Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey.

IN3038.1 Laboratorio de diseño y optimización de operaciones

Proyecto Final de Ciencia de Datos
Semestre Febrero-Junio 2021

Integrantes:

Laura López Reyes	A01363564.
Ricardo Tapia Martínez	A01364132
Paula Camila Nieto Naranjo	A00819174
Daniela Gomez Ortega	A01364823

Fecha de Entrega: 02 de Junio de 2021

Introducción

Hoy en día se generan millones de datos en todo el mundo, de los cuales aproximadamente sólo el 0.5% de ellos se analizan para generar información valiosa [1]. En la actualidad, las organizaciones han mostrado mayor interés por analizar los grandes volúmenes de datos generados por las personas mediante la Ciencia de Datos, campo que utiliza métodos, algoritmos, sistemas, además de que requiere que las personas de este campo tengan conocimientos estadísticos, informáticos, matemáticos, empresariales y de storytelling [1]. El propósito de un proyecto de Ciencia de Datos es analizar los datos con el fin de convertirlos en conocimiento que les ayude a las empresas a tomar mejores decisiones, tener una mayor precisión del comportamiento futuro de sus servicios o productos, incrementar la productividad, conocer con precisión las necesidades y gustos de los usuarios, entre otros, para poder generar ventajas competitivas y diferenciadores en comparación de sus competidores.

De acuerdo a información obtenida por el Instituto Federal de Telecomunicaciones (IFT), desde junio 2013 a septiembre 2020 ha existido un crecimiento de 15.8% de las líneas totales de telefonía celular [2] y al cierre del segundo trimestre de 2020 se concluyó que hay 95 líneas de telefonía celular por cada 100 habitantes [3], por lo tanto se concluye que la telefonía móvil se ha clasificado como una de las tecnologías más consumidas por los usuarios mexicanos.

El objetivo de este proyecto es implementar un Proyecto de Ciencia de Datos en una empresa situada en el sector de telefonía celular, Motorola. En este proyecto se requiere predecir la demanda en los diferentes puntos de venta de todos los productos que ofrece la empresa y así evitar costos adicionales o excesivos y ofrecer un alto nivel de servicio al cliente. Además se desarrollan las diferentes etapas que comprenden la elaboración de un Proyecto de Ciencia de Datos mediante el uso de la metodología CRISP-M.

Etapa 1: Comprensión del Negocio

1) Descripción de la situación actual.
Motorola Inc fue una empresa especializada en telecomunicaciones y electrónica, en el año 2011, la compañía decidió dividirse en 2 firmas independientes, Motorola Mobility y Motorola Solutions. La compañía Motorola Mobility está encargada de la comercialización de dispositivos móviles mientras que Motorola Solutions se enfocó en la venta de soluciones de telecomunicaciones a empresas y gobiernos.[4]

Tras una serie de eventos, en el año del 2014 Lenovo adquirió Motorola Mobility con el fin de desafiar a 2 de los competidores más fuertes de la industria, Apple y Samsung. Esta adquisición fue estratégica ya que Lenovo identificó que Motorola tenía presencia fuerte en los mercados occidentales y mayores conocimientos de la tecnología móvil y así mejorar su oferta de teléfonos inteligentes. Motorola decidió entrar al mercado de gama media y precios accesibles que generaron un gran interés por parte de los consumidores. Estos esfuerzos lograron que Motorola sea el segundo fabricante con mayor intervención en el

mercado mexicano, con 17.1% de participación de acuerdo al ranking de la consultora *The Competitive Intelligence Unit*.[5]

La importancia de este proyecto radica en el análisis de datos para la toma de decisiones dentro de las empresas logrando mayores ganancias, liderazgo en la industria, entre otros beneficios. Un Ingeniero Industrial es capaz trabajar con ciencia de datos ya que entiende sistemas complejos y tiene fuertes bases en ciencias que le permiten proponer soluciones para mejorar e innovar procesos y sistemas mediante el uso de herramientas tecnológicas las cuales facilitan el manejo eficiente de los datos a analizar.[1]

2) Entender y describir la problemática (en términos del negocio).

Motorola y otras empresas del sector de la telefonía celular se han dado cuenta de la demanda creciente de sus productos, ya que el 75.1% de la población mexicana cuenta con un celular con fines comunicativos. A pesar de que la mayoría de la población cuenta con un dispositivo móvil, no todas las personas buscan las mismas características al adquirir uno. Por lo que Motorola debe pronosticar el número de unidades móviles a vender de cada uno de sus productos en sus diferentes puntos de venta. De no hacerlo de manera correcta sus utilidades, satisfacción al cliente y logística podrían verse afectadas.

3) Entender y describir la problemática (en términos de ciencia de datos).

Para iniciar un Proyecto de Ciencia de Datos hay que comenzar con una pregunta,en nuestro caso es ¿Cuál será la demanda de los productos de la marca de telefonía celular Motorola en sus diferentes puntos de venta, al siguiente mes de registro?.
Una vez realizado este paso, se requiere de analizar el tipo de tarea de nuestro proyecto, en ese caso es de regresión ya que el atributo, que en este caso es la demanda, no es de naturaleza discreta sino continua. Al estar lidiando con periodos de tiempo, se puede llamar de previsión.
Para la obtención de información de este proyecto se cuentan con grandes volúmenes de datos estructurados que serán analizados a través de software especializado para poder ejecutar las diferentes etapas de nuestro Proyecto de Ciencias de Datos.

4) Plasmar los objetivos.

- Etapa 1: Tener un completo entendimiento sobre la problemática a la que se enfrenta, tener claros los objetivos generales y específicos del proyecto, así como el alcance del mismo.
- Etapa 2: Conocer y explorar los datos que se nos presentan y así identificar los problemas que impiden un correcto manejo de los mismos.
- Etapa 3: Realizar las modificaciones necesarias a los datos para un uso óptimo de ellos.
- Etapa 4: Aplicar las herramientas y metodologías aprendidas a lo largo del curso para hacer una predicción de demanda para períodos futuros con un mínimo porcentaje de error.
- Etapa 5: Realizar una correcta presentación e interpretación de los resultados para un total entendimiento.

5) Estructurar el proyecto.

Anexo 1: Diagrama de Gantt con plan preliminar del proyecto final

Etapa 2: Comprensión de los Datos

1) **Recolectar Datos**
   Motorola nos proporcionó un dataframe en formato CSV que cuenta con diferentes registros que son de utilidad para cumplir con nuestro objetivo. En general, el dataframe muestra las ventas que ha tenido motoral entre el 01 Junio del 2018 hasta el 30 de Marzo del 2019.

