

**LABORATORIO DE DISEÑO Y OPTIMIZACIÓN DE OPERACIONES**

**ENTREGA FINAL DE PROYECTO**

**ETAPA 5**

**EQUIPO 3**

Mauricio Rivas A01367206

Kayla Pierson A01364762

Carlos Gálvez A01362234

Carlos Aarón González A01365656

**INTRODUCCIÓN**

Se desarrolló un proyecto de ciencia de datos, basándonos en la metodología CRISP-DM, en el cual la empresa que analizamos fue Hisense, nuestro principal objetivo fue determinar las ventas totales que daban como resultado y hacer una predicción en cuanto a la cantidad de modelos a vender de telefonía para que sean favorables a la empresa, generando la respuesta a la problemática inicial que resaltaba.

Mediante las herramientas vistas en clase, llevamos a cabo las 5 etapas que concierne la metodología, obteniendo la mayor parte de la informaciónmediante la programación en el software de Rstudio. Para este estudio fue importante la parte del desarrollo que se tuvo con las predicciones, en las que pudimos proponer y crear modelos. Para esto se toma en cuenta que la predicción creada fue hecha a partir del modelo que escogimos; árbol de decisiones.

Trabajamos con medio millón de observaciones en excel, teniendo como soporte la herramienta para realizar el estudio en el software, con esto se resalta la importancia realizar este tipo de proyectos con herramientas de programación, ya que cuando se trabaja con big data no es posible hacer un análisis en excel y mucho menos que sea robusto. Con ayuda de la ciencia de datos podemos extraer esta big data que se nos proporciona para realizar estudios de cualquier tipo y emplearlos para poder hacer predicciones, mejoras, etc.

**ETAPA 1: Comprensión del negocio**

Pregunta a responder: ¿Cuántas unidades de los modelos seleccionados de la marca Hisense se venderán en el siguiente periodo de ventas?

1. Descripción de la situación actual

Hisense es un fabricante multinacional de electrodomésticos y productos electrónicos con sede en Qingdao, China. Hisense fue patrocinador de la Eurocopa 2016, la Copa FIFA confederaciones 2017, la Copa Mundial de Fútbol de 2018 y la Eurocopa 2020. Tiene 13 plantas de fabricación alrededor del mundo incluyendo a México.

A pesar de que sus principales productos de manufactura son televisores de pantalla plana, esta empresa está apostando por formar parte del mercado de telecomunicaciones por telefonía celular, ofreciendo a sus clientes mexicanos equipos de bajo costo.

La empresa señala que a pesar de las complicadas proyecciones, espera crecer entre un 15 y 20% en ingresos brutos y participar de manera más visible en el mercado nacional.

Con el fin de lograr posicionar a la empresa fuertemente en el mercado nacional, aplicaremos la metodología CRISP-DM. Al mismo tiempo, cumpliremos con las 6 fases y así alcanzar los objetivos de la firma.

* Business understanding
* Data understanding
* Data preparation
* Modeling
* Evaluation
* Deployment

1. Entender y describir la problemática

Según datos de STATISTA y el INEGI el uso de la telefonía celular ha ganado una gran importancia y se estima que en el 2020 había casi 81 millones de mexicanos que tenían un smartphone, con un crecimiento de casi un 14% en comparación al 2018, para el 2025 se proyecta que habrá 95 millones de smartphones.  
  
De esta forma es de suma importancia saber analizar los datos anteriores para lograr un pronóstico favorable para HISENSE que entra a esta categoría y desea mantenerse. Si este análisis no se hace de forma satisfactoria puede generar millones de dólares en pérdidas y posiblemente poner en riesgo la rentabilidad de la marca.

La empresa se presenta con miles de ventas y necesita estimar la cantidad correcta de smartphones para fabricar por los próximos años, de esta forma, no generar costos innecesarios cómo, exceso de inventario o falta de inventario, relaciones frágiles con proveedores, warehousing, transportación y diversos factores que se utilizan para satisfacer la demanda y a raíz de esto se generan más costos que están involucrados en la cadena de suministro, volviéndola ineficiente.

1. Entender y describir la problemática

3.1 Ciencia de datos

En la ciencia de datos se involucran diferentes áreas, tales como; estadística, data mining, análisis de datos, etc. Todo esto utilizado para analizar datos de clientes, compras, en plataformas digitales, mercado, etc. En este caso teniendo un enfoque en la industria electrónica de aparatos de uso electrodoméstico.

La empresa fue fundada en China y tiene sedes en diferentes partes alrededor del mundo, tales como; Francia, Pekín, Egipto, Argelia, México, Hungría, etc. Pero el enfoque principal del proyecto se encuentra en las plantas que están ubicadas en México.

El enfoque principal que se tiene de las variables a analizar son; el punto de venta, fecha, mes, año, número de venta, sku, gamma, costo promedio, zona, Estado, Ciudad, latitud y longitud. Las cuales nos guiaran en la parte del análisis para llegar a la meta a partir de estos datos.

3.2 Tipos de tareas

Existen distintos tipos de tareas que surgen a lo largo de la elaboración de un proyecto , una de los aspectos que se necesitan llevar a cabo es el proceso para tener predicciones exactas; como en el aspecto financiero, haciendo uso de la regresión.

Un evento que podría ocurrir sería tener que hacer clasificaciones o agrupaciones con los datos obtenidos a lo largo del proyecto dependiendo de los resultados que se obtengan para generar un reporte más preciso y detallado, pero este ya no sería tomado en cuenta la parte inicial de clasificar y agrupar datos.

3.3 Tipo de dato

Los datos proporcionados y sobre los cuáles se van a trabajar son estructurados ya que son fáciles de almacenar y clasificar por subgrupos, además de poder ser interpretados con mayor facilidad. Estos datos se manejan como una base de datos en tablas por su tipo de clasificación.

3.4 Data people

El equipo de proyecto se conforma por cuatro integrantes, de los cuáles cada uno tiene un área específica en la que se tiene mayor conocimiento. Dichas áreas son; big data, estadística, matemáticas y programación.

Esto hace que el proyecto sea más fácil de manejar y se puedan tener inspecciones y concentraciones más específicas para verificar que se cumple con el objetivo del proyecto y se llevará al éxito.

