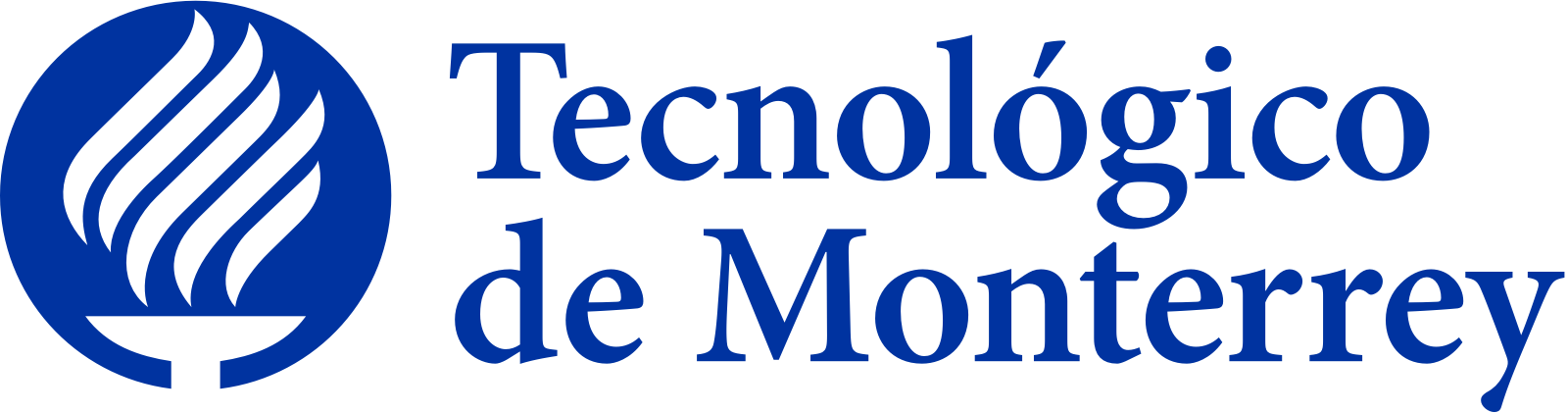
Instituto Tecnológico y de Estudios Superiores de Monterrey

Entrega Final de proyecto



*Materia: Laboratorio de diseño y optimización de operaciones*

Profesores:

Ana Luisa Masetto

Integrantes:

A01363654 María Verónica Cisneros Salgado

A01366094 Liliana Laura Arce Cruz

A01363506 Zaide Pale Bautista

20 Noviembre de 2020

Índice

[**Etapa 1: Comprensión del negocio**](#_u346nrxj6d69) **3**

[Descripción de la situación actual.](#_4fuec1uc8iw8) 3

[Entender y describir la problemática (en términos del negocio).](#_x5u9t5s0ztkd) 3

[Entender y describir la problemática (en términos de ciencias de datos).](#_1wfstejpbkut) 3

[Plasmar los objetivos (por etapas)](#_wa1u7ebuwqs6) 4

[Estructurar el proyecto y hacer un plan preliminar.](#_458lgdyts49o) 4

[**Etapa 2: Comprensión de los datos**](#_r1bepdnoz0ps) **5**

[Describir los datos crudos.](#_kkpwnxdrbbck) 5

[Detectar problemas de calidad](#_v1ar59mw33t2) 5

[**Etapa 3: Preparación de los datos**](#_pjuhoxporjoq) **6**

[Limpiar los datos:](#_vydtr61pfc8w) 6

[Análisis Exploratorio](#_vp3tapdmqcbk) 7

[Seleccionar y construir (ingeniería de características) variables para la etapa de modelado.](#_ga73wibejgox) 8

[**Etapa 4: Modelado**](#_2gkog8krrwe0) **9**

[Promedios móviles](#_napzb48fwkqr) 9

[Consideraciones](#_5yo92vamlk5s) 9

[Limitantes](#_njci784i2z2l) 10

[Resultados](#_6d46yunhoonf) 10

[**Etapa 5: Evaluación**](#_1k346pos5ggy) **11**

[Resultados](#_nved8h2myio0) 11

[Conclusiones](#_217ihp3gjsd4) 11

[**Bibliografía**](#_7sgcre7da6vy) **11**

[**Anexos**](#_kkpocuw5ezxe) **12**

[Anexos A](#_9v2998iemh0n) 12

[Anexos B](#_40h0j4ad2xaw) 14

[Anexos C](#_61l6phkogqt4) 16

# Etapa 1: Comprensión del negocio

## Descripción de la situación actual.

Motorola es una compañía estadounidense de telecomunicaciones que se dedica, principalmente, a la producción, venta y distribución de teléfonos inteligentes y otros dispositivos móviles. Actualmente, ‘Motorola Mobility’ forma parte de la empresa Lenovo que con el paso de los años ha ido perdiendo presencia ante las grandes compañías que manufacturan teléfonos inteligentes al día de hoy (como Apple y Samsung). Motorola contaba para finales de año con un 15.5% de participación en el mercado de telefonía y por lo tanto la evolución del negocio ha buscado construir y ejecutar estrategias como la consolidación de su portafolio de productos y hacer sus precios bastante asequibles en comparación con los competidores. Una proyección de la demanda supondría las bases para una estrategia sólida de mercado que le permitirá a la compañía establecer objetivos y estrategias claras de comercialización para asegurar su futuro en la industria de telecomunicaciones; ahí radica la relevancia de generar proyectos que agreguen valor a las decisiones de un negocio para asegurar el crecimiento y desarrollo del mismo aplicando conocimientos de estadística que, como ingenieros industriales, se adquieren con la práctica y el manejo de datos así como la información de se obtiene de los mismos.

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

Motorola se ve en la necesidad de reinventarse pues de ser líder del mercado pasó a competir con los gigantes como Apple y Samsung hasta perder terreno; para esto, es necesario redefinir su estrategia comercial y el enfoque para el desarrollo de nuevos productos. Motorola ha realizado movimientos arriesgados en el mercado, como el lanzamiento de productos sin analizar la demanda de los mismos y las necesidades de su cartera de clientes actuales, obteniendo como resultado productos sin éxito y que quedaron en el olvido a pocos meses de su lanzamiento. Esto supone un reto para exceder las habilidades del equipo de trabajo en el análisis de los datos que ya existen de la compañía para colaborar en la definición de las estrategias que se desplieguen a partir del análisis de datos y el desarrollo del presente trabajo. La empresa estadounidense se encuentra en un estancamiento, lo que le ha llevado a perder su identidad, la confianza de sus clientes y el prestigio en la industria; sin embargo, México se ha mantenido como un mercado potencial para su crecimiento, a través del lanzamiento de la familia **Moto G,** la compañía ha conseguido despuntar nuevamente en el mercado, por lo que se vuelve esencial, analizar detalladamente las demanda de sus clientes en cada uno de sus puntos de venta para continuar ofreciendo productos diferenciados y que satisfagan las necesidades de sus clientes.

