Laboratorio de Diseño y optimización de operaciones.

PROYECTO FINAL

Entrega Final

Equipo 4

DATALENTED

Integrantes:

Daniela Monserrat García SoteloA01365499
Jorge Abraham Sanchez Mora A01364653

Introducción y Etapa 1: Comprensión del Negocio

1) Descripción de la situación actual

La industria de las telecomunicaciones se ha desarrollado de manera drástica durante la última década. Si bien es cierto, la globalización ha sido un factor determinante en el incremento de la demanda en este ámbito, especialmente en el sector de telefonía celular. Conforme pasa el tiempo, la gente requiere el acceso a tecnologías que le brinden acceso a información y herramientas para su desempeño, posicionándose en la búsqueda de aquellos dispositivos que se adecuen más a sus necesidades. Debido a esto, las empresas se han visto obligadas a mejorar e incrementar sus procesos de producción y logística.

Apple, una de las empresas más exitosas del mundo , productora de equipos electrónicos, software y servicios, es víctima de esta problemática y busca desarrollar la metodología adecuada para satisfacer las demandas netas de los clientes.

Desde el año 2012, Apple empezó a tener un peso muy importante en el mercado mexicano, convirtiéndose en la empresa con el mayor porcentaje de demanda en el país, superando a marcas como Huawei y Samsung. (nótese en la gráfica siguiente).

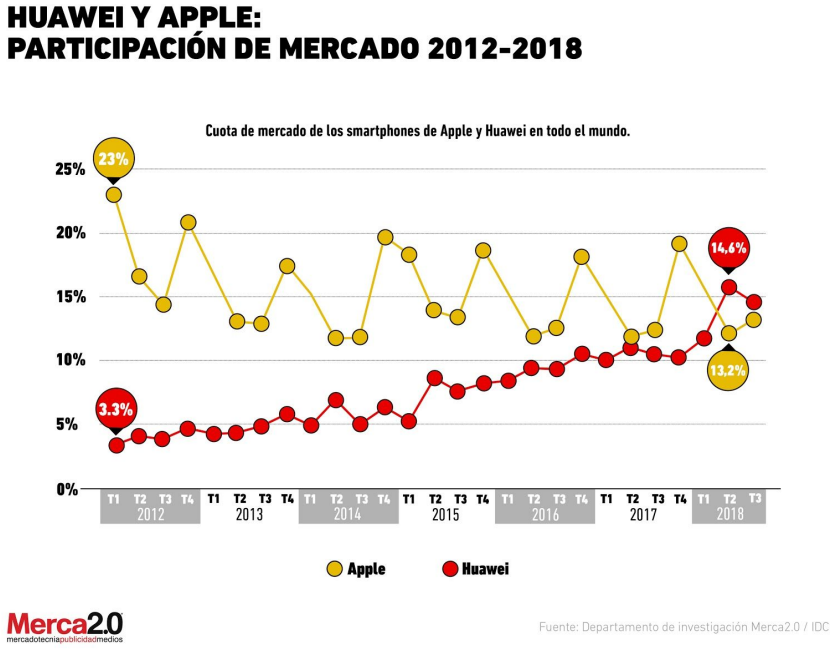


Figura 1. Participación de Mercado (Merca 20, 2018)

En la figura 1, se puede observar claramente que el comportamiento de la demanda de productos Apple se ha mantenido constante durante estos años, teniendo picos positivos y negativos. Este tipo de patrón se debe a distintos motivos que ha tenido que enfrentar esta marca a lo largo del tiempo, entre ellos, está que la gente adoptó una ideología diferente respecto a estos dispositivos, puesto que llegó un punto, mencionado por el mismo Steve Jobs, en que las personas ya no eligen esta marca por utilizar un teléfono, sino por la mezcla entre su aspecto, su fama, su belleza y su modus operandi. Por ello, Apple se vió obligado a realizar diversas modificaciones en sus productos, tratando de adecuarse a las especificaciones demandadas por los clientes, lo cual para ninguna empresa ha sido fácil.

Consecuentemente, el pronosticar la demanda de los clientes, respecto a las revoluciones tecnológicas, actualizaciones y nuevos atractivos, es una tarea muy complicada para los productos en este sector. Por ello, los ingenieros industriales son los responsables de realizar este tipo de tareas, empleando herramientas estadísticas, conocimientos de logística y sistemas, para determinar posibles futuras demandas en cualquier mercado, dando a las empresas un estudio sustentado en el que puedan tomar decisiones con mayores beneficios y menores riesgos de fracaso y/o pérdida.

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

Como se mencionó anteriormente, Apple presenta una de las mayores problemáticas dentro de su industria, la predicción de demanda. Existen diversos factores que propician los cambios drásticos en las necesidades de las personas y su enfoque en la búsqueda de dispositivos celulares que las satisfagan. Actualmente, las personas ya no solo buscan un teléfono cualquiera, se trata de desarrollar un dispositivo que esté al alcance de cualquier usuario, que pueda interactuar con él de manera inteligente. Por otro lado, debido a la alta competencia en este sector, las empresas también deben ser capaces de producir elementos que, en términos financieros, pueda ser accesible a la mayor parte de los usuarios y así obtener mayores clientes. No obstante, esta no es una característica que se le pueda evidenciar a Apple, pues, a pesar de tener precios altamente elevados en sus teléfonos, ellos han decidido enfocarse en el teléfono más “preciso, inteligente, dinámico y estético” para su clientela, lo cual lo ha ayudado a mantener un ranking elevado de demandas.

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

Como equipo, tenemos la tarea de analizar una gran cantidad de datos referentes a los historiales de demanda de productos Apple en diversos puntos de venta, alrededor de México. Nos enfrentamos a un problema de regresión para pronósticos futuros, además de un conjunto de datos estructurados. Como se vió en clase, los datos estructurados son aquellos fáciles de almacenar, ordenar y trabajar con ellos, que pueden ser colocados en tablas; gracias a que nuestros datos son estructurados y no “no estructurados” es que nuestro trabajo se vuelve más sencillo, puesto que si fueran no estructurados (como imágenes, videos,etc.), tendríamos que utilizar otras técnicas de inteligencia artificial.
Nuestro conjunto de datos se debe estudiar, limpiar y modificar de manera correcta para lograr un uso adecuado de estos; una vez realizada esta etapa, utilizando la metodología CRISP-DM y herramientas de programación como Rstudio, se propondrá y desarrollará un modelo que nos permita resolver un problema de construcción de portafolios de productos, es decir, predecir las demandas de telefonía celular de la marca Apple, también denominados iPhone, en los próximos meses.