2) **Describir Datos**
   Se utilizó el software R para poder visualizar y familiarizarse con los datos. De esta manera, se supo que el dataframe en formato CSV cuenta con 302,561 observaciones distribuidas en 14 variables de la siguiente manera:

Variable	Descripción
Punto_de_venta	Expone los diferentes puntos de venta en donde Motorola vende sus productos. Actualmente existen 1900 diferentes puntos de venta alrededor de México, entre ellos están poniente, acayucan y benito juarez. Finalmente, esta variable es tipo character(chr.).
Fecha	Muestra el registro de la fecha en la que se realizó una compra. Todas están en formato día/mes/año por ejemplo la primera venta fue realizada el 01/06/2018 y la variable es de tipo character (chr.).
Mes	Detalla el mes en el que se realizó una compra. Tiene formato numérico nominal. Es decir si la venta se realizó en Julio el registro lo muestra con el número 7 y así sucesivamente. Es una variable tipo character (chr.).
Anio	Despliega el año en el que se registró una compra. Es de valor numérico y se muestra de dos diferentes maneras, ya sea en formato “aaaa” como 2018 o 2019 o solamente los dós últimos dígitos del año, es decir, 18 o 19. Esta variable es de tipo entero (int).
Num_ventas	Esta variable representa el número de ventas que se realizó. En este caso todos los registros muestran el valor de 1 lo que quiere decir que las personas compran como mínimo y máximo una unidad. Esta variable es de tipo entero (int.).
SKU	El SKU es el código de referencia del producto. Es una secuencia de números y letras que representa un único producto. Por ejemplo, N.MZ2PLYG. Esta variable es de tipo character (chr.)

Marca	Esta variable muestra la marca del dispositivo que se está comprando. En este sentido, todos los registros tienen el nombre de “motorola” ya que es la empresa a la que se le está comprando. Es una variable tipo character (chr.)
Gamma	Refleja los diferentes tipos de gamma que maneja motorola. En este registro sólo se encontraron dos categorías agrupadas, entre ellas están celulares de gama baja y media. Es una variable tipo character (chr.)
Costo_promedio	Muestra el precio promedio del producto. Se encontraron registros con valores mínimos de 0 y máximos de 7143. Todos los precios detallados en esta lista están en pesos.mexicanos y es una variable tipo numérica (num)
Zona	Detalla la zona en la que se encuentra el punto de venta y se compró el producto. En total existen 9 zonas diferentes. Entre ellas están centro sur, golfo de méxico y norte. Es una variable tipo character (chr.)
Estado	Registra el estado en donde se realizó una compra. Se encontraron 35 estados registrados como puebla, tehuacán, michoacán, veracruz, entre otros. Es una variable tipo character (chr.)
Ciudad	Expone la ciudad en la que se realizó una compra. Se encontraron 228 ciudades diferentes por ejemplo, tehuacán, zamona, oluta, león, entre otras. Es una variable tipo character (chr.)
Latitud	Esta variable es un registro del Sistema de Coordenadas Geográficas de la latitud registrada. Tiene un valor mínimo de 14.9 y valor máximo de 1793999. Sus unidades están dadas en grados con decimales. Es una variable tipo numérica (num)
Longitud	Muestras coordenadas de la longitud registrada. Valor mínimo de -949106.0 y valor máximo de -86.8.Sus unidades están dadas en grados con decimales. Es una variable tipo numérica (num).

Tabla 1. Descripción de las variables existentes en la dramafrase.

3) Explorar Datos

Existen varias variables relacionadas entre sí. Por ejemplo el punto de venta se relaciona con la ciudad, el estado, la zona, la latitud y longitud. Por otro lado, el tipo de gamma se relaciona con el precio promedio de la venta. Ya que hay muchos datos de localización de la venta podríamos posteriormente realizar una gráfica en donde se muestran los lugares más recurrentes para vender los productos. Además con los datos se podrían realizar pronósticos de la demanda

4) **Verificar la Calidad de los Datos**

- **Punto de Venta:** (Faltas de ortografía) Existen 5 puntos de venta escritos de manera errónea. Esto afecta el análisis de datos ya que puede incrementar el número de puntos de ventas que tiene Motorola cuando la realidad es otra.
- **Mes:** (Datos no homogéneos) Esta variable pertenece a otro tipo de clase, se la debe cambiar a una numérica. La clase afecta en el uso de fórmulas dentro del análisis de datos ya que algunas de ellas sólo funcionan con cierto tipos de clases. Además existen valores mal registrados en lugar de números, son letras y todos deben convertirse en numéricos por el tipo de clase con el que se está trabajando.
- **Año:** (Datos no homogéneos/Valores faltantes) Debe seguir el formato de un valor numérico de 4 dígitos (valores faltantes) de lo contrario existirán datos homogéneos y serán tomados como otro año o fecha, lo cual puede alterar al reporte.
- **Marca:** (Faltas de Ortografía)Existen datos escritos de forma errónea y al no corregirlos podrán ser expresados una marca extra lo cual es erróneo porque sólo existe una.
- **Zona:** (Faltas de Ortografía) Existe una zona que está mal escrita. Es un error de calidad porque se puede expresar como una zona extra y perjudicar en el análisis de datos.
- **Estado:** (Incongruencia de datos) Hay 3 estados más de los que en realidad existen. Es un error de calidad, ya que es información mal proporcionada. Es decir se estaría alterando una realidad.
- **Latitud y Longitud:** En ambos casos, existen valores que están fuera de rango. Es un problema de calidad ya que afecta a la manera en la que se expresan los mismos, además que afecta posteriormente a su manipulación y análisis. Se recomienda homogeneizarlos.

Etapa 3: Preparación de los Datos

¿Cómo se hizo la limpieza de datos?

Una vez analizados los datos se identificaron diferentes problemas de calidad en los mismos. Para que exista una congruencia y estandarización en ellos, se realizaron modificaciones por lo se utilizó el comando “str_replace” para familiarizarnos con los mismos y poder manipularlos correctamente.

Punto de Venta: Se encontraron 5 puntos de venta escritos de manera errónea, los cuales se corrigieron ya que se encontraban mal escritos. A continuación, se muestran los puntos de venta con la corrección realizada.