## Entender y describir la problemática (en términos de ciencias de datos).

Apegarse a los cambios y necesidades del cliente se vuelve difícil cuando el mercado se vuelve cambiante y la satisfacción del cliente se vuelve un tema complejo. La proyección de la demanda es un enfoque muy efectivo para proveer al negocio de una herramienta que le permita redireccionar sus objetivos y fijar estrategias inteligentes con la confianza que el conocimiento sobre los datos que se retoman en este proyecto generen valor para la empresa. Por lo tanto la problemática que enfrenta la compañía Motorola corresponde a un problema de regresión, ya que se busca obtener un modelo que permita predecir el número de unidades de cada producto que se venderán en cada punto de venta donde la compañía tiene presencia. Para llevar a cabo la solución de la problemática se deberán realizar distintas fases en donde debemos realizar un trabajo de limpieza, identificación, clasificación y proyección de datos para cumplir con una estimación y desarrollar la solución adecuada para el negocio. Motorola ya tiene establecido su portafolio de productos; sin embargo, la proyección de demanda con base en los datos existentes de ventas resulta en una estrategia que le permita al negocio enfocarse en aquellos productos y puntos de venta estratégicos que promuevan el cambio para bien dentro de la gestión de la compañía. Entender las preferencias de los clientes a través de este análisis de datos también le permite generar valor a la empresa al darle un enfoque centrado en el cliente y en sus necesidades. Lo anterior también complementa la estrategia de negocio que se pueda poner en marcha a partir de la manipulación de los datos y de la extracción relevante que se haga sobre los más de 350,000 datos que se trabajarán en el desarrollo de este proyecto y en el que se responderá a la pregunta:

¿Cuántas unidades de cada producto de la marca Motorola, se van a vender en cada punto de venta de la compañía al siguiente mes de registro?

## Plasmar los objetivos (por etapas)

***Etapa 1- Objetivo de equipo:*** Entender en un contexto amplio la situación de la empresa, tanto sus ventajas competitivas como sus limitaciones y problemáticas, y su visión a corto plazo para establecer actividades clave a partir de los datos proporcionados de la demanda.

***Etapa 2- Objetivo de equipo:*** Realizar un análisis de los datos existentes y realizar las actividades de limpieza y segmentación de datos relevantes para la generación de información excluyendo cualquier impureza del análisis para asegurar la calidad del proyecto y la información que se genere.

***Etapa 3- Objetivo de equipo:*** Entender la situación actual a través del estudio y análisis de datos, así como generar propuesta sobre las posibles correcciones de la misma serie de datos. Tener una base sólida para el planteamiento del modelado.

***Etapa 4- Objetivo de equipo:*** Generar un mapa estratégico sobre los puntos de venta para predecir las compras del cliente sobre los puntos de venta más relevantes para la generación de estrategias encaminadas a la construcción y comparación de modelos de predicción.

***Etapa 5- Objetivo de equipo:*** Evaluar el desempeño de los modelos para generar conclusiones que mantengan la visión de la empresa bajo estudio.

***Etapa 6- Objetivo de equipo:*** Realizar el documento y presentación de la entrega final desplegando los resultados obtenidos que esperamos sean favorables aplicados a la realidad del negocio.

Al ***final*** del proyecto el objetivo principal es: Resolver un problema de construcción de portafolios de productos, es decir, predicción de la demanda de Motorola, basándonos en las compras de los últimos dos años (2018-2019), así como su información demográfica (SKU, gamma, zona, estado, ciudad, longitud y latitud), al igual que el costo promedio del artículo.

## Estructurar el proyecto y hacer un plan preliminar.

|  |  |  |  |
| --- | --- | --- | --- |
| **Etapa Proyecto** | **Líder de etapa** | **Fechas Límites y recursos** | **Actividades relevantes** |
| **1** | **Liliana Arce** | **Semana 4:** Inmersión en el entendimiento de empresa y proyecto. Análisis de datos y detección de problemas de calidad. | **-** Primer avance proyecto  - Investigación del negocio  -Planeación general proyecto |
| **2** | **Verónica Cisneros** |
| **3** | **Zaide Pale** | **Semana 6:** Limpiar, dar formato y depurar datos de problemas de calidad. | -Act 6: limpieza de datos  - Segunda entrega de proyecto (Segundo parcial)  - Generar modelo. |
| **4** | **Liliana Arce** | **Semana 10:** Diseño de modelado y comparación para su evaluación. |
| **5** | **Verónica Cisneros** | **Semana 16:** Evaluación y despliegue de resultados finales de acuerdo a modelación y entregas previas. | -Evaluación e interpretación de resultados  -Documento final  - Presentación final |
| **6** | **Zaide Pale** |

Tabla 1. Plan preliminar

# Etapa 2: Comprensión de los datos

## Describir los datos crudos.

**Diccionario de columnas:**  
Punto de venta: muestra los distintos lugares en donde la empresa cuenta con una tienda.  
Fecha: muestra día/ mes/ año.  
Mes y año: siendo estas dos columnas que muestran valores tanto numéricos como de texto.  
Número de ventas: cantidad de ventas que han tenido por fecha, teniendo un único valor (1).  
SKU: Stock Keeping Unit, el cual representa al código de referencia, presentándose 37 tipos.  
Marca: se está trabajando con la marca Motorola.  
Gamma: hace referencia a los niveles o categorías del producto (baja y media).  
Costo promedio: este es el costo total dividido entre el número de unidades vendidas/ producidas.  
Zona: ubicaciones aproximadas, o bien, áreas donde se encuentra la tienda.  
Estado y ciudad: siendo estas dos columnas, hacen referencia a las ubicaciones más generales de las tiendas.  
Latitud y longitud: siendo estas dos columnas, son ubicaciones más específicas, de un punto exacto de la tienda, por medio de un conjunto de números.

