Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Toluca

Laboratorio de diseño y optimización de operaciones

(Gpo 1)

Entrega Final - Samsung

Equipo 1: LO YODAS

Members:

Ricardo Alfonzo García Aponte	**IIS**	**A01363988**
Marylin Castañeda Navarrete	**IIS**	**A01363657**
Sharon Patricia Perea Vilchis	**IIS**	**A01365656**
Alfonso Enrique Varela Rosales	**IIS**	**A01362809**

Dia de Entrega:
02/06/2021

Introducción

Durante este proyecto se está analizando a la empresa Samsung, esto mediante una base de datos que contiene información importante para la toma de decisiones. Por lo que se estará observando el cómo se hizo para conseguir información de esta base de datos, graficando para volver esta información en algo más fácil de visualizar y práctico para la toma de decisiones, pero antes pongamos un poco de contexto de la empresa Samsung.

La empresa Samsung actualmente es una distribuidora de productos de tecnología, los cuales pueden ir desde electrodomésticos, pantallas de televisión, celulares y monitores.

Con una participación de mercado del 21%, Samsung lidera las ventas mundiales de smartphones en el 2018, de acuerdo con Gartner, seguido de su principal competidor que tiene una participación del 14,1%.

Según Gartner, se vendieron casi 384 millones de teléfonos inteligentes en el primer trimestre de 2018, lo que representa el 84% del total de teléfonos móviles vendidos. Samsung está vendiendo por hora 42.000 teléfonos celulares.

En términos generales, los resultados de Samsung Electronics registran unos US$56.250 millones, marcando un crecimiento interanual de más de un 20%. En cuanto a los beneficios operativos, Samsung ha ganado unos US$14.500 en este primer trimestre del año, una subida récord que se sitúa en torno al 58% respecto al mismo período del año pasado.

En base a lo que se mencionó anteriormente, creemos que es importante realizar el proyecto porque podremos utilizar varios conceptos vistos en esta clase y otros conocimientos adquiridos durante la carrera como aplicar herramientas y conceptos de ciencia de datos.

Bajo ciencia de datos hacer una transformación de datos digitales, el cual nos va ayudar a poder visualizar de mejor manera los datos que la empresa ha ido recopilando de manera histórica y de esta manera trabajar con datos certeros y datos duros, para una toma de decisión ya sea nuevo producto o campañas publicitarias.

Etapa 1 - Comprensión del Negocio

- **Entender y describir la problemática**

Como reto y proyecto que tenemos como equipo, es desarrollar o construir un portafolio de productos (predicción de la demanda), para los diferentes puntos de ventas de Samsung.

Del cual estaremos trabajando con una base datos, el cual nos va a ayudar a generar lo que estamos buscando, para que de ahí se puedan tomar acciones decisiones de las estrategias que a futuro la empresa Samsung podría realizar, dando predicción de demanda y/o conteo de productos dentro de la República Méxicana.

Con la ya mencionando el problema es que la compañía no tiene el pronóstico de producción o de las posibles ventas para los próximos años, por lo que se le ayudará a la empresa para conseguir dichos datos, reduciendo el error en pronósticos futuros y evitando pérdidas monetarias por falta de inventario o por sobregirar este mismo.

Es importante enfocarnos en este problema porque será aquello que nos de una buena reputación entregando al cliente siempre a tiempo y con el mejor servicio, sin exceder inventarios por cada tienda que se tiene y sobre cada modelo de celular.

- **Describir la problemática**

El equipo recibió una base de datos no estructurados que se pretende estructurar y depurar. Además de poder aplicar herramientas estadísticas como la regresión y los promedios móviles para poder aportar en un área de oportunidad. Además de poder darle un sentido y estructura a la información provista desde un Libro de Excel®. Considerando la finalidad de este proyecto, los datos se enfrentan a problema de regresión pues se desea crear un modelos con datos limpios que permitan predecir las ventas para el siguiente mes.

Así que, el propósito de llevar a cabo este proyecto es: ¿Cuál sería el pronóstico de las ventas del siguiente mes?

- **Objetivos**

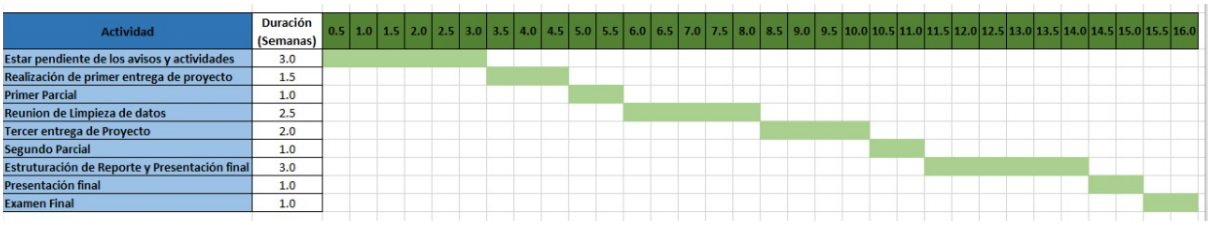
Como objetivos del proyecto es que mediante los conocimientos adquiridos por otras clase y de esta materia (Laboratorio de Diseño y Optimización), se buscará obtener buenos resultados con el análisis de datos de la empresa Samsung, en cuál nos sirva para comprender mejor lo que estamos aprendiendo y obtener los resultados que el proyecto está necesitando.