Punto de venta erróneo	Corrección
#272 5 mayo zmm	#277 5 de mayo zmm
#625 ace BENITO JUAREZ	#591 ace benito juarez
#894 ace FRACChidalgo	#840 ace fracchidalgo
#50295 colosio hrmosillo	#50290 colosio hermosillo

#50581 cruz dl sr	#50410 cruz del sur

Mes: (Datos no homogéneos) Se encontraron variables pertenecientes a otro tipo de clase (nominal), se la debe cambiar a una numérica. La corrección realizada se muestra en la tabla siguiente:

Mes erróneo	Corrección
FEB	2
JUL	7
AGOSTO	8
NOV	11
DIC	12

Año: (Datos no homogéneos/Valores faltantes) Debe seguir el formato de un valor numérico de 4 dígitos (valores faltantes). Para esta variable se modificó lo siguiente:

- Año erróneo: 18
- Corrección: 2018

Marca: La marca se encontraba escrita de forma errónea

- Marca Errónea: motorola-motorola, mmotorola
- Corrección: motorola

Zona: Había una zona que estaba mal escrita.

- Zona Errónea: GOLFO DE MEX
- Corrección: golfo de mexico

Estado: Se encontró una incongruencia de los datos ya que hay 3 estados más de los que en realidad existen.

Estado erróneo	Corrección
tehuacan	puebla
toluca	estado de mexico
acapulco	guerrero

Latitud y Longitud: En ambos casos, se encontraron valores que están fuera de rango.

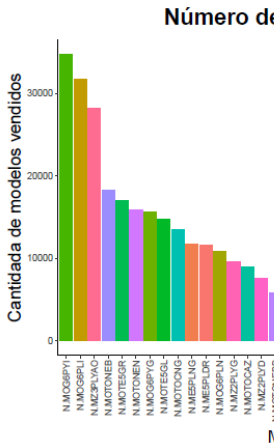
Latitud	Corrección
1793999	17.93999

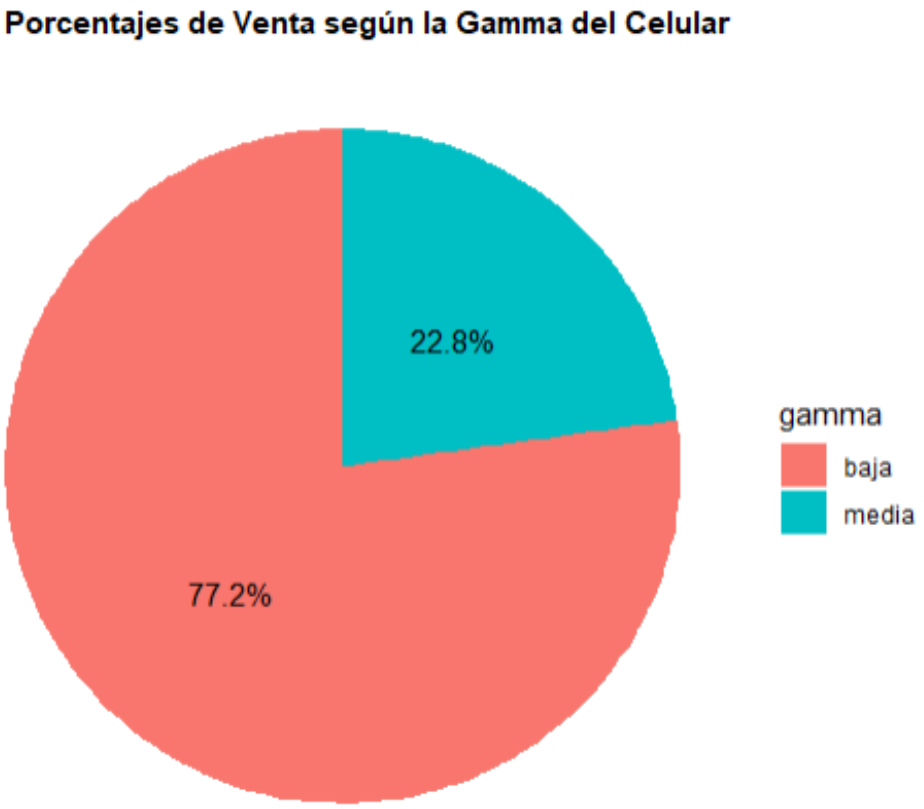
Longitud	Corrección
-949106.00000	-94.9106
-949106	-94.9106

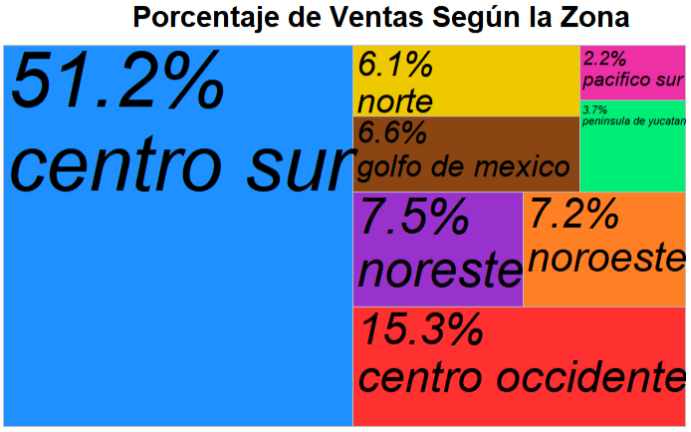
2) ¿Cuál es la situación actual?

A partir de la limpieza de datos se pudo contar con datos de calidad para su análisis correspondiente al proyecto de ciencia de datos en las siguientes etapas. Posteriormente se realizó el análisis exploratorio con la finalidad de llevar a cabo una búsqueda de gráficos visuales que asistan a comprender los datos y que se pueda extraer información valiosa.

