Dibujo en blanco y negro

Descripción generada automáticamente con confianza baja

Dibujo de un animal

Descripción generada automáticamente con confianza baja

Proyecto N° 2

ANÁLISIS PREDICCIONES DE VUELO

Integrantes:

López, Valeria

Peña, Zunida

Profesor:

Tobar, José Miguel

Cátedra:

Data Analytics

**TABLA DE CONTENIDO**

[INTRODUCCIÓN 3](#_Toc89539024)

[I. DESCRIPCIÓN DEL PROBLEMA DE NEGOCIO 4](#_Toc89539025)

[II. SUPUESTOS UTILIZADOS 7](#_Toc89539026)

[III. TÉCNICA O TÉCNICAS APLICADAS SOBRE LOS DATOS 8](#_Toc89539027)

[IV. RESULTADOS DE TÉCNICAS APLICADAS 19](#_Toc89539028)

[V. CUANTIFICACIÓN DE LA MEJORA CON MODELO REALIZADO 24](#_Toc89539029)

[VI. PROBLEMAS ENCONTRADOS 25](#_Toc89539030)

[VII. PROPUESTAS DE APLICACIÓN 26](#_Toc89539031)

[BIBLIOGRAFÍA 27](#_Toc89539032)

# **INTRODUCCIÓN**

*“La falta de puntualidad es el ladrón del tiempo”*

*Oscar Wilde*

Los retrasos en los vuelos se han transformado en un tema importante para el transporte aéreo en todo el mundo debido a las pérdidas financieras ocasionadas por esta situación.

Estos retrasos no solo causan inconvenientes a las aerolíneas sino también a los pasajeros. El resultado es un aumento en el tiempo de viaje que aumenta los gastos asociados con la comida y el alojamiento y, en última instancia, causa estrés entre los pasajeros. Las aerolíneas son víctimas de costos adicionales asociados a sus tripulaciones, reposicionamiento de los aviones, consumo de combustible al tratar de reducir los tiempos transcurridos y muchos otros factores que, en conjunto, alteran su reputación y, a menudo, resultan en una pérdida de demanda por parte de los pasajeros.

Las razones de estos retrasos tienen diversas fuentes, desde la congestión del aire hasta las condiciones climáticas, problemas mecánicos, dificultades para abordar a los pasajeros y simplemente la incapacidad de las aerolíneas para manejar la demanda dada su capacidad.

Por lo tanto, ¿qué se pueden hacer los pasajero para evitar retrasos en los vuelos? ¿Es posible saber si su vuelo se retrasará antes de que aparezca en los paneles de salida o antes de ingresar al avión? La respuesta a estas preguntas se realizará a través de los modelos de Machine Learning: Support Vector Machine (SVM) y Árboles de decisión.

En este proyecto, analizamos diferentes técnicas / algoritmos de Machine Learning para tratar de predecir si un vuelo se retrasará o no antes de que se anuncie en los tableros de salida debido a ciertos factores.

Como se podrá apreciar a continuación, este proyecto se realizó con datos de 2020 para los vuelos nacionales de EE. UU. No tiene en cuenta los efectos de COVID-19 ya que es más un ejercicio para ver si esto se puede hacer o no, y con qué grado de precisión.

# **DESCRIPCIÓN DEL PROBLEMA DE NEGOCIO**

La puntualidad de los vuelos es una medida importante de la calidad del servicio de los aeropuertos y de las compañías aéreas.

Según Alvarado, Cerna & Spolmann (2009), en Estados Unidos, los retrasos del transporte aéreo son perjudiciales para su economía. Sólo para los vuelos nacionales en el año 2007 se estima que los pasajeros sufrieron una demora de 320 millones de horas, con un costo para la economía estadounidense de más de US$41 mil millones. Asimismo, indican que en este país el tratamiento de estas materias se encuentra desregulado, lo cual explica la ausencia de un marco regulatorio específico para el tratamiento de situaciones como los retrasos y cancelaciones de vuelo.

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

Asimismo, cada país cuenta, particularmente, con una estructura de penalidades económicas al no cumplir niveles mínimos de cancelaciones y retrasos en los vuelos programados. En Estados Unidos, las aerolíneas que operan en este país están sujetas a cumplir los lineamientos y estándares de calidad determinados por la *International Air Transport Association* (IATA).

Esta entidad corresponde a la asociación comercial del 83% del tráfico aéreo (290 aerolíneas). Su misión es representar, liderar y servir a la industria de las aerolíneas. Entre sus propósitos está el emitir estándares de recomendaciones de operación que, a su vez, son empleadas para construir un marco regulatorio ajustado para cada país.

