemos que los autores no explotaron todas las posibles fuentes de datos que podrian estar disponibles para planear los vuelos con la antelacion suficiente con mayor precision.

En resumen, nuestra solucion es la siguiente: Procesado de data: Partimos de los datos historicos de vuelos de aviones. Solo necesitamos los siguientes features:

'TAIL NUMBER' : numero de cola del avion

'ORIGIN AIRPORT' : aeropuerto de origen

'DESTINATION AIRPORT' : aeropuerto de llegada

'AIRLINE' : aerolinea

'SCHEDULED DEPARTURE' : hora de salida planificada

'DEPARTURE TIME' : hora de salida acontecida

'DEPARTURE DELAY' : retraso en la salida

'SCHEDULED ARRIVAL' : hora de llegada planificada

'ARRIVAL TIME' : hora de llegada acontecida

'ARRIVAL DELAY' : retraso en la llegada

'AIRLINE DELAY' : cual parte del retraso fue provocada por la aerolinea

'WEATHER DELAY' : cual parte del retraso fue provocada por el tiempo

'AIR SYSTEM DELAY' : cual parte del retraso fue provocada por el sistema aereo

'SECURITY DELAY' : cual parte del retraso fue provocada por motivos de seguridad

'LATE AIRCRAFT DELAY' : cual parte del retraso fue provocada por retraso del avion en llegar.

La parte de la data que describe a que se debio cada porcion del retraso, no se usa directamente para caracterizar cada datapoint y entrenar al modelo,

sino para caracterizar aerolineas, vuelos y fechas temporales, para realizar un encoding de ellas.

La limpieza de la data la abordaremos en las particularidades de nuestra implementacion, ya que al tratar de replicar lo que hicimos pueden presentarse retos distintos. Una vez que se tienen todos los vuelos limpios con todos los campos relevantes, dividimos cada vuelo en dos datapoints, la departure y el arrival. Procedemos luego a ordenarlos por numero de cola del avion y como segundo criterio de ordenamiento por la fecha planificada de ocurrencia. De esta manera, podremos para cada avion modificar una a una las filas incluyendo data relativa a sus vuelos anteriores que creamos necesaria. Ahora calcularemos, al igual que en el paper de Mahmout y Ezzat, el retraso de la operacion anterior realizada por el avion y el tiempo previsto a transcurrir entre la operacion anterior y la actual, llamados PFD y TFD, parametros que ya fueron explicados anteriormente. Llega el momento de codificar la data categorica. Ahora mismo los datos categoricos a codificar son los siguientes:

'TAIL NUMBER'
'ORIGIN AIRPORT'
'DESTINATION AIRPORT'
'AIRLINE'

Queremos ademas codificarlo de manera que en cada datapoint quede incluida la mayor cantidad de informacion relevante a la fecha de manera que podemos luego eliminarla, y cada datapoint sea independiente temporalmente, o sea, nuestro modelo no necesite tener los datapoints en orden para realizar una prediccion (manteniendo el modelo simple). Los encodings usados en los papers que pudimos leer fueron One-Hot-Encoding y Label Encoding. No nos gusto ninguno de los dos parza nuestro modelo. One Hot Encoding eleva demasiado la dimensionalidad, y la cantidad de columnas que tenemos con informacion relevante paara cad datapoints no nos parecen la suficiente como para defenderse semejante gula dimensional (Hay 23 aerolineas en el dataset que usamos pero en la vida real puediera ser mucho mas, por no hablar de la cantidad de aeropuertos). Por otro lado eso conllevaria complejizar en demasia el modelo. Label Encoding no nos gusta, pues introduce un sesgo de ordinalidad ajeno al conocimineto que tenemos del negocio. En cambio escogimos usar Target Encoding, tecnica que sin elevar la dimensionalidad de la matriz pretende reflejar la influencia de cada categoria sobre la data. Target Encoding es la sustitucion de cada categoria por el promedio que alcanza la variable prevista en los datapoints que