4) Plasmar objetivos

Nuestros objetivos, como industria, es satisfacer a nuestros clientes, sus necesidades y demandas, ya que, estamos conscientes de la gran variedad de competencias que existen en este sector, que cada una de las empresas ofrece elementos muy entretenidos para el cliente y sabemos que las demandas serán muy variadas dependiendo de sus gustos y satisfacciones, entre otros factores.Es por ello que consideramos estos principales objetivos:

a) Satisfacer los gustos y necesidades del cliente.
b) Crear y desarrollar productos innovadores que se adapten a las nuevas tendencias tecnológicas.
c) Ofrecer un producto que pueda combinar belleza, inteligencia, innovación y accesibilidad al mismo tiempo.

“Datalented” se ve obligado a trabajar con datos estructurados de historiales de demanda de productos marca Apple para poder pronosticar, mediante el uso de la metodología CRISP-DM y herramientas de software como RStudio, las demandas futuras para una de las empresas con la telefonía celular más desarrollada en la actualidad, mejor conocida como iPhones. Nuestros principales objetivos son:

a) Familiarizarse, a profundidad, con el uso de la herramienta de software Rstudio

b) Estudiar a detalle la metodología CRISP-DM, para trabajar en base a las fases contenidas en este modelo.
c) Realizar la limpieza necesaria de nuestros datos estructurados.
d) Desarrollar el procedimiento adecuado solicitado por nuestro asesor y en base a nuestra metodología.
e) Pronosticar la demanda de productos marca Apple, para los próximos meses.
f) Obtener conclusiones acerca del proyecto realizado.
g) Exponer, mediante el uso correcto de un storytelling, la información final recabada.

5) Estructurar el proyecto y hacer un plan preliminar

La estructura y plan preliminar del proyecto se muestra en el documento de excel anexo.

Etapa 2. Comprensión de los datos

1) **Describir los datos crudos**

Se nos dio una tabla de datos estructurados con 102 530 datos, dentro de los cuales, en cada columna se encuentra la siguiente información:

Puntos de venta: Refieren al lugar dónde se ubican los productos apple.
Fecha: Refiere al día que se entregó el producto.
Anio: Es el año en el que el producto se vendió.
Número ventas: Es el número de unidades vendidas que se obtuvo.
Costo promedio: Costo total promedio del producto.
Marca: Refiere a la marca del producto, es decir, el modelo del mismo.
Gamma: Indica el nivel del producto.
Zona: Zona o lugar donde se vendió el producto.
Estado: Estado, del país, donde se vendió el producto.
Ciudad: Ciudad donde se vendió el producto.
Latitud: Latitud se refiere a aquella donde se encuentra el punto de venta
Longitud: La longitud es el punto geográfico del punto de venta.

2) **Detectar los problemas de calidad**

- **Fecha:** Los registros están limpios.
- **Año:** Debe tomar valores de 4 dígitos
- **Número de ventas:** Los registros están limpios.
- **Puntos de venta:** Existen 5 puntos de venta de manera errónea, ya sea por mayúsculas, acentos, etc.

- **Variable Mes:** Es una variable numérica, hay que cambiarla. Hay 5 valores mal registrados, en lugar de números, hay letras.
- **Variable Sku**: Los registros están limpios.
- **Variable Marca:** Hay 5 marcas escritas de forma errónea.
- **Variables Costo, Ciudad y Gamma:** Registros limpios.
- **Variable Zona**: Hay un registro mal escrito
- **Variable Estado**: Hay tres estados más de los que en realidad existen.
- **Variables Latitud y Longitud**: Hay un valor fuera de rango.

Etapa 3: Preparación de los datos

Al ya tener los conocimientos suficientes aprendidos durante la primera entrega, durante este segundo periodo, como equipo, nos enfocamos en cuatro principales aspectos: la limpieza de los datos, el análisis exploratorio de ellos, la ingeniería de características y promedios móviles simples. Dando seguimiento a nuestra ideología base, CRISP-DM, y empleando herramientas de programación como R Studio e implementando una nueva herramienta llamada Jupyter Notebook, nos dispusimos a realizarlo.

Limpieza de Datos

Previo a la limpieza de los datos, tuvimos que estudiar y analizar de manera detallada las características de los datos, es decir, cuántos de estos existían y cuántos de estos realmente eran significativos puesto que muchos de ellos eran repetitivos. Esto se debía a que, al registrar los datos, los operadores pudieron tener errores de distracción o descuidos, ocasionando una escritura errónea de los nombres en las filas y/o columnas, cambio de letras por números o viceversa (como fue en el caso de los meses), faltas de ortografía, tildes inapropiadas, etc.

Más específicamente, en la variable de año, existían algunos valores numéricos de dos dígitos y otros de cuatro dígitos, por ello tuvimos que analizar estos datos y conocer cuáles eran los valores mal escritos y corregirlos, igualmente en los puntos de venta y las marcas existieron palabras que estaban escritas en mayúscula, y debido a esto el programa con el que trabajamos lo consideraba como un elemento diferente a los mismos puntos de venta escritos en minúsculas. Personalmente, como equipo consideramos a estas variables las más complicadas de limpiar debido a que era una gran cantidad de nombres, por ende eran varios errores que

modificar, no obstante gracias a nuestras herramientas de software, logramos emplear funciones de programación que nos permitían identificar los datos “únicos”, es decir el programa buscaba entre todos los datos, aquellos que no se repitieran y nos los mostraba, así, podríamos ver aquellos mal escritos o escritos de manera diferente. Variables como el mes, eran variables numéricas, las cuales tuvimos que cambiar por letras para un mejor manejo de estos.

Cabe resaltar que al momento de corregir los estados, existían variables que realmente no eran estados, es decir, se trataban de zonas dentro de estos estados, para ello también tuvimos que recurrir a una investigación geográfica para localizar estos lugares y poder cambiar esta variable por el estado correspondiente.

Por otro lado, dentro de las variables longitud y latitud existían valores que se encontraban fuera de rango, para ello recurrimos rápidamente a nuestro excel, como una manera de identificarlos, añadimos dos columnas más y en estas ordenamos de mayor a menor nuestros datos de longitud y latitud, estando hasta abajo o hasta arriba aquellos valores fuera de rango. Una vez identificados, regresamos a nuestro programa e igualmente identificamos estos valores con funciones de programación, corroborando estos mismos. Para modificar estos valores, nos fijamos en el punto de venta de estos, puesto que, lógicamente tendrían que tener los mismos valores los mismos puntos de venta, y así, cambiamos estos valores por los de otros cuya longitud y latitud estuvieran correctas.