Algunos de los descubrimientos más importantes que se identificaron durante esta etapa fueron los siguientes:



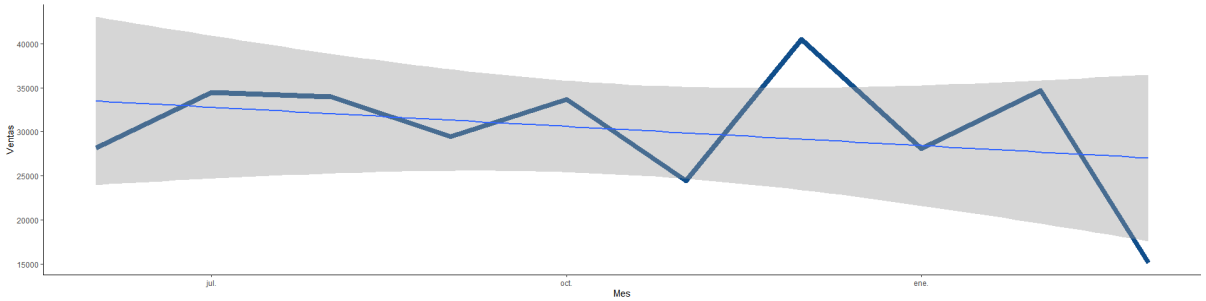




- Los modelos más demandados durante el periodo 2018-2019 fueron el N.MOG6PYI, N.MOG6PLI y N.MZ3PLYAO, mientras que los menos demandados fueron el N.MOTOEGRB, N.MOTOEDRB, N.MZ2PLYGB.
- Por otra parte, la gráfica de pastel muestra que hay una mayor demanda de la telefonía de gama baja.
- Finalmente, la última gráfica demuestra que la mayor cantidad de ventas de cualquier tipo de celular se concentra en el centro sur con un 51.2%. Por el contrario, el que genera menor cantidad de ventas es la zona pacifico sur con un 2.2% de ventas.

3) ¿Comportamiento de datos en el tiempo? etc. (Incluir gráficas) Seleccionar y construir (ingeniería de características) variables para la etapa de modelado

Con el objetivo de entender de manera visual la demanda a través del tiempo y así generar un pronóstico se realizó una gráfica en donde se refleja la cantidad de productos vendidos entre finales de 2018 y principios del 2019.



Gráfica 4. Series de tiempo

De esta manera, la línea azul oscura representa el número de ventas realizadas durante un año. Se puede observar que no existen patrones por lo que los valores varían de manera independiente. En cambio, la línea gris muestra el área que pueden tomar los valores de la

demanda. Se podría decir que en diciembre existen valores de demanda atípicos ya que se encuentran fuera de esta área.

Una vez realizadas las actividades anteriores procedimos a realizar la ingeniería de características en la que se le asignó un índice a cada registro y después se crearon nuevas características como, “ventas por tienda”, “ventas por producto” y crear rezagos con la finalidad de mejorar el desempeño de los algoritmos propuestos a la solución problemática sobre cumplir con la demanda en los diferentes puntos de venta de motorola.

Para realizar el pronóstico con el método de promedios móviles simples se consideró que es necesario realizar el análisis utilizando todas las variables y seleccionando los rezagos de los 3 tiempos disponibles con la finalidad de comparar cual es el mejor pronóstico de acuerdo a las métricas de desempeño.

Etapa 4: Modelado

Modelo de Promedio Móviles
Los promedios móviles son promedios calculados a partir de subgrupos artificiales de observaciones consecutivas. En las gráficas de control, se puede crear una gráfica de promedio móvil para los datos de tiempo ponderados. (Minitab, 2019).

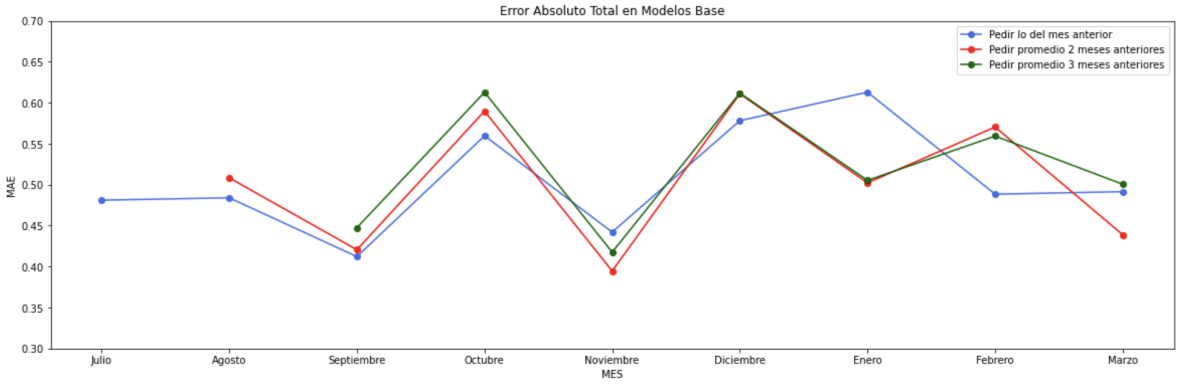
El pronóstico de promedio móvil es óptimo para patrones de demanda aleatorios o nivelados donde se pretende eliminar el impacto de los elementos irregulares históricos mediante un enfoque en períodos de demanda reciente.[6]

La utilización de promedios móviles implica la elección arbitraria de su longitud u orden, es decir, del número de observaciones que intervienen en el cálculo de cada media móvil. Cuanto mayor sea la longitud, mejor se eliminarán las irregularidades de la serie, ya que al intervenir más observaciones en su cálculo se compensarán las fluctuaciones de este tipo, pero por el contrario, el coste informativo será mayor. Por el contrario, cuando la longitud es pequeña, la media móvil refleja con mayor rapidez los cambios que puedan producirse en la evolución de la serie. Es conveniente, pues, sopesar estos factores al decidir la longitud de la media móvil.[7]

Las limitantes que tiene este método es que solamente pueden llegar a pronosticar un período más y suelen ser simplificaciones reales y no garantizan las variables influyentes en el futuro de los pronósticos que se encuentren incluidos en el modelo de dicho pronóstico.

Para realizar los pronósticos utilizando esta metodología se utilizó el programa Jupyter Notebook y el lenguaje de programación Python para gestionar el gran volumen de la base de datos.
Se usaron 3 modelos diferentes: usando solo lo del mes anterior, calculando un promedio de los dos meses anteriores y por último, calculando un promedio de tres meses anteriores. Tomando en cuenta estos 3 pronósticos, se calculó el Error Absoluto Total, comparando las predicciones que los modelos arrojaban contra los datos reales y así ser capaces de tomar una decisión respecto al mejor modelo a utilizar.