1. Objetivos:

4.1 Llevar a cabo la metodología del CRISP-DM de una manera correcta en el proyecto.

4.2 Entender la importancia de la Ciencia de Datos hoy en día e implementarla, debido al desarrollo que tiene hoy en día.

4.3 Saber como ejecutar un proyecto de Ciencia de Datos desde cero.

4.4 Saber cómo analizar e interpretar correctamente las variables y los resultados obtenidos.

4.5 Analizar correctamente los datos obtenidos, involucrando la empresa dada.

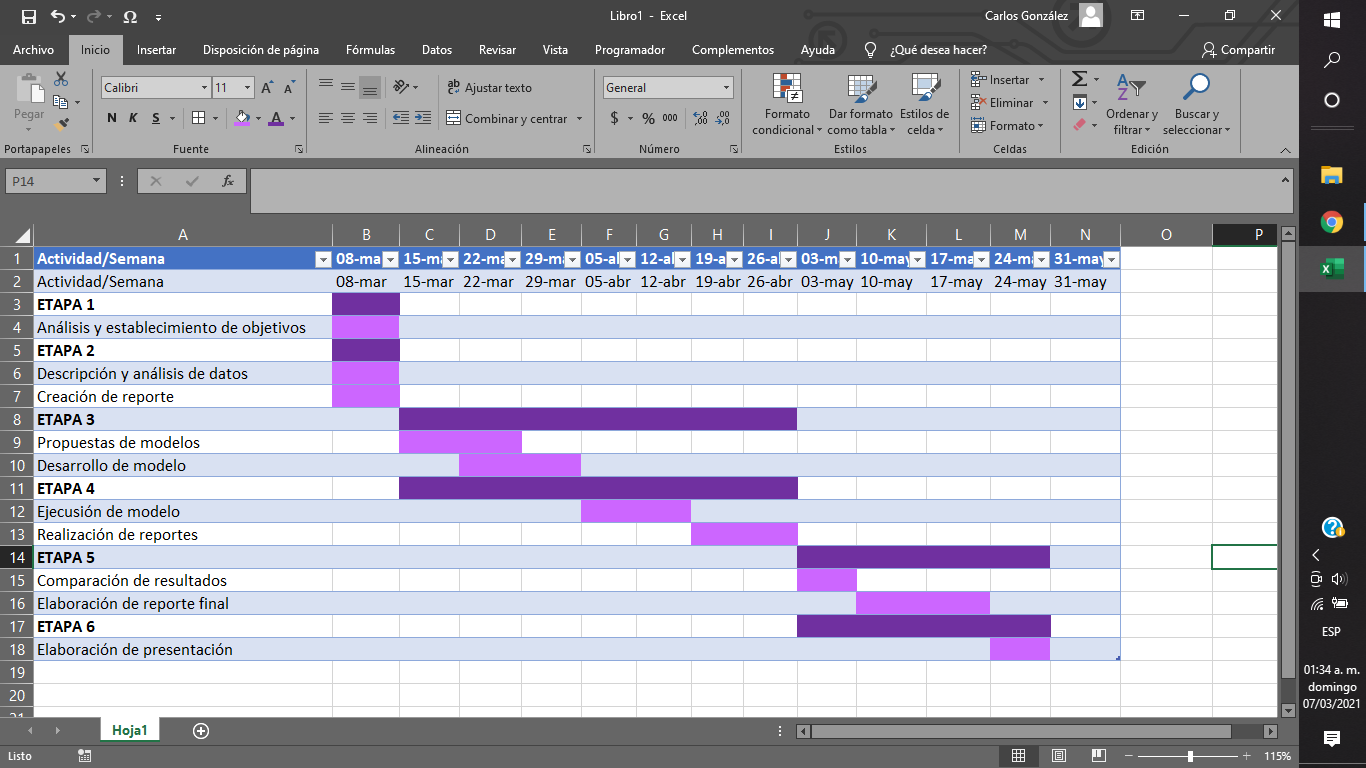
4.6 Desarrollar el campo de programación con un enfoque en la Ciencia de Datos.

1. Estructurar el proyecto y hacer un plan preliminar

El proyecto comprenderá las siguientes etapas hasta ser finalizado:

* Etapa 1: Comprensión del Negocio: Lo que se busca es realizar un análisis del contexto y las necesidades del negocio en cuestión para lograr esclarecer los objetivos del proyecto.
* Etapa 2: Comprensión de los datos: Posteriormente se analizarán los datos de manera que su composición y posibles defectos puedan ser comprendidos por todo el equipo.
* Etapa 3: Preparación de los datos: En esta etapa se realizará un análisis más detallado de los datos así como una preparación de los mismos para ser analizados a través del planteamiento inicial de un modelo.
* Etapa 4: Modelado: En esta etapa se analizarán los datos con el modelo ya desarrollado previamente.
* Etapa 5: Evaluación (Resultados): Con la obtención de los resultados de la etapa anterior se analizará el desempeño de los modelos para que sus resultados puedan ser validados con un buen nivel de confianza.
* Etapa 6: Implementación: Esta etapa consistirá en la realización de una presentación en la que se corroboren todos los resultados y los pasos llevados a cabo anteriormente.

GANTT

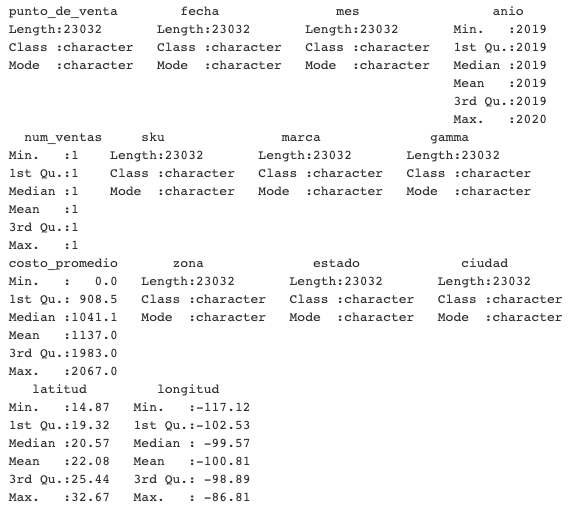


Para las actividades descritas en el diagrama de Gantt, en todas participarán los miembros del equipo de manera equitativa.