Para lo cual se tendrá que hacer un análisis detallado de la situación actual de la empresa Samsung, de forma visual, para entender mejor el tema a tratar.

Para que una vez obtenida esta información empezar a proponer y crear modelos de predicción de demanda. Esto bajo las herramientas de programación vistas en clase y storytelling para la construcción y exposición de los resultados.

Concluyendo con un reporte técnico, junto con una presentación ejecutiva que permita explicar y a entender cada unos de los resultados conseguidos de manera práctica, sencilla y concisa.

A continuación el Diagrama de Gantt que explica la duración y fases del presente proyecto:



Etapa 2 - Comprensión de los Datos

- **Descripción de Datos Crudos**

Como datos crudos hablamos de la base datos de Samsung, en donde podemos observar diferentes variables, como podría ser gamma, años, números de ventas, estados y entre otros. Con estos datos son con los que vamos a empezar a trabajar.

```
punto_de_venta        fecha              mes                anio        num_ventas
Length:148575      Length:148575      Length:148575      Min.   : 19   Min.   :1
Class :character   Class :character   Class :character   1st Qu.:2018   1st Qu.:1
Mode  :character   Mode  :character   Mode  :character   Median :2018   Median :1
                                                         Mean   :2018   Mean   :1
                                                         3rd Qu.:2018   3rd Qu.:1
                                                         Max.   :2019   Max.   :1
    sku               marca              gamma           costo_promedio      zona
Length:148575      Length:148575      Length:148575      Min.   :    0   Length:148575
Class :character   Class :character   Class :character   1st Qu.: 2426   Class :character
Mode  :character   Mode  :character   Mode  :character   Median : 2761   Mode  :character
                                                         Mean   : 4758
                                                         3rd Qu.: 5082
                                                         Max.   :15498
   estado             ciudad             latitud            longitud
Length:148575      Length:148575      Min.   :     14.9   Min.   :-1009514.0
Class :character   Class :character   1st Qu.:     19.3   1st Qu.:    -101.6
Mode  :character   Mode  :character   Median :     19.7   Median :     -99.2
                                      Mean   :     38.6   Mean   :    -107.2
                                      3rd Qu.:     21.9   3rd Qu.:     -99.1
                                      Max.   :2546176.0   Max.   :     -86.8
```

Cabe mencionar un factor importante es que esta tabla que se está presentado contiene errores que fueron corregidos, pero estas correcciones son visibles más adelante.

En la *Figura 1*, se puede apreciar los niveles para la variable “gamma” la cual habla de las diferentes gamas de productos existentes para los dispositivos de Samsung

gamma <chr>
baja
media
premium
alta

Figura 1: Niveles para la variable “Gamma”

En la *Figura 2*, se denotan los meses de análisis de caso siendo doce meses. Para el 2019 solo se cuenta con el periodo 1, 2 y 3 (siendo los meses de enero, febrero y marzo, respectivamente). Para el año 2018 solo se cuentan con el periodo 4, 5, 6, 7, 8, 9, 10, 11, 12, siendo del mes de (abril a diciembre, respectivamente). Además, en la misma *Figura 2*, también se aprecia que solo hay dos niveles para la variable de año, siendo: 2018 y 2019

anio <fctr>	mes <fctr>
2018	10
2018	11
2018	12
2018	6
2018	7
2018	8
2018	9
2019	1
2019	2
2019	3

Figura 2: Niveles para variable de "mes" y "anio"

Finalmente, en la *Figura 3* se puede apreciar la variable "*zona*": para identificar cuales obtuvieron un mayor número de ventas presentes y cuáles son los costos promedios en las mismas.

zona <fctr>
centro occidente
centro sur
golfo de mexico
noreste
noroeste
norte
pacifico sur
peninsula de yucatan

Figura 3: Niveles para la variable "zona"

- **Detectar Problemas de Calidad**

A continuación, se enlistan problemas de calidad hallados en la base de datos:

- En la columna "año" se hallaron datos tenía un formato erróneo, pues estaban aforados en 2 dígitos, cuando en realidad debían contar con 4 dígitos.
- En la columna "mes" se hallaron datos que estaban escritos con letras en vez de estar aforados con número.
- Respecto a la columna "nombre de la marca" había valores que estaban mal escritos, algunos con letras en mayúsculas, otros con letras repetidas o con la palabra entera repetida.

- Por otro lado, uno de los valores de la columna “zonas”, estaba escrito en mayúsculas.
- En la columna “estado” había nombres de ciudades que en vez de tener nombres de estados.
- Finalmente, en las columnas de “latitud” y “longitud” fueron hallados valores que estaban fuera de rango.