La cantidad de datos que nos fueron proporcionados era muy grande, por lo tanto, analizar estos errores de manera directa y manual nos llevaría demasiado tiempo. Si bien es cierto, nuestro asesor pedagógico menciona que, generalmente, al tratarse de este tipo de proyectos, los analíticos o especialistas en ciencia de datos invierten una gran parte de su tiempo en esta limpieza. Por consiguiente, y como se mencionó anteriormente, se utilizó R Studio, una herramienta de programación que nos permitía observar de manera simple este tipo de errores; con simples funciones como “*unique”* o *“select”,* ya éramos capaces de identificar posibles equivocaciones en los datos.

Finalmente, empleando diversas funciones y programaciones, logramos limpiar de manera correcta cada uno de los datos, reduciendo completamente los errores de registros y así dar seguimiento al proyecto y tener información cien por ciento verídica y confiable. Para corroborar que todos nuestros datos estuvieran correctos, dentro de este mismo programa en RStudio, volvimos a correr nuestras funciones donde mostraba los datos y efectivamente observamos que las únicas variables que aparecían en cada una de las columnas eran realmente las que tenían que estar, años con 4 dígitos, únicamente estados de la república, marcas y puntos de venta escritas correctamente, sin acentos y sin mayúsculas, meses escritos en letras y no en números y finalmente longitudes y latitudes dentro de los rangos apropiados.

ANÁLISIS EXPLORATORIO DE DATOS

Al finalizar la limpieza de nuestros datos, desarrollamos un análisis exploratorio de estos, el cual permite tener una perspectiva o una visualización de las circunstancias de nuestros datos en la realidad, permitiéndonos entender de manera más profunda la situación.

Para ello, recurrimos a nuestra herramienta R Studio nuevamente, dentro de la cual implementamos nuevas librerías y funciones que nos permitían agrupar los datos de acuerdo con los elementos que quisiéramos analizar y poder graficarlos para una mejor representación. La librería que utilizamos fue ggplot2, dentro de la cual existían numerosas maneras de representar los datos, mediante diversos gráficos, como gráficos de caja, gráfico de barras, histogramas, de puntos, etc.

Cabe resaltar que, a pesar de agrupar los datos para reducir las variables que se fueran a graficar y evitar múltiples datos representados, algunas de estas segmentaciones eran aún muy grandes, como en el caso de los puntos de ventas.

Primeramente, quisimos explorar nuevos modelos y no empezar de lleno con los que ya conocemos, por ello realizamos un gráfico bin2d mostrado a continuación, el cual representa todos los puntos de venta alrededor del país dependiendo del Estado de la República en el que se encuentre. Consideramos importante reconocer estos elementos en una gráfica para conocer y concientizar a qué estados se debería enviar más mercancía debido a la cantidad de puntos de venta en él. Aunque no se aprecian todos los nombres de los distintos puntos de venta en la imagen, se puede observar en un primer plano, qué estados cuentan con más tiendas con gran facilidad, por ejemplo, la Ciudad de México y el Estado de México, puesto que son los renglones con más cajas azules. La figura 2 es anexada en la parte final de este documento para su mejor entendimiento.

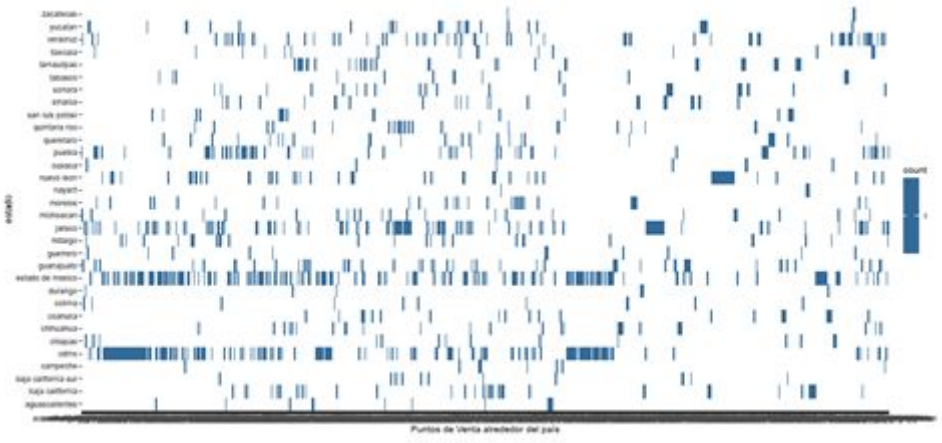


Figura 2. Gráfico Bin2d (Puntos de venta vs. Estados)

Posteriormente, realizamos un gráfico de puntos y líneas donde estudiamos las variables gamma del teléfono y el costo promedio referente a ello (mostrado en la parte inferior del párrafo). Esta información nos permite tener una perspectiva de la cantidad de valor que tienen este tipo de dispositivos, al ser una empresa de gran calidad, solo cuentan con una gamma alta o gamma premium, por lo tanto, los precios en los teléfonos van desde los 12000 pesos hasta los 25000 aproximadamente. Además, se observa una varianza casi constante entre los costos en cada gamma, no obstante, se aprecia un ligero “montón” de datos en la gamma alta, tocando los 15000 pesos, esto quiere decir que hay más cantidad de estos productos en gamma alta con estos costos, nuevamente dándonos una idea de las posibles demandas a cumplir.

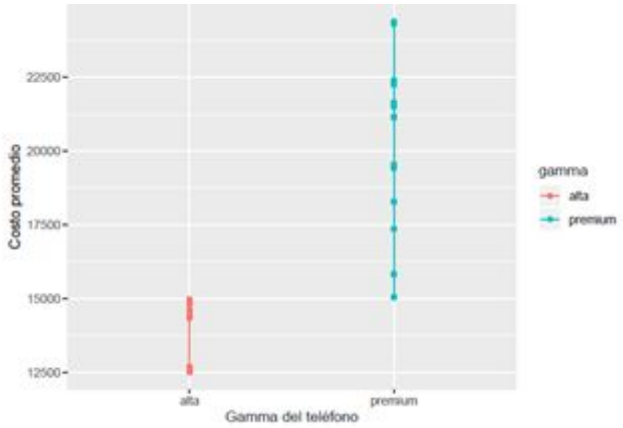


Figura 3. Gráfico de puntos y líneas (Gamma del teléfono vs. Costo de producto)

Finalmente construimos un gráfico boxplot (mostrado en la parte inferior) para realizar un análisis más detallado respecto a la distribución de las variables “costos promedios” y “Estados”, con el fin de saber si realmente el hecho de estar en una región diferente ha ocasionado un aumento o disminución en los costos, lo cual es muy usual (aumento) cuando se trata de tiendas situadas en Polanco, Antea, etc., dando un ejemplo contextual.

