Tecnológico de Monterrey

Campus Toluca

Entregable final del Proyecto

Materia: IN3038.1 Laboratorio de Diseño y Optimización de Operaciones

Profesora: M. en C. Ana Luisa Masetto Herrera

Equipo número 3:

Alejandra Velazquez Bastida A01368039
Arlin San Juan Pezat A01368226
Yessica Vidal Castellanos A01366760
Alejandro Gabriel Hernandez A01367757
Juan Carlos Robles Guicho A01368398

Fecha de entrega: Viernes 19 de noviembre del 2021

Semestre Agosto- Diciembre 2021

0. Introducción.

La industria de las telecomunicaciones ha crecido radicalmente en los últimos años gracias a los múltiples adelantos tecnológicos que se han alcanzado; en esta industria está una sub área enfocada enfocado a la telefonía celular, sector que además ha presentado un crecimiento destacable en su demanda de productos.

Pese a que la necesidad de tener un dispositivo móvil es de la mayor parte de los individuos, no todos buscan las mismas propiedades en dichos dispositivos. Es por esa razón, por lo cual las compañías enfocadas a la comercialización de dichos productos afrontan como desafío primordial el profetizar el número unidades a vender de cada producto en sus varios puntos de comercialización, debido a que de no hacerse propiamente esto podría producir diferentes inconvenientes, como lo son precios logísticos desmesurados o insatisfacción de consumidores.

En el transcurso de este proyecto desarrollaremos portafolios de venta que nos permitan predecir la demanda de los diferentes puntos de venta para la marca *Hisense*. Todo esto con el desarrollo de un proyecto de ciencia de datos, empleando la metodología y la estructura CRISP-DM. Emplearemos la base de datos de la compañía, que previamente se nos ha brindado con el fin de contestar nuestra pregunta objetivo:

1. Etapa 1- Comprensión del Negocio.

1.1 Descripción de la situación actual.

Hisense llega a México en 2011 y en tan solo una década, el país se posicionó en el tercer mercado más importante para la empresa, debido a la relación costo beneficio sus diferentes productos electrodomésticos y de línea blanca. Se sabe que entre el 20% y 25% de la producción se queda para consumo local [1]. Derivado de la pandemia del COVID-19, la escasez de chips y el aumento en los precios del cobre, aluminio y plástico, la cadena de suministro de Hisense se ha visto afectada drásticamente. La empresa ha sufrido una disminución en la participación del mercado de venta de smartphones, cayendo del 0.05% al 0.022% del 2020 a 2021. Aún con todo esto la empresa ha priorizado a México como uno de los principales mercados para sus productos (Statista Research Department, 2021).

Por contexto como el anterior nace la importancia de realizar un proyecto como este, un ingeniero industrial puede tomar decisiones con base en datos analizados estadística y matemáticamente con programación y no solo por intuición. Esto constituye uno de los perfiles profesionales más relevantes en la actualidad; el poder entender y analizar sistemas complejos

1.2 Entender y describir la problemática.

Es de suma importancia que la empresa comience a generar estrategias que le permitan aprovechar el creciente mercado de celulares, debido a su gran desempeño el año pasado pese

a la pandemia del COVID-19. Además, una correcta alineación de la producción con la demanda le permitirá a Hisense afrontar los escenarios presentes. Para este problema un ingeniero industrial con conocimientos en ciencia de datos puede auxiliar a la empresa en el proceso de toma de decisiones.

1.3 Entender y describir la problemática. **(en términos de ciencia de datos)**
La empresa Hisense requiere la construcción de portafolios de productos (predicción de demanda) para sus diferentes puntos de venta. Para nuestro caso de estudio tenemos un problema de regresión, pues vamos a pronosticar un valor numérico con base en datos **estructurados** que se nos han brindado. Con este proyecto buscamos dar respuesta a la interrogante: **¿Cuántos celulares se venderán en cada punto de venta?** Generando así una reducción en los costos logísticos y un aumento en la satisfacción del cliente.

1.4 Plasmar los objetivos. Nuestro objetivo como equipo es lograr el entendimiento y la correcta aplicación de un proyecto de ciencia de datos, con la comprensión y estructura de la metodología CRISP-DM.

Etapa	Objetivos	Indicador de éxito
Businesss Understanding	Entender el contexto, historia, objetivos, criterios de éxito, riesgos, recursos, terminología y la problemática de la empresa.	Reporte y desglose de sucesos relevante de la empresa.
	Identificar los objetivos del data mining.	Tabla de objetivos e indicadores por fase.
	Definir plan inicial.	Diagrama de Gant.
Data Understanding	Recolección/ solicitud de datos.	Base de datos completa.
	Comprensión de la terminología interna de la empresa.	Diccionario de las variables a utilizar.
	Descripación de datos.	División y clasificación de variables.
	Verificar calidad de dato/7Indetificar el tamaño y tipo de archivo de nuestros datos.	Lectura de archivo CSV, reporte de calidad datos.
Data preparation	Limpieza correcta de datos.	Datos estandarizados, congruentes y en orden.
	Selección del software o plataforma para el manejo y análisis de datos.	Software, plataforma instalada.
	Archivo preparado para el software a utilizar (CSV, XML, HTML, etc.)	Archivo listo.
Modeling	Selección correcta de algoritmo para pronosticar y técnicas a emplear.	Notas y ecuaciones listas.
	Generar test design.	Fracción de datos a utilizar en test design.
	Construcción efectiva del modelo	Código funcional.
	Definición de parámetros del modelo.	Métrica de eficiencia del modelo.
Evaluation	Evaluación efectiva de resultados.	Resultados alineados a criterios de evaluación de la empresa.
	Revisión y mapeo de proceso.	Checklist de proceso.
	Definir next steps.	Planeación de cierre.
Deployment	Definir plan de documentación.	Checklist de etapas del proceso a documentar.
	Definir plan de monitoreo y control.	Criterios / actividades de monitoreo y control definidos.
	Documentación de experiencia.	Reporte final.

