



UNIVERSIDAD TÉCNICA DE MACHALA
UNIDAD ACADÉMICA DE INGENIERÍA CIVIL

PROPUESTA METODOLÓGICA Y TECNOLÓGICA AVANZADA
EN OPCIÓN AL TÍTULO DE MAGÍSTER EN SOFTWARE

TEMA

ANÁLISIS PREDICTIVO DE DATOS APLICADO AL SECTOR AUTOMOTRIZ

AUTOR

ING. JUAN CARLOS JARAMILLO PONTÓN

TUTORA

ING. BERTHA EUGENIA MAZÓN OLIVO

COTUTOR

ING. DIXYS LEONARDO HERNÁNDEZ ROJAS

MACHALA

2021

PENSAMIENTO

“Estudia el pasado si quieres pronosticar el futuro”

Confucio, Filósofo

“Ya no estamos en la era de la información. Estamos en la era de
la gestión de la información”

Chris Hardwick, Actor

DEDICATORIA

Dedico este trabajo principalmente a Dios, por haberme dado la vida y permitirme el haber llegado hasta este momento tan importante de mi formación profesional. A mis padres, por ser el pilar más importante y por demostrarme siempre su amor, cariño y apoyo incondicional en todo momento. A mi esposa por siempre estar a mi lado, por su apoyo incondicional en mi vida, brindarme su amor, respaldo y ayudarme a alcanzar mis objetivos. A mis hijas por el motor de mi vida y lo que me motivan a seguir adelante día a día.

AGRADECIMIENTO

Agradezco primeramente a Dios por regalarme la vida, por guiarme a lo largo de mi existencia, por ser el apoyo y fortaleza en aquellos momentos de dificultad y de debilidad.

A toda mi familia, a mi esposa, mis hijas, mi padre, mi madre y mis hermanas por ser el pilar fundamental en cada paso que doy.

De manera especial a mi tutora de tesis Mgs. Bertha Mazón Olivo, por haberme guiado, en la elaboración de este trabajo de titulación. También a los maestros, maestras y coordinadora de la maestría que compartieron su conocimiento en cada módulo recibido.

A mis compañeros y amigos con los que compartí muchas cosas que han permitido mi crecimiento personal y profesional dentro y fuera de las aulas.

RESPONSABILIDAD DE AUTORÍA

Por medio de la presente declaro ante el comité Académico de la Maestría en Software de la Universidad Técnica de Machala. Que el trabajo de titulación titulado “**ANÁLISIS PREDICTIVO DE DATOS APLICADO AL SECTOR AUTOMOTRIZ**”, de mi propia autoría, no contiene material escrito por otra persona al no ser referenciado debidamente en el texto, parte de ella o en su totalidad no ha sido aceptada para el otorgamiento de cualquier otro diploma de una institución nacional o extranjera.

Juan Carlos Jaramillo Pontón

C.I. 0702585761

Machala, 2021/06/18

REPORTE DE SIMILITUD TURNITI

ANÁLISIS PREDICTIVO DE DATOS APLICADO AL SECTOR AUTOMOTRIZ

INFORME DE ORIGINALIDAD

9%

ÍNDICE DE SIMILITUD

9%

FUENTES DE INTERNET

1%

PUBLICACIONES

4%

TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1

docs.microsoft.com

Fuente de Internet

2%

2

www.coursehero.com

Fuente de Internet

<1%

3

repositorio.usil.edu.pe

Fuente de Internet

<1%

4

docplayer.es

Fuente de Internet

<1%

5

Submitted to Universidad Andina del Cusco

Trabajo del estudiante

<1%

6

www.fcecon.unr.edu.ar

Fuente de Internet

<1%

7

www.natacionmasteralbacete.es

Fuente de Internet

<1%

8

tesis.ucsm.edu.pe

Fuente de Internet

<1%

9

dspace.udla.edu.ec

Fuente de Internet

CERTIFICADO DEL TUTOR

Por medio de la presente apruebo que el trabajo de titulación, titulado “**ANÁLISIS PREDICTIVO DE DATOS APLICADO AL SECTOR AUTOMOTRIZ**”, del autor Juan Carlos Jaramillo Pontón, en opción al título de Magíster en Software, sea presentado al Acto de Defensa.

Ing. Bertha Mazón Olivo, Mgs

C.I. 0603100512

Machala, 2021/06/18

CESIÓN DE DERECHOS DE AUTORÍA

Yo, **Juan Carlos Jaramillo Pontón**, en calidad de autor del presente trabajo titulado **“ANÁLISIS PREDICTIVO DE DATOS APLICADO AL SECTOR AUTOMOTRIZ”**, Autorizo a la UNIVERSIDAD TÉCNICA DE MACHALA la publicación y distribución en el Repositorio Digital Institucional de la Universidad Técnica de Machala.

El autor declara que el contenido que se publicará es de carácter académico y se enmarca en las disposiciones definidas por la Universidad Técnica de Machala.

Juan Carlos Jaramillo Pontón

C.I. 0702585761

Machala, 2021/06/18

RESUMEN

En la actualidad los sistemas de información representan una herramienta de apoyo a la toma de decisiones en toda organización, esto debido al rápido desarrollo de las tecnologías y al crecimiento continuo de la información, las mismas que generan una gran cantidad de datos todos los días. Estos datos nos brindan la oportunidad de generar conocimiento, esto debe ser aprovechados para obtener información oculta, que sirva como oportunidades de negocios para las empresas e instituciones, para ello podemos aplicar diversas técnicas de minería de datos, análisis de datos y modelos predictivos de datos. La predicción necesita descubrir la correlación interna del objeto y predecir tendencias futuras, por esto la predicción se vuelve cada vez más importante. En este trabajo se analizó segmentos de ventas de vehículos en el sector automotriz aplicando modelos de minería de datos predictivos para la determinación de tendencias, se propone predecir el crecimiento del parque automotor en el Ecuador, con este propósito la estrategia seguida fue utilizar información histórica de venta de vehículo en unidades a nivel nacional comprendida de enero 2007 a julio 2020, la metodología aplicada fue Proceso de Ciencia de Datos en Equipo (TSDP), se comenzó con la limpieza y depuración de variables, para posteriormente aplicar técnicas de análisis de información y minería de datos; estos resultados se almacenaron en un repositorio para facilitar el análisis de datos y obtener resultados; como siguiente paso se realizó el análisis predictivo utilizando series temporales con varios modelos predictivos: STLM (Descomposición estacional y de tendencias usando Loess con múltiples períodos estacionales), STLFI (Pronóstico de carga a corto plazo), Holt Winters, STLM ARIMA 3.1.6, EST (Error, Tendencia, Estacional), Auto Arima (Método ARIMA ajustado automáticamente en cada serie), NNETAR (Pronósticos de Series de Tiempo de Redes Neuronales) y TBATS (Estacionalidad trigonométrica, Transformación Box-Cox, Errores ARMA, Componentes de tendencia y estacionales), para predecir el crecimiento del parque automotor en el Ecuador, tomando en consideración un período de 36 meses y así conocer el comportamiento en un futuro. Los modelos fueron comparados mediante sus métricas con el propósito de identificar cuál es el más eficiente para su uso. Se concluye que los modelos STLM ARIMA 3.1.6 (Modelo híbrido de STLM y ARIMA “Modelo autorregresivo integrado de media móvil”), y STLM son los más óptimos con las métricas más bajas en MAE (Error Absoluto Medio) y MAPE (Error Porcentual Absoluto Medio). Cabe mencionar que al

realizar una comparación con los resultados obtenidos con los datos reales de los últimos siete meses el modelo que más se acerca a ellos es NNETAR con un 6.24% de diferencia. Esto se puede asociar a la pandemia del Covid 19 que afectó a nuestro país y el mundo entero, donde los últimos meses las ventas de en este sector han tenido un comportamiento atípico.

PALABRAS CLAVES: Minería de Datos, Ciencia de Datos, Análisis Predictivo, Series Temporales, Sector Automotriz.

ABSTRACT

Currently, information systems represent a tool to support decision making in any organization, due to the rapid development of technologies and the continuous growth of information, which generate a large amount of data every day. These data give us the opportunity to generate knowledge, this must be used to obtain hidden information, which serves as business opportunities for companies and institutions, for this we can apply various techniques of data mining, data analysis and predictive data models. Prediction needs to discover the internal correlation of the object and predict future trends, so prediction is becoming increasingly important. In this work we analyzed segments of vehicle sales in the automotive sector by applying predictive data mining models to determine trends, we propose to predict the growth of the vehicle fleet in Ecuador, for this purpose the strategy followed was to use historical information of vehicle sales in units nationwide from January 2007 to July 2020, the methodology applied was Team Data Science Process (TSDP), we began with the cleaning and debugging of variables, and then apply information analysis techniques and data mining; These results were stored in a repository to facilitate data analysis and obtain results; as a next step, predictive analysis was performed using time series with several predictive models: STLM (Seasonal and Trend Decomposition using Loess with multiple seasonal periods), STLF (Short Term Load Forecast), Holt Winters, STLM ARIMA 3. 1.6, EST (Error, Trend, Seasonal), Auto Arima (ARIMA method automatically adjusted in each series), NNETAR (Neural Network Time Series Forecasts) and TBATS (Trigonometric Seasonality, Box-Cox Transformation, ARMA Errors, Trend and Seasonal Components), to predict the growth of the vehicle fleet in Ecuador, taking into consideration a 36-month period and thus know the behavior in the future. The models were compared by means of their metrics in order to identify which one is the most efficient to use. It is concluded that the STLM ARIMA 3.1.6 (Hybrid model of STLM and ARIMA "Moving Average Integrated Autoregressive Model"), and STLM are the most optimal models with the lowest metrics in MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). It is worth mentioning that when comparing the results obtained with the real data of the last seven months, the model that comes closest to them is NNETAR with a 6.24% difference. This can be associated to the Covid 19 pandemic that affected our country and

the whole world, where in the last months the sales in this sector have had an atypical behavior.

KEY WORDS: Data Mining, Data Science, Predictive Analysis, Time Series, Automotive section.

ÍNDICE GENERAL

Pág.

INTRODUCCIÓN.....	24
CAPÍTULO 1	31
MARCO TEÓRICO REFERENCIAL	31
1.1 ANTECEDENTES HISTÓRICOS DE LA INVESTIGACIÓN.....	31
1.1.1 PREGUNTAS DE INVESTIGACIÓN.....	31
1.1.2 OBJETIVOS.....	32
1.1.2.1 OBJETIVO GENERAL.....	32
1.1.2.2 OBJETIVOS ESPECÍFICOS.....	32
1.1.3 PROCESO DE BÚSQUEDA.....	32
1.1.4 CRITERIO DE INCLUSIÓN Y EXCLUSIÓN	33
1.1.4.1 CRITERIO DE INCLUSIÓN	33
1.1.4.2 CRITERIO DE EXCLUSIÓN	33
1.1.5 GRUPO DE CONTROL.....	33
1.1.6 CADENA DE BÚSQUEDA	34
1.1.7 SELECCIÓN DE ESTUDIOS	36
1.1.8 RESULTADOS DE LA REVISIÓN – ESTADO DEL ARTE.....	36
1.2 ANTECEDENTES CONCEPTUALES	41
1.2.1 HIPÓTESIS DE INVESTIGACIÓN	41
1.2.2 RED DE CATEGORÍAS DE LAS VARIABLES	41
1.2.2.1 VARIABLES DEPENDIENTES	42
1.2.2.2 VARIABLES INDEPENDIENTES.....	42
1.2.3 FUNDAMENTACIÓN TEÓRICA DE LA VARIABLE INDEPENDIENTE	42
1.2.3.1 INTELIGENCIA DE NEGOCIOS.....	42
1.2.3.2 BI COMPONENTES.....	42
1.2.3.3 PROCESO ETL.....	43
1.2.3.4 DATA WAREHOUSE	44
1.2.3.5 DATA MART.....	45
1.2.3.6 HERRAMIENTAS DE VISUALIZACIÓN	45
1.2.3.7 CUADRANTE MÁGICO DE GARTNER.....	46

1.2.3.8 ANALÍTICA DE DATOS	47
1.2.3.9 ANÁLISIS PREDICTIVO.....	48
1.2.3.10 CICLO DE VIDA DEL ANÁLISIS PREDICTIVO	48
1.2.3.11 MINERÍA DE DATOS.....	50
1.2.3.12 TÉCNICAS DE MINERÍA DE DATOS	51
1.2.3.13 ALGORITMOS DE MINERÍA DE DATOS	52
1.2.3.14 ANÁLISIS DE REGRESIÓN	52
1.2.3.15 ANÁLISIS DE VARIANZA Y COVARIANZA	53
1.2.3.16 SERIES TEMPORALES	53
1.2.3.17 MÉTODOS BAYESIANOS	54
1.2.3.18 ANÁLISIS DISCRIMINANTES	54
1.2.3.19 ALGORITMO DE ÁRBOL DE DECISIÓN	54
1.2.3.20 REDES NEURONALES	55
1.2.3.21 METODOLOGÍA CRISP-DM	55
1.2.3.22 METODOLOGÍA KDD.....	56
1.2.3.23 METODOLOGÍA SEMMA.....	57
1.2.3.24 METODOLOGÍA TDSP	58
1.2.4 FUNDAMENTACIÓN TEÓRICA DE LA VARIABLE DEPENDIENTE.....	59
1.3 ANTECEDENTES CONTEXTUALES.....	60
1.3.1 DELIMITACIÓN DEL CONTEXTO DEL ESTUDIO.....	60
1.3.2 PROPUESTA DE SOLUCIÓN Y CONTRIBUCIONES.....	61
CAPÍTULO 2	62
MATERIALES Y MÉTODOS	62
2.1 TIPO DE ESTUDIO O INVESTIGACIÓN REALIZADA	62
2.2 EL PARADIGMA O ENFOQUE DEL TRABAJO	64
2.2.1 CUANTITATIVO	64
2.2.2 CUASI EXPERIMENTAL.....	66
2.3 CALCULO DE LA POBLACIÓN Y MUESTRA	67
2.4 MÉTODOS TEÓRICO CON MATERIALES UTILIZADOS.....	68
2.5 MÉTODOS EMPÍRICOS CON MATERIALES UTILIZADOS	70
2.6 TÉCNICAS ESTADÍSTICAS PARA LA INVESTIGACIÓN	70
2.7 HERRAMIENTAS UTILIZADAS.....	71
CAPÍTULO 3	72

PROPUESTA DE ANÁLISIS DE DATOS EN EL SECTOR AUTOMOTRIZ APLICACIÓN DE TÉCNICA DE MINERÍA DE DATOS.....	72
3.1 FUNDAMENTACIÓN TEÓRICA DE LA PROPUESTA.....	72
3.1.1 METODOLOGÍAS DE MINERÍA DE DATOS.....	72
3.1.2 COMPARACIÓN DE METODOLOGÍAS MINERÍA DE DATOS	72
3.1.3 SELECCIÓN DE LA METODOLOGÍA	74
3.1.4 METODOLOGÍAS TDSP	75
3.1.4.1 COMPONENTES DEL TDSP	75
3.1.4.2 CICLO DE VIDA TDSP	75
3.1.4.3 ESTRUCTURA DE PROYECTO ESTANDARIZADA	83
3.1.4.4 INFRAESTRUCTURA Y RECURSOS	84
3.1.4.5 HERRAMIENTAS Y UTILIDADES.....	85
3.1.4.6 ROLES Y TAREAS	85
3.1.5 SERIES TEMPORALES	87
3.1.5.1 MODELO STLM	87
3.1.5.2 MODELO STLF.....	87
3.1.5.3 MODELO HOLT WINTERS	88
3.1.5.4 MODELO STLM ARIMA 3,1,6	88
3.1.5.5 MODELO EST SUAVIZACIÓN EXPONENCIAL	88
3.1.5.6 MODELO AUTO.ARIMA	89
3.1.5.7 MODELO NNETAR	89
3.1.5.8 MODELO TBATS.....	89
3.2 PROPUESTA METODOLÓGICA.....	89
3.2.1 APLICACIÓN METODOLOGIA TDSP.....	90
3.2.1.1 ENTENDIMIENTO DEL NEGOCIO	90
3.2.1.2 ADQUISICIÓN Y COMPRENSIÓN DE DATOS	91
3.2.1.3 EXPLORACIÓN DE DATOS - MODELADO	92
CAPITULO 4	100
PRESENTACIÓN Y DISCUSIÓN DE RESULTADOS OBTENIDOS	100
4.1 ESTADÍSTICA DESCRIPTIVA EN MINERÍA DE DATOS.....	100
4.2 TÉCNICA DE MINERÍA DE DATOS UTILIZADA.....	105
4.2.1 SERIES TEMPORALES	105
4.2.1.1 MODELO STLM	107
4.2.1.2 MODELO STLF.....	108

4.2.1.3 MODELO HOLTWINTERS	109
4.2.1.4 MODELO STLM ARIMA 3,1,6	111
4.2.1.5 MODELO EST SUAVIZACIÓN EXPONENCIAL	112
4.2.1.6 MODELO AUTO.ARIMA	113
4.2.1.7 MODELO NNETAR	115
4.2.1.8 MODELO TBATS.....	116
4.3 INTERPRETACIÓN DE RESULTADOS.....	117
4.3.1 COMPARACIÓN RESULTADOS DE MODELO DE SERIES TEMPORALES	118
4.3.2 DATOS DE PREDICCIÓN DE LOS MODELOS	118
4.3.3 COMPARACIÓN DATOS PREDICTIVOS CON DATOS REALES.....	120
4.4 TRABAJOS FUTUROS.....	130
CONCLUSIONES	131
RECOMENDACIONES	133
BIBLIOGRAFÍA.....	134

LISTA DE ILUSTRACIONES Y TABLAS

FIGURAS

	Pág.
Figura 1. Crecimiento parque automotor, Agencia Nacional de Transito 2008-2018.....	27
Figura 2. Vehículos por Segmento	28
Figura 3. Número de artículos encontrados por Base de Datos	35
Figura 4. Red de Categorías de las Variables	41
Figura 5. Componentes BI.....	43
Figura 6. Proceso ETL.	44
Figura 7. Arquitectura Data Warehouse.....	45
Figura 8. Cuadrante mágico de Gartner - Plataformas de análisis BI, 2020.	46
Figura 9. Cuadrante mágico de Gartner DS y aprendizaje automático, 2020.	47
Figura 10. Ciclo de vida del análisis predictivo.	49
Figura 11. Clasificación de las técnicas de Minería de Datos	51
Figura 12. Fases del modelo de proceso de la metodología CRISP-DM [55]	56
Figura 13. Etapas de KDD.....	57
Figura 14. Fases de SEMMA.....	58
Figura 15. Ciclo de vida metodología TDSP.....	59
Figura 16. Alcance y tipo de Investigación Descriptiva.....	63
Figura 17. Rasgos de la Investigación Descriptiva.	64
Figura 18. Enfoque Investigación Cuantitativo.	65
Figura 19. Fases de proceso Cuantitativo.	66
Figura 20. Diseño Cuasi-Experimental	67
Figura 21. Estructura del Dataset.	68
Figura 22. Componentes claves de TDSP.....	69
Figura 23. Fases de TDSP [53]	69

Figura 24. Comparación Metodologías.....	73
Figura 25. Ciclo de vida TDSP.....	76
Figura 26. Tareas y artefactos. Ciclo TDSP y Roles.....	77
Figura 27. Estructura del directorio TDSP.....	84
Figura 28. Tareas según de los roles.....	86
Figura 29. Estructura completa dataset “datos”.....	90
Figura 30. Arquitectura de solución.....	91
Figura 31. Estructura completa dataset “DatosGV”.....	94
Figura 32. Evolución mensual en número de vehículos 2007 - 2020	95
Figura 33. Evolución mensual en número de vehículos – Grafica.....	96
Figura 34. Evolución de las unidades.....	96
Figura 35. Grafica Serie No Estacional.....	97
Figura 36. Grafica Serie no Estacional – Traza de subseries.....	97
Figura 37. Grafica Diferencia – Traza en cajas.....	98
Figura 38. Grafica Eliminar el componente estacional.....	99
Figura 39. Grafica descomposición de series de tiempo multiplicativas.....	99
Figura 40. Ventas por mes periodo enero 2007 a julio 2020	101
Figura 41. Unidades por mes periodo enero 2007 a julio 2020 – Generada con R Studio.....	102
Figura 42. Total de unidades por año	103
Figura 43. Top 10 Ventas por Provincia.....	103
Figura 44. Top 10 Ventas por Marca.....	104
Figura 45. Top 10 Ventas por país de procedencia.....	104
Figura 46. Top 10 Ventas por modelo.....	105
Figura 47. Unidades Vendidas por Mes – Visualizadas en la Herramienta R Studio.....	106
Figura 48. Aplicación Modelo STL.....	107
Figura 49. Aplicación Modelo STLF.....	108
Figura 50. Aplicación HoltWinters.....	110
Figura 51. Aplicación STLM ARIMA 3,1,6.....	111

Figura 52. Aplicación Modelo EST suavización exponencial.....	112
Figura 53. Aplicación Modelo Auto.Arima.....	114
Figura 54. Aplicación NNETAR.	115
Figura 55. Aplicación modelo TBATS.....	116
Figura 56. Comparación de modelo de series temporales vs datos reales	120
Figura 57. Modelo TSLM vs Datos Reales.	121
Figura 58. Modelo TSLF vs Datos Reales.....	122
Figura 59. Modelo Holt Winters vs Datos Reales.	123
Figura 60. Modelo STLM arima 3,1,6 vs Datos Reales.	124
Figura 61. Modelo EST arima 3,1,6 vs Datos Reales.	125
Figura 62. Modelo Auto Arima vs Datos Reales.....	126
Figura 63. Modelo NNETAR vs Datos Reales.	127
Figura 64. Modelo TBATS vs Datos Reales.....	128
Figura 65. Comparación de modelos - Margen de error.	129

TABLAS

	Pág.
Tabla 1. Vehículos por Segmento - agosto 2020	27
Tabla 2. Búsqueda de Número de Artículos	35
Tabla 3. Variables Independiente y Dependiente	67
Tabla 4. Comparación de metodología minería de datos.....	74
Tabla 5. Corrección datos campo Marca	92
Tabla 6. Corrección datos campo Segmento	93
Tabla 7. Corrección datos campo Provincia.....	93
Tabla 8. Corrección datos campo País Origen	94
Tabla 9. Ventas unidades por mes años 2007 – 2020	100
Tabla 10. Ventas por año 2007 – 2020	101
Tabla 11. Información Dataset número de unidades.	106
Tabla 12. Predicciones Modelo STL.	107
Tabla 13. Predicciones Modelo STLF.	109
Tabla 14. Predicciones Holt Winters.....	110
Tabla 15. Predicciones Modelo STLF ARIMA 3,1,6.	111
Tabla 16. Predicciones Modelo EST suavización exponencial.	113
Tabla 17. Predicciones Modelo Auto.Arima.	114
Tabla 18. Predicciones Modelo NNETAR	115
Tabla 19. Predicciones Modelo TBATS.....	117
Tabla 20. Comparación modelos serie temporales resultados Accuracy.....	118
Tabla 21. Comparación modelos serie temporales vs datos reales de unidades.....	119
Tabla 22. Datos de predicción Modelo STLM arima 3,1,6, - Mejores resultados.....	119
Tabla 23. Comparación Modelo STLM vs Datos reales.....	121
Tabla 24. Comparación Modelo STLF vs Datos reales.	122
Tabla 25. Comparación Modelo Holt Winters vs Datos reales	123

Tabla 26. Comparación Modelo STLM arima 3,1,6 vs Datos reales.....	124
Tabla 27. Comparación Modelo EST vs Datos reales.....	125
Tabla 28. Comparación Modelo Auto.Arima vs Datos reales.	126
Tabla 29. Comparación Modelo NNETAR vs Datos reales	127
Tabla 30. Comparación Modelo TBATS vs Datos reales	128
Tabla 31. Comparación general de modelos	129

ABREVIATURAS Y SÍMBOLOS

AEADE.- Asociación de Empresas Automotrices del Ecuador.

ICE.- Impuesto a los Consumos Especiales.

INEC.- Instituto Nacional de Estadísticas y Censos.

ANT.- Agencia Nacional de Transito.

SLR.- Systematic Literature Reviews – Revisión Sistemática de la Literatura.

BI.- Business Intelligence –Inteligencia de negocios.

IoT.- Internet Of Things - Internet de las cosas.

OLAP.- On-Line Analytical Processing - Procesamiento analítico en línea.

DM.- Data Mining – Minería de datos.

KDD.- Knowledge Discovery in Databases - Descubrimiento de conocimiento en bases de datos.

TI.- Tecnologías de la Información.

BD.- Base de Datos.

TDSP.- Team Data Science Process - Proceso de ciencia de datos en equipo.

RTD.- Random Decision Tree - Árbol de decisión aleatorio.

ID3.- Induction Decision Trees - Árboles de decisión de inducción.

CRISP-DM.- Cross Industry Standard Process for Data Mining - Proceso estándar de la industria para la minería de datos.

SPSS.- Producto de Estadística y Solución de Servicio.

ETL.- Extract, Transform and Load - Extracción-Transformación-Carga.

DS.- Data Science - Ciencia de Datos.

SVM.- Support vector machine - Máquinas de vectores soporte

ANOVA.- ANalysis Of VAriance

ANCOVA.- ANalysis of COVAriance

LDA.- Linear Discriminant Analysis

QDA.- Quadratic Discriminant Analysis

SEMMA.- Sample, Explore, Modify, Model, Assess

VCS.- Version Control System - Sistema de control de versiones

TFS.- Team Foundation Server

ONNX.- Open Neural Network Exchange - Intercambio de redes neuronales abiertas

ARIMA.- Autoregressive Integrated Moving Average - Modelo autorregresivo integrado de media móvil

STLM.- Descomposición estacional y de tendencias usando Loess con múltiples períodos estacionales

STLF.- Pronóstico de carga a corto plazo

EST.- Error, Tendencia, Estacional

NNETAR.- Pronósticos de Series de Tiempo de Redes Neuronales

TBATS.- Estacionalidad trigonométrica, Transformación Box-Cox, Errores ARMA, Componentes de tendencia y estacionales.

ME.- Mean Error - Error Medio

RSME.- Root Mean Squared Error - Error Cuadrático Medio

MAE.- Mean Absolute Error - Error Absoluto Medio

MAPE.- Mean Absolute Percentage Error - Error Porcentual Absoluto Medio.

MPE.- Mean Percentage Error - Error porcentual medio

MASE.- Mean Absolute Scaled Error - Error escalado absoluto medio.

ACF1.- Autocorrelation of errors at lag 1 - Autocorrelación de errores en el desfase 1.

INTRODUCCIÓN

En la actualidad, debido a la gran transformación de la era digital de los últimos años y a la tasa en la que crecieron los datos, la Inteligencia de Negocios (BI), Big Data y Minería de Datos, se han vuelto cada vez más importantes en las empresas; [1]-[2]. De este universo en expansión se derivan varias oportunidades para empresas comerciales para poder conocer a sus consumidores, al utilizar los datos como nuevas formas de hacer negocios.

El análisis de datos en los últimos años se viene utilizando con mayor frecuencia en las diferentes áreas para generar conocimiento, convirtiéndose en un componente principal en las investigaciones de disciplinas como *Big Data*, Internet de las cosas (*IoT*), Minería de Datos, aprendizaje automático, entre otras [3]. Las empresas comerciales por lo general se encargan de recopilar grandes volúmenes de datos multimodales, donde se incluyen las transacciones de sus clientes, gestión de ventas, gestión de inventarios, publicidades, relaciones con los clientes, preferencias y deseos de los clientes, datos financieros, etc. [4]. Todo esto para conseguir estabilidad a largo plazo y poseer una superioridad competitiva en los negocios. La ciencia de datos incluye: análisis estadístico, inteligencia de negocios (que a su vez incluye: informes, visualizaciones, procesamiento analítico en línea (OLAP), Minería de Datos), Aprendizaje Automático, entre otros [1]; con el fin de proporcionar soluciones y ayudar en la toma de decisiones. Debido al rápido crecimiento de las tecnologías, se han minimizado los tiempos de procesamiento de información de una manera significativa, por lo que es posible procesar grandes volúmenes de datos en poco tiempo[5], con el propósito de descubrir conocimiento, que por lo general está oculto en la data para ser utilizado en futuras acciones y decisiones [6].

Las técnicas de minería de datos se empiezan a conocer alrededor de la década de los sesenta, para ese entonces los estadísticos utilizaban a menudo los términos de *Data Fishing* (pesca de datos) *Data Mining DM* (procesamiento de datos) o *Data Archaeology* (arqueología de datos) con el objetivo de examinar relaciones fuera de un razonamiento previo de las bases de datos. A inicios de los años ochenta varios investigadores como Robert Blum, Gregory Piatetsky-Shapiro, Rakesh Agrawal, Gio Wiederhold entre otros comenzaron a fundamentar los términos de Minería de Datos y *KDD* (descubrimiento de conocimiento en bases de datos) [7]. Estas técnicas se convirtieron en un

buen instrumento para las personas que se desempeñan en ambiente de los negocios y el ámbito educativo. Las técnicas de minería de datos con el tiempo han evolucionado en sus herramientas, de las cuales se dividen en cuatro etapas primordiales que son: en el año 1960, las Bases de Datos; en 1980, el Acceso de Datos; en 1989-1990, los Almacenes de Datos para el apoyo de Decisiones; y, por último, como Minería de Datos Inteligente en transcurso de la década de 1990.

La Minería de Datos se centra en encontrar las relaciones y además de modelos integrales que se utilizan por lo general en bases de datos extensas los mismos que están ocultos a la infinidad de datos [7]. Se conoce que la Minería de Datos tiene relación directa con la Inteligencia de Negocios. BI abarca un grupo de estrategias y elementos para acceder a transformar los datos que se obtiene de Sistemas de Procesamiento Transaccional o formatos de diferentes fuentes, la arquitectura *Business Intelligence* es conformada de 3 componentes importantes principales que son: fuentes de datos, almacenes de datos y herramientas de análisis [8].

Las empresas están tomando un giro trascendental en sus negocios, para la comercialización de sus productos es posible predecir tendencias en los consumidores (ventas) [9]. El Análisis Predictivo es resultado de los nuevos conocimientos que se han incorporado con respecto al uso de los datos, el cual consiste en la facultad real de procesar y almacenar extensos volúmenes de datos. Gracias a los grandes avances generados por las TI que han permitido gestionar archivos con gran volumen de todo tipo de datos, los mismos que son aptos de ser analizados para las búsquedas de preferencias. El Análisis Predictivo requiere de técnicas y herramientas informáticas que permiten identificar patrones y tendencias, conocimiento que es útil para apoyar las decisiones empresariales.

Después de los antecedentes teóricos analizados, el **objeto** de estudio de este trabajo consiste en el análisis de las ventas del sector automotriz del Ecuador. El **campo** específico a tratar, es la Minería de Datos Predictiva aplicada al sector automotriz. Y, el **problema** que se aborda en este trabajo, es la necesidad de analizar el comportamiento de las ventas de vehículos a futuro, en base a los datos históricos. A continuación, se describe el contexto del problema.

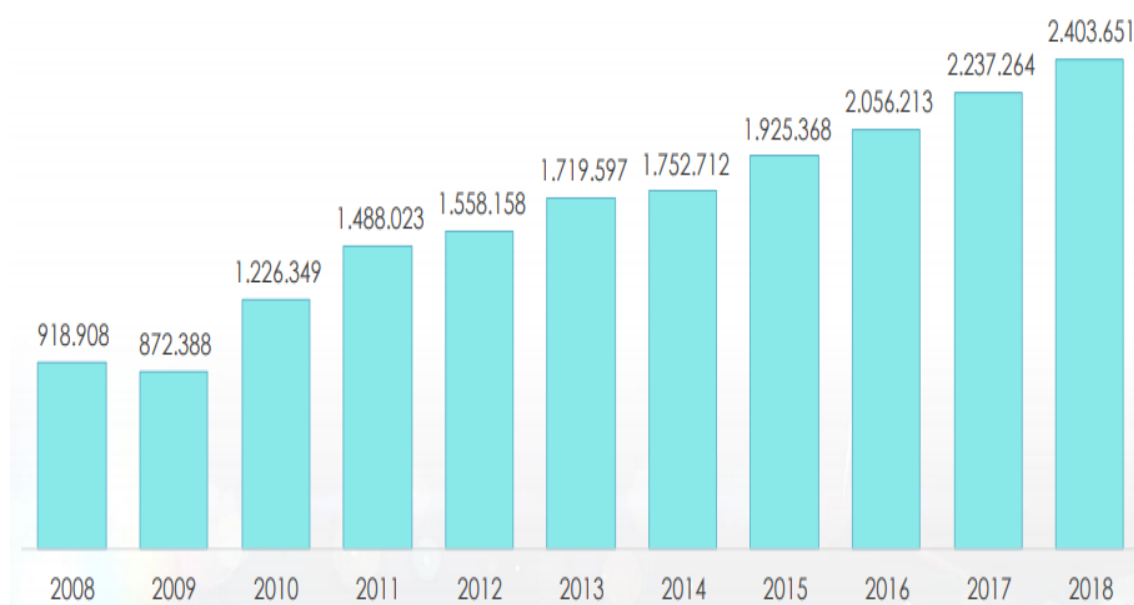
Actualmente el sector automotriz es uno de los que se ha visto más afectados por la emergencia sanitaria que atraviesa nuestro país y el mundo entero por los efectos causados

sobre la economía a consecuencia del virus SARS-CoV2 “Covid-19”. Por este motivo la Asociación de Empresas Automotrices del Ecuador (AEADE), ha presentado un plan para ayudar a su reactivación. La propuesta plantea ocho objetivos para fortalecer el sector que se está viendo muy afectado, la principal es promover el uso de un vehículo al considerarse un medio de transporte seguro y confiable para evitar el contagio de virus SARS-CoV2. También proponen una reforma en lo tributario que permita disminuir aranceles e impuestos como el ICE, automatizar los trámites de importación, adicional desarrollar sitios web para la comercialización de vehículos [10].

Gerardo Baldeón principal directivo de la AEADE, indica que el objetivo principal de su propuesta es levantar y fortalecer al sector automotriz, el mismo que hace énfasis en la necesidad de adquirir un vehículo propio a la ciudadanía, ya que se convierte en la opción más segura de movilización por la situación de la pandemia que se está atravesando. Según estadísticas en el Ecuador alrededor del 70% de los ciudadanos se movilizan en vehículos de transporte de uso público. Es importante recalcar que, en los 3 primeros meses de pandemia y emergencia sanitaria, el sector automotriz estima la pérdida 300 millones de dólares en ingresos y unos 890 millones de dólares en facturación. Debido que en los meses de mayor crisis (marzo, abril y mayo) la disminución en ventas fue de un 77,7% nunca antes visto en este sector. En estos meses solo se pudo facturar 7.180 vehículos a nivel nacional, mientras que en los mismos meses del año 2019 se vendieron 32.311 unidades. El gremio automotriz a inicios del 2020 estimó un descenso de por lo menos el 10%, pero por los factores y circunstancias que está atravesando nuestro país se proyecta que la caída del sector sea entre un 50 a 55% [11].

Por otro parte el crecimiento del parque automotor en el Ecuador ha sido considerable en los últimos 10 años, esto se debe en gran parte a la necesidad de movilización de las personas en ciudades grandes que son las que en mayor número mueven este sector. Según las estadísticas proporcionadas por el INEC, los cuales procesan información sobre la matriculación vehículos a nivel nacional, con datos proporcionados por la Agencia Nacional de Tránsito, en su último informe publicado con datos en referencia a información hasta el 2018. En la Figura 1 se muestra el número de vehículos matriculados registrados aumenta con el tiempo, incrementado en un 7,4% entre 2017 y 2018 [12].

Figura 1. Crecimiento parque automotor, Agencia Nacional de Transito 2008-2018



Fuente: Elaboración propia en base a los datos obtenidos de [12].

El parque automotor matriculado en Ecuador creció en casi de 1,4 millones de vehículos en una década, lo que situó la cifra por sobre los 2,4 millones de unidades a 2018, esto fue informado el viernes 1 de noviembre de 2019 por el INEC, información obtenida del Anuario Estadístico de Transporte [12].

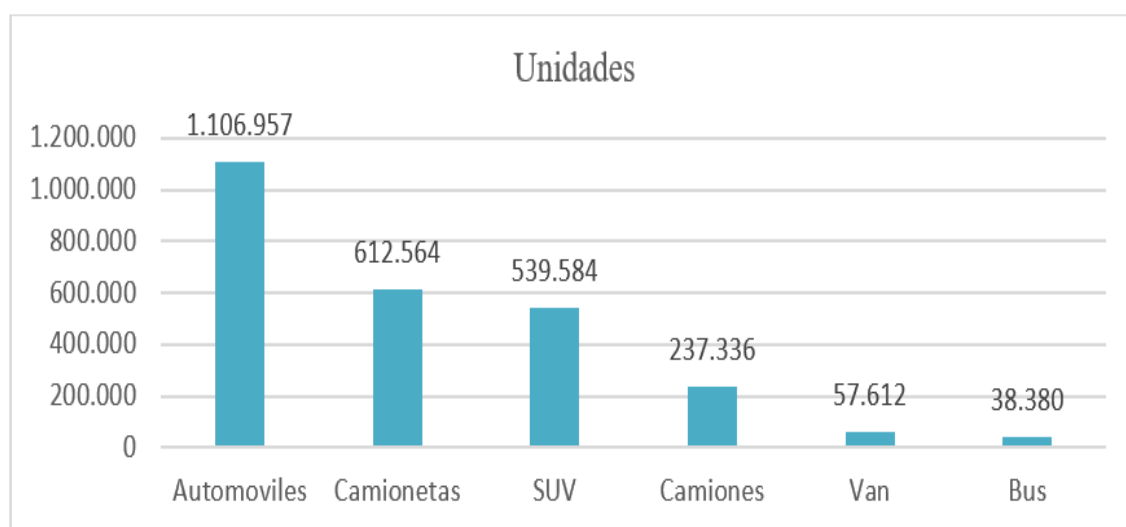
Actualmente el parque automotor en el Ecuador consta de 2.592.433 con información cortada a agosto 2020, divididos en sus segmentos como se observa en la Tabla 1 y Figura 2, esto según la AEADE [13].

Tabla 1. Vehículos por Segmento - agosto 2020

Segmentos	Unidades
Automóviles	1.106.957
Camionetas	612.564
SUV	539.584
Camiones	237.336
Van	57.612
Bus	38.380
Total	2.592.433

Fuente: Elaboración propia en base a los datos obtenidos de [13].

Figura 2. Vehículos por Segmento



Fuente: Elaboración propia en base a los datos obtenidos de [13], AEADE Boletín – Sector en cifras 48, corte a agosto 2020.

Luego de analizar el contexto del sector automotriz ecuatoriano, se formula la siguiente pregunta de investigación: ¿Cuál será el comportamiento (tendencia) de las ventas de vehículos en el Ecuador, durante los siguientes tres años (2021-2023)?

El objetivo principal de este trabajo es analizar los modelos de análisis predictivo para la búsqueda de tendencias en los segmentos de vehículos en el sector automotriz. Se estudió varios modelos predictivos de datos, luego se los evaluó en base a distintas métricas, siendo posible determinar el modelo más eficiente de predicción. Es decir, lo que se busca es implementar algoritmos para determinar la tendencia de crecimiento del parque automotor en el país en los próximos años, utilizando técnicas de análisis predictivo de datos, teniendo como referencia información histórica de ventas en unidades. Además, se requiere identificar la información y datos precisos que se necesitan para realizar predicciones por segmentos o provincias en el sector automotriz. Se estudió los modelos de análisis predictivo de datos, para analizar sus ventajas y falencias, y poder seleccionar el modelo de análisis predictivo más adecuado. Además, se propone aplicar el modelo más efectivo para realizar el análisis predictivo en el sector automotriz. Por último, se analizó y evaluó los resultados obtenidos con el análisis realizado.

Al poder determinar el crecimiento del parque automotor en el Ecuador, se podría establecer formas para disminuir la contaminación ambiental por el incremento de vehículos circulando en el país, también al llevar una estadística automatizada de la antigüedad de los vehículos se podría ayudar a establecer si es factible la chatarrización de vehículos en el Ecuador.