Etapa 3 - Preparación de Datos

- **Limpieza de Datos**

Summary Corregido

```
summary(SAMSUNG_DATOS)
```

```
 punto_de_venta        fecha               mes                anio        num_ventas
 Length:148575      Length:148575      Length:148575      Min.   : 19   Min.   :1
 Class :character   Class :character   Class :character   1st Qu.:2018   1st Qu.:1
 Mode  :character   Mode  :character   Mode  :character   Median :2018   Median :1
                                                          Mean   :2018   Mean   :1
                                                          3rd Qu.:2018   3rd Qu.:1
                                                          Max.   :2019   Max.   :1

     sku               marca              gamma           costo_promedio      zona
 Length:148575      Length:148575      Length:148575      Min.   :    0   Length:148575
 Class :character   Class :character   Class :character   1st Qu.: 2426   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median : 2761   Mode  :character
                                                          Mean   : 4758
                                                          3rd Qu.: 5082
                                                          Max.   :15498

             estado         ciudad             latitud          longitud
 cdmx            :28878   Length:148575      Min.   :   1.0   Length:148575
 estado de mexico:20663   Class :character   1st Qu.: 439.0   Class :character
 jalisco         :10369   Mode  :character   Median : 825.0   Mode  :character
 nuevo leon      : 9024                      Mean   : 861.7
 guanajuato      : 8700                      3rd Qu.:1292.0
 veracruz        : 7215                      Max.   :1751.0
 (other)         :63726
```

Summary con Errores

```
```{r}
summary(SAMSUNG_DATOS)
```
```

```
 punto_de_venta       fecha               mes                anio          num_ventas
 Length:148575      Length:148575      Length:148575      Min.   : 19   Min.   :1
 Class :character   Class :character   Class :character   1st Qu.:2018   1st Qu.:1
 Mode  :character   Mode  :character   Mode  :character   Median :2018   Median :1
                                                          Mean   :2018   Mean   :1
                                                          3rd Qu.:2018   3rd Qu.:1
                                                          Max.   :2019   Max.   :1
     sku                marca              gamma           costo_promedio       zona
 Length:148575      Length:148575      Length:148575      Min.   :    0   Length:148575
 Class :character   Class :character   Class :character   1st Qu.: 2426   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median : 2761   Mode  :character
                                                          Mean   : 4758
                                                          3rd Qu.: 5082
                                                          Max.   :15498
    estado             ciudad             latitud            longitud
 Length:148575      Length:148575      Min.   :     14.9   Min.   :-1009514.0
 Class :character   Class :character   1st Qu.:     19.3   1st Qu.:     -101.6
 Mode  :character   Mode  :character   Median :     19.7   Median :      -99.2
                                       Mean   :     38.6   Mean   :     -107.2
                                       3rd Qu.:     21.9   3rd Qu.:      -99.1
                                       Max.   :2546176.0   Max.   :      -86.8
```

- **Análisis exploratorio**

Ahora bien, analizando la Figura 4 se puede avistar la frecuencia de adquisición de dispositivos Samsung según su gama. Siendo la gama “baja” la más adquirida en el mercado durante los años 2018 y 2019. Por lo que se puede describir que la atención debe de centrarse en productos de baja gama, pues un mal pronóstico podría significar oportunidades de venta perdidas.

