UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Sistema geolocalizado de pronóstico de casos de personas desaparecidas

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE:

Ingeniero de Sistemas

AUTORES:

Jaime Cordova, Jhordy Yosshi (ORCID: 0000-0003-1103-7012)

Paz Sanchez, Cristian Alexander (ORCID: 0000-0003-1824-6183)

ASESOR:

Dr. Alfaro Paredes, Emigdio Antonio (ORCID: 0000-0002-0309-9195)

LÍNEA DE INVESTIGACIÓN:

Sistema de información y comunicaciones

LÍNEA DE RESPONSABILIDAD SOCIAL UNIVERSITARIA:

Desarrollo sostenible y adaptación al cambio climático

LIMA – PERÚ

2020

Dedicatoria

A nuestras familias, por habernos apoyado en todo momento y por los consejos diarios que siempre nos han brindado.

Les agradecemos por los valores, la motivación constante y por la perseverancia que les caracteriza y que siempre nos han inculcado para lograr concluir el presente trabajo de investigación.

Agradecimiento

En esta oportunidad aprovecharemos para agradecer a Dios por permitirnos llegar a este punto. También a los profesores de la Escuela Profesional de Ingeniería de Sistemas por sus grandes enseñanzas brindadas a lo largo de la carrera. Además, queremos expresar nuestro mayor agradecimiento al Dr. Emigdio Antonio Alfaro Paredes por el asesoramiento continuo para lograr esta investigación.

Índice de contenidos

Índice de tablas

Índice de figuras

Índice del Abreviaturas

N°	Sigla	Significado	Pág.
1	RNPED	Registro Nacional de Datos de Personas Extraviadas o desaparecidas (Gobierno de México, 2020, π. 1)	15
2	UNODC	United Nations Office on Drugs and Crime u Oficina de las Naciones Unidad Contra la Droga y el Delito en Colombia (UNODC, 2010, p.19)	4
3	Red Social VK	Vkontakte versión rusa de Facebook (Severiukhina et al., 2020)	9
4	ME	Mean Error o Error Medio (Vandeput, 2019, π. 2)	79
5	RMSE	Mean Square Error o Error Cuadrático Medio (Vandeput, 2019, π. 2)	79
6	ETL	Estraction, Transformation, Loading o Extracción, Transformación, Carga (Power Data,2018, π. 1)	16
7	MAE	Mean Absolute Error o Error absoluto medio (Vandeput, 2019, π. 2)	79
8	MPE	Mean Percentage Error o Error porcentual medio (Vandeput, 2019, π. 2)	79
9	MAPE	Mean Abolute Percent Error o Error porcentual absoluto medio (Vandeput, 2019, π. 2)	79
10	SEMMA	Sample, Explore, Modify, Modify, Model, Evaluate o Muestra, Explorar, Modificar, Modelo, Evaluar (Hernández et al., 2009, p. 83)	14
12	Catalyst o P3TQ	Product, Place, Price, Time, Quantity o Producto, Lugar, Precio, Tiempo, Cantidad (Aquino et al., 2015, p. 280)	15
13	KDD	Knowledge Discovery in Databases O Descubrimiento de Conocimiento en Base de datos (Timarán et al., 2016, p. 64)	13
14	MASE	Mean Aboslute Scaling Error o Error escalado absoluto medio (Vandeput, 2019, π. 2)	79
15	ACF1	First-Order Autocorrelation Coefficient o Coeficiente de auto correlación de primer orden (Vandeput, 2019, π. 2)	79
16	BBC	British Broadcasting Coporation o Corporación Británica de Radiofusión (Vandeput, 2019, π. 2)	4
17	KPI	Key Performance Indicator or Performance Meter o Indicador Clave de Desempeño o Medidor de Desempeño (Vandeput, 2019, π. 2)	19

Resumen

El problema de la investigación fue ¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el uso de la regresión logística binaria y modelos de series de tiempo? Asimismo, el objetivo de la investigación fue determinar la precisión del pronóstico de casos de personas desaparecidas con la regresión logística binaria y los modelos de series de tiempo como ARIMA, Holt-Winters y ETS.

La investigación realizada fue de tipo aplicada, de enfoque cuantitativo y de diseño transversal correlacional causal, además se usó un diseño longitudinal de tendencia debido a que se analizaron los datos de casos de personas desaparecidas de México en el periodo de 2007-2018. Los datos son de acceso público y fueron obtenido mediante el RNPED. Asimismo, para el desarrollo de la investigación se usó el proceso de minería de datos implementando la metodología CRISP-DM. Además, de un análisis de los diferentes tipos de sistemas de geolocalización, identificando las técnicas para la innovación y dando a conocer la información para tener una mejor precisión en el pronóstico.

Los resultados fueron satisfactorios por lo que finalmente se concluyó que el sistema de geolocalización de pronóstico de personas desaparecidas se puede usar en cualquier entorno siempre y cuando existan base de datos reales para lograr un pronóstico fiable. En este sentido, con los resultados obtenidos dentro de la investigación en los modelos ARIMA y ETS los cuales llegaron a ser los más resaltantes con diecisiete y dieciséis modelos respectivamente de precisión de pronósticos más acertados, tomando como ejemplo al estado de Chiapas con una precisión del 99.31% a comparación de las otras series de tiempo con un índice de precisión del 99.28% en el modelo Holt-Winters y 99.17% en el modelo ARIMA. Finalmente, se recomendó ampliar este sistema en otras plataformas interactivas como el Tableau o Dundas BI para que junto al Power Bi se pueda llegar a una cantidad mayor de usuarios.

Palabras Clave: Sistema de pronóstico, Personas desaparecidas, Metodología CRISP-DM, Modelo ARIMA y ETS.

Abstract

The research problem was how much was the improvement in the accuracy of missing persons case forecasting with the use of binary logistic regression and time series models? Also, the research objective was to determine the accuracy of missing persons case forecasting with binary logistic regression and time series models such as ARIMA, Holt-Winters and ETS.

The research conducted was of an applied type, quantitative approach and causal correlational cross-sectional design, in addition, a longitudinal design of trend was used because the data of cases of missing persons in Mexico in the period 2007-2018 were analyzed. The data is publicly available and was obtained through the RNPED. Likewise, for the development of the research, the data mining process was used by implementing the CRISP-DM methodology. In addition, an analysis of the different types of geolocation systems, identifying techniques for innovation and providing information for better forecasting accuracy.

The results were satisfactory, so it was finally concluded that the geolocation system for missing persons forecasting can be used in any environment as long as there are real data bases to achieve a reliable forecast. In this sense, with the results obtained within the research in the ARIMA and ETS models, which became the most outstanding with seventeen and sixteen models respectively of more accurate forecast accuracy, taking as an example the state of Chiapas with an accuracy of 99.31% compared to the other time series with an accuracy rate of 99.28% in the Holt-Winters model and 99.17% in the ARIMA model. Finally, it was recommended to extend this system to other interactive platforms such as Tableau or Dundas BI so that, together with Power Bi, a larger number of users can be reached.

Keywords: Prognostic System, Missing Persons, CRISP-DM Methodology, Model ARIMA and ETS.

I. INTRODUCCIÓN

En la actualidad, los casos de personas desaparecidas son muy alarmantes debido a que los resultados de la búsqueda no han sido los esperados, ya sea por falta de medios, de organización o de conocimiento de los recursos (Castellano y López, 2017, p. 565). Por estas razones, el presente estudio tiene como prioridad presentar un sistema geolocalizado, en este sentido lograr contribuir al pronóstico de los posibles puntos de riesgo y los casos de personas desaparecidas. Asimismo, este sistema tiene una orientación novedosa, ya que, busca el desarrollo de modelos de propuestas para la ejecución de proyectos de aprendizaje basados en la geolocalización con una voluntad de servicio a la comunidad (Gros y Forés, 2013, p. 42).

En los siguientes párrafos se detallan algunos antecedentes que aportaron las ideas para el estudio y determinar si son útiles, destacando los resultados obtenidos por otros investigadores (Arias, 2020, p. 17). Por ello, se realizó un análisis de los datos recogidos, para detectar posibles valores que han beneficiado al estudio (Fabara et al., 2019, p. 74). Un sistema de geolocalización sigue teniendo un papel innovador, ya que es una de las herramientas más utilizadas para las diferentes funciones que ofrece el mismo (Beltrán, 2015, p. 97). Asimismo, este sistema informará a las personas pertinentes sobre los puntos de riesgo y el pronóstico que se dará en los próximos años de los casos de personas desaparecidas, a través de un portal web accesible desde una computadora (Díaz, 2011, p. 25).