Se formuló la siguiente hipótesis: “El análisis de minería de datos predictivo de ventas de vehículos permitirá determinar tendencias que orienten la toma de decisiones en las empresas en el sector automotriz”. Se debe aplicar y validar en la empresa comerciales de venta de vehículos, al conocer las tendencias futuras en ventas puedan desarrollar estrategias de negocio y marketing para atraer a más clientes, por otro lado, también sirve de apoyo para la toma de decisiones en las empresas.

La información analizada en este trabajo se centra en sector automotriz del Ecuador; se obtuvo de los boletines mensuales de la AEADE. La variable dependiente tratada es la cantidad de vehículos vendidos por ciudad y marca en todo el Ecuador, y la variable independiente es el tiempo, el periodo analizado fue de enero 2007 a Julio 2020. Se realizó una exploración de los datos mediante estadística descriptiva para comprender sus características y para aplicar limpieza, transformación e imputación de datos con la finalidad de depurar la data y dejarla lista para posteriores análisis.

En esta investigación se utilizó TDSP [14], que es una metodología nueva, se presenta flexible y abierta; además, es ampliamente utilizada para analizar grandes volúmenes de datos y descubrir información valiosa. La metodología TDSP, consta de cinco fases: Comprensión empresarial, Adquisición y comprensión de datos, Modelado, Despliegue y Aceptación del cliente.

Al aplicar las técnicas de series temporales se realizó las predicciones a 36 meses, comprendidos de agosto 2020 a julio 2023; en las observaciones de los resultados que proporciono cada una de ellas, la técnica que genero el mínimo error en base a las métricas MAE y MAPE, es STLM Arima 3.1.6. Vale la pena señalar que, al realizar una comparación con los datos reales de las ventas obtenidas de agosto 2020 a marzo 2021, se encontró que el modelo que generó resultados más aproximados fue NNETAR con un 6.24% de diferencia.

En el presente documento está organizado de la siguiente forma, a continuación, se procede a detallar lo que abarca el proyecto tecnológico, en el **capítulo 1** se puntualiza la

parte del marco teórico referencia del trabajo con antecedentes: históricos, conceptuales y contextuales de las bibliografías analizadas, también conocido como el estado del arte. El **capítulo 2** se describe la metodología y materiales utilizados para realizar el trabajo. En el **capítulo 3** se presenta la propuesta metodológica de análisis de datos en el sector automotriz aplicación de técnica de minería de datos. En el **capítulo 4** se realiza la presentación de los resultados obtenidos del estudio realizado y discusión de los mismos, terminados con las conclusiones y recomendaciones.

CAPÍTULO 1

MARCO TEÓRICO REFERENCIAL

En este capítulo se presentan los antecedentes de la investigación, los cuales se usaron para estudiar y analizar investigaciones previas relacionadas, obteniendo como resultado la recopilación de definiciones de distintos autores. Se detalla el marco teórico científico, donde se muestran los conceptos, métodos y técnicas que se emplearon en este trabajo, además de los conceptos utilizados.

1.1 ANTECEDENTES HISTÓRICOS DE LA INVESTIGACIÓN.

En esta sección se describe información relevante para la presente investigación, además de aportar importantes conceptos, teorías e investigaciones prácticas en los estudios literarios, se relaciona con el entorno técnico que constituye la base del desarrollo de toda investigación [15].

1.1.1 PREGUNTAS DE INVESTIGACIÓN

Para el presente trabajo se formularon las siguientes preguntas de investigación:

RQ1 ¿Se aplica análisis predictivo en el sector automotriz en el Ecuador?

RQ2 ¿Qué utilizan las empresas del sector automotriz para determinar proyección de ventas?

RQ3 ¿Qué tipos técnicas de análisis predictivo son más eficiente?

RQ4 ¿Qué herramientas son las más eficientes para aplicar análisis predictivo?

RQ5 ¿Qué tipo de métodos y técnicas se utilizan en el análisis predictivo de datos?

1.1.2 OBJETIVOS

1.1.2.1 OBJETIVO GENERAL

Analizar segmentos de ventas de vehículos en el sector automotriz aplicando modelos de minería de datos predictivos para la determinación de tendencias que orienten la toma de decisiones.

1.1.2.2 OBJETIVOS ESPECÍFICOS

- Realizar búsqueda científica de trabajos relacionados para la elaboración del estado del arte.
- Identificar la información de segmentos de ventas en el sector automotriz de Ecuador.
- Explorar datos de ventas de vehículos, mediante estadística descriptiva básica y técnicas de pre procesamiento de datos.
- Seleccionar la metodología y técnica para realizar el análisis predictivo de ventas en el sector automotriz.
- Aplicar algoritmos de modelos predictivos de series temporales a los datos de unidades vendidas de vehículos para luego compararlos y seleccionar el más eficiente en generar tendencias de ventas futuras.

1.1.3 PROCESO DE BÚSQUEDA

La búsqueda de información científica para el presente trabajo se realizó en bases de datos relevantes y confiables, estas se han convertido en una alternativa importante para el desarrollo de investigaciones, esto forma parte del método científico. El proceso de búsqueda se realizó de acuerdo a las normas establecidas para el marco teórico, utilizando fuentes primarias y secundarias, de artículos científicos de revistas indexadas en bases de datos como Scielo, Web Of Science, EBSCO, ScienceDirect, IEEEExplore entre otras.

Dentro de los resultados encontrados se consideraron artículos científicos, libros y secciones de libros que hacen referencia al tema de investigación propuesto.

1.1.4 CRITERIO DE INCLUSIÓN Y EXCLUSIÓN

1.1.4.1 CRITERIO DE INCLUSIÓN

Se realizó búsqueda de información en inglés y español de los últimos años del periodo comprendido del 2016 al 2021, esto debido que en algunas bases de datos ya existe información 2021, utilizando varios términos para la búsqueda de información, los cuales se detallan a continuación:

- Término 1: Análisis Predictivo “*Predictive Analysis*”
- Término 2: Minería de Datos “*Data Mining*”
- Término 3: Inteligencia de Negocios “*Business Intelligence*”
- Término 4: Sector Automotriz “*Automotive Sector*”
- Término 5: Análisis de Regresión “*Regression Analysis*”
- Término 6: Árboles de Decisión “*Decisions Tree*”
- Término 7: Series Temporales “*Temporal Series*”

1.1.4.2 CRITERIO DE EXCLUSIÓN

Para esta investigación se excluye información anterior al año 2016, y artículos científicos que no estén publicados en revistas o congresos de alto impacto indexadas en las bases de datos según el reglamento establecido por la institución. También se excluyen artículos de opiniones, casos únicos y de editoriales no confiables.

1.1.5 GRUPO DE CONTROL

Se procedió a buscar en varias bases de datos más reconocidas como (Google Scholar, IEEEExplore, Web of Science, Elsevier, Springer, Science Direct,), para cada una de ellas se utilizó palabras claves.

Palabras claves: Predictive Analysis, Data Mining, Business Intelligence, Automotive Sector, Regression Analysis, Decisions Tree, Temporal Series.

Al obtener el resumen de las primeras búsquedas realizadas en cada una de las bases de datos, de aquí se parte y se debe realizar búsquedas más específicas y exactas para tener una base precisa de los antecedentes, y poder realizar la revisión sistemática de la literatura para elaborar el marco teórico y estado del arte.

Posterior a la búsqueda de información, se realizó la depuración de la misma eliminando artículos repetidos y seleccionando los que se tiene mayor aporte significativo e interés para la investigación.

Para la selección del material bibliográfico del presente trabajo de investigación se lo realizó aplicando los siguientes filtros:

- Por título del artículo.
- Por resumen.
- Lectura completa.

1.1.6 CADENA DE BÚSQUEDA

Para el análisis bibliográfico se ha utilizado la siguiente cadena de búsqueda en varias bases de datos, con las palabras claves en inglés ("***Predictive Analysis***" AND "***Data mining***" AND "***Business intelligence***" AND ("***Regression analysis***" OR "***Decisions Tree***" OR "***Temporal series***") OR "***Automotive Sector***"), en español se establece de la siguiente manera ("**Análisis predictivo**" AND "**Minería de datos**" AND "**Inteligencia Empresarial**" AND ("**Análisis de regresión**" OR "**Árbol de decisiones**" OR "**Serie Temporal**") OR "**Sector Automotriz**"), en un periodo de los ultimo cinco años comprendido entre 2016 y 2021. Se obtuvo como resultado de la cadena de búsqueda, 910 artículos en total, de los cuales se seleccionó los artículos más relevantes para esta investigación, y así poder determinar el estado del arte del presente documento.

En **Google Académico** se aplicó la cadena de búsqueda encontrando 236 artículos.

En **IEEEExplore** se utilizó la cadena antes expuesta y se encontraron aproximadamente 173 resultados.

En **Web of Science** se utilizó la siguiente cadena **TS=("**Predictive Analysis**" AND "**Data Mining**")** y se encontraron 36 documentos.

En **ScienceDirect** se encontramos 188 resultados utilizando la siguiente cadena de búsqueda: **Predictive Analysis AND Data Mining AND Business Intelligence AND Automotive AND (Regression analysis OR Decisions Tree OR Temporal series)**

En **Elsevier** se aplicó la cadena de búsqueda encontrado un total de 145 artículos de revistas en el área de ciencia de la computación.

En **Springer** se aplicó la siguiente cadena de búsqueda **Predictive Analysis AND Data Mining AND Business Intelligence AND Automotive AND (Regression analysis OR Decisions Tree OR Temporal series)** y se encontraron un total de 132 artículos.

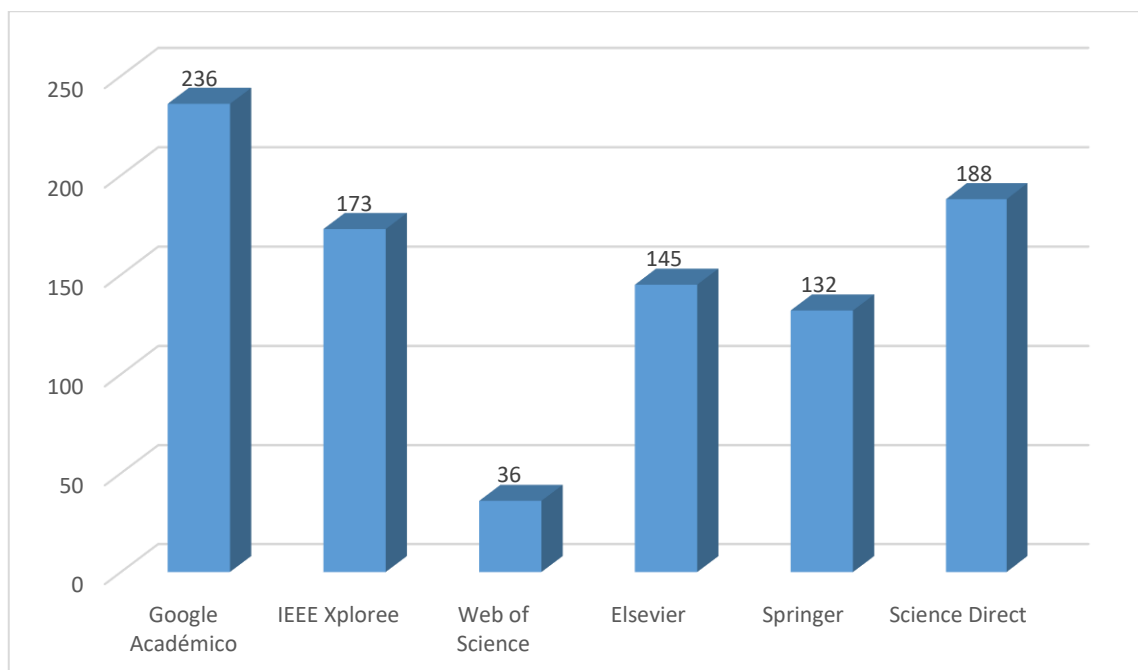
En la Tabla 2 y Figura 3 se detalla el número de artículos encontrados en algunas de las bases de datos consultadas.

Tabla 2. Búsqueda de Número de Artículos

Base de Datos	Número de Artículos
Google Académico	236
IEEEExplore	173
Web of Science	36
Elsevier	145
Springer	132
Science Direct	188

Fuente: Elaboración propia

Figura 3. Número de artículos encontrados por Base de Datos



Fuente: Elaboración propia

1.1.7 SELECCIÓN DE ESTUDIOS

Para realizar la selección de estudios se utilizó la metodología Revisiones Sistemáticas de Literatura (SLR) de Kitchenham [16], se escogieron los artículos que se consideraron que brindan un aporte importante a la presente investigación.

Como primer filtro se tomó en cuenta los artículos que se encuentren en revisas de bases de datos confiables y de renombre; luego a los artículos que pasaron el primer filtro se realizó una lectura analítica para determinar si es relevante y aporta a la investigación, de ellos se obtuvo las bases para el presente trabajo.

1.1.8 RESULTADOS DE LA REVISIÓN – ESTADO DEL ARTE

De los artículos analizados se determinó que es importante el manejo de análisis predictivo utilizando la minería de datos en las organizaciones e instituciones. Primero se requiere buscar los datos, seleccionar y organizar la información, para con ella realizar los análisis según el problema a resolver. Para el caso de estudio se analizó métodos y modelos de predicción de datos para aplicarlos en el sector automotriz con la finalidad de determinar la tendencia de consumo. A continuación, se detallan varios de los artículos analizados que aportan conocimiento a este trabajo.

La Minería de datos está involucrada en todas partes donde se encuentre flujos de información por esto en el artículo de Jayamalini y Ponnaivaikko [17], determinaron que la web es la base de datos más grande del universo, ya que contiene un gigantesco conjunto de datos de diversos tipos; estos datos son invaluable para las organizaciones si estos son extraídos de una manera eficaz. La minería web es un procedimiento cuyo objetivo principal es encontrar información y conocimiento beneficioso de los contenidos de las páginas web. En su investigación exponen cómo se realiza la extracción de datos para obtener conocimiento y el uso de los diversos tipos de métodos de la minería web. Además, indican que los contenidos web ofrecen nuevos desafíos para los algoritmos tradicionales de minería de datos, que por lo general funciona en datos planos.

Por otra parte, la minería de datos también está muy involucrada con datos de redes sociales, como se explica en el artículo de Stenberg et al [18], analizaron el compromiso de los clientes de Turkish Airlines con grandes datos sociales, en la cual su área de investigación se centró en aprovechar la Big Data Social, aprovechando la información generada de las redes sociales, para este caso específico, utilizaron la página de Facebook de la aerolínea Turkish; y descubrieron cómo se puede mejorar las estrategias de

marketing enfocadas a los clientes que manejan redes sociales. Manifiestan que la obtención de datos, está aumentando a un ritmo exponencial, primordialmente en publicaciones y comentarios en sitios de redes sociales. esto ayuda a las organizaciones a conocer el comportamiento de compras de los clientes, y a manejar la inteligencia de negocio para crear un marketing más efectivo dirigido a sus futuros compradores y fomentar su lealtad. Con esto la empresa puede mejorar sus ingresos en un corto plazo.

Otros autores también han utilizado el análisis predictivo con datos de redes sociales, como lo indican Desai y Patil [19], la predicción y categorización de datos de las redes sociales se ha vuelto tendencia en el área de investigación para ayudar a la recopilación de información y conocimiento empresarial . El análisis de los datos de las redes sociales se puede manejar para muchos propósitos, como recomendaciones de personas, recomendaciones de servicios o productos, entre otras cosas. En su investigación los autores concluyeron que el algoritmo de árbol de decisión aleatorio (RTD: Random Decision Tree, desarrollado por Breiman y Cutler en 1995), funciona mejor comparándolo con diferentes algoritmos de árboles de decisión (ID3 y C4.5 desarrollados por Quinlan en 1983 y 1993 respectivamente), siendo RTD el más conveniente y preciso para datos de las redes sociales, porque es flexible y proporciona resultados más precisos que otras técnicas. También identifican las diferentes fases de la minería de datos en redes sociales y sugieren la selección y análisis de datos como fase primordial.

La publicación realizada por Sanjay y Alamma [5], expresan una visión de analítica de big data, métodos y aplicaciones. Ellos indican que el desarrollo de la tecnología en los últimos años, ha permitido que los investigadores creen datos e información a un ritmo muy precipitado; siendo complicado recopilar los datos de forma eficaz. Después del descubrimiento de las computadoras, la selección, el almacenamiento y el procesamiento de datos se realizada de una forma tediosa. Por esto ellos discuten como la analítica de datos tradicional difiere de la analítica de big data, y, de igual forma las aplicaciones de big data con respecto a aplicaciones empresariales, científicas y sociales.

Para Sanjay y Alamma [5], el análisis de datos es un procedimiento de revisión, limpieza, transformación y modelado de los datos con el propósito de indicar comprensión, crear soluciones y favorecer a la toma de decisiones. También indican que cuando la información está disponible, se utilizan métodos de análisis de datos para encontrar información interesante y útil para las organizaciones. Se conoce que existen tres tipos de

análisis de datos que son: análisis descriptivo, análisis predictivo y análisis prescriptivo. El análisis de Big Data brinda a los científicos la posibilidad de manejar grandes volúmenes de datos de una manera más ágil, proporcionan el acceso a los mismos; además, el intercambio de información, para encontrar modelos o patrones previamente ocultos en los conjuntos de datos (data sets).

Según lo expuesto por Vilorio et al [20], señalan que en investigaciones realizadas en los últimos años las empresas que utilizan análisis predictivos obtuvieron un rendimiento promedio del 145%, en comparación con un promedio del 89% mostrado por otros métodos inteligentes no predictivos. Los autores propusieron generar estimaciones de ventas empleando modelos de análisis predictivos para que la producción cosmética; con la finalidad de obtener modelos que admitan generar pronósticos más confiables en diferentes categorías de productos, para facilitar el trabajo de marketing de la empresa al tomar decisiones que contribuyan a las futuras ventas, además, generaron dos modelos de análisis predictivos con algoritmos de minería de datos de los cuales escogieron el que ofrecía los mejores resultados con información confiables. Y a partir de los resultados obtenidos, se generaron campañas de ventas especiales.

Por otro parte, R. McCarthy et al [21], en su libro titulado “Introducción a la Analítica Predictiva”, definen el análisis de negocios, big data y análisis predictivo. Exponen las diferencias entre el análisis descriptivo y el análisis predictivo; para los autores, el análisis predictivo proporciona la capacidad de innovar la forma de cómo operan los negocios con la evolución de la tecnología. El análisis predictivo comprende fundamentalmente tres grandes técnicas que son: análisis de regresión, árboles de decisiones y redes neuronales. En su documento exponen algunos ejemplos de cómo se puede y debe utilizar el análisis predictivo en las organizaciones para fortalecer los negocios con el correcto manejo de la información que disponen las empresas.

El ejemplo más notable es del gigante de Amazon, que es el minorista número uno en ventas en línea, ellos utilizan en análisis predictivo para realizar marketing dirigido a sus usuarios, usando algoritmos que analizan las transacciones y búsquedas que realizan sus clientes para encontrar patrones de compras ocultos y relaciones entre productos. Recopilan datos de todo tipo: transacciones de compra, información de listas de deseos de sus clientes y productos más buscados y que revisaron los clientes, además poseen un sistema de recomendaciones donde indican a sus clientes que compraron un determinado

artículo, también puede adquirir tal artículo adicional o que por lo general son comprados juntos. Todo esto lo realizan usando algoritmos muy eficaces de análisis predictivo. Amazon fue el pionero y patentó su algoritmo predictivo que anticipa que es lo que el cliente comprará [7]. Otra empresa que utiliza estos métodos es Netflix, utiliza sistemas de recomendaciones a sus usuarios en base al comportamiento de visualización previo, para sugerir qué nuevas series o películas pueden ver [12].

Kambatla [8] indican que el análisis predictivo es utilizado en varios ámbitos en la actualidad, tales como en la medicina [22], estafas [23], comercialización e industria de producción. Por esto, esta técnica debe ser mayormente explotada para ayudar en las empresas a interpretar de mejor forma la información que poseen en sus historiales transaccionales y registros para ser transformados en conocimiento, y con ello aportar en la toma de decisiones efectivas y obtener ventajas competitivas. Las técnicas de modelado para realizar análisis predictivo se dividen en dos categorías principales que son: técnicas de regresión y técnicas de aprendizaje automático.

Los trabajos [21], [24]–[29] abordan conceptos primordiales como: preparación de la información, análisis predictivo y sus diferentes métodos, útiles para esta investigación.

R. McCarthy et al [24] en su documento “Know Your Data: Data Preparation”, indican que una vez que se identifican el problema comercial, se desarrollan las hipótesis y se recopilan los datos; el siguiente paso en el proceso es analizar los datos y prepararlos para el modelado predictivo. La mayoría de los datos sin procesar se consideran "sucios" o "ruidosos" porque los datos pueden tener información incompleta, información redundante, valores atípicos o errores. Por lo tanto, los datos deben analizarse y "limpiarse" antes del desarrollo del modelo.

Para R. McCarthy et al en su documento [25], consideran el Big 3 en análisis predictivo: regresión, árboles de decisión y redes neuronales. Sin embargo, estos no son los únicos métodos disponibles. Ya se han desarrollado otros métodos y su uso ha empezado a generalizarse. El panorama del análisis predictivo ha tenido un crecimiento significativo a medida que se presentan más oportunidades para aplicar estas técnicas en aplicaciones nuevas e interesantes. Las prácticas comerciales de antaño hicieron hincapié en dar al cliente lo que quiere y por esto las prácticas comerciales del mañana se centrarán en brindarle al cliente lo que necesita.

En el documento [26] de R. McCarthy et al, describen los modelos predictivos utilizando árboles de decisión. Los árboles de decisión son una herramienta importante del modelado predictivo, no por su complejidad sino por su simplicidad. Por lo general se usan para lograr facilitar un método fácil para establecer qué variables de entrada tienen un impacto importante en una variable objetivo.

Según R. McCarthy et al [27], los modelos predictivos que utilizan redes neuronales, son una de las técnicas de análisis predictivo más poderoso. La definición de red neuronal tiene más de cincuenta años; pero, es el avance reciente en la velocidad de cómputo, la memoria y el almacenamiento de datos lo que ha permitido su uso generalizado más actual. Además, se describe una variedad de arquitecturas de redes neuronales diferentes, y se presenta un análisis de cómo optimizar y evaluar las redes neuronales, seguido del uso de un árbol de decisión para mostrar cómo describir una red neuronal. Como parte final, aplican múltiples redes neuronales al conjunto de datos de seguros de automóviles para determinar qué red neuronal proporciona el modelo que mejor se ajusta.

En la publicación “Predictive Models Using Regression” de R. McCarthy et al [28], especifican los modelos predictivos que usan regresión, indican que después de efectuar un análisis descriptivo y de organizar los datos, el siguiente paso es montar el modelo predictivo. Los modelos de regresión pueden usarse como modelos predictivos. Los modelos de regresión populares incluyen regresión lineal, regresión logística, regresión con componentes principales y mínimos cuadrados parciales. Además, explican cuando es conveniente usar los diversos modelos de regresión. Se discuten los supuestos de regresión para cada tipo. Se examinan las métricas de evaluación para determinar el ajuste del modelo, incluidos R^2 , y R^2 ajustado y p-value. También discuten las técnicas de selección de variable (hacia adelante, hacia atrás y paso a paso) y el examen de coeficientes del modelo.

R. McCarthy et al [29], indican que antes de analizar cualquier conjunto de datos, es importante comprender primero los datos. Las estadísticas descriptivas presentan datos de una forma significativa con la intención de entender que se necesitará hacer con los datos para prepararlos para el análisis. Hay muchas pruebas estadísticas que se pueden utilizar. Este estudio se centra en una revisión del análisis estadístico descriptivo que se utiliza para preparar y respaldar el análisis predictivo. Analizaron los métodos para

garantizar que los datos estén preparados para el análisis, así como los métodos para limpiar o reducir variables para optimizar los resultados del análisis predictivo.

1.2 ANTECEDENTES CONCEPTUALES

1.2.1 HIPÓTESIS DE INVESTIGACIÓN

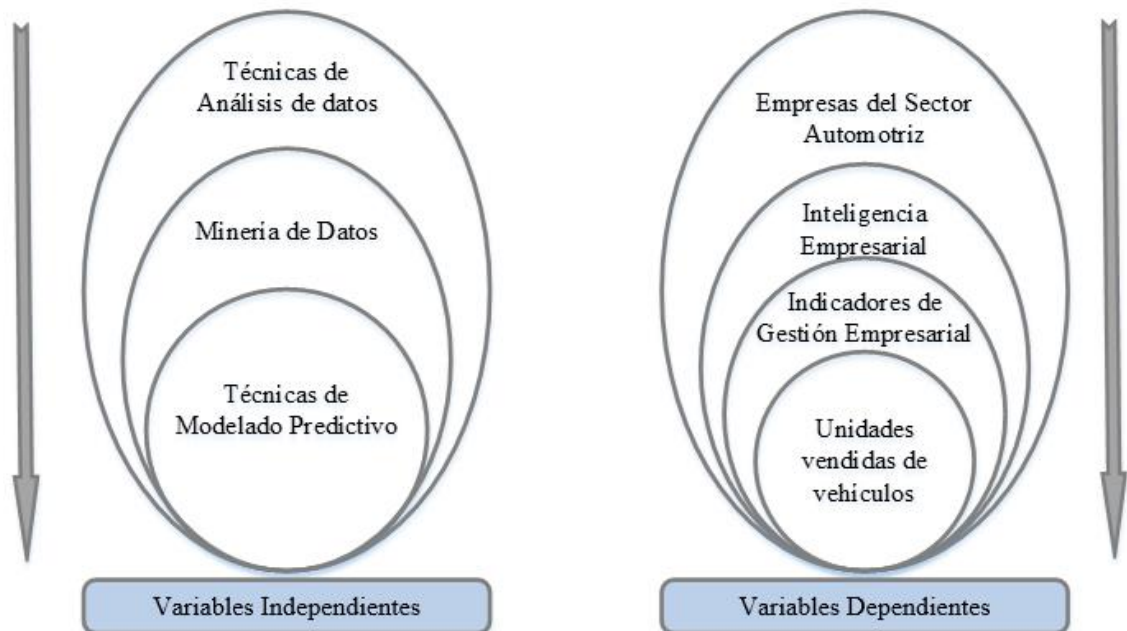
Para la presente investigación se planteó la siguiente hipótesis:

El análisis de minería de datos predictivo de ventas de vehículos permitirá determinar tendencias que orienten la toma de decisiones en las empresas en el sector automotriz.

1.2.2 RED DE CATEGORÍAS DE LAS VARIABLES

La Figura 4 muestra la red de categorías de las variables independientes y dependientes para la investigación.

Figura 4. Red de Categorías de las Variables



Fuete: Elaboración propia.

1.2.2.1 VARIABLES DEPENDIENTES

La predicción de tendencias en **Unidades vendidas de vehículos** a nivel nacional, permitirá a su vez, mejorar las oportunidades de negocios a las empresas comercializadoras de nuevos vehículos; debido a que se podrá entender y conocer el comportamiento del sector en los próximos años.

1.2.2.2 VARIABLES INDEPENDIENTES

Las técnicas de análisis predictivo permiten brindar información oportuna al sector automotriz en cuanto al crecimiento del parque automotor en el Ecuador. Se estudió algoritmos de minería de datos para determinar el más adecuado aplicando minería de datos como una técnica de análisis de datos.

1.2.3 FUNDAMENTACIÓN TEÓRICA DE LA VARIABLE INDEPENDIENTE

En esta sección detalla las bases teóricas conceptuales utilizadas para la presente investigación.

1.2.3.1 INTELIGENCIA DE NEGOCIOS

Inteligencia de negocios o BI se define como un conjunto de tecnologías para transformar datos en información significativa. Básicamente la expresión inteligencia empresarial tiene dos significados diferentes cuando se relaciona con la inteligencia. El primero es la inteligencia humana o la facultad de un cerebro normal aplicado a los asuntos comerciales. BI se ha convertido en una novedad, las aplicaciones del entendimiento humano y las nuevas tecnologías como la inteligencia artificial se utilizan para la gestión y la toma de decisiones en diferentes problemas relacionados con el negocio. El segundo es la información que ayuda a incrementar el capital en los negocios. El entendimiento inteligente adquirido por expertos y la tecnología eficaz en la gestión de negocios individuales y empresariales [30].

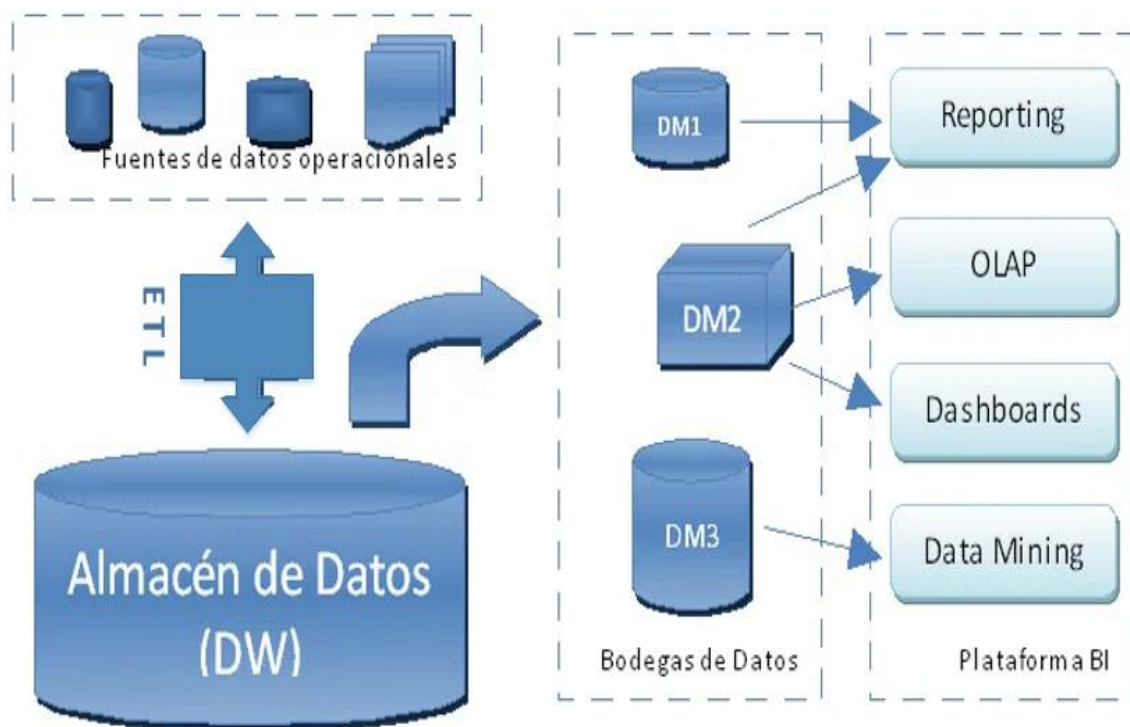
1.2.3.2 BI COMPONENTES

BI comprende un enfoque complejo y eficaz de los datos que influye en la toma de decisiones estratégicas correctas, y por lo tanto afecta la calidad de la organización, que debe tener una estructura bien organizada en cuanto a sus componentes [31], ver Figura 5.

Los componentes típicos de BI son:

- Fuentes de datos operacionales
- Procesos ETL
- Almacén de datos (Data Warehouse)
- Bodegas de datos (Data Mart)
- Plataforma BI
 - Reporting
 - OLAP
 - Dashboard
 - Data Mining

Figura 5. Componentes BI.

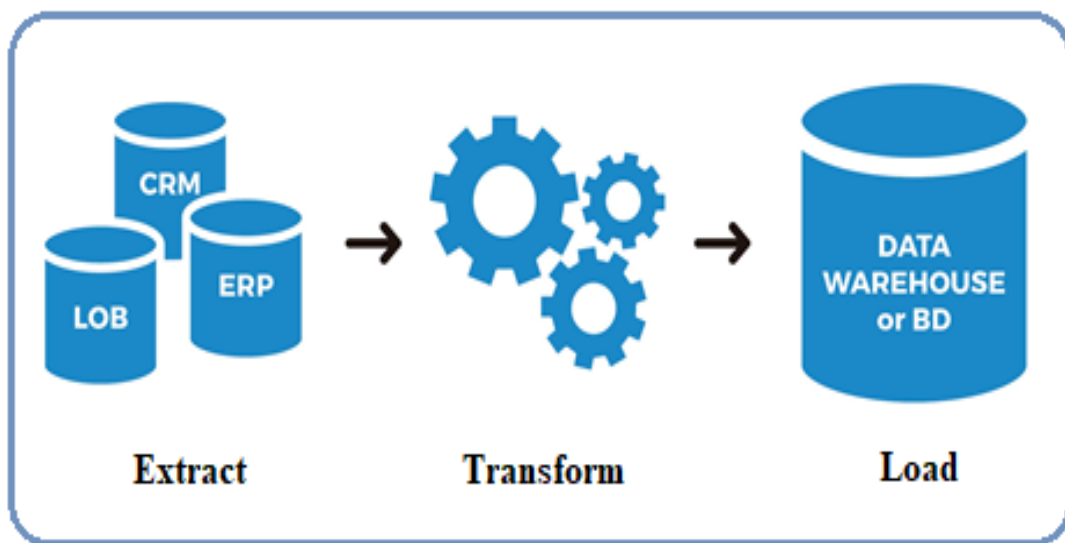


Fuente: Elaboración propia.

1.2.3.3 PROCESO ETL

El Proceso ETL, Extracción-Transformación-Carga: significa extraer datos de diferentes fuentes, transformarlos en un formato estándar y cargarlos en un repositorio de datos [32], en la Figura 6 se describe el proceso ETL.

Figura 6. Proceso ETL.



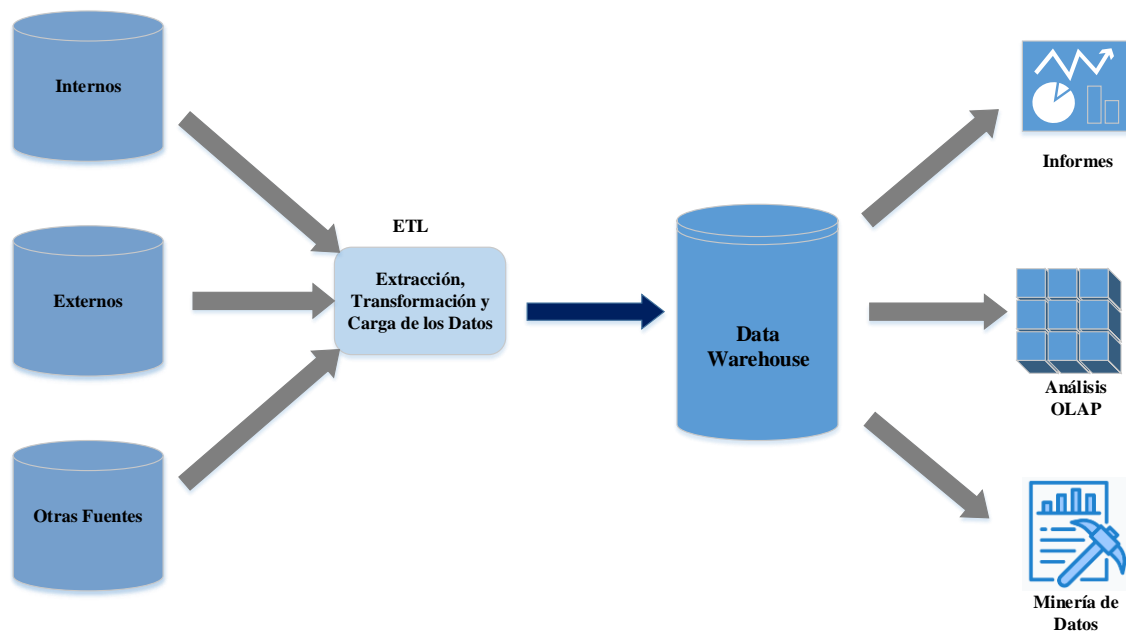
Fuente: Elaboración propia.

Las herramientas ETL se utilizan para transferir datos a almacenes físicos. Este proceso ETL comienza con “Extracción”, para extraer información de distintas fuentes de datos para posterior proceder a la “Transformación”, que consiste en limpiar, fusionar y convertir en el formato requerido por el almacén de datos y por ultimo “Carga”, la información en el Data Warehouse [32]. El proceso ETL es un factor importante en el éxito de un proyecto BI. La calidad de los datos cargados y almacenados en el *Data Warehouse*, afectan en gran manera en los resultados obtenidos para las tareas de análisis que se realicen con los datos por analistas de negocios [33].

1.2.3.4 DATA WAREHOUSE

Los Data Warehouse son repositorios de datos que almacenan grandes cantidades de información, para luego ser analizada de manera eficiente, gestionar información histórica y de manera adecuada [34]. Al utilizar esta tecnología se incrementa la oportunidad para consolidar información valiosa y a partir de ella generar conocimiento útil para las empresas [33], la arquitectura Data Warehouse se describe en la Figura 7.

Figura 7. Arquitectura Data Warehouse.



Fuente: Elaboración propia

1.2.3.5 DATA MART

Los Data Mart son un almacén de datos departamentales que se orientan a temas específicos. El análisis de datos en el Data Mart debe ser eficaz en la realización de la extracción de datos basada en los modelos de datos. Las tablas se almacenan como tablas de hechos y dimensiones. La tabla de hechos está en el centro de la estructura y posee el atributo de granularidad. La tabla de dimensiones es el conjunto de datos. Tanto la tabla de hechos como la tabla de dimensiones son para construir una estructura de organización multidimensional para formar el cubo de datos [35].

1.2.3.6 HERRAMIENTAS DE VISUALIZACIÓN

Las principales de herramienta de trabajo y visualización en analítica de datos, entre gratuitas y de pago se encontró las siguiente: SAP Business Intelligence, IBM Cognos, Tableau, SAS Business Intelligence, MicroStrategy, Domo, Sisense, Yellowfin BI, TIBCO Spotfire, Dundas BI, QlikView, Hevo Data, lear Analytics, Microsoft Power BI, Oracle BI [36].

1.2.3.7 CUADRANTE MÁGICO DE GARTNER.

Gartner Inc. es una empresa con sede en Stamford, Connecticut, Estados Unidos. Se dedica a la consultoría además de investigación de las tecnologías de la información. Según Gartner [36], en su Cuadrante Mágico 2020 indican que Microsoft es líder en Inteligencia de Negocios y Analítica; ya que dispone de una gama de productos. Microsoft con Power BI ofrece preparación de los datos, visualización en paneles interactivos., la versión Desktop, es una herramienta muy utilizada al ser gratuita para realizar análisis de datos de manera personal. Como se observa en la figura **Microsoft, Tableau, Qlik y ThoughtSpot** mantienen su posicionamiento privilegiado en el cuadrante de los líderes para el año 2020, como se observa en la Figura 8.

Figura 8. Cuadrante mágico de Gartner - Plataformas de análisis BI, 2020.



Fuente: Gartner [36]

Las plataformas de ciencia de datos y aprendizaje automático también poseen en espacio en esta consultoría, el cuadrante mágico de Gartner [37] ver Figura 9, en el cual los científicos de datos expertos y otros profesionales que trabajan en las áreas de ciencia de datos se guían para adquirir capacidades para obtener datos, construir modelos y poner en funcionamiento conocimientos de aprendizaje automático [38]. En el cual muestra una

visión en competencia y reflejan un mercado saludable que está madurando rápidamente. En cuanto a ciencia de datos y aprendizaje automático Gartner evalúa las plataformas de diversos proveedores definiendo una plataforma DSML como un producto central, esta plataforma es una mezcla entre funcionalidades básicas y avanzadas esenciales para crear soluciones DSML, principalmente modelos predictivos y prescriptivos. En primer lugar está la empresa Alteryx [39], líder en automatización de procesos analíticos, tiene su sede en Irvine, California, Estados Unidos la misma que ofrece cuatro productos de software que componen su plataforma DSML: Alteryx Connect, Alteryx Designer, Alteryx Server y Alteryx Promotion. Alteryx Designer es el producto principal,

Figura 9. Cuadrante mágico de Gartner DS y aprendizaje automático, 2020.