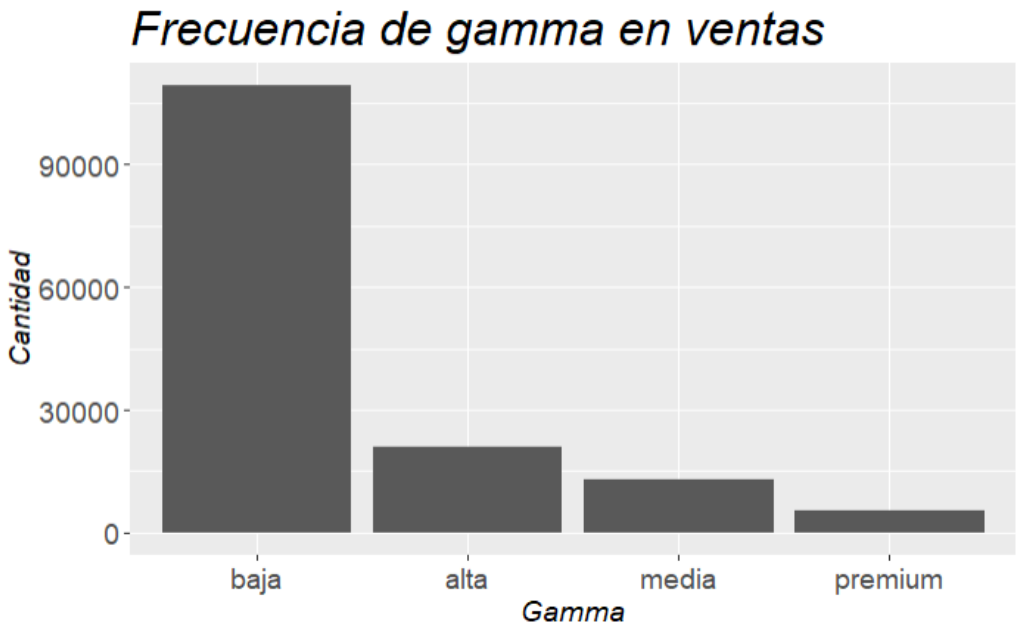


Figura 4: Gráfica de frecuencias para las gamas vendidas de 2018 a 2019

Analizando los meses del año 2019, se puede observar en la *Figura 5*, que en el mes de mayor ventas fue el mes de febrero, tomando en cuenta el costo promedio de los dispositivos adquiridos en dicho periodo. Es importante considerar este dato en caso de que se desee llevar futuras predicciones.

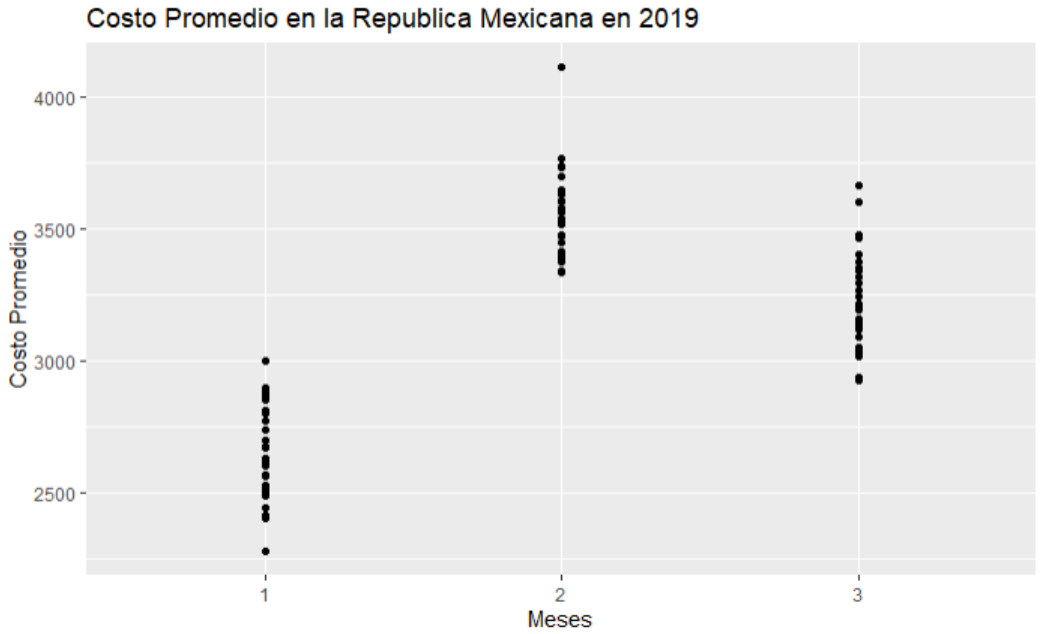


Figura 5: Gráfica de costo promedio respecto a meses analizados en el año 2019

En la siguiente gráfica, en la *Figura 6*, se analiza el costo promedio de cada una de las ocho zonas, siendo: “centro sur”, “noreste” y centro occidente las que obtuvieron los mayores costos promedio de este análisis. Sin embargo, las zonas de menor costo promedio son: “noroeste”, “golfo de méxico” y “norte” las tres zonas con promedios de venta menores por lo que se muy probable que los promedios respecto al pronóstico se mantengan con valores similares en el pronóstico que se realizará en este proyecto

