Laboratorio de diseño y optimización de operaciones

Proyecto Semestral

Maestra Ana Luisa Masetto Herrera

Equipo #1

Silvana Armas Sámano A01364792
Guillermo Candelario Rivera Contreras A01365068
Alejandra Guadalupe Salinas Marín A01362661

Noviembre 20, 2020.

Índice

Introducción

En este trabajo se presenta el desarrollo de nuestro proyecto de ciencia de datos. Este consistió en un proceso de creación de modelos de aprendizaje de máquina, con el objetivo de comparar con Promedios Móviles y así pronosticar la demanda de cada uno de los diferentes modelos de celulares que la marca Huawei ofrece en los diferentes puntos de venta alrededor de nuestro país. Tomamos como base la metodología CRISP para realizar su estructuración, haciendo un elaborado análisis en cada una de las etapas.

I. Etapa 1: Comprensión del Negocio

1. *Descripción de la situación actual*

Huawei es un proveedor de infraestructura de tecnologías y comunicaciones y dispositivos inteligentes fundada en 1987. Es una empresa comprometida con llevar la tecnología digital a personas, hogares y organizaciones para lograr un mundo inteligente y totalmente conectado. Creen firmemente que el beneficio de la tecnología digital no debería de estar disponible únicamente para ciertas personas que se lo pueden permitir. Su portafolio punta-a-punta de productos, soluciones y servicios es competitivo y seguro.

La firma llegó a México en 2002 en el sector de las telecomunicaciones, en 2012 incursionó en el área de los smartphones y hasta 2018 la empresa habría vendido 9.39 millones de dispositivos [b]. En la primera mitad de 2019, la marca alcanzó una participación de mercado de 12.1% con presencia en todas las gamas (alta, media y baja) [c]. Entre 2018 y 2019, se vendieron alrededor de 369,617 celulares en nuestro país.

De acuerdo con la página oficial de Huawei México, actualmente la empresa ofrece 24 modelos diferentes de teléfonos en el país [1] y posee diversas formas de llevar sus productos al consumidor. Huawei cuenta con varios puntos de venta de accesorios [2], sitios autorizados de venta en línea, además de disponibilidad de productos en tiendas departamentales o de electrónica de otras empresas [3].

2. *Problemática en términos de negocio*

Huawei tiene la ambiciosa filosofía de ampliar los beneficios de la tecnología digital a todas las personas, en todas las partes del mundo. Sin embargo, esta ideología en términos de negocio resulta bastante compleja de llevar a cabo. Alcanzar lo que la empresa se plantea, conlleva tener un preciso control de todos los puntos de venta y un perfecto abastecimiento de los productos demandados. Es crucial para Huawei contar con personas capacitadas para hacer un profundo análisis de las ventas. Como ingenieros industriales es una gran ventaja conocer cómo tomar los datos, limpiarlos e utilizarlos, en este caso para poder pronosticar las demandas futuras de cada tipo de smartphone en cada establecimiento y así poder siempre ofrecer al cliente lo que necesita.

3. *Problemática en términos de ciencia de datos*

Todas las tiendas que comercializan los productos de Huawei generan registros sobre las ventas realizadas. Al considerar el número de sucursales alrededor del país y toda la información que producen, estamos hablando de una cantidad abrumadora de datos. Además, los encargados de tomar esos registros muchas veces no toman en cuenta el formato en el cual se deben de escribir. Esto definitivamente afecta mucho tanto en la calidad de los datos, como en la facilidad del análisis de los mismos. Por eso, la ciencia de datos juega un rol muy importante en la depuración de la información obtenida y en su estratégica utilización a favor de la marca.
En este proyecto estamos lidiando con datos estructurados, ya que tienen un formato adecuado para poder utilizarlos. Además de que tratamos con una tarea de Regresión, porque nuestro objetivo es realizar la predicción de valores numéricos. Necesitamos conocer a profundidad la naturaleza de los datos con los que estamos trabajando. Debemos estudiarlos, graficarlos, entender cómo se están comportando para poder elegir un método adecuado para el pronóstico de demanda futura. Esperamos que con nuestros conocimientos de Ingeniería Industrial y de Ciencia de Datos, podamos contestar la pregunta: *¿Cuántas unidades de cada uno de los productos de Huawei se van a vender en cada punto de venta, en el siguiente mes de registro?*

4. *Objetivos*

De acuerdo con la página oficial de la empresa, el objetivo general es crear la mejor experiencia para el usuario, en el que se pueda ofrecer canales un poco más extensos, inteligentes y confiables con mejor rendimiento y sin tiempo de espera.

Introducción a los objetivos del proyecto:

- Comprender e investigar la situación en la que se encuentra la empresa, la problemática y los objetivos de la misma para así poder estructurar de forma adecuada el plan preliminar del proyecto.
- Comprender y detectar los problemas de calidad con respecto a los datos proporcionados .
- Construir modelos que permitan obtener la solución del problema, utilizando promedios móviles y aprendizaje de máquina.
- Evaluar los modelos y concluir con respecto a su desempeño.
- Presentar los resultados al equipo docente y alumnado.

5. *Plan preliminar*

Actividades	Fecha de inicio	Duracion(dias)	Fecha Fin	Porcentaje completados	Días completados
Descripcion de situacion actual	31/08/2020	7	07/09/2020	100.00%	7.00
Descripcion del problema (negocio y ciencias de datos)	31/08/2020	7	07/09/2020	100.00%	7.00
Plasmar objetivos y estructuracion del proyecto	31/08/2020	7	07/09/2020	100.00%	7.00
Descripcion de datos crudos	31/08/2020	7	07/09/2020	100.00%	7.00
Detecion de problemas de calidad	31/08/2020	7	07/09/2020	100.00%	7.00
Limpieza de datos	07/09/2020	7	14/09/2020	0.00%	0.00
Realizacion de analisis exploratorio	14/09/2020	7	21/09/2020	0.00%	0.00
Construccion de variables	21/09/2020	8	29/09/2020	0.00%	0.00
Codificacion de promedios moviles	29/09/2020	10	09/10/2020	0.00%	0.00
Codificacion de modelo de aprendizaje de maquina	09/10/2020	10	19/10/2020	0.00%	0.00
Construccion de graficas para evaluar desempeño de modelos	19/10/2020	10	29/10/2020	0.00%	0.00
Construccion de conclusiones	29/10/2020	8	06/11/2020	0.00%	0.00
Creacion de reporte final	06/11/2020	5	11/11/2020	0.00%	0.00
Creacion de presentacion final	11/11/2020	5	16/11/2020	0.00%	0.00

Figura 1. Tabla de relación de actividades, tiempos y porcentajes.