Fuente: Gartner [37]

1.2.3.8 ANALÍTICA DE DATOS.

La analítica de datos puede ser definido como el proceso de descubrir inferencias mediante el uso de métodos estadísticos, sistemas de software (minería de datos,

aprendizaje automático), visualización para identificar los modelos y patrones en los datos. El procedimiento de análisis consta de datos que pueden ser registros simples o una recopilación de archivos de datos en una base de datos. A medida que crecen los datos, se debe incrementar el tamaño de base de datos o tiene que ser almacenado en otros lugares [3]. En los últimos años, la analítica de datos ha sido utilizada en varios proyectos para descubrir respuesta a varias preguntas. La Analítica favorece a corroborar las siguientes preguntas y no referentes a la base de suponer ¿Qué pasó?, ¿Cómo o por qué sucedió?, ¿Qué está pasando ahora?, ¿Qué es probable que suceda después?; la analítica de datos se la clasifica en tres grandes grupos: Análisis descriptiva, Análisis predictivo y Análisis de perspectiva. Para este trabajo se utilizarán técnicas, modelos y algoritmos para análisis predictivo [4].

1.2.3.9 ANÁLISIS PREDICTIVO

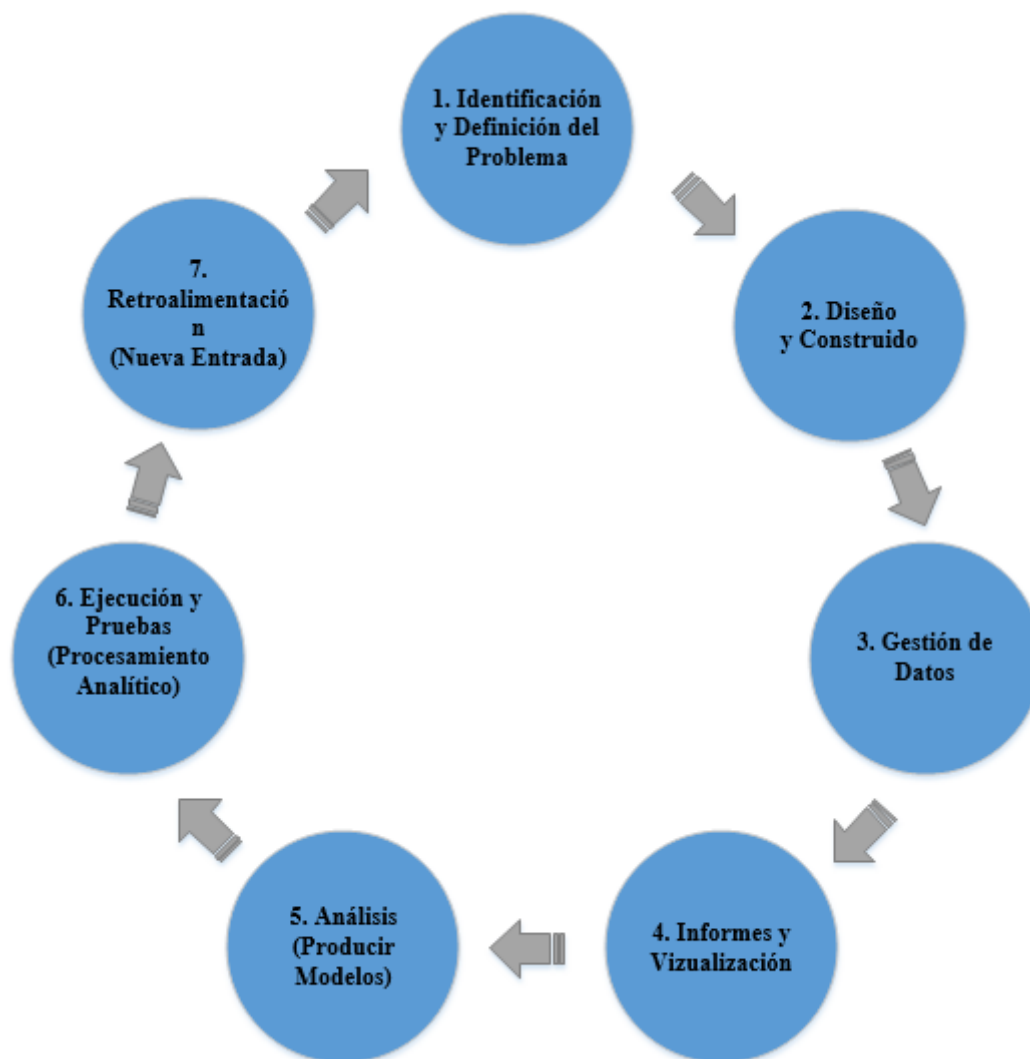
El análisis predictivo utiliza información o datos históricos para pronosticar eventos futuros. La pregunta principal para el análisis predictivo es ¿Qué pasara? El análisis predictivo utiliza estadísticas avanzadas además de otras técnicas de aprendizaje automático. Es indispensable que los datos históricos sean representativos de las tendencias futuras para que la analítica predictiva sea efectiva. Las técnicas de análisis predictivo se pueden majear para anunciar un valor, por ejemplo, se desea saber ¿Cuánto tiempo puede trabajar el motor de este avión antes de demandar mantenimiento o para considerar una posibilidad? O ¿Qué tan posible es que un cliente incumpla el pago de un préstamo hipotecario? Las técnicas de análisis predictivo además se pueden aplicar para seleccionar una categoría: ¿qué marca de zapatos comprará un cliente: Adidas, Nike, Puma o New Balance? El análisis predictivo utiliza algoritmos basados en datos para elaborar modelos. Los algoritmos llevan los datos a través de una serie de pasos para deducir los resultados predicativos. Evidentemente, los algoritmos de análisis predictivo automatizan el proceso de expresar ideas como modelos, patrones, tendencias en los datos [21].

1.2.3.10 CICLO DE VIDA DEL ANÁLISIS PREDICTIVO

El éxito de la analítica predictiva se basa más en una mejor toma de decisiones. Anteriormente, con un volumen de datos bajo, la toma de decisiones intuitiva aún era

exitosa. Pero debido a que el tamaño de los datos ha alcanzado proporciones extraordinarias en aumento, la capacidad humana para tomar decisiones intuitivas se ha reducido por completo. Como resultado, la toma de decisiones basada en datos se interpreta más como una garantía de un camino razonable para una mejor toma de decisiones. Estas decisiones impulsada por datos que por lo general se basa en modelos cuantitativos que se realizan mediante procesos de ciclo cerrado, comúnmente denominados ciclos [40].

Figura 10. Ciclo de vida del análisis predictivo.



Fuente: Elaboración propia a partir de [40]

En la Figura 10 se representa el ciclo de vida del análisis predictivo. Se puede observar que el ciclo de vida comienza con la identificación y definición del problema y luego continúa con el diseño y construcción de un marco analítico. Después de eso, pasa a las

etapas de gestión de datos, informes y visualización. Para más información, se llevan a cabo análisis para producir modelos, ejecución y pruebas. El ciclo termina con comentarios que luego se convierten en otra entrada para el próximo proyecto de análisis predictivo [40].

1.2.3.11 MINERÍA DE DATOS

Minería de datos, con el tiempo ha ido evolucionando en sus herramientas; se encontró cuatro etapas primordiales que son: en el año 1960, la colección de datos, en 1980 el acceso de datos, entre 1989-1990, los Almacenes de Datos para el apoyo de Decisiones y por último como Minería de Datos Inteligente en transcurso de la década de 1990 [9]. La Minería de Datos específicamente hace referencia al proceso de extraer información oculta interesante de Bases de Datos (BD) de las empresas o instituciones, que sería imposible conseguirlo de manera manual. Estas definiciones de variables se pueden adjudicar a la presentación de la expresión “Descubrimiento de Conocimiento en Bases de Datos” en el primer estudio de KDD en 1989 [7]. A partir de aquel tiempo, los investigadores y autores han relacionado KDD con Data Mining, y algunos afirma que uno y otro tiene significado similar, principalmente debido a que el conocimiento es el fruto obtenido de la Minería de Datos y su explotación[7]. La Minería de Datos no es más que un proceso de identificar información relevante obtenida de grandes volúmenes de datos, con el único objetivo de evidenciar patrones, modelos y tendencias. Estructurando la información proporcionada con un carácter comprensible para utilizarla posteriormente [41]. Minería de Datos y la Inteligencia de negocios son dos disciplinas de las ciencias de la computación que se centran en el análisis de datos y aportan en gran medida valor para la toma de decisiones [8], [42].

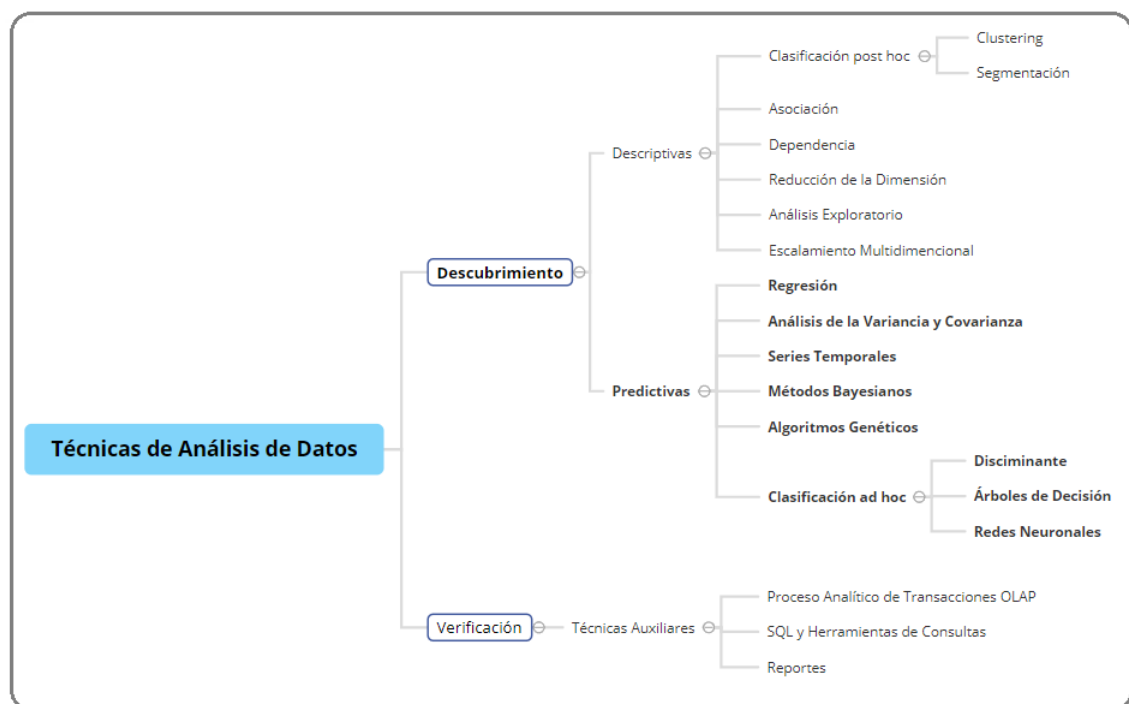
La Minería de Datos está conformada por un grupo de métodos y técnicas que brindan la posibilidad de estudiar grandes bases de datos, de forma automática o semiautomática, con el propósito de hallar modelos repetitivos que interpreten la conducta de estos datos. La minería de datos se creó con el objetivo de ayudar a interpretar grandes volúmenes de datos, y que estos pudieran ser analizados para extraer conocimiento y así contribuir en el progreso y desarrollo de los negocios, el mejoramiento de las ventas o fidelización de clientes. Además, los datos son el medio o la base para alcanzar las conclusiones de acuerdo al análisis realizado y transformar estos datos en información relevante, para que

las empresas puedan incorporar mejoras y soluciones que les ayuden a alcanzar sus objetivos [41].

1.2.3.12 TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de la minería de datos son algoritmos que provienen de la Estadística y de la Inteligencia Artificial. Estas técnicas poseen un nivel de sofisticación y que se aplican sobre un grupo de datos para obtener unos resultados. La clasificación inicial de las técnicas de minería de datos, distingue entre técnicas descriptivas, predictivas y auxiliares. En la Figura 11 se muestra la clasificación de las técnicas de minería de datos.

Figura 11. Clasificación de las técnicas de Minería de Datos



Fuente: Elaboración propia a partir de [43].

Las técnicas **descriptivas**, en este caso, todas las variables tienen el mismo estado, y el patrón se crea automáticamente de acuerdo con el cálculo; y, el proceso de combinar y clasificar objetos, se conoce como técnica de clasificación [43].

Las técnicas **predictivas**, este modelo está diseñado para procesar datos con conocimiento previo y las variables se pueden dividir en dependientes e independientes. Básicamente el producto utilizado en esta tecnología debe seguir los siguientes pasos:

Identificar el objetivo, indica qué reglas de datos se utilizan para encontrar el mejor producto que se ajusten a los datos. Evaluación, realiza el cálculo de los parámetros del producto seleccionado para los datos a un nivel aceptado. Diagnóstico, un método para comparar la calidad del diseño a realizar. Y, Predicción, utiliza estimaciones y evaluaciones aplicando el modelo identificado, para realizar predicciones de valores futuros de las variables [43].

Las técnicas **auxiliares** son herramientas de apoyo muy limitadas, se basan en modelos de estadística descriptiva, informes y consultas. Se enfocan por lo general en la verificación, además, OLAP es para realizar análisis multidimensional de la data [43].

Las técnicas descriptivas y predictivas están básicamente diseñadas para descubrir conocimiento basado en los datos [43].

1.2.3.13 ALGORITMOS DE MINERÍA DE DATOS

Los algoritmos de la minería de datos se aplican sobre un grupo de datos para obtener información. Entre las técnicas predictivas más representativas encontramos:

- Análisis de regresión.
- Análisis de varianza y covarianza.
- Series temporales.
- Métodos bayesianos.
- Algoritmos genéricos.
- Análisis discriminantes.
- Árboles de decisión.
- Redes neuronales.

1.2.3.14 ANÁLISIS DE REGRESIÓN

En minería de datos se utilizan una variedad de modelos para analizar datos, entre los cuales se tiene los tipos de modelo estándar de minería de datos que involucran los análisis de regresión normal, para predicción y la regresión logística para la clasificación [44].

El análisis de regresión se realiza a menudo en la minería de datos para estudiar o estimar la relación entre varios predictores; luego, usa el método de mínimos cuadrados para estimar el parámetro que mejor se ajusta. El modelo se verifica usando uno o más pruebas de hipótesis [45].

Regresión Lineal, es un modelo que predice una variable basada en otra utilizando un método estadístico. Consiste en crear una línea de mejor ajuste que minimice la suma de residuos al cuadrado. Se utiliza para establecer la relación entre dos variables o para encontrar una posible relación estadística entre las dos variables [45]. **Regresión de Optimización de Secuencia Mínima** [44]. Este algoritmo se encarga de resolver el problema de la programación cuadrática. El algoritmo continúa ejecutando el bucle de iteración para alcanzar la combinación de pares más prometedora para optimizar los pesos. El modelo también optimiza el pronóstico de series de tiempo al reducir el tiempo de ejecución de la operación; sin embargo el algoritmo Support Vector Machine (SVM), evita realizar programación cuadrática [45].

1.2.3.15 ANÁLISIS DE VARIANZA Y COVARIANZA

ANOVA son las siglas en ingles de ANalysis Of VAriance ANOVA agrupa modelos estadísticos y sus técnicas asociadas, donde la varianza es particionada en algunos componentes. Este tipo de análisis utiliza tipos de pruebas paramétricas y deben cumplirse una serie de supuestos para poder aplicarla [46].

El análisis de **covarianza** (ANCOVA) es un método que trata de dos o más variantes medidas, donde las variables independientes medibles no se encuentran a niveles predeterminados; también facilita analizar factores dentro de una población, como en un experimento factorial. ANCOVA utiliza conceptos tanto de ANOVA y de análisis de regresión, ANCOVA proviene de ANalysis of COVAriance [46].

1.2.3.16 SERIES TEMPORALES

Una serie temporal es una secuencia discreta de una función valorada y ordenada a lo largo del tiempo. La predicción de una serie temporal es útil para una variedad de aplicaciones del mundo real. Por ejemplo: la lluvia, la temperatura atmosférica, crecimiento de la población, para determinar las perspectivas futuras de los ciudadanos, el producto interno bruto y similares; donde los datos son medidos en un intervalo regular

de tiempo real. Estos datos sin procesar, que representan las variaciones temporales de la entidad en puntos de tiempo fijos, dentro de un determinado intervalo finito de tiempo, describen una serie de tiempo. La inclusión de intervalo finito de tiempo real en la definición de una serie de tiempo tiene sentido desde la perspectiva de su predicción en un momento no incluido en el conjunto de datos registrados Sin embargo, si la motivación de la serie es preservar solo los datos históricos, ignorando las predicciones futuristas, la restricción del intervalo fijo y finito de la noción de serie temporales puede omitir [47].

1.2.3.17 MÉTODOS BAYESIANOS

Los métodos Bayesianos son una técnica probabilística simple basada en la aplicación del teorema de Bayes con supuestos independientes entre variables, se asigna probabilidades a cada objeto para cada clase posible [48].

1.2.3.18 ANÁLISIS DISCRIMINANTES

El Análisis Discriminante se dividen en dos grupos que son los siguientes:

El **Análisis Discriminante Lineal** o *Linear Discriminant Analysis (LDA)* este es un método de clasificación controlado por variables elegibles, donde se sabe que dos o más grupos son preferidos y las nuevas observaciones se clasifican en uno según sus características. Usando la teoría bayesiana, LDA considera el valor de los predictores y estima la probabilidad de estas observaciones [49].

El **Análisis Discriminante cuadrático** o *Quadratic Discriminant Analysis (QDA)* es una alternativa al LDA cuando cada clase tiene su propia matriz de covarianza. Al igual que LDA, QDA también asume que las observaciones de cada clase siguen una distribución normal multivariante, así como también introduce las estimaciones de los parámetros en la ecuación del teorema de Bayes para obtener las predicciones [50].

1.2.3.19 ALGORITMO DE ÁRBOL DE DECISIÓN

Los árboles de decisión son modelos de minería de datos de clasificación, estos se basan en diagramas de construcción lógica realizado a partir de base de datos, que permite el análisis de registros que conlleven a facilitar decisiones útiles y lógicas al analista de datos [51]. Un árbol de decisiones es una herramienta de toma de decisiones que utiliza el diseño en forma de árbol de las posibles clases.

1.2.3.20 REDES NEURONALES

Las redes neuronales son una poderosa herramienta analítica que busca imitar las funciones del cerebro humano. Un componente clave es su capacidad para aprender de la experiencia. Dentro del cerebro, las neuronas son el componente que habilita la cognición y la inteligencia. El cerebro está compuesto por un sistema de neuronas que trabajan juntas para formar una unidad cohesiva. Las entradas llegan a cada neurona a través de una conexión llamada dendrita. Las dendritas transmiten su información a la neurona enviando neurotransmisores, a través de una brecha sináptica. Estos neurotransmisores excitan o inhiben la recepción neuronal. Si excitan la neurona receptora, esto se denomina activación de la neurona. Si inhiben la neurona, no se activa. También es importante tener en cuenta que la cantidad de neurotransmisores que se transmiten a través de la brecha sináptica determina la fuerza relativa de la conexión de cada dendrita con su correspondiente neurona [27].

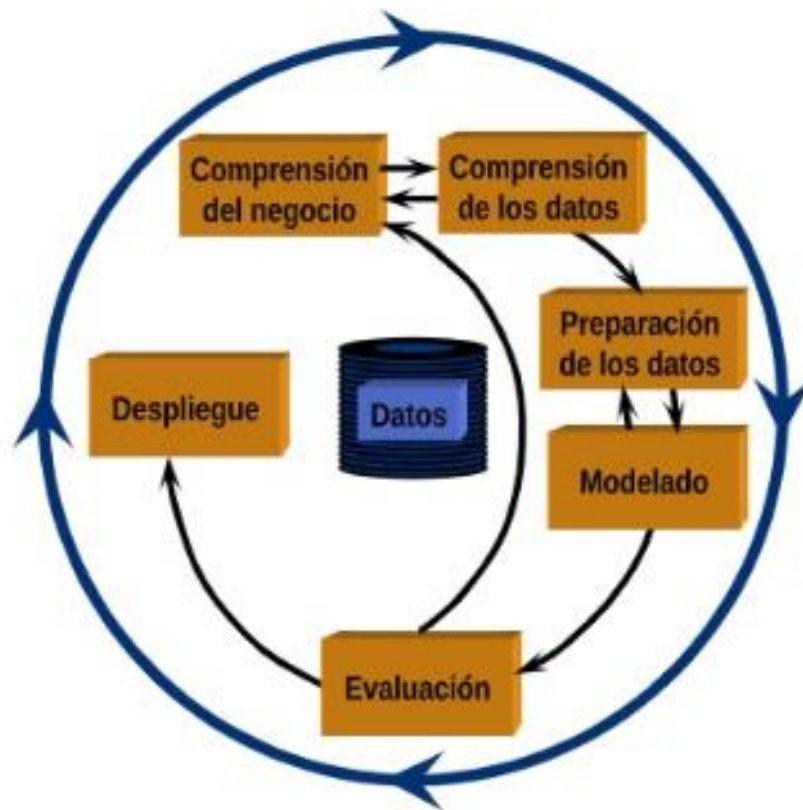
1.2.3.21 METODOLOGÍA CRISP-DM

CRISP-DM es la metodología más empleada para desarrollar las actividades de minería de datos en la educación y la industria. [52].

En la Figura 12 se ilustra el ciclo de vida de un proyecto de minería de datos el cual consta de fases que no son rígidas, permitiéndoles un movimiento bidireccional hacia adelante y hacia atrás entre las fases. Las flechas muestran las interacciones más importantes y frecuentes entre fases. Las seis fases del proceso CRISP-DM que se utilizan en cualquier proyecto de minería de datos, con la definición del objetivo del proyecto que se incluyen en cada una de las fases y estas son: comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación e implementación [52], [53].

El proceso estándar para la minería de datos CRISP-DM es un marco para traducir problemas comerciales en tareas de minería de datos y llevar a cabo proyectos de minería de datos independientes tanto del área de aplicación como de la tecnología utilizada [54]. Es una implementación orientada a la industria ampliamente adoptada del proceso genérico de descubrimiento de conocimientos KD [52].

Figura 12. Fases del modelo de proceso de la metodología CRISP-DM [55]



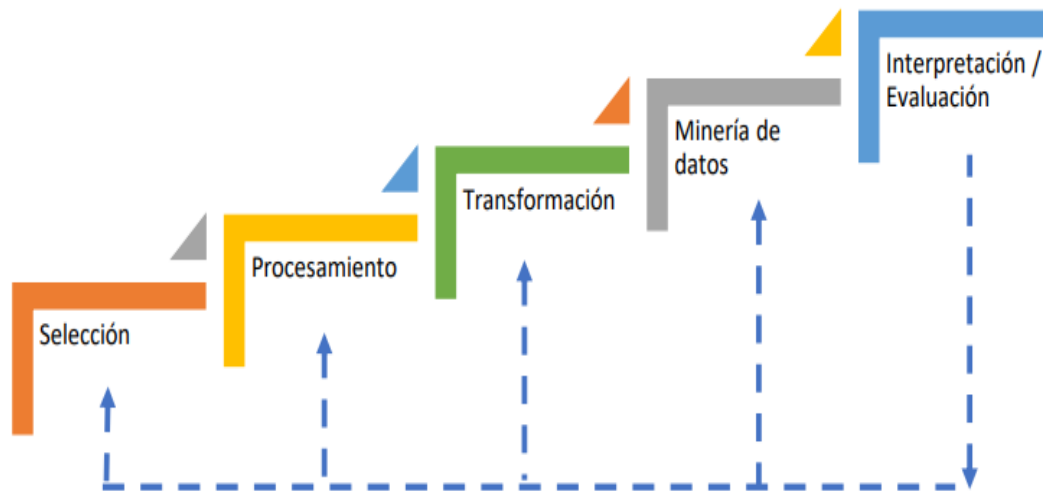
Fuente: C. García-Osorio [56]

1.2.3.22 METODOLOGÍA KDD

KDD es el proceso de usar métodos de minería de datos para extraer lo que se considera conocimiento según la especificación de medidas y entradas, utilizando una base de datos junto con cualquier pre procesamiento, submuestreo y transformación requeridos de la base de datos. Cuando hablamos de grandes cantidades de datos, el Descubrimiento de Conocimiento en Bases de Datos o KDD se refiere al proceso de identificar patrones válidos, novedosos, potencialmente útiles y principalmente entendibles.

A continuación, se describen las etapas de esta metodología, además se muestran en la Figura 13.

Figura 13. Etapas de KDD



Fuente: S. Huber et al, [52]

Selección, esta etapa consiste en crear un conjunto de datos de destino o centrarse en un subconjunto de variables o muestras de datos, en las que se realizará el descubrimiento.

Procesamiento, esta etapa consiste en la limpieza y el pre procesamiento de los datos de destino para obtener datos consistentes.

Transformación, esta etapa consiste en la transformación de los datos utilizando la dimensionalidad, métodos de reducción o transformación.

Minería de datos, esta etapa consiste en la búsqueda de patrones de interés.

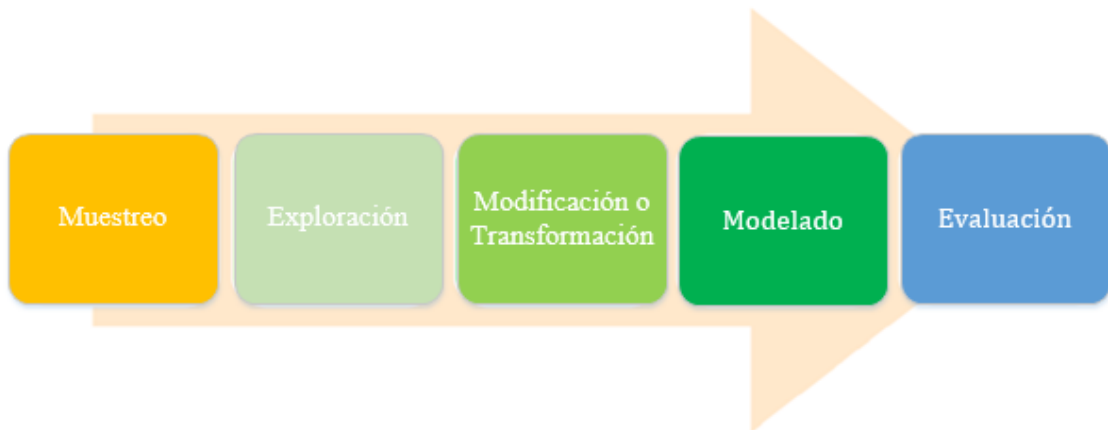
Interpretación / Evaluación, esta etapa consiste en la interpretación y evaluación de los patrones.

El proceso de KDD es interactivo e iterativo, involucra numerosos pasos y se toman muchas decisiones por el usuario.

1.2.3.23 METODOLOGÍA SEMMA

La metodología SEMMA fue desarrollado por el Instituto SAS. El acrónimo SEMMA significa Muestreo, Exploración, Modificación o Transformación, Modelado y Evaluación; se considera un ciclo con 5 etapas para el proceso como se muestra en la Figura 14.

Figura 14. Fases de SEMMA



Fuente: S. Huber et al, [52]

Muestreo, esta etapa consiste en muestrear los datos mediante la extracción de una parte de un gran conjunto de datos, suficiente para contener la información significativa, pero lo suficientemente pequeña para manipular rápidamente.

Exploración, esta etapa consiste en la exploración de los datos mediante la búsqueda de tendencias no anticipadas y anomalías para ganar comprensión e ideas.

Modificación o Transformación, esta etapa consiste en la modificación de los datos mediante la creación, selección y transformación de las variables para enfocar el proceso de selección del modelo.

Modelado, esta etapa consiste en modelar los datos permitiendo que el software busque automáticamente una combinación de datos que predice de manera confiable un resultado deseado.

Evaluación, esta etapa consiste en evaluar los datos mediante la evaluación de la utilidad y confiabilidad de los hallazgos del proceso de extracción de datos y estimar qué tan bien funciona un modelo.

1.2.3.24 METODOLOGÍA TDSP

El proceso de ciencia de datos en equipo (TDSP) es una metodología de ciencia de datos ágil e iterativa que ofrece soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. Utiliza una combinación de Scrum y CRISP-DM [52].

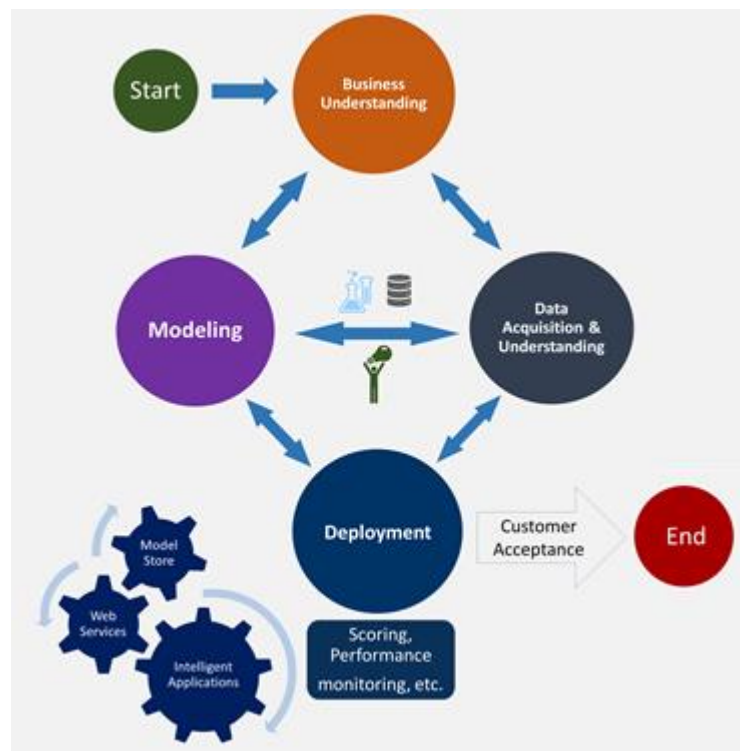
TDSP ayuda a mejorar la colaboración y el aprendizaje del equipo al sugerir cómo los roles del equipo funcionan mejor juntos. TDSP incluye las mejores prácticas y estructuras de Microsoft y otros líderes de la industria para ayudar a lograr una implementación exitosa de las iniciativas de ciencia de datos. El objetivo es ayudar a las empresas a aprovechar plenamente los beneficios de su programa de análisis.

Componentes de TDSP

En la metodología TDSP encontramos los siguientes componentes [14]:

- Ciclo de vida de la ciencia de datos, como se muestra en la Figura 15.
- Estructura de proyecto estandarizada.
- Infraestructura y recursos.
- Herramientas y utilidades.

Figura 15. Ciclo de vida metodología TDSP.



Fuente: Adatado de [14].

1.2.4 FUNDAMENTACIÓN TEÓRICA DE LA VARIABLE DEPENDIENTE

Todo lo que se puede medir, se puede mejorar, esto se puede conseguir al tener una buena estructura para realizar el proceso de análisis de datos. La captura de datos es el primer

paso que se realiza, estos deben ser transformados en información que permita tomar decisiones para mejorar el desempeño de las organizaciones. En segundo lugar, la información debe ser clara sobre los objetivos estratégicos de la organización, con esto se puede realizar mediciones que lleven a la institución a procesos de mejora continuos en el tiempo.

Los Indicadores clave de desempeño (KPI: Key Performance Indicator) son una forma de medir si una empresa, un empleado o departamento, están logrando sus metas y objetivos. En general, las organizaciones utilizan métricas de gestión empresarial en varios niveles, para evaluar el desempeño y el éxito previamente definidos en un marco estratégico, evaluar los resultados y tomar las mejores decisiones en el ámbito empresarial. Los indicadores de alto nivel se enfocan en el rendimiento general de la organización [1]. La inteligencia empresarial también conocida como inteligencia de negocios (BI), combina varios aspectos entre ellos: análisis de negocios, minería de datos, visualización de datos, herramientas e infraestructura de datos, y estas prácticas son recomendadas para ayudar a las organizaciones a tomar decisiones basadas en los datos e información que poseen [2].

Las empresas del sector automotriz en este último año se están viendo afectadas por el decrecimiento en ventas siendo un periodo atípico por la situación de emergencia sanitaria que está pasando el mundo entero y el Ecuador no está excepto por la pandemia causado por el SARS-CoV-2. Por esto se cree conveniente realizar el presente estudio y poder analizar el comportamiento del KPI: “Ventas del mercado automotriz” en el futuro y con esto considerar la nueva realidad de este sector.

1.3 ANTECEDENTES CONTEXTUALES

1.3.1 DELIMITACIÓN DEL CONTEXTO DEL ESTUDIO

Para Viloria [14], nos dice que las predicciones de ventas aportan de manera muy importantes a las empresas, con la finalidad de obtener capital de inversión y pronosticar el rendimiento a corto y largo plazo. Por lo que les permite anunciar los ingresos por ventas y determinar de modo efectivo los recursos que permiten al analista de negocios pronosticar el crecimiento futuro. La predicción de ventas se fundamenta en muchos supuestos, algunos de ellos conectados con el cliente y otros supuestos

conectados con el comportamiento de los competidores, también como con la naturaleza del mercado [58].

Con el Análisis Predictivo se pretende descubrir las tendencias o comportamiento del mercado automotriz. Por lo expuesto, al implementar el análisis predictivo e inteligencia de negocios se busca mejorar las ventas y optimizar las compras del sector automotriz y optimizar su flujo de información.

Con esta investigación se quiere aportar con un granito de arena al sector automotriz en el Ecuador, al realizar minería de datos y explorar sus técnicas y algoritmos para escoger y proponer la más eficiente. Se cuenta con información de venta de vehículos, de los últimos 14 años a nivel nacional para realizar todos los análisis pertinentes para esta investigación.

1.3.2 PROPUESTA DE SOLUCIÓN Y CONTRIBUCIONES

La propuesta de solución consiste en aplicar varios modelos predictivos que muestren la tendencia del crecimiento del parque automotor en el país en los próximos 36 meses, en base a información histórica de ventas de vehículos nuevos en concesionarios a nivel de todo el Ecuador. Luego, seleccionar un modelo de análisis predictivo con el cual se obtengan los resultados más eficaces para predicción de datos en el sector automotriz, y con esto las empresas comercializadoras de vehículo nuevos tengan una idea de cómo se comportarán las ventas, además de poder determinar si crecerán o decrecerán en los próximos años.

CAPÍTULO 2

MATERIALES Y MÉTODOS

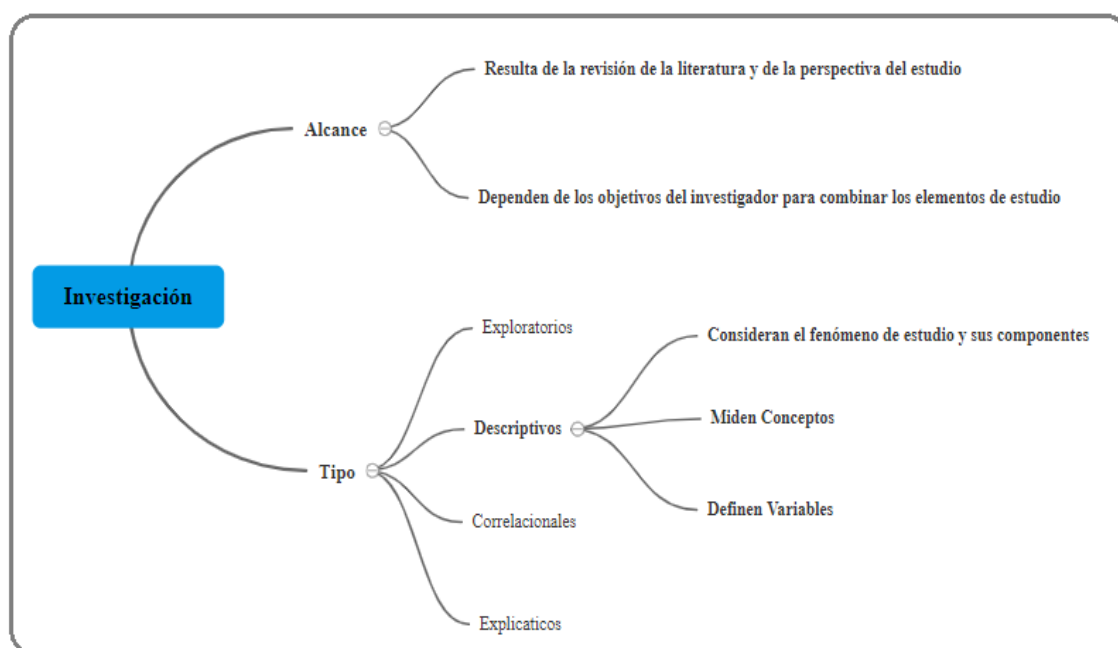
En esta sección, se especifica la metodología y los materiales que se utilizaron en la ejecución de la investigación; se describe el tipo de estudio, el enfoque en el cual se realizó, el cálculo de la población y muestra, los métodos teóricos y empíricos con los materiales utilizados, y por ultimo las técnicas estadísticas para el procesamiento de los datos obtenidos. Para la propuesta de un modelo de análisis predictivo aplicado al sector automotriz, utilizando las técnica, algoritmo y herramienta analizadas y seleccionadas en el Capítulo 1.

2.1 TIPO DE ESTUDIO O INVESTIGACIÓN REALIZADA

Luego haber realizado el estudio de la literatura en publicaciones y libros en las principales bases de datos de búsqueda de información científica, se determina que este trabajo es importante y puede ser elaborado. El siguiente paso es definir y visualizar el alcance que tendrá el presente trabajo por ende se debe seleccionar en qué tipo de investigación encaja [35].

El tipo de estudio que se utilizó para este trabajo es: la Investigación Descriptiva, la misma que busca determinar características y propiedades importantes en el fenómeno que se va analizar, además de describir tendencias de un grupo o población. Estos estudios son útiles para analizar cómo es y cómo se manifiesta un fenómeno y sus componentes, miden conceptos y se definen variables, como se muestra en la Figura 16.

Figura 16. Alcance y tipo de Investigación Descriptiva.

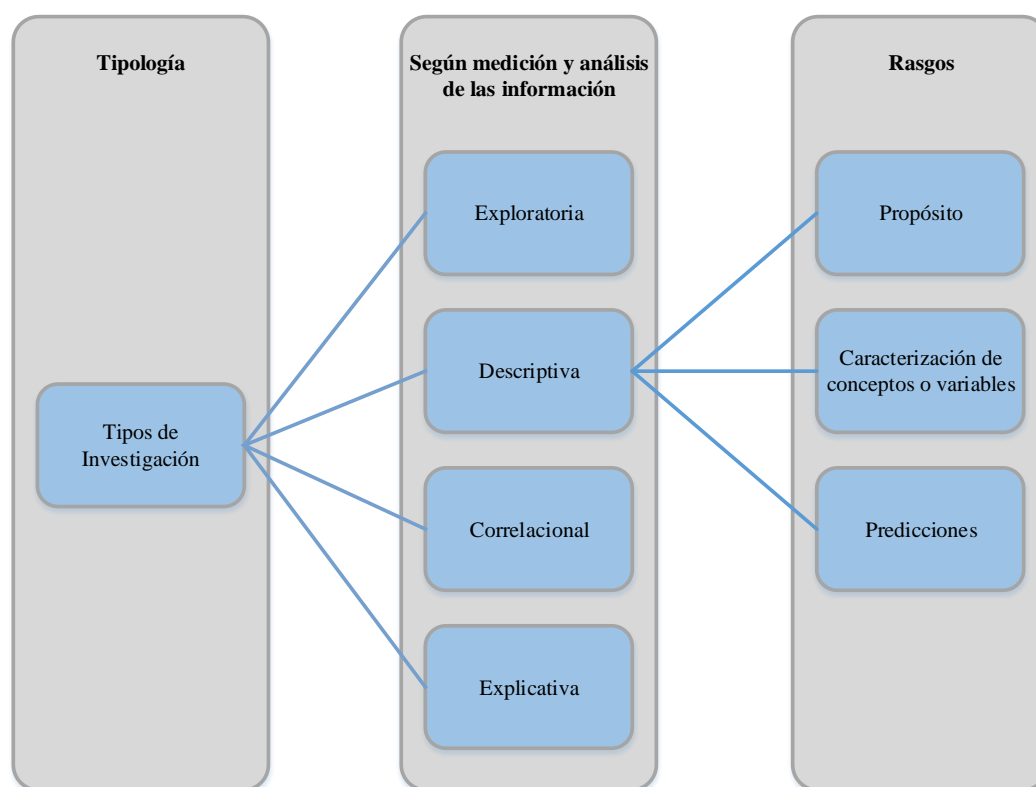


Fuente: Adaptado de [59]

La investigación descriptiva, por lo general describe los datos, los mismos que tienen un impacto en las vidas de las personas que le rodea. El principal objetivo de la investigación descriptiva consiste en poder conocer las situaciones, costumbres y actitudes predominantes a través de la representación exacta de las actividades, objetos, procesos y personas. Su propósito no se limita a la recolección de datos, sino a la predicción e identificación de las relaciones que existen entre dos o más variables, ver Figura 17. Es utilizada por los investigadores para recoger los datos sobre la base de la hipótesis planteada, exponer y resumir la información de manera metódica, para luego analizar minuciosamente los resultados, a fin de extraer generalizaciones significativas que contribuyan al conocimiento [59]. Por estas razones se aplica el tipo de investigación descriptiva para este trabajo.