Arias (2020) citó a Hernández y Mendoza (2018), quienes señalaron: “Una vez desarrollada la idea del tema de estudio, es importante revisar otras investigaciones, estudios y/o trabajos de investigación anteriores para saber qué es lo que ya se ha estudiado con respecto al tema en desarrollo”. Añadiendo, Hernández et al. (2018) concluyeron: “Que los antecedentes sirven para traer nuevas ideas a nuestro estudio y si son útiles para compartir y aprender sobre los descubrimientos hechos por otros investigadores” (p. 17).

Este sistema se enfrenta a una serie de problemas internos que afectan a la fiabilidad y la precisión de los datos de geolocalización recabados, especialmente en lo que respecta a los errores de nivel unitario y otros errores (Arana et al., 2018, p. 2). Es necesario realizar una serie de análisis de los datos

reunidos, tanto univariantes como multivariantes, para detectar posibles valores atípicos e identificar las relaciones entre las variables (Fabara, et al*.,* 2019, p. 74). Asimismo, Cabrera y De León (2018) indicaron: “Una vez conocida esta información, se procede a indicar si se pueden estimar los errores menores de pronóstico” (p. 26).

Por lo que no es necesario retroceder mucho en el tiempo para darse cuenta del inmenso desarrollo que ha experimentado el campo de la geolocalización. Al respecto, Cabello (2017) explicó: “Para conocer la ubicación de un pueblo, presuntos delincuentes; las unidades de investigación sólo tenían el seguimiento policial tradicional, lo que significaba un gran despliegue de personas y medios económicos” (p. 285). Asimismo, Cabello (2017) concluyó: “Que hoy en día, la tecnología pone en sus manos un conjunto de herramientas para la investigación de lo ilícito y el alcance de la geolocalización no se ha quedado relegado” (p. 285).

Beltrán (2015) indicó: “Los sistemas de geolocalización siguen desempeñando un papel innovador, es decir, que es una de las herramientas más utilizadas por los geógrafos para colocar personas u objetos en el espacio utilizando sus coordenadas” (p. 97). Al mismo tiempo, el fenómeno de compartir información sobre cada lugar e individuo se ha ido incrementando paulatinamente en los medios de comunicación (Beltrán, 2015, p. 97). Este sistema informará a las personas pertinentes sobre la geolocalización de los puntos de riesgo y las posibles características de una persona a través de una interfaz web accesible desde los ordenadores portátiles o de escritorio, por lo que se tomarán estudios que corroboran la exactitud de técnicas en la construcción de un sistema de geolocalización para pronosticar los casos de personas desaparecidas (Díaz, 2011, p. 25).

La incertidumbre por la desaparición de algún familiar llevará a estos a empezar una investigación para su posible ubicación, y dar con su paradero desconocido (Schulz y Salazar, 2018, p. 103). Asimismo, los especialistas de la BBC (2020) precisaron: “Durante la cuarentena en el mes de julio se registraron 915 denuncias de personas desaparecidas de las cuales el 70% eran niñas y adolescentes” (π. 14). Asimismo, los especialistas de la BBC (2020) acotaron:

“Durante la cuarentena no existían mecanismos para realizar las denuncias, por lo que las personas no podían salir de sus casas hacia las comisarías, por lo que probablemente la cifra sea aún más alta que la oficial” (π. 24).

Los especialistas de la Fundación Justicia (2020) indicaron “Durante la emergencia sanitaria generada por el virus SARS-CoV2 (COVID 19) la búsqueda de personas desaparecidas no debe de parar, por lo que cada día es una oportunidad de encontrar a ese ser querido” (π. 1). El propósito de la geolocalización es analizar los diferentes lugares donde es posible ubicar el proyecto, buscándose en establecer un lugar que ofrezca los máximos beneficios, los mejores costos, es decir, donde se obtenga el máximo beneficio, si se trata de un proyecto social (Corrillo y Guitiérrez, 2016, p. 29).

El sistema de geolocalización de pronóstico es una nueva tecnología que se basa en las características y posibles puntos de riesgo por lo que, cuando se trabaja con nuevas tecnologías; la reducción del tiempo y el riesgo son puntos importantes dentro del desarrollo. Las características del sistema, permitió un amplio abanico de posibilidades y reducción en el tiempo de desarrollo, por lo que, si se considera continuar con las actualizaciones en un futuro, se considera impulsar el mayor número de veces posible, probando nuevos requerimientos y adaptándose a las necesidades (Araneda y Gatica, 2013, p. 25).

Los especialistas de la UNODC que se describe como la oficina de Naciones Unidas Contra la Droga y el Delito (2010) indicaron: “La desaparición forzada de personas consiste en la privación de libertad de una o más personas por cualquier medio (aprehensión, detención o secuestro)” (p. 19). Además, Guzmán (2017) evaluó “Un sistema de geolocalización es un sistema que, mediante las coordenadas, se pueda ubicar geográficamente un punto en el mapa; proporcionando una información exacta de la ubicación (p. 20).

Por tanto, la implementación del sistema geolocalizado de pronóstico permitirá reducir el porcentaje de casos de personas desaparecidas al dar a conocer los puntos de riesgo de cada Estado o sus posibles características que hagan que sean una persona vulnerable. Al respecto, Arias (2020) concluyó: “Para comenzar con una investigación es fundamental conocer lo que se va a

estudiar, un acercamiento a la realidad y a la situación que se intenta analizar, mejorar y/o resolver; todo esto nace de una idea y/o una situación problemática, que puede provenir de diferentes fuentes o medios" (p 12).

Por lo tanto, el problema general de la investigación fue: ¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el uso de la regresión logística binaria y modelos de series de tiempo? El problema específico de la investigación fue:

PE1: ¿Cuáles son las relaciones entre el rango de edad, la presencia de acné y verrugas, los bigotes, la complexión, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares con los casos de desaparición de varones?

PE2: ¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el modelo ARIMA?

PE3: ¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el modelo Holt-Winters?

PE4: ¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el modelo ETS?

El objetivo general de la investigación fue determinar la precisión del pronóstico de casos de personas desaparecidas con la regresión logística binaria y los modelos de series de tiempo. Los objetivos específicos fueron los siguientes:

OE1: Determinar las relaciones entre el rango de edad, la presencia de acné y verrugas, los bigotes, la complexión, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares con los casos de desaparición de varones.

OE2: Determinar la precisión del pronóstico de casos de personas desaparecidas con el modelo ARIMA.

OE3: Determinar la precisión del pronóstico de casos de personas desaparecidas con el modelo Holt-Winters.

OE4: Determinar la precisión del pronóstico de casos de personas desaparecidas con el modelo ETS.

La hipótesis general de la investigación fue: “El uso de la regresión logística binaria y modelos de series de tiempo mejoró la precisión del pronóstico de casos de personas desaparecidas.”. Bazán (2020) citó a Mun (2016) quien indicó: “Cuando se cuenta con datos históricos, es recomendable hacer una aproximación estadística o cuantitativa; mientras que, si se carece de estos datos, el único recurso es un juicio de valor o un acercamiento cualitativo” (p. 207). Además, Menacho (2013) concluyó: “El análisis de series de tiempo tiene como propósito estudiar el comportamiento de una variable cuyos datos son registrados cronológicamente en el tiempo y predecir sus valores futuros para apoyar la toma de decisiones” (p. 245). Las hipótesis específicas fueron las siguientes:

H1: La complexión, el rango de edad, la presencia de acné y verrugas, los bigotes, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares están relacionados con los casos de desaparición de varones.

Delgado et al. (2014) emplearon la técnica de regresión logística binaria en el rendimiento académico y el modelo resultante tuvo una estimación correcta del 80.5% de su muestra que fue del total de 298 (p. 310). Además, Pérez (2017) usó el modelo de la regresión logística en su predicción de riesgo crediticio, el cual fue aceptable con una sensibilidad del 70%, una muestra de 155 organizaciones y una población de 678 (p. 241). Asimismo, Reyes et al. (2007) evaluaron la regresión logística aplicada en el rendimiento estudiantil donde obtuvieron 71.4% (rendimiento general) y 75% (rendimiento con la aprobación de cuatro cursos a más), culminando el análisis y con esos resultados indican que es un buen procedimiento para predecir (p. 119).

H2: El uso del modelo ARIMA mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 95%.

Sosa et al. (2019) aplicaron modelos de pronósticos ARIMA, consiguiendo una precisión aproximada del 95% para el monitoreo de parámetros ambientales (p. 55). Además, Sánchez et al. (2014) evaluaron que con la aplicación de los modelos ARIMA, la cual obtuvo un pronóstico con errores absolutos porcentuales por debajo del 15% lo que se aproxima a un 85% en su pronóstico de producción de leche (p. 217).