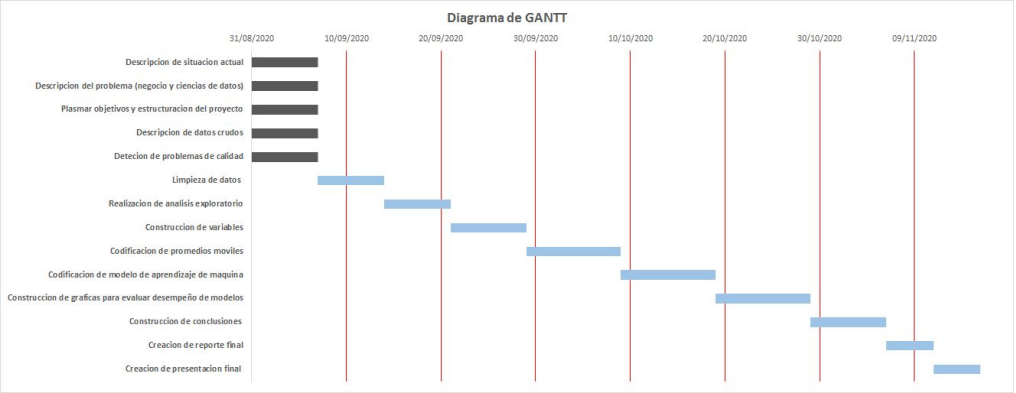


Figura 2. Diagrama de Gantt de acuerdo a los datos de la tabla anterior.

Actividades	Fecha de inicio	Duracion(dias)	Fecha Fin	Porcentaje completados	Días completados
Descripcion de situacion actual	31/08/2020	7	07/09/2020	100.00%	7.00
Descripcion del problema (negocio y ciencias de datos)	31/08/2020	7	07/09/2020	100.00%	7.00
Plasmar objetivos y estructuracion del proyecto	31/08/2020	7	07/09/2020	100.00%	7.00
Descripcion de datos crudos	31/08/2020	7	07/09/2020	100.00%	7.00
Detecion de problemas de	31/08/2020	7	07/09/2020	100.00%	7.00
Limpieza de datos	07/09/2020	7	14/09/2020	100.00%	7.00
Realizacion de analisis exploratorio	14/09/2020	7	21/09/2020	100.00%	7.00
Construccion de variables	21/09/2020	8	29/09/2020	100.00%	8.00
Codificacion de promedios moviles	29/09/2020	10	09/10/2020	100.00%	10.00
Codificacion de modelo de aprendizaje de maquina	09/10/2020	10	19/10/2020	100.00%	10.00
Construccion de graficas para evaluar desempeño de modelos	19/10/2020	10	29/10/2020	100.00%	10.00
Construccion de conclusiones	29/10/2020	8	06/11/2020	100.00%	8.00
Creacion de reporte final	06/11/2020	5	11/11/2020	100.00%	5.00
Creacion de presentacion final	11/11/2020	5	16/11/2020	100.00%	5.00

Figura 3. Tabla de relación de actividades, tiempos y porcentajes, exponiendo todas las actividades hechas al 100%.

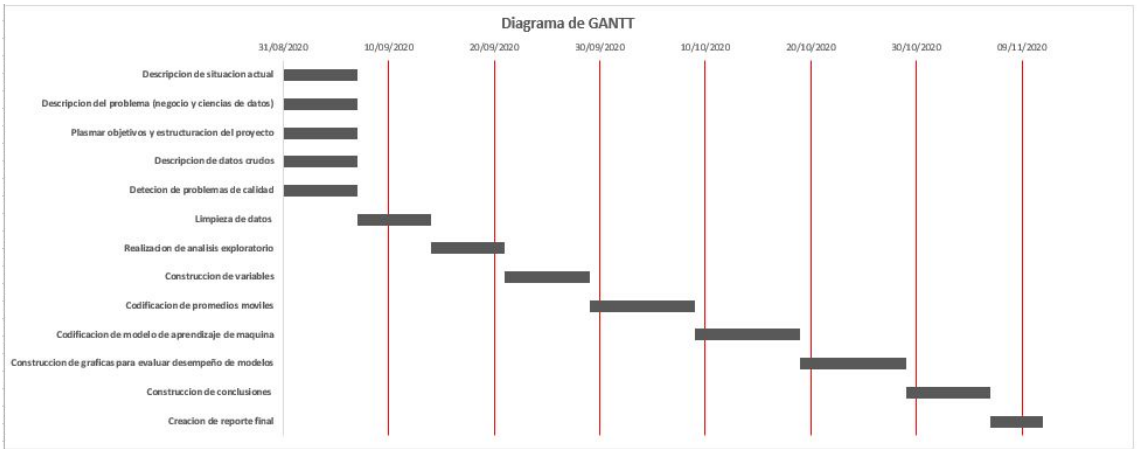


Figura 4. Diagrama de Gantt de acuerdo a los datos de la tabla anterior, exponiendo todas las actividades hechas al 100%.

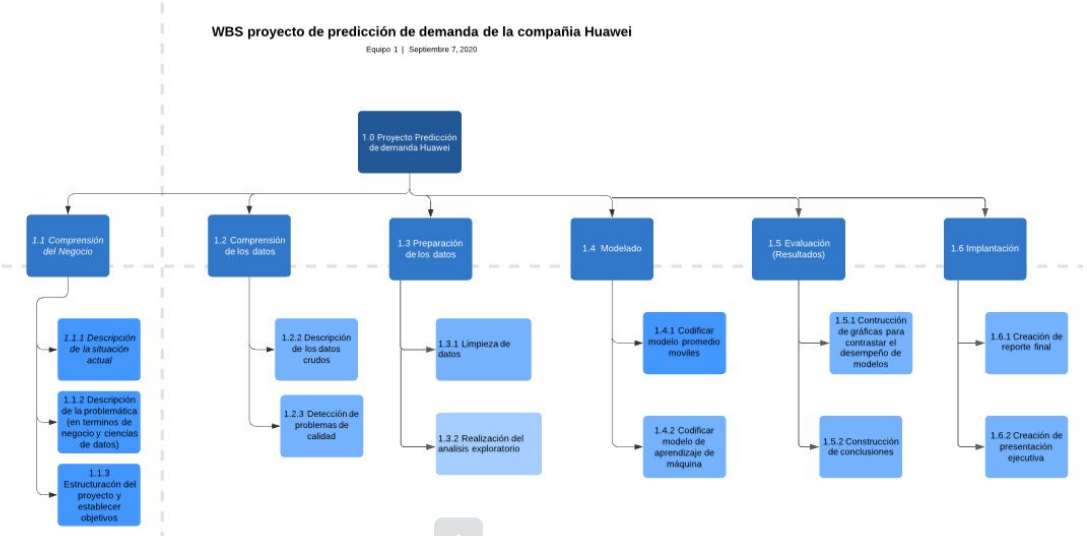


Figura 5. WBS del proyecto

II. Etapa 2: Comprensión de los Datos

1. *Describir los datos crudos y problemas de calidad*

Para la realización de este proyecto contamos con una base de datos de 14 variables y 369,617 observaciones, con esta cantidad de datos se necesitaría una gran capacidad de procesamiento en programas de análisis de datos convencionales como Microsoft Excel, es por esto que utilizamos R Studio, un programa y lenguaje especializado en el manejo y análisis de datos con un enfoque estadístico. Durante esta etapa, es crucial identificar cualquier problema de calidad que exista en los datos, ya que pueden generar conflictos en etapas posteriores. Un pequeño cambio en esta etapa puede provocar grandes complicaciones en la etapa de análisis y profundas implicaciones al momento de la resolución del modelo, un ejemplo sería un incorrecto pronóstico.