Por otra parte, la investigación descriptiva con análisis predictivo se caracteriza por anticipar situaciones futuras. Los estudios de preferencia y más encontrados en este tipo de investigaciones son escenarios que se llevan a cabo en las áreas como economía, medicina, planificación, ingeniería, entre otras.

Figura 17. Rasgos de la Investigación Descriptiva.



Fuente: Elaboración propia.

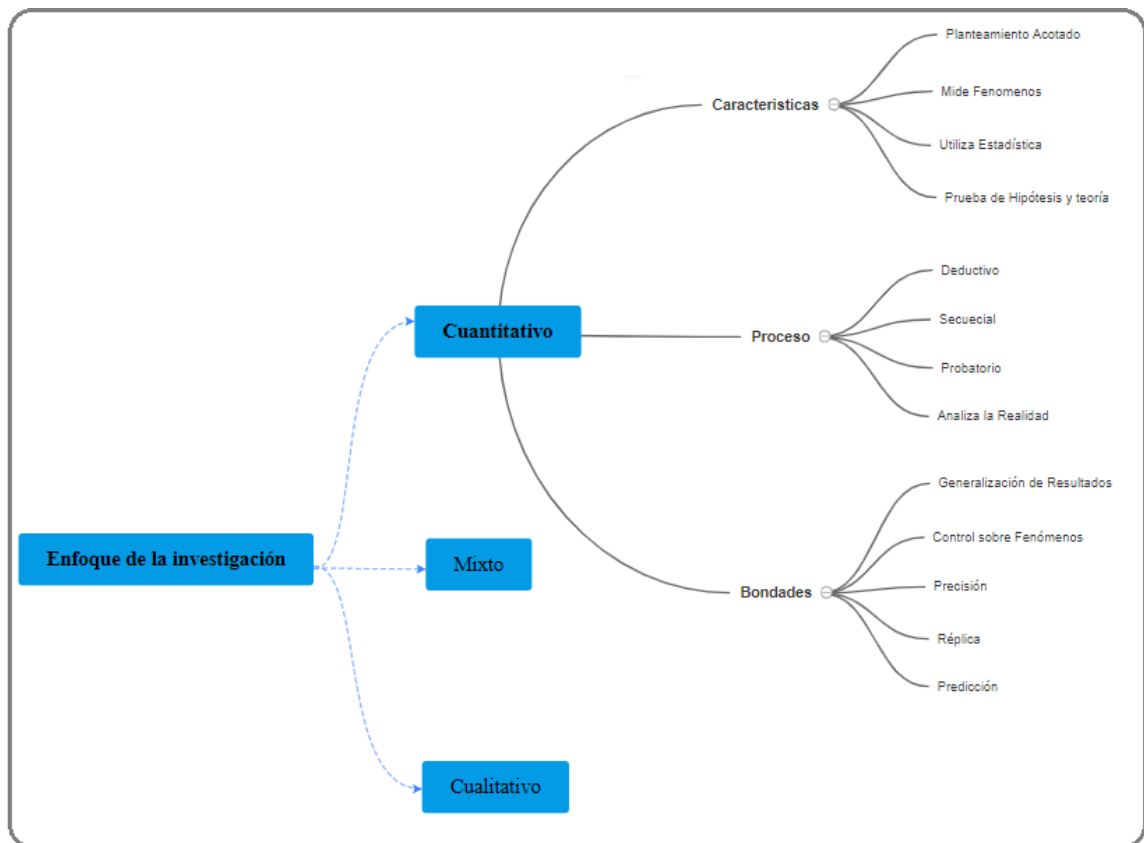
2.2 EL PARADIGMA O ENFOQUE DEL TRABAJO

El paradigma que se determinó para el presente trabajo de investigación, basado en el libro de metodología de investigación según R. Hernández Sampieri [59], se lo define como Cuantitativo – Cuasi Experimental.

2.2.1 CUANTITATIVO

El enfoque cuantitativo utiliza la recopilación de datos para evidenciar una hipótesis, basándose en la medición numérica y también en el análisis estadístico, con el fin de establecer patrones de comportamiento y certificar teorías. El enfoque cuantitativo comprende un conjunto de procesos: deductivo, secuencial, probatorio y análisis de la realidad objetiva. Entre las bondades están: la generalización del resultado, control sobre fenómenos, precisión, replica y predicción, ver Figura 18.

Figura 18. Enfoque Investigación Cuantitativo.

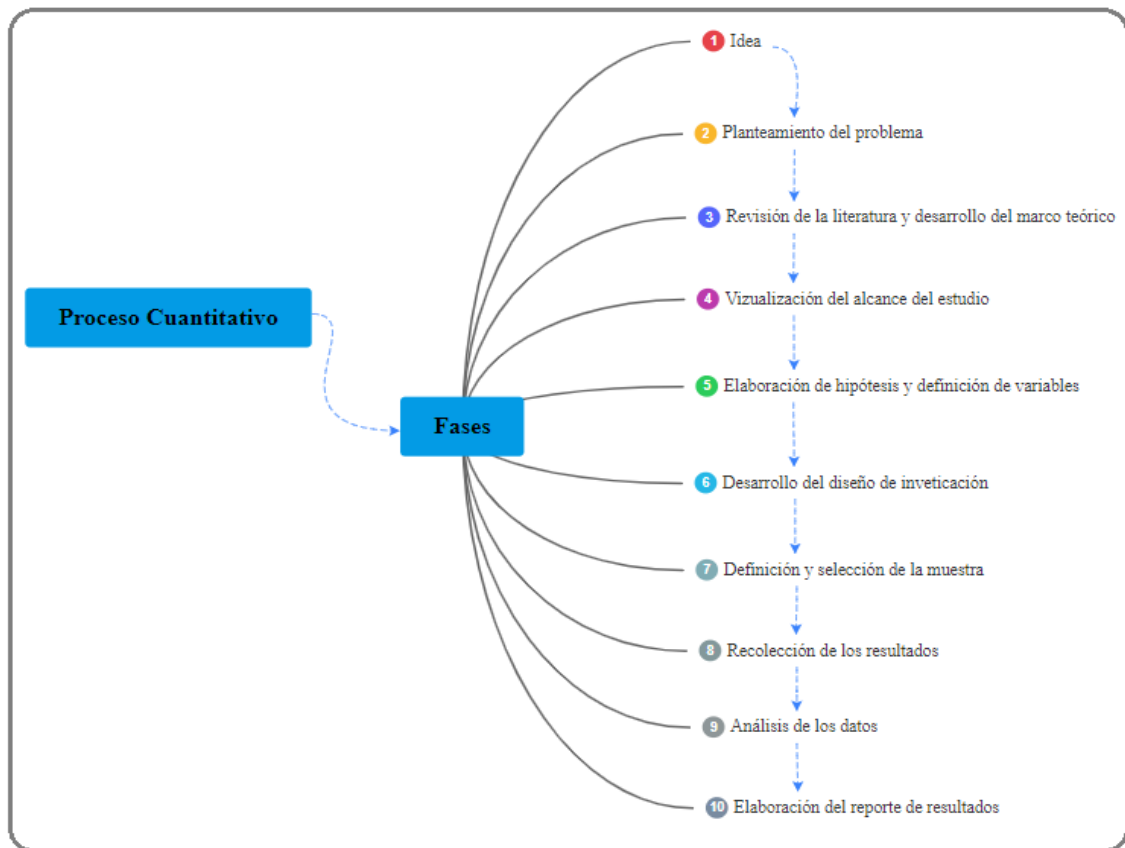


Fuente: Adaptado de [59]

Cada una de las etapas precede a la siguiente y no se puede saltar o eludir alguna fase, su orden se lo debe de seguir de manera rigurosa. Se parte de la idea principal luego se delimita, para derivar en los objetivos y preguntas de investigación, posterior se revisa la literatura y con ella se elabora el marco teórico, esto se realizó en el capítulo anterior. Posteriormente de las preguntas de investigación se establece la hipótesis y se determinan las variables, para esto se debe crear una estrategia para probarlas, también se examinan las mediciones que se obtienen utilizando métodos de estadística básica para finalmente obtener algunas conclusiones con respecto a la hipótesis planteado en un principio.

Las fases que se deben seguir en el proceso cuantitativo con las siguientes tal como se detallan en la Figura 19.

Figura 19. Fases de proceso Cuantitativo.



Fuente: Adaptado de [59]

2.2.2 CUASI EXPERIMENTAL

Una vez definido el planteamiento del problema, se debe definir el alcance inicial de la investigación y la formulación de la hipótesis; posteriormente se visualiza la manera práctica de contestar las preguntas de investigación y se debe cumplir con los objetivos que se fijaron al principio. Para esto se debe desarrollar uno o varios diseños de investigación y aplicarlos al trabajo planteado, y con esto poder obtener la información que se requiere para responder al planteamiento del problema. Para el enfoque cuantitativo, se debe utilizar los diseños para el análisis de la certeza de la hipótesis, de esta manera aportar con evidencias al respecto de la investigación.

En el diseño cuasi experimental no puede hacerse asignación aleatoria de sujetos (variable independiente) a grupos o condiciones, porque los grupos ya existen [59]. Se manipula a propósito, por lo menos una variable independiente para de esta forma observar su efecto sobre una o más variables dependientes.

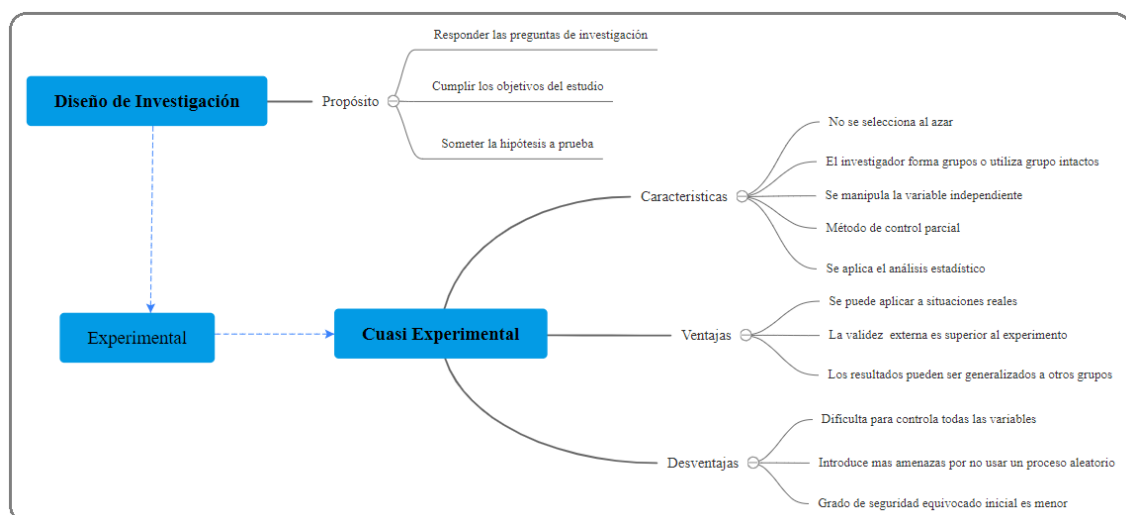
Por lo expuesta anteriormente el diseño de la presente investigación es de tipo cuasi experimental, ver Tabla 3.

Tabla 3. Variables Independiente y Dependiente

Variable Independiente	Variable Dependiente
Técnicas predictivas de minería de datos (Series de tiempo)	Unidades vendidas (del parque automotor en el Ecuador en los próximos años)

Una de las características que comparten los diseños cuasi experimentales con el experimental es que puede manipular la variable independiente, estos dos tipos de diseño su propósito principal es el estudio del efecto de la variable independiente sobre la dependiente de la investigación, ver Figura 20.

Figura 20. Diseño Cuasi-Experimental



Fuente: Adaptado por el autor a partir de [59]

2.3 CALCULO DE LA POBLACIÓN Y MUESTRA

Para la presente investigación no se trabajó con población o individuos, por ende, no contamos con cantidades en población, puesto que se trabajó con un dataset. Un dataset es conjunto de datos en el cual se almacena los contenidos de una única tabla de base de datos o también una única matriz de datos de estadística, en el cual cada columna de la tabla representa una variable en particular, y cada fila representa una observación

determinada del conjunto de datos que se está tratando. El dataset que se utilizó es de una base de datos con un formato específico y estructurado donde la información pasó por un proceso de ETL. El dataset cuenta con aproximadamente 250.000 registros, la información almacenada es sobre las ventas de vehículo nuevos de las diferentes marcas que se comercializan en el Ecuador. Está estructurado con los siguientes datos: año, mes, marca, modelo, familia, segmento, provincia, precio unitario, cilindraje, país de origen, unidades y monto. Se cuenta con un registro histórico desde el 1 de enero del 2007 al 31 de julio del 2020. En la Figura 21 se muestra la estructura del dataset utilizado.

Figura 21. Estructura del Dataset.



	Año	Mes	Marca	Modelo	Familia	Segmento	Provincia	Precio	Cilindraje	País Ensamblaje	Unidades	Total
1	2010	ENE	GEELY	D8CK 1.5 GS	SERIE CK	AUTOMOVIL	GUAYAS	7140	1.500	CHINA	1	7140
2	2010	ENE	GEELY	D8CK 1.5 GS	SERIE CK	AUTOMOVIL	PICHINCHA	7140	1.500	CHINA	7	49980
3	2010	ENE	GEELY	D8CK 1.5 GS	SERIE CK	AUTOMOVIL	SANTO DOMINGO	7140	1.500	CHINA	1	7140
4	2010	FEB	GEELY	D8CK 1.5 GS	SERIE CK	AUTOMOVIL	COTOPAXI	7140	1.500	CHINA	1	7140
5	2010	FEB	GEELY	D8CK 1.5 GS	SERIE CK	AUTOMOVIL	GUAYAS	7140	1.500	CHINA	3	21420

Fuente: Propia a partir de la base de datos de la asociación AEADE.

2.4 MÉTODOS TEÓRICO CON MATERIALES UTILIZADOS

La metodología de minería de datos permite explotar los datos con el objetivo de generar modelos que posibiliten describir, encontrar patrones, establecer agrupaciones, clasificar, segmentar o asociar productos, clientes o cualquier otra entidad objeto de obtener conocimiento y ser aplicados a otros nuevos [51].

Se utilizó la metodología TDSP que utiliza la colaboración en equipo y ayuda a mejorar el aprendizaje del mismo. Posee una combinación de las mejores prácticas y estructuras de Microsoft y otros en la industria de tecnologías, facilitando una implementación exitosa en iniciativas de ciencia de datos.

Al combinar software moderno y las buenas prácticas ágiles, esto con el ciclo de vida de la ciencia de datos, TDSP se integra de cuatro componentes principales, ver Figura 22.

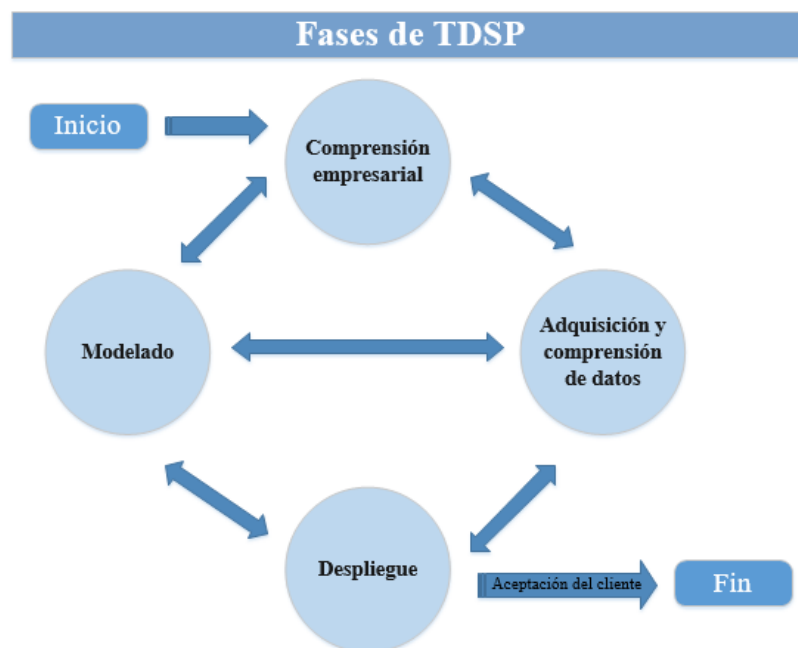
Figura 22. Componentes claves de TDSP



Fuente: Adaptación a partir de [57]

La metodología TDSP establece la minería de datos como una secuencia de fases: Comprensión empresarial, Adquisición y comprensión de datos, Modelado y Despliegue, como se muestra en la Figura 23.

Figura 23. Fases de TDSP [53]



Fuente: Adaptación a partir de [57]

2.5 MÉTODOS EMPÍRICOS CON MATERIALES UTILIZADOS

Los métodos empíricos se utilizaron para la ejecución de la investigación son:

- Modelos Matemáticos, se utiliza para analizar la relación entre dos o más variables para que la verdad se pueda interpretar a través de cálculos matemáticos, variables, asociaciones y parámetros descritos. Examinar la complejidad de múltiples sistemas en situaciones en las que es difícil observar hechos.
- Casos de Estudio, esta técnica de investigación basada en el entorno del mundo real para comparar o mejorar datos, conectar y encontrar formas; es ampliamente utilizado en ingeniería de software para describir métodos y herramientas. Este método tiene la ventaja de ser más fácil de planificar que de experimentar.
- Método de revisión de la literatura SLR [16], fue utilizado en el capítulo I, para la selección de artículos científicos con los cuales se pudo realizar la sección de antecedentes históricos de la investigación.

2.6 TÉCNICAS ESTADÍSTICAS PARA LA INVESTIGACIÓN

Las técnicas estadísticas utilizadas para la presente investigación son:

- Dispersión, la misma que presenta la variabilidad que se encuentren en las estadísticas realizadas. Esta técnica es importante debido a la naturaleza de los errores que ocurren al evaluar la escala de desviación central, que son cálculos estadísticamente significativos destinados a evaluar como la variación de los datos difieren entre sí. Las medidas diferenciales proporcionan información sobre la variabilidad. Y está destinada a resumir las diferencias que tiene el conjunto de datos. Los factores de propagación más comunes son: Rango de variación, Varianza, Desviación estándar, Coeficiente de variación.
- Técnicas Predictivas y de clasificación que se utilizaron son: regresión lineal, regresión cuadrática, series temporales y redes neuronales concurrentes, árboles de decisiones, además de técnicas de minería de datos descriptivas, que se analizaron en los antecedentes conceptuales en el capítulo I.

2.7 HERRAMIENTAS UTILIZADAS

Las herramientas utilizadas en el presente trabajo de investigación son para el manejo y modelado de datos son las siguientes:

Microsoft Excel, para el manejo de estadística descriptiva básica generación de gráficos de barras y además que actualmente dispone de un complemento de Minería de Datos en Excel, tras instalarlo mostrará una nueva pestaña en el menú de opciones con el nombre de MINERÍA DE DATOS, con las herramientas y asistentes que nos permitirán trabajar en ella, este beneficio está disponible desde la versión 2016 en la que incorpora un conjunto de funciones llamadas Obtener y Transformar que facilita la recopilación y organización de los datos. El mecanismo es el siguiente, primero establece una conexión a los datos externos a Excel (Bases de datos, software). Se extraen los datos a Excel y se les da formato, es decir, puede eliminar columnas o unir tablas para satisfacer mejor las necesidades. También puede combinar datos de varias fuentes en un solo modelo y usar herramientas como tablas dinámicas para ver mejor los datos [60].

Lenguaje R, es un lenguaje y un entorno para la computación y los gráficos estadísticos, proporciona una amplia variedad de técnicas estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, etc.) y técnicas gráficas. R permite realizar fácilmente gráficos con un excelente diseño y calidad de publicación, permite incluir los símbolos y fórmulas matemáticas de ser necesario. R está disponible como software libre, se compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows, MacOS, Linux, y sistemas similares [61].

CAPÍTULO 3

PROPUESTA DE ANÁLISIS DE DATOS EN EL SECTOR AUTOMOTRIZ APLICACIÓN DE TÉCNICA DE MINERÍA DE DATOS

En esta sección se presenta la elección de la mejor técnica, herramienta y algoritmo para la selección del modelo de análisis predictivo, producto del análisis de la literatura realizado en los capítulos anteriores. Para identificar un modelo que cumpla con los objetivos del negocio y de minería de datos, para esto se realizaron varias pruebas, en la cuales se obtuvieron diferentes parámetros. A continuación, se muestran los resultados obtenidos en cada uno de ellos.

3.1 FUNDAMENTACIÓN TEÓRICA DE LA PROPUESTA

Como parte de la fundamentación teórica para el presente trabajo, se analizaron metodologías de minería de datos.

3.1.1 METODOLOGÍAS DE MINERÍA DE DATOS

La minería de datos es el proceso que permite extracción de información significativa de grandes bases de datos, información que revela inteligencia del negocio, a través de datos ocultos, correlaciones y tendencias para proporcionar la facilidad de realizar predicciones que permitan resolver problemas del negocio otorgando así una ventaja competitiva [56].

Las metodologías de minería de datos analizadas fueron cuatro consideradas las más utilizadas entre las cuales tenemos: KDD, CRISP-DM, SEMMA y TDSP.

3.1.2 COMPARACIÓN DE METODOLOGÍAS MINERÍA DE DATOS

Para realizar la comparación de las metodologías de minería de datos analizadas se basó basamos en el Figura 24, la misma que nos muestra cada una de las fases de cada metodología [62].

Figura 24. Comparación Metodologías.

CRISP-DM	KDD	SEMMA	TDSP
Compresión del negocio			Compresión del negocio
Comprensión de los datos	Selección	Muestra ----- Explore	Adquisición y comprensión de datos
Preparación de los datos	Pre-procesado ----- Transformación		
Modelado	Minería de Datos	Modelo	Modelado
Evaluación	Interpretación / Evaluación	Evaluar	
Exploración			Implementación y aceptación del cliente

Fuente: Elaboración propia adaptado de [62]

Tomando en cuenta las fases y que comprende cada una de ellas, se realizó una comparación, otorgando un punto por cada característica cumplida en las metodologías y se procederá a la tabulación como se indica en la Tabla 4, los parámetros evaluados están inmersos en cada una de las fases y proceso de desarrollo de las metodologías, por esta razón se consideraron los siguiente:

- Entendimiento del negocio
- Muestra de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Implementación
- Despliegue
- Aceptación

Tabla 4. Comparación de metodología minería de datos.

Características	Metodología			
	KDD	SEMMA	CRISP-DM	TDSP
Entendimiento del Negocio	0	0	1	1
Comprensión de los datos	1	1	1	1
Preparación de los datos	1	1	1	1
Modelado	1	1	1	1
Evaluación	1	1	1	1
Implementación	1	1	1	1
Despliegue	0	0	1	1
Aceptación	0	0	0	1
Total de Medición	5	5	7	8

Fuente: Elaboración Propia adaptado de [62]

3.1.3 SELECCIÓN DE LA METODOLOGÍA

De lo analizado SEMMA y CRISP -DM son metodologías basadas en KDD, de acuerdo al análisis realizado se observó que CRISP–DM es una metodología que abarca una etapa muy importante en el entendimiento del negocio; si no se comprende el negocio, no se puede formular hipótesis adecuadas, que luego sean discernidas mediante la aplicación de modelos [62].

Por otro parte TDSP es una metodología para ciencia de datos que ofrecer resolver métodos de análisis predictivo mejorados y aplicaciones inteligentes, esto permite mejorar el aprendizaje y colaboración del equipo sugiriendo roles que funcionan mejor juntos. Dentro de ella hay un apartado dedicado al desarrollo ágil de proyectos en Ciencias de Datos. El mismo explica los pasos a seguir para planificar un sprint, agregar ítems de trabajo a uno entre los cuales tenemos (características, historias de usuario, tareas) y crear una plantilla de ítem de trabajo dentro de las etapas del ciclo de vida. TDSP también incorpora como aspecto fundamental la comprensión del negocio, un aspecto clave es que incorpora la característica de aceptación del cliente (no cubierta explícitamente por CRISP-DM).

Ante todo, lo expuesto de las metodologías analizadas y basados en el punto anterior en la comparación de las metodologías, TDSP obtuvo ocho puntos de los parámetros evaluados, además considerando que es una metodología nueva en minería de datos que está ganado posicionamiento siendo utilizada en muchos proyectos, en el presente trabajo se utilizó la metodología TDSP.

3.1.4 METODOLOGÍAS TDSP

TDSP utiliza una combinación de Scrum y CRISP-DM, lanzada en el 2016 Microsoft. TDSP se ha convertido en una "solución de ciencia de datos de ritmo rápido para ofrecer análisis de visión mejorados y aplicaciones de valor agregado" [14]. TDSP ayuda a mejorar la colaboración y el aprendizaje en equipo. Contiene una mezcla de las mejores prácticas y estructuras de Microsoft y otros en la industria que facilitan la implementación exitosa de iniciativas de ciencia de datos.

La metodología TDSP es una forma rápida y rentable de ofrecer análisis y aplicaciones mejoradas. TDSP ayuda a aumentar el compromiso y el aprendizaje del equipo al sugerir las mejores formas de trabajar juntos como equipo. El TDSP incluye las mejores prácticas y planes de Microsoft y otros líderes de la industria para implementar con éxito el modelo de educación científica. El objetivo principal es que los analistas de datos y las empresas puedan aprovechar de los beneficios de esta metodología [14].

3.1.4.1 COMPONENTES DEL TDSP

TDSP posee cuatro componentes claves en su estructura y son los siguientes:

- Ciclo de vida de la ciencia de datos
- Estructura de proyecto estandarizada
- Infraestructura y recursos
- Herramientas y utilidades

3.1.4.2 CICLO DE VIDA TDSP

El ciclo de vida de la metodología TDSP [14], facilita planificar el desarrollo de trabajos de ciencia de datos. El ciclo de vida representa el proceso completo del proyecto para el éxito profesional del mismo.

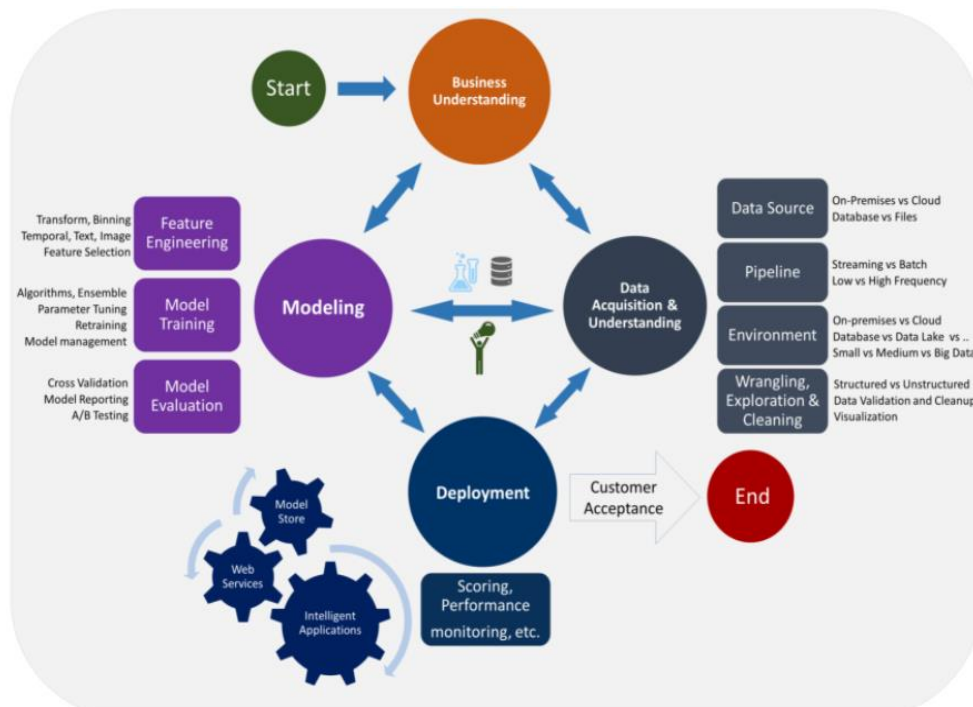
Esta metodología se presenta como parte de una aplicación inteligente. Esta herramienta aplica productos de inteligencia artificial o aprendizaje automático para análisis predictivo. Un proyecto de ciencia de datos o un proyecto de investigación de rutina puede utilizar esta función al usar este proceso.

El ciclo de vida describe los pasos claves que deben desarrollarse en el proyecto, por lo general de forma interactiva:

- **Comprensión empresarial:** defina objetivos e identifique fuentes de datos.
- **Adquisición y comprensión de datos:** ingiera datos y determine si pueden responder a la pregunta de presentación (combina de manera efectiva la *comprensión* y la *limpieza de datos* de CRISP-DM).
- **Modelado:** ingeniería de características y entrenamiento de modelos (combina modelado y evaluación).
- **Implementación:** implementar en un entorno de producción.
- **Aceptación del cliente:** validación del cliente si el sistema satisface las necesidades comerciales (una fase no cubierta por CRISP-DM).

A continuación, se muestra el ciclo de vida del TDSP observar Figura 25.

Figura 25. Ciclo de vida TDSP.

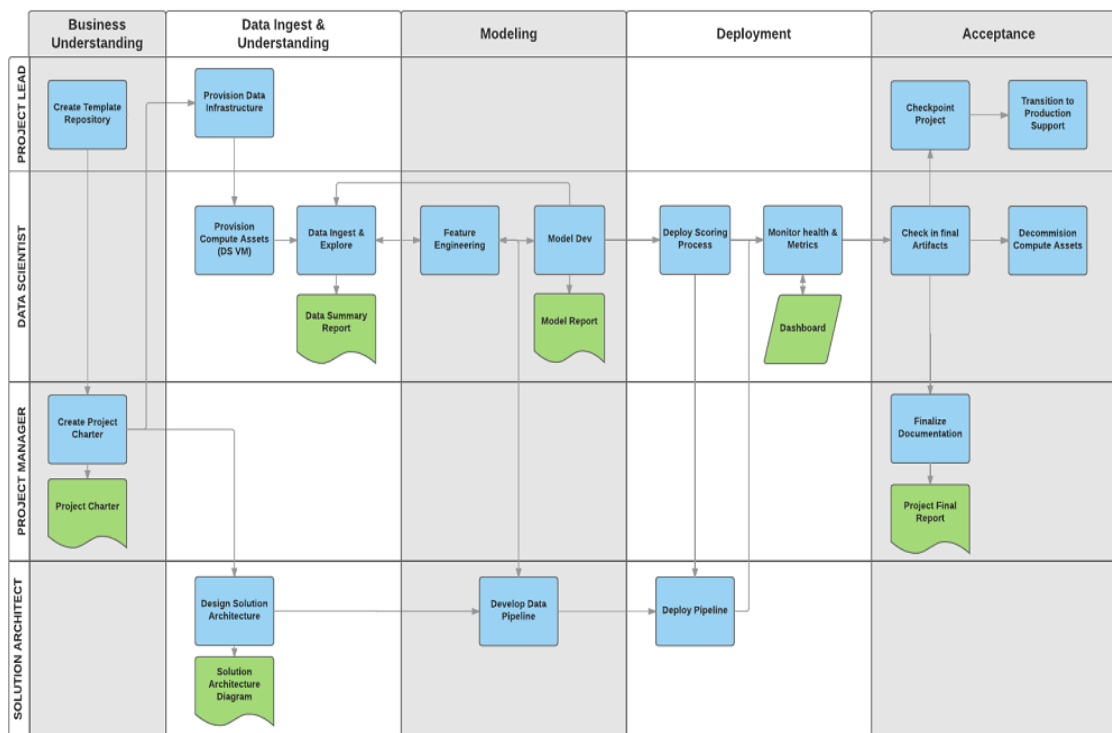


Fuente: Tomado de [14]

Las tareas y artefactos para la documentación de las etapas del ciclo de vida del TDSP ver Figura 26, estas están asociados con roles de proyecto:

- Arquitecto de soluciones
- Gerente de proyecto
- Ingeniero de datos
- Científico de datos
- Desarrollador de aplicaciones
- Líder del Proyecto

Figura 26. Tareas y artefactos. Ciclo TDSP y Roles



Fuente: Tomado de [14]

COMPENSIÓN EMPRESARIAL

Es la primera fase del ciclo de vida de TDSP.

- Especificar las variables claves que deben servir como objetivos del modelo y cuyas métricas relacionadas se utilizan para determinar el éxito del proyecto.
- Identificar las fuentes de datos relevantes a las que la empresa o proyecto tiene acceso o necesita obtener.

Existen dos tareas primordiales que se deben desarrollar en esta etapa:

- **Definir objetivos:** en este punto se trabaja en conjunto con el cliente y otras partes interesadas para comprender e identificar los problemas comerciales. Se formulan preguntas que definan los objetivos comerciales de las que se puedan dirigirse las técnicas de ciencia de datos.
 - Identificar un objetivo central, este paso permite las variables comerciales clave que el análisis necesita predecir.
 - Definir los objetivos del proyecto, formulando preguntas precisas, que sean específicas y relevantes.
 - Se define el equipo para el proyecto especificando responsabilidades y roles para cada miembro, además se desarrolla un plan con los hitos importantes, que se repiten cada vez que se descubre nueva información.
 - Por último, se definen métricas de éxito, las mismas que deben ser inteligentes.
- **Identificar fuentes de datos:** Determinar los datos relevantes que te ayuden a responder las preguntas que definen los objetivos del proyecto.
 - Datos que son relevantes para la pregunta. Deben poseer medidas del objetivo y características relacionadas con el objetivo.
 - Información que sea una medida precisa para el producto final y las características de interés.

Artefactos

Estos son los entregables en esta fase.

Documentos de estatus: proporciona una plantilla para la definición de la estructura del proyecto TDSP, es el documento de constitución y se actualiza a cada vez que se realizan nuevos descubrimientos.

Fuentes de datos: esta sección, determinará el origen y el destino de sus datos. El siguiente paso es completar detalles adicionales, como los scripts para mover los datos al entorno de aprendizaje.

Diccionario de datos: este documento facilita las descripciones de los datos proporcionados por el cliente. Esta información incluye descripciones sobre el esquema, además de los diagramas entidad-relación, si existieran.

ADQUISICIÓN Y COMPRENSIÓN DE DATOS

En esta segunda fase las metas a conseguir son las siguientes:

- Producir un conjunto de datos limpio y de alta calidad cuya relación con las variables objetivo se comprenda. Administrar el conjunto de datos en el entorno de análisis más apropiado para que esté listo para su modelación.
- Desarrollar una arquitectura de solución donde se canalicen los datos que actualice y califique los datos con regularidad.

Existen tres tareas principales que se cumplir en esta etapa:

- **Ingiera los datos** en el entorno analítico de destino.
 - Configurar el proceso para mover los datos desde las ubicaciones de origen a las ubicaciones de destino donde ejecuta operaciones de análisis, como entrenamiento y predicciones.
- **Explorar los datos** para determinar si la calidad de los datos es adecuada para responder la pregunta.
 - Para entrenar el modelo, se debe desarrollar una comprensión sólida de los datos. Los conjuntos de datos del mundo real a menudo son ruidosos, les faltan valores o tienen una serie de otras discrepancias.
 - Al estar conforme con la calidad de los datos depurados, el siguiente paso es comprender mejor los patrones relacionados a los datos.
- **Configurar una canalización de datos** para obtener datos nuevos o actualizados periódicamente.

Artefactos

Los siguientes artefactos son los entregables para esta etapa:

Informe de calidad de los datos: el informe incluye resúmenes de datos, las relaciones entre cada atributo y el objetivo, clasificación de variables y más.

Arquitectura de la solución: puede ser un diagrama o una descripción de la canalización de datos que se utilizaron para ejecutar puntuaciones o predicciones en datos nuevos después de haber creado un modelo.

Decisión de punto de control: antes de comenzar la ingeniería de funciones completas y la construcción de modelos, puede reevaluar el proyecto para determinar si el valor esperado es suficiente para continuar con el mismo. Con esto se puede continuar, recopilar más datos o abandonar el proyecto si la información encontrada no responde a la pregunta realizada en la fase anterior.

MODELADO

Es la tercera fase, las metas a conseguir son las siguientes:

- Determinar las características de datos óptimas para el modelo de aprendizaje automático.
- Crear un modelo informativo de aprendizaje automático que prediga el objetivo con mayor precisión.
- Crear un modelo de aprendizaje automático que sea adecuado para la producción.

Se debe cumplir con tres tareas principales que se desarrollan en esta etapa:

- **Ingeniería de características:** aquí se crean características de datos a partir de los datos sin procesar para facilitar el entrenamiento del modelo.
 - La ingeniería de características implica la inclusión, agregación y transformación de variables en bruto para crear las características utilizadas en el análisis. Si desea obtener información sobre lo que impulsa un modelo, se debe comprender cómo se relacionan las características entre sí y cómo los algoritmos de aprendizaje automático deben usar esas características.
 - En este paso se requiere una combinación creativa de experiencia en el dominio y la información obtenida del paso de exploración de datos. La ingeniería de características es un acto de equilibrio para encontrar e incluir variables informativas, pero al mismo tiempo, trata de evitar demasiadas variables no relacionadas. Las variables informativas mejoran su resultado; las variables no relacionadas introducen información innecesaria en el modelo. Como resultado,

la generación de estas características solo puede depender de los datos disponibles en el momento de la valoración del mismo.

- **Entrenamiento de modelos:** para encontrar el modelo que responda a la pregunta con mayor precisión comparando sus métricas de éxito.
 - Se deben dividir los datos de entrada de forma aleatoria para modelarlos en un conjunto de datos de entrenamiento y un conjunto de datos de prueba.
 - Crear los modelos utilizando el conjunto de datos de entrenamiento.
 - Evaluar el entrenamiento y el conjunto de datos de prueba. Se utiliza una serie de algoritmos de aprendizaje automático en competencia junto con los diversos parámetros de ajuste asociados que están orientados a responder la pregunta de interés con los datos actuales.
 - Determinar la mejor solución para responder la pregunta comparando las métricas de éxito entre métodos alternativos.
- Determinar si el modelo es el más **adecuado para la producción**.

Artefactos

Los artefactos que se obtienen en esta etapa incluyen:

- **Conjuntos de características:** las características desarrolladas para el modelado, estas se describen en la sección conjuntos de características del informe de definición de datos.
- **Informe del modelo:** cada modelo que se prueba, se debe elaborar un informe estándar basado en plantillas que proporcionan detalles sobre cada experimento.
- **Decisión del punto de control:** para evaluar si el modelo funciona lo suficiente para la producción. Algunas preguntas clave para hacer son:
 - ¿Responde el modelo a la pregunta con suficiente confianza, dados los datos de la prueba?
 - ¿Debería probar algún enfoque alternativo?
 - ¿Debería recopilar datos adicionales, hacer más ingeniería de funciones o experimentar con otros algoritmos?

DESPLIEGUE

En esta fase se debe implementar los modelos obtenidos para que el usuario final los acepte.

La principal tarea de esta etapa:

- Poner en funcionamiento el modelo: implementar el modelo obtenido y la canalización en un entorno de producción o similar a la producción para el consumo de aplicaciones.

Ejecutar modelo

Una vez que se obtiene un conjunto de modelos y funcionan bien, se los debe poner en funcionamiento para que los consuman otras aplicaciones. Dependiendo de los requisitos del proyecto, las predicciones se realizan en tiempo real o por lotes. Para implementar modelos se los presenta con una interfaz abierta. La interfaz permite que el modelo se consuma fácilmente desde varias aplicaciones, como: sitios web en línea, hojas de cálculo, cuadros de mando, aplicaciones de línea de negocio, aplicaciones back-end.