Palomino et al. (2020) aplicaron el modelo ARIMA en las variables ambientales (temperatura, humedad, velocidad del aire), posteriormente evaluaron el ruido blanco donde obtuvieron su probabilidad menor al 5% (p. 9). Adicionalmente, Córdova et al. (2021) concluyeron: “Los resultados obtenidos con el modelo ARIMA comparados con los datos observados, muestran un ajuste adecuado de los valores; y aunque este modelo, de fácil aplicación e interpretación, no simula el comportamiento exacto en el tiempo este puede considerarse una herramienta simple e inmediata para aproximar el número de casos”, en los casos de COVID-19 en Perú (p. 73).

H3: El uso del modelo Holt-Winters mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 85%.

Roque et al. (2016) quienes estudiaron la técnica de predicción Holt-Winters y su modelo final obtuvo un 84.42% de grado de confianza en el pronóstico de ventas (p. 126). Asimismo, Mariño et al. (2021) concluyeron que aplicando el modelo de suavizamiento Holt-Winters se presentó menor nivel de error, en términos porcentuales alrededor del 3% en la demanda eléctrica del sector de explotación de minas y canteras en Colombia (p. 19). Al respecto, Banda et al. (2014) concluyeron que el método Holt-Winters es una herramienta que permite predecir los flujos de efectivo cuando estos presentan

estacionalidad, como ejemplo: un día de la semana, una semana al mes, un mes al año, un bimestre del año, etc. (p. 19).

H4: El uso del modelo ETS mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 90%.

Roque (2016) estudió la técnica de predicción ETS, cuyo modelo final obtuvo un 90.51% de grado de confianza en el pronóstico de ventas (p. 127). Además, Del Pilar et al. (2018) añadieron: “El modelo de ETS sin tendencia y Sarima fue un 99,999912% mejor, el modelo ETS con tendencia y Holt-Winters fue un 99,99983% mejor y el modelo ARIMA un 99,99997% mejor” (p .44). Asimismo, Del Pilar et al. (2018) concluyeron que en el modelo ETS obtuvo un 17% de mejor resultado con tendencia en las metodologías para el pronóstico (p. 45).

II. MARCO TEÓRICO

En este segundo capítulo se muestran todos los antecedentes de fuentes confiables, por lo que cada una de estas contiene su objetivo de investigación, su diseño metodológico, sus conclusiones y sus recomendaciones. Posteriormente se detallarán las teorías relacionadas con el tema, tales como la metodología, las series de tiempo y diversas herramientas.

Los antecedentes sirven para aportar nuevas ideas al estudio y son útiles para compartir y aprender sobre los hallazgos, hechos por otros investigadores (Arias, 2020, p. 17). Además, Arias (2020) concluyó: “Uno de los primeros pasos más importante al realizar un estudio de investigación es revisar los antecedentes existentes, es decir, otros estudios similares al nuestro, para familiarizarse con el tema y la teoría disponible en el área de interés” (p. 17).

Barbulescu et al. (2021) evaluaron el pronóstico del consumo de carga diaria, esta es una cuestión bastante fundamental para cada uno de los operadores de los sistemas de distribución; dentro de sus objetivos se utilizaron varios procedimientos: redes neuronales artificiales y tres procedimientos convencionales; estos últimos se basan en la aproximación lineal, el ajuste de la curva y la decisión (p. 3).

Barbulescu et al. (2021) evaluaron que el punto de partida está representado por un monumental conjunto de datos procedentes de un operador real del sistema de distribución en nuestro territorio, donde se ha desarrollado una herramienta de programa, dentro de Matlab, para el procedimiento basado en la IA (inteligencia artificial); estudiaron una enorme cantidad de datos y con esto se realizó la predicción para todas las ramas de distribución pertenecientes a ese operador (p. 3). Asimismo, calcularon varios índices para poder proporcionar comentarios relacionados, por lo que, realizadas todas las predicciones, se culminó con un análisis detallado; asimismo, también se ha proporcionado una jerarquía basada en las características de rendimiento (Barbulescu et al., 2021, p. 4).

Severiukhina et al. (2020) evaluaron que un sistema basado en una mezcla de varios componentes para la modelización y predicción en paralelo de procesos en redes con asimilación de datos reales de la red (p. 5). Además, la

principal novedad de este trabajo radica en la asimilación de datos para la predicción de procesos en redes sociales, lo que permite mejorar la calidad de la predicción (Severiukhina et al., 2020, p. 5). Estos consideraron la red social VK como fuente de información para decidir los tipos de entidades y los parámetros del modelo (Severiukhina et al., 2020, p .5).

Severiukhina et al. (2020) indicaron que los resultados del pronóstico no deberían perder su relevancia a lo largo de los cálculos, evaluando a fin de obtener el resultado del pronóstico para redes con cientos de miles de nodos en un tiempo razonable, estableciendo un paralelismo en el proceso de simulación, por lo que utilizaron la métrica del MAE para la microescala, el criterio de Kolmogorov-Smirnov para la dinámica agregada (p. 8). La calidad en el régimen operacional también se considera por el número de lotes con datos asimilados a fin de tener la precisión requerida y la interacción de tiempo de cálculo en los marcos del período de previsión (Severiukhina et al., 2020, p. 9).

Además, los resultados incluyen estudios experimentales de las propiedades funcionales, la escalabilidad y, por consiguiente, el rendimiento del sistema. Mammadova et al. (2020) evaluaron una estrategia para pronosticar la población utilizando el modelo de series temporales difusas, esta técnica basada en series temporales, para la predicción del número de habitantes (población total, población sin discapacidad, población económicamente activa, población de diferentes grupos de edad, mortalidad y nacimientos, etc.) de la población total de Azerbaiyán y estos se comparan con los resultados de otros modelos de predicción (p. 8).

Mammadova et al. (2020) evaluaron la muestra, el comienzo de la gestión del sistema la cual lleva a cabo su esquema servible y explica el comienzo del rendimiento de cada bloque (p. 9). La base de conocimientos del sistema se basa en las reglas de producción y apoya las elecciones prospectivas en materia de política demográfica remitiéndose al estudio de los resultados previstos de diversos indicadores demográficos (Mammadova et al., 2020, p. 9).

Li et al. (2020) analizaron la predicción con imágenes de series temporales, detallando que las representaciones de series cronológicas basadas

en características han atraído una atención considerable en una amplia gama de métodos de análisis de las mismas (p. 15). Además, Li et al. (2020) evaluaron la utilización de características de series cronológicas para el promediar modelos de pronóstico los cuales han sido un nuevo foco de investigación en la comunidad de pronosticadores, no obstante, la mayoría de los enfoques existentes dependen de la elección manual de un conjunto apropiado de características (p. 15).

Li et al. (2020) evaluaron primero transformar las series temporales en gráficos de recurrencia, de los que se pueden extraer características locales utilizando algoritmos de visión artificial; las características extraídas se utilizan para el promedio del modelo de pronóstico (p. 16). Asimismo, Li et al. (2020) concluyeron que los experimentos que muestran la predicción basada en características extraídas automáticamente, con menos intervención humana y una visión más completa de los datos brutos de las series temporales, arroja rendimientos muy comparables con los mejores métodos en el mayor conjunto de datos de la competición de predicción y supera los mejores métodos en el conjunto de datos de la competición de predicción (p. 16).

Blaconá et al. (2012) explicaron: “Holt-Winters y ETS utilizan diferentes criterios de optimización, mientras Holt-Winters emplea valores heurísticos para el Estado inicial y luego estima los parámetros de optimización por el error medio mínimo cuadrático (MSE), mientras que en los modelos ETS el Estado inicial y la optimización de los parámetros de suavizado se realiza mediante de la función de verosimilitud” (p. 24). Este análisis realizado en el estudio mediante estas técnicas y teniendo en cuenta el modelo ARIMA permitieron hacer el pronóstico en cada método y su posterior análisis de resultados con los datos trabajados en esta investigación.

Rodríguez et al. (2020) evaluaron las estrategias del Estado Mexicano para minimizar los feminicidios y proporcionaron datos y análisis actualizados sobre ese fenómeno con el objetivo de demostrar la existencia de estrategias de las autoridades para subestimar los hechos y culpar de ellos a las víctimas. Rodríguez et al. (2020) evaluaron trabajos seminales de género, los cuales se apoyan en el trabajo documental de datos oficiales de la Fiscalía de Ciudad

Juárez, así como en informes de organizaciones no gubernamentales. Rodríguez et al. (2020) utilizaron el enfoque cualitativo etnográfico, con observación a los participantes y entrevistas personales.