**ETAPA 2: Comprensión de los datos**

1. Describir los datos crudos:

Los datos se encuentran en formato de excel “.csv”. Se utilizó Rstudio para visualizarlos y pudimos observar que se cuenta con 23,032 entradas las cuales se dividen en 14 variables de diferente clases, como caracteres y enteras. a continuación se presenta un resumen de los datos obtenido en Rstudio:



Con el objetivo de facilitar la comprensión de los datos y el significado de cada variable, se realizó un diccionario que incluye características como la clase de la variable y su significado dentro del proyecto, el diccionario es el siguiente:

|  |  |  |
| --- | --- | --- |
| **Nombre de Variable:** | **Clase:** | **Significado:** |
| punto\_de\_venta | Caracteres | Puntos de venta de la Hisense |
| fecha | Caracteres | Fecha en la que se vendió el producto |
| mes | Caracteres | Mes en el que se vendió el producto |
| anio | Entero | Año en el que se vendió el producto |
| num\_ventas | Entero | Cantidad del producto vendido |
| sku | Caracteres | “Stock Keeping Unit” (Para identificar el producto) |
| marca | Caracteres | Marca del producto (Hisense) |
| gamma | Caracteres | Segmento de mercado a la que pertenece el producto |
| costo\_promedio | Número | Costo del producto |
| zona | Caracteres | Zona de ubicación del punto de venta |
| estado | Caracteres | Estado en el que se ubica el punto de venta (P.V) |
| ciudad | Caracteres | Ciudad en la que se ubica el P.V |
| latitud | Número | Latitud de la ubicación del P.V |
| longitud | Número | Longitud de la ubicación del P.V |

1. Detectar problemas de calidad:

Algunas variables serán modificadas para un mejor manejo de los datos, por lo que el *tipo* de variable cambiaria en los siguientes casos: mes de character a factor, zona de character a factor, estado de character a factor y ciudad de character a factor. Esto debido a que las consideramos variables cualitativas.

Además, existen algunos errores en la entrada de datos de algunas variables que hay que modificar; estos se presentarán a continuación: Punto\_de\_venta: 5 puntos de venta escritos de manera errónea, Mes: 5 meses registrados con letras en lugar de números, hay que cambiar las letras. Anio: Los años deben de registrarse con 4 dígitos, hay al menos un registro con solo 2 dígitos que se debe(n) modificar. Marca: Existen 5 marcas escritas de manera errónea. Zona: Hay un registro mal escrito. Estado: Existen 3 estados más de los que debería. Latitud y Longitud: Hay 1 valor fuera de rango. Dichos errores se consideran problemas en la calidad de los datos debido a que dificultan el tratado correcto de los datos siendo que el programa utilizado para analizarlos podría no considerarlos, clasificarlos en categorías erróneas, o bien, podrían influir en los resultados finales de manera significativa si los valores que poseen se encuentran fuera de los rangos aceptables debido a un error de escritura de los mismos.

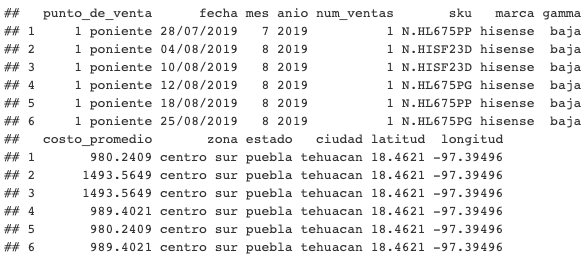
Fuera de lo ya mencionado se considera que los datos se encuentran en buen estado, una vez realizada la limpieza de datos se podrá pasar a la siguiente etapa.

**ETAPA 3: Limpieza de Datos**

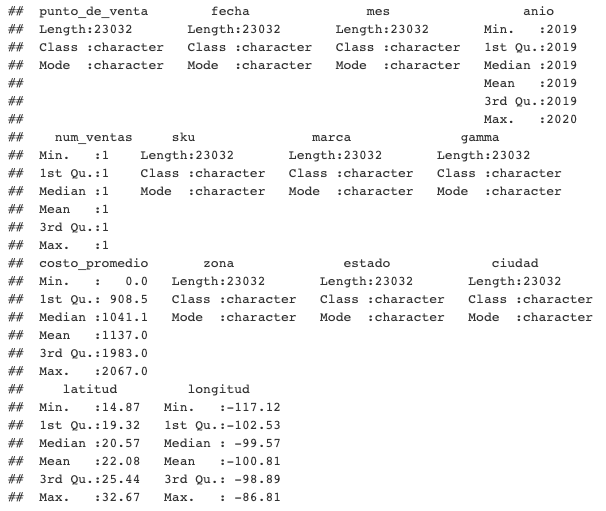
**3.1 Procedimiento para la limpieza de datos**

En esta etapa identificamos la cantidad de observaciones y la composición de sus variables así mismo el tipo al que pertenecen y de esta forma proceder al ajuste de datos ingresados incorrectamente.

Continuamos identificando los formatos que se repiten con una función “unique” para utilizar los valores que aportan al análisis y corregir los que no. Para algunos casos no fue complicado encontrar cuáles eran las variables incorrectas, ya que algunas estaban marcadas muy fuera del rango que se tenía, hablando de las numéricas. Por otro lado, puede ser más tedioso encontrar el error cuando se tienen faltas de ortografía.  
  
De igual forma se cambian las variables de “chr” a numéricas para utilizar los meses como números y no como caracteres o texto.

Se muestra las diferentes variables que hay, la clasificación que tienen y el nombre o carácter que se asigna para cada una, una vez que ya se terminó de hacer la limpieza de datos. ****

Después se muestra de manera más específica, mostrando el resumen de los datos con su clasificación final y los rangos de los datos.