toman su valor. Nosotros, sin embargos no decidimos usar el promedio. Nos parece que el promedio de los delays no es una variable lo suficientemente expresiva, ya que da tanto peso a las categorias que producen pocos retrasos pero de larga duracion como a las categorias que retrasan mucho pero cortos periodos. Ademas, dado que la mayoria de los vuelos no se retrasan o retrasan muy poco o cantidad negativa se distorisionaria la data y ademas esta estadistica no nos permite trazar una linea clara entre retrasos reales (aquellos mayores a 15 minutos) y retrasos por ejemplo de 14 o 13 minutos que no cuentan realmente como retrasos acorde al sistema estadounidense de aviacion. En cambio, nuestro target encoding atendiendo a las limitacines del target encoding promediado, sustituye cada columna categorica a la que se lo aplicaremos por dos columnas: la proporcion de los vuelos con esa categoria que retrasan, y el promedio de los retrasos, cuando sea que estos ocurren, de manera que se incluyen tanto los datos de la cantidad de retrasos y promedio de estos cuando son, sin distorsionarse ni afectarse mutuamente. De hecho, comprobamos la covarianza entre ambas variables en las ocasiones en que la usamos, siempre encontrando un valor negativo medio (0.6, 0.5), de manera que no introducimos colinearidad en la data al usarlas. Vayamos caso a caso con el encoding

Aerolineas:

Sustituimos el nombre de cada aerolinea por la proporcion de los vuelos de esa aerolinea que se retrasan por causas inducidas por la propia aerolinea, asi como por el promedio de duracion de tales retrasos.

Aeropuertos:

Para codificar los aeropuertos, empleamos target encoding, pero no usamos solo dos columnas, sino 5, debido a que los aeropuertos agrupan en si varios factores que influyen directamente en el retraso de los aviones. Esta codificacion nos permite ademas, deshacernos de las columnas temporales for good. Sin mas adentremonos en ellas. La calidad operacional de los aeropuertos varia, por diversos motivos, como el disenho del aeropuerto o el interes socioneconomico en el que permita o no modernizarlo. La calidad operacional de un aeropuerto es independiente de la etapa del anho, y viene expresada por los vuelos que son explicitamente categorizados como retrasados por el Sistema de Control del aeropuerto. Dada la atemporalidad de esta data, podemos describirla de la misma manera que describimos los retrasos inducidos por las aerolineas, como la proporcion de los vuelos retrasados por el sistema de control del aeropuerto y cuanto duraron en promedio esta clase de retrasos. La localizacion del aeropuerto, influye a traves del factor cli-

matico. El clima varia de mes a mes de formas diferentes en aeropuertos diferentes, razon por la cual esta data no es atemporal. Luego, para cada mes y cada aeropuerto computamos la proporcion de retrasos taggeados como relacionados con el clima y el promedio de la duracion de estos. Un ultimo parametro relacionado con los aeropuertos y que no queremos dejar fuera, y cuya inclusion consideramos novedosa de nuestro enfoque es la saturacion que presenta el aeropuerto en el momento de realizar la operacion en cuestion. Definamosla. Separando los eventos del dataset en intervalos de 5 minutos poodemos saber para cada aeropuerto cuantos aviones hay realizando operaciones de despegue o aterrizaje en este preciso momento. La saturacion de un momento se define como cuan a tope esta de aviones el aeropuerto y es el cociente de la cantidad de aviones que realizan operaciones en el aeropuerto en el momento en cuestion dividida por el maximo valor que ha tenido tal estadistica. Pero no quisimos dejarlo asi, y calculamos la saturacion para una operacion X como una ponderacion de cuan saturado estaba el aeropuerto en el intervalo de 5 minutos el momento de la operacion en cuestion y la saturacion en algunos otros intervalos alrededor del intervalo, sobre todo los anteriores y el inmediatamente posterior (este ultimo en el caso de que el vuelo nos quedara casi en el limite de su intervalo). Se pudiera pensar que este dato no esta facilmente disponible para entrenar el modelo, o que mucho menos para predecir un dato que esta lejano en el futuro. Pero por el contrario. La satruracion para caa aeropuerto en cada intervalo de 5 minutos es trivial de calcular para vuelos ya ocurridos si se dispone de la data historica. Por otro lado, para vuelos en el futuro, se pueden adoptar dos enfoques para el calculo de la saturacion. El primero de ellos y mas naural parte del hecho de que la prediccion de los vuelos futuros se realizarta uno a uno, en orden de su fecha de salida planificada. A medida que se vayan prediciendo sus retrasos se puede mantener un estimado de cuanto esperamos que se retrase y por tanto de cuando esperamos que lleguie a su aeropuerto de destino, por lo que bastaria mantener los datos para cada aeropuerto de cuantos aviones tenemos previsto que lleguen o salgan en cada intervalo de tiempo. Si por otro lado, se quisiera predecir los futuros retrasos para un solo avion, sin tener en cuenta ningun otro vuelo, no seria posible tener calculado el dato de la saturacion del aeropuerto, pero dado que la cantidad de aviones entrando y saliendo del aeropuerto en cada intervalo de tiempo es una serie de tiempo clasica, para aeropuertos muy grandes pudiera ser predicha muy eficientemente, por lo cual este dato tambien estaria accesible si alimentaramos el modelo de prediccion de series de tiempo con la data del aeropuerto en cues-