**Información adicional:**Se presentan 14 columnas, con 302562 datos en total, en donde de acuerdo con los datos presentados en excel, existen 1901 puntos de venta, así como la fecha, y también el mes de manera individual, abarcando los diez meses del año, y una cuarta columna con los dos años a evaluar (2018 y 2019). Los costos promedio, van en un rango de entre 193.9889 hasta 7142.5830, sin embargo, se encuentran casos con un costo de 0, siendo estos 18927 quiénes lo presentan. Dentro de la columna zona, se cuentan con 8. En la sección de estado, 32, y 228 ciudades. La latitud va de entre 14.8741 hasta 32.6657, y la longitud de -117.1187 a -86.8075.

## Detectar problemas de calidad

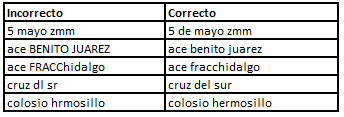
Dentro de los problemas generales en la calidad de los datos, es que estos se encuentran no homogéneos; se presentan incongruencias entre estados y ciudades, así como valores repetidos por faltas de ortografía, es decir, se hace referencia a un mismo caso, no obstante, lo expresan de diferente manera. Iniciando por el punto de venta, se presentan cinco situaciones en donde ya sea por falta de letras, mala ortografía, o bien, la palabra está escrita tanto en mayúsculas como en minúsculas y se repiten, como se muestra en la *Figura 1 (Anexo A)*. Asimismo, los meses están escritos tanto en número como en letra, como se observa en la *Figura 2 (Anexo A)*, con el mes de diciembre. Para el caso de la columna “año”, solo 25 de los datos están escritos con dos cifras, mientras que el resto de ellos lo están con cuatro.

Por otra parte, a pesar de que se esté trabajando con una sola marca, se presentan cinco tipos de escritura de la misma, siendo estas las de la *Figura 4 (Anexo A)*. Con respecto a la columna “zona”, en esta se presentan nueve diferentes, debido a que “Golfo de México” está escrito de dos maneras, uno con mayúsculas y otro con minúsculas, como se observa en la *Figura 5 (Anexo A)*; haciendo que esto arroje una zona adicional a la realidad, donde son 8. Dentro de la columna “estado”, se presentan 35 diferentes; sin embargo, no todos son precisamente estados, sino también se encuentran ciudades, como por ejemplo en la *Figura 6 (Anexo A)* que se cuenta con una categoría de Estado de México, y otra de Toluca, siendo que Toluca debe de estar en la siguiente columna. Esto se reduce a 32 estados. Tanto en latitud como en longitud, se presenta un caso en donde una fila de la categoría “acayucan” en la columna punto de venta, omiten un punto, provocando un descontrol de datos, e incoherencia, debido a la magnitud entre los posibles resultados (*Figura 7, Anexo A)*.

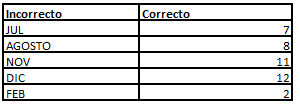
# Etapa 3: Preparación de los datos

## Limpiar los datos:

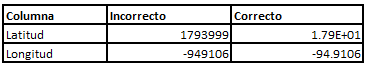
La limpieza de datos consiste en la detección y corrección de registros de datos que se encuentran erróneos, en este caso de la base de datos de Motorola. Este proceso permite identificar aquella información que no es pertinente y debe de ser sustituida o eliminada, dependiendo el caso, con el objetivo de brindar información de calidad y poder continuar con el análisis del mismo. Con base en la información presentada en la etapa dos, se pudo elaborar una limpieza de los datos detectados anormales, o no homogéneos. Iniciamos con el punto de venta, en donde cinco lugares compartían información con otro, sin embargo, al estar escritos con tan solo una mayúscula de diferencia, el lenguaje los identifica como diez distintos. Dentro de los cinco que se reemplazaron fueron las siguientes, las cuales presentaban problemas de mayúsculas, lo cual fue importante para evitar tener duplicidad de puntos de venta para la parte de modelado.

   
Tabla 2. Puntos de venta modificados

Por otra parte, en la sección del mes, se reconocieron cinco meses que eran identificados tanto con letra como con número, y de igual forma, se aplicó un reemplazo a todos aquellos que fueron escritos con letra para poder tener un mejor manejo de datos y un formato estándar para las fechas.

  
Tabla 3. Meses modificados

Para el caso del año, existen 2, 2018 y 2019. No obstante, el 2018 también fue escrito como “18”, el cual también debió de modificarse para evitar problemas en la parte de modelado. Como marca se presentaron diferentes tipos: Motorola, Motorola-motorola, MOTOROLA, mMotorola, Motorola-Motorola, y motorola. Y para facilitar la comprensión del mismo en R, debido a la sensibilidad del lenguaje, optamos por la opción de “motorola”, escrito todo en minúsculas. En las zonas, solo existía uno en donde todo estaba escrito en mayúsculas (GOLFO DE MEX) y el otro en minúsculas (golfo de mexico), y por la misma razón anterior, que se aplicó en Marca, se eligió la opción de minúsculas. Para los Estados, se encontraron 3 ciudades dentro de esta sección, las cuales eran Toluca, que representa al Estado de México, Tehuacán, que se ubica en Puebla, y Acapulco, que es Guerrero. Finalmente, tanto en latitud como en longitud, el error fue identificado en el mismo punto de venta, el cual, ignoraba el punto decimal que permitía encontrarse dentro del rango, para lo cual se completó correctamente la serie de tiempo.

  
Tabla 4. Latitud y longitud modificados

## Análisis Exploratorio

El análisis exploratorio permite tener una visualización de lo que ocurre, esta etapa se lleva a cabo posterior a la limpieza de datos con la finalidad de evitar incongruencias. Para el análisis de datos se llevaron a cabo distintos tipos de gráficas que nos permitieron tener un mejor comprendimiento de la situación actual de la empresa Motorola. Como se muestra en la Figura 1 se realizó una gráfica de barras para evaluar la frecuencia por zona en el entendido de que la frecuencia de registros está ligada al número de ventas. Con base en la gráfica se pudo obtener que la zona con más ventas es la zona centro, mientras que la zona pacífico sur es la que presenta menor volumen de ventas. Con esta información se puede tener una visualización de las posibles estrategias en ventas con respecto a cada zona, es decir, trabajar en estrategias para concentrar las actividades en la zona centro que es la más redituable por el número de ventas y la zona del pacífico sur como un foco rojo para atender la estrategia que ayude a incrementar las ventas en esas zonas. Consideramos como un factor la densidad de población de la zona como una influencia para el resultado de las ventas en las distintas regiones; concentrándose, por ello, en la zona centro.