****

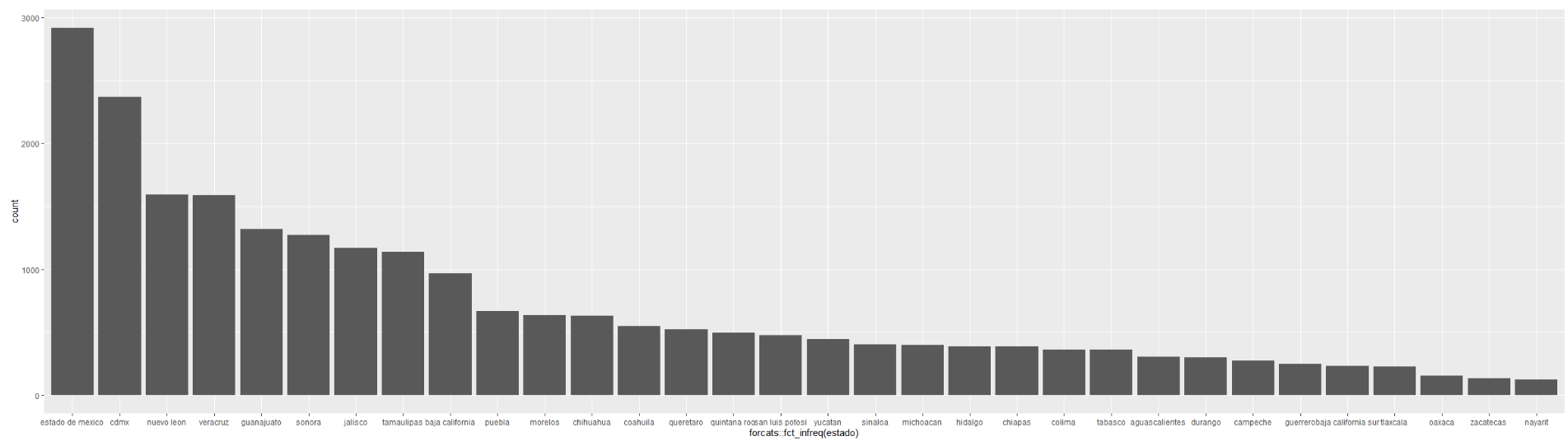
**3.2 Situación actual**

Una vez que ya se concluyó con la limpieza de datos, se pudo seguir con las demás etapas analíticas del proyecto sin presentar problemas por esta parte.

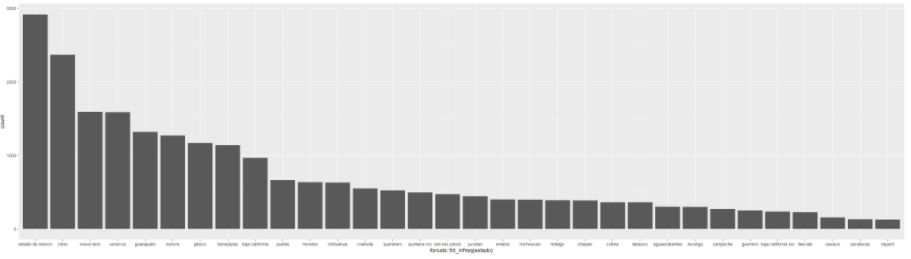
Se hizo un análisis exploratorio de datos, ingeniería de datos y un modelado de aprendizaje de máquina. De esta manera ya se pueden estudiar mejor los datos, acomodarlos en diferentes categorías y hacer un análisis robusto para cumplir con las especificaciones y brindar resultados.

La etapa más tediosa del proyecto que era la limpieza de datos, hace una vez concluida más fácil el proyecto, corrigiendo los datos y asegurando la parte de calidad en el proyecto.

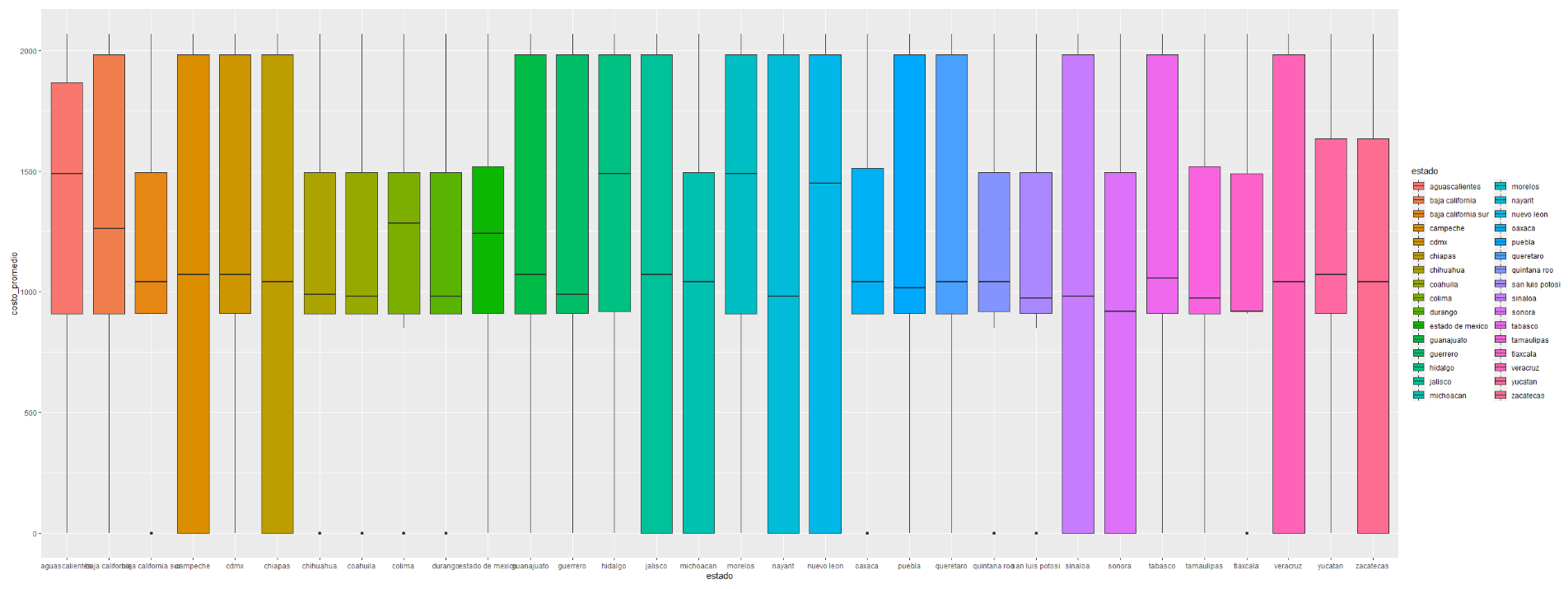
**3.3 Comportamiento de datos con el tiempo**

Al analizar el comportamiento de los datos logramos comprender la frecuencia con la que se adquieren los dispositivos de Hisense, del mismo modo logramos distinguir cual es el estado de la república más competitivo y el menos adecuado para vender o al que más le debemos prestar atención.  
  