tion. De hecho, lo intentamos con prophet, pero no obtuvimos resultados satisfactorios, dado que los intervalos de tiempo de 5 minutos parecen ser un intervalo muy corto.

Este parametro saturacion, realmente nos ahorra tener que dar al modelo datos como el mes, el dia y la hora, ya que, en lugar el modelo tener que buscar la relacion relacionada entre las fechas y los retrasos directamente le estamos dando la saturacion, causa directa de los retrasos y directamente relacionada con la hora, asi como anteriormente le habiamos dado la data climatica que abstrae la relacion de los meses y el clima. Este parametro de la saturacion es especialmente importante para los retrasos de vuelos, y no leimos ningun paper que hiciera uso de el o mencionara su uso en alguna referencia al historia del arte. Falta por mencionar respecto a la codificacion de los aeropuertos, que hay aeropuertos muy grandes y aeropuertos muy pequenhos. Para los aeropuertos pequenhos calcular los valores de proporcion de retrasos y promedio de retrasos es un riesgo, porque se tiene poca data de ellos y el resultaqdo muy probablemente este distorsionada. Por tanto decidimos establecer un threshold tal que, para los aeropuertos que tuvieran un numero de vuelos mayor que el threshold fueran calculadas estas estadisticas de manera individual, mientras que para todos los que tuvieran un numero de vuelos menor que este, calcular su valor como conjunto y asignarle el resultado a todo. Para seleccionar el threshold procedimos a la inspeccion visual de graficos de dispersion realizados sobre la data y el numero de vuelos disponibles para cada aeropuerto, y mediante la inspeccion visual pudimos distinguir claramente un limite a partir del cual los valores se desequilibraban. Este limite fue de 25000 vuelos para el calculo de las variables promediadas anuales, como es el caso de la proporcion y duracion promedio de los retrasos para cada aeropuerto provocados por el propio aeropuerto, asi como 2500 para los datos mensuales de proporcion y promedio de vuelos retrasados por el clima. Tiempo Creemos que la mayor parte si no toda la influencia del tiempo queda correctamente descrita, con las sustituciones que hicimos relativas al clima y la saturacion de los aeropuertos, puesto que, intuitivamente que otra razon podria haber, la industria aeropuortuaria y mas a la escala donde estamos trabajando (solo con las aerolineas quye transportan maws del 1La variable tail number no la necesitamos durante el entrenamiento luego de la codificacion de la data, ya que no se la pasaremos al modelo. Habra otro campo llamado ORIENTATION que contiene el tipo de operacion a ralizar DEPARTURE o ARRIVAL, que se usara para separar la data en caso de que se quiera entrenar un modelo para las departure y otro para los arrival (lo

cual fue en efecto lo que hicimos). Al concluir la parte de limpieza de la data y Feature Engineering quedaran para entrenar al modelo y para describir a los futuros datapoints a describir los siguientes features:

TFD

PFD

SATURATION : saturacion del aeropuerto donde se produce la operacion en cuestion

AIRLINE DELAY PROB : proporcion de retrasos inducidos por la aerolinea en cuestion

AIRLINE AVG DELAY : promedio de duracion de los retrasos inducidos por esta aerolinea

ORIGIN AIRPORT DELAY PROB : proporcion de retrasos inducidos por el aeropuerto de partida

ORIGIN AIRPORT AVG DELAY : promedio de duracion de los retrasos inducidos por el aeropuerto de partida

ORIGIN WEATHER DELAY PROB : proporcion de retrasos inducidos por el clima en el aeropuerto de salida en el mes en cuestion.

ORIGIN AVG WEATHER DELAY : promedio de duracion de los retrasos inducidos por el clima en el aeropuerto de partida en el mes en cuestion

DESTINATION AIRPORT DELAY PROB : proporcion de retrasos inducidos por el aeropuerto de llegada

DESTINATION AIRPORT AVG DELAY : promedio de duracion de los retrasos inducidos por el aeropuerto de llegada

DESTINATION WEATHER DELAY PROB : proporcion de retrasos inducidos por el clima en el aeropuerto de salida en el mes en cuestion

DESTINATION AVG WEATHER DELAY : promedio de duracion de los retrasos inducidos por el clima en el aeropuerto de llegada

Prediccion de resultados de vuelos futuros:

El modelo que se use (los modelos que probamos dan muy buenos resultados), estara disenhado para dadas las categorias mencionadas, predecir el delay del vuelo, independientemente del modelo en que se haya producido el vuelo o de la secuencia de vuelos precedente. De hecho, de eso se trata el procesamiewnto de la data, de embeber en cada datapoint toda la informacion relativa a los eventos temporales de manera que se vuelva independiente. Por tal motivo pudimos probar el modelo con datapoints aleatorios separados de la data justo al iniciar. Pero el hecho de que hayamos abstraido la temporalidad, no implica que nuestro modelo no sea capaz de predecir series de vuelos, por el contrario, esta disenhado para ello. Como predeciriamos,

por ejemplo, el retraso de todos los vuelos del mes siguiente?. Primero que todo, hariamos un preprocesaddo sobre lo que sabemos de esos vuelos, sus fechas de salida y llegada y sus aeropuertos de salida y llegada y la aerolinea. Al igual que en el preprocesado original, los dividiriamos en dos operaciones y luego los ordenariamos por numero de cola del avion y por fecha. Dado que nuestro modelo fue entrenado con la data de todo un anho, tenemos los datos climaticos necesarios y las demas datas categoricas ya sabemos como codificarlas. Solo dejariamos el feature Saturacion y el feature PFD sin calcular (el TFD puede ser calculado usando las horas previstas solamente dato que tendriamos desde el principio). Nos ampararíamos en la data que ya tenemos sobre el ultimo vuelo realizado por el avion en cuestion para calcular el TFD y el PFD para la primera operacion que aparezca para cada avion en la data a predecir. Si el avion no lo teniamos en nuestro dataset anterior, sencillamente asignamos cero a ambos campos. Por otro lado para calcular la saturacion, mantendremos siempre la informacion de cuantos aviones estan realizando una operacion en cada instante de tiempo en cada aeropuerto, partiendo de que tal data que ya fue computada para el dataset historico y la iiremos incrementando vuelo a vuelo a medida que realicemos las predicciones en orden cronologico. O sea tanto el PFD como la saturation de cada data point a predecir se calculan en el momento de su prediccion, a partir de los datos historicos precomputados incrementados con las nuevas predicciones que hemos realizado. De esta manera se puede predecir hasta N vuelos despues de la ultima fecha en los datos historicos, con la correspondiente perdida de precision a medida que avanza el tiempo por los errores en calculos de delay acumulados, que sin embargo estara amortiguada, por el hecho de que la programacion de los vuelos, esta disenhada de tal manera que muy rapido un avin alcanzara una pausa larga donde los delays se resetean por decirlo de alguna manera, por lo que los errores en predicciones de delay quedaran acotados por las pausas 'largas' que se toman los aviones. No era nuestro objetivo primario probar nuestro modelo bajo estas situaciones, por motivos de tiempo principalmente, sino presentar la introduccion de los nuevos features y el como era posible obtener muy buenos resultados de precision usando estos features. No obstante, remarcamos, la eficacia de esta forma de predecir vuelos estara dada mayormente por el modelo para predecir datapoints inividuales independientes, y este si pudimos probarlo y obtuvo resultados muy buenos.