Por otro lado, las causas de los retrasos aéreos que se identifican superan las 50 de ellas. Sin embargo, Díaz & Sandoval (2020) identifican cinco causas principales, las cuales se describen a continuación:

1. **Acumulación de tráfico aéreo.** Entre el aterrizaje y el despegue todas las aerolíneas deben mantener un *tiempo de escala[[1]](#footnote-1)* que debe ser determinados por las mismas. En este tiempo están incluidas todas las operaciones que deben hacer tanto las empresas, como los usuarios. Entonces, con que en un vuelo no cumpla con el tiempo determinado, generará que las demás operaciones se retrasen.
2. **Verificación de los aparatos.** Ligado al *tiempo de escala* ya que este tipo de causa debe estar incluido en este. Esta causa está relacionada a comprobaciones obligatorias, reparaciones, componentes que aseguran la calidad del vuelo, entre otros.
3. **Gestión aeroportuaria.** Hace referencia a todos los controles que los aeropuertos tuvieron que crear debido a situaciones que pongan en riego la integridad de los trabajadores y pasajeros. En estos se incluyen los registros en aduana, registro de equipaje y revisión de equipajes de mano. Sin embargo, este aspecto tiene mayor presencia en temporadas altas.
4. **Navegación aérea y pasajeros.** Esta 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, lo cual ocurre cuando no cumplen con la hora de abordaje establecida y generan incumplimiento en la hora de despegue.
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, entre otros.

Las causas anteriormente mencionadas afectan directa o indirectamente a usuarios, personal del aeropuerto y las empresas prestadoras de servicios aéreos (aerolíneas).

Por lo tanto, las predicciones precisas de los retrasos de los vuelos seguirán siendo fundamentales para ayudar a los aeropuertos y a las compañías aéreas a ofrecer un servicio de alta calidad.

# **SUPUESTOS UTILIZADOS**

Para desarrollar este trabajo, se indican a continuación ciertos parámetros que son excluidos de este estudio:

* *Mercado antes de pandemia:* Debido a las condiciones sanitarias que actualmente se está viviendo en todo el mundo, el mercado de las aerolíneas ha sido uno de los más afectados, es decir, gracias a las restricciones impuestas por cada gobierno, la movilidad de las personas se ha visto reducida, generando una disminución en la compra de tickets aéreos. No obstante, el “efecto pandemia” no se considera al realizar la investigación, ya que, durante el mes de enero del 2020, en Estados Unidos aún no existían restricciones relacionadas a esta situación.

Cabe recalcar que el *dataset* está compuesto por observaciones para el mes de enero del 2020 y es de origen estadounidense.

* *Datos perdidos:* Al encontrarse con datos nulos (NA) que seguramente se perdieron o no se midieron, supondremos que todas maneras la falta de ellos no afectará nuestro análisis y se podrá utilizar el *dataset* para los objetivos impuestos.

# **TÉCNICA O TÉCNICAS APLICADAS SOBRE LOS DATOS**

Para la realización de este trabajo se utilizó la siguiente herramienta la cual sirvió para desarrollar un modelo predictivo.

## **SUPPORT VECTOR MACHINE (SVM)**

### **¿Qué es SVM?**

"Support Vector Machine" (SVM) es un algoritmo de aprendizaje automático supervisado relativamente simple que se utiliza para clasificación y / o regresión. Es más preferido para la clasificación, pero a veces también es muy útil para la regresión. Básicamente, SVM encuentra un hiperplano que crea un límite entre los tipos de datos. En el espacio bidimensional, este hiperplano no es más que una línea. En SVM, se traza cada elemento de datos en el conjunto de datos en un espacio N-dimensional, donde N es el número de características / atributos en los datos. A continuación, busca el hiperplano óptimo para separar los datos. Entonces, con esto, SVM solo puede realizar una clasificación binaria (es decir, elegir entre dos clases). Sin embargo, existen varias técnicas que se pueden utilizar para problemas de varias clases.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Los vectores de apoyo son simplemente las coordenadas de una observación individual. El clasificador SVM es una frontera que segrega mejor las dos clases (hiperplano/ línea).

### **¿Cómo funciona SVM?**

* **Identificar el hiperplano correcto (Escenario-1):** Se tienen tres hiperplanos (A, B y C), entre los cuales se debe identificar el hiperplano correcto para clasificar las estrellas y los círculos, para este ejemplo.

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente

Ilustración 1: Escenario 1

Es necesario recordar una regla general para identificar el hiperplano correcto: "Seleccione el hiperplano que segregue mejor las dos clases". En este escenario, el hiperplano "B" ha realizado esta tarea.