Al momento de construir la gráfica con los MAEs de los 3 modelos (gráfica 5), podemos observar que este método en general es muy poco fiable en cuanto a predecir las fluctuaciones de mes a mes. En un plano general, podemos ver que de un mes a otro el error aumenta y disminuye de forma importante. En los meses de Octubre y Diciembre es en donde, en promedio, se encuentra el mayor error. En cuanto a Diciembre, se podría atribuir a las festividades que tienen lugar en este mes, se puede asumir que las ventas aumentan en este mes ya que es cuándo la gente tiene más dinero que gastar. En Enero, se puede observar que entre modelos hay gran variación, siendo el azul (pedir lo del mes anterior) el que presenta mayor error. Esto se puede atribuir a que, aunque las ventas siguen siendo altas, no lo son como lo son en Diciembre, por lo que tomar un promedio entre las ventas altas de Diciembre y los números estables de los meses anteriores, dará un pronóstico más acertado a las ventas reales del mes de Enero. Se puede observar que el modelo azul es el que tiene mayor diferencia con las ventas reales del mes, por lo que puede ser descartado. Comparando los modelos que toman un promedio de los 2 y 3 meses anteriores, en la gráfica se puede observar que, más a menudo, el modelo que toma el promedio de los 2 meses anteriores presenta un menor Error Absoluto Total. Este modelo se utilizó para compararlo con el modelo de aprendizaje de máquina de árboles de decisión con validación cruzada, y así ser capaces de tomar una decisión, evaluando el desempeño de los distintos modelos.



Gráfica 5. Error Absoluto Total en Modelos Base para 3 modelos de promedios móviles

Resultado de Desempeño

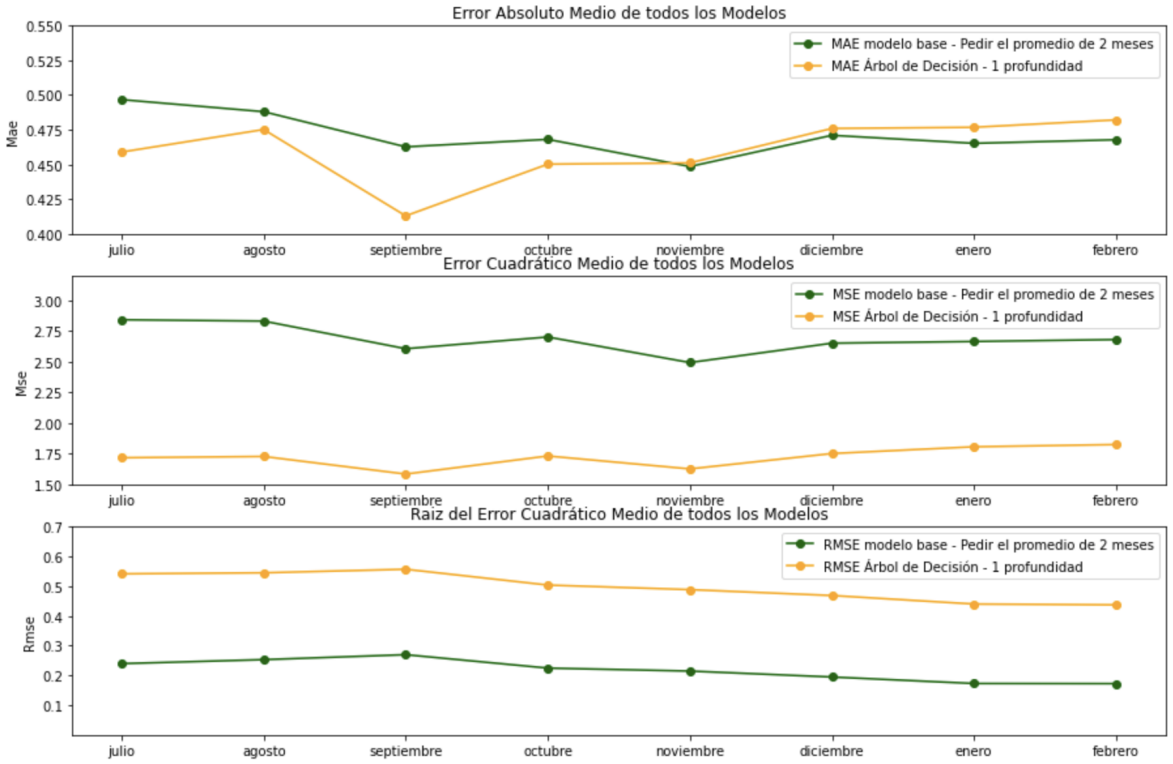
El método de Promedios Móviles no dio excelentes resultados en este caso, sin embargo, hay una posibilidad de que esto sea debido a que los datos presentan un tipo de estacionalidad. Este método podría ser mucho más acertado si se contaran con datos de ventas de varios años, para así, en lugar de comparar un mes con el mes anterior (es), comparar un mes con el mismo mes pero de distinto año. Por ejemplo, si se conoce que las ventas cada Diciembre aumentan de forma importante, podríamos usar los promedios móviles para sacar un promedio de las ventas de Diciembre de 2 o 3 años anteriores y así predecir las ventas del siguiente Diciembre.

Árbol de Decisión con Validación Cruzada para Series de Tiempo

La validación cruzada es una técnica que funciona para evaluar el desempeño de modelos de Machine Learning mediante el entrenamiento de diferentes conjuntos de datos. Esta