Como se mencionó en el párrafo anterior cada una de las observaciones cuenta con 14 variables que son:

1. **Punto de venta**- la tienda o locación en la que se realizó la venta, esta variable es de tipo factor y se cuenta con 1904 niveles o dicho en otras palabras 1904 tiendas donde se vendieron estos equipos. Existen 5 puntos de venta incorrectos y hay que corregirlos.
2. **Fecha**- La fecha de compra compuesta por día mes y año, esta variable es de tipo factor y se cuenta con 301 niveles o sea 301 días durante los años 2018 y 2019, todos los registros están limpios, siguen el mismo formato. No requiere correcciones.
3. **Mes**- Mes de la compra, Esta variable es numérica, hay valores mal registrados (en lugar de números, son letras). Será necesario cambiar los 5 meses que están registrados con letras.
4. **Año**- Año de la compra, esta variable debe de seguir el formato de un valor numérico de 4 dígitos, existen registros que no están bien hechos
5. **Num_ventas** - Cuántas unidades se vendieron en esa transacción, todos los registros están limpios. No requiere correcciones.
6. **SKU**- *Stock Keeping Unit* son uno de los elementos fundamentales para llevar el control y gestionar el stock en el almacén. SKU es el número de referencia único de un producto,

según aparece registrado en el sistema de la empresa. Todos los registros están limpios. No requiere correcciones.

7. **Marca**- Esta variable es la marca de celular, en realidad todos los productos de nuestro registro son marca Huawei, Hay 5 marcas que están escritas de forma errónea, hay que corregirlas.
8. **Gama**- Esta variable define la gama en la que se encuentra el celular.Todos los registros están limpios. No requiere correcciones.
9. **Costo_promedio**- Indica el precio por el cual el cliente obtuvo el producto. Todos los registros están limpios. No requiere correcciones.
10. **Zona**- Es la zona geográfica del país en la que se realizó la venta, hay 1 zona que está mal escrita, hay que corregirla.
11. **Estado**- Es el estado de la república mexicana en la que se realizó la venta del equipo celular. Hay 3 estados más de los que en realidad existen, detectarlos y corregirlos.
12. **Ciudad**- Ciudad de la república mexicana en la que se realizó la venta del equipo celular. Todos los registros están limpios. No requiere correcciones.
13. **Latitud**- Latitud de la coordenada geográfica donde se realizó la venta del equipo. Hay 1 valor fuera de rango. Corregirlo.
14. **Longitud**- Longitud de la coordenada geográfica donde se realizó la venta del equipo. Hay 1 valor fuera de rango. Corregirlo.

III. Etapa 3: Preparación de los Datos

1. Limpieza de Datos

Sabemos que una de las partes más importantes en los proyectos de ciencia de datos es la limpieza de ellos, pues de este paso depende la calidad del análisis y los resultados que se podrán tener.
Para nuestro conjunto de datos tuvimos que hacer modificaciones importantes en los siguientes conjuntos:
***Puntos de venta*:** se identificaron cinco puntos escritos incorrectamente.
Mes: tuvimos grandes problemas pues además de que había cinco meses que estaban registrados con letra en lugar de número, habían 3 errores entre la fecha y el mes, pues existían tres datos que tenían registro del mes de marzo (3) del 2018 pero en la fecha tenían el mes de junio (6), sin duda este paso fue el más retador de toda la limpieza de datos pues tuvimos que identificar en qué posición estaban estos tres datos y modificar el mes manualmente para que coincidieran con la fecha.
Año: tuvimos que cambiar algunos valores para seguir el formato de valor numérico de cuatro dígitos.
Marca: tuvimos que hacer cambios pues existían registros que estaban escritos de manera errónea y a pesar de que todos los datos pertenecen a una misma marca había pequeños errores en la manera de escribirlos.
Zona: tuvimos problema con un registro pues estaba mal escrito y a pesar de que únicamente se trataba de un espacio, ese espacio hacía que 12 registros se les considerara en una zona diferente a pesar de ser la misma.
Estado: encontramos tres datos que no existían, pues el registro se hizo confundiendo la ciudad, un ejemplo de esto es que se registró Mérida en vez de Yucatán.

Latitud y longitud: tuvimos valores fuera de rango a pesar de que la escala de latitudes es fija pues todos los registros se hicieron dentro del territorio de México. Se encontraron errores en la colocación de puntos, un ejemplo es que un registro en lugar de 19.41 estaba escrito como 1941.

El proceso de limpieza de datos es vital para poder hacer un buen análisis en pasos siguientes, un ejemplo claro de esto fue que no nos percatamos del error mencionado entre mes y fechas, y al momento de hacer las gráficas en el análisis exploratorio de los datos nos dimos cuenta que existían tres registros en el mes 3 del 2018 a pesar de que todos deberían de empezar en el mes de julio, es por eso que nos tuvimos que regresar a hacer las modificaciones pertinentes en la limpieza de datos para no arrastrar errores en las siguientes etapas del proceso.

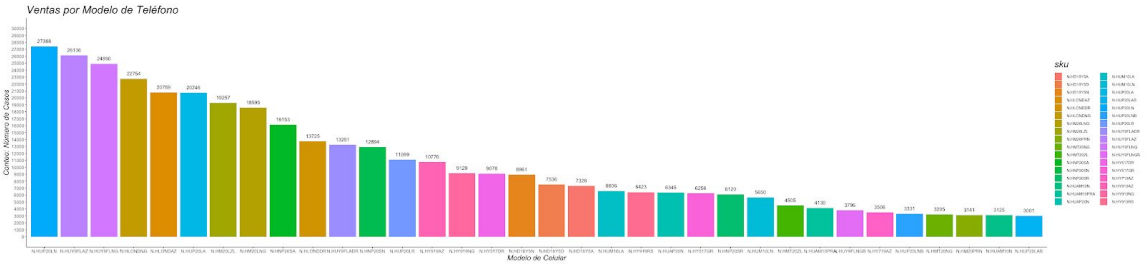
2. *Análisis Exploratorio*

Dentro de la etapa de Análisis Exploratorio se realizó un más profundo estudio y entendimiento de los datos que, en la etapa anterior, ya habían sido depurados. Básicamente, consistió en plantear preguntas que nos llevaran a obtener información relevante sobre nuestros valores; contestándolas con gráficas de diversos tipos. Todo esto ayudando a darle un significado visual a lo que antes sólo representaban cifras. A continuación exponemos las preguntas principales que fueron respondidas con el Análisis Exploratorio de los datos.