* **Identificar el hiperplano correcto (Escenario 2):** Se tienen nuevamente tres hiperplanos (A, B y C) y todos segregan bien las clases. Ahora, ¿cómo se puede identificar el hiperplano correcto?

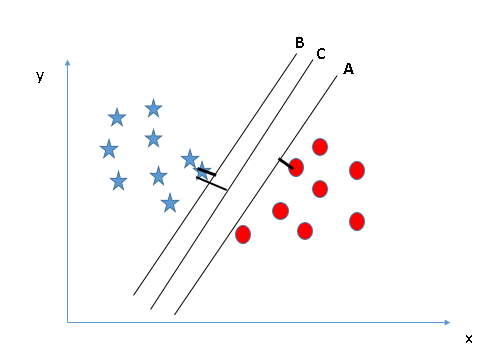
Ilustración 2: Escenario 2.1

Interfaz de usuario gráfica, Gráfico

Descripción generada automáticamente con confianza media

Aquí, la maximización de las distancias entre el punto de datos más cercano (cualquier clase) y el hiperplano ayudará a decidir el hiperplano correcto. Esta distancia se denomina Margen.

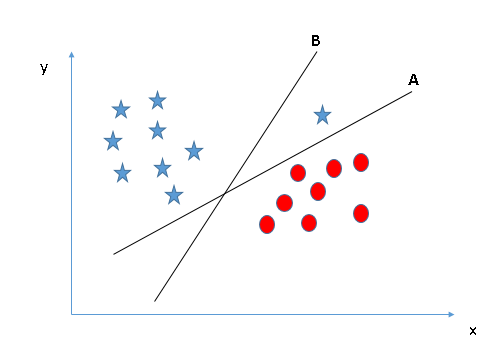
Ilustración 3: Escenario 2.2



En la imagen anterior, se puede ver que el margen del hiperplano C es más grande en comparación con A y B. Por lo tanto, nombramos el hiperplano correcto como C. Otra razón para seleccionar el hiperplano con mayor margen es la robustez. Si seleccionamos un hiperplano con un margen bajo, hay muchas posibilidades de que se produzcan errores de clasificación.

* **Identifique el hiperplano correcto (Escenario 3):** Sugerencia: Se deben utilizar las reglas discutidas en el escenario anterior para identificar el hiperplano correcto.

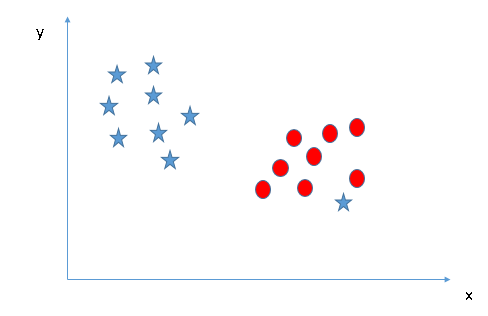
Ilustración 4: Escenario 3



Algunos seleccionarán el hiperplano B ya que este tiene mayor margen en comparación con A. Sin embargo, SVM selecciona el hiperplano que clasifica las clases con precisión antes de maximizar el margen. Aquí, el hiperplano B tiene un error de clasificación y A ha clasificado todo correctamente. Por lo tanto, el hiperplano correcto es A.

* **¿Podemos clasificar dos clases (Escenario 4)?:** A continuación, no se puede segregar las dos clases utilizando una línea recta, ya que una de las estrellas se encuentra en el territorio de la otra clase (círculo) como un valor atípico.

Ilustración 5: Escenario 4.1



Como se indica anteriormente, una estrella en el otro extremo es como un valor atípico para la clase de estrellas. El algoritmo SVM tiene una característica para ignorar los valores atípicos y encontrar el hiperplano que tiene el máximo margen. Por lo tanto, se puede decir que la clasificación SVM es robusto a los valores atípicos.

Ilustración 6: Escenario 4.2

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

* **Encuentre el hiperplano para segregar a las clases (Escenario-5):** En el escenario siguiente, no se puede tener un hiperplano lineal entre las dos clases, así que ¿cómo clasifica la SVM estas dos clases?