García y Cunjama (2020) estudiaron la desaparición forzada de personas en México, apuntes para un análisis crítico criminológico. Además, García et al. (2020) explicaron que las sociedades actuales, denominadas modernas o posmodernas, presentan una gran cantidad de problemas políticos, sociales, culturales y económicos importantes. Asimismo, García et al. (2020) concluyeron que lo más importante de ellos está relacionado con el conflicto, la violencia, ya sea criminal (como se define en una norma prescriptiva) o no.

García et al. (2020) evaluaron aquellos hechos individuales, colectivos o institucionales que socavan el curso de la civilización hacia la perfección individual y social, situación que expresa un error en la cadena que se supone que los seres humanos deben seguir para su felicidad. Asimismo, García et al. (2020) concluyeron que es más complejo cuando estos actos de violencia no se producen sólo entre individuos (como peleas, robos, secuestros, homicidios, violaciones, suicidios), sino que tienen como punto de origen instituciones, ya sea el Estado o el mercadeo (torturas, homicidios, desapariciones forzadas, detenciones arbitrarias, violaciones, genocidios, robos, entre otros) (García et al., 2020).

Frigolé (2019) estudió la violencia, el riesgo y la incertidumbre de la desaparición forzada en México a través de las voces de las madres. Asimismo, Frigolé (2019) añadió que la desaparición forzada genera una incertidumbre profunda y una preocupación enorme en el entorno familiar, que el proceso de búsqueda trata de desviar o minimizar. Asimismo, Frigolé (2019) indicó que las madres son las principales responsables de la búsqueda, a veces acompañadas por sus maridos o por uno de sus hijos, o incluso reemplazadas por los hijos si una enfermedad u otra causa grave se lo impide.

Frigolé (2019) indicó que la desaparición de los hijos y la búsqueda de las madres transforman la relación privada entre madre e hijo en una relación pública, lo que da un carácter político a su voz. Asimismo, Frigolé (2019) indicó

que el dolor de las madres transformado en energía social genera solidaridad, lo que se traduce en la creación de numerosos colectivos y organizaciones que reclaman la aparición de los desaparecidos, los buscan activamente y critican la política de guerra del gobierno y la impunidad imperante.

El marco teórico es un proceso de investigación que implica la búsqueda científica del investigador, éste debe hacer una investigación exhaustiva en textos, artículos científicos, tesis, foros, informes de organizaciones gubernamentales y no gubernamentales, informes de patentes, materiales audiovisuales e incluso sitios web alineados con su situación problemática, objetivos, cuestiones y el tema del estudio en particular (Arias, 2020, p. 18).

A continuación, se detallaron las diferentes metodologías que se pueden aplicar en una investigación similar como son la CRISP-DM la cual se divide en seis fases, la KDD contando con cinco etapas, la metodología SEMMA con cinco etapas y Catalyst con dos modelos circunstancias de negocios y explotación de información, posteriormente se dará una pequeña definición de cada uno de estas. Sifuentes (2018) indicó: “La Metodología CRISP-DM (Proceso estándar industrial para la minería de datos) incluye seis fases las cuales se apreciarán en el anexo 8” (p. 48).

Vialardi et al. (2011) evaluaron que la metodología CRISP-DM se centra en el desarrollo de métodos y técnicas para dar sentido a los datos (p. 222). La fase de modelización es la más relevante, en la cual se analizan los datos, por lo que el reto más importante es trabajar con una gran cantidad de datos que pueden presentar sus propios problemas como: la falta de los mismos datos, causando volatilidad, etc. (Vialardi et al., 2011, p. 222).

Timarán et al. (2016) detallaron que la metodología KDD (Knowledge Discovery in Databases: Descubrimiento de Conocimiento en Base de datos) es básicamente un proceso automático en el que se combinan descubrimiento y análisis y que el proceso consiste en extraer patrones en forma de reglas o funciones a partir de los datos para que el usuario los analice, las cuales están en el anexo 9 (p. 64). Esta metodología se basa en su mayor parte en su proceso y análisis de datos.

Hernández y Dueñas (2009) detallaron que la metodología SEMMA (muestra, explorar, modificar, modelo, evaluar) fue propuesta por el SAS Institute. En la metodología SEMMA se define el proceso de selección, exploración y modelado aplicado a cantidades significativas de datos almacenados que permitan el descubrimiento de patrones como herramientas de apoyo para el negocio con un conjunto de herramientas funcionales enfocadas hacia los aspectos de desarrollo propio de un modelo de minería, los cuales se están en el anexo 10 (Hernández y Dueñas, 2009, p. 83).

Catalyst es una metodología también llamada como P3TQ (producto, lugar, precio, tiempo, cantidad) y está compuesta por dos modelos: un modelo de negocio y un modelo de explotación de la información (Aquino et al., 2015, p. 280). La metodología Catalyst ofrece una guía para el desarrollo y la construcción de un modelo con el objetivo de abordar un problema o una oportunidad de negocio ofreciendo una guía para la realización y ejecución de modelos de explotación de información (Aquino et al., 2015, p. 280). Los modelos de la metodología Catalyst están en el anexo 11.

Tras una investigación exhaustiva de las diferentes metodologías, se seleccionó la metodología CRISP-DM, de libre utilización, en la que no se necesitan más que las bases teóricas para poner en práctica el sistema de pronóstico geolocalizado. En el anexo 12 se aprecia un formato donde se detalla la selección de la metodología a trabajar.

A continuación, se detalla las diferentes plataformas que se pueden aplicar en una investigación similar como son Power BI y Tableau. Power-BI es una plataforma de Microsoft por lo que tiene mayor presencia en el mercado o usuarios y tiene una visualización de datos que puede interactuar entre sí mismo y un mismo dashboard (Microsoft Power BI, 2020). Tableau ofrece análisis de datos para comprender los datos de una manera más visual, por lo que es una plataforma de análisis de un extremo a otro más potente, segura y flexible (Tableau, 2020, π. 2). Asimismo, la plataforma Dundas BI es “Una plataforma revolucionaria para todas sus necesidades de inteligencia empresarial” (Dundas BI, 2020).

Tras la comparación realizada con las tres plataformas descritas brevemente y analizando los aspectos tanto técnicos como internos se seleccionó la plataforma de Power BI para desarrollar el Front End (visual) del sistema. En el anexo 13 se aprecia una tabla con los criterios de selección de la plataforma a trabajar.

A continuación, se definen los términos precisión, efecto, pronóstico e impacto. Estos términos fueron definidos por los especialistas de la Real Academia Española (RAE) de la siguiente manera:

- Precisión: (a) calidad de la precisión y (b) dicho a un aparato, una máquina, un instrumento, etc. Construido con singular cuidado para obtener los mejores resultados posibles (Real Academia Española, 2020a, π. 1).
- Efecto: (a) el que sigue en virtud de una causa y (b) técnicas de algunos obstáculos, trucos para provocar ciertas impresiones (Real Academia Española, 2020b, π. 1).
- Pronóstico: la predicción de algo futuro desde el principio (Real Academia Española, 2020c, π. 1).
- Impacto: Efecto producido en la opinión pública por un acontecimiento, una disposición de la autoridad, una noticia, etc. (Real Academia Española, 2020d, π. 1).

La minería de datos se define como el proceso de extraer una gran cantidad de datos en patrones ocultos, para ello, la predicción o previsión entra en el foco de la minería de datos, por lo que esta toma varias disciplinas como la estadística, la informática y las matemáticas con el fin de encontrar patrones que son útiles y conocer los grandes bloques de datos (Yavuz et al., 2019, p. 3715).

Hurtado (2005) citó a Mamani et al. (2017), quienes concluyeron: “El clustering es una de las principales técnicas de modelado de la minería de datos la cual consiste en dividir la información en grupos diferentes, internamente los miembros de cada grupo son muy similares unos de otros y disímiles respecto a los miembros de los otros grupos. Los grupos o clusters pueden ser usados para clasificar nuevos datos” (p. 123). Los especialistas de Power Data (2018)

detallaron “El proceso ETL (Extracción, Transformación, Carga) tiene como objetivo el facilitar el movimiento de los datos y la transformación de los mismos” (π. 1).

El Registro Nacional de Desaparecidos y Desaparecidas (RNPED) de México integra los datos de personas no localizadas obtenidos a partir de las denuncias presentadas ante la autoridad ministerial correspondiente (Gobierno de México, 2020, π. 1). Este registro solo incluye a las personas que en el momento de la audiencia judicial permanecieron en paradero desconocido (Gobierno de México, 2020, π. 1). Además, la base de datos RNPED del fuero común incluye investigaciones previas iniciadas por desaparición, por Estado, las que permanecen sin localizar a la fecha del tribunal correspondiente (Gobierno de México, 2020, π. 1).