¿Cuáles fueron las ventas por modelo de celular?

En esta pregunta utilizamos una gráfica de barras para poder observar mejor cada modelo con su respectivo número de ventas.

Este tipo de información es importante ya que, al saber el modelo más demandado en nuestro país, Huawei puede asegurarse de siempre tenerlo en existencia.



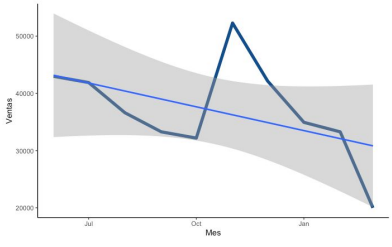
A continuación incluimos una tabla con los 5 modelos con mayores ventas así como los 5 modelos con menores ventas.

5 modelos con más ventas	
N.HUP20LN	27388
N.HUY9FLAZ	26136
N.HUY9FLNG	24890
N.HLONDNG	22754
N.HLONDAZ	20789

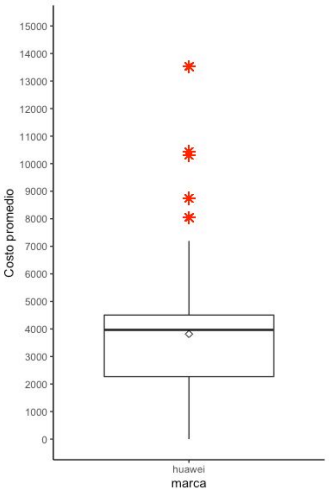
5 modelos con menos ventas	
N.HUP20LNB	3331
N.HMT20NG	3205
N.HM20PRN	3141
N.HUAM10N	3125
N.HUP20LAB	3001

¿Cuales fueron las ventas por mes?

Utilizamos diagramas de dispersión para analizar cómo se comportan la suma de las ventas por mes, considerando los dos años 2018 y 2019. Existe un claro incremento en los últimos meses del año 2018 y enero del 2019,



después una caída en los siguientes meses del 2019. Esto también podría ayudar a la compañía a conocer en qué temporadas las ventas son más fuertes. Además podemos observar la línea de tendencia con sus intervalos para cada punto de la tendencia, como podemos ver, en el mes de noviembre observamos un repunte de las ventas, puede darse por promociones en el Buen Fin o el inicio de la temporada navideña.



¿Cuál es el costo promedio de los teléfonos celulares?

Utilizamos un diagrama de caja para representar el costo promedio de los celulares de la marca vendidos. Se puede observar que la media no pasa los $5,000 pesos. Esto quiere decir que el segmento de mercado objetivo son personas de clase media y que está muy por debajo de los dispositivos de la competencia al que tenemos, algunos datos atípicos que llegan hasta los $13500 pesos.

3. Ingeniería de Características

Ingeniería de Características se utiliza creando características adicionales para un conjunto de datos, con la finalidad de proporcionar información que ayude a diferenciar mejor sus patrones. Considerando que la base de datos del proyecto, contiene una amplia cantidad de valores, aplicar Ingeniería de Características es una forma de posibilitar su fácil utilización.

Para la primera parte de este proceso se crearon índices, cuyo objetivo era facilitar el manejo de las variables cualitativas. Se asignaron números a cada punto de venta dentro del listado, a cada tipo de producto, así como a los meses para distinguir cada periodo de tiempo de registro. Posteriormente, utilizando la función de *left join,* la cual une al dataset principal con el nuevo dataset con índices tomando como referencia la columna del índice. En este paso, es importante recalcar que una buena limpieza de datos favorece el proceso de combinación de dataframes.

A continuación, se agruparon las cantidades de los registros de número de ventas que anteriormente estaban escritos como valores individuales, generando una nueva variable llamada *Ventas Totales*. También, utilizando la función *merge,* combinamos columnas asignando los datos que correspondieran juntos. Quedando una fila dando información sobre Punto de Venta, Mes, Producto y Ventas Totales. Para este paso, se asignó NA a los puntos donde no existiera algún registro de ventas.

Se generó una Variable Respuesta, ya que en realidad nuestro objetivo era decidir cuántas unidades pedir de un producto al mes siguiente. Después, buscamos generar nuevas características que ayudaran a entender más el comportamiento de los datos con respecto a la demanda y ventas; por ejemplo, el promedio de ventas por producto en cada tienda, ventas totales de hace 3 meses, etc.

Mantuvimos la base de datos creada en clase debido a que es bastante completa, las variables creadas nos son útiles para comprender los datos y organizarlos de una forma mucho más manipulable.

Al final obtuvimos un dataset con 24 columnas o variables y 645,659 filas o registros, a continuación listamos las 24 variables incluidas en el dataset:

$ pdv_id
$ mes_id
$ sku_id
$ ventas_totales
$ y_ventas_siguiente_mes
$ ventas_totales_en_tienda_de_cada_mes
$ ventas_promedio_en_tienda_de_cada_mes
$ ventas_totales_en_tienda_de_cada_sku
$ ventas_promedio_en_tienda_de_cada_sku
$ ventas_totales_1_mes_pasado
$ ventas_totales_2_meses_pasados
$ ventas_totales_3_meses_pasados

$ ventas_totales_tienda_y_mes_del_mes_pasado
$ ventas_totales_tienda_y_mes_2_pasado
$ ventas_totales_tienda_y_mes_3_pasado
$ ventas_promedio_tienda_y_mes_del_mes_pasado
$ ventas_promedio_tienda_y_mes_2_pasado
$ ventas_promedio_tienda_y_mes_3_pasado

$ ventas_totales_tienda_y_sku_del_mes_pasado
$ ventas_totales_tienda_y_sku_2_pasado
$ ventas_totales_tienda_y_sku_3_pasado
$ ventas_promedio_tienda_y_sku_del_mes_pasado
$ ventas_promedio_tienda_y_sku_2_pasado
$ ventas_promedio_tienda_y_sku_3_pasad

IV. Etapa 4: Modelado

1. *Promedios Móviles*

Un pronóstico es una predicción de una situación futura. Basados en datos pasados, ofrecen un vistazo más aterrizado sobre lo más probable que puede ocurrir y por esa razón, se utilizan modelos que extrapolan el patrón de datos anteriores.
Los promedios móviles se utilizan suponiendo que todas las observaciones de serie de tiempo tienen relevancia en cuanto a la estimación del parámetro a pronosticar. Además que es muy útil cuando se tiene información no desagregada.
Se tiene tres tipos de promedios:

-Simple
Se utiliza cuando se tienen datos estacionarios y funciona calculando la media de un cierto número de valores anteriores al periodo al que se desea pronosticar.
-Móviles dobles
Se utiliza generalmente cuando los datos tienen cierto tipo de tendencia.
-Móviles ponderados
En este modelo se le puede asignar pesos a ciertos periodos en el pronóstico según sea requerido.