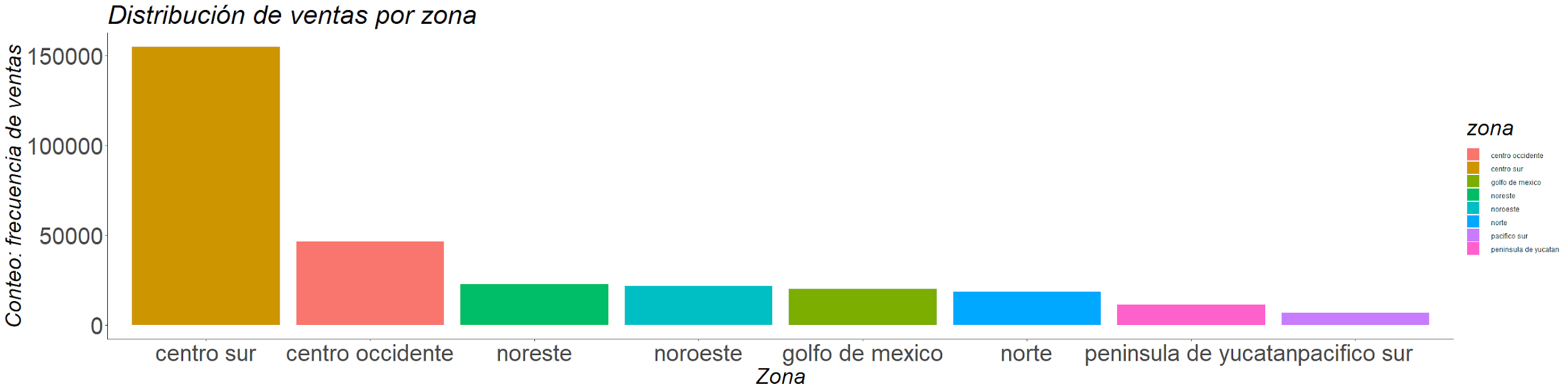


Figura 1. Distribución de ventas por zona

Posteriormente se realizó un diagrama de caja el cual buscaba identificar si el costo promedio de los productos influenciaba la intención de compra por zona. Lo que refleja la Figura 2 es que la variación de los costos promedio por zona es constante, basándonos en la ausencia de outliers en los datos. Esto puede deberse a la estabilidad de la demanda en los últimos dos años, es decir, la fluctuación de las ventas para cada zona se han mantenido constante indicando que las ventas no han subido. Lo anterior es de relevancia para el estudio puesto que indica que lejos de buscar una estabilidad de la demanda se busca aumentarla satisfaciendo las necesidades del mercado.

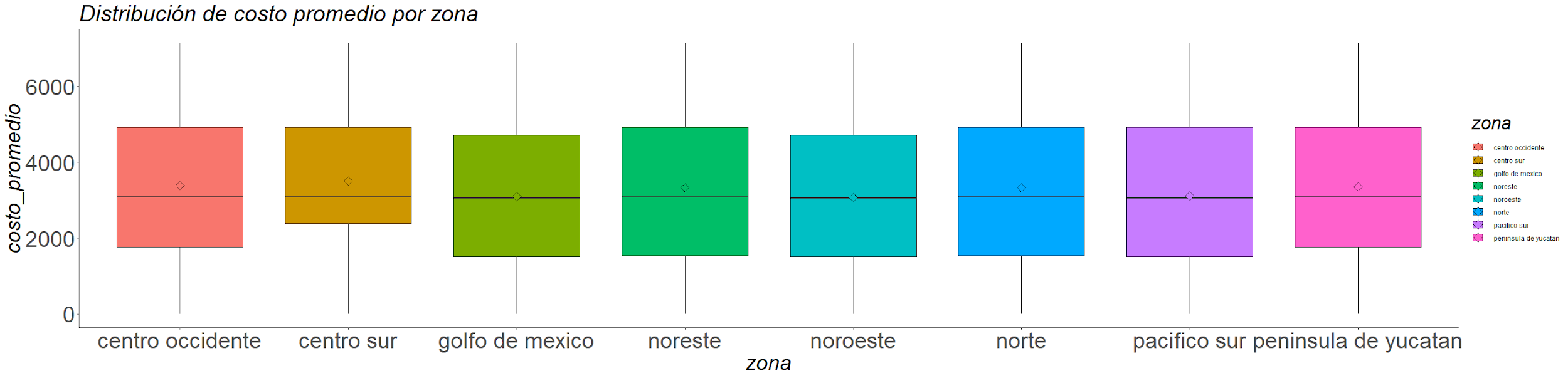


Figura 2. Distribución de costo promedio por zona

Así mismo, se buscó obtener el comportamiento de las ventas durante el año, para lo cual se realizó una gráfica de línea de tendencia como se observa en la Figura 3, la cual permitió apreciar que esta se encuentra en estado constante, ya que es más hacía horizontal que sesgada. Esto quiere decir, que de acuerdo a lo comentado en la Figura 2, la demanda resulta ser estable, siendo claro que, por el mes de diciembre, se encuentra el pico más alto de ventas, considerando que se presentan días festivos. Por otra parte, al finalizar las vacaciones de enero, se observa una disminución considerable de la misma, posicionándose fuera o cerca de un posible límite inferior. Elaborar esta gráfica es de suma relevancia, ya que nos permite identificar el grado de estacionalidad en las ventas, para poder generar los mejores pronósticos con los métodos adecuados.

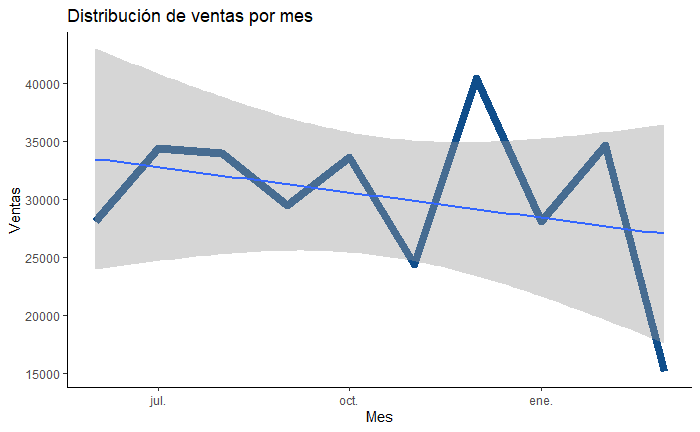


Figura 3. Distribución de ventas por mes

## Seleccionar y construir (ingeniería de características) variables para la etapa de modelado.

Después del análisis exploratorio y antes de revisar los modelos de regresión para la creación del modelo. En esta parte del proyecto se trabajó con la serie de datos que se guardó después de la limpieza, para aplicar ingeniería de características en las que se crearon variables con la intención de agregar valor al modelo. El primer paso fue crear índices con respecto a las variables cualitativas de los datos que fuesen más relevantes y así facilitar el manejo con respecto a los puntos de venta, fecha y modelo (SKU). Además se agregaron columnas al set data con cada índice que se creó. Para ello se utilizó la función “left\_join” en el que se adjuntan los data frames creados (con los índices) para que se vinculen con el dataset original. La función se aplica para los índices realizados anteriormente con los puntos de venta, fecha y modelo (SKU).