Fig 1. Tabla de objetivos por etapa. Ver completo: https://docs.google.com/spreadsheets/d/1s1dyiiofhoxb-HrmjxUCh9EiTJFuIDLk/edit?usp=sharing&ouid=101688482481224037026&rtpof=true&sd=true

1.5 Estructurar el proyecto y hacer un plan preliminar.
A lo largo de este documento se desarrollará un proyecto de ciencia de datos enfocado a dar solución al proceso de generación de portafolios de productos, una predicción de la demanda para los diferentes puntos de venta de la empresa en estudio. El proyecto se estructurará tomando como base la metodología CRISP-DM y la aplicación de diversas herramientas de ciencias de datos, para ello se seguirá el calendario de actividades plasmadas en el siguiente diagrama de Gantt: Anexo. Diagrama de Gantt: https://docs.google.com/spreadsheets/d/1N83SSjAv8x7GcwoVv-xdbqd3mV9uykCN/edit?usp=sharing&ouid=101688482481224037026&rtpof=true&sd=true

2. Etapa 2- Comprensión de los datos.

2.1 Describir los datos crudos.

La empresa nos proporcionó una base de datos en la cual tenemos información de las unidades vendidas desde el 1ro de junio del 2019 hasta el 31 de marzo del 2020.

Tipo de archivo: **Excel.csv**

Dimensión de los datos: **14 variables con 23032 registros en total.**

No.	Variable	Descripción	Tipo de dato	Valor min	Valor max	Nivel	Ejemplo
1	Punto_de_venta	Puntos de venta donde la empresa vende	Character			1538	"1 poniente", "5 de mayo zmm", "acayucan"
2	Fecha	Fecha de la venta del equipo	Character			301	31/12/19
3	Mes	Mes de la venta del equipo	Character			11	"AGOST","FEB, "1", "10"
4	Anio	Año de la venta del equipo	Number	2020	2019	2	"19", "2019", "2020"
5	Num_ventas	Unidades vendidas del equipo	Number	1	1	1	1
6	Sku	Número de serie del equipo vendido	Character			28	"N.HIF24AZ", "N.HIF24NG", "N.HIL675B"
7	Marca	Marca del equipo vendido	Character			5	"Hisense-Hisense", "hisense"
8	Gamma	Categoria del equipo en el mercado	Character			1	"baja"
9	Costo_promedio	Costo promedio de los equipos vendidos	Number	0	2067	22	"1982.981979", "2067.007825"
10	Zona	Zona donde se vendió el equipo	Character			9	"centro occidente","centro sur"
11	Estado	Estado donde se vendió el equipo	Character			35	"aguascalientes","baja california"
12	Ciudad	Ciudad donde se vendió el equipo	Character			210	"zitacuaro", "zumpango"
13	Latitud	Latitud de la ciudad donde se vendió el equipo	Number	14.9	1958275	1515	"21.0302", "21.03042","21.03051"
14	Longitud	Longitud de la ciudad donde se vendió el equipo	Number	-990229	-86.8	1473	"-99.07311","-99.06958","-99.05718"

Fig 2. Diccionario de variables.

2.2 Detectar problemas de calidad.

No.	Variable	Problema de calidad	¿Por qué?
1	Punto_de_venta	Coherencia	En la sucursal de acayucan se encuentra otra sucursal capturada como acayuckan, por lo que se ve incongruencia en la captura del nombre de las sucursales.
2	Fecha		
3	Mes	Coherencia	Se usa diferente notación para la identificación del mes.
4	Anio	Coherencia	Se usa diferente notación para la identificación del año.
5	Num_ventas		
6	Sku	Representación	El número de caracteres en la serie no es congruente.
7	Marca	Unicidad	El nombre de la marca no está estandarizado para todos los registros.
8	Gamma		
9	Costo_promedio	Completitud	Hay valores faltantes.
10	Zona	Coherencia	Se usan mayúsculas para la identificación de la zona del golfo de México.
11	Estado	Precisión	Existen registros de ciudades como estados.
12	Ciudad		
13	Latitud	Representación	El número de decimales no es igual en todos los datos.
14	Longitud	Representación	El número de decimales no es igual en todos los datos.

Fig 3. Tabla de problemas de calidad.