Nos propusimos seguir los vuelos realizados por cada avion en particular, de manera que a la hora de predecir el retraso de una operacion aeropor-

turaria, debemos conocer (o sustituir el valor por una prediccion previa que hayamos realizado) cual fue el retraso del avion en su ultima operacion, y ademas cuan saturado estara el aeropuerto donde se realiza la operacion. Pero empece Entonces, como sustituir las datas categoricas. Escogimos el target encoding, que consideramos mas adecuado a nuestro proposito de mantener el modelo simple, asi como lo empleamos de manera que pudimos introducir data propia del conocimiento del negocio al hacerlo de la manera particular en que lo hicimos. El target encoding es una tecnica de procesamiento de datos donde se sustituye la variable categorica en cuestion, por el valor promedio obtenido por la variable objetivo para esa categoria. Evita disparar la dimensionalidad sin dejar de expresar la influencia de la variable a sustituir sobre los datos en general. Sin embargo dada la naturaleza del problema decidimos no usar el promedio como metrica en ninguna de las categorias en que realizamos el target encoding. Esto se debe a que el promedio de los delays no nos parece la medida mas exacta para representar la influencia que tiene un aerpuerto o una Aerolinea sobre el retraso de los vuelos, ya no podriamos diferenciar entre la frecuencia de los retrasos y la duracion de estos. En su lugar, los targets encoding que usamos convierten las columnas categoricas en dos columnas, uno para la proporcion de retrasos para la categoria correspondiente y la otra para el promedio de duracion de sus retrasos. Las relaciones temporales las sustituimos completamente, extrayendo la informacion relevante al momento en que ocurrieron las operaciones y colocandola en cada datapoint, evitando asi al modelo que usemos tener que seguir una sucesion de vuelos y siguiendo nuestra filosofia de mantener el modelo sencillo a traves de un pre procesamiento

Usamos el dataset provisto en el siguiente enlace https://www.kaggle.com/datasets/usdot/flight-delays, porque en

Nos propusimos realizar un PreProcesamiento de la data que nos permitiera mantener el modelo simple, y que fuera lo mas expresivo relativo a los datos disponibles sobre cada vuelo con la anterioridad que sea. Nuestro modelo, sabiendo apensas unos features que estan disponibles previamente al vuelo, puede predecir el retraso del vuelo en cuestion, y a la hora de predecir un largo volumen de vuelos futuros, cada vez que predizcamos el retraso de uno de los vuelos, este se tomara en cuenta para calcular el retraso de los proximos. De esta manera, se pueden predecir hasta n vuelos futuros, pudiendo ser n tan grande como se quiera, solo sabiendo su origen y su destino y sus horarios previstos de salida y de llegada.

Kaggle (https://www.kaggle.com/datasets/usdot/flight-delays)

4 Experimentos y resultados

Se realizaron experimentos para evaluar la factibilidad de la introduccion de los datos ya mencionados, para esto se entrenaron diversos modelos con el 80 porciento de los datos con los que se contaba y se realizo test con el 20 porciento restante.