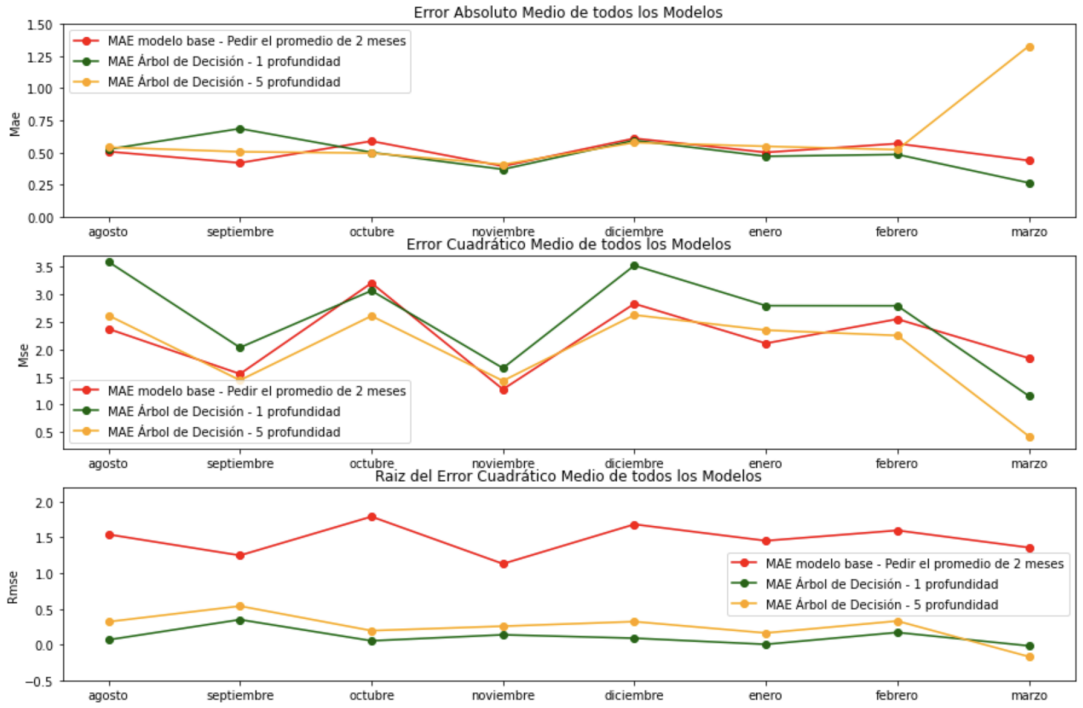
técnica ayuda a comparar y posteriormente seleccionar el modelo en el aprendizaje de máquina aplicado.
La validación cruzada k-fold significa que el conjunto de datos se divide en un número K. Divide el conjunto de datos en el punto en el que el conjunto de pruebas utiliza cada pliegue.
Las limitantes de este modelo es la alta probabilidad de la existencia de un sobreajuste o un infra ajuste. El modelo puede generar predicciones precisas con los nuevos datos cuando el modelo se ajusta perfectamente al conjunto de datos. Un algoritmo adecuado para el conjunto de datos entrenado puede ayudar a entrenar el nuevo conjunto de datos. Por otra parte, si el modelo de aprendizaje automático se basa en un proceso de entrenamiento no ajustado, no generará datos precisos ni predicciones adecuadas. Por lo tanto, el modelo no podrá procesar los patrones importantes de los conjuntos de datos.
Si el modelo se detiene durante el proceso de entrenamiento, se producirá un infra ajuste. Esto indica que los datos necesitan más tiempo para ser procesados completamente. Esto afectará al rendimiento del modelo para los nuevos datos. El modelo no producirá resultados precisos y no será útil.[8]
Para realizar los pronósticos utilizando esta metodología se utilizó el programa Jupyter Notebook y el lenguaje de programación Python para gestionar el gran volumen de la base de datos.
Se generaron dos conjuntos de datos diferentes que se utilizaron para el entrenamiento del modelo y para su prueba. Posteriormente, se dividieron ambos conjuntos en variables de entrada (x) y variable de salida o de respuesta (y). Se repitieron estos dos pasos para las 8 particiones que se realizaron.
Para entrenar el modelo, se realizaron ocho particiones. Para estas, se construyó el árbol con una profundidad máxima de 1. Posteriormente se entrenó el modelo con los conjuntos de datos que anteriormente se generaron. Se calculó el error de entrenamiento que se presenta y se calculan los límites máximos y mínimos de la predicción. Por último se juntan datos para finalmente calcular los errores MAE, MSE y R^2.
Se repitieron los pasos anteriores para las ocho particiones que se realizaron.
Para probar los modelos, en cada partición se siguió el siguiente proceso:
Primeramente, se calculó el error de prueba que se presenta. Posteriormente, se calculan los límites máximos y mínimos de la predicción para después generar la predicción y obtener una respuesta a nuestra pregunta inicial (¿Cuántas unidades se venderán en el mes de prueba (en todos los puntos venta y considerando todos los productos))?
Como en el entrenamiento, posteriormente se juntan los datos para calcular el error, se calculan los errores MAE, MSE y R^2, y se repite este proceso para las ocho particiones.
Al obtener los resultados de los modelos, estos se registraron en Excel. Posteriormente, al leerlos en Python, se separan entre datos de entrenamiento y prueba, así como por métrica (MAE, MSE, R2).
Finalmente, se graficaron los errores de los datos de entrenamiento y de prueba.

Resultado de Desempeño



Gráfica 6. Error Absoluto Medio de los modelos de entrenamiento

En las gráficas 6, se puede observar que aquella que presenta un menor rango y mayor exactitud es el MAE ya que sus valores se encuentran entre 0.4 y 0.5 mientras que las demás métricas muestran una mayor amplitud entre su modelado base y árbol de decisión o tienen valores elevados, arriba de 1.5 o 3 lo cual no es confiable. Además, la variación entre el modelo base vs el árbol de decisión del MAE es muy parecido, a excepción de septiembre que muestra un valor atípico.



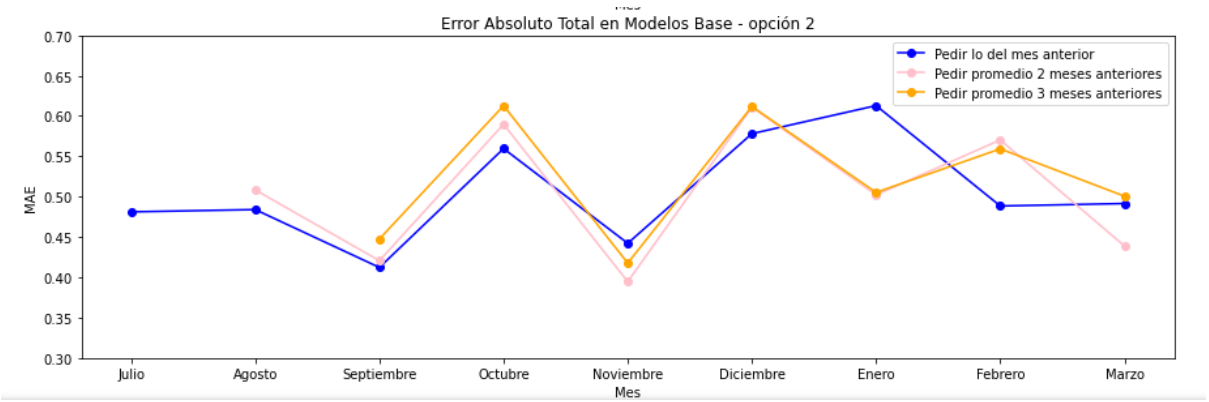
Gráfica 7. Error Absoluto Medio de los modelos de entrenamiento

En las gráficas 7, se observa que la métrica más adecuada para utilizar en este modelo es la del MAE ya que sus valores entre su modelo base y los árboles de decisiones de diferentes profundidades son muy similares a través del tiempo. Por otro lado el MSE, al igual que el MAE, presenta valores parecidos en sus modelos a través del tiempo sin embargo los mismos están entre 1.5 y 3 lo cual no son muy confiables. Finalmente, la gráfica del RMSE presenta una amplitud muy alta entre su modelo base y árboles de decisión de diferentes profundidades.