Artefactos

- Un panel de estado, el que muestra el estado del sistema y las métricas clave.
- Un informe de modelado final con detalles de implementación.
- Un documento de arquitectura de solución final.

ACEPTACIÓN DEL CLIENTE

Es la última fase del ciclo de vida del TDSP.

Finalizar los entregables del proyecto: se confirma la validación del modelo y su implementación en un entorno de producción que satisfaga los objetivos del proyecto.

Hay dos tareas principales que se deben realizar en esta etapa:

- **Validación del sistema:** confirmar que el modelo este implementado y validado, además que satisfacen las necesidades del proyecto.

- Se debe validar que el sistema satisface las necesidades comerciales y del proyecto, y que responde a las preguntas con una precisión aceptable para implementar el sistema en producción.
- **Traspaso del proyecto:** entregar el proyecto a la entidad que ejecutará el sistema en producción.

Artefactos

El principal artefacto producido en esta etapa final es el **informe de salida del proyecto para el cliente**. Este informe técnico contiene todos los detalles del proyecto que son útiles para aprender a operar el sistema. TDSP proporciona una plantilla de informe de salida.

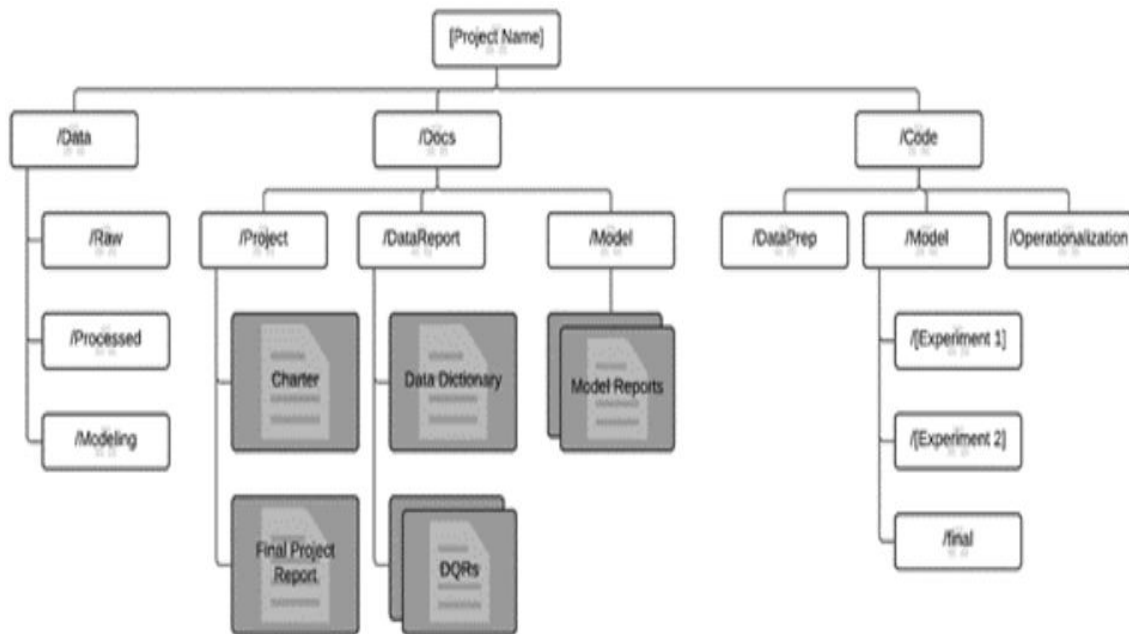
3.1.4.3 ESTRUCTURA DE PROYECTO ESTANDARIZADA

Los proyectos comparten una estructura índice y utilizan plantillas de trabajo para proporcionar a los integrantes del equipo información sobre el proyecto. El código y la documentación se almacén en sistemas de gestión de versiones VCS y subversión, lo que permite la colaboración en equipo. Las funciones de seguimiento y las funciones en los sistemas de seguimiento del sistema como Wait, Rally y Azure DevOps permiten un seguimiento completo del sistema de características individuales.

TDSP recomienda crear un repositorio separado para cada proyecto de VCS para administrar la versión, la seguridad de los datos y la colaboración. El marco central de todas las actividades ayuda a proporcionar conocimientos formales dentro del equipo.

Este sistema de carpetas contiene funciones de análisis y extracción de datos, así como también organiza archivos que registran contenido actualizado. Estas plantillas facilitan que a los integrantes del equipo comprendan el trabajo realizado por otros ver Figura 27.

Figura 27. Estructura del directorio TDSP.



Fuente: Tomato de [14]

3.1.4.4 INFRAESTRUCTURA Y RECURSOS

TDSP sugiere administrar la infraestructura de almacenamiento y análisis compartidos, tales como:

- Sistemas de archivos en la nube.
- Bases de datos
- Clústeres de big data
- Servicio de aprendizaje automático

La infraestructura de análisis y almacenamiento, donde se almacenan los conjuntos de datos sin procesar y procesados, puede estar en la nube o en las instalaciones. Esta infraestructura permite un análisis reproducible. También evita la duplicación, que puede generar incoherencias y costos de infraestructura innecesarios. Se proporcionan herramientas para aprovisionar los recursos compartidos, rastrearlos y permitir que cada miembro del equipo se conecte a esos recursos de forma segura.

3.1.4.5 HERRAMIENTAS Y UTILIDADES

Las herramientas para implementar el proceso de ciencia de datos y el ciclo de vida ayudan a reducir las barreras y aumentar la consistencia de su adopción. TDSP proporciona un conjunto inicial de herramientas y scripts para impulsar la adopción de TDSP dentro de un equipo. Estos recursos pueden luego ser aprovechados por otros proyectos dentro del equipo o la organización. Microsoft proporciona amplias herramientas dentro de Azure Machine Learning admite tanto el código abierto de Python, R y marcos de aprendizaje profundo comunes como también las herramientas propias de Microsoft.

3.1.4.6 ROLES Y TAREAS

Estructura de los grupos y equipos de ciencia de datos.

Las funciones de ciencia de datos en las organizaciones a menudo se organizan en la siguiente jerarquía:

- Grupo de ciencia de datos
 - Equipo(s) de ciencia de datos dentro del grupo

En esta estructura, hay líderes de grupo y líderes de equipo. Por lo general, un proyecto de ciencia de datos lo realiza un equipo de ciencia de datos. Los equipos de ciencia de datos tienen líderes de proyectos para la gestión de proyectos y tareas, y científicos e ingenieros de datos individuales para realizar las partes de ciencia de datos e ingeniería de datos del proyecto. La configuración del proyecto inicial la realizan el grupo, el equipo o los líderes del proyecto.

Definición y tareas para los cuatro roles de TDSP

Al realizar un proyecto de ciencia de datos con TDSP, este consta de equipos dentro de un grupo, hay cuatro roles distintos para el personal:

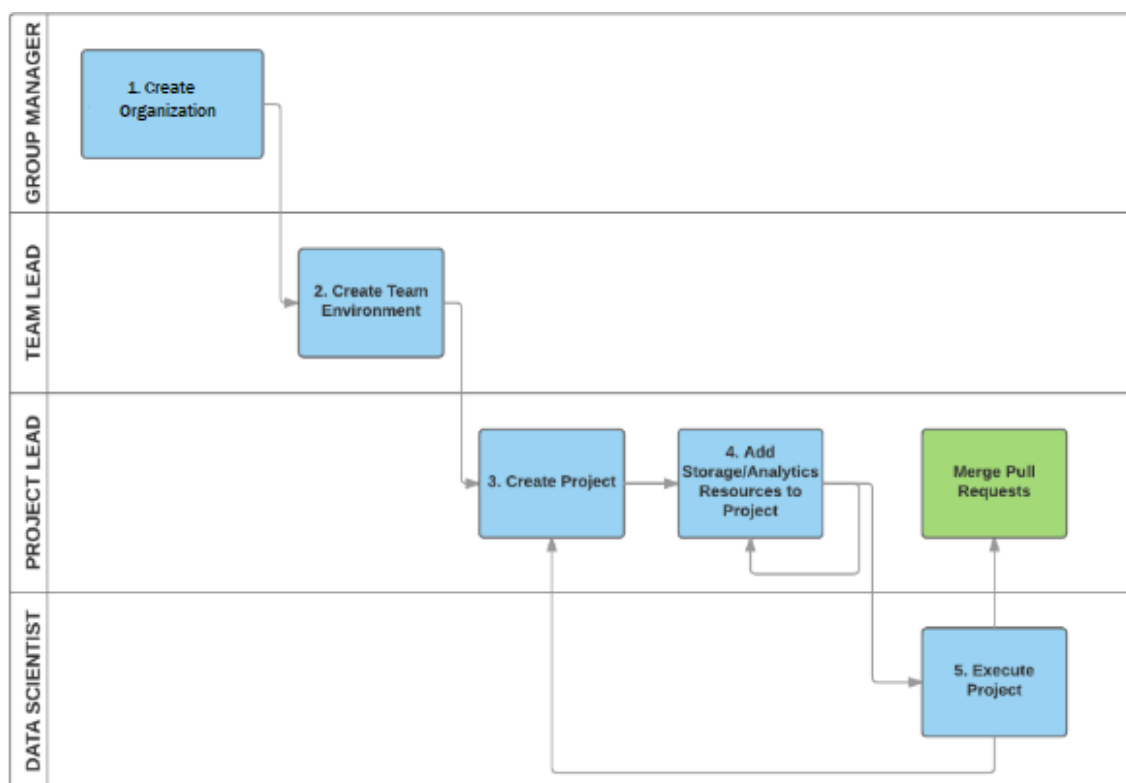
- **Administrador de grupo:** es el asignado en administrar toda la unidad de ciencia de datos del proyecto o de una empresa. Una unidad de ciencia de datos puede tener varios equipos, Un administrador de grupo puede delegar sus tareas a un sustituto, pero las tareas asociadas con el rol no cambian.

- **Líder de equipo:** administra al equipo en la unidad de ciencia de datos de un proyecto, un equipo puede tener varios científicos de datos. Para una unidad de ciencia de datos pequeña, el administrador del grupo y el líder del equipo pueden ser la misma persona.
- **Jefe de proyecto:** es el encargado de gestionar las actividades diarias de los científicos de datos individuales en el proyecto de ciencia de datos específico.
- **Colaboradores individuales del proyecto:** son científicos de datos, analistas de negocios, ingenieros de datos, arquitectos y otros que ejecutan un proyecto de ciencia de datos.

Tareas que deben completar los cuatro roles.

En la Figura 28 se muestra las tareas de nivel superior para cada rol de Proceso de ciencia de datos en equipo. Este esquema y el siguiente esquema más detallado de tareas para cada rol de TDSP pueden ayudarlo a elegir el tutorial que necesita en función de sus responsabilidades.

Figura 28. Tareas según de los roles



Fuente: Tomado de [14]

3.1.5 SERIES TEMPORALES

Las series temporales son una sucesión de datos, observaciones o valores, los que son medidos en determinados momentos y ordenados de forma cronológica. La predicción de series de tiempo requiere que solo se puedan usar datos pasados para predecir datos futuros [63] .

El modelo de series temporales a evolucionado en el tiempo, el estadístico británico Yule propuso el modelo autorregresivo (modelo AR) en 1927. En 1931, Varg estableció un modelo de media móvil (modelo MA) y un modelo de media móvil autorregresiva (modelo ARMA). Después de eso, apareció un modelo de promedio móvil integrado autorregresivo (modelo ARIMA). Con el rápido desarrollo del aprendizaje automático, se han aplicado muchos métodos en el aprendizaje automático a la predicción [63], estas han sido creadas para resolver el problema de la predicción de series temporales.

Es un método robusto y versátil para aplicar descomposición de series temporales. STL es un acrónimo de “Descomposición estacional y de tendencias con Loess”, siendo Loess un método para estimar relaciones no lineales. Este método STL fue creado por Cleveland, McRae y Terpenning.

3.1.5.1 MODELO STLM

El modelo de predicción de eventos basada en STLM. Los modelos STLM se están utilizando con éxito para modelar series de tiempo con la ayuda del vector de estado oculto, lo que permite resumir en el estado oculto información del pasado más distante [64]. STLM obtiene la entrada actual (evento) y actualiza sus estados ocultos. El estado oculto genera señales para el siguiente estado oculto, así como predicciones para la ocurrencia de eventos en el siguiente paso de tiempo [65].

3.1.5.2 MODELO STLF

El modelo STL fue creado por Cleveland, McRae y Terpenning. Es un Método robusto y versátil, se lo utiliza para descomposición de series temporales. STL es un acrónimo de “Descomposición estacional y de tendencias con Loess”. siendo Loess un modelo utilizado para estimar relaciones no lineales.

3.1.5.3 MODELO HOLT WINTERS

El método Holt Winters es una adición del método Holt que considera solo dos exponentes suavizantes. Este método considera nivel, tendencia y estacional de una determinada serie de datos de tiempos. Este método tiene dos principales modelos, dependiendo del tipo de estacionalidad que son el modelo multiplicativo estacional y el modelo aditivo estacional. El método de Holt Winters se utiliza para modelar cada serie de tiempo y el pronóstico final se obtiene agregando el conjunto de pronósticos [66]. Los métodos de Holt Winters son los métodos más apropiados para las predicciones a largo plazo [67].

3.1.5.4 MODELO STLM ARIMA 3,1,6

El modelo ARIMA permite describir valores en función de líneas de datos pasados y errores de riesgo. Además, puede incluir eventos ocasionales o estacionales. Es decir, debe contener todos los elementos necesarios para describir el problema, Box y Jenkins recomiendan al menos 50 observaciones en una serie de eventos. Son algoritmos importantes para series de tiempo datos. Asume una relación lineal y necesita muchos datos para producir resultados precisos [68].

3.1.5.5 MODELO EST SUAVIZACIÓN EXPONENCIAL

El modelo EST que se puede utilizar para generar predicciones es una simulación simple que incorpora predicciones de errores pasados. Los criterios para este modelo son:

$$F_{t+1} = F_t + \alpha \cdot (A_t - F_t) = \alpha \cdot A_t + (1 - \alpha) \cdot F_t$$

Donde:

A_t : Valor real de la serie para el período t .

F_t : Predicción realizada para el período t .

α : Constante de suavización, con $0 \leq \alpha \leq 1$.

F_{t+1} : Predicción para el período $t+1$.

El valor de α puede lograr utilizando un modelo optimizado. Puede ver que esto reduce la percepción errónea en el entrenamiento porque puede aplicar la percepción errónea en el entrenamiento, como error de predicción se puede utilizar el MAPE.

3.1.5.6 MODELO AUTO.ARIMA

Es un Modelo Autorregresivo-Integrado de Medias Móviles de orden p, d, q , o abreviadamente ARIMA (p,d,q) , es el modelo ARIMA (p,q) , aplicado al conjunto de grado complejo d , es decir, es necesario el tiempo d para eliminar la condición

3.1.5.7 MODELO NNETAR

El modelo NNETAR es una función en el paquete de pronóstico para R que ajusta un modelo de red neuronal a una serie de líneas de tiempo que tiene la misma cantidad de retraso que la entrada. Por lo tanto, es un proceso directamente auto agresivo y el análisis predictivo no se puede obtener mediante análisis, entonces usa la simulación [64].

3.1.5.8 MODELO TBATS

Un modelo TBATS requiere la estimación de $2 \dots (k_1 + k_2 + \dots + k_T)$ valores iniciales, Por lo general, menos que el número de versiones semilla en el modelo BATS. Otra ventaja es que se puede utilizar funciones trigonométricas para modelar frecuencias incomparables [68]. Algunos de los principales beneficios del modelo TBATS son:

- Admite un espacio de parámetros grande con posibilidad de mejores pronósticos.
- Permite múltiples componentes estacionales anidadas y no anidadas.
- Trata características no lineales que a menudo se presentan en series de tiempo reales
- Permiten que automáticamente se tenga en cuenta cualquier autocorrelación que se presente en los residuos.
- Involucra un método de estimación más simple y eficiente.

3.2 PROPUESTA METODOLÓGICA

Para el desarrollo de este trabajo se consideró utilizar TDSP que es una metodología flexible, esta puede adaptar y personalizar fácilmente, además que permite crear un modelo de minería de datos que se acople a las necesidades específicas del negocio.

3.2.1 APLICACIÓN METODOLOGIA TDSP

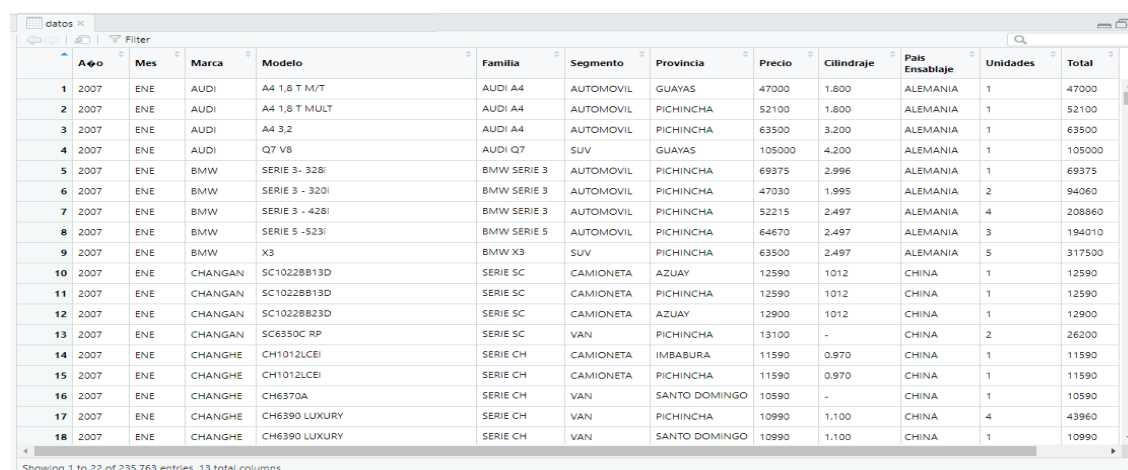
En este punto se describirá el desarrollo de las etapas de Entendimiento del negocio y Adquisición de datos, que comprenden el ciclo de vida de la metodología TDSP Team Data Science Process. Al desarrollar estas estepas permiten identificar información clave, que se tiene que utilizar en el proceso de entrenamiento del modelo y predecir el comportamiento del parque automotor en el Ecuador, con ello identificar si el número de unidades nuevas anualmente incrementa o decremento.

3.2.1.1 ENTENDIMIENTO DEL NEGOCIO

Para el desarrollo de esta etapa del proyecto, se solicitó a la jefatura del departamento de sistemas de una empresa automotriz el acceso a los datos almacenados. Los datos proporcionados corresponden a una copia de su Base de datos de SQL Server generados por la herramienta de exportación de Microsoft. Se realizó un barrido de información para poder determinar la que sea importante y relevante para aplicar en el proyecto.

En esta base de datos se encuentran registradas todas las ventas de vehículos nuevos de todos los concesionarios a nivel nacional la información del periodo comprendido del 1 de enero del 2007 al 31 de julio del 2020, con varios campos que son para cada registro que sirven como indicadores para poder realizar algunas consultas y depuraciones para obtener una data limpia para la aplicación de las técnicas de aprendizaje de series temporales. La estructura del dataset cuenta con los siguientes campos Año, Mes, Marca, modelo, Familia, Segmento, Provincia, Precio, Cilindraje, País Ensamblaje, Unidades y Total, como se puede observar en la Figura 29, el dataset cuenta con un total de registros de 235763.

Figura 29. Estructura completa dataset “datos”.



	Año	Mes	Marca	Modelo	Familia	Segmento	Provincia	Precio	Cilindraje	País Ensamblaje	Unidades	Total
1	2007	ENE	AUDI	A4 1,8 T M/T	AUDI A4	AUTOMOVIL	GUAYAS	47000	1.800	ALEMANIA	1	47000
2	2007	ENE	AUDI	A4 1,8 T MULT	AUDI A4	AUTOMOVIL	PICHINCHA	52100	1.800	ALEMANIA	1	52100
3	2007	ENE	AUDI	A4 3,2	AUDI A4	AUTOMOVIL	PICHINCHA	63500	3.200	ALEMANIA	1	63500
4	2007	ENE	AUDI	Q7 V8	AUDI Q7	SUV	GUAYAS	105000	4.200	ALEMANIA	1	105000
5	2007	ENE	BMW	SERIE 3- 328i	BMW SERIE 3	AUTOMOVIL	PICHINCHA	69375	2.996	ALEMANIA	1	69375
6	2007	ENE	BMW	SERIE 3 - 320i	BMW SERIE 3	AUTOMOVIL	PICHINCHA	47030	1.995	ALEMANIA	2	94060
7	2007	ENE	BMW	SERIE 3 - 428i	BMW SERIE 3	AUTOMOVIL	PICHINCHA	52215	2.497	ALEMANIA	4	208860
8	2007	ENE	BMW	SERIE 5 -523i	BMW SERIE 5	AUTOMOVIL	PICHINCHA	64670	2.497	ALEMANIA	3	194010
9	2007	ENE	BMW	X3	BMW X3	SUV	PICHINCHA	63500	2.497	ALEMANIA	5	317500
10	2007	ENE	CHANGAN	SC1022BB13D	SERIE SC	CAMIONETA	AZUAY	12590	1012	CHINA	1	12590
11	2007	ENE	CHANGAN	SC1022BB13D	SERIE SC	CAMIONETA	PICHINCHA	12590	1012	CHINA	1	12590
12	2007	ENE	CHANGAN	SC1022BB23D	SERIE SC	CAMIONETA	AZUAY	12900	1012	CHINA	1	12900
13	2007	ENE	CHANGAN	SC6350C RP	SERIE SC	VAN	PICHINCHA	13100	-	CHINA	2	26200
14	2007	ENE	CHANGHE	CH1012LCEI	SERIE CH	CAMIONETA	IMBABURA	11590	0.970	CHINA	1	11590
15	2007	ENE	CHANGHE	CH1012LCEI	SERIE CH	CAMIONETA	PICHINCHA	11590	0.970	CHINA	1	11590
16	2007	ENE	CHANGHE	CH6370A	SERIE CH	VAN	SANTO DOMINGO	10590	-	CHINA	1	10590
17	2007	ENE	CHANGHE	CH6390 LUXURY	SERIE CH	VAN	PICHINCHA	10990	1.100	CHINA	4	43960
18	2007	ENE	CHANGHE	CH6390 LUXURY	SERIE CH	VAN	SANTO DOMINGO	10990	1.100	CHINA	1	10990

Fuente: Elaboración Propia, datos tomados de [11]

Es necesario aclarar que los datos de referencia que se utilizaron desde la ingesta hasta la construcción y evaluación del modelo corresponden a la totalidad de ventas generadas por los concesionarios autorizados en el país. Con estos datos se puede determinar los KPI que se deseen analizar utilizando la misma mecánica de aquí en adelante cualquier tipo de parámetro que se desee analizar.

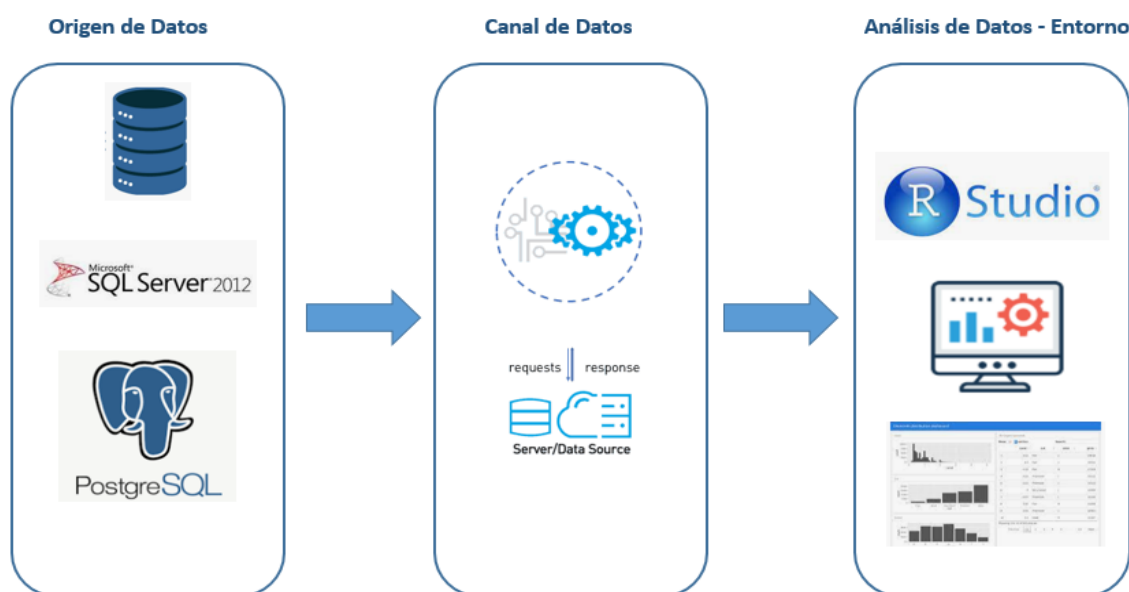
3.2.1.2 ADQUISICIÓN Y COMPRENSIÓN DE DATOS

Los datos disponibles se utilizan para realizar los pasos de entrada Ingesta de datos, pre procesamiento y exploración. Estos tres pasos son primordiales en la preparación de datos a utilizarse en el modelo de aprendizaje predictivo.

Ingesta de datos, en este punto se realizó la restauración de la copia de la base de datos, en PostgreSQL que es compatible con SQL y con RStudio que la encuentra en la aplicación web Jupyter. La herramienta RStudio importamos las bibliotecas de código abierto para poder realizar el entorno para el análisis y desarrollar los modelos de aprendizaje automático predictivos.

Canalización de datos, luego se realiza el canal de datos para obtener la información de la base de datos y enviarla la data source de donde se va analizar los datos, además que se convierte en nuestro entorno de aprendizaje como se observa en la Figura 30.

Figura 30. Arquitectura de solución.



Fuente: Elaboración propia.

3.2.1.3 EXPLORACIÓN DE DATOS - MODELADO

Considerando como referencia las tareas propuestas por la metodología TDSP para este paso, se describe las actividades realizadas durante el trabajo de preparación de datos y los procedimientos utilizados con la información.

Pre procesamiento, en este paso se limpia y se realizó la depuración de la información, revisando que no contengan información no valida, que no falten valores en los campos. Se realizó una auditoria de los datos para verificar que estén conformes a lo requerido y poder aplicar los modelos de análisis.

Con funciones de SQL se eliminan espacio es blanco al inicio del dato y al final del mismo corrigiendo un total de 64785 registros de la base de datos, también se comparan datos en las columnas encontrando datos errados que no correspondan, por esto se procedió a corregirlos, esto se encontró en la columna Marca un total de 32415 registros ver Tabla 5.

Tabla 5. Corrección datos campo Marca

Campo: **Marca**

Incorrecto	Correcto	Registros Afectados
AUDY	AUDI	2563
CHAMGAN	CHANGAN	978
BMVV	BMW	1023
CADILAC	CADILLAC	86
CHANCHE	CHANGHE	203
HYUNDAY	HYUNDAI	6952
ZOTIE	ZOTYE	321
MITSUBICHI	MITSUBISHI	1545
JEP	JEEP	3774
KYA	KIA	12452
QCM	QMC	3
VOLCSWAGEN	VOLKSWAGEN	1765
VOLWO	VOLVO	163

RENAUL	RENAULT	587
--------	---------	-----

Fuente: Elaboración propia

En la columna Segmento se corrigieron 73828 registros ver Tabla 6.

Tabla 6. Corrección datos campo Segmento

Campo: Segmento

Incorrecto	Correcto	Registros Afectados
VAM	VAN	12352
SUW	SUV	24019
AUTO MOVIL	AUTOMOVIL	37457

Fuente: Elaboración propia

Para la columna Provincia se corrigió un total de 29848 registros ver tabla

Tabla 7. Corrección datos campo Provincia

Campo: Provincia

Incorrecto	Correcto	Registros Afectados
AZUY	AZUAY	2357
ELORO	EL ORO	4568
PICHICHA	PICHINCHA	9541
GUAYAQUIL	GUAYAS	8475
LOSRIOS	LOS RIOS	1853
CHIMBORASO	CHIMBORAZO	874
TUNGURAGUA	TUNGURAHUA	1422
COTOPACI	COTOPAXI	758

Fuente: Elaboración propia

Por ultimo en la columna País Ensamblaje se corrigieron 6768 como se muestra en detalle en la Tabla 8.

Tabla 8. Corrección datos campo País Origen

Campo: País Origen

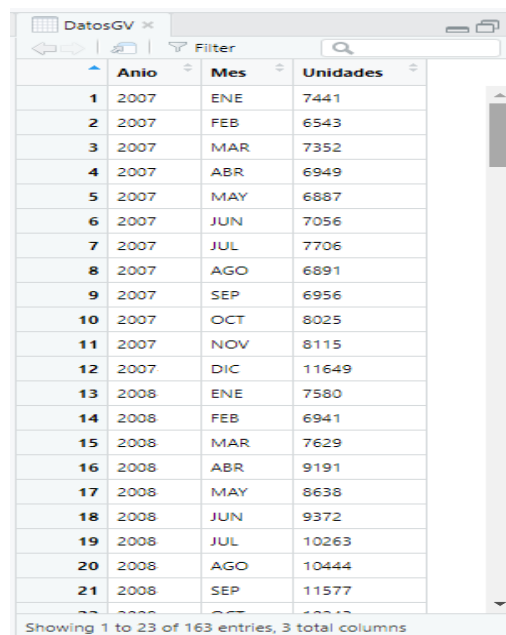
Incorrecto	Correcto	Registros Afectados
ITALYA	ITALIA	267
BRAZIL	BRASIL	2654
EEUU	ESTADOS UNIDOS	789
CHIMA	CHINA	3058

Fuente: Elaboración propia

También se encontró campos en la columna precio que estaba vacíos sin valores, los cuales tuvieron que ser depurados y realizar un script SQL para completar esta información, se actualizaron un total de 42.658 registro que estaba en blanco esta columna.

Para el proyecto se va a analizar las ventas totales por mes año, para lo cual se realizó un resumen de información generando un nuevo dataset de nombre DatosGV, Se utilizó el método de agregación, para la reducción de datos mediante sentencias SQL, se realiza la agrupación de los campos: años, mes y unidades, ver Figura 31.

Figura 31. Estructura completa dataset “DatosGV”.



	Anio	Mes	Unidades
1	2007	ENE	7441
2	2007	FEB	6543
3	2007	MAR	7352
4	2007	ABR	6949
5	2007	MAY	6887
6	2007	JUN	7056
7	2007	JUL	7706
8	2007	AGO	6891
9	2007	SEP	6956
10	2007	OCT	8025
11	2007	NOV	8115
12	2007	DIC	11649
13	2008	ENE	7580
14	2008	FEB	6941
15	2008	MAR	7629
16	2008	ABR	9191
17	2008	MAY	8638
18	2008	JUN	9372
19	2008	JUL	10263
20	2008	AGO	10444
21	2008	SEP	11577
22	2008	OCT	10012
23	2008	NOV	10012

Showing 1 to 23 of 163 entries, 3 total columns

Fuente: Elaboración propia.

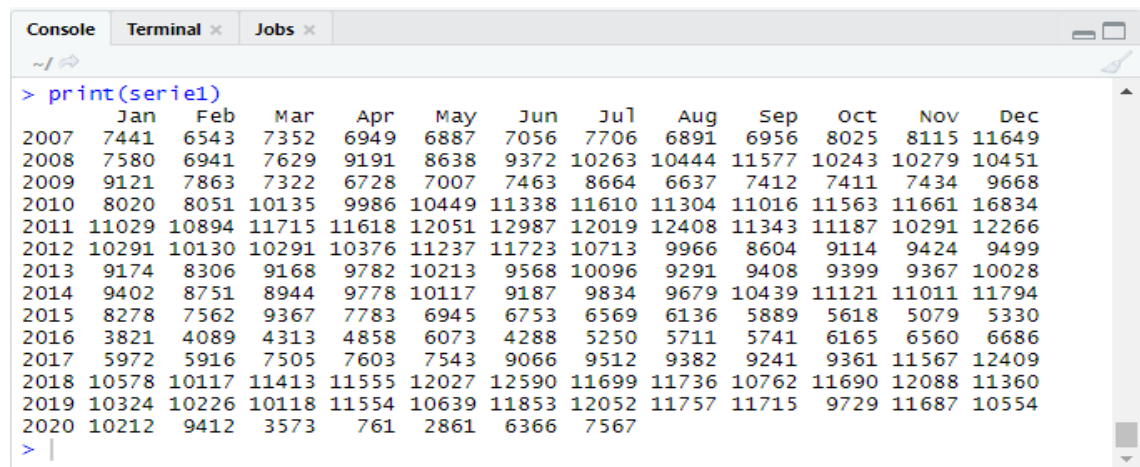
Análisis exploratorio, para el desarrollo del análisis de los datos utilizamos la herramienta RStudio, con el que se desarrollaron algunos análisis exploratorios:

El propósito es analizar la muestra y proporcionar datos del número de vehículo colocados mensualmente en el parque automotor del país, por mes durante los últimos 14 años. Lo primero que se realizó es leer los datos, organizarlos, transformarlos en una línea de tiempo y mostrarlo gráficamente.

Trazamos la serie de tiempo **serie1**, tendencia ascendente, no es estacionaria en la media, para observar la evolución de unidades nuevas colocadas por mes de la data general, ver Figura 32.

```
serie1<-ts(DatosGV$Unidades, start = c(2007,1), frequency = 12)
print(serie1)
```

Figura 32. Evolución mensual en número de vehículos 2007 - 2020



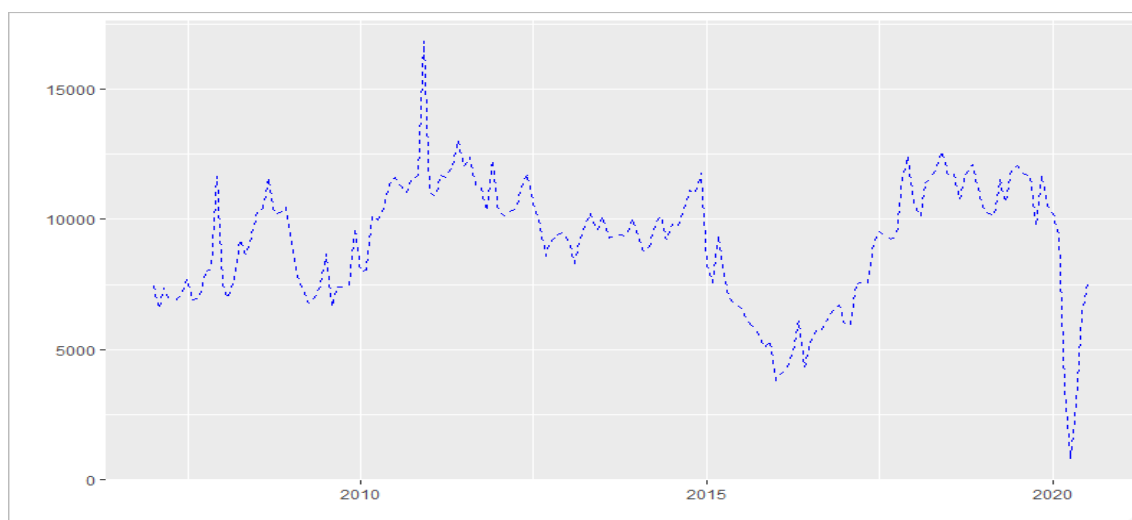
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	7441	6543	7352	6949	6887	7056	7706	6891	6956	8025	8115	11649
2008	7580	6941	7629	9191	8638	9372	10263	10444	11577	10243	10279	10451
2009	9121	7863	7322	6728	7007	7463	8664	6637	7412	7411	7434	9668
2010	8020	8051	10135	9986	10449	11338	11610	11304	11016	11563	11661	16834
2011	11029	10894	11715	11618	12051	12987	12019	12408	11343	11187	10291	12266
2012	10291	10130	10291	10376	11237	11723	10713	9966	8604	9114	9424	9499
2013	9174	8306	9168	9782	10213	9568	10096	9291	9408	9399	9367	10028
2014	9402	8751	8944	9778	10117	9187	9834	9679	10439	11121	11011	11794
2015	8278	7562	9367	7783	6945	6753	6569	6136	5889	5618	5079	5330
2016	3821	4089	4313	4858	6073	4288	5250	5711	5741	6165	6560	6686
2017	5972	5916	7505	7603	7543	9066	9512	9382	9241	9361	11567	12409
2018	10578	10117	11413	11555	12027	12590	11699	11736	10762	11690	12088	11360
2019	10324	10226	10118	11554	10639	11853	12052	11757	11715	9729	11687	10554
2020	10212	9412	3573	761	2861	6366	7567					

Fuente: Elaboración propia.

```
autoplot(serie1, ts.colour = "blue", ts.linetype = "dashed")
```

Realizada la traza de la serie1 se observa que posee una tendencia variable en el tiempo como lo muestra la Figura 33. Además, podemos determinar que la serie no es estacionaria, también se aprecia marcados pico en la traza.

Figura 33. Evolución mensual en número de vehículos – Grafica.



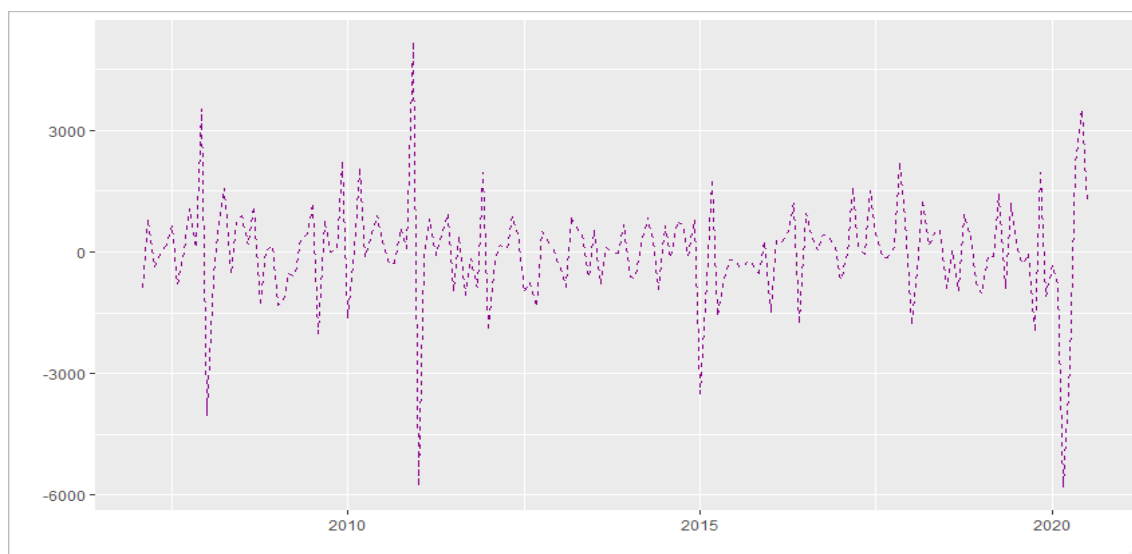
Fuente: Elaboración propia.