El segundo paso dentro de la aplicación de ingeniería de característicos se hizo a partir del número de ventas (“num\_ventas”) en donde se agruparon los registros de manera que así pudiésemos detectar cuántos registros de ventas se hicieron en cada periodo. Como paso siguiente se completaron las series de tiempo de manera que se pudieran preparar los datos para la aplicación del modelo del proyecto. Como primer paso se determinó la cantidad de registros con los que se trabajaría para completar los datos; para esta parte se utilizó la función “Merge” en donde se crean las combinaciones entre fecha y modelos que agranda la dataset y permite completar la serie de tiempos. La secuencia y código de los pasos descritos se muestra más a detalle en el anexo A. La etapa de completar serie de tiempos permite tener claras las ventas por modelo y periodo y ahí radica la aplicación de la herramienta merge para enriquecer el conocimiento de los datos en general y retroalimentar el proyecto con la información desplegada como producto de las combinaciones; esto, a su vez, también provee de un trabajo previo a la aplicación del modelo para asegurarnos de trabajar con datos que ya se completaron después de la limpieza y la exploración analítica de los mismos. También se usó para identificar aquellos puntos de ventas en los que no había registros y a partir de ello identificarlo como cero en términos de venta para poder trabajar con dichos datos.

Después de completar las series, definimos la variable respuesta enfocada al objetivo de predicciones propuesto inicialmente en el proyecto; por lo tanto, se registra la columna de predicciones con base en los datos con los que se cuentan. A partir de esto se crean nuevas características en los que se tomaron en cuenta promedios, conteos y rezagos como el promedio de ventas de un punto de venta por producto para enriquecer el valor de los datasets. Las dimensiones de nuestro dataset finales son 209,414 ya que decidimos quedarnos con las variables de Punto de Venta\_Id, SKU\_Id, Mes\_Id y Ventas Totales puesto que consideramos que son las variables que pueden aportarnos mayor información para nuestro modelo y dentro de las cuales están implícitas las variables como Zona, Marca, Gamma, Ciudad, entre otras.

# Etapa 4: Modelado

## Promedios móviles

Los promedios móviles son un cálculo utilizado para analizar un conjunto de datos en modo de puntos para crear series de [promedios](https://es.wikipedia.org/wiki/Promedio). Así las medias móviles son una lista de números en la cual cada uno es el promedio de un subconjunto de los datos originales. Por ejemplo, si se tiene un conjunto de 100 datos el primer valor de la serie de medias móviles podría ser el promedio de los primeros 25 términos, luego el promedio de los términos 2 al 26, el tercer elemento de los términos 3 al 27 y así, hasta por último el promedio de los últimos 25 números del 76 al 100.

Una serie de medias móviles puede ser calculada para cualquier [serie temporal](https://es.wikipedia.org/wiki/Serie_temporal). Se usa para demanda estable, sin tendencia ni estacionalidad; suaviza las fluctuaciones de plazos cortos, resaltando así las [tendencias](https://es.wikipedia.org/wiki/Tendencia) o ciclos de plazos largos.

Hay varios tipos de promedios móviles:

* Promedio móvil simple: es el promedio de los “*n”* datos anteriores. En esta técnica elemental de predicción, cuanto más grande sea “*n”*, mayor será la influencia de los datos antiguos. En contrapartida, si se selecciona una “*n”* baja, se tendrán en cuenta datos más recientes para nuestra predicción.
* Promedio móvil ponderado: es una media multiplicada por ciertos factores, que le dan determinado peso a determinados datos. La media móvil ponderada desarrolla y mejora las aplicaciones de la media móvil simple. Se trata de la media aritmética de los “*n”* valores anteriores ponderados según diferentes criterios.
* Promedio móvil exponencial: es una media móvil ponderada exponencialmente. Se trata de la media aritmética de los “*n”* valores anteriores con factores de ponderación que decrecen exponencialmente. La ponderación para cada punto de datos más antiguo decrece exponencialmente, nunca llegando a cero.

## Consideraciones

Dentro de las principales consideraciones que se presentan en la selección de un método de elaboración de pronósticos es la precisión del mismo, es decir, se pretende utilizar el que tiene errores menores.

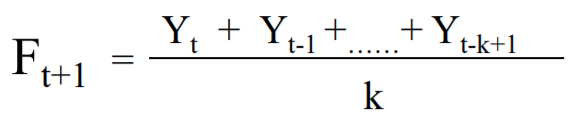
Para trabajar con promedios móviles, se deben tomar las siguientes tres variables:

Yt: son las observaciones en el período.

Ft: pronóstico para el período t

K: valores de datos más recientes en la serie de tiempo como el pronóstico para el siguiente período.

Una vez teniendo dichas variables, podemos obtener Ft de la siguiente forma:

****

Es importante recordar que sólo se promedian las últimas observaciones deseadas, así como que el orden se determina a priori. Asimismo, el tener un orden grande permite eliminar los picos, es decir, por medio del método de suavización, mientras que el orden pequeño sigue muy de cerca los cambios de corto plazo.

Es primordial considerar que este método se utiliza cuando se requiere dar más valor a los conjuntos de datos más recientes, con el objetivo de obtener una previsión. Otra gran ventaja de este, es que es óptimo para la manipulación de patrones de demanda, ya sean aleatorios o nivelados, buscando eliminar el efecto que los datos irregulares históricos pudieran presentar, mediante el enfoque ya mencionado, de basarnos en los últimos registros de demanda

## Limitantes

El modelo de promedios móviles es uno de las herramientas más utilizadas para el pronóstico de ventas porque permite una aplicación sencilla; sin embargo, la aplicación del modelo y la eficiencia del mismo dependerá de los objetivos internos para los que está destinada su aplicación así como de las limitaciones que conlleva su uso y aplicación como herramienta de pronóstico. A continuación se listan algunas de las limitantes más relevantes para propósitos del proyecto.