Características:

- Es indispensable tener ordenado cronológicamente todos los periodos. La información que llegase a hacer falta debe ser suplida.
- El procedimiento se basa en el valor promedio de la variable calculada durante un número específico de periodos pasados. Un promedio móvil nos proporciona información sobre la tendencia.
- Facilita la toma de decisiones.

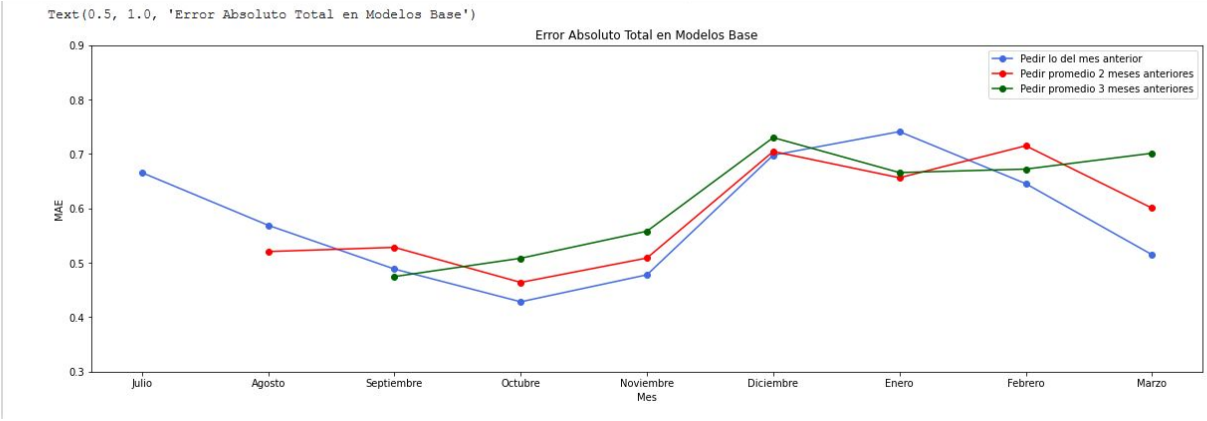
Limitantes
- No permite predecir con mayor confianza
- Pronostica solo un periodo extra

Para el presente trabajo, codificamos en python el método de promedios móviles simple y decidimos trabajar con tres modelos diferentes: pedir promedio del mes anterior, pedir promedio de dos meses anteriores y pedir promedio de tres meses anteriores, esto con la finalidad de tener tres modelos para poder compararlos con mayor facilidad de forma visual y decidir cual es el más adecuado y preciso de los tres. Utilizamos nuestros datos limpios y preparados que estuvimos trabajando desde la actividad 6 con la limpieza, la actividad 7 con el análisis exploratorio y la actividad 8 con la ingeniería de características.

El procedimiento que se llevó a acabo en la programación del método fue: importación de librerías, lectura de datos, eliminación de columnas innecesarias en promedios móviles, construccion de primer modelo(pedir lo del mes anterior), construcción del segundo modelo (pedir lo de dos meses anteriores), construcción de tercer modelo (pedir lo de tres meses anteriores) y cálculo de errores (dividir en conjunto de datos por mes, calcular error més/modelo, crear data/frame con errores y crear gráfica)

	Mes	mae_pedir_anterior	mae_promedio_2_meses_anteriores	mae_promedio_3_meses_anteriores
0	Julio	0.665629	NaN	NaN
1	Agosto	0.568906	0.520591	NaN
2	Septiembre	0.488492	0.528343	0.474672
3	Octubre	0.428352	0.463928	0.508333
4	Noviembre	0.477914	0.508743	0.558028
5	Diciembre	0.697798	0.704527	0.730049
6	Enero	0.740994	0.656197	0.665717
7	Febrero	0.645371	0.714997	0.672124
8	Marzo	0.514869	0.600417	0.701246

Tabla promedio de MAE de cada modelo.



Dataset del promedio de MAE de cada modelo.

De los tres modelos, el que presenta un mejor comportamiento con respecto a la minimización de los errores y que puede ser atractivo para la empresa en cuestión, (además de que el promedio de las MAE sea el menor dentro de los tres modelos utilizados), es el segundo ya que se registra un comportamiento más estable, a comparación del primero que está representado por tener fuertes variaciones mes con mes en el que al principio se observa un decremento, sin embargo después hay un crecimiento un poco gradual hasta llegar al máximo y de ahí se empieza a observar nuevamente una disminución de errores, esto provoca que haya mucha incertidumbre y la empresa en cuestión no estaría muy de acuerdo en adoptar un modelo que se comportara de dicha manera . El último modelo que creíamos podría ser el mejor en un principio, sufre de un incremento en sus errores para llegar al punto maximo y despues se va decrementando poco a poco, sin embargo se puede visualizar nuevamente que después vuelve a haber un crecimiento en la cantidad de errores, por lo que este modelo es muy poco confiable ya que su comportamiento es bastante inestable.
Teniendo estos resultados, posteriormente se busca experimentar con modelos de aprendizaje de máquina, con el objetivo de superar los resultados obtenidos con el modelo de promedio de dos

meses anteriores. Creemos que estos modelos al ser más especializados y complejos puedan ofrecernos una aproximación más acertada en sus pronósticos.

2. *Modelos Aprendizaje de Máquina*

Árbol de decisión

El modelo de aprendizaje de máquina que se eligió y desarrolló para este caso fue el de: árboles de decisión, que consiste en ser un tipo de algoritmo de aprendizaje supervisado y se usa principalmente en problemas de clasificación.

Las variables de entrada y salida pueden ser en especie continuas o categóricas. Divide el espacio de predictores en regiones no sobrepuestas y diferentes.

Entre las limitantes que caracterizan a este modelo, se encuentran las siguientes:

- Pérdida de información al catalogar variables continuas.
- El desarrollar divisiones estratégicas puede alterar la precisión del árbol.
- Se puede modificar la estructura del árbol de manera muy significativa, con un pequeño cambio de datos, suele presentarse un poco de inestabilidad.
- Las máquinas de vectores de soporte, que es otro tipo de modelo de aprendizaje de máquina, generalmente tienen tasas de error 30% más bajas que los árboles de decisión.

Para la etapa de este trabajo, se llevó a cabo el desarrollo del código en el lenguaje de programación: python. Se decidió trabajar con el modelo que expone mejor comportamiento de la subetapa de modelado: promedios móviles, igual desarrollada en python y ésta fue: pedir promedio de dos meses anteriores.