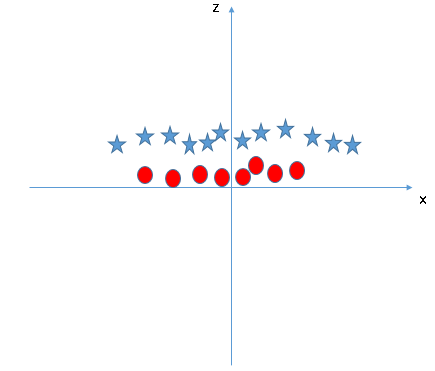
Ilustración 7: Escenario 5.1

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

La SVM puede resolver este problema fácilmente. Este se resuelve mediante la introducción de una característica adicional. Aquí, se añade una nueva característica z = x2+y2. Ahora se trazan los puntos de datos en el eje x y z:

Ilustración 8: Escenario 5.2

****

Los puntos a considerar del gráfico anterior son los siguientes:

* Todos los valores de z serán siempre positivos porque z es la suma cuadrada de x e y.
* En el gráfico original, los círculos rojos aparecen cerca del origen de los ejes x e y, lo que lleva a un valor más bajo de z y la estrella relativamente alejada del origen da como resultado un valor más alto de z.

En el clasificador SVM, es fácil tener un hiperplano lineal entre estas dos clases. Pero, otra pregunta que surge es: ¿se debe añadir esta característica manualmente para tener un hiperplano?. La respuesta es no, el algoritmo SVM tiene una técnica llamada *el truco del kernel*. El *kernel* de la SVM es una función que toma un espacio de entrada de baja dimensión y lo transforma en un espacio de mayor dimensión, es decir, convierte un problema no separable en un problema separable. Es muy útil en los problemas de separación no lineal. En otras palabras, realiza algunas transformaciones de datos extremadamente complejas y, a continuación, encuentra el proceso para separar los datos basándose en las etiquetas o salidas que se han definido.

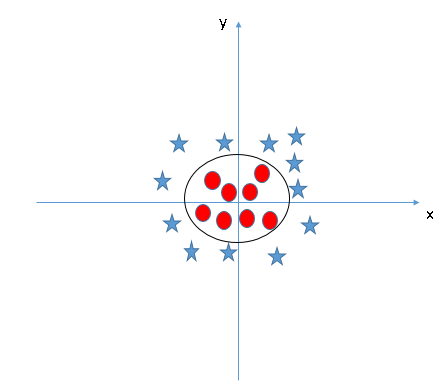


Ilustración 9: Escenario 5.3

### **Kernel SVM**

Los algoritmos SVM utilizan un conjunto de funciones matemáticas que se definen como el núcleo. La *función del kernel* es tomar datos como entrada y transformarlos en la forma requerida. Los diferentes algoritmos de SVM utilizan diferentes tipos de funciones del *kernel*. Estas funciones pueden ser de diferentes tipos. Por ejemplo, lineal, no lineal, polinomial, función de base radial (RBF) y sigmoide.

### **Ventajas y desventajas del modelo**

Al igual que varios modelos, el SVM posee ventajas y desventajas al utilizarlo, las cuales se describen a continuación:

Tabla 1: Ventajas y desventajas del modelo SVM

|  |  |
| --- | --- |
| Ventajas | Desventajas |
| * SVM funciona relativamente bien cuando hay un claro margen de separación entre clases. | * El algoritmo SVM no es adecuado para grandes conjuntos de datos. |
| * SVM es más eficaz en espacios de gran dimensión. | * SVM no funciona muy bien cuando el conjunto de datos tiene más ruido, es decir, las clases se superponen. |
| * SVM es eficaz en los casos en que el número de dimensiones es mayor que el número de muestras. | * En los casos en los que la cantidad de características para cada punto de datos exceda la cantidad de muestras de datos de entrenamiento, la SVM tendrá un rendimiento inferior. |
| * SVM es relativamente eficiente en memoria. | * Como el clasificador de vectores de soporte funciona poniendo puntos de datos, por encima y por debajo del hiperplano de clasificación no hay una explicación probabilística para la clasificación. |

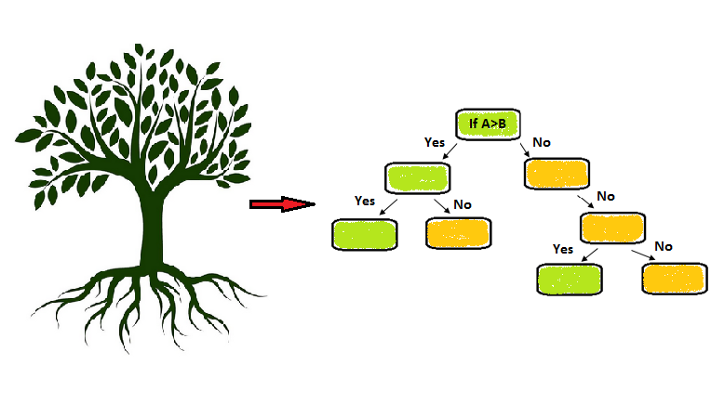
Fuente: Elaboración propia

## **ÁRBOLES DE DECISIÓN**

### **¿Qué es un árbol de decisión?**

Un árbol de decisión es una estructura ramificada que muestra las diferentes opciones y sus consecuencias. Los puntos en los que hay que tomar decisiones se muestran como nodos, las ramas unen estos nodos y las decisiones últimas son las hojas, donde el camino termina (también se denominan nodos terminales) .