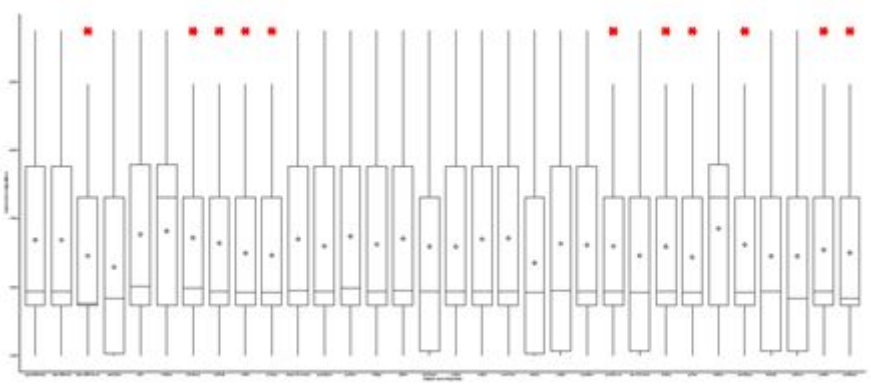


Figura 4. Gráfico boxplot (Costos promedios vs. Estados)

Ingeniería de Características

Para esta etapa, es necesario recordar algunos conceptos. La ingeniería de características busca tomar un conjunto de datos y con estos construir variables características o explicativas que se puedan emplear posteriormente en un modelo de aprendizaje de máquinas, con el fin de resolver cualquier problema de predicción o clasificación. Empleando R Studio, utilizamos índices en nuestras variables cualitativas más relevantes, con el fin de manejarlas de una manera más accesible, como fueron “mes_id”, “pvd_id”, “sku_id”, sin embargo, las variables como marca o la gamma no fueron necesarias puesto que son variables de una sola característica y otras variables como la zona, el estado o la ciudad estaban implícitas en la tienda (punto de venta). Después agrupamos los datos de las variables, es decir, si existía más de una venta del mismo producto, en el mismo punto de venta, en la misma fecha, etc, con el fin de tener una mejor perspectiva de nuestros datos, pues con esos datos sabríamos las ventas totales por zonas. Cabe resaltar que, tuvimos que remover conjuntos de datos para evitar exceder valores que en un futuro se implementarían en la validación cruzada de series de tiempo, específicamente en los entrenamientos y pruebas, pues en caso de excederse necesitaríamos nuestra variable y, la cual evidentemente no tenemos. Finalmente creamos características de conteos, como ventas promedio por mes, por tienda y por producto, elementos necesarios para nuestra siguiente etapa. Todas estas variables, como equipo, las consideramos muy útiles para el resto del proyecto, nos guían un paso más hacia nuestra variable de respuesta, que es la predicción de la demanda, puesto que al tener agrupados y contados estos datos con estas variables características, se vuelve información significativa y accesible para nuestras siguientes etapas.

Finalmente, las dimensiones finales resultantes en nuestro dataframe fueron las siguientes 24 columnas ya con las nuevas variables agregadas y son 446311 renglones

Etapa 4: Modelado

1) **Construir (codificar) el modelo de promedios móviles, cambiando 2 veces el periodo móvil.**

Como se puede observar en la gráfica que obtuvimos, la línea de color rojo es la más estable por lo que es la que elegiríamos, es la más estable debido a que aunque comienza con un poco de más error no brinca tanto como la azul o la verde que suben en demasía.

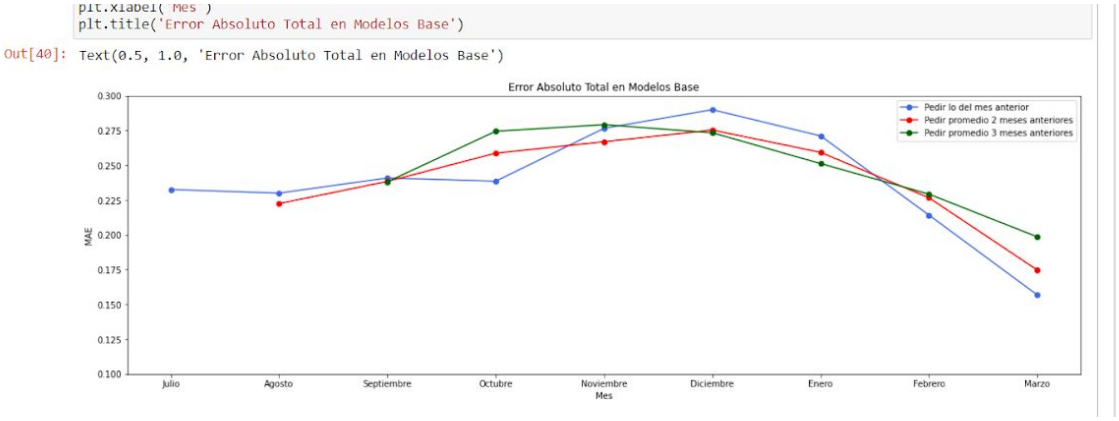


Figura 5. Error absoluto total en modelos base.

Se necesita consistencia y la línea roja es la que me puede ofrecer eso, por ejemplo, se puede visualizar que la línea verde tiene un aumento considerable en los meses de octubre y noviembre lo que me afecta bastante a mi como empresa. Si tomáramos la línea azul ocurre lo mismo y hasta un poco peor ya que aumenta mucho más , tiene un pico grande en el mes de diciembre.

Analizando la línea roja, mes por mes y comparándola con las otras líneas: en el mes de agosto se comporta mejor que la línea azul, en el mes de septiembre las 3 líneas tienen un comportamiento similar, en octubre aunque no es el mejor no hay un brinco tan grande como el verde, al siguiente mes la línea azul tiene un gran incremento, pero la línea roja siempre se mantiene estable, no da grandes saltos de errores como las otras dos.