Los modelos utilizados fueron Linear Regression (LR), Random Forest (RF), Gradient Boosting (GB), KNN y Ridge Regression (RR), las métricas utilizadas fueron el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE) y la precisión con respecto a tipos de retraso, definiendo como resultado positivo si la medición real y la predicción son menores o mayores al mismo tiempo que un valor k (Accuracy ¡ k).

Métrica	LR	RF	GB	KNN	RR
MAE	13.36	11.15	17.44	12.6	13.36
MSE	670.43	528.13	1075.13	688.44	670.43
Accuracy (< 15)	88.96	87.38	83.85	89.28	88.96
Accuracy (< 60)	96.48	97.05	95.03	96.95	96.48
Accuracy (< 90)	97.83	98.41	97.26	98.27	97.83

Como se puede observar los modelos presentan un error de poco más de 10 minutos de predicción, sin embargo la identificación del tipo de retraso es muy efectivo y permite estar preparado en una gran cantidad de casos.

Pero queriamos identificar si la separación de la data en distintas clases nos permitia obtener mejores resultados, por lo que separamos los datos en "Arrivals" y "Departures" y analizamos los resultados en todos los modelos probados.

	LR		RF		GB		KNN		RR	
Métrica	Arrival	Departure	Arrival	Departure	Arrival	Departure	Arrival	Departure	Arrival	Departure
MAE	9.00	14.20	9.30	12.38	18.52	16.39	11.87	10.95	9.00	14.20
MSE	163.88	915.58	176.02	840.55	1098.09	1041.61	283.81	870.36	163.88	915.58
Accuracy (< 15)	93.20	84.67	92.36	82.86	84.21	84.39	92.05	88.57	93.20	84.68
Accuracy (< 60)	98.68	95.06	98.55	95.86	95.06	95.08	98.37	96.53	98.68	95.06
Accuracy (< 90)	99.33	97.21	99.27	97.76	97.23	97.28	99.14	97.98	99.33	97.21

Luego de separar en clases decidimos profundizar en los hiperparametros de los modelos utilizados.

En el caso de *Random Forest* se aumentaron la cantidad de estimadores:

Métrica	10 Estimadores	20 Estimadores
MAE	11.15	10.82
MSE	528.13	503.42
Accuracy (< 15)	87.38	87.88
Accuracy (< 60)	97.05	97.2
Accuracy (< 90)	98.41	98.49

Por otro lado en *Gradient Boosting* se aumentaron distintos hiperparametros por separado:

Métrica	Normal	Iteraciones	Profundidad	Aprendizaje
MAE	17.44	16.65	17.4	14.65
MSE	1075.1	1000.38	1091.29	790.86
Accuracy (< 15)	83.85	86.38	84.05	87.97
Accuracy (< 60)	95.03	95.06	95.03	95.65
Accuracy (< 90)	97.26	97.27	97.26	97.35

Finalmente en KNN se aumentaron la cantidad de vecinos observados:

Métrica	5 Vecinos	10 Vecinos
MAE	12.6	12.02
MSE	688.44	687.93
Accuracy (< 15)	89.28	89.26
Accuracy (< 60)	96.95	96.96
Accuracy (< 90)	98.27	98.29

5 Discusión de los resultados

El trabajo con distintos tipos de métricas (métricas de error o precisión) nos resultó necesario dado el estudio de trabajos similares donde normalmente se presentaba uno solo de estos, imposibilitando totalmente una comparación de resultados entre trabajos (si bien es cierto que muchos se realizan con distintas bases de datos, la diferenciacion de métricas separaba más esta comparación).

Al abrir los resultados en las clases Arrivals y Departures se pudo identificar que es más complicado para nuestro modelo determinar el retraso en las entradas a un aeropuerto, por ende estos son los datos que hacen que las métricas con toda la base de datos (Primera Imagen) aumenten los errores y disminuyan la precision con respecto a las salidas.

Por otro lado los cambios de hiperparametros representaron un mayor costo computacional pero no reflejaron un gran cambio en las métricas utilizadas (excepto en el caso de Gradient Boosting) por lo que no se considera un experimento factible para el proceso.