2.3 Actividades a futuro.

Conforme a lo detallado en el diagrama de Gantt previamente anexado, como equipo debemos enfocarnos en la preparación de los datos a través de la limpieza, análisis y selección de variables relevantes para avanzar a las siguientes etapas. Con esto en mente, es importante agendar periódicamente sesiones de trabajo donde estemos presentes todos los integrantes del equipo y colaboremos activamente a garantizar datos confiables. De esta forma, tendremos conocimiento del proceso que lleve a acciones directas que apoyen a las estrategias y acciones competitivas para Hisense.

3. Etapa 3- Preparación de los datos

3.1 Limpieza de datos.

Limpiar los datos es uno de los pasos más relevantes para comenzar con un proyecto de ciencia de datos, se debe efectuar con el propósito de crear una cultura en torno a la toma de decisiones de datos de calidad. En este primer reporte describiremos los pasos que nuestro equipo siguió para realizar el proceso de limpieza de datos.

3.1.1 Comprensión de los datos.

Antes de comenzar a depurar nuestro conjunto de datos, un paso importante es comprender con qué variables estamos trabajando, esto nos ayudará a realizar una mejor limpieza de los mismos. En este caso nuestros datos son el cotejo del número de ventas de celulares de la marca Hisense con la información específica de cada variable.

Una vez comprendidos los datos que se nos brindaron, comenzamos a utilizar R studio para su limpieza. Para ello empleamos algunos comandos y librerías, como requerimientos iniciales, dichos comandos fueron los siguientes: **library(tverse)**: Llamar a la librería tidyverse / **getwd()**: Obtener la dirección donde estamos trabajando en nuestra PC. / **datosE3<- read.csv()**:Leer el archivo CSV, con el conjunto de datos.

3.1.2 Análisis de datos.

Con el fin de profundizar en la comprensión de nuestro datos, se utilizaron tres comandos los cuales nos ayudan a visualizar las dimensiones, y un sumario de cada variable. Para lo anterior los comandos utilizados fueron los siguientes: **dim(datosE3)**: Muestra las dimensiones de nuestra base de datos (Renglones y columnas) / **str(datosE3**): Nos brinda una lista de los tipos de datos con los que estamos trabajando. / **summary(datosE3)**: Nos brinda un resumen de cada variable, valores mínimos, máximos, medias, etc.

3.1.3 Detección de problemas de calidad.

Después de hacer un primer análisis de nuestros datos y de leerlos en la plataforma de R studio, el paso siguiente es detectar problemas de calidad. Para ello utilizamos las métricas universales del data quality, las cuales buscan eliminar problemas de: unicidad, validez, precisión, entre otros. Tomando esto en consideración, los problemas de calidad detectados fueron los siguientes:

- #1. En la variable punto de venta hay 5 puntos de venta escritos de manera errónea, este es un error de calidad de tipo coherencia.
- #2. En la variable mes hay valores mal registrados (en lugar de número, son letras). Cambiar los 5 meses que están registrados con letras, es un problema de tipo coherencia.
- #3. La variable de año no sigue un formato de valor numérico de 4 dígitos, es un error de tipo coherencia.

Entre otros errores que se describieron en el entregable.

3.1.4 Corrección de errores.

Debido a que cada error presentaba requerimientos diferentes, fue necesario emplear herramientas diferentes en cada problema. Algunos de los comandos utilizados para limpiar nuestro datos fueron: **datosE3[_,_]<- "_____"**: Se empleó para reemplazar algún dato de una columna y renglon especificos. / **datosE3$___<- tolower(datosE3$____)**: Se empleó para convertir en minúsculas los nombres de variables que se encontraban en mayúsculas. / **datosE3$mes<- str_replace(datosE3$__, "__", "_")**: Se utilizó para reemplazar todos los datos de un valor o carácter específico de una columna. /**datosE3 %>% select(___) %>% unique()**: Con este comando observamos los valores unicos de cada variable.

3.1.5 Exportar nueva base de datos.

Una vez que realizamos todas las correcciones de nuestra base de datos, el paso siguiente era escribirlos en un nuevo archivo CSV, con el fin de tenerlos actualizados con respecto a la base de datos inicial. Para este último paso se empleó el comando: **write.csv(datosE3, file= "datoslimpioshisense.csv")**: Con este comando exportamos nuestra nueva y actualizada base de datos.

3.2 Análisis exploratorio.

Posterior a la limpieza de los datos, el análisis de los datos para observar qué es lo que los datos representan es indispensable, para ello se recurrió al uso de herramientas gráficas, para que podamos visualizar la relación entre ellos y al mismo tiempo confirmar que no existan inconsistencias. Se plantearon las siguientes preguntas y se usaron los códigos a continuación:

1) ¿Cómo están las ventas distribuidas por los estados de la república?