El procedimiento que se llevó a cabo para el desarrollo del método fue: lectura de datos (importar librerías, leer los datos), validación cruzada-preparación de los datos(dividir los datos en entrenamiento y prueba, generación de índices en python, división de ambos conjuntos en variables de entrada y variable respuesta, visualización de lo que hizo anteriormente), modelación (determinación de un rango de valores, importación de función para construcción de árboles), entrenamiento de modelo (construcción de modelo, entrenamiento de modelo, cálculo error de entrenamiento, calculo de maximos y minimos de predicción, redondeo y ajustes de valores, cúmulo de datos para cálculo de error, calculo de errores con funciones de paquetería, MAE), prueba de modelo(calculo error prueba, encontrar los maximos y minimos, redondeo y ajuste de valores, obtención respuesta a la pregunta principal, juntar datos para encontrar el error, calculo de errores con funciones de paquetería) registro de datos manualmente en un archivo de excel, en una nueva hoja de python se leyeron nuevamente los datos pero ahora con el registro de los errores, tal y como se muestra en el anexo 2, se procedió al filtrado de datos por entrenamiento y prueba y los conjuntos anteriores por métrica, se graficó los errores de entretenimiento y de prueba.

Random forest

Este modelo de aprendizaje de máquina es útil para la regresión y clasificación.

- Sirve para la reducción de dimensionalidad; en un archivo de muchísimas variables de entrada, puede manejar las más significativas.
- Generación de múltiples árboles
- Detección de datos atípicos
- Estimación de valores faltantes mediante métodos efectivos.

Limitantes:

- Se llega a perder claridad de interpretación
- Las predicciones no suelen ser de naturaleza continua
- No se pueden llegar a predicciones más allá de los valores del conjunto de entrenamiento.

La elaboración de este tipo modelo se constituye de esta manera:

- Se toma una muestra de N datos que se seleccionaron con reemplazo. La muestra suele ser ocupada para construir el árbol, ya que este será el conjunto de entrenimiento.
- En dado caso que exista M variables de entrada, m variables se elige del conjunto M de manera aleatoria. La división que presentó mejor comportamiento se utiliza para ramificar el árbol; m se debe mantener constante durante la generación de este modelo.
- Las instancias surgen a partir de la agregación de predicciones de los árboles.

Este método de aprendizaje se puede interpretar como “cajas negras”, ya que no se puede comprender o racionalizar el porqué se otorga cierta predicción a algún conjunto de variables predictoras. Además nos proporciona el entendimiento, con respecto a una variable explicativa, de la medición del error. De igual forma, se hace un promedio o representación de dos casos cuando se encuentran en el mismo nodo terminal un número elevado de veces.

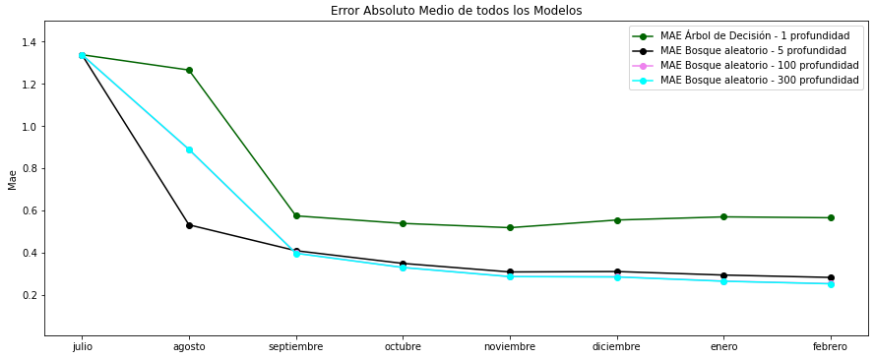
V. Etapa 5: Evaluación

Cada una de las etapas de este proyecto tuvo su nivel de dificultad, desde la identificación de los datos su limpieza su categorización, empezar a conocer que nos decían cada uno de los números hasta empezar aplicar modelos con el objetivo de predecir la demanda para esta empresa.

Cómo aclaramos en la introducción, el objetivo principal de este proyecto es predecir las ventas para el siguiente mes de la empresa Huawei para cada uno de esos celulares comenzando con un modelo de promedios móviles simples, a partir de este modelo pudimos tener una predicción que nos diera una referencia del posible resultado para las ventas futuras, sin embargo según la teoría no es el mejor método para predecir ventas futuras ya que estos datos cuenten con estacionalidad, además de una clara tendencia es aquí donde se tiene que empezar a explorar modelos más complejos para realizar predicciones, entonces llegamos a los árboles de decisión realizamos un árbol con profundidad 1 y nos dio mejores resultados que los promedios móviles simples (modelo base) sin embargo sabíamos que existían otros modelos que se podrían acoplar y predecir mejor nuestros datos. De esta manera, llegamos a los bosques aleatorios como se describió en la etapa anterior, para los bosques aleatorios comenzamos con un bosque de profundidad cinco que nos arrojó un resultado ligeramente menor que los promedios móviles de dos meses, sin embargo a medida que íbamos aumentando la profundidad estos árboles van reduciendo el error tanto en el entrenamiento como en la prueba.

A continuación se presenta la gráfica del error absoluto medio para cada uno de los modelos en la etapa de entrenamiento, cabe aclarar que en el entrenamiento no es posible graficar el modelo de promedios móviles simples, pues no le corresponde una etapa de entrenamiento porque no es un modelo de aprendizaje de máquina; se puede observar que comienzan con un error absoluto medio por arriba de 1.2, a medida que se van realizando más entrenamientos las predicciones para cada uno de los siguientes meses van reduciendo significativamente su error absoluto hasta llegar a

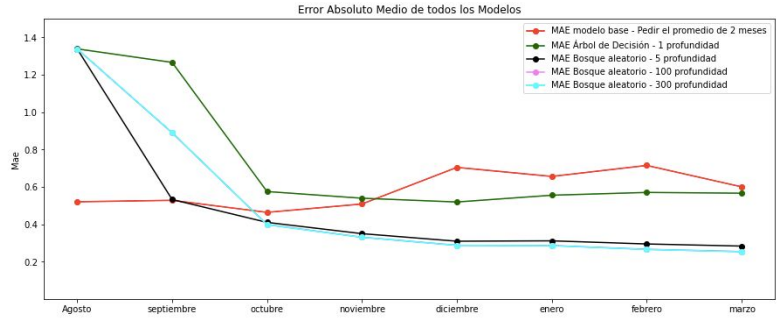
mínimos cercanos de 0.35. Podemos ver que para el árbol de decisión se estanca en mínimos de 0.5 mientras que se nota una ligera tendencia a la baja para los modelos de aprendizaje de máquina.



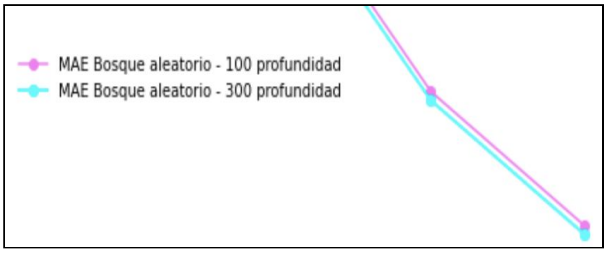
Gráfica error absoluto medio de todos los modelos. MAE: Árbol de Decisión.