Tomando como hipótesis que si nosotros fuéramos una empresa, nos convendría mucho más la línea roja porque aunque no siempre es la que tiene más bajos los errores las otras dos tienen picos bastante considerables que nos harían perder dinero, se necesita algo estable y seguro para que la empresa esté segura.

Para este análisis nos basamos en lo que fue la estadística tradicional o estadística descriptiva, ya que se basa en la precisión de los datos y propone el análisis detallado de las variables dedicadas para luego pasar a una descripción a fondo de

los datos. Además tiene como objetivo organizar y establecer una clasificación de los datos obtenidos de un grupo de población por ejemplo.

Para obtener estos resultados el programa fue desarrollado en python con ayuda de Jupyter, la cual es una herramienta para python que ayuda al procesamiento de datos, en diferentes ventanas usando el promedio de diferentes meses.

Gracias a este análisis se puede observar como cambian los resultados dependiendo de los meses y promedios que elijamos, nosotros nos decantamos por elegir estos meses porque con base a nuestra experiencia nos dimos cuenta que unas líneas iban a tener pequeños picos y una de ellas iba a ser la más estable, se nos hace muy interesante que gracias a esta tecnología podemos predecir y tomar mejores decisiones para nuestra empresa o el lugar donde trabajemos.

Promedios móviles

Los promedios móviles son promedios calculados a partir de subgrupos artificiales de observaciones consecutivas. Se utiliza cuando se quiere dar más importancia a conjuntos de datos más recientes para obtener la previsión. Cada punto de una media móvil de una serie temporal es la media aritmética de un número de puntos consecutivos de la serie, donde el número de puntos es elegido de tal manera que los efectos estacionales y / o irregulares sean eliminados.

Este tipo de promedio por sí solo, no produce predicciones precisas referentes al movimiento de precios. Por lo tanto, se deben combinar medidas a corto y a largo plazo para que produzcan señales o se combinen con otros indicadores que si puedan lograr una apropiada medición.

Existen diversos tipos de promedios móviles, estos son:

- **Promedio móvil simple**
    - Esta técnica se usa cuando se desea darle mayor importancia a un conjunto de datos recientes y de esa manera lograr obtener un pronóstico. Cada vez que se adquiera una nueva observación, esta se debe agregar al conjunto de datos y se elimina la observación que tenga más antigüedad.
- **Promedio móvil ponderado**
    - En general se difiere que los diversos puntos de datos se pueden ponderar o asignar a un punto concreto de gran importancia. La media móvil ponderada tiene la capacidad de agregarle importancia a los puntos de datos que estén más recientes.

Por ejemplo, una compañía de suministro de productos de oficina monitorea los niveles de inventario cada día. La compañía desea utilizar los promedios móviles de

longitud 2 para rastrear los niveles de inventario con el fin de suavizar los datos. Durante 8 días, se recolectan datos correspondientes a uno de sus productos.

Día	1	2	3	4	5	6	7	8
Nivel de inventario	4310	4400	4000	3952	4011	4000	4110	4220
Promedio móvil	4310	4355	4200	3976	3981.5	4005.5	4055	4165

Figura 6. Tabla de datos recolectados

El primer promedio móvil es de 4310, el cual representa el valor de la primera observación. El siguiente promedio móvil es el promedio de las dos primeras observaciones (4310 + 4400) / 2 = 4355. El tercer promedio móvil es el promedio de las observaciones 2 y 3, (4400 + 4000) / 2 = 4200, y así sucesivamente.

Acciones a realizar posteriormente:

Para poder llegar a los resultados necesitamos ir desarrollando el proyecto en python con ayuda de Jupyter al igual que en estudio R, ya que gracias a estas herramientas vamos a poder procesar los datos de manera más eficiente, rápida y segura del procedimiento que se va a llevar a cabo con este proyecto. Gracias a estos análisis que hemos llevado a cabo a lo largo de estas 4 etapas podemos saber que nuestro proyecto se dirige a detectar cuales son los datos que nos van a servir para la toma de decisiones a futuro para nuestra propia empresa o lugar en donde estemos trabajando.

MODELO DE APRENDIZAJE DE MÁQUINAS

Primero que nada, nos gustaría abarcar un poco el concepto de aprendizaje de máquinas para entrar de lleno en este tema. Aprendizaje de máquinas se trata de métodos computacionales que puedan desarrollar técnicas que les permitan a las máquinas aprender de datos, no a memorizarlos, reconocer patrones, tendencias o relaciones y así desarrollar reglas para mejorar el desempeño de un proceso. Entre estos métodos, se encuentran los bosques aleatorios, presentados a continuación.

BOSQUES ALEATORIOS. TEORÍA QUÉ MODELO, QUÉ ES PARA QUE SIRVE, SUS VENTAJAS DESVENTAJAS

Los bosques aleatorios es un algoritmo de machine learning flexible y fácil de usar que produce incluso sin ajuste de parámetros, un gran resultado la mayor parte del tiempo. Es un gran algoritmo para entrenar temprano en el proceso de desarrollo del modelo, para ver cómo se desempeña y es difícil construir un mal modelo con este algoritmo debido a su simplicidad.

Por lo tanto, en Bosques Aleatorios, el algoritmo para dividir un nodo sólo tiene en cuenta un subconjunto aleatorio de las características. Incluso puede hacer que los árboles sean más aleatorios, mediante el uso adicional de umbrales aleatorios para cada función en lugar de buscar los mejores umbrales posibles, como lo hace un árbol de decisión normal.

Ventajas:

Es un algoritmo muy útil y fácil de usar ya que los parámetros predeterminados a menudo producen un buen resultado de predicción. S

Si hay suficientes árboles en el bosque, el algoritmo no se adaptará al modelo, evitando el sobreajuste.

Se pueden utilizar en ambos métodos, clasificación y regresión.

Maneja los valores perdidos y mantiene la precisión con la falta de datos.

Manejar grandes conjuntos de datos con mayor dimensionalidad.

Genera predicciones más robustas.

Desventajas:

Una gran cantidad de árboles puede hacer que el algoritmo sea lento e ineficiente para las predicciones en tiempo real.

Es una herramienta de modelado predictivo y no una herramienta descriptiva.

No funciona tan bien con los problemas de regresión. No predice más allá del primer dataset.

No se tiene mucho control sobre lo que realiza el modelo.

BOSQUES ALEATORIOS. PROCESO DE PROGRAMACIÓN.