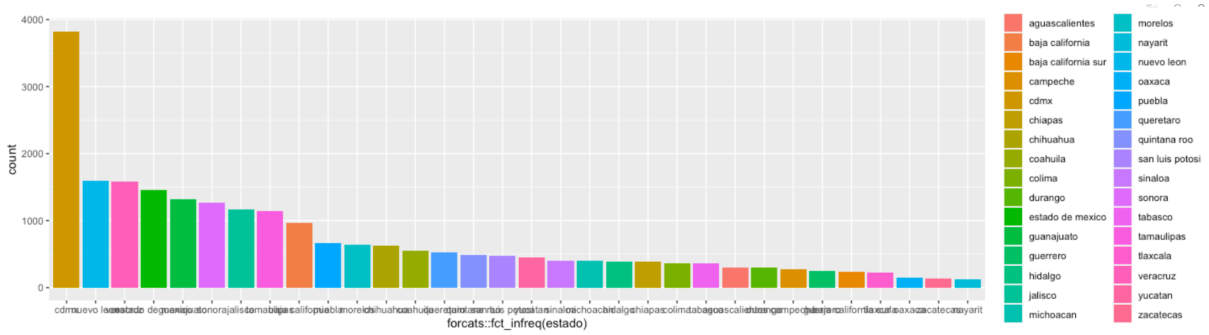

Fig 4. Gráfica de las ventas por estado..

Se observa que se tienen ventas en las 32 entidades del país, sin embargo estás son en diferente cantidad para cada una.

2) ¿Cuáles son los estados con mayores ventas a 1500?

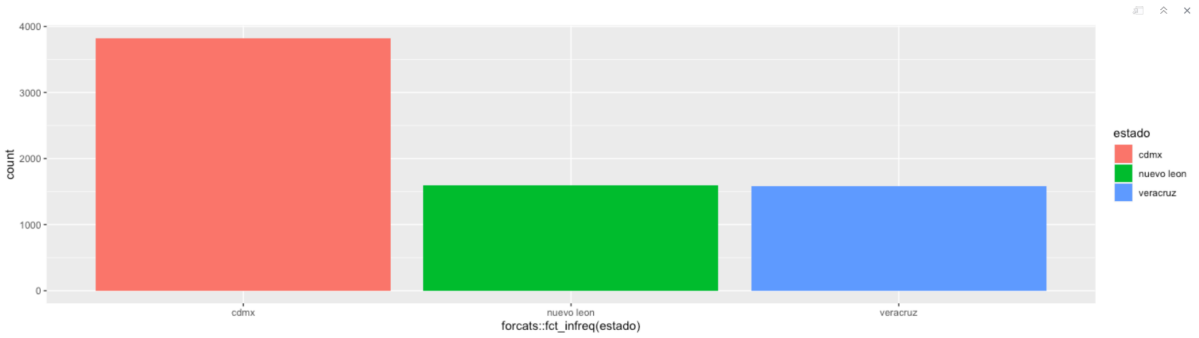

Fig 5. Gráficas de estados con mayor venta.

Los estados con mayores ventas son: Cdmx, Veracruz y Nuevo León considerando ventas arriba de 1500 u.

3) Divide las ciudades del estado que más ventas tiene.

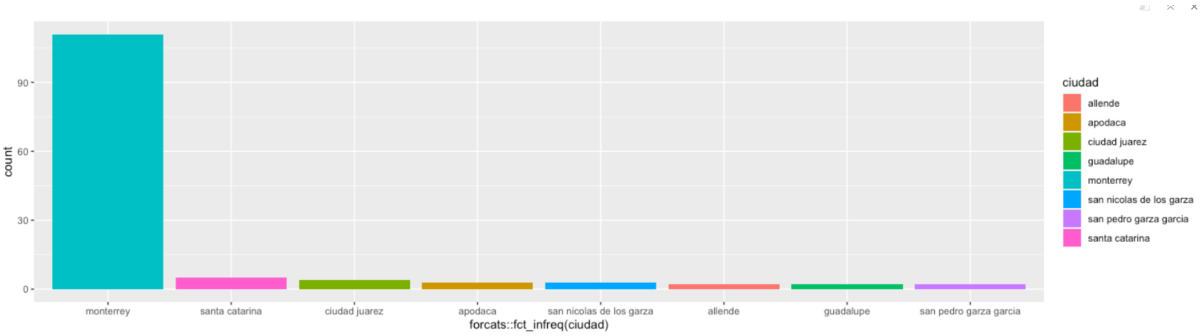

Fig 6. Figura de ciudades con ventas elevadas.

Las ciudades con mayores ventas se enlistan a continuación, 8 ciudades pertenecientes al estado de Nuevo León, segundo estado con mayores ventas.

3.3 Ingeniería de características.

La ingeniería de características es el proceso por el cual a partir de nuestro conjunto de datos limpios, podemos generar nuevas variables que impactan o nos ayuden a predecir de una mejor manera lo que buscamos en nuestro proyecto.

3.3.1 Leer los datos.

Como paso inicial para esta etapa de ingeniería de características, volvemos a leer nuestros datos en R studio, para ello utilizamos el comando **datos <- read.csv("___")**, para corroborar sus dimensiones y un pequeño resumen de los mismos, utilizamos los comandos **dim(datos), str(datos) y summary(datos).**

Como paso adicional en esta fase inicial, asignamos algunas de nuestras variables al tipo de variable específica, que haga más fácil su manejo; ejemplo de ello es convertir las variables de carácter a numéricas si es necesario. Para esta etapa el comando empleado fue el siguiente:

- datos$VARIABLE <- as.NUEVA VARIABLE(datos$VARIABLE)

3.3.2 Índices para variables consideradas.