Una vez concluida la fase de entrenamiento es preciso pasar a la fase de prueba en la cual cada uno de los modelos, incluyendo ahora si el modelo base prueban sus capacidades de predicción; las gráficas nos muestran que durante los primeros meses, el modelo base es más preciso que los otros. Sin embargo, a partir de Noviembre, el error medio del mismo comienza a aumentar hasta mantenerse estable por encima del 0.6 mientras que los otros modelos reducen enormemente su valor hasta alcanzar en el caso de Bosque aleatorio 5 profundidad y Bosque aleatorio 300 profundidad valores menores a 0.4.

A partir de esto podemos concluir que los modelos de aprendizaje de máquina son significativamente mejores que los modelos tradicionales.
Podemos ver que un bosque aleatorio de profundidad 5 es muy parecido a un bosque aleatorio de profundidad 100, pero además de eso el bosque aleatorio de 100 de profundidad no se puede observar claramente pues siempre se mantiene con valores extremadamente parecidos a los de profundidad 300; es por eso que en la gráfica entre el bosque aleatorio 100 y 300 podemos ver que la diferencia es muy pequeña a partir de esto, nos nace la duda si existen diferencias significativas entre estos tres modelos.



Gráfica error absoluto medio de todos los modelos. MAE: Modelo Base.



Representación gráfica de MAE Bosque aleatorio 100 y 300 profundidad

Para resolver estas dudas realizamos un análisis de varianza (ANOVA) para determinar si existen diferencias significativas entre los tres bosques aleatorios, consideramos desde el mes de octubre hasta el mes de marzo, poniendo como factores el mes y la profundidad del árbol de decisión. Pudimos demostrar dos de los tres supuestos ya que el recibo contra el orden fue significativo porque se trata de datos tomados cronológicamente.

Este estudio se puede encontrar en el anexo número dos y a partir de él podemos concluir que sí existe una diferencia significativa en el error absoluto medio en alguno de los modelos, es por esto que realizamos una prueba de Tukey para determinar cuáles son aquellos que demuestran diferencias significativas y pudimos concluir que el modelo de bosque aleatorio de profundidad cinco es diferente al de 100 y de 300, mientras que los últimos dos no son estadísticamente diferentes, si quisiéramos predecir las ventas reduciendo el error lo más que se pueda sin importar la capacidad computacional recomendaríamos realizar un bosque aleatorio de profundidad 300. Sin embargo por la rapidez de la realización del modelo así como la cantidad de recursos computacionales necesarios para realizarlo, si no se cuenta con estos dos factores y se busca una respuesta certera, recomendaremos un bosque aleatorio de profundidad 100 puesto que no representa diferencias con el modelo de 300 estadísticamente, sugerimos que como siguientes pasos se pudieran realizar más pruebas de las diferencias entre las distintas profundidades del modelo.

A partir de todos los conocimientos obtenidos en este proyecto podemos exponer determinadamente que un modelo de aprendizaje de máquina de bosque aleatorio con una **profundidad de 300** fue la mejor manera de predecir las ventas futuras para nuestra marca.

Anexos

Anexo 1

[1] Lista de teléfonos que ofrece la marca

- Huawei P40 Pro+
- Huawei P40 Pro
- Huawei P40
- Huawei Mate Xs
- Huawei Mate 30 Pro
- Huawei P40 lite
- Huawei Y5p
- Huawei Y6p
- Huawei Y7p
- Huawei Y8p
- Huawei Y6s
- Huawei Y8s
- Huawei Y9s
- Huawei Y9 Prime 2019
- Huawei nova 5T
- Huawei P30 Pro
- Huawei P30
- Huawei P30 lite
- Huawei Mate 20 Pro
- Huawei Mate 20
- Huawei smart 2019
- Huawei Mate 20 lite
- Huawei Y7 2019
- Huawei Y6 2019

Obtenido de :
https://consumer.huawei.com/mx/phones/

[2] Lista de puntos de venta de accesorios en México

- Tienda Huawei Masaryk
- Tienda Huawei Forum Buenavista
- Tienda Huawei Oasis Coyoacán
- Tienda Huawei Metepec
- Tienda Huawei Santa Fé
- Tienda Huawei Toreo
- Tienda Huawei Tezontle
- Tienda Huawei Andares
- Tienda Huawei Paseo Acoxpa
- Tienda Huawei Parque las Antenas
- Tienda Huawei Antea