Ilustración 10: Esquema de un árbol de decisión



En el campo del aprendizaje automático, hay distintas maneras de obtener árboles de decisión. La que se utilizará para el presente trabajo es conocida como CART: *Classification And Regression Trees*, la cual es una técnica de aprendizaje supervisado. Se tiene una variable objetivo (dependiente) y la meta es obtener una función que permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.

La implementación particular de CART que se utiliza en este trabajo es *Recursive Partitioning and Regression Trees* o RPART. Este algoritmo permite encontrar la variable independiente que mejor separa las observaciones contenidas en la base de datos en grupos, las cuales tienen relación con las categorías de la variable objetivo. Una característica muy importante de este algoritmo es que una vez que alguna variable ha sido elegida para separar los datos, ya no es utilizada nuevamente en los grupos que ha creado. Se buscan variables distintas que mejoren la separación de los datos.

### **Ventajas y desventajas del modelo**

A continuación, a través de la siguiente tabla se presentan las ventajas y desventajas de este modelo:

Tabla 2: Ventajas y desventajas de los árboles de decisión

|  |  |
| --- | --- |
| Ventajas | Desventajas |
| * Fácil de interpretar e implementar. | * Es un tipo de clasificación débil ya que sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo. |
| * No es demandante en poder de cómputo comparado con procedimientos más sofisticados. | * Fácil de sobre ajustar, esto es, hacerlos excelentes para clasificar datos que conocemos, pero deficientes para datos conocidos. |
| * Tiende a dar buenos resultados de predicción para muchos tipos de datos. |  |
|  |  |

Fuente: Elaboración propia

Las variables que se utilizarán para realizar este análisis son del tipo cuantitativa. Estas variables se encuentran en la base de datos *flight\_prediction*, la cual cuenta con la descripción de todos los vuelos realizados en los Estados Unidos durante el mes de enero en 2020. Las variables que constituyen la base de datos que se utilizará para predecir el modelo se presentan y describen en la siguiente tabla:

Tabla 3: Descripción de variables utilizadas para la predicción de atrasos de vuelos

|  |  |
| --- | --- |
| FEATURES | DESCRIPCIÓN |
| *DAY\_OF\_MONTH* | Variable cuantitativa que indica el día del mes |
| *DAY\_OF\_WEEK* | Variable cuantitativa que indica el día de la semana |
| *OP\_UNIQUE\_CARRIER* | Variable cualitativa que indica el código del transportista. |
| *OP\_CARRIER\_AIRLINE\_ID* | Variable cuantitativa que señala el número de identificación asignado por el Departamento de Transporte de EE.UU. para identificar una línea aérea única. |
| *OP\_CARRIER\_FL\_NUM* | Variable cuantitativa que indica el número de vuelo. |
| *ORIGIN\_AIRPORT\_ID* | Variable cuantitativa que señala el Aeropuerto de origen, ID de aeropuerto. Número de identificación asignado por el Departamento de Transporte de Estados Unidos para identificar un único aeropuerto. |
| *DEST\_AIRPORT\_ID* | Variable cuantitativa que indica el Aeropuerto de destino, ID de aeropuerto. Un número de identificación asignado por el Departamento de Transporte de Estados Unidos para identificar un único aeropuerto. |
| *DEP\_TIME* | Variable cuantitativa que señala la Hora de salida real (hora local: hhmm) |
| *DEP\_DEL\_15* | Variable categórica del tipo numérico que es un Indicador de retraso en la salida, 15 minutos o más (1=Sí, 0=No) |
| *ARR\_DEL15* | Variable categórica del tipo numérico que representa un Indicador de retraso de llegada, 15 minutos o más (1=Sí, 0=No) |
| *BLOQUE* | Variable cuantitativa creada que indica el bloque de horario de salida (toma valores del 0 al 19 ). |
| *CANCELLED* | Variable cuantitativa que es un Indicador de vuelo cancelado (1=Sí, 0=No) |
| *DIVERTED* | Variable categórica del tipo numérico que representa a un Indicador de vuelo desviado (1=Sí, 0=No) |
| *DISTANCED* | Variable cuantitativa que señala la Distancia entre aeropuertos (millas) |

Fuente: Elaboración propia

De estas variables, la que se busca predecir, y, por ende, se justa al problema de negocio es la variable DEP\_DEL 15, la cual indica si existe un retraso de 15 minutos en la salida del vuelo. Esta variable es del tipo discreta ya que toma valores 1 (posee el atributo ) y 0 (no posee el atributo).