Durante la última etapa de este proyecto, fue requerido implementar un modelo de aprendizaje de máquinas para culminar este trabajo, obtener resultados, y compararlos con nuestro modelo base, el cual fue el de promedios móviles, con el objetivo de determinar cuál de ellos sería el óptimo para predecir nuestras demandas en los siguientes meses para los productos Apple en la República Mexicana.

Datalented se dió a la tarea de investigar, analizar y determinar los distintos métodos de aprendizaje de máquinas vistos en el curso, optando por Bosques Aleatorios como el método a desarrollar.

Una vez estudiado el modelo,implementamos una metodología llamada Validación Cruzada para series de tiempo, un método computacional para producir estimaciones internas del error de predicción, utilizando los datos y partiéndose en dos segmentos principales, entrenamiento y pruebas (respetando su secuencia temporal). La validación cruzada se emplea cuando los registros tienen una secuencia periódica, esto se hace con el fin de evitar sobre ajustes en los modelos y comprobar su efectividad en observaciones NO vistas anteriormente. Dentro de esta metodología ya se ha determinado un modelo base (el modelo de promedios móviles) y se propone otro modelo de aprendizaje de máquinas para comparar el desempeño entre estos dos.

Así bien, una vez entendido el concepto, empezamos a desarrollarlo en una herramienta de programación llamada Python, específicamente implementamos Jupyter Notebook para ello. Primeramente, se hizo una lectura de los datos para comprobar que efectivamente se utilizaban los adecuados y corroborando la correcta limpieza de los mismos. Posteriormente, en la fase 2, comienza la validación cruzada, donde preparamos los datos pues se realizarán ocho particiones, y dentro de cada una se dividirán los datos en entrenamientos y en pruebas. Una vez definiendo las variables en estos dos segmentos para x y para y, la cual es nuestra variable respuesta, en la fase 3 de modelado se empiezan a entrenar los datos implementando ahí mismo el modelo de bosques aleatorios, es decir, ahora se prepara, ajusta y entrena este modelo de bosques para realizar las predicciones y a su vez obtener el error de predicción del mismo, tomando en cuenta rangos de predicción con máximos y mínimos, redondeando y ajustando los valores.

Hacemos un paréntesis para dar explicación del por qué se eligió el Error Cuadrático Medio (MSE) para esto. El MSE es el criterio de evaluación más utilizado para problemas de regresión, especialmente cuando se emplea un modelo de aprendizaje supervisado, es decir, un modelo que contiene variables de entradas y una de salida, la cual en nuestro caso es la predicción de la demanda para productos Apple el siguiente mes.

Por otro lado, calculamos el error al cuadrado, en lugar del error simple, para que el error siempre sea positivo. De esta forma sabemos que el error perfecto es 0. Si no elevásemos el error al cuadrado, unas veces el error sería positivo y otras negativo, complicando de esta manera los modelos, su análisis y comparación.

Ya que se calcularon los errores cuadráticos medios de cada una de las ocho particiones en la fase tres de entrenamiento, ahora hay que probar si realmente este entrenamiento es correcto y/o bueno. Para ello, igualmente en ocho particiones, se preparan los datos de pruebas, considerando los mismos elementos de rangos, máximos y mínimos, y valores ajustados, obteniendo sus respectivos errores cuadráticos.

Cabe resaltar que, un parámetro muy importante a considerar dentro de este programa fue el valor de la profundidad del modelo, puesto que este elemento determina la cantidad de ramas dentro de nuestro conjunto de árboles de decisión y las reglas que debe haber en cada uno de los nodos que desembocan. En esta ocasión, decidimos ajustar un modelo de bosques con una profundidad de 1 y otro con una profundidad de 8, para observar y comparar el rendimiento y comportamiento del modelo ante estos atributos.

Finalmente, registramos nuestros valores de MSE, de cada partición, tanto de entrenamientos como de pruebas en un nuevo archivo de datos, para posteriormente leer este archivo y realizar gráficos de punto y línea a través de programación plt.plot.

En el siguiente esquema se puede observar todo lo mencionado anteriormente, las ocho particiones, los entrenamientos y las pruebas dependiendo de los meses, como debe ser en una serie de tiempo.

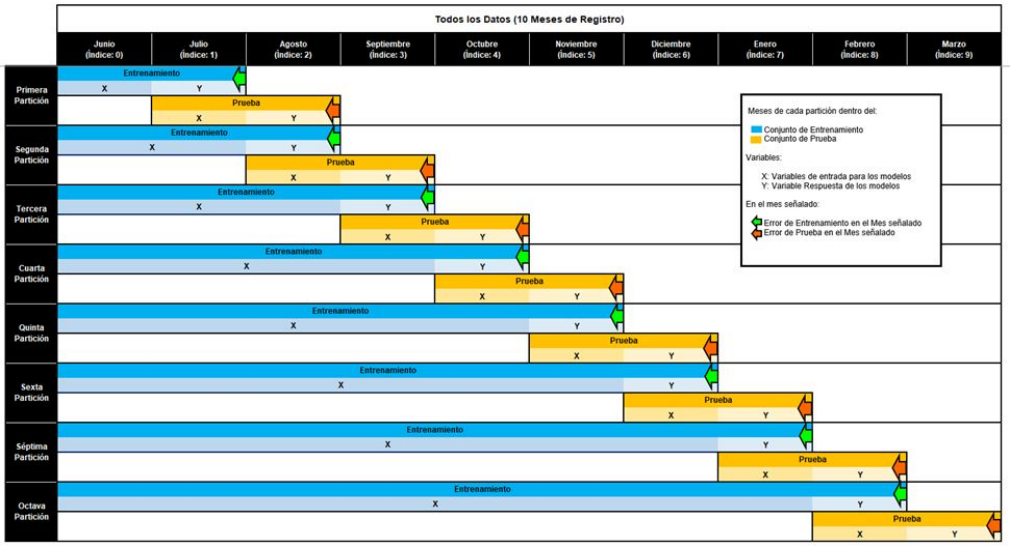


Figura 7 Esquema de Validación cruzada.

ETAPA 5 RESULTADOS Y CONCLUSIONES

Haciendo hincapié en lo mencionado en la etapa 4, los datos obtenidos de los errores cuadráticos medias tanto de entrenamientos como de pruebas, se plasmaron en dos gráficas respectivamente, mostradas a continuación (Figura 8 y 9).