Obtenido de :
https://consumer.huawei.com/mx/storelist/

[3] Lista de sitios oficiales en línea

- Huawei Store
- Linio
- Amazon
- Mercado Libre

Lista de tiendas físicas distribuidoras

- Sanborns
- A-móvil
- Best Buy
- Coppel
- Liverpool

Obtenido de :
https://consumer.huawei.com/mx/support/content/es-us00375282/

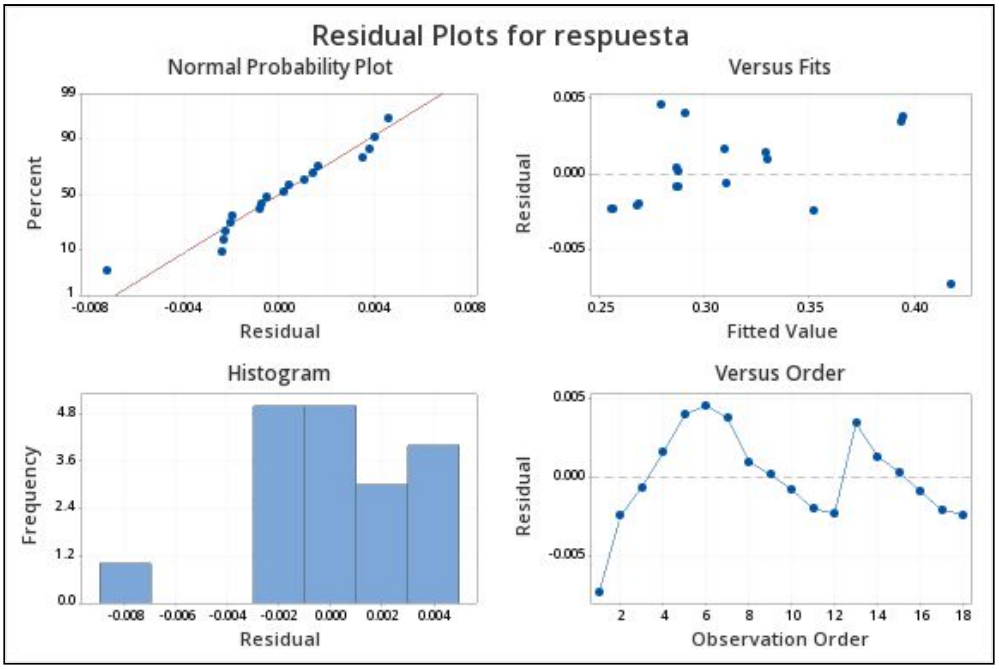
Anexo 2

Resultados de modelo base, árboles de decisión y bosques aleatorios

metrica	conjunto	mes	modelo_base	dt_1_profundidad	rf_1_profundidad_5	rf_1_profundidad_100	rf_2_profundidad_300
mae	entrenamiento	julio	NaN	1.338553	1.338351	1.338351	1.338351
mae	entrenamiento	agosto	NaN	1.266115	0.532796	0.889206	0.890809
mae	entrenamiento	septiembre	NaN	0.57538	0.410355	0.398941	0.397877
mae	entrenamiento	octubre	NaN	0.53961	0.350029	0.330987	0.330604
mae	entrenamiento	noviembre	NaN	0.51939	0.309646	0.287941	0.287371
mae	entrenamiento	diciembre	NaN	0.55565	0.311627	0.28674	0.285912
mae	entrenamiento	enero	NaN	0.57061	0.29491	0.266421	0.265615
mae	entrenamiento	febrero	NaN	0.56645	0.283611	0.254248	0.253458
mae	prueba	Agosto	0.52059	1.33855	1.338351	1.338351	1.338351
mae	prueba	septiembre	0.52834	1.26612	0.532796	0.889206	0.890809
mae	prueba	octubre	0.46393	0.57538	0.410355	0.398941	0.397877
mae	prueba	noviembre	0.50874	0.53961	0.350029	0.330987	0.330604
mae	prueba	diciembre	0.70453	0.51939	0.309646	0.287941	0.287371
mae	prueba	enero	0.6562	0.55565	0.311627	0.28674	0.285912
mae	prueba	febrero	0.715	0.57061	0.29491	0.266421	0.265615
mae	prueba	marzo	0.60042	0.56645	0.283611	0.254248	0.253458

ANEXO 3

ANOVA y Tukey



Comprobación de supuestos, normalidad, independencia y varianza.

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
mes	Fixed	6	3, 4, 5, 6, 7, 8
profundidad	Fixed	3	5, 100, 300

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
mes	5	0.039115	0.007823	525.84	0.000
profundidad	2	0.002091	0.001045	70.27	0.000
Error	10	0.000149	0.000015		
Total	17	0.041354			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0038571	99.64%	99.39%	98.83%

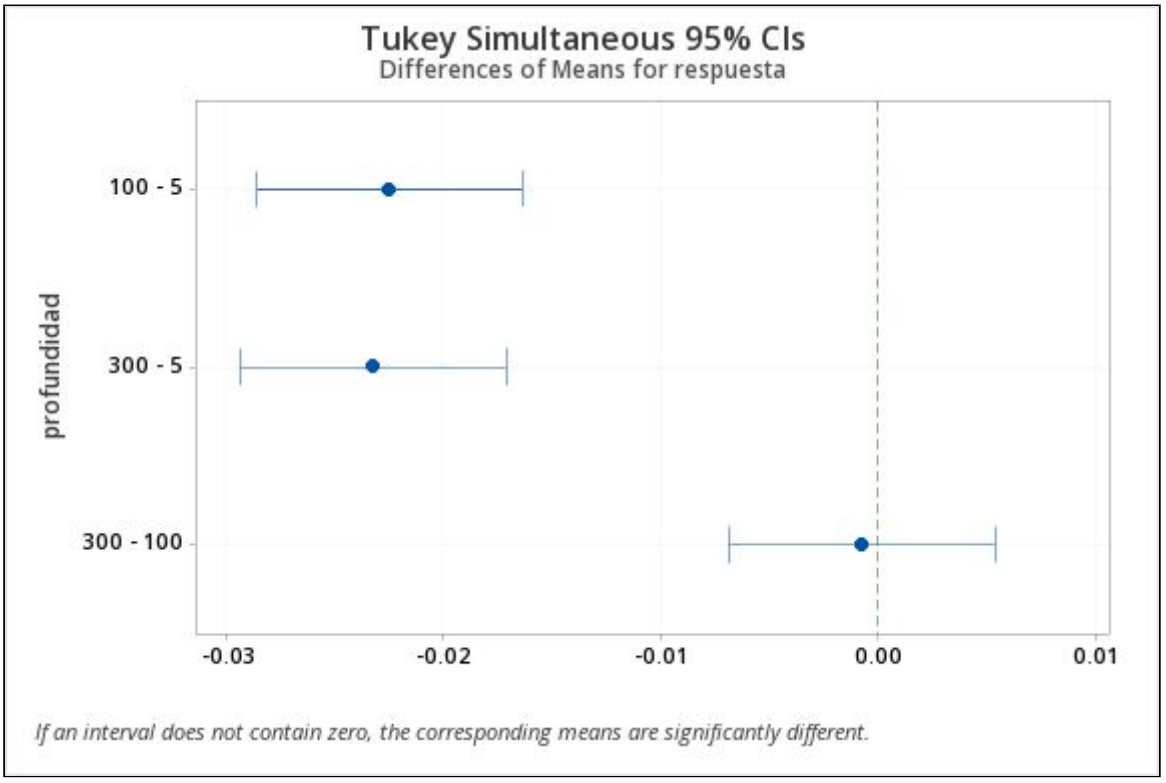
Análisis de varianza (ANOVA)

Tukey Pairwise Comparisons: profundidad

Grouping Information Using the Tukey Method and 95% Confidence

profundidad	N	Mean	Grouping
5	6	0.326696	A
100	6	0.304213	B
300	6	0.303473	B

Means that do not share a letter are significantly different.



Comparación Tukey

Bibliografía

[a] Devore, J.. (2018). Probabilidad y Estadística para Ingeniería y Ciencias. México: CENGAGE Learning.

[b] El Financiero. (2018, April 19). *Huawei, y el sigiloso plan para "comerse" el mercado en México... y el mundo*. https://www.elfinanciero.com.mx/bloomberg-businessweek/huawei-quiere-venderte-todo-para-tu-smartphone-y-no-solo-un-telefono

[c] El Financiero. (2019, August 19). *Estas 5 firmas lideran el mercado de los smartphones en México*.

https://www.elfinanciero.com.mx/empresas/estas-5-firmas-lideran-el-mercado-los-smartphones-en-mexico

[d] Microsoft. (2020, May 14). *Ingeniería de características en ciencia de datos: proceso de ciencia de datos en equipo*. Microsoft Docs. https://docs.microsoft.com/es-es/azure/machine-learning/team-data-science-process/create-features

[e] Orellana, J.. (2018). Árboles de decisión y Random Forest. noviembre 18, 2020, de Ucuenca Sitio web: https://bookdown.org/content/2031/

[f] Bosques Aleatorios. noviembre 18, 2020, de coursera Sitio web: https://es.coursera.org/lecture/big-data-procesamiento-analisis/bosques-aleatorios-zjCqX