Para un mejor manejo, convertimos nuestras variables cualitativas en índices. Este proceso se realizó ordenando por orden alfabético algunas variables para después, asignar una columna extra de ID con una secuencia de número enteros, todo esto para las variables consideradas.

3.3.3 Agrupar conjunto de datos.

Como segundo paso, se realizará la agrupación de datos de nuestro nuevo data frame con nuestros datos originales basándonos en la columna de ID. Este proceso se realizó con la función left join, para todas las variables consideradas en los puntos anteriores. Todo lo anterior se describe en el siguiente código: **datos <- left_join(datos, VARIABLE_id, by="VARIABLE") head(datos)**

3.3.4 Agrupas ventas totales.

Con el fin de observar las ventas adicionales que hay en los puntos de venta, en la misma fecha, se realizó la agrupación de ventas totales. Para realizar esto se tomó en cuenta la sugerencia de quitar la información adicional implícita en el punto de venta. Lo anterior se realizó con el comando: **datos <- datos %>% #quitamos fecha porque vamos a hacer el análisis por mes group_by(pdv_id, sku_id, mes_id)%>% summarise(ventas_totales = sum(num_ventas))**

3.3.5 Completar serie de tiempo.

Con el paso anterior podemos observar que nuestra serie de tiempo está incompleta, pues hay periodos en donde no hay ventas registradas. Para resolver esto se construyeron 3 conjuntos nuevos de índices con las variables designadas. Posteriormente se realizó la combinación de estos conjuntos de datos empleando la función merge ().

Una vez realizada la combinación, se empleó nuevamente la función left join () para obtener las ventas totales en cada punto de venta. Con esta combinación se puede ver los meses en donde no hubo ventas de productos, para los cuales se colocó un índice 0.

3.3.6 Construcción de variable de respuesta (Y).

Para nuestro modelo de pronósticos de ventas, y considerando la información histórica que tenemos, se decidió emplear un código, el cual desplaza la información una columna creando una nueva variable (Y), esto nos permitirá saber las ventas para el siguiente mes.

3.3.7 Crear nuevas características.

Este paso nos permite crear nuevas variables que agreguen valor a nuestro análisis. Para nuestro proyecto se realizaron agrupaciones y conteos que permitan realizar las predicciones de una mejor manera. En esta etapa creamos las características de ventas promedio por mes, tienda y producto y ventas totales con las cuáles se crean las características que necesitamos de manera rezagada más adelante para nuestro modelo de predicción.

Realizado el paso anterior, se agruparon con nuestra base de datos inicial, en complemento con las fases anteriores.

3.3.8 Rezagos.

Como último paso con los siguientes comandos: **library(zoo) datos_completos<-na.locf(datos_completos, fromLast = TRUE) head(datos_completos)** se realizaron los valores faltantes NA con 0, con el fin de tener una base de datos completa. Para finalizar este paso, se escribió nuestra nueva base de datos en un nuevo archivo de CSV, con el siguiente código: **write.csv(datos_completos, file="datos_completosE3.csv", row.names = FALSE)**

3.3.9 Conclusiones de ingeniería de características.

Después de realizar el proceso con ingeniería de características, podemos observar que el cambio de dimensiones de nuestro nuevo archivo son: 386316 renglones con 24 columnas. En cuanto al archivo trabajado anteriormente antes del proceso de ingeniería de características donde las dimensiones eran de 23032 renglones con 14 columnas.

4. Etapa 4- Modelado.

4.1 Promedios móviles.

El promedio móvil es un indicador de tendencias que se usan para realizar análisis de datos anteriores con la finalidad de formar una serie de medidas que provengan de diversos subconjuntos de datos de precios, por lo tanto, tienen la capacidad de examinar las medidas de precios que disminuyen en un período de tiempo.

Los promedios móviles se deben calcular después de las observaciones consecutivas de los subgrupos artificiales. Estos se pueden utilizar en las gráficas de control para crear gráficas de promedios para los datos en determinados tiempos programados. Cuando se realizan los análisis de series de tiempo, se usa el promedio móvil y de esa forma se pueden suavizar los datos y disminuir las dudas aleatorias en una determinada serie de tiempo.

4.1.1 Tipos de promedios móviles.

4.1.1.1 Promedio móvil simple.

El modelo de promedio móvil funciona mejor con datos horizontales (datos sin tendencia). Un promedio móvil se obtiene encontrando la media de un conjunto específico de valores y aplicándolo después para pronosticar el siguiente periodo.

4.1.1.2 Promedio móvil ponderado.

En general se difiere que los diversos puntos de datos se pueden ponderar o asignar a un punto concreto de gran importancia. La media móvil ponderada tiene la capacidad de

agregarle importancia a los puntos de datos que estén más recientes. Dentro del período estipulado, a cada uno de los puntos se le asignará un multiplicador de datos reciente para que luego vaya descendiendo ordenadamente. Después cuando se le añade al principio un nuevo punto, se eliminará el punto de datos que contenga mayor antigüedad.