En este primer gráfico, Figura 8, podemos observar únicamente dos líneas a través del tiempo, las cuales son de el modelo bosques aleatorios con profundidad de 1 y el modelo del mismo con profundidad 8, no existe en este gráfico la línea del modelo base puesto que en un modelo de promedios móviles no existen los entrenamientos, simplemente las pruebas.

Ahora bien, los mejores desempeños de un modelo se basan en un valor de MSE menor y una mayor estabilidad a lo largo del tiempo. Observando el gráfico y comparando valores, podemos observar que, en cuanto a los entrenamientos, el mejor comportamiento es el del modelo bosques aleatorios con una profundidad de 8, lo cual quiere decir que al ser una profundidad mayor, requiere un modelo más complejo con mayores reglas en nuestro esquema, es decir, al ser una profundidad de ocho, indica este número de ramas que tendremos entre nuestro conjunto de árboles de decisión, obligando a tener 8 reglas para cada una de ellas (como se mencionó anteriormente), por lo tanto se concluye que el modelo requiere de reglas más complejas para poder obtener mejores resultados en cuanto a las predicciones.

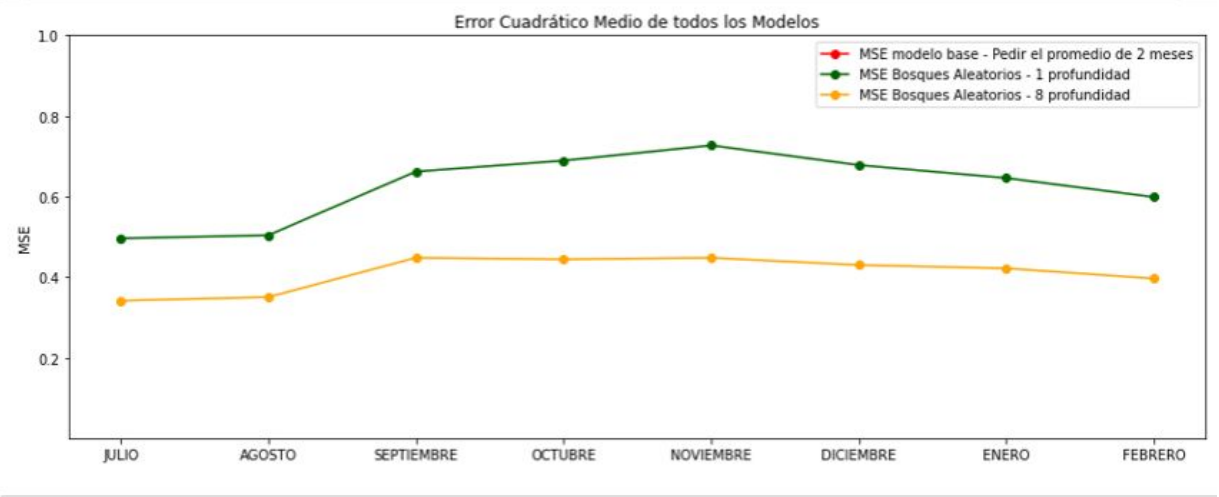


Figura 8 Gráfico de entrenamientos

Por otro lado, en el gráfico 9 se observan los comportamientos de los errores cuadráticos en el segmento de prueba, en el cual ya existen valores de nuestro modelo base a comparar, los promedios móviles. Mientras que la línea verde y amarilla son los modelos bosques aleatorios con profundidades 1 y 8 respectivamente. Podemos observar una inestabilidad drástica en el modelo bosques profundidad 1; en el mes de diciembre, al encontrarse un valor menor a todos los demás puntos, quiere decir que hubo un mejor comportamiento del modelo en este mes, teniendo un mínimo error de predicción, al igual que los meses de agosto y septiembre, mientras que en octubre, noviembre, enero y febrero los errores cuadráticos son mucho más grandes, siendo un peor desempeño del modelo.

Por otro lado, podemos observar un comportamiento similar en el modelo de bosques, profundidad 8, siendo sus mejores meses agosto, septiembre y diciembre.

Además, claramente se observa que la línea que más se acerca al 0 ideal de error dentro del gráfico es la roja, la del modelo base de promedios móviles, lo cual concluye que el mejor modelo para predecir nuestra demanda de productos Apple para el siguiente mes es este, con valores de MSE menores a 0.3, con una gran estabilidad, puesto que todos estos se encuentran en un valor promedio de 0.2, siendo inclusive en un futuro un modelo con mejor desempeño, obteniendo el mínimo valor de entre todos los modelos, 0.17 error de predicción.