Para el primer problema (SVM) se utiliza una parte del dataset original (5.000 observaciones) ya que este tipo de modelo no funciona bien en grandes conjuntos de datos.

En cuanto al segundo modelo, árboles de decisión, se realizarán dos árboles. Para el primero, se utiliza la base de datos completa, donde el 70% de ésta será utilizada como un set de entrenamiento y el 30% restante será utilizado como test de prueba. En cuanto al segundo árbol, se utilizará una submuestra de la base de datos original (5.000), donde los set de entrenamiento y prueba tendrán la misma proporción antes descrita (70% y 30% respectivamente).

En relación con el cómo se determinará cual modelo se comporta mejor, se utilizará el *accuracy* que arroje cada uno de ellos.

# **RESULTADOS DE TÉCNICAS APLICADAS**

## **Modelo 1: SUPPORT VECTOR MACHINE (SVM)**

El dataset utilizado (flight\_prediction) contiene datos que especifican cada vuelo realizado durante el mes de enero, en Estados Unidos, en el año 2020. La variable que representa el problema de negocio (atrasos de vuelos) es DEP\_DEL15, la cual indica si hubo o no un retraso de 15 minutos o más en la salida del vuelo (1 = con atraso, 0 = sin atraso).

Para obtener el mejor hiperplano mediante el modelo SVM, en primer lugar, se trabaja con una base de datos de 5.000 observaciones, puesto que la data original es de gran tamaño y se demora en correr los comandos necesarios. Transformamos la variable DEP\_DEL15 en factor para poder realizar la clasificación y modificamos el data frame a tibble. Luego realizamos la búsqueda de parámetros y encontramos los mejores para el modelo, estos se explican a continuación.

Tabla 4: Resultados de parámetros

|  |  |
| --- | --- |
| **Parámetros** | |
| Kernel | Polynomial |
| Degree | 3 |
| Cost | 4.4 |
| Gamma | 0.911 |
| Mmce.test.mean | 0.1127774 |

Fuente: Elaboración propia

Como resultado se obtuvo un *kernel* polinomial de grado 3, con el hiperparámetro de 4,4 que representa el peso de cada observación a la hora de clasificar y se usa para controlar el error y con el hiper parámetro gama de 0,911 que representa la distancia entre las observaciones y se usa para dar peso de curvatura al límite de decisión.

Por último, de acuerdo con los resultados, el modelo utilizado acierta en un 89,12% (accuracy) respecto al total.

## **Modelo 2: Árboles de decisión**

En este modelo, el primer árbol indica cual es el porcentaje de aquellos vuelos que salieron sin atraso del aeropuerto de origen que efectivamente llegan al aeropuerto de destino con atraso.

De este se puede apreciar que el 14% de los vuelos que salieron del aeropuerto de origen no contaba con 15 minutos de retraso en su salida. Sin embargo, un 75% de esos vuelos (ARR\_DEL15 ≠ 0) sí presentó un retraso en su llegada al aeropuerto de destino (véase ilustración 11).

Ilustración 11: Árbol de decisión 1

Diagrama

Descripción generada automáticamente

Fuente: Elaboración propia

En cuanto al segundo árbol, éste cuenta con un nodo más que el anterior, lo cual permite obtener una descripción más detallada del problema de negocio.

El segundo árbol, al igual que el anterior, indica cual es el porcentaje de aquellos vuelos que salieron sin atraso del aeropuerto de origen que efectivamente llegan al aeropuerto de destino con atraso. Pero, además, permite conocer a qué hora de salida del vuelo es más probable que presente un atraso en el arribo al aeropuerto de destino.

De este se obtiene que el 14% de los vuelos que salieron del aeropuerto de origen no contaba con 15 minutos de retraso en su salida, pero un 78% de esos vuelos (ARR\_DEL15 ≠ 0) sí presentó un retraso en su llegada al aeropuerto de destino. Además, se observa que la mayoría de los retrasos se concentraron después de las 08:15 de la mañana (aproximadamente).