Etapa 5: Evaluación

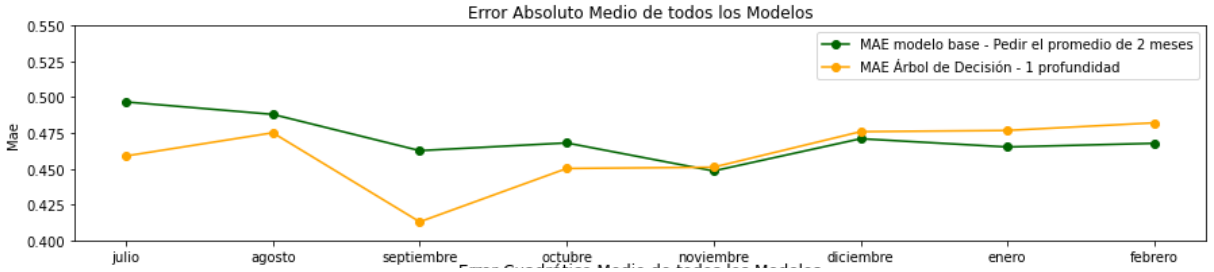
El propósito del proyecto consiste en predecir la demanda de unidades por punto de venta de la telefonía celular Motorola, a través de un proyecto de ciencia de datos donde se realizó una comparación de modelos tradicionales contra modelos de aprendizaje de máquina con la finalidad de obtener un mejor pronóstico. A lo largo del semestre se realizaron diferentes actividades que abarcaron la comprensión del negocio, comprensión y preparación de de los datos y modelado de los mismos.

Durante la etapa de modelado se realizó una comparación de las ventas registradas utilizando el método de los promedios móviles simples, modelo óptimo para la predicción de patrones aleatorios de demanda empleando datos históricos de periodos de demanda reciente. Para este proyecto se utilizaron promedios móviles simples con periodos de uno a tres meses. Asimismo, se estableció en el modelo como métrica primaria a utilizar el error absoluto medio, el cual es calculado mediante la diferencia entre un valor observado medido y su valor verdadero. Después de analizar la gráfica se decidió seleccionar el promedio móvil simple de 2 meses anteriores ya que mostraba un menor error y variación en comparación de los demás.



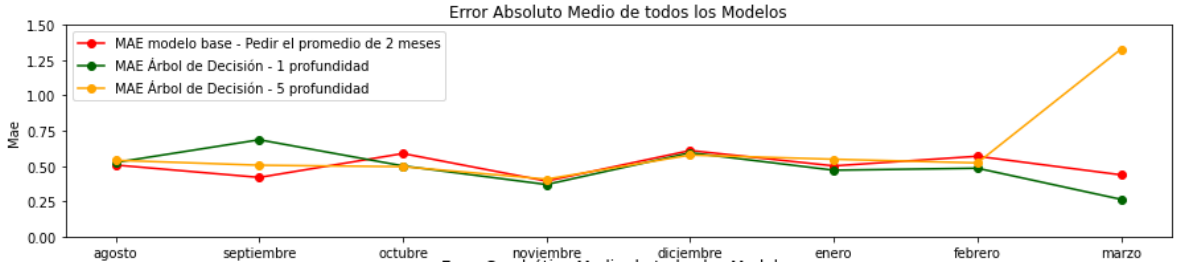
Gráfica 8: Comparación del método de promedio móvil simple.

Posteriormente, se realizó la implementación de modelos de aprendizaje de máquina con validación cruzada para series de tiempo, la cual ayuda a evitar el sobreajuste de los datos del proyecto realizando una técnica de ocho particiones calculando los errores de entrenamiento y prueba respetando la temporalidad de los mismos. De igual manera, se realizó una regresión con el modelo de aprendizaje de máquina de árboles de decisión con uno y cinco nodos de profundidad para realizar la predicción.



Gráfica 9: MAE de entrenamiento de PMS y árbol de decisión.

De la misma manera que con el método de promedios móviles simples el error absoluto medio fue seleccionado como métrica ya que presentaba un menor rango de error y exactitud con el modelo del árbol de decisión. Se puede observar que en el conjunto de entrenamiento el modelo de árbol de decisión con nodo 1 de profundidad muestra un menor error que el modelo base en la mayoría de sus puntos. Mientras que en el conjunto de prueba se muestra de la siguiente manera:



Gráfica 10: MAE de prueba de PMS y árbol de decisión.

Se puede observar cada uno de los modelos graficados con su respectivo error. El modelo base (promedios móviles simples) muestra un error muy parecido al del árbol de decisión con un nodo de profundidad ya que los dos se encuentran en rangos de 0.25 a un poco más de 0..50, mientras que el modelo del árbol de decisión con nodo 5 de profundidad muestra en general errores similares a los otros dos modelos a excepción del último mes donde aumenta drásticamente obteniendo un rango de error por arriba de 0.25 a más de 1.25.