El análisis previo revela que tenemos que eliminar la tendencia y la estacionalidad de la serie para que pueda ser estacionaria. El modelo utilizado es ARIMA, que distingue la serie a destacar. Se utilizó las funciones `ndiffs` y `nsdiffs`, estas calculan el número de diferencias periódicas y estacionales, respectivamente que requiere una implementación para que la serie sea estacional

En la Figura 34 los cálculos nos muestran que la serie necesita una diferenciación regular (tendencia regular con la media) y otra estacional (estructura que se repite en un periodo).

```
diff.serie1<-autoplot(diff(serie1), ts.linetype = "dashed", ts.colour = "darkmagenta")
diff.serie1
```

Figura 34. Evolución de las unidades

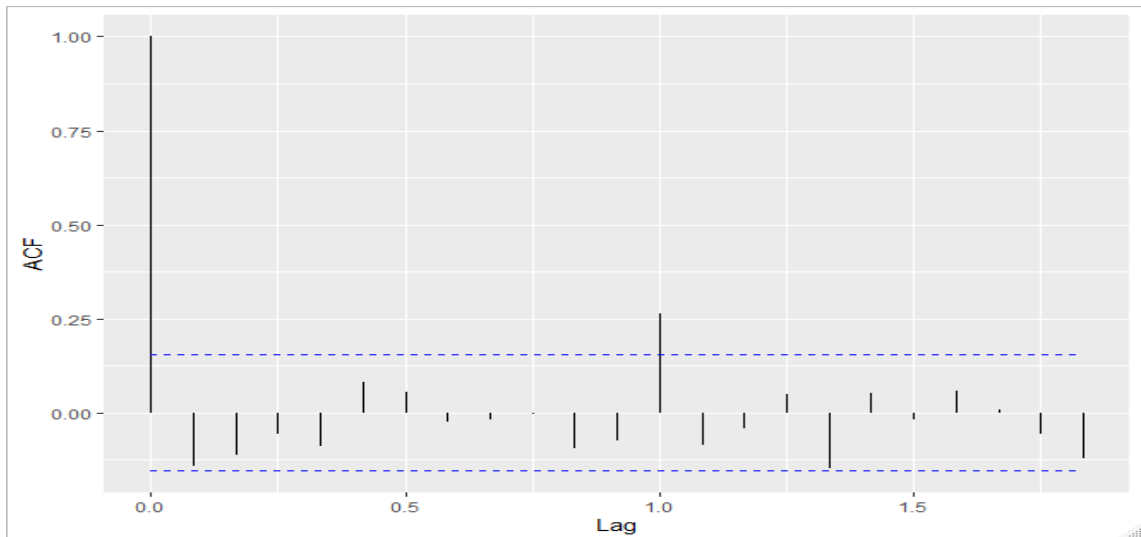


Fuente: Elaboración propia

Al ejecutar la siguiente línea de código y podemos observar el grafico que la serie no es estacional, más adelante se muestra el análisis de los compones por separado, para esto descomponemos la serie, observar Figura 35.

```
autoplots(acf(diff(serie1), plot = FALSE))
```

Figura 35. Grafica Serie No Estacional.

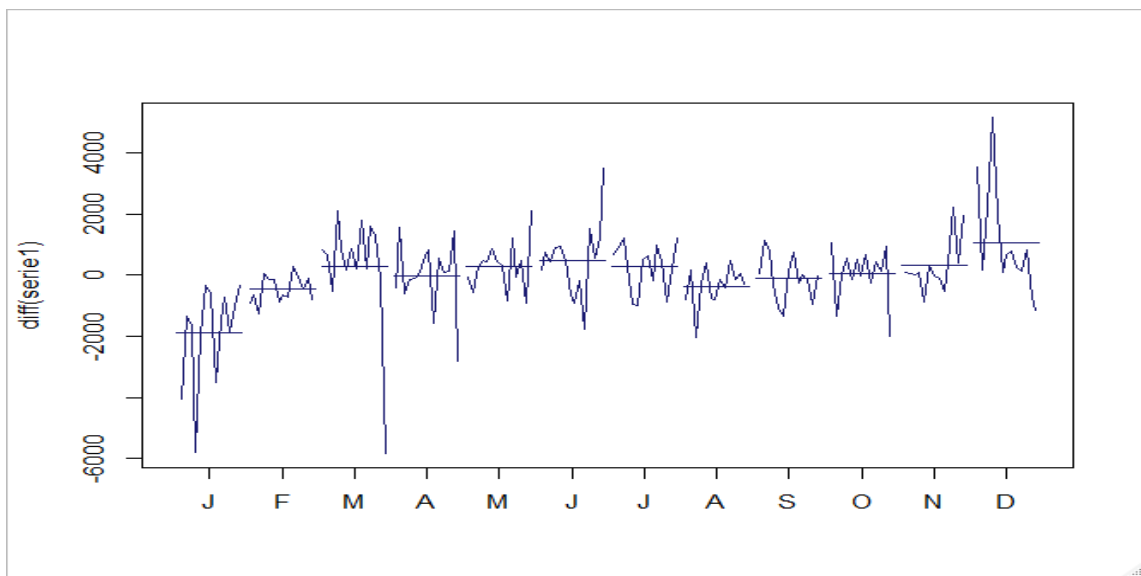


Fuente: Elaboración propia.

La serie ya no es estacionaria, se traza una subseries estacionales de una serie de tiempo, ver Figura 36.

```
monthplot(diff(serie1), col = "midnightblue")
```

Figura 36. Grafica Serie no Estacional – Traza de subseries



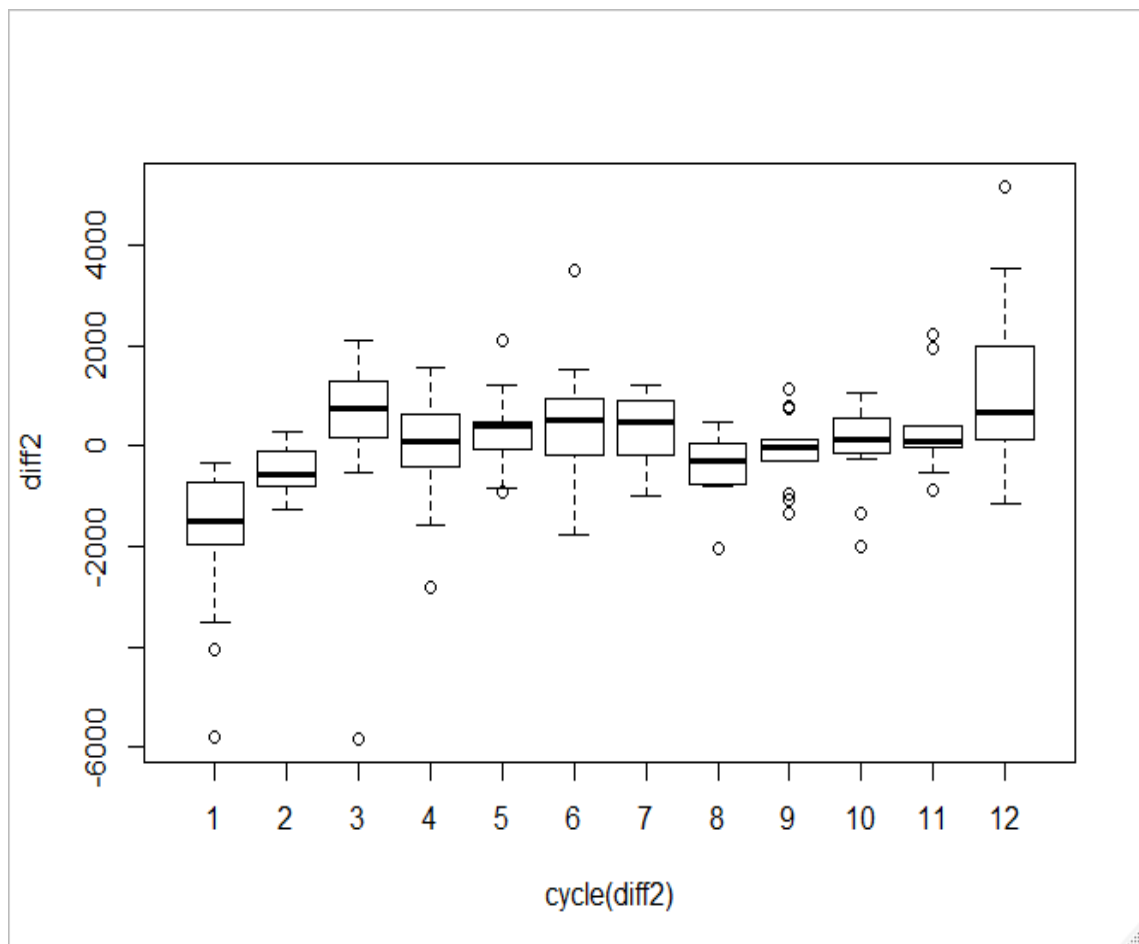
Fuente: Elaboración propia

La diferencia también se las puede presentar en caja como muestra la Figura 37.

```
diff2<-diff(serie1)
```

```
boxplot(diff2~cycle(diff2))
```

Figura 37. Grafica Diferencia – Traza en cajas



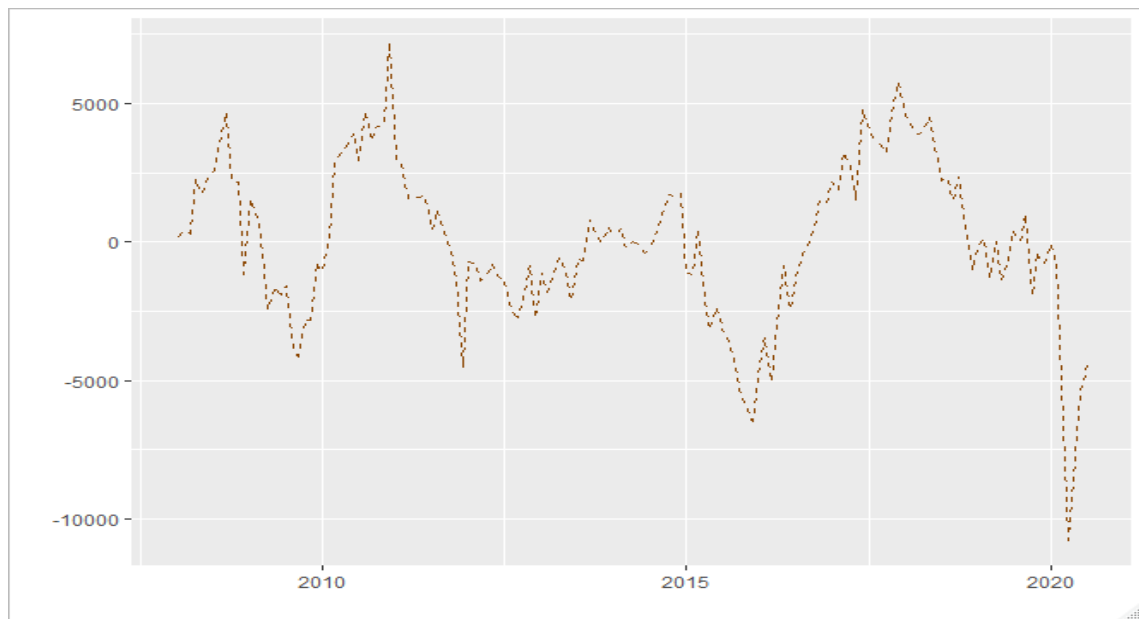
Fuente: Elaboración propia.

Al eliminar el componente estacional tenemos la Figura 38, esto se realiza con el siguiente código

```
diff.serie1.12<-diff(serie1, lag = 12)
```

```
autoplot(diff.serie1.12, ts.colour = "darkorange4", ts.linetype = "dashed")
```

Figura 38. Grafica Eliminar el componente estacional.

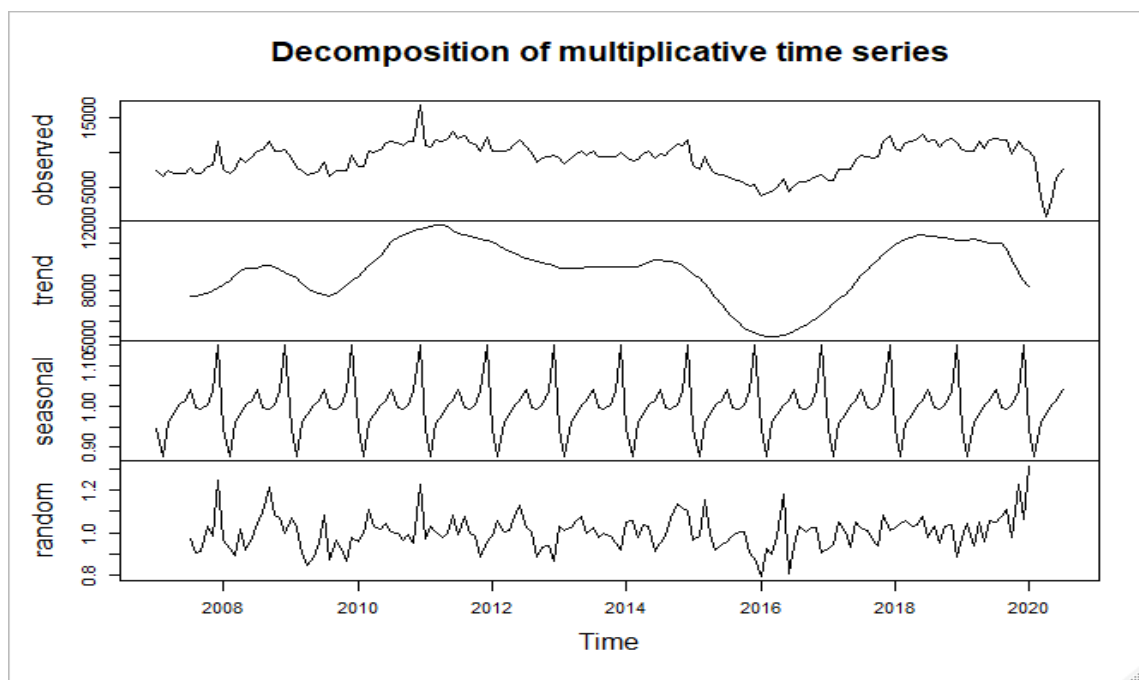


Fuente: Elaboración propia.

Con el siguiente código se realizó la descomposición de la serie que se observa en la Figura 39.

```
deSerie1 <- decompose(serie1, type="multiplicative")  
plot(deSerie1)
```

Figura 39. Grafica descomposición de series de tiempo multiplicativas



Fuente: Elaboración propia.

CAPITULO 4

PRESENTACIÓN Y DISCUSIÓN DE RESULTADOS OBTENIDOS

En esta sección se detalla la presentación ya discusión de los resultados obtenidos en la ejecución del modelo y sus variantes establecido en el capítulo anterior, los resultados son analizados e interpretados en base a los modelos de predicción planteados con el objetivo de proponer el que genere mejores resultados. Para proponer un modelo que cumpla con el objetivo del negocio y de la minería de datos, se realizaron varias pruebas las mismas que se describen a continuación.

4.1 ESTADÍSTICA DESCRIPTIVA EN MINERÍA DE DATOS

Con los datos que se tiene se realizó un análisis estadístico de la información, en la Tabla 9 se muestra las ventas de unidades por mes de los años comprendidos desde enero 2007 hasta julio 2020. En la Tabla 10 se pueden ver las ventas por año del mismo periodo.

Tabla 9. Ventas unidades por mes años 2007 – 2020

	<i>Unidades Mensual</i>	<i>Total Ventas Mensual</i>
Media	8956,49	241290195,71
Error típico	246,70	7058569,76
Mediana	9046	246389974
Moda	8835	#N/A
Desviación estándar	3110,71	89005178,07
Coeficiente variación	34,73	36,89
Varianza de la muestra	9676545,52	7,92192E+15
Curtosis	6,9144	8,2766
Coeficiente de asimetría	1,6736	2,0053
Rango	22554	639653934
Mínimo	754	23601122
Máximo	16834	663255056
Suma	1424082	38365141118
Número de meses	163	163

Fuente: Elaboración propia.

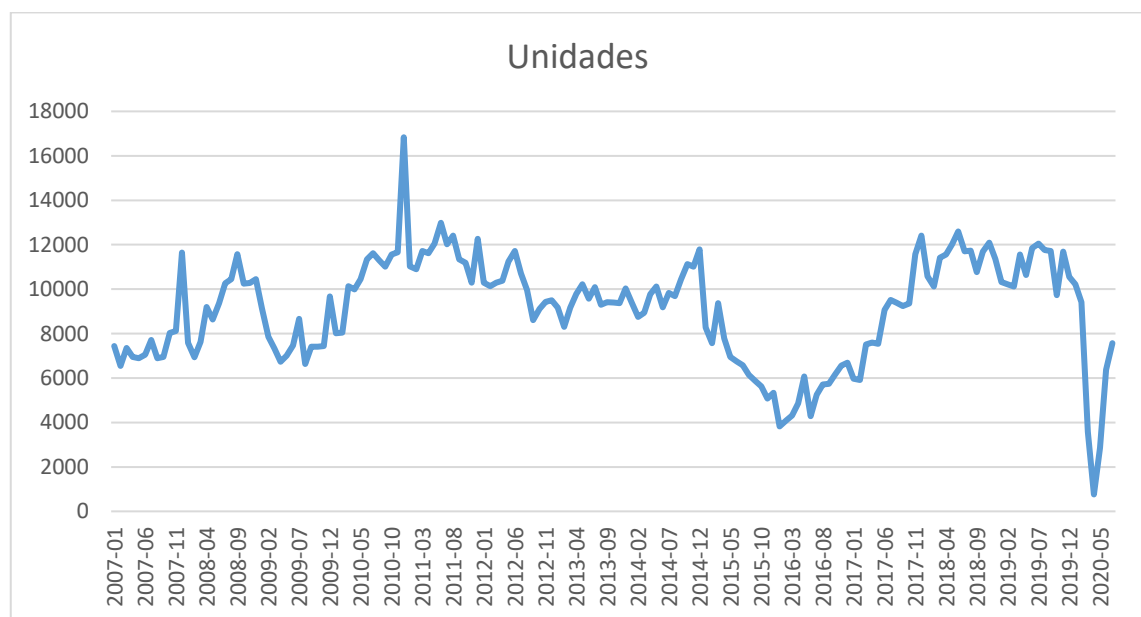
Tabla 10. Ventas por año 2007 – 2020

	<i>Unidades Anuales</i>	<i>Total Ventas Anuales</i>
Media	101720,14	2740367222,71
Error típico	7496,85	207308168,82
Mediana	109044	3052038936
Moda	#N/A	#N/A
Desviación estándar	28050,66	775676141,19
Coefficiente variación	27,58	28,31
Varianza de la muestra	786839666,44	6,01673E+17
Curtosis	0,1168	-0,3262
Coefficiente de asimetría	-0,8855	-0,6234
Rango	93153	2656812285
Mínimo	40119	1122881124
Máximo	133272	3779693409
Suma	1424082	38365141118
Número de Años	14	14

Fuente: Elaboración propia.

Se realizó un gráfico estadístico que se muestra en la Figura 40, en el cual se muestra las ventas por mes del periodo comprendido de enero 2007 a julio 2020.

Figura 40. Ventas por mes periodo enero 2007 a julio 2020

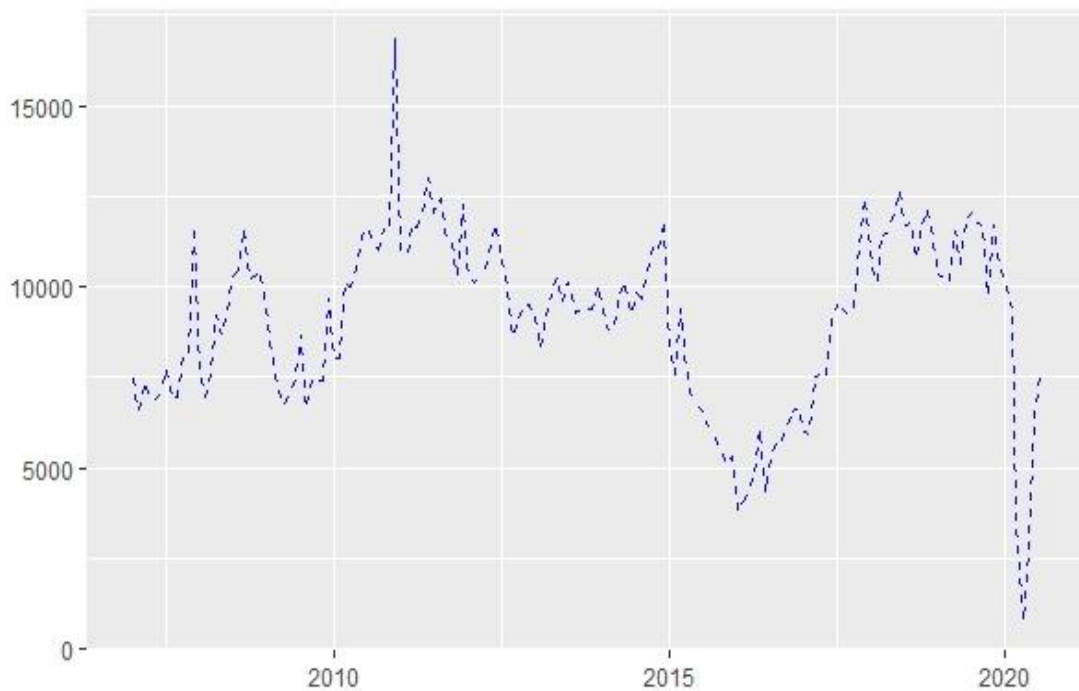


Fuente: Elaboración propia.

Se realizó el mismo gráfico con la herramienta R Studio donde se puede observar que la gráfica es similar a la generado con la técnica descriptiva, ver Figura 41.

```
autoplot(serie1, ts.colour = "blue", ts.linetype = "dashed")
```

Figura 41. Unidades por mes periodo enero 2007 a julio 2020 – Generada con R Studio.

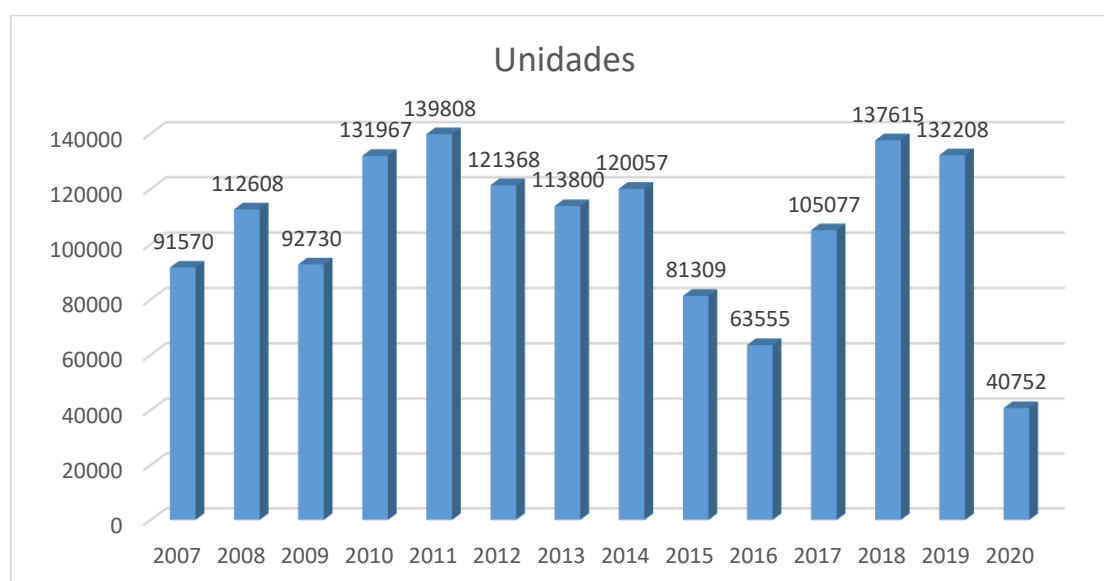


Fuente: Elaboración propia.

Con la información se realizó algunas estadísticas descriptivas de varios parámetros analizados, los que se detallan a continuación.

En la Figura 42 se puede observar el número de ventas por año en el periodo comprendía de enero 2007 a julio 2020, en la gráfica se puede observar que en el año 2011 fue el de mayor venta de vehículo nuevos con un total de 139808 unidades, en segundo lugar, tenemos en año 2018 con 137615 unidades y en tercer lugar el año 2019 con 132208 unidades.

Figura 42. Total de unidades por año

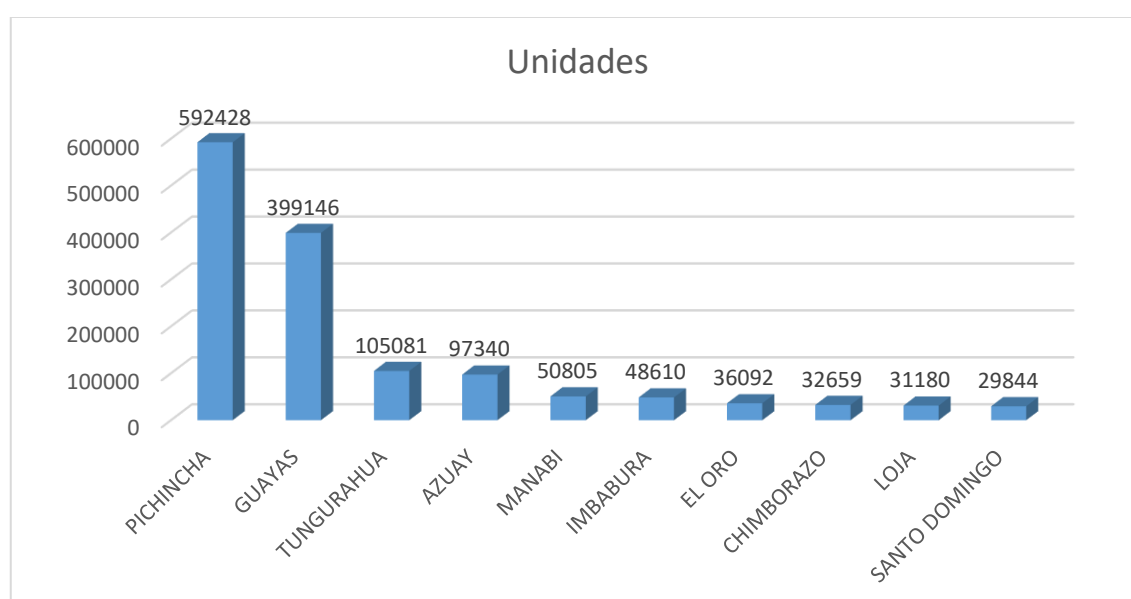


Fuente: Elaboración propia

En los siguientes gráficos se muestra los tops 10 de algunas estadísticas

En la Figura 43 se puede observar las 10 provincias que generaron más ventas de vehículos nuevos en el periodo comprendido de enero 2007 a julio 2020, ver figura 31, como se muestra la provincia de Pichincha está en primer lugar con 592,428 unidades, seguida de Guayas con 399,146 siendo las 2 provincias con mayor participación en el mercado automotriz.

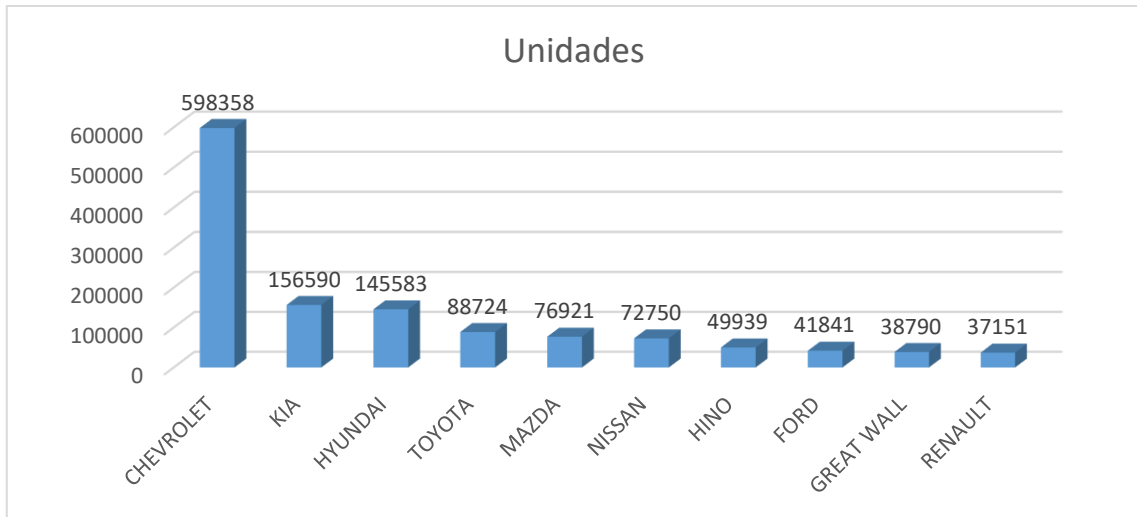
Figura 43. Top 10 Ventas por Provincia.



Fuente: Elaboración propia.

En la Figura 44, se muestra el top 10 de las marcas con mayor participación en el mercado ecuatoriano tomando en cuenta el mismo periodo de tiempo, observado en primer lugar la marca Chevrolet con un total de 598358 unidades, en segundo puesto Kia con 156690 y en tercer lugar Hyundai con 145583. Podemos observar que la marca Chevrolet es la de mayor aceptación en el país, y que las 2 marcas que le siguen están compitiendo muy de cerca 2 por el segundo lugar.

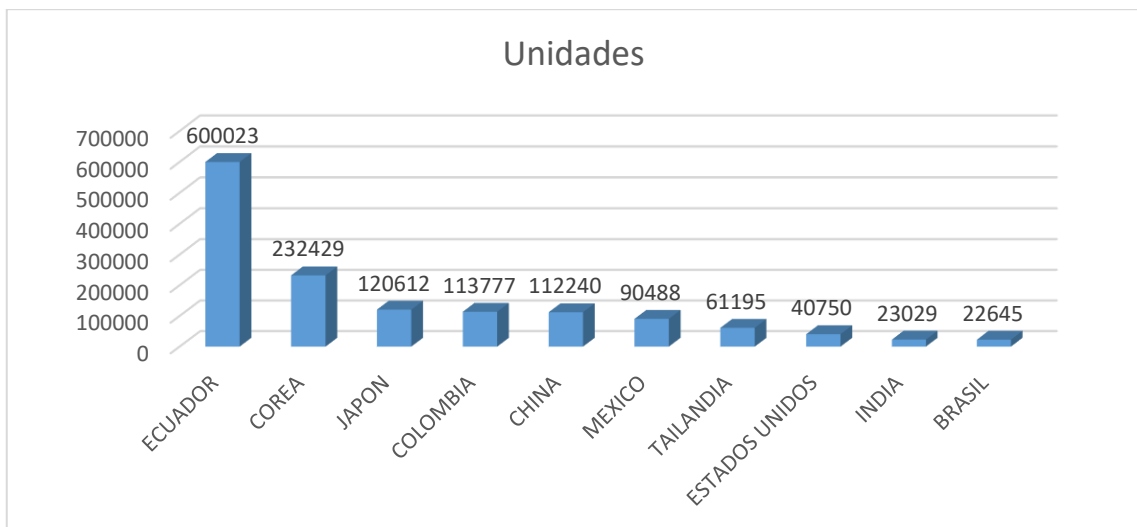
Figura 44. Top 10 Ventas por Marca.



Fuente: Elaboración propia.

En cuanto al país de origen de los vehículos tenemos que los vehículos más vendidos son de ensamblados en el Ecuador con un total de 600023, seguido de Corea con 232429, esto en el periodo de enero 2007 a julio 2020, como se muestra en la Figura 45.

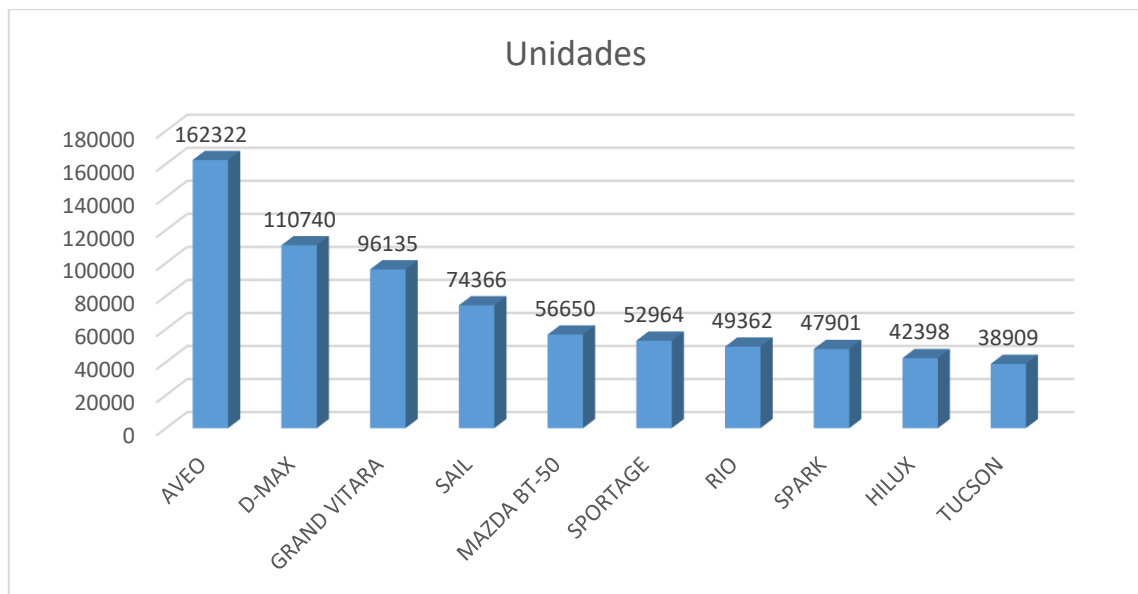
Figura 45. Top 10 Ventas por país de procedencia.



Fuente: Elaboración propia.

Por último, se muestra el top 10 de los modelos más vendidos en este mismo periodo que se tiene la información encontrando el siguiente análisis, que el modelo Aveo fue el más vendido con un total de 162322 unidades, en segundo puesto D-Max con 110740 unidades. En este análisis se observa la tendencia de preferencia de la Marca Chevrolet que coloca 5 modelos (Aveo, D-Max, Grand Vitara, Sail y Spark) en el top 10 de ventas. La marca Kia coloca 2 modelos como (Sportage y Rio), Hyundai un modelo que es Tucson, esto se puede observar en la Figura 46.

Figura 46. Top 10 Ventas por modelo.



Fuente: Elaboración propia.

4.2 TÉCNICA DE MINERÍA DE DATOS UTILIZADA

Se procedió a realizar el análisis de los datos con varias técnicas y sus modelos para realizar las predicciones en el sector automotriz, para esto determinó utilizar series temporales.

4.2.1 SERIES TEMPORALES

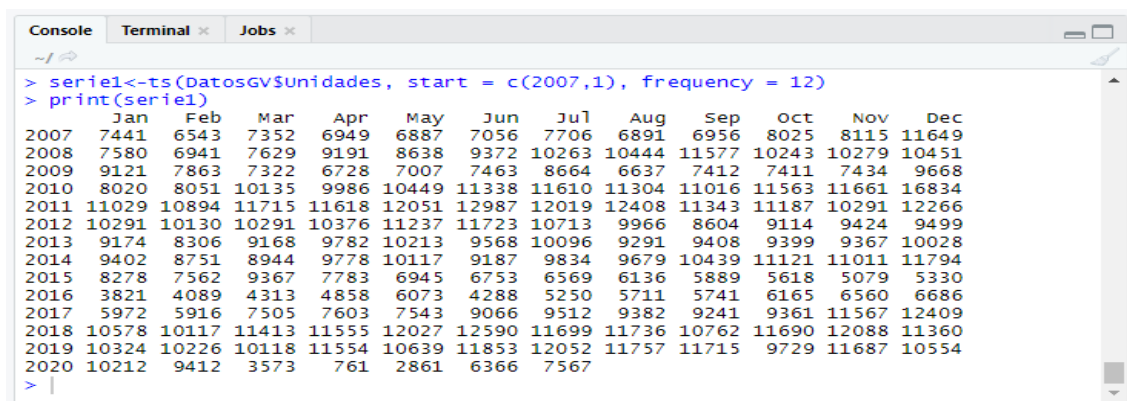
Utilizando la herramienta R Studio aplicamos varios modelos de series temporales para poder realizar el análisis predictivo para la presente investigación.

Una **serie temporal** es una secuencia de datos, observaciones o valores, medidos en determinados momentos y ordenados cronológicamente.

Primero realizamos un resumen de la información en número de unidades vendidas por mes en el Ecuador como se muestra en la Figura 47.

```
serie1<-ts(DatosGV$Unidades, start = c(2007,1), frequency = 12)
print(serie1)
```

Figura 47. Unidades Vendidas por Mes – Visualizadas en la Herramienta R Studio.



Fuente: Elaboración propia.

En la Tabla 11 se puede ver la información resumida almacenada en el dataset.

Tabla 11. Información Dataset número de unidades.

Año / Mes	ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SEP	OCT	NOV	DIC
2007	7441	6543	7352	6949	6887	7056	7706	6891	6956	8025	8115	11649
2008	7580	6941	7629	9191	8638	9372	10263	10444	11577	10243	10279	10451
2009	9121	7863	7322	6728	7007	7463	8664	6637	7412	7411	7434	9668
2010	8020	8051	10135	9986	10449	11338	11610	11304	11016	11563	11661	16834
2011	11029	10894	11715	11618	12051	12987	12019	12408	11343	11187	10291	12266
2012	10291	10130	10291	10376	11237	11723	10713	9966	8604	9114	9424	9499
2013	9174	8306	9168	9782	10213	9568	10096	9291	9408	9399	9367	10028
2014	9402	8751	8944	9778	10117	9187	9834	9679	10439	11121	11011	11794
2015	8278	7562	9367	7783	6945	6753	6569	6136	5889	5618	5079	5330
2016	3821	4089	4313	4858	6073	4288	5250	5711	5741	6165	6560	6686
2017	5972	5916	7505	7603	7543	9066	9512	9382	9241	9361	11567	12409
2018	10578	10117	11413	11555	12027	12590	11699	11736	10762	11690	12088	11360
2019	10324	10226	10118	11554	10639	11853	12052	11757	11715	9729	11687	10554
2020	10212	9412	3573	761	2861	6366	7567					

Fuente: Elaboración propia.

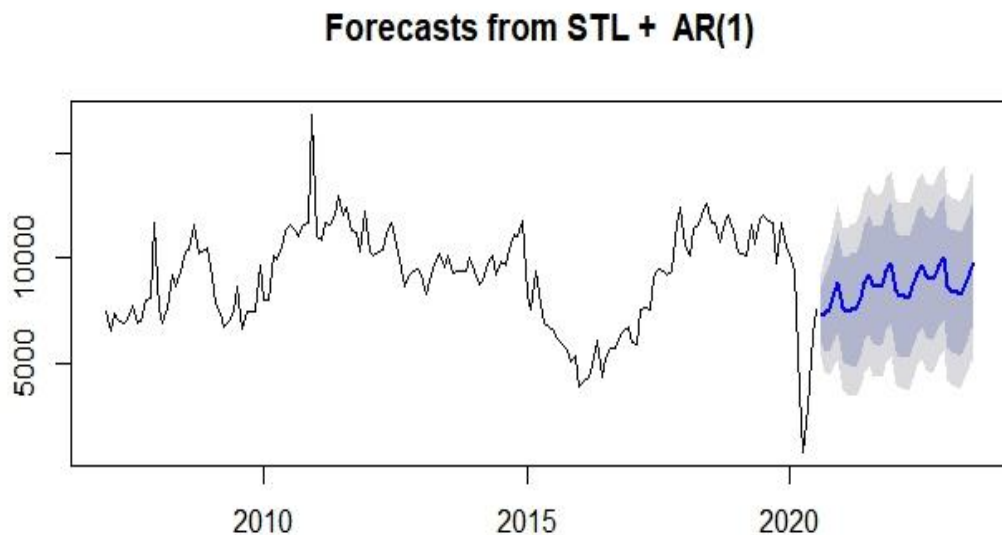
A continuación, se aplicaron los algunos de los modelos que posee la técnica de series temporales.

4.2.1.1 MODELO STLM

Se realizó el análisis en RStudio para el modelo STLM con obtenido los siguientes resultados ver Figura 48.

```
fitDG1 <- stlm(serie1, modelfunction=ar)
fcDG1<- forecast(fitDG1, h=36)
plot(fcDG1)
```

Figura 48. Aplicación Modelo STL.



Fuente: Elaboración propia.

En la Tabla 12 se muestran los resultados de las predicciones generados por el modelo STL.

Tabla 12. Predicciones Modelo STL.