4.1.2 Métricas: MAE,RMSE,MSE.

Las métricas nos ayudan a medir el desempeño del modelo, además de ayudarnos a contrastar el Modelo actual vs el Modelo propuesto. Mediante las métricas podemos obtener resultados interpretables y medibles.

- MAE: Error absoluto medio
- MSE: Error cuadrado medio
- RMSE: Error medio

4.2 Construcción del modelo de promedios móviles.

A continuación se presenta de manera breve la construcción del modelo de promedios móviles utilizado para la base de datos de la compañía Hisense, empresa bajo estudio durante este proyecto.

4.2.1 Lectura de datos.

El código del modelo se realizó en Jupyter, descargando las librerías correspondientes y cargando el archivo con nuestros datos.

4.2.2 Descartar columnas de datos.

A través del comando *datos.drop (),* se descartaron algunas columnas de nuestros datos, dejando la columna ***ventas_totales*** y ***y_ventas_siguiente_mes*** como nuestras variables x y y respectivamente para realizar el modelo de promedios móviles.

```
In [10]: datos_E3.head(10)
```

Out[10]:

	pdv_id	mes_id	sku_id	ventas_totales	y_ventas_siguiente_mes
0	1	0	1	0	1
1	1	1	1	1	1
2	1	2	1	1	1
3	1	3	1	1	0
4	1	4	1	0	0
5	1	5	1	0	0
6	1	6	1	0	0
7	1	7	1	0	0
8	1	8	1	0	0
9	1	0	2	0	0

Fig 7. Datos con columnas descartadas

4.2.3 Pedir datos.

Se procede a construir una columna que muestre lo vendido el mes anterior como base para la predicción de ventas posterior. Se calcula otra columna que muestre el promedio de ventas de los dos meses anteriores. De esta forma cambiamos dos veces el periodo móvil.

	pdv_id	mes_id	sku_id	ventas_totales	y_ventas_siguiente_mes	m1_pedir_lo_del_mes_pasado	m2_promedio_de_dos_meses_anteriores
0	1	0	1	0	1	0	NaN
1	1	1	1	1	1	1	0.5
2	1	2	1	1	1	1	1.0
3	1	3	1	1	0	1	1.0
4	1	4	1	0	0	0	0.5
5	1	5	1	0	0	0	0.0
6	1	6	1	0	0	0	0.0
7	1	7	1	0	0	0	0.0
8	1	8	1	0	0	0	0.0
9	1	0	2	0	0	0	NaN
10	1	1	2	0	2	0	0.0
11	1	2	2	2	0	2	1.0
12	1	3	2	0	0	0	1.0
13	1	4	2	0	0	0	0.0
14	1	5	2	0	0	0	0.0
15	1	6	2	0	0	0	0.0

Fig 8. Datos con 2 cambios de periodo móvil

4.2.4 Cálculo de MAE.

Se realiza el cálculo del indicador MAE respecto al error del promedio calculado con los datos reales de las ventas por mes. Se realiza el cálculo manual mostrando el nombre del mes y el MAE tanto para el promedio de 2 meses anteriores como para 3. Finalmente, se muestra de manera gráfica el error absoluto total en el modelo.

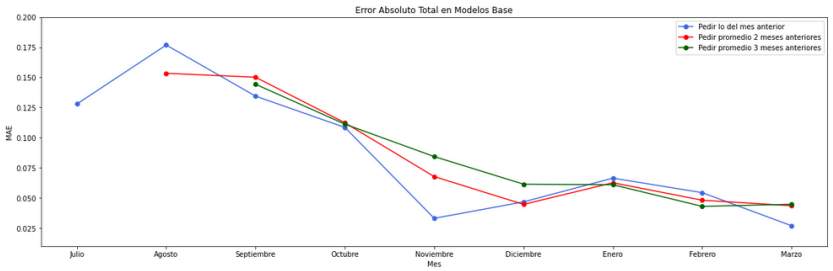

Fig 9. Gráfica del Error Absoluto Total en el modelo.

5. Etapa 5- Evaluación (Resultados).

Llegando al final de nuestro proyecto, es momento de evaluar los resultados que nuestros modelos han arrojado. Para nuestro proyecto, decidimos evaluar el resultado de 6 modelos diferentes, tres de ellos de promedios móviles considerando 1,2,3 meses y los otros tres del

modelo de árbol de decisiones, considerando 1,2,5 de profundidad. Todos ellos fueron corridos en Jupyter Notebook, con códigos de Python.

Aunado a esto, se programaron los códigos necesarios para que cada modelo nos mostrará tres métricas importantes, las cuales son: mean absolute error (MAE), mean square error (MSE) y root mean square error (RMSE). Mismos que graficamos y utilizaremos como punto de referencia para evaluar el comportamiento de los modelos. Todo esto se realizará para las métricas obtenidas en las iteraciones del conjunto de prueba y entrenamiento.