Diagrama

Descripción generada automáticamente

Ilustración 12: Árbol de decisión 2

Fuente: Elaboración propia

Como se puede apreciar, ambos modelos presentan distintos resultados en los primeros nodos. Esto se debe a que el primer árbol fue realizado con el total de las observaciones que contiene la base de datos. Por su parte, el segundo árbol fue realizado con una submuestra de 5.000 observaciones de la base de datos original. Por ende, en base a estos se corrobora una de las desventajas que presenta este modelo (árboles de decisión), la cual indica, en síntesis, que los resultados obtenidos dependerán de la cantidad de datos utilizados.

Asimismo, se creó para ambos árboles dos set, uno de entrenamiento para generar un modelo predictivo, y un set de prueba mediante el cual se comprueba la eficacia del modelo para realizar predicciones correctas. El set de entrenamiento corresponde al 70% del total de la base de datos utilizada, y el set de prueba contempla un 30% de estos.

También, para este modelo, y cada uno de los árboles generados, se pudo obtener las siguientes matrices de confusión:

Tabla 5: Matriz de confusión árbol de decisión 1

|  |  |
| --- | --- |
| Sin retraso | Con retraso |
| Sin retraso | 148.291 | 6.043 |
| Con retraso | 6.243 | 18.573 |

Fuente: Elaboración propia

Tabla 6: Matriz de confusión árbol de decisión 2

|  |  |
| --- | --- |
| Sin retraso | Con retraso |
| Sin retraso | 1.315 | 81 |
| Con retraso | 21 | 130 |

Fuente: Elaboración propia

Esta matriz de confusión nos permite visualizar el rendimiento del algoritmo, en este caso uno de aprendizaje supervisado (en el aprendizaje no supervisado generalmente se denomina matriz coincidente). Cada fila de la matriz representa las instancias en una clase predicha, mientras que cada columna representa las instancias en una clase real (o viceversa). El nombre se deriva del hecho de que hace que sea fácil ver si el sistema confunde dos clases (es decir, comúnmente etiquetar incorrectamente una como otra).

Por otro lado, para conocer cuál de los modelos se ajusta mejor al problema de negocio se comparan sus valores de *accuracy*, los cuales se indican en la siguiente tabla:

Tabla 7: Comparación de modelos

|  |  |
| --- | --- |
| Modelo | Accuracy |
| Support Vector Machine (SVM) | 0.8912 |
| Árbol de decisión 1 | 0.9317 |
| Árbol de decisión 2 | 0.9341 |

Fuente: Elaboración propia

El modelo SVM presenta un *accuracy* de 0.8912, pero los árboles de decisión se comportan mejor, siendo el segundo árbol el cual sobresale de los demás. De estos resultados se puede extraer que existe un mejor comportamiento cuando la cantidad de datos utilizados es menor. Cabe recalcar que el segundo árbol de decisión fue realizado con 5.000 observaciones, es decir, una submuestra de la base de datos original.

# **CUANTIFICACIÓN DE LA MEJORA CON MODELO REALIZADO**

Los modelos anteriormente realizados tendrían un impacto positivo tanto para las compañías de transporte aéreo como para las personas que utilizan este servicio. Estos permitirán tener información de los atrasos d ellos vuelos, pero, además, obtener predicciones futuras de estos atrasos y las razones de éstos (clima, día y horario en el cual se realiza el vuelo, entre otros).

Es gracias a este tipo de tecnología aplicada que las molestias que experimentan los viajeros se reducen, las cuales se producen debido a los retrasos de vuelos e inconvenientes en los aeropuertos (los cuales también son parte de estos retrasos).

Para mejorar estos modelos se propone usar y analizar, además de los datos disponibles, los datos operacionales para cada una de las aerolíneas (programación de tripulación, mantenimiento de los aviones, número de pasajeros y sillas disponibles, etc.) las cuales también son determinantes en el atraso de un vuelo.

Asimismo, se espera que el accuracy de estos modelos fluctúe entre un 80% a 90% para que se ajuste a resultados reales que se puedan obtener a través de algoritmos de Machine Learning.

Cabe destacar que los algoritmos de Machine Learning están diseñados para observar un conjunto de datos que cambian de manera rápida y encontrar patrones. A partir de ello, los algoritmos realizan predicciones y aprenden a hacer nuevas predicciones y tomar decisiones.