III. METODOLOGÍA

En el tercer capítulo se detalla el tipo y diseño de investigación, asimismo los métodos o tecnologías utilizadas para reunir la información para el desarrollo de este estudio. Asimismo, se definirá la variable con su contenido, la matriz de operacionalización, definición conceptual, definición operacional, indicadores y el instrumento.

3.1 Tipo y diseño de investigación

Esta investigación es de tipo aplicada, porque depende de los resultados y avances de la investigación; es decir, toda investigación aplicada requiere un marco teórico, aunque lo que interesa son las consecuencias prácticas (Muntané, 2010, p. 221). El enfoque de la investigación fue cuantitativo y se denomina así porque trata de fenómenos que pueden medirse mediante el uso de técnicas estadísticas para el análisis de los datos reunidos, su propósito más importante radica en la descripción, explicación, predicción y control objetivo de sus causas (Sánchez, 2018, p. 104). El diseño de investigación se centra en el diseño no experimental. Al respecto, Hernández et al. (2014) indicaron:

> En un estudio no experimental no se genera ninguna situación, sino que se observan situaciones existentes, no provocadas intencionalmente en la 26 investigación por las personas que la realizan. En la investigación no experimental ocurren variables independientes y no es posible manipularlas, no tenemos control directo sobre estas variables ni podemos influir en ellas, porque ya han sucedido, así como sus efectos (p. 152). Dicho de otro modo, los diseños no experimentales se pueden clasificar en transaccionales (transversal) y longitudinales. (p. 152)

Los diseños transeccionales (transversales) se dividen en tres: exploratorios, descriptivos y correlacionales-causales (Hernández et al*.,* 2014, p. 157). Esta investigación tuvo un diseño correlacional-causal, el cual describe relaciones entre dos o más categorías, conceptos o variables en un momento determinado; por lo tanto, el diseño debe limitarse a establecer relaciones entre variables sin precisar sentido de causalidad o pretender analizar relaciones causales (Hernández et al*.,* 2014, p. 157).

Dentro de los diseños longitudinales se dividen en tres tipos: diseños de tendencia, diseño de análisis evolutivo de grupos y diseños panel. De los cuales esta investigación acogerá el diseño longitudinal de tendencia, son aquellos que analizan cambios al paso del tiempo en categorías, conceptos, variables o sus relaciones de alguna población en general, su característica distintiva es que la atención se centra en la población o universo (Hernández et al*.,* 2014, p. 160).

Teniendo en cuenta lo dicho, la presente investigación cuantitativa cumple con garantizar la consistencia del trabajo y la búsqueda de información en fuentes confiables de la base de datos académica EBSCO, ProQuest, etc., donde se extraen artículos indexados como en Web of Science, Scielo, Scopus y cumpliendo con tener menos de cinco años de antigüedad.

3.2 Variables y operacionalización

La variable del estudio fue el efecto del uso del sistema Geolocalizado de pronóstico de casos de personas desaparecidas. Asimismo, Ñaupas et al. (2014) indicaron que la operacionalización de variables es un procedimiento lógico que consiste en transformar las variables teóricas en variables intermedias, luego éstas en variables o indicadores empíricos y finalmente elaborar índices (p. 191). En el Anexo 1 se muestra la matriz de operacionalización de la variable.

En el presente estudio el error en la precisión será casi nulo, además, Zita (2018) concluyó que la precisión indicada se refiere al grado de proximidad o cercanía de los resultados de diferentes mediciones entre sí. Es decir, cuanto se repite una serie de medidas o acciones, siempre que se utilicen instrumentos similares se mide lo mismo y en las mismas condiciones (π. 8-9).

Legarretaetxebarria (2011) concluyeron que la precisión (o error de ubicación) suele ser el requisito más importante en los sistemas de posicionamiento. Asimismo, Legarretaetxebarria (2011) evaluó que en general, el error de distancia medio se toma como la métrica de rendimiento, que viene dada por la media de la distancia entre la posición estimada y la posición real.

Legarretaetxebarria (2011) concluyó que esta característica única a menudo se tiene en cuenta al evaluar un sistema, cuanto mayor sea la precisión,

mejor será el sistema, pero suele haber un equilibrio entre la precisión y otras mediciones. Además, Legarretaetxebarria (2011) concluyeron que es necesario medir sobre otras características del sistema para buscar ese equilibrio necesario (p. 49).

López y Fachelli (2016) evaluaron que el cálculo del PseudoR^2 es determinar la bondad del ajuste; por lo que quiere decir que cuando un modelo se mejora es en relación así mismo, expresado en un porcentaje (p. 25). Además, López et al. (2016) concluyeron: “Los PseudoR^2 se validan con los resultados de Cox y Snell y del de Nagelkerke” (p. 25).

Vandeput (2019) concluyó que “Medir la precisión (o el error) del pronóstico no es una tarea fácil ya que no existe un indicador único para todos. Solo la experimentación mostrará qué indicador clave de rendimiento (KPI) es mejor para usted” (π. 2). Asimismo, con respecto al MAE, Vandeput (2019) indicó que el error absoluto medio es un KPI muy bueno para medir la precisión del pronóstico, como su nombre lo indica, es la medida del error absoluto (π. 13).

3.3 Población, muestra y muestreo

A continuación, se detalló los conceptos asociados a la población y a la muestra:

A: Población

La población que se toma como modelo para la realización del proyecto de investigación proviene de una base de datos gratuita de México (RNPED) la cual contiene un total de 36156 datos. La mayoría de los países europeos están incorporando mediante la creación de portales de datos abiertos y los correspondientes mecanismos de licencia mediante los cuales se da acceso a las bases de datos públicas (Ramos, 2016, p. 5). Se debe precisar lo siguiente:

- Criterios de inclusión: Total de estados, años del 2007 al 2018, sexo, complexión, señas particulares, dependencia.

- Criterios de exclusión: Fecha, Hora, Estatura, Etnia y Discapacidad.

Estos criterios se detallarán dentro de los resultados en el desarrollo de la metodología escogida para el desarrollo.

B: Muestra

La muestra que se utilizó para el trabajo de investigación proviene de una base de datos gratuita (RNPED). La minería de datos es el término más tradicional para designar el uso de la tecnología y la arquitectura diseñadas para detectar información accionable y obtener así el rendimiento económico y social del gran volumen de información depositada en las bases de datos y que es accesible directamente o a través de Internet, ámbito conocido como datos masivos, para el que se utilizan herramientas de extracción, transformación y carga de diferentes elementos con el fin de obtener nuevos conocimientos (Ramos, 2016, p. 3).

Por lo que, se usó el total de 24,476 datos previamente seleccionados y depurados. Estos datos se pasaron a través de los criterios de inclusión y exclusión. En donde se analizó todos los campos y enlazando los que tienen relación entre sí, revisando los requerimientos y datos que se llegó a necesitar para la elaboración del sistema de pronóstico. En la fase 3.1, pág. 52, limpieza de datos se detalló todos los criterios de exclusión y el porqué del mismo.

C: Muestreo

Este estudio tiene un enfoque del muestreo no probabilístico por lo que interviene el criterio para seleccionar de acuerdo a ciertas características que se han requerido (Ñaupas et al., 2018). El muestreo por cuotas se toma en consideración características de la población con las que se está trabajando. De acuerdo a los datos que se trabajaron en la investigación con un total de 36,156 datos de los cuales se

delimitan por las características desarrolladas en la metodología aplicada (Ñaupas et al., 2018).

3.4 Técnicas e instrumentos de recolección de datos

La colección se basa en instrumentos estandarizados. Es uniforme para todos los casos. Los datos se obtienen mediante observación, medición y documentación. Se utilizan instrumentos que han demostrado ser válidos y confiables en estudios anteriores, o se generan nuevos basados en la revisión de la literatura, se prueban y ajustan. Los ítems o indicadores utilizados son específicos con posibilidades de respuesta o categorías predeterminadas (Hernández et al*.,* 2014, p. 12).

3.5 Procedimientos

El análisis de datos utilizado es de naturaleza cuantitativa, por lo que los datos pueden ser analizados numéricamente, la recopilación de datos se utilizó para probar la hipótesis basada en la medición numérica y el análisis estadístico. Utilizando series de tiempo de pronóstico como los modelos ARIMA, Holt-Winters, ETS y a su vez aplicando la regresión logística binaria definiendo una variable dependiente e independiente. Esto está precisado en el Anexo 3.

3.6 Método de análisis de datos

López et al*.* (2016) evaluaron que la técnica de regresión logística binaria se aplica cuándo se pretende explicar una característica o suceso dicotómico (p. 6). Además, López et al*.* (2016) concluyeron que se considera dos sucesos (dicotómico) de un fenómeno o variable, que se codifican con valores 0 y 1 (p. 11).