Los resultados de este proceso, se muestran a continuación:

5.1 Evaluación de modelos de entrenamiento.

Como podemos observar en las gráficas, no encontramos la presencia de las métricas de promedios móviles, esto es por que este modelo no cuenta con resultados del conjunto de prueba; por ello, el único modelo que se muestra en la gráfica es el de árbol de decisiones.

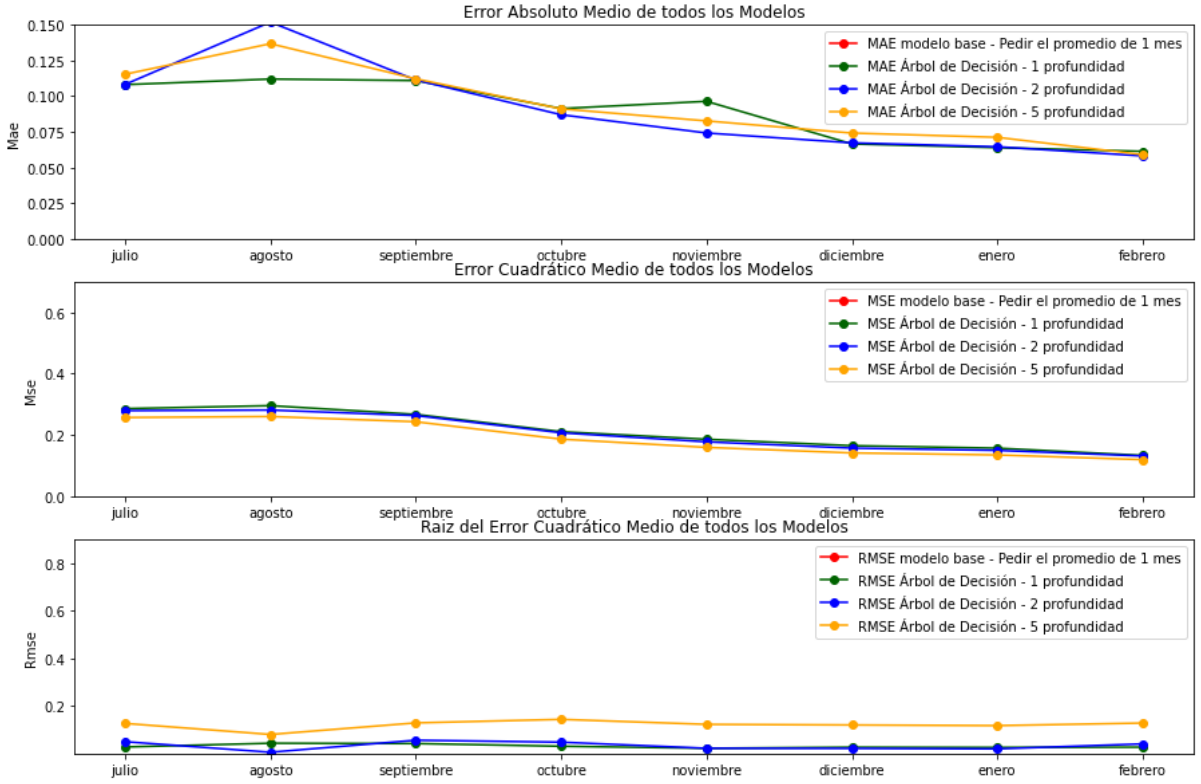

Fig 10. Gráficas de Métricas MAE, MSE, RMSE de árbol de decisiones.

Como podemos observar en las gráficas de la figura 7 el modelo que tiene los valores de MAE, MSE y RMSE más bajos es el de: árbol de decisión con profundidad 2. Sabemos que esta decisión es preliminar, pues aún tenemos que evaluar los resultados del conjunto de pruebas.

5.1 Evaluación de modelos de prueba.

Como podemos observar en la figura 11, tenemos la presencia de las gráficas de los 6 modelos. A simple vista podemos ver que las gráficas roja y morada, tiene los desempeños más aceptables, sin embargo, debido a la precisión y a los valores constantes que muestra la gráfica morada, se seleccionó como el mejor modelo. Por lo tanto el mejor modelo del conjunto de prueba es el modelo de: promedios móviles con 3 meses.