Mes	2020	2021	2022	2023
Ene		7629,318	8464,394	8679,392
Feb		7437,923	8183,717	8375,728

Mar		7521,378	8187,434	8358,916
Abr		7528,794	8123,638	8276,786
May		8074,708	8605,954	8742,727
Jun		8821,488	9295,935	9418,085
Jul		9212,783	9636,504	9745,595
Ago	7313,134	8782,957	9161,375	
Sep	7332,635	8645,311	8983,270	
Oct	7567,834	8740,163	9041,989	
Nov	8351,212	9398,200	9667,756	
Dic	8860,319	9795,367	10036,103	

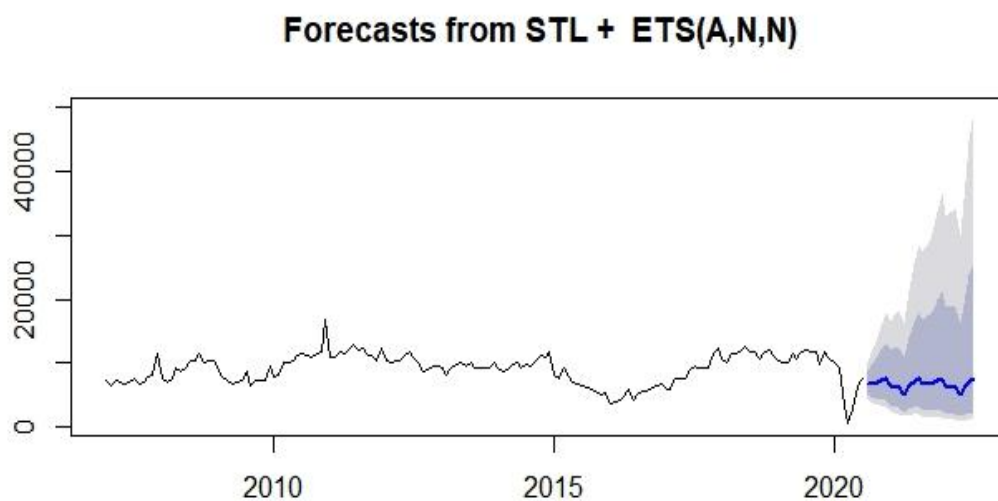
Fuente: Elaboración propia.

4.2.1.2 MODELO STLF

Se realizó el análisis en R para el modelo STLF del cual tenemos estos resultados, ver Figura 49.

```
fitDG2 <- stlf(serie1, lambda=0)
fcDG2<- forecast(fitDG2, 36)
plot(fcDG2)
```

Figura 49. Aplicación Modelo STLF.



Fuente: Elaboración propia.

En la Tabla 13 se observan los resultados de las predicciones generados por el modelo STLF en números unidades por mes, para los próximos 36 meses.

Tabla 13. Predicciones Modelo STLF.

Mes	2020	2021	2022	2023
Ene		6498,158	6498,158	6498,158
Feb		6332,216	6332,216	6332,216
Mar		6129,871	6129,871	6129,871
Abr		5170,335	5170,335	5170,335
May		6376,421	6376,421	6376,421
Jun		7126,249	7126,249	7126,249
Jul		7566,915	7566,915	7566,915
Ago	6868,456	6868,456	6868,456	
Sep	6767,429	6767,429	6767,429	
Oct	6838,829	6838,829	6838,829	
Nov	7261,293	7261,293	7261,293	
Dic	7554,686	7554,686	7554,686	

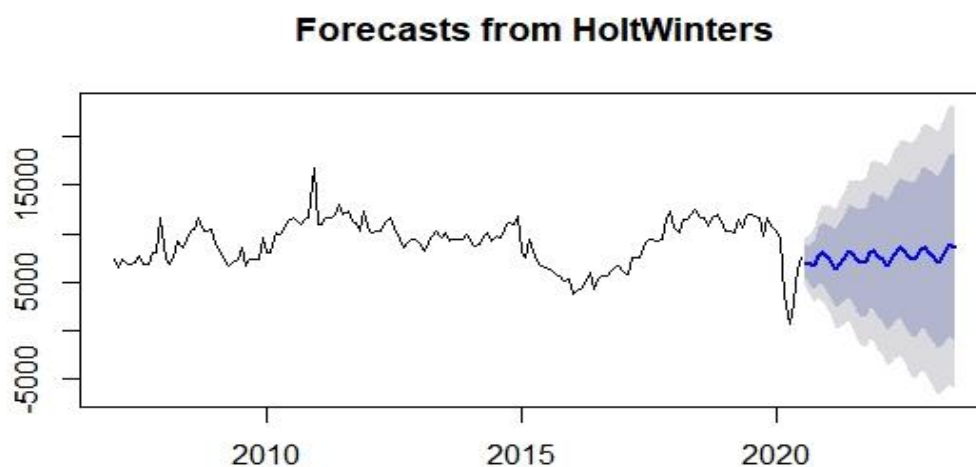
Fuente: Elaboración propia.

4.2.1.3 MODELO HOLTWINTERS

Al realizar el análisis en RStudio para el modelo HoltWinters se obtuvo los siguientes resultados, ver Figura 50.

```
fitDG3 <- HoltWinters(serie1)
fcDG3 <- forecast(fitDG3, 36)
plot(fcDG3)
```

Figura 50. Aplicación HoltWinters.



Fuente: Elaboración propia.

En la Tabla 14, se puede ver los resultados de las predicciones generados por el modelo Holt Winters en números unidades para los próximos 36 meses.

Tabla 14. Predicciones Holt Winters.

Mes	2020	2021	2022	2023
Ene		7490,613	7813,217	8135,821
Feb		7048,298	7370,903	7693,507
Mar		6386,628	6709,232	7031,837
Abr		6847,661	7170,265	7492,870
May		7551,186	7873,791	8196,395
Jun		8224,423	8547,028	8869,632
Jul		7889,604	8212,209	8534,813
Ago	7096,440	7419,044	7741,648	
Sep	6788,491	7111,095	7433,700	
Oct	6735,151	7057,755	7380,360	
Nov	7739,383	8061,987	8384,591	
Dic	7993,841	8316,446	8639,050	

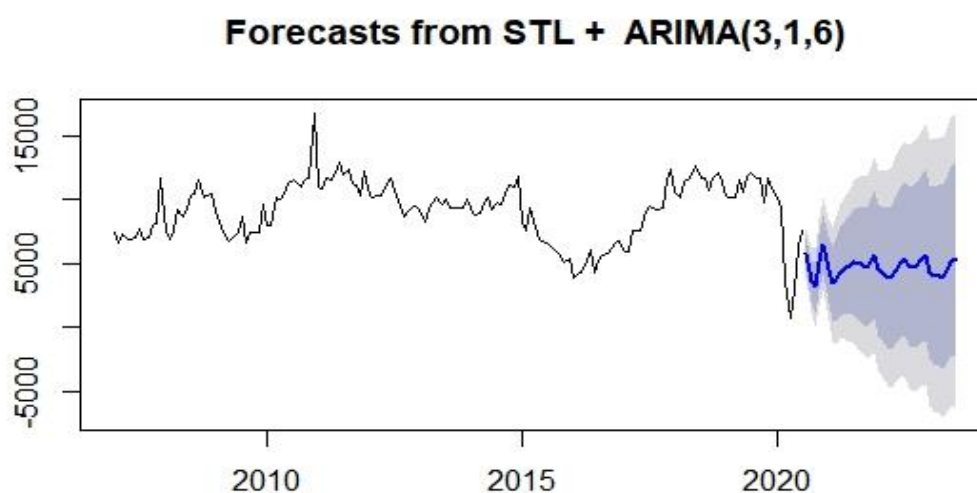
Fuente: Elaboración propia.

4.2.1.4 MODELO STLM ARIMA 3,1,6

Con la herramienta RStudio se realizó el análisis para el modelo STLM ARIMA 3.1.6 con los siguientes resultados obtenidos, observar Figura 51.

```
fitDG4 <- stlm(serie1, modelfunction=Arima, order=c(3,1,6))  
fcDG4 <- forecast(fitDG4, 36)  
plot(fcDG4)
```

Figura 51. Aplicación STLM ARIMA 3,1,6.



Fuente: Elaboración propia.

En la Tabla 15 se muestran los resultados obtenidos de las predicciones generados por el modelo STLM ARIMA 3.1.6, para los siguientes 36 meses en números de unidades

Tabla 15. Predicciones Modelo STLF ARIMA 3,1,6.

Mes	2020	2021	2022	2023
Ene		4533,927	4512,268	4345,597
Feb		3385,294	4093,987	4069,379
Mar		3692,175	3906,012	4028,814
Abr		4342,925	3863,021	3909,356

May		4723,157	4459,703	4375,516
Jun		4811,033	5125,401	5070,958
Jul		5067,734	5342,356	5395,012
Ago	5814,294	5011,333	4831,402	
Sep	3612,249	4967,547	4716,391	
Oct	3112,795	4712,633	4795,269	
Nov	5111,855	5145,950	5357,777	
Dic	6511,724	5695,678	5679,357	

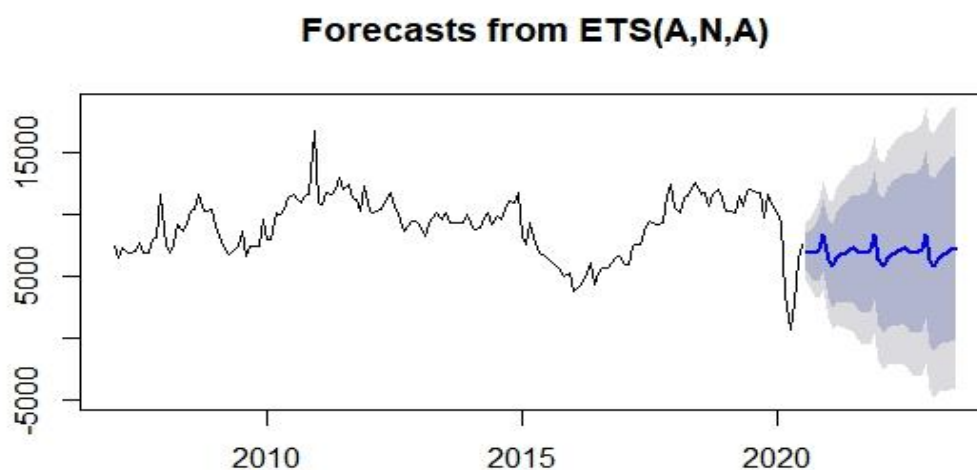
Fuente: Propia.

4.2.1.5 MODELO EST SUAVIZACIÓN EXPONENCIAL

Al realizar el análisis del modelo EST suavización exponencial en RStudio se obtuvo los siguientes resultados, ver Figura 52.

```
fitDG5 <- ets(serie1)
fcDG5<- forecast(fitDG5, h=36)
plot(fcDG5)
```

Figura 52. Aplicación Modelo EST suavización exponencial.



Fuente: Elaboración propia.

En la Tabla 16 se puede ver los resultados de las predicciones generados por el modelo EST suavización exponencial en números unidades para los próximos 36 meses.

Tabla 16. Predicciones Modelo EST suavización exponencial.

Mes	2020	2021	2022	2023
Ene		6409,187	6409,187	6409,187
Feb		5876,246	5876,246	5876,246
Mar		6500,357	6500,357	6500,357
Abr		6770,045	6770,045	6770,045
May		6917,078	6917,078	6917,078
Jun		7144,543	7144,543	7144,543
Jul		7293,120	7293,120	7293,120
Ago	6994,492	6994,492	6994,492	
Sep	6863,876	6863,876	6863,876	
Oct	6922,005	6922,005	6922,005	
Nov	7241,471	7241,471	7241,471	
Dic	8339,327	8339,327	8339,327	

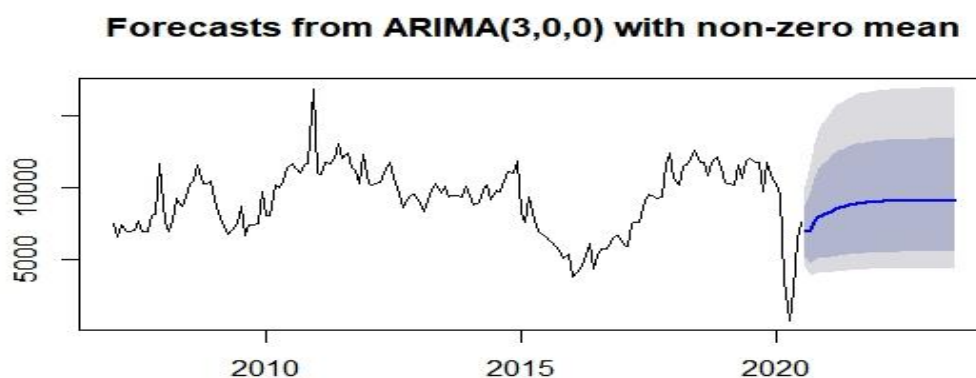
Fuente: Elaboración propia.

4.2.1.6 MODELO AUTO.ARIMA

Empleando la función Autoarima de R, del paquete forecast, variante del modelo ARIMA que sugiere R se obtuvieron los siguientes resultados, ver Figura 53.

```
fitDG6 <- auto.arima(serie1, lambda=0, biasadj=TRUE)
fcDG6<- forecast(fitDG6, h=36)
plot(fcDG6)
```

Figura 53. Aplicación Modelo Auto.Arima.



Fuente: Elaboración propia.

La Tabla 17 se muestran los resultados de las predicciones generados por el modelo Auto.Arima en números unidades por mes, para los próximos 36 meses.

Tabla 17. Predicciones Modelo Auto.Arima.

Mes	2020	2021	2022	2023
Ene		8194,169	8999,281	9087,332
Feb		8354,485	9015,843	9089,176
Mar		8486,018	9029,639	9090,712
Abr		8586,321	9041,128	9091,991
May		8669,737	9050,696	9093,057
Jun		8741,583	9058,665	9093,944
Jul		8801,700	9065,302	9094,683
Ago	6877,067	8851,240	9070,829	
Sep	6997,225	8892,406	9075,432	
Oct	7551,955	8926,821	9079,266	
Nov	7895,373	8955,524	9082,459	
Dic	8046,191	8979,403	9085,118	

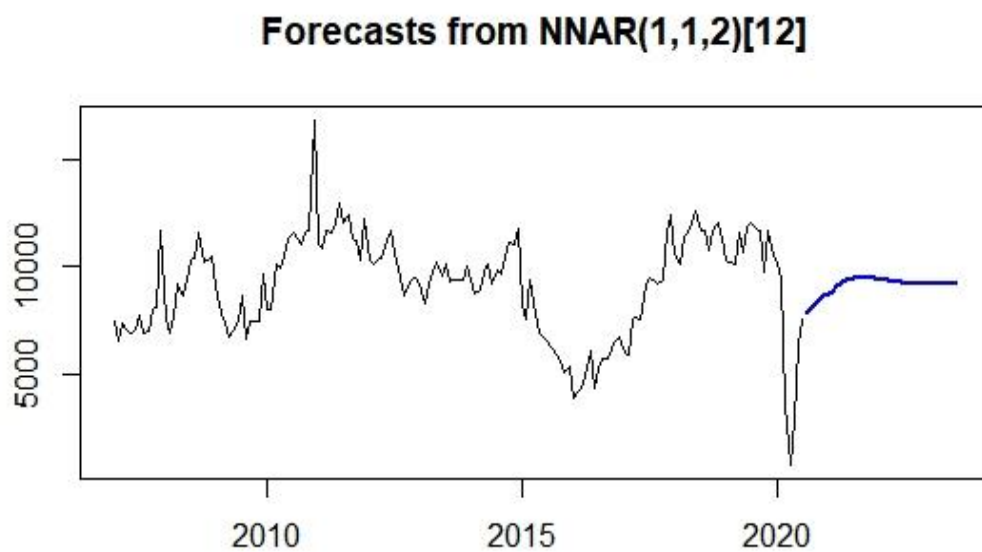
Fuente: Elaboración propia.

4.2.1.7 MODELO NNETAR

Al realizar el análisis del modelo NNETAR suavización exponencial en RStudio se obtuvo los siguientes resultados, observar la Figura 54.

```
fitDG7 <- nnetar(serie1)
fcDG7 <- forecast(fitDG7, h=36)
plot(fcDG7)
```

Figura 54. Aplicación NNETAR.



Fuente: Elaboración propia.

En la Tabla 18 se visualizan los resultados de las predicciones generados por el modelo NNETAR para los próximos 36 meses en número de unidades.

Tabla 18. Predicciones Modelo NNETAR

Mes	2020	2021	2022	2023
Ene		8747,803	9424,127	9186,697
Feb		8833,918	9393,499	9182,405
Mar		9090,117	9343,817	9178,077
Abr		9208,113	9304,396	9173,950

May		9346,676	9275,256	9170,189
Jun		9426,395	9253,568	9166,860
Jul		9465,112	9237,057	9163,967
Ago	7817,342	9454,242	9223,644	
Sep	8066,893	9445,707	9212,782	
Oct	8253,773	9439,862	9204,022	
Nov	8473,824	9436,598	9196,996	
Dic	8626,694	9433,566	9191,358	

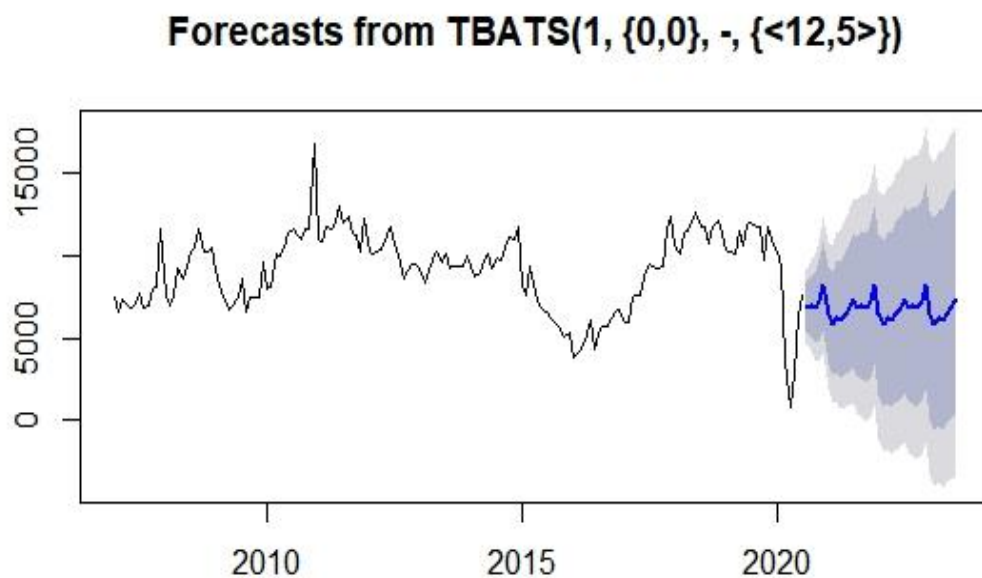
Fuente: Elaboración propia.

4.2.1.8 MODELO TBATS

Se aplica este modelo para nuestros datos en la figura se muestra el grafico generado por la aplicación, ver Figura 55.

```
fitDG8 <- tbats(serie1, biasadj=TRUE)
fcDG8<- forecast(fitDG8, h=36)
plot(fcDG8)
```

Figura 55. Aplicación modelo TBATS.



Fuente: Elaboración propia.

En la Tabla 19 se muestran los resultados obtenidos de las predicciones generados por el Modelo TBATS, para los siguientes 36 meses en números de unidades

Tabla 19. Predicciones Modelo TBATS.

Mes	2020	2021	2022	2023
Ene		6430,604	6430,604	6430,604
Feb		5830,764	5830,764	5830,764
Mar		6255,319	6255,319	6255,319
Abr		6092,617	6092,617	6092,617
May		6525,11	6525,11	6525,11
Jun		6880,532	6880,532	6880,532
Jul		7297,468	7297,468	7297,468
Ago	6892,874	6892,874	6892,874	
Sep	6946,146	6946,146	6946,146	
Oct	6833,736	6833,736	6833,736	
Nov	7281,591	7281,591	7281,591	
Dic	8205,713	8205,713	8205,713	

Fuente: Elaboración propia.

4.3 INTERPRETACIÓN DE RESULTADOS

Para interpretar los resultados se realizó los análisis de datos, se utilizaron métodos matemáticos que permites obtener resultados comprensibles y válidos para comparar que modelo es el más efectivo para su aplicación, para esto se consideraron las técnicas predictivas como STLM, STLF, Holt Winters, STLM ARIMA 3.1.6, EST, Auto Arima, NNETAR y TBATS. Con estos modelos de desarrollaron los análisis para determinar cuál puede ser el comportamiento en las ventas mensuales en el sector automotriz en el Ecuador. Para seleccionar el modelo de datos más efectivo se tomaron a consideración los siguientes parámetros MAE (Error medio absoluto) que calcula la función de error absoluto medio para el pronóstico y los resultados posibles y MAPE (El error porcentual absoluto medio) que calcula la función de error (desviación) promedio, porcentaje absoluto para los pronósticos y los resultados eventuales. Con respecto a la minería de datos se consideró utilizar la metodología TDSP con cada una de las etapas del ciclo de vida para el proyecto y lograr una buena ejecución, estas fases son: entendimiento del

negocio, adquisición y entendimiento de los datos, modelado, despliegue y aceptación del cliente.

4.3.1 COMPARACIÓN RESULTADOS DE MODELO DE SERIES TEMPORALES

La comparación de los **accuracy** de los ocho modelos analizados en la técnica de series temporales podemos observar que la que muestra mejores resultados en el modelo **STLM arima 3,1,6** que es el que género menor valor en MAE y MAPE parámetro generados de las Accuracy de cada modelo, ver Tabla 20.

Tabla 20. Comparación modelos serie temporales resultados Accuracy

Accuracy	Training set						
Modelo	ME	RSME	MAE	MPE	MAPE	MASE	ACF1
STLM	-4,827153	1005,956	696,9607	-3,917063	11,01758	0,3168499	-0,4334499
STLF	-5,946314	1088,442	758,5781	-2,345014	10,98814	0,3448593	-0,1305483
HoltWinters	-129,4626	1265,373	891,4715	-5,024545	14,28879	0,4052777	0,03951072
STLM arima 3,1,6	-19,10044	947,8423	659,1263	-3,202483	10,37173	0,2996498	0,00138171
EST	-15,22073	1114,854	744,6866	-4,081202	12,48604	0,3385469	0,01668561
Auto.Arima	-40,976	1279,692	882,7532	-4,944755	13,30052	0,4013143	-0,1604401
NNETAR	0,3549693	1150,861	776,7189	-4,373725	12,07662	0,3531093	0,08537357
TBATS	-9,504907	1103,917	735,7844	-3,786421	12,03909	0,3344998	0,01607409

Fuente: Elaboración propia

El modelo **STLM Arima 3,1,6** genero un **MAE** de **659,1263** y un **MAPE** de **10,37173** por estos parámetros se considera el modelo más óptimo para la predicción del análisis realizado para ventas de numero unidades mensuales en el sector automotriz.

4.3.2 DATOS DE PREDICCION DE LOS MODELOS

Se realizó una comparación de los ocho modelos de series temporales con los cuales se analizaron los datos los mismos que los comparamos con las unidades de ventas reales de agosto a noviembre 2020, para observar que modelo arrojo mejores resultado versus la información real a la actualidad, ver Tabla 21.

Tabla 21. Comparación modelos serie temporales vs datos reales de unidades

Modelos de Series Temporales										
Año	Mes	STLM	STLF	Holt Winters	STLM arima 3,1,6	EST	Auto.Arima	NNETAR	TBATS	Ventas Real
2020	Ago	7313	6868	7096	5814	6994	6877	7817	6893	7686
	Sep	7333	6767	6788	3612	6864	6997	8067	6946	8665
	Oct	7568	6839	6735	3113	6922	7552	8254	6834	9924
	Nov	8351	7261	7739	5112	7241	7895	8474	7282	9941
	Dic	8860	7555	7994	6512	8339	8046	8627	8206	9550
2021	Ene	7629	6498	7491	4534	6409	8194	8748	6431	8491
	Feb	7438	6332	7048	3385	5876	8354	8834	5831	8100
	Mar	7521	6130	6387	3692	6500	8486	9090	6255	9789

Fuente: Elaboración propia

En cambio, según los datos generados por el MAE y MAPE, el modelo que mejor predicción género es STLM ARIMA 3,1,6. Ver Tabla 22.

Tabla 22. Datos de predicción Modelo STLM arima 3,1,6, - Mejores resultados

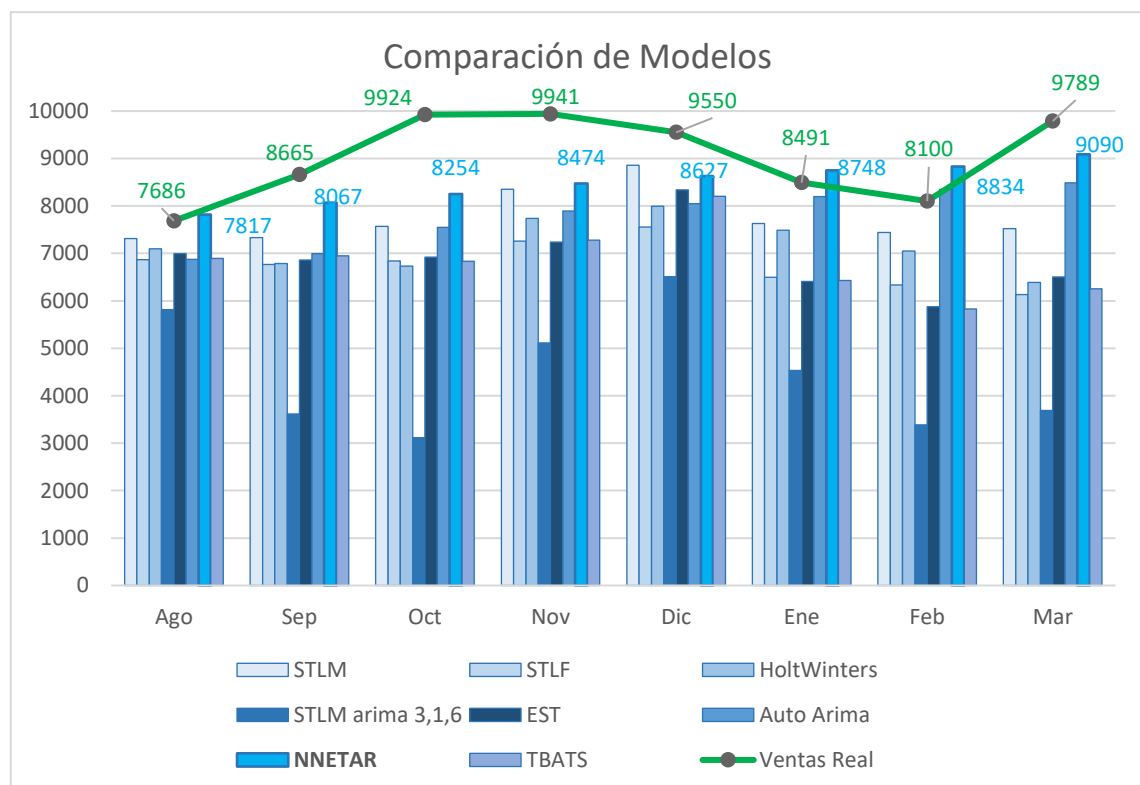
Mes	2020	2021	2022	2023
Ene		4534	4512	4346
Feb		3385	4094	4069
Mar		3692	3906	4029
Abr		4343	3863	3909
May		4723	4460	4376
Jun		4811	5125	5071
Jul		5068	5342	5395
Ago	5814	5011	4831	
Sep	3612	4968	4716	

Oct	3113	4713	4795	
Nov	5112	5146	5358	
Dic	6512	5696	5679	

Fuente: Elaboración propia

En la Figura 56 se muestra las comparaciones de todos los modelos con las ventas reales donde se observa que el modelo que más se acerca con los datos de predicción es el modelo **NNETAR**.

Figura 56. Comparación de modelo de series temporales vs datos reales



Fuente: Elaboración propia

4.3.3 COMPARACIÓN DATOS PREDICTIVOS CON DATOS REALES

En este punto presentamos las comparaciones de los datos de predicciones generados por cada modelo versus los datos reales de ventas obtenidos de los meses de agosto 2020 a marzo 2021, para comprobar que modelo tuvo mayor exactitud.

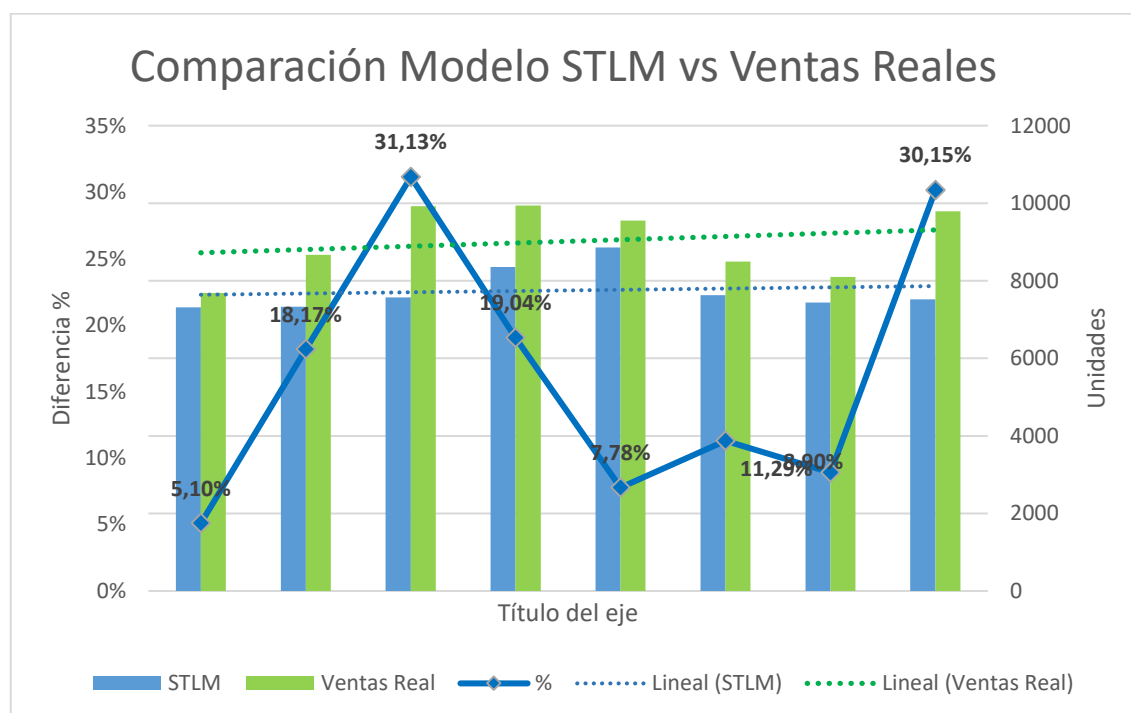
En la Tabla 23 se muestra la comparación del modelo STLM con los datos reales, también se muestra una gráfica de estos datos, ver Figura 57.

Tabla 23. Comparación Modelo STLM vs Datos reales.

Año	Mes	STLM	Ventas Real	Diferencia	%
2020	Ago	7313	7.686	373	5,10%
	Sep	7333	8.665	1332	18,17%
	Oct	7568	9.924	2356	31,13%
	Nov	8351	9.941	1590	19,04%
	Dic	8860	9.550	690	7,78%
2021	Ene	7629	8.491	862	11,29%
	Feb	7438	8.100	662	8,90%
	Mar	7521	9.789	2268	30,15%

Fuente: Elaboración propia.

Figura 57. Modelo TSLM vs Datos Reales.



Fuente: Elaboración propia.

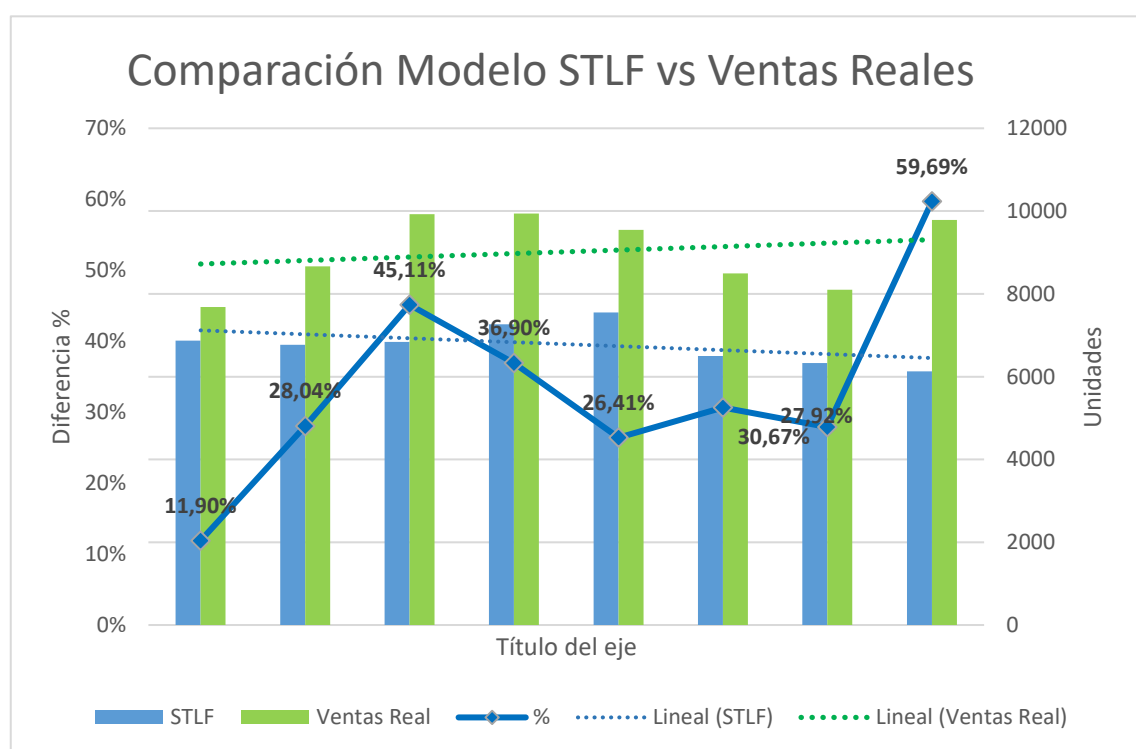
En la Tabla 24 se muestra la comparación del modelo STLTF con los datos reales, además se muestra una gráfica con esta comparación, esto se muestra en la Figura 58.

Tabla 24. Comparación Modelo STLTF vs Datos reales.

Año	Mes	STLTF	Ventas Real	Diferencia	%
2020	Ago	6868	7.686	818	11,90%
	Sep	6767	8.665	1898	28,04%
	Oct	6839	9.924	3085	45,11%
	Nov	7261	9.941	2680	36,90%
	Dic	7555	9.550	1995	26,41%
2021	Ene	6498	8.491	1993	30,67%
	Feb	6332	8.100	1768	27,92%
	Mar	6130	9.789	3659	59,69%

Fuente: Elaboración propia.

Figura 58. Modelo TSLF vs Datos Reales.



Fuente: Elaboración propia.

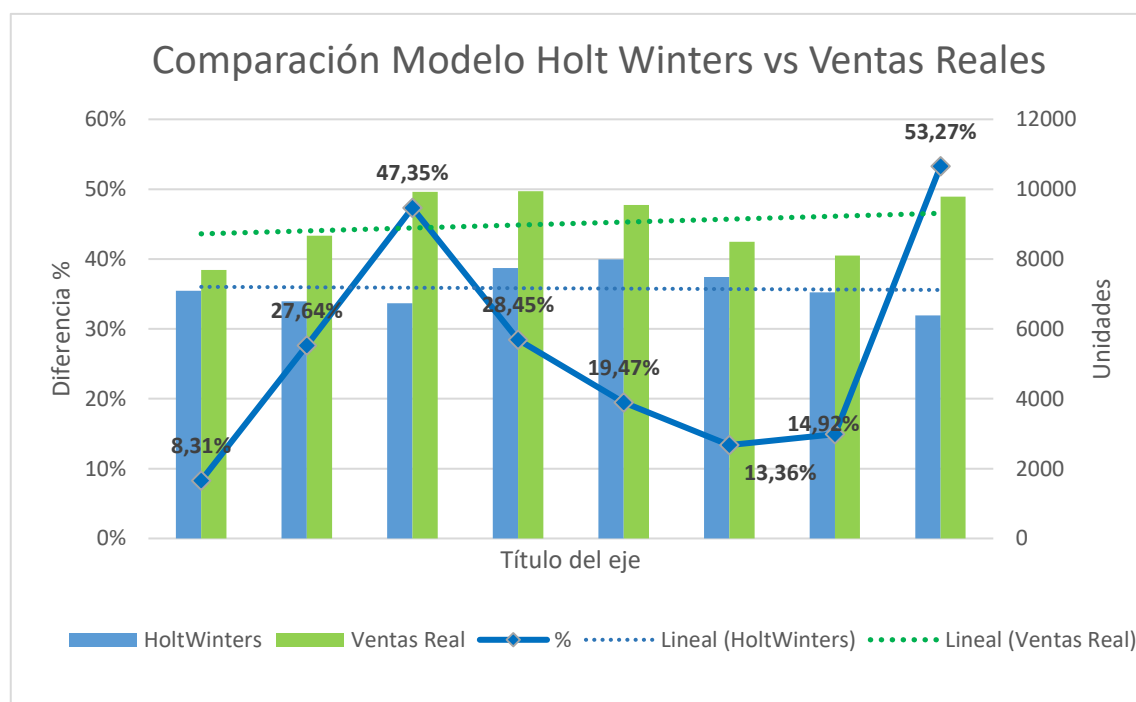
En la Tabla 25 se muestra la comparación del modelo Holt Winters con los datos reales, además se muestra una gráfica con esta comparación, esto se muestra en la Figura 59.

Tabla 25. Comparación Modelo Holt Winters vs Datos reales

Año	Mes	Holt Winters	Ventas Real	Diferencia	%
2020	Ago	7096	7.686	590	8,31%
	Sep	6788	8.665	1877	27,64%
	Oct	6735	9.924	3189	47,35%
	Nov	7739	9.941	2202	28,45%
	Dic	7994	9.550	1556	19,47%
2021	Ene	7491	8.491	1000	13,36%
	Feb	7048	8.100	1052	14,92%
	Mar	6387	9.789	3402	53,27%

Fuente: Elaboración propia.

Figura 59. Modelo Holt Winters vs Datos Reales.



Fuente: Elaboración propia.

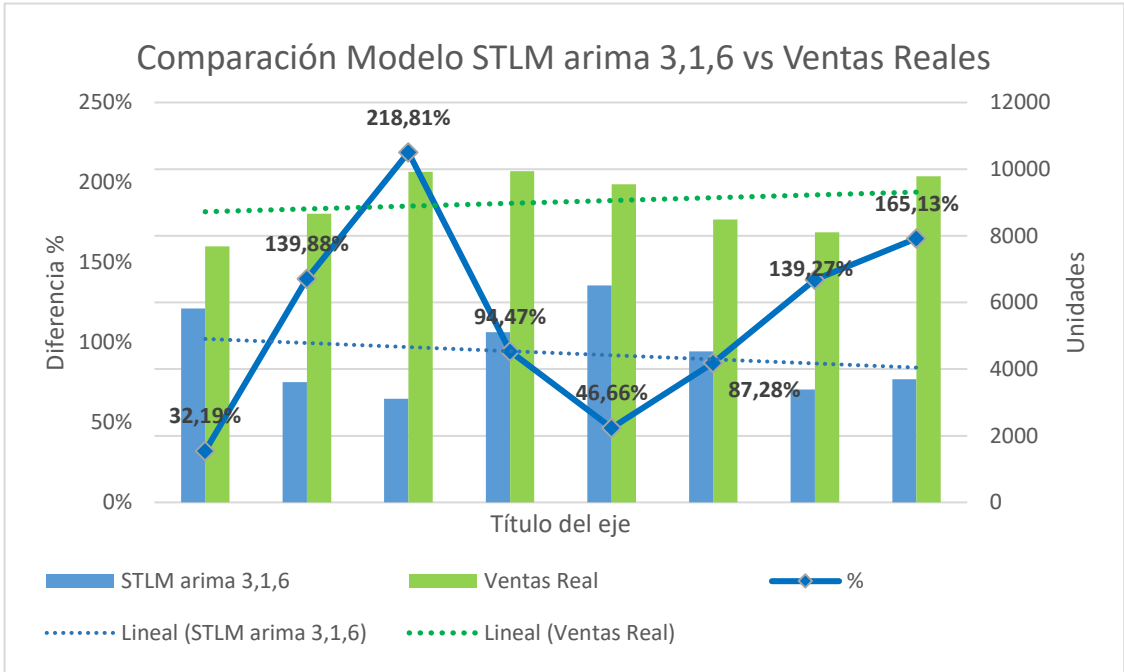
En la Tabla 26 se muestra la comparación del modelo STLM arima 3,1,6 con los datos reales, adicional tenemos una gráfica de estos datos, ver Figura 60.