* Solamente pueden llegar a pronosticar un periodo más.
* 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.
* Es un método sencillo en la aplicación; sin embargo, sólo resulta útil y eficiente si los datos no presentan estacionalidad o tendencias ascendentes o descendientes.
* Para propósitos prácticos en su uso se recomienda utilizar una mayor cantidad de periodos y la falta de registros (en términos de proyecto) puede incidir cuando se tienen cambios recientes dentro de la temporalidad del modelo.
* Aún cuando el modelo cuenta con variantes que se usan para series de tiempo con tendencias, en la práctica le quita relevancia a los datos más recientes de ventas lo que afecta el pronóstico inmediato.

## 

## Resultados

A partir de las predicciones que se realizaron con promedios móviles para los periodos elegidos (9 de Julio a Marzo) se calcularon también los errores por cada periodo a partir del promedio de 1 a 3 periodos anteriores. Sobre esto se hizo la gráfica mostrada en la Figura 4, con los resultados por cada modelo realizado. Con los comportamientos sobrepuestos en la gráfica se puede apreciar mejor los contrastes entre cada uno para tomar decisiones sobre el proyecto, de los tres el que consideramos más adecuado, teniendo el contexto de promedios móviles abordado en este reporte, se mantuvo en tres periodos considerando que el error tiene un comportamiento más estable y que en algunos casos los registros de ventas para cada periodo son de cero y tomar un modelo de sólo un periodo o dos no reflejaría una predicción tan confiable.

En la Figura 4, se puede apreciar como el pronóstico con un periodo móvil logra aproximarse en mayor medida a las ventas reales de este año. Asimismo, si de origen la variable a evaluar no presenta grandes variaciones, se recomienda el uso de una n mayor. Mientras que si muestra patrones cambiantes se recomienda el uso de un valor más pequeño para n periodos, como lo consideramos para este caso, pudiendo predecir con mayor confianza.

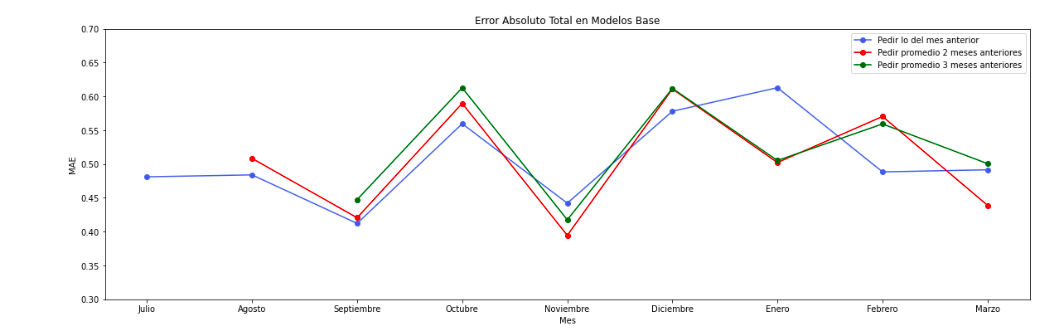
****

Figura 4. Gráfica del Error Absoluto Total

# Etapa 5: Evaluación

## Resultados

Siguiendo con el análisis de datos y los métodos de regresión vistos en clase, decidimos que la aplicación de Bosques Aleatorios como algoritmo de aprendizaje de máquina para los pronósticos de venta de los datos completos que trabajamos anteriormente y que corresponden a los registros en los diferentes puntos de venta de la compañía Motorola. En esta etapa se hizo la división de los datos entre datos de entrenamiento y de prueba como se muestra en la Figura.15 dentro del anexo. A partir de la construcción tanto de los modelos de promedios móviles y de los bosques aleatorios con parámetros de n=10 y n=300 los resultados obtenidos se concentran en las gráficas mostradas a continuación.

En la Figura 5, se muestra la comparación entre el modelo entrenamiento con el modelo de *bosques aleatorios* con 10 árboles de decisión y el modelo entrenado con 300 árboles. Así mismo, se incluyó en esta comparación el error absoluto medio de entrenamiento tomando tres periodos de datos, es decir, 3 meses de promedio.

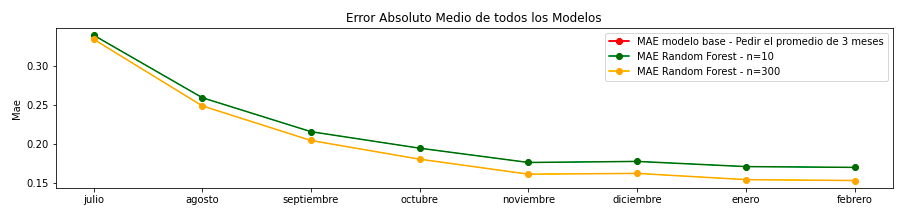


Figura 5. Gráfica de errores de entrenamiento.

Por otra parte, se construyó una gráfica similar de comparación para los errores de prueba en los tres modelos construidos: promedios móviles, bosque aleatorio con 10 árboles y bosque aleatorio con 300 árboles. Los resultados son evidencia de la excelente aplicación de los bosques aleatorios para las predicciones que genera y la ventaja que tiene el uso de un número mayor de árboles como parámetro decisivo para disminuir el error tanto de prueba como de entrenamiento.

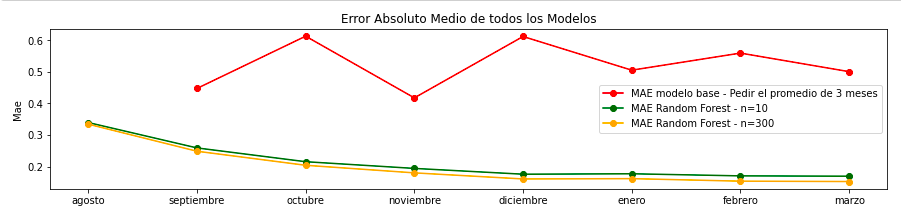


Figura 6. Gráfica de errores de prueba.

En la comparación de los resultados obtenidos a partir de la medición del error absoluto medio, podemos ver el grado de mejora en comparación con el modelo de promedios móviles y la diferencia también entre el cambio de parámetro n (número de árboles de decisión) para los modelos de *bosques aleatorios* ; incluso se puede ver también cómo a pesar de parecer una escasa diferencia entre el modelo con 10 y 300 árboles sí se aprecia que entre más grande sea el parámetro para ‘n’ menor será el error entre el valor obtenido y las predicciones y por lo tanto el bosque aleatorio de 300 árboles representa para efectos de los resultados, y considerando el uso del error medio absoluto como criterio, el resultado con un menor error en las predicciones.