3.7 Aspectos éticos

Esta investigación se realiza con una base de datos gratuita que permite disponer de los datos para su uso en el sistema. Además, esta investigación cumple con todos los aspectos del código de ética de la investigación de la Universidad César Vallejo (2020), tal como se aprecia a continuación:

- En el artículo 1 se indica que esta investigación respeta y cumple los máximos estándares de rigor científico, responsabilidad y honestidad, para asegurar la precisión del conocimiento científico, mediante las referencias y fuentes confiables donde se recabaron los documentos garantizando el máximo rigor en la investigación.
- En el artículo 3 se indica el principio ético de probidad, lo que en esta investigación se garantiza con el rigor científico en la realización de todo el proceso de investigación hasta su publicación y con la presentación fidedigna de los resultados; además, no se ha incorporado a autores que no han tenido un aporte significativo en la investigación.

Por lo expuesto, esta investigación garantiza en su totalidad la comprobación de todas las fuentes bibliográficas seleccionadas, citadas con sus respectivos autores. Cabe resaltar que los especialistas del Colegio de Ingenieros del Perú (2020) precisaron en el artículo 8 del Código de Ética lo siguiente: “la conducta profesional del ingeniero y su comportamiento deben ser acordes con los objetivos y fines de la Institución” (p. 2). En el estudio se tiene las bases establecidas para lograr un estudio innovador como es el sistema geolocalizado de pronóstico de casos de personas desaparecidas este tiene una visión innovadora en todos los aspectos.

Asimismo, en el Artículo 14 se indica: “los ingenieros están al servicio de la sociedad. Por consiguiente, tienen la obligación de contribuir al bienestar humano, dando importancia primordial a la seguridad y adecuada utilización de los recursos en el desempeño de sus tareas profesionales” (p. 3). Por lo expuesto, este estudio tiene la prioridad de dar a conocer los puntos de riesgo para aportar con los conocimientos brindados y plasmarlos en el sistema para ayudar a la sociedad con un tema muy delicado que es la desaparición de personas mediante pronósticos de series de tiempo.

IV. RESULTADOS

En el presente capítulo se logró explicar la aplicación de la metodología CRISP-DM para el desarrollo en el proyecto de minería de datos (Espinoza, 2019, p. 2). A continuación, se detalla cada fase de las seis que tiene y cómo se ha acoplado al sistema geolocalizado de pronóstico de casos de personas desaparecidas, como la compresión del negocio, la comprensión de los datos, la preparación de los datos, el modelado, la evaluación y el despliegue.

IV.1 Fase 1: Comprensión del problema o negocio

IV.1.1 Identificación del problema

Consiste en la descripción de los problemas que se han encontrado como la falta de un sistema de personas desaparecidas y menos si se habla de un sistema geolocalizado orientado en el pronóstico. Por lo que en los capítulos anteriores se ha detallado la problemática de lo que conlleva a lo que es personas desaparecidas y mediante un sistema de pronóstico usando métodos de pronóstico y series de tiempo se proyectará en qué estados son los más riesgosos en el país de México. A partir del uso de la base de datos libre que permitió realizar el estudio al RNPED del fuero común (Gobierno de México, 2020).

IV.1.2 Determinación de objetivos

En el presente estudio el objetivo es determinar el efecto del sistema geolocalizado de pronóstico de casos de personas desaparecidas apoyándonos con la metodología CRISP-DM, técnicas de regresión y series de tiempo. A continuación, se detallarán objetivos que se quieren lograr con el sistema:

- Elaborar un modelo eficaz de extracción de datos para la elaboración de un pronóstico más acertado mediante las técnicas presentadas.
- De acuerdo a las técnicas elegidas se determinará la dependencia en la cual es más probable que ocurra una desaparición, a la vez se realizó el estudio de diferentes características con las cuales también se elaboran la mayor precisión.
- La predicción se trabaja bajo el entorno de RStudio y los resultados se plasmarán mediante la plataforma de Power BI.

IV.1.3 Evaluación de la situación actual

La base de datos descargada tiene un tamaño de 6.94 MB (6,9440 KB) y está alojada en formato Microsoft Excel. Dentro de esta existe un total de 36,156 datos y cuentan con diferentes campos que ayudaran al estudio, esta incluye la fecha que se le vio por última vez, hora en que se le vio por última vez, país en donde se le vio por última vez, entidad en la que se le vio por última vez, municipio en la que se le vio por última vez, localidad en la que se le vio por última vez, nacionalidad, estatura, complexión, sexo, edad, descripción, etnia, discapacidad, dependencia que se envió la información. Para la realización del estudio se usó los recursos de software como los que se detallaron a continuación:

En la tabla 1 se describe lo especificado por los especialistas de Microsoft (2020) quienes indicaron los siguientes requisitos del software de Microsoft SQL Server Management Studio:

Tabla 1 *Requisitos de Microsoft SQL Server*

Componente	Requisito
Disco Duro	▪ SQL Server requiere un mínimo de 6 GB de espacio disponible en el disco duro
Monitor	▪ SQL Server requiere Super-VGA (800x600) o un monitor de mayor resolución.
Internet	▪ Requiere acceso a Internet
Memoria	▪ Mínimo: Ediciones Express: 512 MB ▪ Recomendado: Ediciones Express: 1 GB
Velocidad del procesador	▪ Mínimo: x64 Procesador: 1,4 GHz ▪ Recomendado: 2,0 GHz o más rápido
Tipo de procesador	▪ Procesador x64: AMD Opteron, AMD Athlon 64, Intel Xeon con soporte Intel EM64T, Intel Pentium IV con soporte EM64T
Sistema Operativo	▪ Windows 10 TH1 1507 o Superior ▪ Windows Server 2016 o Superior

En la tabla 2 se describe lo especificado por los especialistas de RStudio (2020) quienes indicaron los siguientes requisitos como son la conexión a internet, un mínimo de memoria RAM, y el tipo de compatibilidad del sistema operativo del software de RStudio:

Tabla 2 *Requisitos de RStudio*

Componente	Requisito
Internet	▪ Requiere conexión a la red
Memoria	▪ Requiere un mínimo de 32 MB
Sistema Operativo	▪ Windows 9x/ME/NT4.0/2000/XP/2003/Vista/7/8/2012 Server/8.1/10

En la tabla 3 se describe lo especificado por los especialistas de Microsoft Excel (2020) quienes indicaron las especificaciones mínimas que se debe de tener, asimismo los requisitos del software son los siguientes:

Tabla 3 *Requisitos de Microsoft Excel*

Componente	Requisito
Disco Duro	▪ SQL Server requiere un mínimo de 6 GB de espacio disponible en el disco duro
Monitor	▪ SQL Server requiere Super-VGA (800x600) o un monitor de mayor resolución.

En la tabla 4 se describe lo especificado por los especialistas de Microsoft Power BI (2020) quienes indicaron las especificaciones mínimas que se debe de tener, asimismo los requisitos del software son los siguientes:

Tabla 4 *Requisitos de Power BI*

Componente	Requisito
Sistema Operativo	▪ Windows 7/ Windows Server 2008 R2 o Posterior
Memoria	▪ Mínimo: 1GB ▪ Recomendable: 1.5 GB o más
CPU	▪ 64 bits (x64)
Pantalla	▪ 1440 x 900 o 1600 x 900

En las siguientes tablas 5 y 6 se describe la información de los equipos que se han utilizado para el desarrollo del estudio precisando las características de hardware.