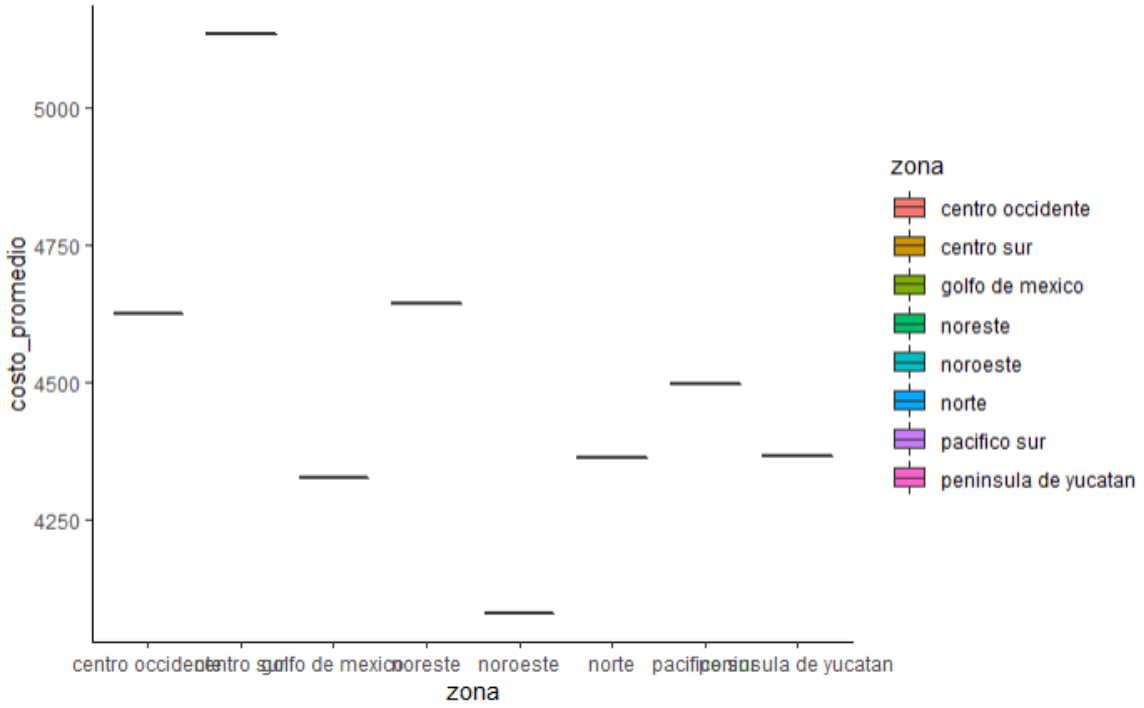


Figura 6: Gráfica de análisis de costo promedio respecto a las zonas

Finalmente, en la gráfica de la *Figura 7*, se observan los costos promedio respecto a los meses de este análisis. Es importante tener en cuenta esta tendencia de comportamiento, pues podría ser indicador del comportamiento de los datos en un futuro.

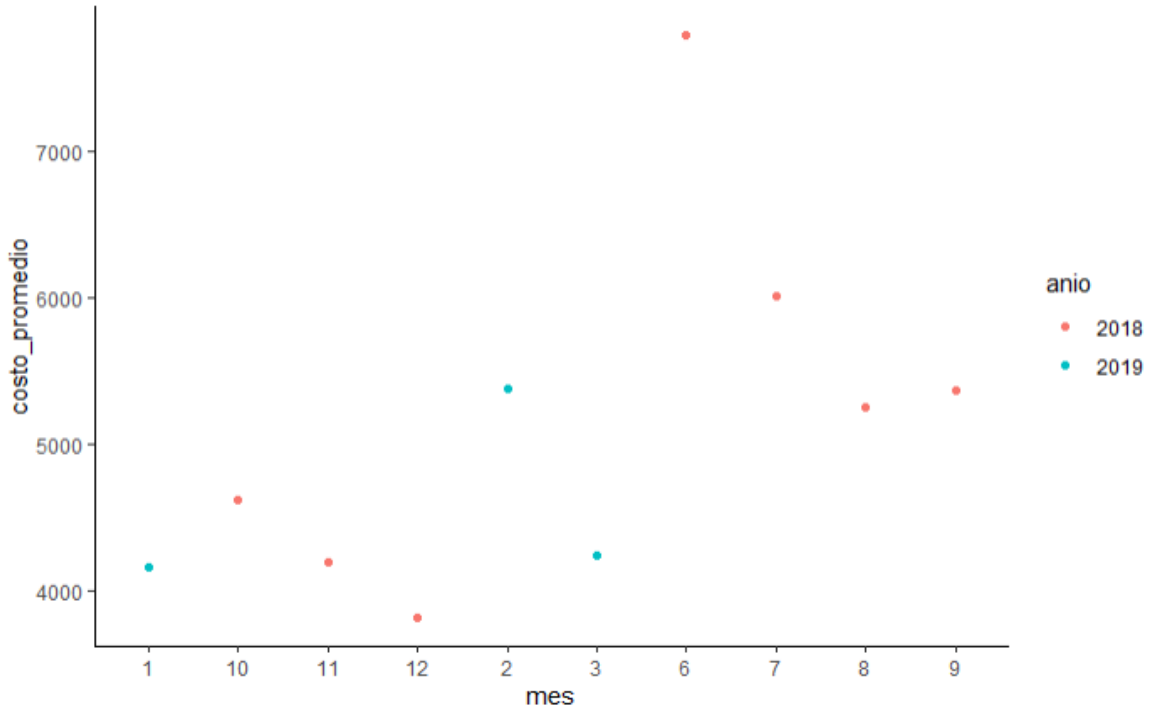


Figura 7: Gráfica de puntos de costo promedio y periodicidad de estudio del caso

- **Ingeniería de Características**

gamma <chr>	sku <chr>	repeticiones <dbl>
alta	N.SAMS9GR	2513
alta	N.SAMS9MR	1797
alta	N.SAMS9NG	3116
alta	N.SAMS9PG	3186
alta	N.SAMS9PM	2930
alta	N.SAMS9PN	5390

6 rows

Costo promedio no - No es variable cualitativa

Extra: El costo promedio esta implícito en la gamma.