También se hizo un análisis para ver cuál era el Estado en el que se vendía más el producto, junto con los Estados que tenían un número mucho menor de ventas, con un aproximado de 1000 unidades cada uno. Teniendo como los más populares la Ciudad de México y el Estado de México.



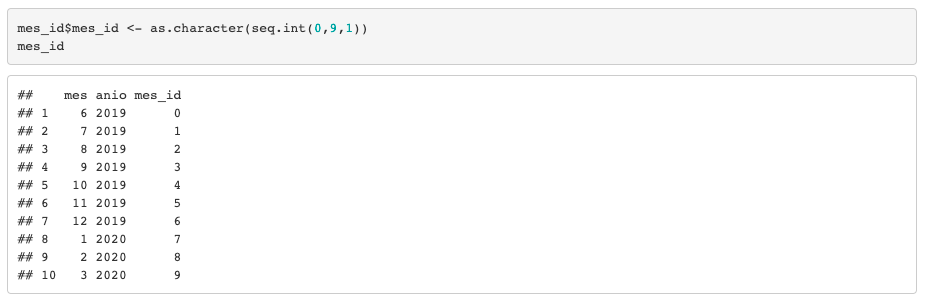
Las gráficas de caja nos ayudaron a corroborar el costo promedio más alto por estado, siendo Aguascalientes, Hidalgo y Morelos los que se llevan el mayor de esta clasificación.

****

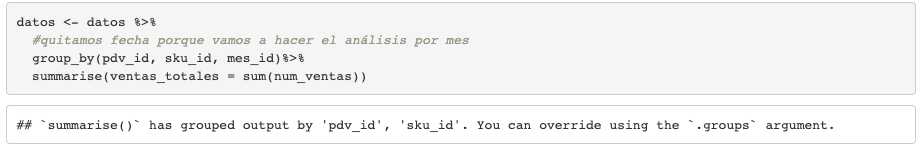
**3.4 Ingeniería de características**

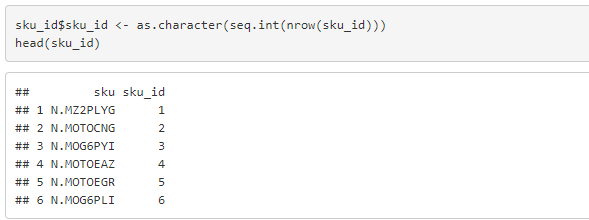
Con la Ingeniería de características lo que buscamos es generar características adicionales a las existentes, posterior a esto se comienzan a eliminar las irrelevantes o las que no sean de mucha ayuda en cuanto a la variabilidad de las respuestas que haya.

Para este apartado lo primero que se tiene que hacer es crear los índices por separado de las variables cualitativas más importantes y de esta forma darle un seguimiento durante todo el proyecto sin perder la capacidad de análisis.

****

Después vamos a comenzar con la lectura de estos datos ya clasificados de manera más específica para poder agrupar las ventas totales en un apartado.

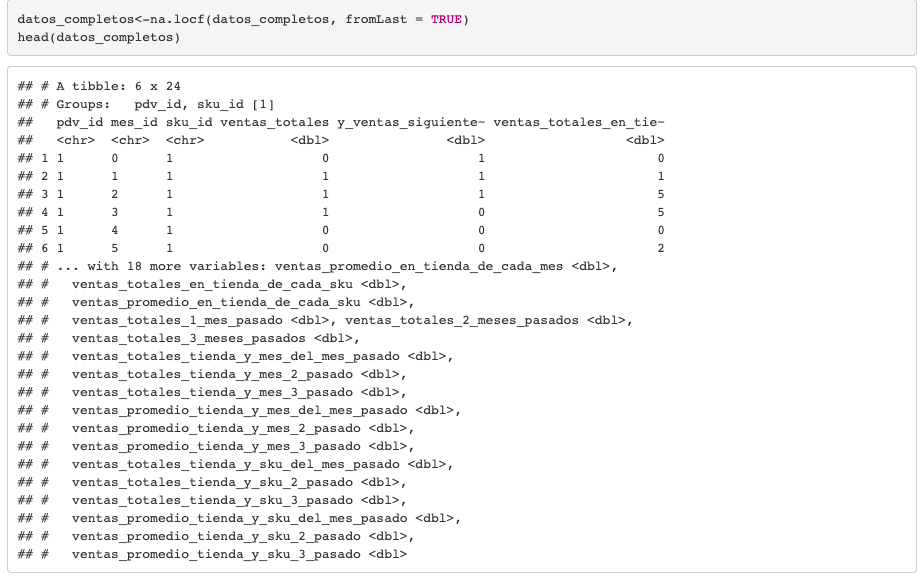
****

****

Continuamos asignando el índice al Stock Keeping Unit (SKU) de los productos de la empresa Hisense.

Es importante no asignarle un índice a todas las variables que se encuentran como: marca, gamma, costo promedio, zona, estado, ciudad, latitud, longitud, ni ventas totales.

Una vez que ya se hayan completado las series de tiempo se puede llegar a la variable respuesta; que en este caso son las ventas. A partir de esto, ya se generan las nuevas características que buscamos, como lo son; conteos, rezagos y promedios.