Tabla 5 *Información del equipo, cumpliendo los requisitos de hardware para los productos de software del equipo 1*

Información de Equipo	
	Paz Sanchez, Cristian Alexander
Equipo	▪ Laptop Lenovo
Disco Duro	▪ Disco C: 490 GB ▪ Disco D: 440 GB ▪ Disco E: 930 GB
Monitor	▪ Resolución de Escritorio: 1366 x 768 ▪ Resolución de señal activa: 1366 x 768 ▪ Frecuencia de actualización: 60Hz ▪ Profundidad en bits: 8Bits ▪ Formato de color: RGB ▪ Espacio de colores: Rango dinámico estándar (SDR) ▪ AMD Radeon(TM) R5 Graphics
Internet	▪ PING ms: 12 ▪ DESCARGA Mbps: 19.56 ▪ Carga Mbps: 5.43
Memoria	▪ 8 GB
Tipo de procesador	▪ AMD A8-6410 APU with AMD Radeon R5 Graphics 2.00GHz
Sistema Operativo	▪ 64 bits
Procesador	▪ X64
Edición Windows	▪ Windows 10 Home Single Language

Tabla 6 *Información del equipo, cumpliendo los requisitos de hardware para los productos de software del equipo 2*

Información de Equipo	
	Jaime Cordova, Jhordy Yosshi
Equipo	▪ PC
Disco Duro	▪ Disco C: 222 GB ▪ Disco D: 931 GB
Monitor	▪ Resolución de Escritorio: 1920 x 1080 ▪ Resolución de señal activa: 1920 x 1080 ▪ Frecuencia de actualización: 60Hz ▪ Profundidad en bits: 8Bits ▪ Formato de color: RGB ▪ Espacio de colores: Rango dinámico estándar (SDR) ▪ NVIDIA GeForce GTX 1050 Ti

Información de Equipo	
	Jaime Cordova, Jhordy Yosshi
Internet	▪ PING ms: 9 ▪ DESCARGA Mbps: 106.80 ▪ Carga Mbps: 9.85
Memoria	▪ 16 GB
Tipo de procesador	▪ Intel(R) Core(TM) i5-8400 CPU @ 2.800GHz 2.81GHz
Sistema Operativo	▪ 64 bits
Procesador	▪ X64
Edición Windows	▪ Windows 10 Pro

En la tabla 7 se muestran los requisitos de hardware de los equipos (1 y 2) para lograr un mejor rendimiento en el desarrollo del sistema.

Tabla 7 *Tabla cruzada del cumplimiento de los requisitos de hardware para los productos de software de las computadoras disponibles por los investigadores*

Tabla cruzada entre requisitos y nuestros equipos			
		Paz	**Jaime**
		Laptop Lenovo	PC
SQL Server	**Disco Duro**	✓	✓
	Monitor	✓	✓
	Internet	✓	✓
	Memoria	✓	✓
	Velocidad del procesador	✓	✓
	Tipo de procesador	✓	✓
	Sistema Operativo	✓	✓
RStudio	**Internet**	✓	✓
	Memoria	✓	✓
	Sistema Operativo	✓	✓
Excel	**Disco Duro**	✓	✓
	Monitor	✓	✓
Power BI	**Sistema Operativo**	✓	✓
	Memoria	✓	✓
	CPU	✓	✓
	Monitor	✗	✓

IV. 2 Fase 2: Comprensión de los datos

IV.2.1 Recolección de datos

En el estudio se utilizaron los datos abiertos del “Registro Nacional de Datos de Personas Extraviadas o Desaparecidas (RNPED)” del país de México. No se tuvo problemas alguno como se detalló anteriormente ya que son datos abiertos y son libres, sólo se tuvo unos inconvenientes al separar los datos.

En la figura 1 se muestra la página oficial del Gobierno de México de donde se consiguió la recolección de datos para trabajar.

Figura 1. Pantalla de los datos abiertos del RNPED

En la figura 2 se muestra los datos que se usaron para el estudio, los cuales fueron seleccionados del fuero común por la cantidad de datos.

Figura 2. Pantalla de la base de datos del RNPED del fuero común

En la tabla 8 se describe una información adicional que fue obtenida desde la página oficial donde se recolectó los datos del fuero común.

Tabla 8 *Información adicional del RNPED del fuero común*

Información Adicional Fuero Común	
Campo	**Valor**
Última actualización	hace 1 año
Formato	CSV
Licencia	LIBRE USO MX
DataStore Activo	True
Estado	Activo
Fecha de última modificación de datos	2018-05-31
Periodo de actualización	R/P1M
Periodo cubierto por los datos	De 2006-12-01 a 2018-04-30
Herramientas para visualizar los datos	Excel
Espacio geográfico	México

En la figura 3 se muestran los datos ya descargados y exportados al programa de Microsoft Excel el cual en su totalidad arrojó un total 36,156 datos, pero de los cuales se encontraron algunos inconvenientes.

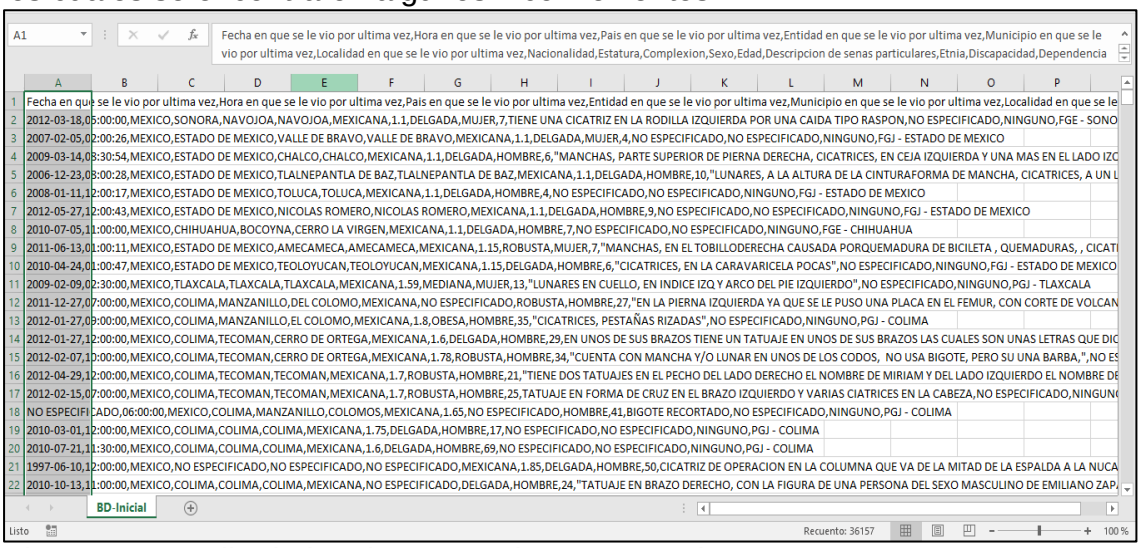


Figura 3. Pantalla de los datos en Microsoft Excel

En la figura 4 se muestra el delimitado de datos lo significa que se separó por columnas. Posteriormente en la figura 5 se selecciona la delimitación por coma, este separa los campos en comas que se encuentren. Luego, en la figura 6 se muestran los valores de las fechas ya que uno de los campos a trabajar es con la fecha de desaparición.

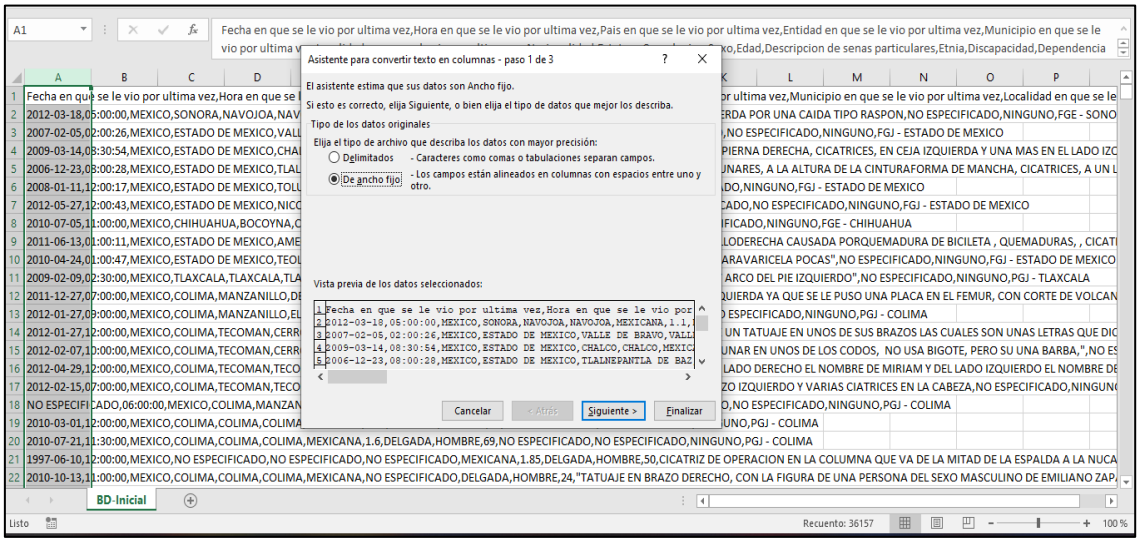


Figura 4. Pantalla de delimitado de datos (A)