# **PROBLEMAS ENCONTRADOS**

A continuación, se destacan algunos problemas que surgieron en relación con el desarrollo del modelo de las predicciones de vuelos:

* **Base de datos grande.** El principal problema con el que nos encontramos tiene relación con el tamaño de la base de datos. Al utilizar una data con aproximadamente 600.000 observaciones, el programa R se demoraba demasiado tiempo en correr los comandos. En solución a este problema solo trabajamos con 5.000 observaciones para poder obtener resultados a tiempo, específicamente para el modelo SVM.
* **Máximo de iteraciones**. Al momento de correr el modelo SVM arroja una alerta de un “*máximo de iteraciones*”, por lo que solo pudimos utilizar 5.000 observaciones para poder obtener resultados.
* **Variables cualitativas.** Para el modelo SVM sólo deben utilizarse variables del tipo numéricas, y aquellas que no lo sean se deben representar numéricamente. Sin embargo, la base de datos, por ejemplo, cuenta con más de 500 rutas posibles, por lo que asignarle un número a cada una de ellas tendría un gran costo (tiempo). Por lo tanto, se decide sacar este tipo de variables de la base de datos y trabajar exclusivamente con las variables cuantitativas contenidas en el dataset.

# **PROPUESTAS DE APLICACIÓN**

Este tipo de modelos se recomienda que sean aplicados a problemas que cuenten con dos opciones o respuestas. Primero, el modelo SVM podría aplicarse, por ejemplo, en los siguientes casos:

1. **Detección de rostros:** SVM clasifica partes de la imagen como rostros y no rostros y crea un límite cuadrado alrededor del rostro.
2. **Categorización de texto e hipertexto:** las SVM permiten la categorización de texto e hipertexto para modelos inductivos y transductivos. Usan datos de entrenamiento para clasificar documentos en diferentes categorías. Se clasifica en función de la puntuación generada y luego se compara con el valor umbral.
3. **Clasificación de imágenes:** el uso de SVM proporciona una mayor precisión de búsqueda para la clasificación de imágenes. Proporciona una mayor precisión en comparación con las técnicas de búsqueda tradicionales basadas en consultas.
4. **Bioinformática:** Se utiliza SVM para identificar la clasificación de genes, pacientes sobre la base de genes y otros problemas biológicos.

Por su parte, los árboles de decisiones se utilizan en cualquier proceso que implique el tomar una decisión. Estos suelen ser binarios, pero esto no quiere decir que no puedan existir árboles con tres o más nodos. Asimismo, permite tomar decisiones a partir de predicciones más detalladas.

# **BIBLIOGRAFÍA**

Agrawal, S. (Enero de 2020). *Introduction to Support Vector Machine(SVM)*. Obtenido de https://ai.plainenglish.io/: https://ai.plainenglish.io/introduction-to-support-vector-machine-svm-cd0759098471

Alvarado, A., Cerna, M., & Spolmann, S. (24 de Noviembre de 2009). *IMPACTO EN LA INDUSTRIA AERONÁUTICA POR EL PAGO DE INDEMNIZACIONES O MULTAS POR DENEGACIÓN DE EMBARQUE, ATRASOS Y CANCELACIONES DE VUELOS.* Obtenido de ac.gob.cl: http://www.jac.gob.cl/wp-content/uploads/2014/10/ESTUDIOS-004-IMPACTO-EN-LA-INDUSTRIA-AERON%C3%81UTICA-POR-PAGO-DE-INDEMNIZACIONES-2009.pdf

Association), I. (. (2021). *https://www.iata.org/*. Obtenido de https://www.iata.org/en/about/

Chavarro, C., Ramirez, W., & Arias, C. (01 de Junio de 2020). *MODELO PARA IDENTIFICAR LOS VUELOS AFECTADOS POR RETRASOS O CANCELACIONES EN EL AEROPUERTO EL DORADO DE BOGOTÁ, COLOMBIA.* Obtenido de https://repository.javeriana.edu.co/: https://repository.javeriana.edu.co/bitstream/handle/10554/51109/Ramirez%2C%20W.%20Chavarro%2C%20C.%20Arias%2C%20C.%202020.pdf?sequence=1&isAllowed=y

Diaz, S., & Sandoval, A. (2020). *PROPUESTA DE MODELO PREDICTIVO EN RETRASO DE DESPEGUE PARA AEROPUERTOS DE MEDIANA COMPLEJIDAD EN COLOMBIA.* Obtenido de https://repository.usta.edu.co/: https://repository.usta.edu.co/jspui/bitstream/11634/34814/1/2021andressandoval.pdf

Mendoza, J. (23 de Abril de 2018). *Arboles de decisión con R - Clasificación*. Obtenido de RPubs: https://rpubs.com/jboscomendoza/arboles\_decision\_clasificacion

Ray, S. (17 de Septiembre de 2017). *Understanding Support Vector Machine(SVM) algorithm from examples (along with code)*. Obtenido de analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

1. El tiempo de escala es una herramienta que permite visualizar todos los sucesos que han ocurrido, en lo que respecta a este trabajo, en un vuelo. [↑](#footnote-ref-1)