costo_promedio <= 5000: “baja”

costo_promedio > 5000 & costo_promedio<=10000: “media”

costo_promedio > 10000 & costo_promedio<=15000: “alta”

costo_promedio > 15000: “premium”

mes_id <chr>	pdv_id <chr>	ventas_totales_en_tienda_de_cada_mes <dbl>	ventas_promedio_en_tienda_de_cada_mes <dbl>
0	1	2	0.05555556
0	10	1	0.02777778
0	100	0	0.00000000
0	1000	2	0.05555556
0	1001	0	0.00000000
0	1002	0	0.00000000
0	1003	0	0.00000000
0	1004	0	0.00000000
0	1005	6	0.16666667
0	1006	0	0.00000000

1-10 of 16,137 rows Previous 1 2 3 4 5 6 … 100 Next

#En el mes ____ y en el punto de venta _____, se tuvieron _____ ventas totales y se obtuvo un promedio de ventas de _____

mes_id <chr>	sku_id <chr>	ventas_totales_en_tienda_de_cada_sku <dbl>	ventas_promedio_en_tienda_de_cada_sku <dbl>
0	1	741	0.4132738427
0	10	540	0.3011712214
0	11	0	0.0000000000
0	12	0	0.0000000000
0	13	263	0.1466815393
0	14	0	0.0000000000
0	15	0	0.0000000000
0	16	0	0.0000000000
0	17	0	0.0000000000
0	18	0	0.0000000000

1-10 of 324 rows Previous 1 2 3 4 5 6 ... 33 Next

#En el mes ____ y en el punto de venta _____, se tuvieron _____ ventas totales y se obtuvo un promedio de ventas de _____ (3778/1900 productos)

pdv_id <chr>	mes... <chr>	sku_id <chr>	ventas_totales <dbl>	y_ventas_siguiente_mes <dbl>	ventas_totales_en_tienda_de_cada_mes <dbl>
1	0	1	1	0	2
1	1	1	0	0	1
1	2	1	0	0	7
1	3	1	0	0	7
1	4	1	0	0	12
1	5	1	0	0	10

6 rows | 1-6 of 9 columns

pdv_id <chr>	mes... <chr>	sku_id <chr>	ventas_totales <dbl>	y_ventas_siguiente_mes <dbl>	ventas_totales_en_tienda_de_cada_mes <dbl>
1	0	1	1	0	2
1	1	1	0	0	1
1	2	1	0	0	7
1	3	1	0	0	7
1	4	1	0	0	12
1	5	1	0	0	10
1	6	1	0	0	12
1	7	1	0	0	14
1	8	1	0	0	1
1	0	2	1	0	2

1-10 of 20 rows | 1-6 of 24 columns Previous 1 2 Next

Etapa 4 - Modelado

- **Promedios móviles simples**

En esta parte se realizó la parte de modelado de la variables previamente seleccionadas, de las cuales buscaremos sacar el pronóstico y comparar la gráfica del promedio móvil con 2 meses de promedio y con el de 3 meses de promedio y buscar explicar de manera breve qué resultados fueron los que se obtuvieron.

	Mes	mae_pedir_anterior	mae_promedio_2_meses_anteriores	mae_promedio_3_meses_anteriores
0	Julio	0.200301	NaN	NaN
1	Agosto	0.236072	0.248877	NaN
2	Septiembre	0.253966	0.252014	0.265198
3	Octubre	0.236227	0.232610	0.240720
4	Noviembre	0.286701	0.288142	0.295083
5	Diciembre	0.528955	0.504493	0.505149
6	Enero	0.507297	0.451253	0.416853
7	Febrero	0.272371	0.384652	0.373438
8	Marzo	0.151453	0.196342	0.297489

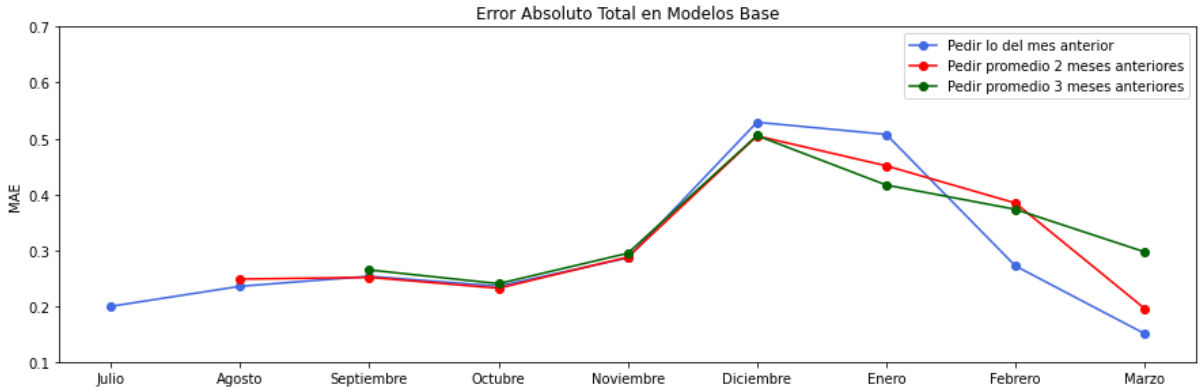
Los promedios móviles son medias calculadas a partir de subgrupos artificiales de observaciones consecutivas. En las gráficas de control, se puede crear gráficas de promedio móvil para los datos de tiempo ponderados. En los análisis de series de tiempo, Minitab utiliza el promedio móvil para suavizar los datos y reducir las fluctuaciones aleatorias en una serie de tiempo.