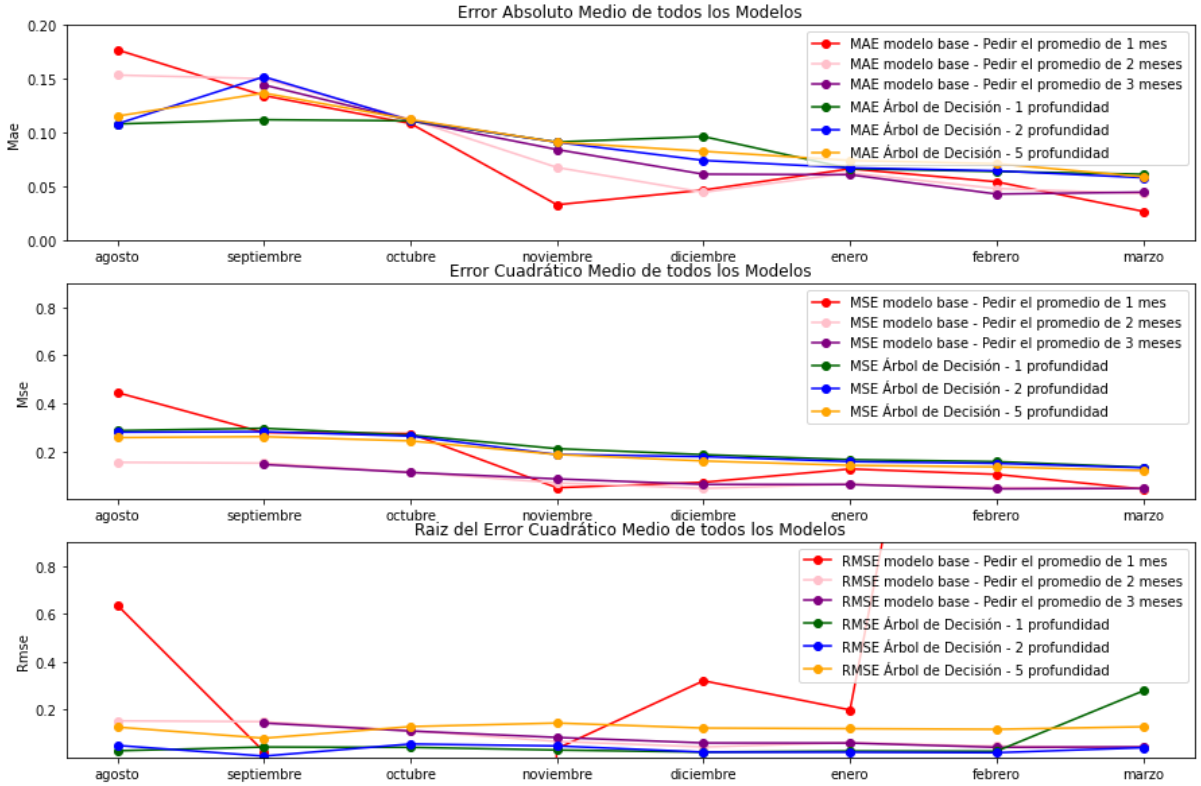

Fig 11. Gráficas de Métricas MAE, MSE, RMSE de promedios móviles y árbol de decisión.

6. Conclusiones.

Con el desarrollo de este proyecto logramos abordar un caso sumamente práctico con herramientas de ciencia de datos. Siguiendo la metodología CRISP-DM, para la ejecución correcta de nuestro modelo, desde entender el problema y el negocio, hasta algo más complejo, que es modelar los programas en R y Python. Nos gustaría resaltar que aunque algunos de nosotros ya teníamos conocimientos en estadistica, matematicas, machine learning y storytelling, nunca los habíamos aplicado en un proyecto como esto, por lo que el aprendizaje no solo de las herramientas, si no de la metodología y los conceptos, fueron bastante gratificantes para todos los miembros del equipo.

7. Referencias Bibliográficas.

[1] (n.d.). Sobre Hisense. Se recuperó el septiembre 01, 2021 de https://www.hisense.es/sobre-hisense/

[2] Cristina O. (2021, julio 17). Hisense ve en México una prioridad en ventas y producción. Periódico digital Milenio, Se recuperó el septiembre 2, 2021 de: https://www.milenio.com/negocios/hisense-ve-mexico-prioridad-ventas-produccion

[3] Statista Research Department (2021). *Marcas de smartphones con mayor cuota de mercado en México 2021*. Disponible en: https://es.statista.com/estadisticas/1077738/participacion-mercado-marcas-smartphones-mexico/

[4] Steve, O. (2020). *Con todo y COVID, los mexicanos nunca habían comprado tantos smartphones de entre 16,000 y 20,000 pesos como en 2020.* Xataka México. Disponible en: https://www.xataka.com.mx/celulares-y-smartphones/todo-covid-mexicanos-nunca-habian-comprado-smartphones-16-000-20-000-pesos-como-2020

[5] (2020, noviembre 13). Data Cleansing. Todo lo que debes saber sobre la 'limpieza de datos'. Se recuperó el octubre 12, 2021 de https://bigdatamagazine.es/data-cleansing-todo-lo-que-debes-saber-sobre-la-limpieza-de-datos

[6] (n.d.). Calidad de Datos. Cómo impulsar tu negocio con los datos.. Se recuperó el octubre 14, 2021 de https://www.powerdata.es/calidad-de-datos

[7] Minitab.(2019). ¿Qué es un promedio móvil?Disponible en: 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/#:~:text=Los%20promedios%20m%C3%B3viles%20son%20promedios%20calculados%20a%20partir,suavizar%20los%20datos%20y%20reducir%20las%20fluctuaciones%20

[8] Pacheco, J. (2021). *¿Qué es el Promedio Móvil?*. Web y Empresas. Disponible en: https://www.webyempresas.com/promedio-movil/

[9] Avilés, E.G. (2021), *Ingeniería Estadística*, [presentación], Disponible en: Mis clases: Ingeniería estadística (Gpo 1) (tec.mx)