**ETAPA 4: Modelado**

1. **Modelo de promedios móviles**

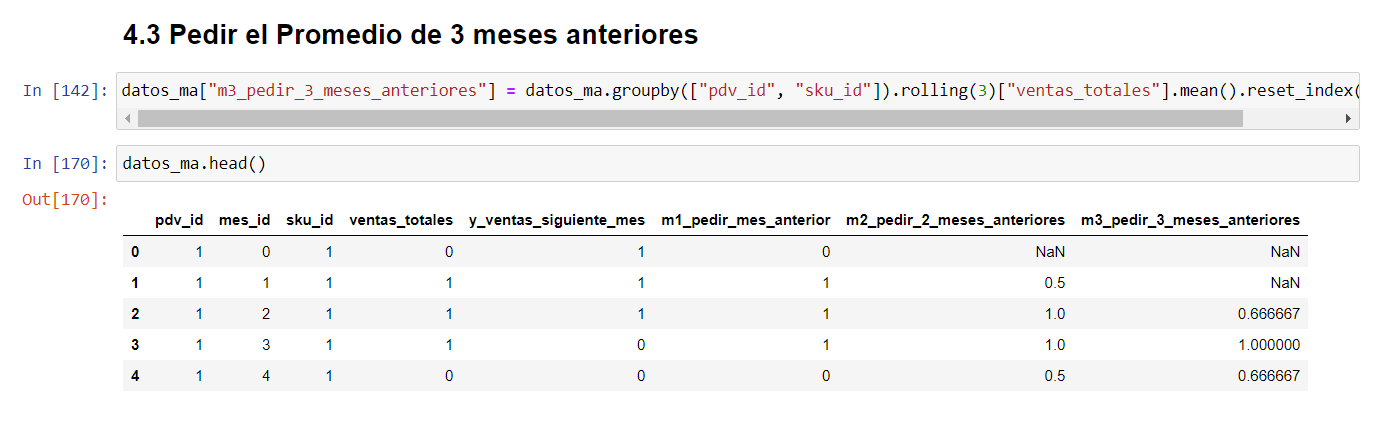
Es un método de pronóstico que utiliza promedios calculados a partir de subgrupos de observaciones consecutivas, donde el tamaño de los subgrupos será definido y constante, esto con el objetivo de que los efectos estacionales o irregulares sean eliminados. Es una herramienta bastante utilizada ya que es de fácil realización, sin embargo hay que tener en cuenta que la demanda no debe de tener tendencia ni estacionalidad.

* 1. **Consideraciones**

El pronóstico de promedio móvil es utilizado cuando existen patrones de demanda aleatorios o nivelados de los cuales se pretende eliminar el impacto de los elementos irregulares utilizando un enfoque de periodos de demanda reciente.

Creemos que por la naturaleza de nuestro proyecto el uso de este pronóstico será de bastante utilidad considerando que se requiere estimar la cantidad de productos smartphone a producir en los próximos años, y sabiendo que contamos con datos históricos de la compañía.

Para el cálculo de promedios móviles se tiene que definir un *periodo móvil (PM),* el cual define el tamaño de los subgrupos utilizados para calcular el promedio y predecir el siguiente periodo. En nuestro proyecto se construyó un modelo que considero 3 periodos móviles diferentes, de tamaños: n=1, n=2 y n=3. Esto quiere decir que para calcular el promedio de n=3, se utilizaron los tres periodos anteriores al que se desea calcular. A continuación se presenta una tabla con los diferentes promedios móviles considerando las primeras 5 entradas de nuestros datos ya limpios:



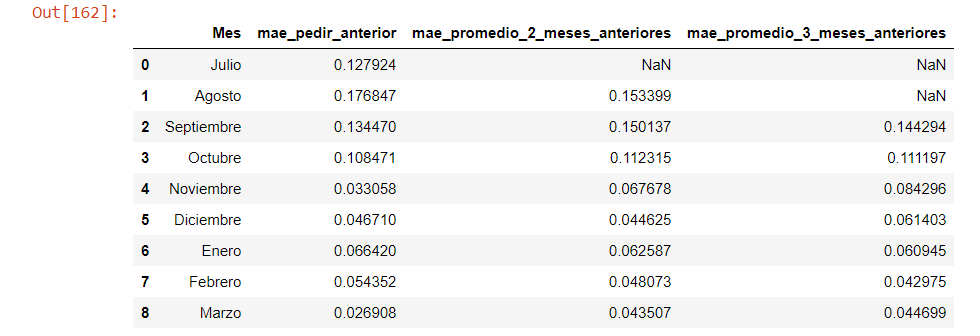
Finalmente es importante considerar que se busca el modelo con el periodo móvil más estable, por lo que más adelante con el uso de Python se analizarán los errores absolutos de cada PM con el fin de encontrar cual de los tres será más certero en el momento de realizar predicciones.

* 1. **Limitantes**

Si bien la herramienta de promedios móviles simples es muy útil para elaborar pronósticos rápidos con un buen nivel de confianza, también presentan algunas limitaciones entre las que destacan:

* Solamente pueden llegar a pronosticar un período más.
* No pueden utilizarse si los datos presentan tendencia o estacionalidad.
* Las medias móviles dibujan las tendencias de la información anterior sin tomar en cuenta los cambios que pueden afectar a futuro, como nuevos competidores, mayor o menor demanda de productos en la industria y cambios en la estructura de gestión de una empresa.
* Presenta una mayor complicación para responder a cambios bruscos.

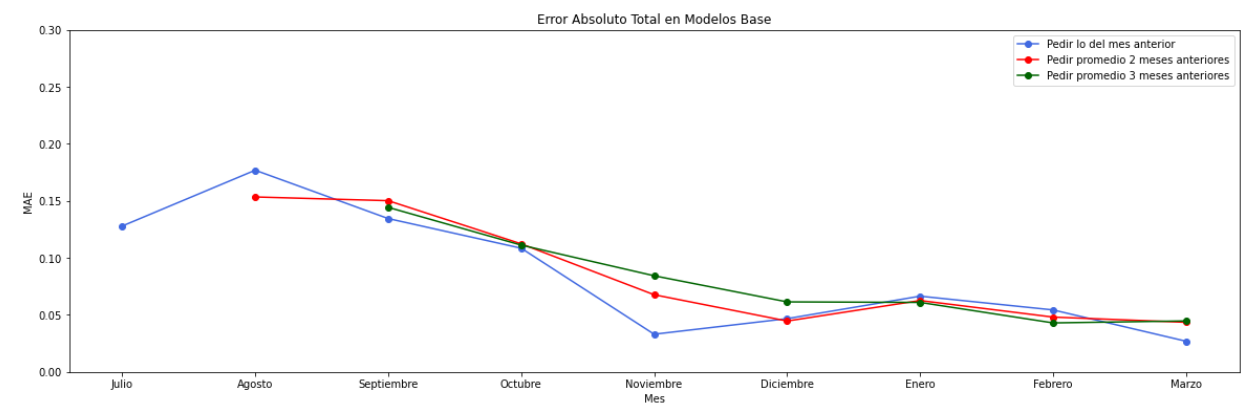
Como ya se explicó anteriormente, para el propósito de este proyecto se utilizaron 3 diferentes periodos móviles, una limitante en el uso de este pronóstico es que es necesario el cálculo de los errores absolutos para saber cuál será de mayor utilidad cuando se quiera hacer alguna predicción. El cálculo de estos errores absolutos se realizó en Python de manera individual para cada mes que se consideró y se acomodo en la siguiente tabla para poder realizar un análisis de desempeño:



Se busca utilizar el modelo con mayor estabilidad para realizar predicciones, no necesariamente el que sea más acertado en ciertos periodos. Más adelante en este reporte se hará el análisis de rendimiento en donde se discutirán estos datos.