Conclusiones

Si bien dentro de los modelos analizados a lo largo del periodo de construcción y evaluación de los modelos, en la metodología de desarrollo seguida, los resultados se acoplan al diseño del modelo y en consecuencia los resultados serán buenos o no a partir del criterio que se requiera cumplir para una problemática. En este caso lo que se busca, esencialmente, es el de obtener un modelo de predicción que permita resolver la predicción de la demanda; por ende, la solución al problema será el modelo estudiado que de una mejor predicción y de acuerdo con los resultados el que tenga el error de prueba más bajo. Dentro de los modelos que se estudiaron si bien los promedios móviles resultan ser útiles cuando no se tiene un historial de datos muy extenso y por esto se elige un periodo de tiempo determinado para hacer la predicción; sin embargo en este caso tenemos una serie de registros muy grande y como se muestra en la Figura 6, los datos que no forman parte del periodo de tiempo elegido no se toman en consideración por ello el error promedio medio resulta más grande que en el resto de los modelos con los que se le comparan en la gráfica. Las dos líneas que representan los resultados del modelo de bosques aleatorios se hicieron tomando en cuenta un número pequeño de árboles de decisión esto para ver el efecto de contraste entre distintos modelos (promedios móviles y bosques aleatorios). Como se puede apreciar en las gráficas de resultados tanto en los resultados de entrenamiento como en los de prueba obtenemos errores muchos menores que en los promedios móviles por lo tanto en cuanto a modelos los bosques aleatorios resultan mucho mejores para la predicción de la demanda de acuerdo con los resultados obtenidos para el error absoluto medio.

Si bien los bosques aleatorios suelen ser identificados como un modelo eficiente debido a que “los parámetros utilizados a menudo producen un buen resultado de predicción”, para fines prácticos de este proyecto sí resultaron ser una herramienta de predicción muy útil ya que debido al volumen de registros de los datos que utilizamos; a manera de comparar la eficiencia con base en el parámetro ‘n’ (número de árboles de decisión) se corrió el modelo con n=10 y n= 300 encontrando incluso errores menores en el bosque aleatorio con n=300, lo que quiere decir que a mayor número de árboles, más precisa será la predicción. Por lo anterior, y contemplado todo el análisis realizado en el presente estudio, podemos asegurar que el mejor modelo para la predicción de la demanda de Motorola resulta ser el algoritmo de bosques aleatorios con parámetro n=300; dicho modelo nos da una predicción de volumen de ventas que se ve reflejado en las predicciones obtenidas para los meses que se muestran en la Figura 7.

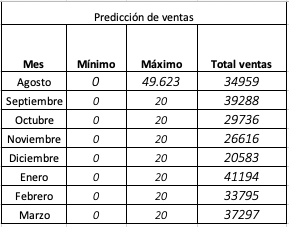


Figura 7. Predicciones de volúmen de ventas modelo de Bosque aleatorio (n=300)

A partir del modelo de bosques aleatorios, y en la parte de prueba del modelo, se rescataron las predicciones para los volúmenes de venta de unidades; esto con la intención de que a partir de un modelo confiable, con base en los resultados del error medio, se hicieran las recomendaciones de portafolio a Motorola a partir de las predicciones de dicho modelo. Lo obtenido, y mostrado en la Figura 7, nos indica que el punto más alto en términos de volúmenes de venta es en el periodo correspondiente al mes de Enero, y los periodos más bajos de Octubre a Diciembre; por lo anterior y con nuestro modelo como respaldo, lo que se aconseja es considerar una propuesta de Mercadotecnia para finales de año aprovechando las fiestas Navideñas, enfocada en elevar los volúmenes de venta para el periodo más bajo y mantener así la constancia en el volúmen de ventas a lo largo del año, de esta manera la marca se puede beneficiar para no sólo tener un solo mes de ventas altas sino trabajar en que todos los periodos sean exitosos y en consecuencia de las predicciones se tomen acciones que beneficien a la empresa económicamente y a la imagen de la marca.

A lo largo del presente trabajo se desarrolló un proyecto con bases de ciencia de datos que nos abrió un panorama completamente nuevo y profundo de conocimientos para el análisis de datos, y en este caso, sus aplicaciones a las proyecciones de demanda a través de modelos de regresión. Es importante reconocer que las aplicaciones no se limitan a las que se desarrollaron en el presente trabajo pero sí fungieron como una introducción enriquecedora para conocer las aplicaciones de los modelos de regresión y sobre todo aprender a usarlos.

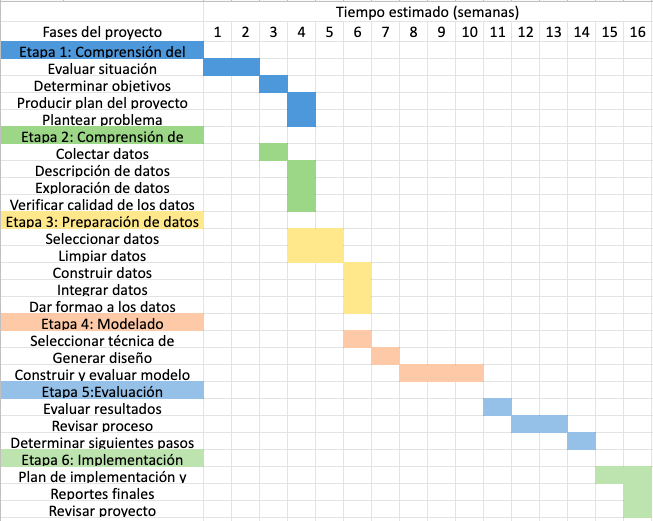
# Bibliografía

* Cahun, A.. (2020). Es oficial, Motorola regresará a la gama alta: su nuevo flagship será presentado en el MWC 2020. Septiembre 03. 2020, de Xalaca México Sitio web: <https://www.xataka.com.mx/celulares-y-smartphones/oficial-motorola-regresara-a-gama-alta-su-nuevo-flagship-sera-presentado-mwc-2020>
* Fernández, C.. (2019). Motorola crece en México gracias Moto G. Septiembre 03, 2020, de Expansión Sitio web: <https://expansion.mx/tecnologia/2019/02/07/motorola-crece-en-mexico-gracias-moto-g>
* FACULTAD DE CIENCIAS ECONÓMICAS, JURÍDICAS Y SOCIALES. (2018). Pronósticos de ventas. Octubre 12, 2020, de UNIVERSIDAD NACIONAL DE SALTA Sitio web: <http://www.marketingetico.com.ar/Pron%C3%B3sticos%20de%20Venta.pdf>
* Villareal, F. (2016). “Introducción a los Modelos de Pronósticos”. Octubre 12, 2020, de Universidad Nacional del Sur- Departamento de Matemática Sitio web: <https://www.matematica.uns.edu.ar/uma2016/material/Introduccion_a_los_Modelos_de_Pronosticos.pdf>
* Wikipedia. (2019). Media móvil. Octubre 12, 2020, de Wikipedia Sitio web: <https://es.wikipedia.org/wiki/Media_m%C3%B3vil>