Tabla 26. Comparación Modelo STLM arima 3,1,6 vs Datos reales

Año	Mes	STLM arima 3,1,6	Ventas Real	Diferencia	%
2020	Ago	5814	7.686	1872	32,19%
	Sep	3612	8.665	5053	139,88%
	Oct	3113	9.924	6811	218,81%
	Nov	5112	9.941	4829	94,47%
	Dic	6512	9.550	3038	46,66%
2021	Ene	4534	8.491	3957	87,28%
	Feb	3385	8.100	4715	139,27%
	Mar	3692	9.789	6097	165,13%

Fuente: Elaboración propia.

Figura 60. Modelo STLM arima 3,1,6 vs Datos Reales.



Fuente: Elaboración propia.

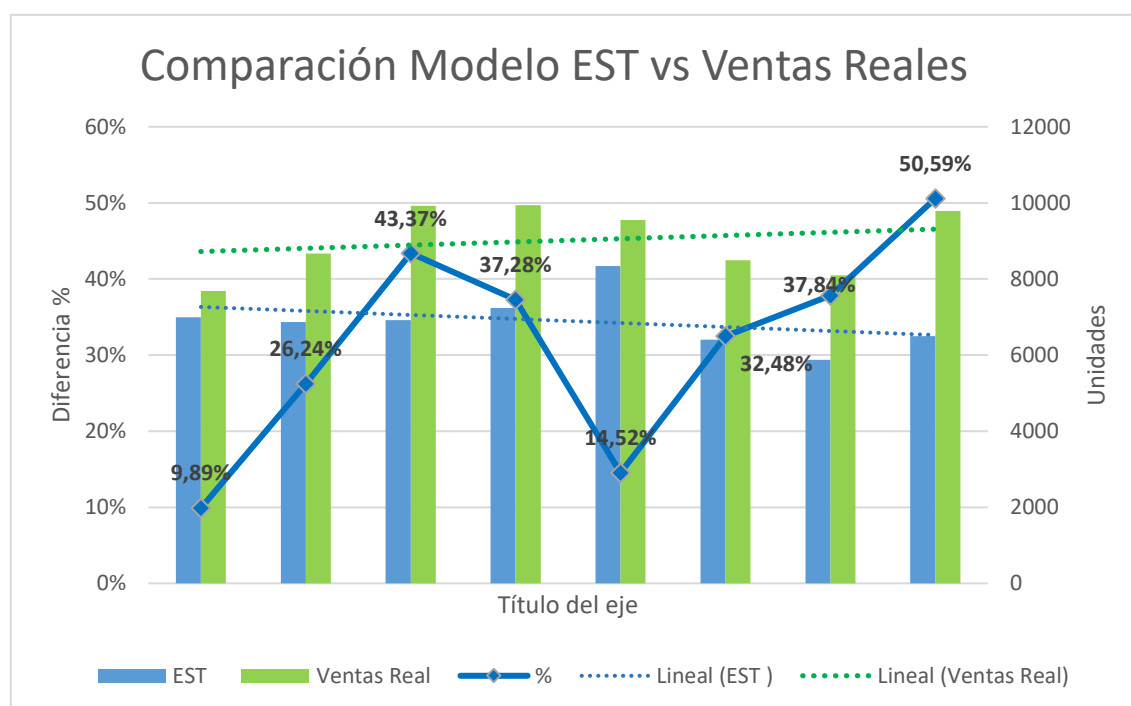
En la Tabla 27 se observa la comparación del modelo EST con los datos reales, también se muestra con una gráfica esta comparación, ver la Figura 61.

Tabla 27. Comparación Modelo EST vs Datos reales.

Año	Mes	EST	Ventas Real	Diferencia	%
2020	Ago	6994	7.686	692	9,89%
	Sep	6864	8.665	1801	26,24%
	Oct	6922	9.924	3002	43,37%
	Nov	7241	9.941	2700	37,28%
	Dic	8339	9.550	1211	14,52%
2021	Ene	6409	8.491	2082	32,48%
	Feb	5876	8.100	2224	37,84%
	Mar	6500	9.789	3289	50,59%

Fuente: Elaboración propia.

Figura 61. Modelo EST arima 3,1,6 vs Datos Reales.



Fuente: Elaboración propia.

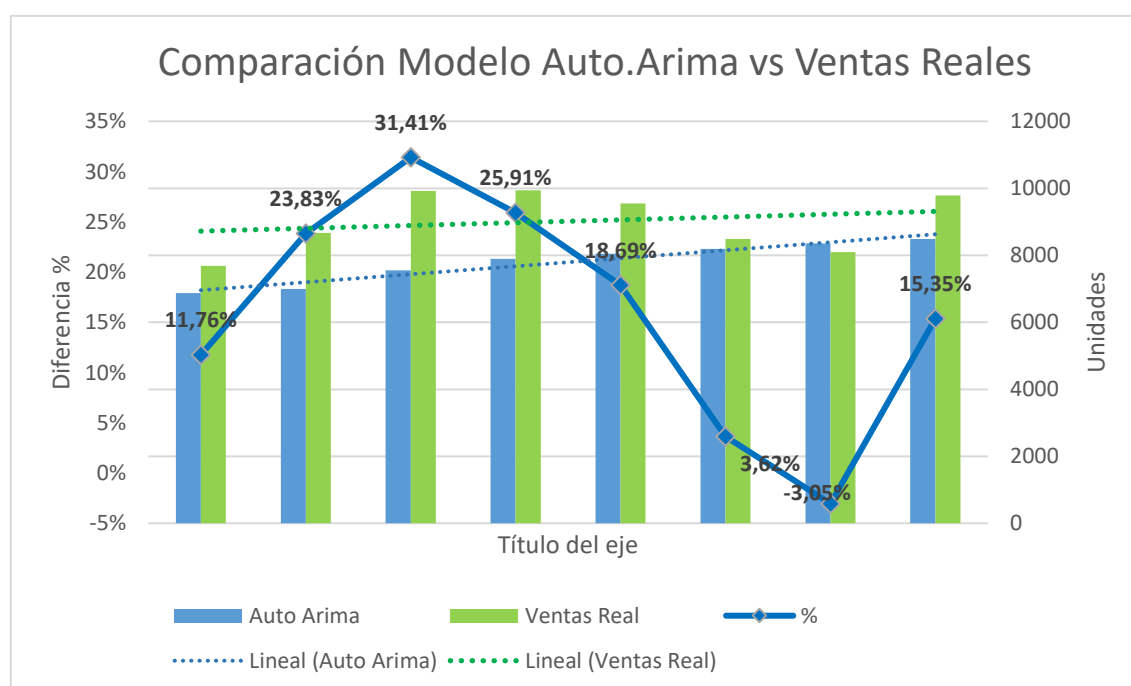
La Tabla 28 se detalla la comparación del modelo Auto Arima con los datos reales, además se observa una gráfica de estos datos, ver Figura 62.

Tabla 28. Comparación Modelo Auto.Arima vs Datos reales.

Año	Mes	Auto Arima	Ventas Real	Diferencia	%
2020	Ago	6877	7.686	809	11,76%
	Sep	6997	8.665	1668	23,83%
	Oct	7552	9.924	2372	31,41%
	Nov	7895	9.941	2046	25,91%
	Dic	8046	9.550	1504	18,69%
2021	Ene	8194	8.491	297	3,62%
	Feb	8354	8.100	-254	-3,05%
	Mar	8486	9.789	1303	15,35%

Fuente: Elaboración propia.

Figura 62. Modelo Auto Arima vs Datos Reales.



Fuente: Elaboración propia.

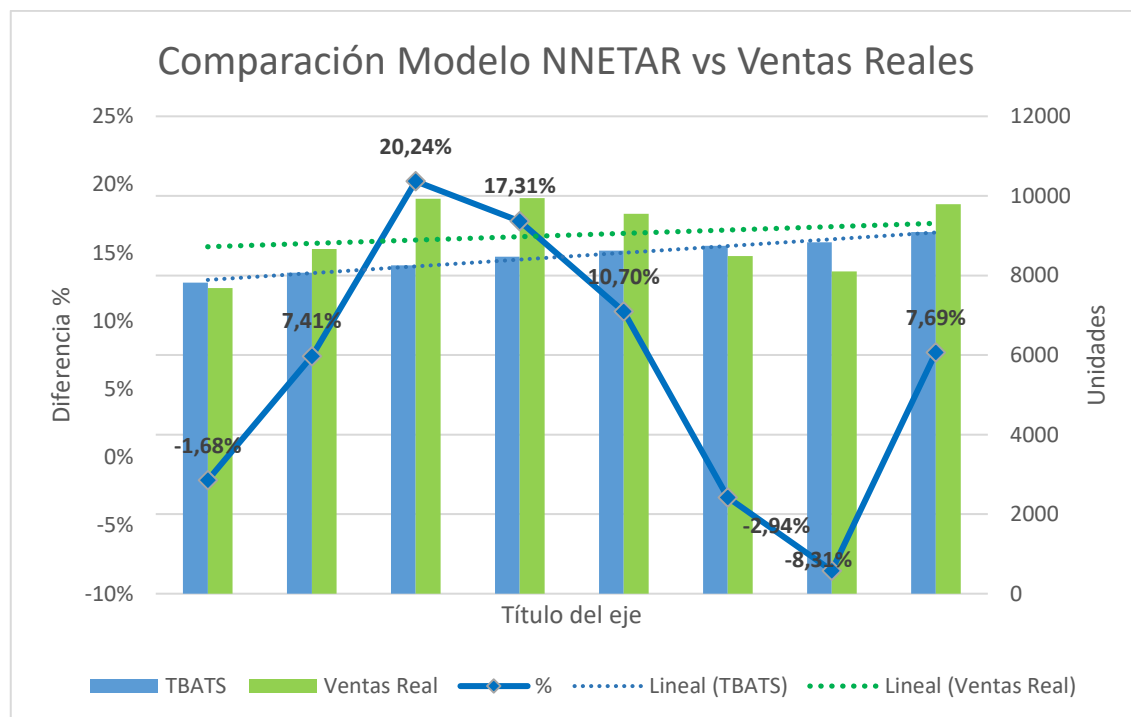
En la Tabla 29 se muestra la comparación del modelo NNETAR con los datos reales, adicional tenemos una gráfica de estos datos, ver Figura 63.

Tabla 29. Comparación Modelo NNETAR vs Datos reales

Año	Mes	NNETAR	Ventas Real	Diferencia	%
2020	Ago	7817	7.686	-131	-1,68%
	Sep	8067	8.665	598	7,41%
	Oct	8254	9.924	1670	20,24%
	Nov	8474	9.941	1467	17,31%
	Dic	8627	9.550	923	10,70%
2021	Ene	8748	8.491	-257	-2,94%
	Feb	8834	8.100	-734	-8,31%
	Mar	9090	9.789	699	7,69%

Fuente: Elaboración propia

Figura 63. Modelo NNETAR vs Datos Reales.



Elaboración propia.

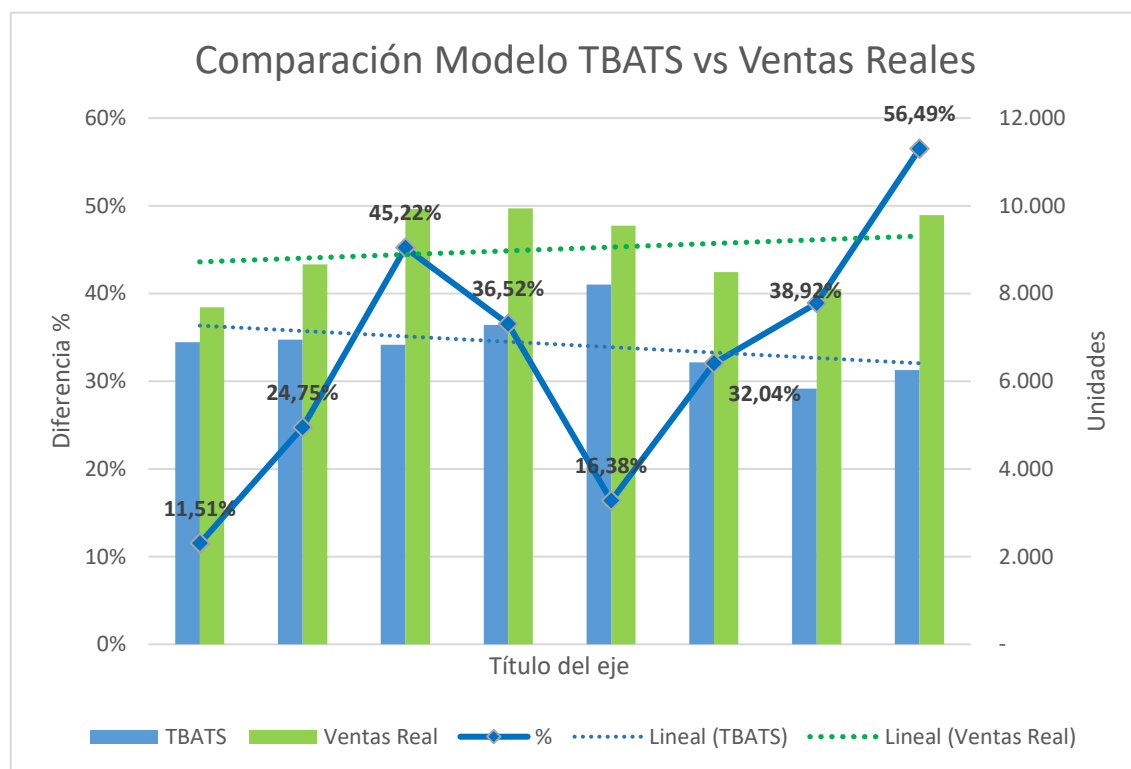
La Tabla 30 se detalla la comparación del modelo TBATS con los datos reales, también se muestra con una gráfica esta comparación, ver la Figura 64.

Tabla 30. Comparación Modelo TBATS vs Datos reales

Año	Mes	TBATS	Ventas Real	Diferencia	%
2020	Ago	6893	7.686	793	11,51%
	Sep	6946	8.665	1719	24,75%
	Oct	6834	9.924	3090	45,22%
	Nov	7282	9.941	2659	36,52%
	Dic	8206	9.550	1344	16,38%
2021	Ene	6431	8.491	2060	32,04%
	Feb	5831	8.100	2269	38,92%
	Mar	6255	9.789	3534	56,49%

Elaboración propia.

Figura 64. Modelo TBATS vs Datos Reales.



Fuente: Elaboración propia.

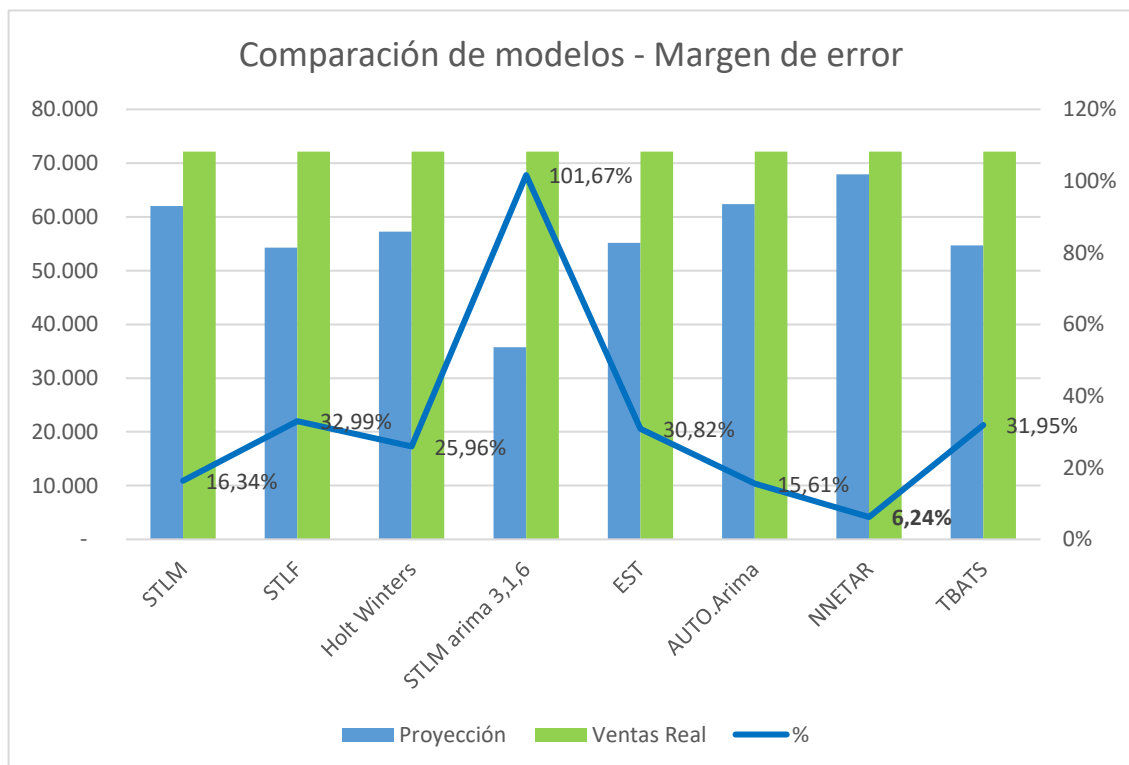
Por ultimo al realizar el análisis global de los resultados obtenidos de cada modelo con las ventas reales generadas en estos últimos 8 meses de agosto 2020 a marzo 2021, podemos observar en la Tabla 31, que el modelo que más se acercó en sus proyecciones fue el modelo de NNETAR con un 6.24% de diferencia ver Figura 65.

Tabla 31. Comparación general de modelos

Modelo	Proyección	Ventas Real	%
STLM	62.014	72.146	16,34%
STLF	54.251	72.146	32,99%
Holt Winters	57.279	72.146	25,96%
STLM arima 3,1,6	35.774	72.146	101,67%
EST	55.147	72.146	30,82%
AUTO.Arima	62.402	72.146	15,61%
NNETAR	67.910	72.146	6,24%
TBATS	54.677	72.146	31,95%

Elaboración propia.

Figura 65. Comparación de modelos - Margen de error.



Fuente: Elaboración propia.

4.4 TRABAJOS FUTUROS

Con la información obtenida se pueden realizar un análisis de datos a más detalle para conocer el comportamiento de como crecer el parque automotor, por provincia, por segmento (automóviles, suv, camiones, buses) o por marca, y poder determinar los vehículos que tiene mayor demanda en cada uno de los segmentos propuestos. De esta forma se puede aportar a las comercializadoras de vehículos nuevos a conocer el comportamiento del mercado en los próximos años y que vehículos van a tener mayor demanda en el futuro en nuestro país, según los análisis históricos de la información. También que es posible publicar los resultados obtenidos de los análisis realizados de la presente investigación y de los trabajos futuros.

CONCLUSIONES

- Luego de revisar el estado del arte de minería de datos predictiva, se destaca su importancia en el descubrimiento de conocimiento oculto y detección de tendencias futuras, que puede ser útil para apoyar la toma de decisiones en las empresas. Las técnicas y modelos de análisis predictivos de datos poseen características que las diferencian y que permiten determinar para qué tipo de análisis de información se las puede aplicar. Las series de tiempo, por ejemplo, son modelos de predicción que se basan en una variable cuantitativa (ejemplo, cantidad de productos fabricados) en función de la variable independiente que es el tiempo (año, mes, día). Estos modelos se han extendido a muchos campos del conocimiento, que requieren descubrir o predecir el comportamiento de una variable de interés en un momento determinado.
- La información analizada en este trabajo se centra en sector automotriz del Ecuador; se obtuvo de los boletines mensuales de la Asociación de Empresas Automotrices del Ecuador (AEADE). La variable dependiente tratada es la cantidad de vehículos vendidos por ciudad y marca en todo el Ecuador, y la variable independiente es el tiempo, el periodo analizado fue de enero 2007 a Julio 2020.
- Se realizó una exploración de los datos mediante estadística descriptiva para comprender sus características y para aplicar limpieza, transformación e imputación de datos con la finalidad de depurar la data y dejarla lista para posteriores análisis.
- Se utilizó la metodología ágil Team Data Science Process (TDSP) desarrollada por Microsoft, para proyectos de ciencia de datos, que incluyen también el análisis predictivo de datos. TDSP fue importante para el desarrollo de las etapas del ciclo de vida de este trabajo; las fases de esta metodología son: entendimiento del negocio, adquisición y entendimiento de los datos, modelado, despliegue y aceptación del cliente. Las técnicas de series temporales fueron utilizadas para el análisis de cantidad de vehículos vendidos en el Ecuador; los modelos seleccionados fueron: STLM (Descomposición estacional y de tendencias usando Loess con múltiples períodos estacionales), STLF (Pronóstico de carga a corto plazo), Holt Winters, STLM

ARIMA 3.1.6 (Modelo híbrido de STLM y ARIMA "Modelo autorregresivo integrado de media móvil"), EST (Error, Tendencia, Estacional), Auto Arima (Método ARIMA ajustado automáticamente en cada serie), NNETAR (Pronósticos de Series de Tiempo de Redes Neuronales) y TBATS (Estacionalidad trigonométrica, Transformación Box-Cox, Errores ARMA, Componentes de tendencia y estacionales). Con cada modelo se realizó el respectivo análisis predictivo de unidades de vehículos vendidos mediante el Lenguaje R.

- Al aplicar las técnicas de series temporales se realizó las predicciones a 36 meses, comprendidos de agosto 2020 a julio 2023; en las observaciones de los resultados que proporciono cada una de ellas, la técnica que genero el mínimo error en base a las métricas MAE (Error Absoluto Medio) y MAPE (Error Porcentual Absoluto Medio), es STLM Arima 3.1.6. Vale la pena señalar que, al realizar una comparación con los datos reales de las ventas obtenidas de agosto 2020 a marzo 2021, se encontró que el modelo que generó resultados más aproximados fue NNETAR con un 6.24% de diferencia.

RECOMENDACIONES

- Es importante contar con información y datos reales históricos y confiables para el desarrollo de cualquier proyecto de minería de datos, para generar conocimiento a través de la información; y para, predecir tendencias futuras para cualquier sector.
- La selección del segmento de la información, es importante para aplicar las técnicas de minería de datos; además, en la fase adquisición y entendimiento de los datos se debe considerar una herramienta o software que facilite la limpieza de los datos, esto con el objetivo de eliminar datos incompletos, erróneos o duplicados. También seleccionar las variables precisas en los datos que vayan acorde a los objetivos planteados desde el inicio de proyecto.
- Existen una gran variedad de técnicas y modelos de análisis de datos predictivos; se recomienda estudiarlas y analizarlas a detalle para determinar la técnica que sea adecuada dependiendo de la información que se posea, de esta forma se puede realizar la construcción de un buen modelado.
- Para todo proyecto de minería de datos, se debe estudiar las metodologías existentes, para poder aplicar la más adecuada. Además, se debe determinar la infraestructura tecnológica, recursos de información y herramientas para el análisis, desarrollo e implementación.
- Se recomienda utilizar los modelos de análisis predictivo, porque nos permiten determinar el comportamiento o tendencia a futuro de determinada variable; lo que nos puede ayudar a tener información oportuna y confiable a considerar como una herramienta poderosa y esencial para la toma de decisiones en una organización o empresa.

BIBLIOGRAFÍA

- [1] L. Atriwal, P. Nagar, S. Tayal, y V. Gupta, «Business intelligence tools for big data», *Journal of Basic and Applied Engineering Research*, vol. 3, n.o 6, pp. 505-509, 2016, doi:10.1081/08873417.2016.1220239
- [2] D. Loshin, «Market and Business Drivers for Big Data Analytics», en *Big Data Analytics*, D. Loshin, Ed. Boston: Morgan Kaufmann, 2016, pp. 1-9. doi: 10.1016/B978-0-12-417319-4.00001-6.
- [3] K. G. Srinivasa, G. M. Siddesh, y H. Srinidhi, «Introduction to Data Analytics», en *Network Data Analytics*, Springer, 2018, pp. 3-28, doi:10.1007/978-3-319-77800-6_1
- [4] K. Kambatla, G. Kollias, V. Kumar, y A. Grama, «Trends in big data analytics», *Journal of Parallel and Distributed Computing*, vol. 74, n.o 7, pp. 2561-2573, jul. 2016, doi: 10.1016/j.jpdc.2014.01.003.
- [5] M. Sanjay y B. H. Alamma, «An insight into big data analytics—Methods and application», en *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 1, pp. 1-5, 10.1109/INVENTIVE.2016.7823269
- [6] Sowmya R y Suneetha K R, «Data Mining with Big Data», en *2017 11th International Conference on Intelligent Systems and Control (ISCO)*, ene. 2017, pp. 246-250. doi: 10.1109/ISCO.2017.7855990.
- [7] F. Zuha y G. Achuthan, «Analysis of Data Mining Techniques and its Applications», *IJCA*, vol. 140, n.o 3, pp. 6-14, abr. 2016, doi: 10.5120/ijca2016909249.
- [8] B. Mazon-Olivo, W. Rivas-Asanza, J. Novillo-Vicuña, y C. Flores-Cabrera, «Análisis de producción avícola mediante técnicas de inteligencia de negocios y minería de datos», *Alternativas*, vol. 19, n.o 2, pp. 80-88, ago. 2018, doi: 10.23878/alternativas.v19i2.203.
- [9] B. Mazon-Olivo, O. Romero-Hidalgo, A. Borja-Herrera, M. Aguirre-Benalcazar, M. Contenido-Segarra, y M. Jaramillo, «Tecnologías de Inteligencia de Negocios y Minería de datos para el análisis de la producción y comercialización de cacao. Business Intelligence and Data Mining Technologies for the analysis of cocoa production and commercialization», *Espacios*, vol. 39, p. 6, ago. 2018.
- [10] P. Alvarado, «Sector automotriz presentó su plan de reactivación frente a una caída en ventas del 77%», *El Comercio*, jun. 05, 2020. <http://www.elcomercio.com/actualidad/sector-automotriz-plan-reactivacion-ventas.html> (accedido jun. 27, 2020).
- [11] «Boletín Sector Automotor en Cifras – AEADE». <https://www.aeade.net/boletin-sector-automotor-en-cifras/> (accedido abr. 25, 2021).
- [12] «Instituto Nacional de Estadística y Censos». <https://www.ecuadorencifras.gob.ec/institucional/home/> (accedido dic. 12, 2020).

- [13] «Boletín Sector Automotor en Cifras – AEADE No. 48». <https://www.aeade.net/wp-content/uploads/2020/10/Boletin-Sector-en-cifras-48-resumen-en-espanol.pdf> (accedido dic. 20, 2020).
- [14] marktab, «What is the Team Data Science Process?». <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview> (accedido mar. 15, 2021).
- [15] Universidad Pedagógica Nacional y R. Guevara Patiño, «El estado del arte en la investigación: ¿análisis de los conocimientos acumulados o indagación por nuevos sentidos?», *Folios*, vol. 1, n.o 44, pp. 165-179, may 2016, doi: 10.17227/01234870.44folios165.179.
- [16] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, y S. Linkman, «Systematic literature reviews in software engineering – A systematic literature review», *Information and Software Technology*, vol. 51, n.o 1, pp. 7-15, ene. 2009, doi: 10.1016/j.infsof.2008.09.009.
- [17] K. Jayamalini y M. Ponnavaikko, «Research on web data mining concepts, techniques and applications», en 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), feb. 2017, pp. 1-5. doi: 10.1109/ICAMMAET.2017.8186676.
- [18] F. Sternberg, K. H. Pedersen, N. K. Ryelund, R. R. Mukkamala, y R. Vatrappu, «Analysing customer engagement of Turkish airlines using big social data», en 2018 IEEE International Congress on Big Data (BigData Congress), 2018, pp. 74-81. doi: 10.1109 / BigDataCongress.2018.00017.
- [19] S. Desai y S. T. Patil, «Efficient regression algorithms for classification of social media data», en 2016 International Conference on Pervasive Computing (ICPC), 2016, pp. 1-5. doi: 10.1109/PERVASIVE.2015.7087040.
- [20] A. Vilorio, J. Li, J. G. Guiliany, y B. de la Hoz, «Predictive Model for Detecting Customer's Purchasing Behavior Using Data Mining», en Proceedings of 6th International Conference on Big Data and Cloud Computing Challenges, 2020, pp. 45-54, doi:10.1007/978-981-32-9889-7_4
- [21] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «Introduction to Predictive Analytics», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 1-25. doi: 10.1007/978-3-030-14038-0_1.
- [22] B. Boukenze, H. Mousannif, y A. Haqiq, «Predictive analytics in healthcare system using data mining techniques», *Computer Science & Information Technology (CS & IT)*, vol. 1, pp. 1-9, 2016, doi: 10.5121/csit.2016.60501.
- [23] R. McCarthy, W. Ceccucci, M. McCarthy, y L. Halawi, «Alpha Insurance: A Predictive Analytics Case to Analyze Automobile Insurance Fraud using SAS Enterprise Miner», *Information Systems Education Journal*, vol. 17, n.o 2, p. 20, 2019, <http://isedj.org/2019-17/>

- [24] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «Know Your Data—Data Preparation», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 27-56. doi: 10.1007/978-3-030-14038-0_2.
- [25] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «Model Comparisons and Scoring», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 175-199. doi: 10.1007/978-3-030-14038-0_7.
- [26] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «Predictive Models Using Decision Trees», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 123-144. doi: 10.1007/978-3-030-14038-0_5.
- [27] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «Predictive Models Using Neural Networks», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 145-173. doi: 10.1007/978-3-030-14038-0_6.
- [28] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «Predictive Models Using Regression», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 89-121. doi: 10.1007/978-3-030-14038-0_4.
- [29] R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, «What Do Descriptive Statistics Tell Us», en *Applying Predictive Analytics: Finding Value in Data*, R. V. McCarthy, M. M. McCarthy, W. Ceccucci, y L. Halawi, Eds. Cham: Springer International Publishing, 2019, pp. 57-87. doi: 10.1007/978-3-030-14038-0_3.
- [30] R. Jain y S. Venkatesan, «Business Analytics And Data Mining Techniques Using Predictive Algorithms To Enhance Business Intelligence», *International Journal of Transformations in Business Management*, vol. 7, pp. 21-33, 2017.
- [31] A. Novák y L. Buřita, «Business Intelligence and ISR Data Processing», en *2019 International Conference on Military Technologies (ICMT)*, may 2019, pp. 1-8. doi: 10.1109/MILTECHS.2019.8870135.
- [32] I. Kholod, M. Efimova, A. Rukavitsyn, y S. Andrey, «Time Series Distributed Analysis in IoT with ETL and Data Mining Technologies», en *Internet of Things, Smart Spaces, and Next Generation Networks and Systems*, Cham, 2017, pp. 97-108. doi: 10.1007/978-3-319-67380-6_9.
- [33] O. Moscoso-Zea, Andres-Sampedro, y S. Luján-Mora, «Datawarehouse design for educational data mining», en *2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)*, sep. 2016, pp. 1-6. doi: 10.1109/ITHET.2016.7760754.
- [34] B. Mazon Olivo, A. Pan, y R. Tinoco Egas, «Inteligencia de negocios en el sector agropecuario», Machala: Universidad Técnica de Machala, 2018. Accedido: jun. 20, 2021. [En línea]. Disponible en: <http://repositorio.utmachala.edu.ec/handle/48000/13330>

- [35] L. Lin, H. Xiu, W. Yunfei, y W. Guangchao, «Research on Overall Equipment Effectiveness for Cigarette Equipments Based on Data Mart and Data Mining», *DEStech Transactions on Engineering and Technology Research*, vol. 0, n.o apetc, 2017, doi: 10.12783/dtetr/apetc2017/11291.
- [36] Gartner, «Reimpresión de Gartner plataformas de análisis y Business Intelligence.» https://www.gartner.com/doc/reprints?id=1-3TXXSLV&ct=170221&st=sb&ocid=mkto_eml_EM597235A1LA1 (accedido nov. 07, 2020).
- [37] Gartner, «Reimpresión de Gartner para plataformas de ciencia de datos y aprendizaje automático.» <https://www.gartner.com/doc/reprints?id=1-1YDUKTC6&ct=200217&st=sb> (accedido nov. 07, 2020).
- [38] I. Ramírez Morales, B. Mazon Olivo, y A. Pan, «Ciencia de datos en el sector agropecuario», en *Análisis de Datos Agropecuarios*, Eds. Machala-Ecuador: Universidad Técnica de Machala, 2018, pp 12-44. Accedido: dic. 20, 2020. [En línea]. Disponible en: <http://repositorio.utmachala.edu.ec/handle/48000/13324>
- [39] «The Alteryx Analytic Process Automation (APA) Platform», Alteryx. <https://www.alteryx.com/products/apa-platform> (accedido nov. 07, 2020).
- [40] Asniar y K. Surendro, «Predictive Analytics for Predicting Customer Behavior», en *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIIT)*, mar. 2019, pp. 230-233. doi: 10.1109/ICAIIIT.2019.8834571.
- [41] H. F. V. Ballesteros, E. G. Iñiguez, y S. R. M. Velasco, «Minería de Datos», *RECIMUNDO: Revista Científica de la Investigación y el Conocimiento*, vol. 2, n.o Extra 1, pp. 339-349, 2018. [En Línea] <https://dialnet.unirioja.es/servlet/articulo?codigo=6732870>
- [42] B. Mazon-Olivo, M.Pita, y F. Redrovan «Desarrollo de competencias en Minería de Datos, una experiencia didáctica», en *Sistematización de experiencias educativas innovadoras - 1ra Edición*. Machala: Editorial UTMACH, 2020. pp383-406 Accedido: ene. 21, 2021. [En línea]. Disponible en: <http://repositorio.utmachala.edu.ec/handle/48000/15219>
- [43] P. L. CESAR y S. G. DANIEL, *Minería de datos. Técnicas y herramientas: técnicas y herramientas*. Editorial Paraninfo, 2017.
- [44] M. I. U. Fassler, A. S. C. Barahona, P. M. M. Naranjo, y H. M. V. Yáñez, «Minería de datos para la toma de decisiones en la unidad de nivelación y admisión universitaria ecuatoriana», *Cumbres*, vol. 4, n.o 2 (Julio-diciembre), pp. 55-67, 2018, [En Línea] <https://dialnet.unirioja.es/servlet/articulo?codigo=6836545>.
- [45] I. Alam, D. Md. Farid, y R. J. F. Rossetti, «The Prediction of Traffic Flow with Regression Analysis», en *Emerging Technologies in Data Mining and Information Security*, Singapore, 2019, pp. 661-671. doi: 10.1007/978-981-13-1498-8_58.
- [46] T. M. O. Diallo, A. J. S. Morin, y H. Lu, «Impact of Misspecifications of the Latent Variance–Covariance and Residual Matrices on the Class Enumeration Accuracy

of Growth Mixture Models», *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 23, n.o 4, pp. 507-531, jul. 2016, doi: 10.1080/10705511.2016.1169188.

[47] A. Konar y D. Bhattacharya, «An Introduction to Time-Series Prediction», en *Time-Series Prediction and Applications: A Machine Intelligence Approach*, A. Konar y D. Bhattacharya, Eds. Cham: Springer International Publishing, 2017, pp. 1-37. doi: 10.1007/978-3-319-54597-4_1.

[48] H. A. Mengash, «Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems», *IEEE Access*, vol. 8, pp. 55462-55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[49] M. K. Rao, K. Veera Swamy, y K. A. Sheela, «Advanced machine learning discriminant analysis models for face retrieval system», en 2018 2nd International Conference on 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), ago. 2018, pp. 609-613. doi: 10.1109/I-SMAC.2018.8653671.

[50] D. Díaz-Vico y J. R. Dorronsoro, «Deep Least Squares Fisher Discriminant Analysis», *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, n.o 8, pp. 2752-2763, ago. 2020, doi: 10.1109/TNNLS.2019.2906302.

[51] M. Macas, L. Lagla, W. Fuertes, G. Guerrero, y T. Toulkeridis, «Data Mining model in the discovery of trends and patterns of intruder attacks on the data network as a public-sector innovation», en 2017 Fourth International Conference on eDemocracy eGovernment (ICEDEG), abr. 2017, pp. 55-62. doi: 10.1109/ICEDEG.2017.7962513.

[52] S. Huber, H. Wiemer, D. Schneider, y S. Ihlenfeldt, «DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model», *Procedia CIRP*, vol. 79, pp. 403-408, ene. 2019, doi: 10.1016/j.procir.2019.02.106.

[53] L. F. Castro R., E. Espitia P., y A. F. Montilla, «Applying CRISP-DM in a KDD Process for the Analysis of Student Attrition», en *Advances in Computing*, Cham, 2018, pp. 386-401. doi: 10.1007/978-3-319-98998-3_30.

[54] S. B. Gómez, M. C. Gómez, y J. B. Quintero, «Business Intelligence Applied to Ecotourism in Colombia», en 2019 14th Iberian Conference on Information Systems and Technologies (CISTI), jun. 2019, pp. 1-6. doi: 10.23919/CISTI.2019.8760802.

[55] F. Peralta, «Proceso de Conceptualización del Entendimiento del Negocio para Proyectos de Explotación de Información», 2017, p. 34. doi: 10.18294/RELAIS.2014.273-306.

[56] F. Schäfer, C. Zeiselmaier, J. Becker, y H. Otten, «Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes», en 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD), nov. 2018, pp. 190-195. doi: 10.1109/ITMC.2018.8691266.

[57] S. A. Mohd Selamat, S. Prakoonwit, y W. Khan, «A review of data mining in knowledge management: applications/findings for transportation of small and medium

enterprises», SN Appl. Sci., vol. 2, n.o 5, p. 818, abr. 2020, doi: 10.1007/s42452-020-2589-3.

[58] M. Abdellatief, E. M. Shaaban, y K. A. Abu-Raya, «Egyptian Case Study-Sales forecasting model for automotive section», en 2019 International Conference on Smart Applications, Communications and Networking (SmartNets), dic. 2019, pp. 1-6. doi: 10.1109/SmartNets48225.2019.9069751.

[59] R. H. Sampieri, C. F. Collado, y P. B. Lucio, Metodología De Investigación 6ta Edición, 6ta ed. McGRAW-WILL, 2019. Accedido: nov. 14, 2020. [En línea]. Disponible en: <https://markainvestigacion.wordpress.com/2019/01/14/libro-de-sampieri-sobre-metodologia-de-investigacion-6ta-edicion/>

[60] «Software de hojas de cálculo Microsoft Excel | Microsoft 365». <https://www.microsoft.com/es-ww/microsoft-365/excel> (accedido jun. 19, 2021).

[61] «R: El Proyecto R para Computación Estadística». <https://www.r-project.org/> (accedido jun. 19, 2021).

[62] F. Martínez-Plumed et al., «CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories», IEEE Transactions on Knowledge and Data Engineering, pp. 1-1, 2019, doi: 10.1109/TKDE.2019.2962680.

[63] Y. Liu, H. Dong, X. Wang, y S. Han, «Time Series Prediction Based on Temporal Convolutional Network», en 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), jun. 2019, pp. 300-305. doi: 10.1109/ICIS46139.2019.8940265.

[64] Z. Pala y R. Atici, «Forecasting Sunspot Time Series Using Deep Learning Methods», Sol. Phys., vol. 294, n.o 5, p. 50, may 2019, doi: 10.1007/s11207-019-1434-6.

[65] J. M. Lee y M. Hauskrecht, «Recent Context-Aware LSTM for Clinical Event Time-Series Prediction», en Artificial Intelligence in Medicine, Cham, 2019, pp. 13-23. doi: 10.1007/978-3-030-21642-9_3.

[66] T. M. Dantas, F. L. Cyrino Oliveira, y H. M. Varela Repolho, «Air transportation demand forecast through Bagging Holt Winters methods», J. Air Transp. Manag., vol. 59, pp. 116-123, mar. 2017, doi: 10.1016/j.jairtraman.2016.12.006.

[67] L. Ferbar Tratar y E. Strmčnik, «The comparison of Holt–Winters method and Multiple regression method: A case study», Energy, vol. 109, pp. 266-276, ago. 2016, doi: 10.1016/j.energy.2016.04.115.

[68] I. A. Abuamra, A. Y. A. Maghari, y H. F. Abushawish, «Medium-term forecasts for groundwater production and rainfall amounts (Deir El-Balah City as a case study)», Sustain. Water Resour. Manag., vol. 6, n.o 5, p. 82, sep. 2020, doi: 10.1007/s40899-020-00446-z.