De la cual ayudan a sacar un pronóstico del siguiente mes, con base a los datos históricos capturados por la compañía, para que de esa manera se puedan tomar decisiones de qué acciones se pueden tomar a futuro.

Se considera que es importante destacar la metodología dependiendo de los datos con los que se cuenta y esto lo se hace bajo el comportamiento de la gráfica cuando se tabula los datos históricos. Pues dependiendo de dicho comportamiento se tomará en consideración la metodología bajo la cual serán evaluados, para tener un resultado más coherente bajo la teoría de promedios móviles.

Ahora para poder hacer una comparación de cada uno de los modelos y verificar cual es mejor, se debe hacer un cálculo del error, el cual de manera teórica conseguiremos un MAE, de lo cuales se buscará comparar este rubro de errores, y

de esta manera obtener el error más pequeño de cada unos de los modelos, para poder mencionar que bajo los datos compartidos, se realizó los cálculos pertinentes y que bajo la comparación de errores se puede decir qué método resulta ser el más adecuado para los datos.



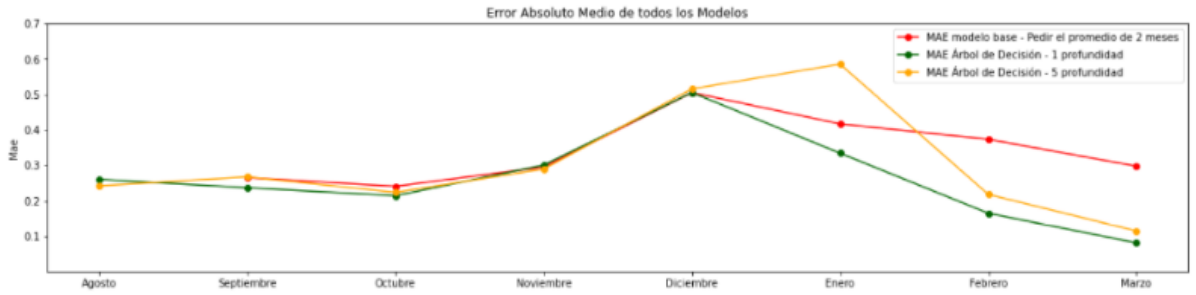
Hilando la explicación anteriormente dada, podemos observar la siguiente gráfica, y empezar a generar conclusiones, como se mencionó con anterioridad: una forma de poder concluir de mejor manera, para saber qué modelo es mejor, será mediante la comparación de los errores, pasando analizar la gráfica de error MAE, de cual podemos observar que al tomar 3 promedios el error se vuelve más pequeño y más constante, de cual se concluye que el mejor método de promedios móviles es el de obtener el promedio de 3 periodos anteriores.

Pero, como en todo, se debe mencionar las limitantes con las que se tienen en este proyecto, pues al momento de sacar el promedio de 2 meses, se observa que se pierden perdiendo datos al momento de avanzar en los cálculos, a diferencia del análisis a partir de los 3 meses, pues el error es pequeño y constante. De igual manera otra limitante es la cantidad de datos que se tiene a través del tiempo, lo cual provoca que se consigan errores más grandes y variación en los resultados. Así que, a mayor cantidad de datos será menor el error cometido en los mismos.

Etapa 5 - Evaluación

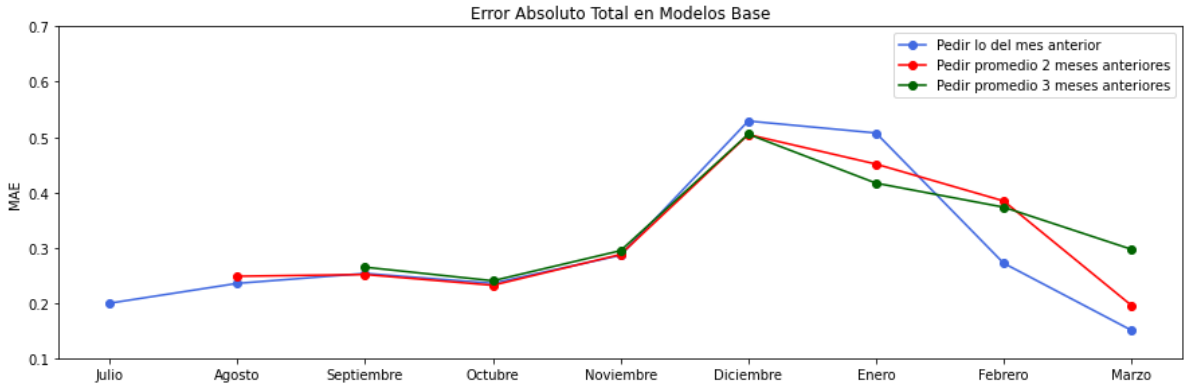
- **Resultados**

En esta etapa se llevó a cabo un modelado con validación cruzada con bosques aleatorios con árboles de decisión para poder ser comparado con el de promedios móviles que se tenía previamente.