# Anexos

## Anexos A

* Cronograma de actividades por fase

****

* Figuras ilustrativas de problemas de calidad

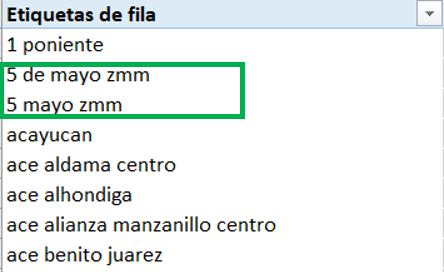


Figura 1. Tabla dinámica de punto de venta

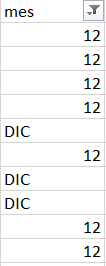


Figura 2. Filtro del mes de diciembre en número y letra

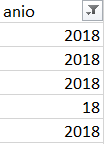


Figura 3. Año 2018 con cuatro y dos cifras

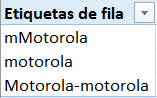


Figura 4. Tabla dinámica con tres tipos de escritura para Motorola

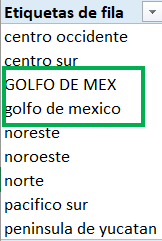


Figura 5. Tabla dinámica de la columna “Zona”

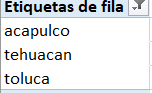


Figura 6. Tabla dinámica de la columna de estados, reflejando 3 ciudades

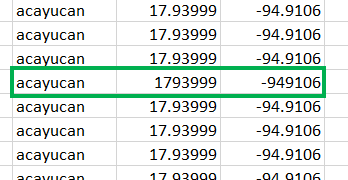


Figura 7. Caso “acayucan” en la columna latitud y longitud

## Anexos B

******

******

******

******

******

******

Figura 8. Creación de índices





Figura 9. Creación de columnas con índices

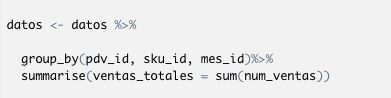
****

Figura 10. Código para agrupación de ventas por periodo

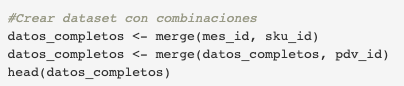








Figura 11. Códigos para completar las series de tiempo

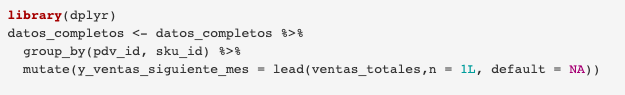
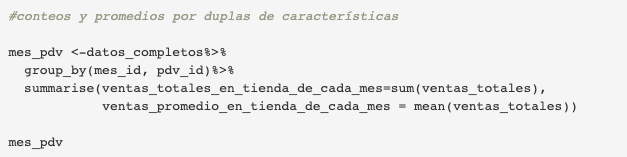




Figura 12. Código de construcción de variable respuesta



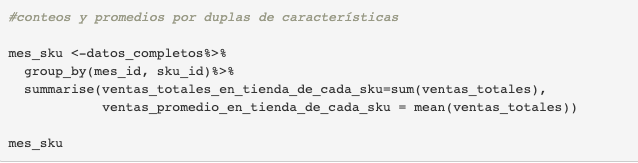


Figura 13. Creación de características.



## Anexos C

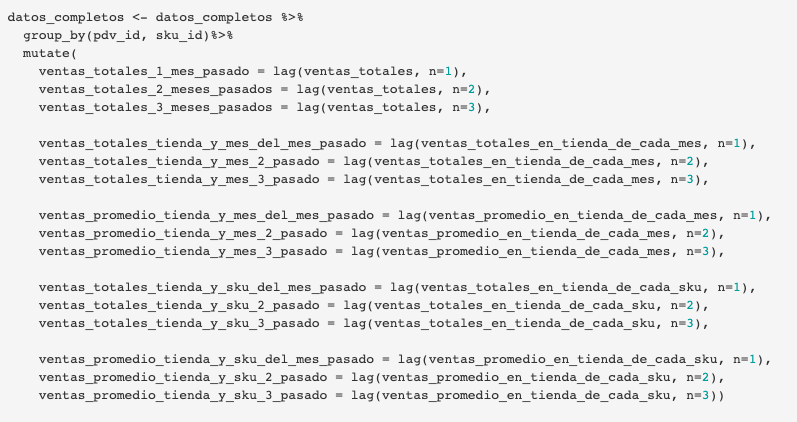




Figura 14. Adición de variables y creación de rezagos por punto de venta y modelo

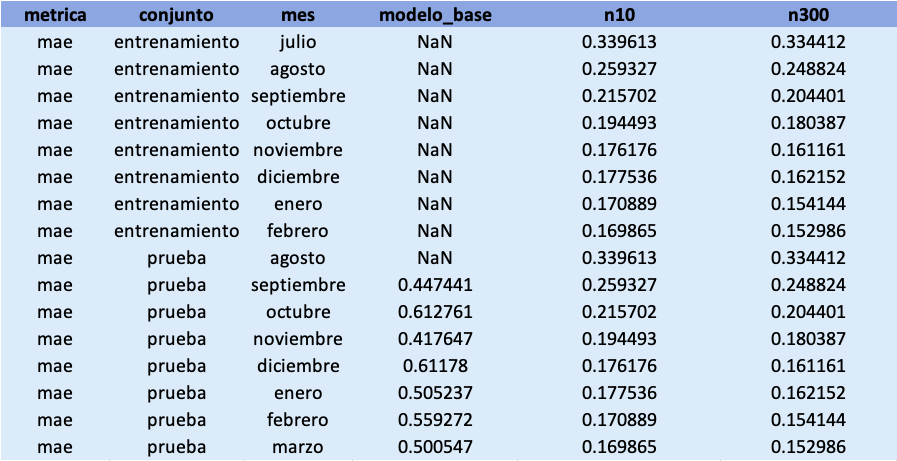


Figura 15. Tabla de resultados de mae para los modelos utilizados (promedios móviles y bosque aleatorio de n=10 y n=300)