Por las limitantes presentadas, se comparará el modelo de promedios móviles con un modelo de aprendizaje de máquina en la siguiente entrega de proyecto. Esto con el propósito de identificar con cuál de los dos modelos se puede obtener un mejor desempeño.

* 1. **Evaluación del rendimiento**

Los tres modelos de promedios móviles tuvieron un desempeño aceptable para todo el año. Si bien el punto de mayor error fue registrado en agosto por el modelo de n=1 así como también fue este modelo el que se mostró con mayores variaciones en cuanto a este valor a lo largo de los meses, este modelo fue el que menor error absoluto promedio mostró de los tres modelos y su variación puede considerarse aún baja para un estándar aceptable. 

Podemos decir que el comportamiento relativamente estable de los tres modelos nos muestra que a pesar de que las ventas de los distintos productos varían mes con mes, no lo hacen de forma brusca con respecto al mes anterior.

Cabe señalar que los tres modelos registraron mayor incidencia de error en los meses de verano y enero, lo cual puede señalar un cambio significativo en las ventas para estos periodos.

Finalmente asumimos que el mejor modelo para predecir en nuestro caso será con un periodo móvil de 3 meses, a pesar de que en algunos meses podemos observar que no tiene el error más pequeño, se da preferencia a un modelo más estable donde los errores se mantengan relativamente constantes.

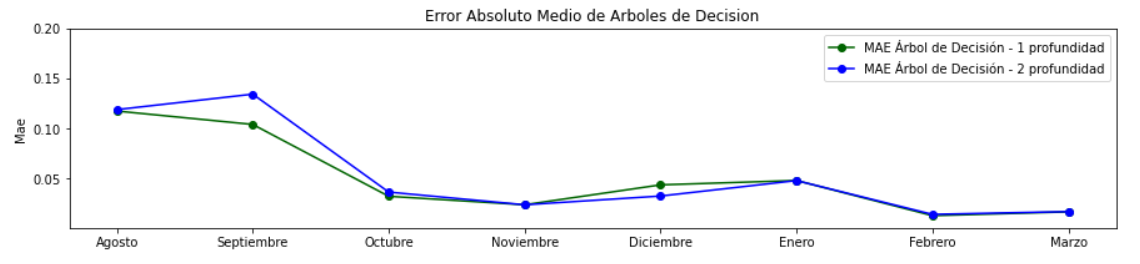
1. **Modelo Arbol de Decision**

Los modelos de árboles de decisión son modelos predictivos conformados por reglas binarias que buscan repartir las observaciones en función de sus atributos para predecir el valor de la variable respuesta. En los modelos de Árbol de decisión el parámetro más importante a considerar es la profundidad del modelo, para este proyecto se realizaron dos modelos de árbol de decisión con profundidades de 1 y 2.

Para estos modelos se utilizó validación cruzada con el objetivo de evitar el sobreajuste, esta validación cruzada tiene que considerar que nuestros datos tienen una secuencia periódica y se tiene que respetar. Para esto realizamos particiones de los datos en 8, antes dividiéndolos en datos de entrenamiento y prueba.

**2.1 Evaluación del rendimiento**

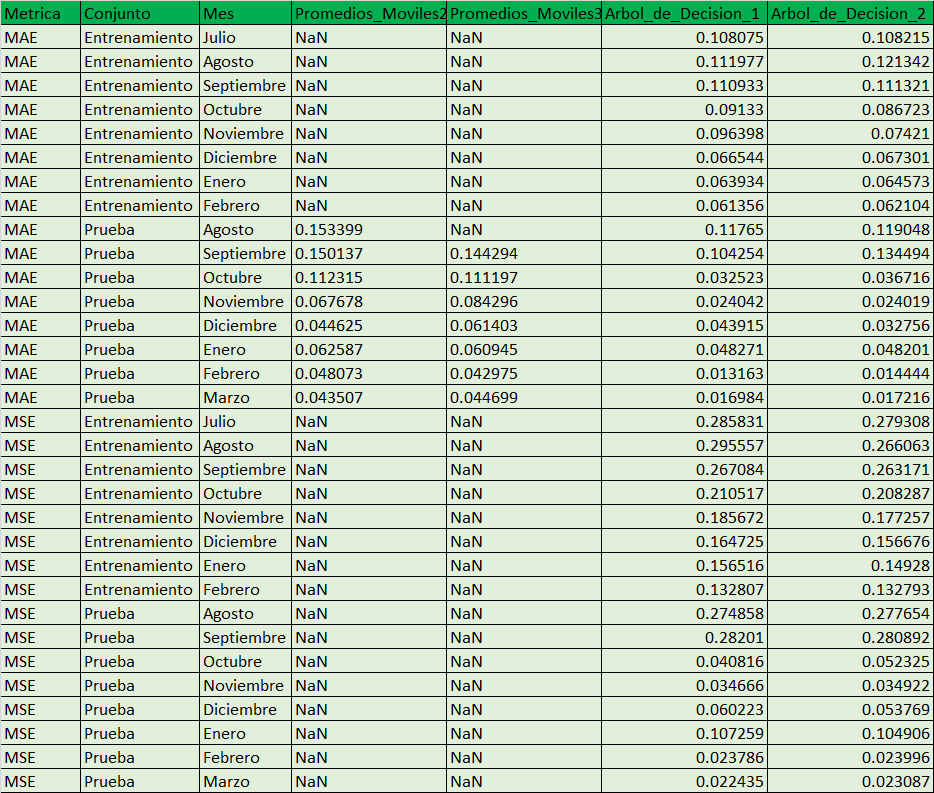
El MAE de prueba obtenido utilizando la profundidad de 1 y 2 con los datos de Hisense son los siguientes:

****

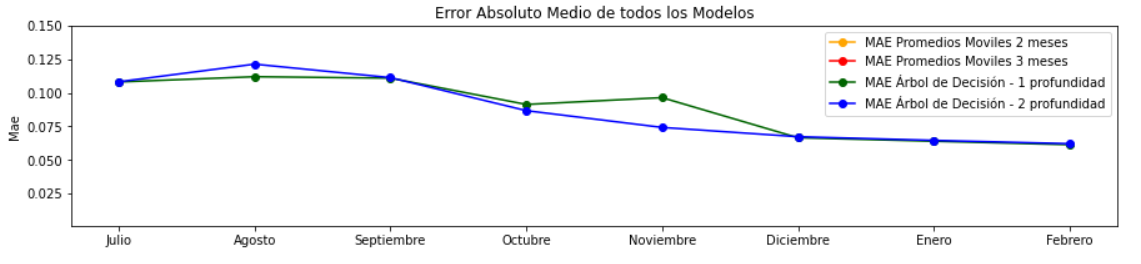
Como podemos observar un Árbol de Decisión de profundidad 2 tiene un error absoluto menor al de un modelo con profundidad de 1, al mismo tiempo este modelo mantiene un error estable lo cual nos da confianza a la hora de hacer predicciones. Es por esto que consideramos que un modelo con una profundidad de 2 es mejor. En la siguiente etapa compararemos los 4 modelos vistos previamente.

**ETAPA 5: Evaluación (Resultados)**

En la siguiente tabla presentamos los errores MAE y MSE de los 4 modelos que se compararon en el proyecto:

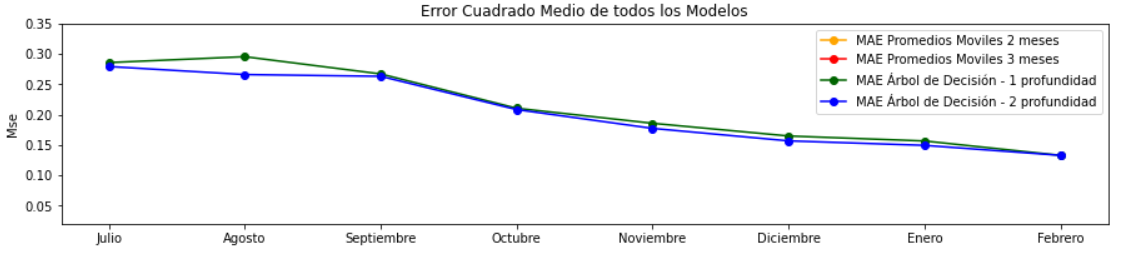


**MAE de entrenamiento de todos los modelos (Solo aplica para árbol de decisión):**



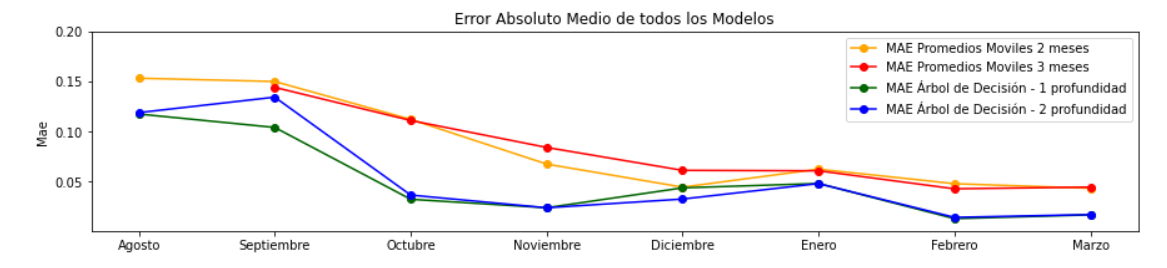
Se observan los errores de entrenamiento MAE de los árboles de decisión con profundidad 1 en color verde y profundidad 2 en color azul. Podemos observar que el modelo de árbol de decisión con profundidad de 2 nos arroja una línea con menos picos lo cual es mejor para hacer predicciones ya que tenemos la certeza de tener predicciones estables, es por esto que concluimos que este modelo es mejor comparado con el árbol de decisión de profundidad 1.

**MSE de entrenamiento de todos los modelos (Solo aplica para árbol de decisión):**



De igual manera se presentan los errores MSE de entrenamiento en verde para profundidad 1 y en azul para profundidad 2. Por la misma razón mencionada en la gráfica anterior, concluimos que la profundidad de 2 es más certera que 1, ya que no observamos picos en las predicciones y al mismo tiempo tiene un error más pequeño que el del modelo de árbol de decisión con profundidad 1.

**MAE de prueba de todos los modelos:**



Para terminar comparamos el error MAE de prueba con todos los modelos y sus variantes y concluimos que el árbol de decisión con profundidad de 2 es el más certero para realizar la predicción de las ventas de Hisense, ya que es el modelo que tiene el error MAE mas pequeño y al mismo tiempo, a pesar de presentar algunos picos, se mantiene relativamente estable a lo largo de los meses lo cual nos favorece para hacer predicciones. Es por esto que el modelo a utilizar para realizar predicciones es el árbol de decisión de profundidad 2.

Finalmente, al utilizar este modelo para realizar las predicciones, la cantidad de producto a vender en el siguiente periodo de venta es la siguiente:



**CONCLUSIONES**

Trabajar con medio millón de datos durante este proyecto fue retador y lleno de mucha información, desde entender la perspectiva del negocio hasta identificar los valores que no aportan al análisis de los datos. El análisis y la corrección de los mismos pueden llegar a ser los pasos más tardados e importantes a los que nos enfrentamos, la ingeniería de características nos ayudó a complementar el análisis y de esta manera considerar las variables cualitativas más relevantes en el dataset. Finalmente la aplicación de dos modelos para lograr diseñar una predicción que fortalezca la decisión de Hisense en qué temporada del año regular producir más unidades.

Los resultados fueron favorables para este proyecto ya que, ambos modelos resultaron con un valor de error bajo. El modelo con el que concluimos el proyecto es el de árbol de decisión con una profundidad de 2 y presentar la predicción de los dispositivos a producir por mes.

La metodología CRISP-DM junto con la jerarquía de sus etapas generaron un abundante flujo de información en los miembros del equipo para lograr el objetivo del proyecto.

**Bibliografía**

Masetto, A. (2021). Laboratorio de diseño y optimización de operaciones. Canvas ITESM. <https://experiencia21.tec.mx/courses/124327>