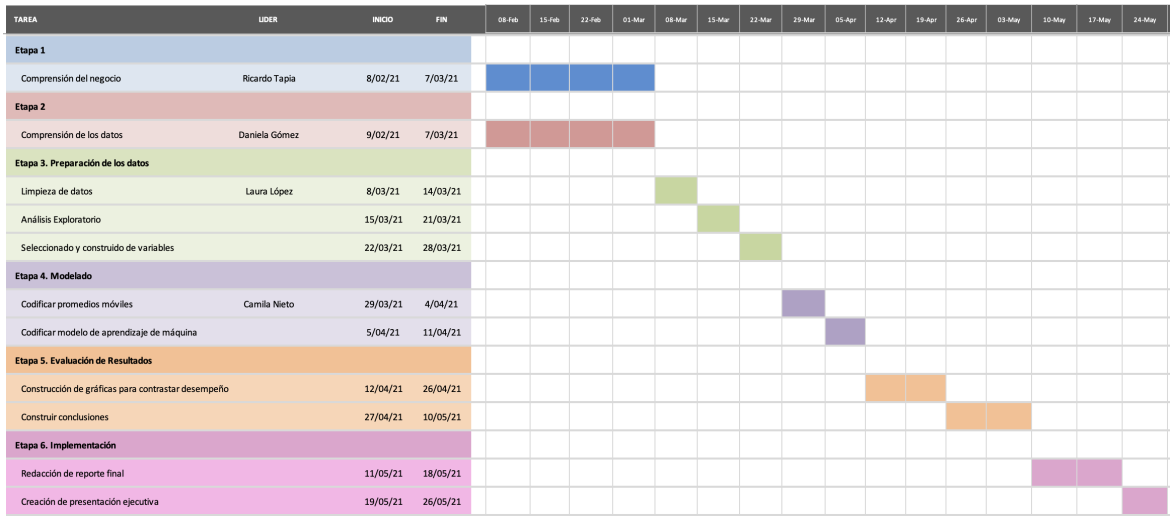
Por lo que se recomienda que Motorola utilice el método de aprendizaje de máquina de 1 nodo de profundidad, con la finalidad de obtener una predicción más certera de la demanda de unidades móviles ya que después de analizarse detalladamente es el método que muestra menor error y variación.

Por último, se realizó una comparación del modelo seleccionado (árbol de decisión con un nodo de profundidad) contra el registro real de ventas. El modelo seleccionado mostró un pronóstico para el mes de marzo de 9,174 unidades vendidas en los diferentes puntos de venta (Anexo 2), mientras que los registros reales mostraron ventas de 15,174 unidades (Anexo 3) abarcando una diferencia de 6,000 unidades , las cuales son explicadas por el error absoluto medio (MAE).

En conclusión, cada etapa de nuestro Proyecto de Ciencia de Datos fue exitosa ya que, finalmente pudimos proponer un modelo cuyo objetivo sea el pronóstico de ventas en todos los puntos de venta de Motorola, el modelo tiene un buen desempeño de acuerdo a las métrica usada y es por eso que se propone a la organización que haga uso del mismo, sin embargo, creemos que se puede mejorar el modelo al tener más registros y con mayor entrenamiento con el fin de tener un mejor desempeño, además también proponemos la posibilidad de usar un modelo de aprendizaje de máquina con mayor robustez para predecir de manera más exacta la demanda futura en los distintos puntos de venta de Motorola y no incurrir en costos adicionales, además de ofrecer un buen servicio al cliente.

Anexos

TAREA	LIDER	INICIO	FIN
Etapa 1			
Comprensión del negocio	Ricardo Tapia	8/02/21	7/03/21
Etapa 2			
Comprensión de los datos	Daniela Gómez	9/02/21	7/03/21
Etapa 3. Preparación de los datos			
Limpieza de datos	Laura López	8/03/21	14/03/21
Análisis Exploratorio		15/03/21	21/03/21
Seleccionado y construido de variables		22/03/21	28/03/21
Etapa 4. Modelado			
Codificar promedios móviles	Camila Nieto	29/03/21	4/04/21
Codificar modelo de aprendizaje de máquina		5/04/21	11/04/21
Etapa 5. Evaluación de Resultados			
Construcción de gráficas para contrastar desempeño		12/04/21	26/04/21
Construir conclusiones		27/04/21	10/05/21
Etapa 6. Implementación			
Redacción de reporte final		11/05/21	18/05/21
Creación de presentación ejecutiva		19/05/21	26/05/21



Anexo 1: Diagrama de Gantt con plan preliminar del proyecto final.

```
In [182]: # ¿Cuántas unidades se van a vender (predicción) en ese mes de prueba (en todos los puntos de venta , considerando todos los
          pred8_prueba.sum()

Out[182]: 9174.0
```

Anexo 2: Unidades pronosticadas a vender en la Variable Y de la octava partición.

mes_id	ventas_totales
9	15174

Anexo 3: Unidades vendidas en Marzo.

Bibliografía

1. Masetto, A. (2021). Tema 1 Introducción [Diapositiva de PowerPoint]. Canvas. https://experiencia21.tec.mx/courses/124327/files/40585026?module_item_id=6437244
2. IFT. (2020). Resumen de Indicadores Trimestrales. mayo 26, 2021, de IFT Sitio web: https://bit.ift.org.mx/SASVisualAnalyticsViewer/VisualAnalyticsViewer_guest.jsp?reportSBIP=SBIP%3A%2F%2FMETASERVER%2FShared%20Data%2FSAS%20Visual%20Analytics%2FReportes%2FResumen%20de%20Indicadores%20Trimestrales(Report)&page=vi1568&sso_guest=true&informationEnabled=false&commentsEnabled=false&alertsEnabled=false&reportViewOnly=true&reportContextBar=false&shareEnabled=false

3. IFT. (2020). SEGUNDO Informe Trimestral Estadístico 2020. mayo 26, 2021, de IFT Sitio web: http://www.ift.org.mx/sites/default/files/contenidogeneral/estadisticas/ite2t2020.pdf
4. Skualo. (Agosto 20, 2018). Motorola: la historia de esta compañía estadounidense.. marzo 06, 2021, de LuisGyG Sitio web: https://luisgyg.com/motorola/
5. Piedras, E. (diciembre 04, 2019). El Mercado de Smartphones Hoy. marzo 06, 2021, de CIU Sitio web: https://www.theciu.com/publicaciones-2/2019/12/4/el-mercado-de-smartphones-hoy

6. Estadística y Machine Learning con R. (Enero 25, 2019).Series Temporales. abril 29, 2021, de Parra, F. Sitio web:https://bookdown.org/content/2274/series-temporales.html
7. Ingeniería Industrial Online. (Junio 30, 2019).Promedio móvil. abril 29, 2021, de Salazar, Bryan Sitio web: https://www.ingenieriaindustrialonline.com/pronostico-de-la-demanda/promedio-movil/
8. Team, D. S., & Team, D. S. (2021, March 24). *Validación cruzada K-Fold - Aprendizaje automático*. DATA SCIENCE. https://datascience.eu/es/aprendizaje-automatico/validacion-cruzada-de-k-fold/.