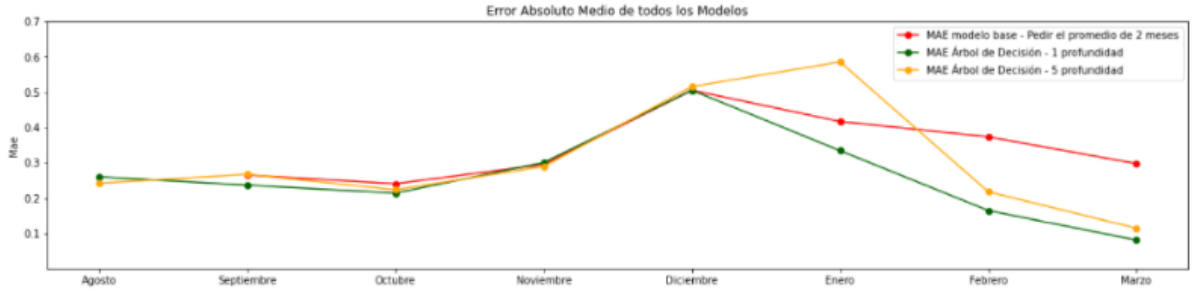


A continuación se realizó un modelo diferente en un programa diferente, Python, en el cual se gráfico de la misma manera el “Error Absoluto Medio” (MAE), de la cual se puede observar 2 modelos de árboles de decisión que son profundidad 1 y profundidad 5, a través del tiempo, en cada tipo de profundidad se obtuvo una variación de error con MAE diferente, que es lo que buscábamos evaluar con cuál profundidad quedaría mejor el pronóstico y con cual modelo entre machine learning y promedios móviles tendríamos un error menor en el futuro.

Se realizó un cálculo del error entre estos 2 modelos para poder analizar y verificar cual es el mejor pronóstico, observando la gráfica podemos concluir que la mejor opción es el MAE 1, ya que dentro de su gráfica no se visualiza mucha variación de cambio con respecto a los meses, a comparación de la otra opción donde desde el mes de Diciembre empieza a subir un mes para luego bajar drásticamente al siguiente. Y con esta explicación se busca que el MAE no solo sea pequeño sino también que no tenga mucha variación.



Promedios Móviles Simples



Machine Learning

Para el siguiente y último paso se realizó una comparación en la obtención de los MAE de las 2 formas diferente o técnicas diferentes, las cuales fueron, promedios móviles simples y machine learning.

Realizando estas 2 técnicas de pronósticos y obteniendo su error, MAE, podemos ir analizando y obteniendo un resultado más certero. Analizando estas 2 gráficas de MAE, concluimos que el error más pequeño y con poca variación del error a través del tiempo es el de machine learning, tomando la opción con profundidad 1 de árboles de decisión, el modelo seleccionado para mejorar a futuro el comportamiento de la empresa en sus pronósticos.

Conclusiones

Dado lo importante que es conocer los datos y su comportamiento. Es que se aplicaron herramientas vistas en el presente curso para poder comprender y, sobretodo, predecir sobre bases de datos.
Por sobre todos los procesos llevados a cabo, destacan: obtención de datos, limpieza de dichos datos para poder ser identificados y usados adecuadamente, realización de gráficas y análisis estadísticos, mediante los cuales se puede tener una justificación sólida sobre las decisiones que a lo largo de la investigación puedan ser tomadas, se procede a interpretar la información obtenida a través de los modelos estadísticos y matemáticos o de las gráficas que arrojarse el modelo y, finalmente, se lleva a cabo la documentación de la investigación puesto que en futuros análisis será de vital importancia contar con las decisiones además de servir como registro en los libros históricos que determinen qué acciones fueron tomadas de acuerdo a las circunstancias que envolvieron a la investigación.
En cuanto se refiere al aprendizaje de máquina, destaca entonces la comparativa de modelos según los datos que se le aportan al modelo. Y será el mejor de los modelos aquel que arroje un nivel de error más pequeño. Dicho análisis deberá ser propiamente documentado y, dentro del programa, estar comentado de tal modo que en futuras investigaciones se comprenda a detalle que se hizo y pueda aportar para ser antecedente de nuevas investigaciones.

Bibliografía

Soporte de Minitab. (2018). ¿Qué es un promedio móvil?. 05/05/2021, 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/

Samsung. (Agosto 2018). Samsung lidera ventas de teléfonos inteligentes. 07/03/2021, de Samsung Sitio web: https://www.samsung.com/latin/news/local/samsung-leads-smartphone-sales/