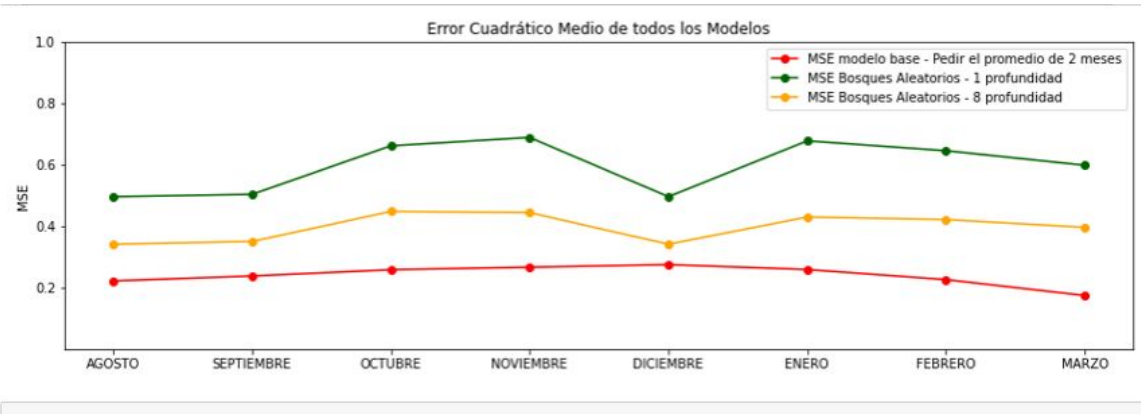


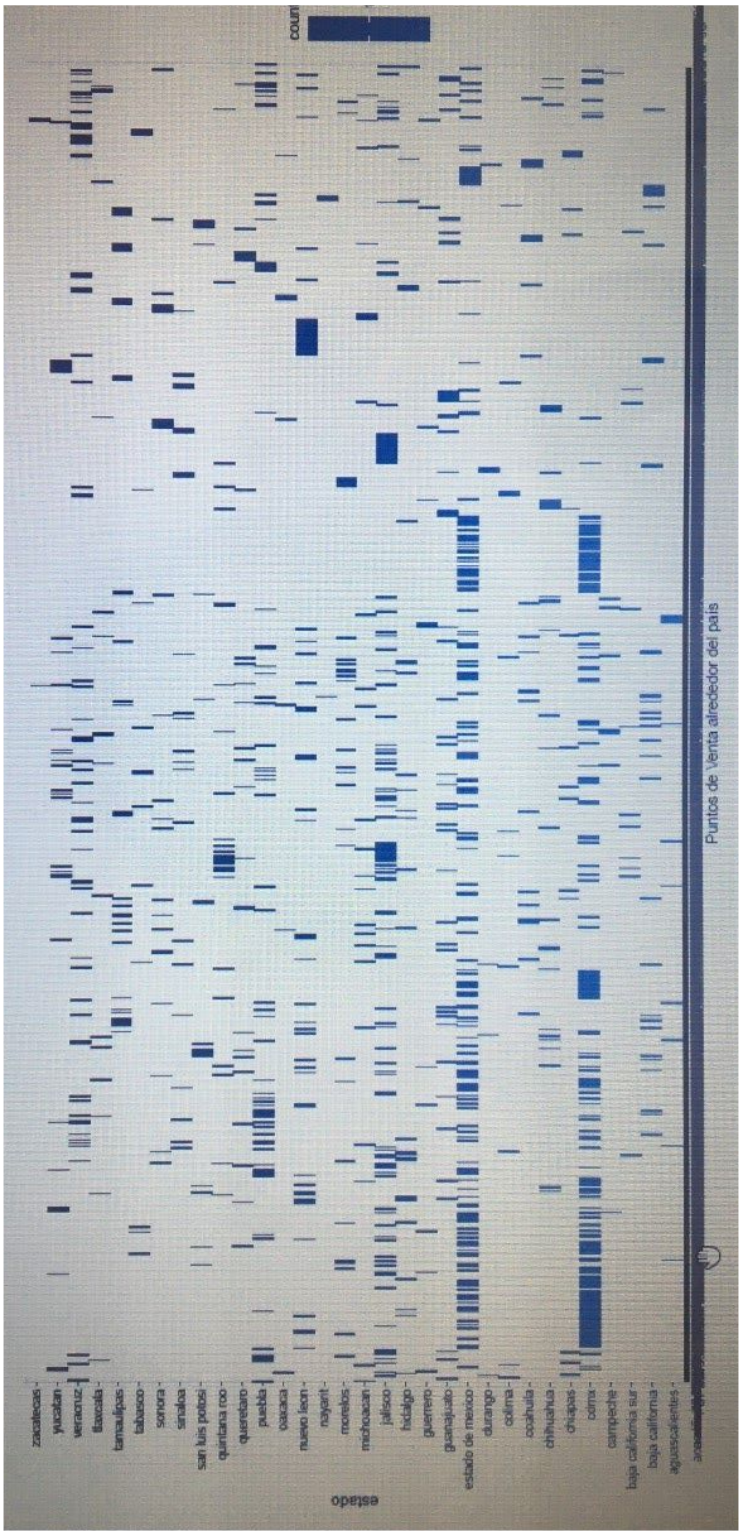
Figura 9. Gráfico de prueba

A continuación se muestra la tabla con todos los valores obtenidos de los errores cuadráticos medios para los distintos modelos.

	MÉTRICA	CONJUNTO	mes	modelo_base	bosques_aleatorios_1	bosques_aleatorios_8
0	MSE	ENTRENAMIENTO	JULIO	NaN	0.496290	0.341520
1	MSE	ENTRENAMIENTO	AGOSTO	NaN	0.503932	0.350534
2	MSE	ENTRENAMIENTO	SEPTIEMBRE	NaN	0.661867	0.448148
3	MSE	ENTRENAMIENTO	OCTUBRE	NaN	0.688899	0.444318
4	MSE	ENTRENAMIENTO	NOVIEMBRE	NaN	0.726792	0.448119
5	MSE	ENTRENAMIENTO	DICIEMBRE	NaN	0.677882	0.430060
6	MSE	ENTRENAMIENTO	ENERO	NaN	0.645833	0.421819
7	MSE	ENTRENAMIENTO	FEBRERO	NaN	0.598457	0.396481
8	MSE	PRUEBA	AGOSTO	0.222434	0.496290	0.341520
9	MSE	PRUEBA	SEPTIEMBRE	0.238355	0.503932	0.350534
10	MSE	PRUEBA	OCTUBRE	0.258661	0.661867	0.448148
11	MSE	PRUEBA	NOVIEMBRE	0.266858	0.688899	0.444318
12	MSE	PRUEBA	DICIEMBRE	0.275287	0.496290	0.341520
13	MSE	PRUEBA	ENERO	0.259226	0.677882	0.430060
14	MSE	PRUEBA	FEBRERO	0.226618	0.645833	0.421819
15	MSE	PRUEBA	MARZO	0.174703	0.598457	0.396481

Figura 10. Valores de MSE en los diferentes modelos.

Anexos:



Anexo 1. Gráfico bin2d (Puntos de venta vs Estados)

REFERENCIAS:

Bryan Salazar López. (2019). Promedio móvil. 15/10/2020, de Ingeniería industrial online Sitio web: https://www.ingenieriaindustrialonline.com/pronostico-de-la-demanda/promedio-movil/

Autor Desconocido. (2019). ¿Qué es un promedio móvil?. 15/10/2020, de Minitab Sitio web: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/time-series/supporting-topics/moving-average/what-is-a-moving-average/

Josefina Pacheco. (2019). ¿Qué es el Promedio Móvil?. 15/10/2020, de Web y Empresas Sitio web: webyempresas.com/promedio-movil/

Apple inc.. (2020). Apple. 06/09/2020, de Apple Sitio web: https://www.apple.com/

Edgar Sánchez. (2019). Participación de mercado de Apple y Huawei 2010-2018. 06/09/2020, de Merca2.0 Sitio web: https://www.merca20.com/mercado-apple-huawei/

Jose Martínez Heras. (10/10/2020). IArtificial. 20/11/2020, de IArtificial.net Sitio web: https://www.iartificial.net/error-cuadratico-medio-para-regresion/

Anónimo. (2018). Bosques Aleatorios Regresión. 20/11/2020, de AprendeIA Sitio web: https://aprendeia.com/bosques-aleatorios-regresion-practica-con-python-machine-learning/