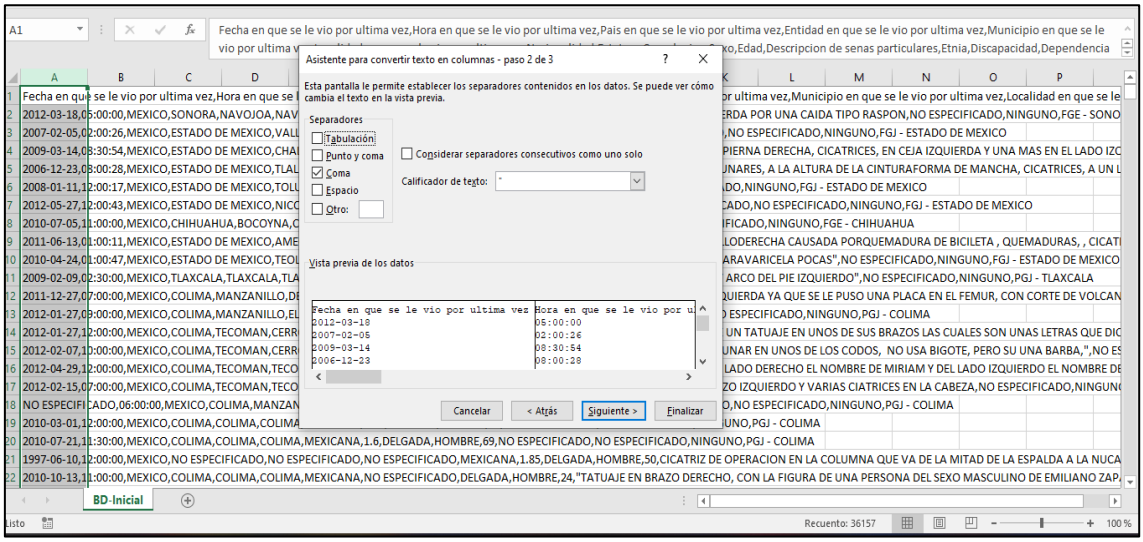


Figura 5. Pantalla de delimitado de datos por coma (A)

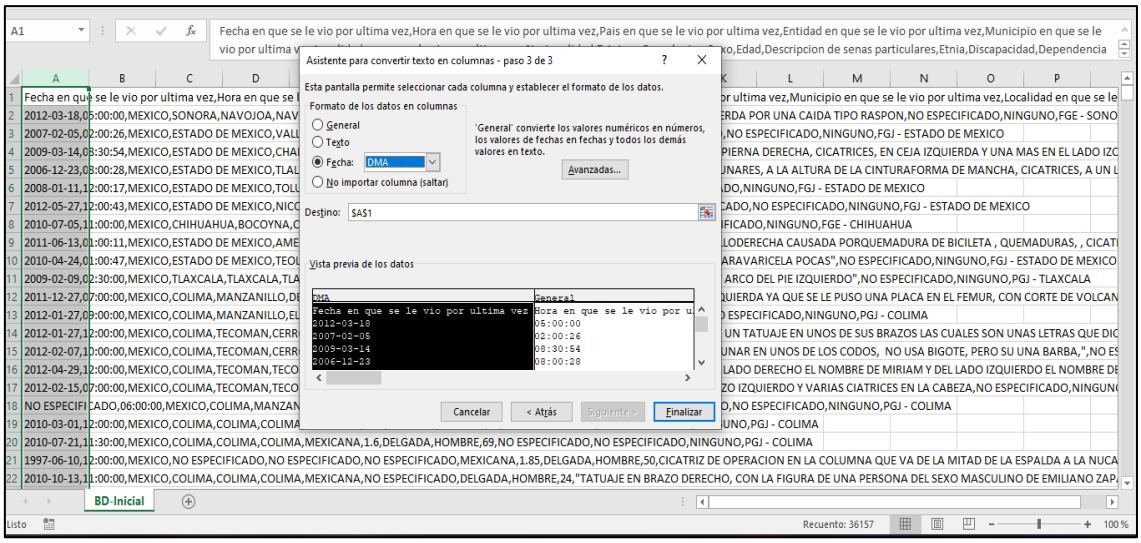


Figura 6. Pantalla de delimitado de datos por fecha (A)

En la figura 7 se detalla que al culminar la delimitación se muestra un error, en el cual los datos presentan valores sobreexpuestos, lo que quiere decir que dentro de los campos había más datos, por lo que no permitía la separación por columnas.

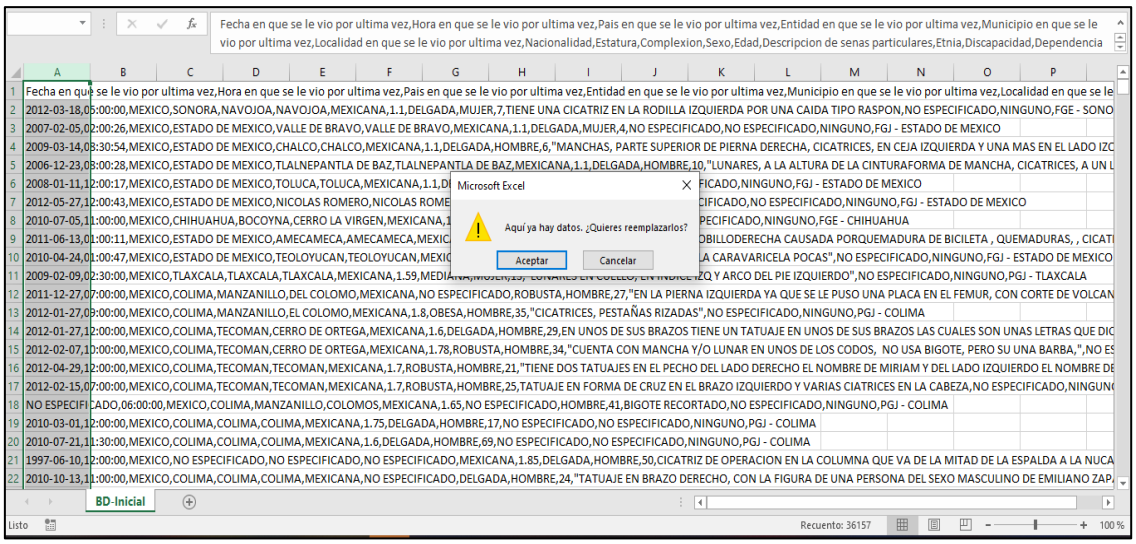


Figura 7. Pantalla de error de delimitación (A)

En la figura 8 se visualiza la razón por la cual no permite delimitar los campos; aplicando un filtro que arrojó que en las columnas se encontraban datos que correspondían a las otras, teniendo 589 datos que no coinciden respectivamente a su columna.

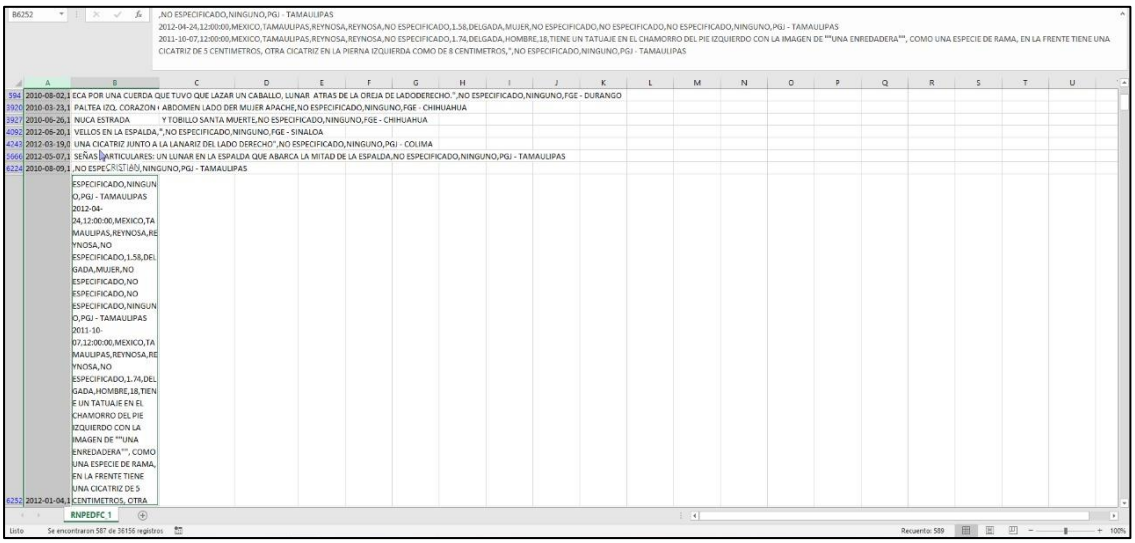


Figura 8. Pantalla de error en los datos (A)

En la figura 9 se muestra que al culminar el análisis de los datos erróneos y analizando los datos originales que era 36,156 datos aumento en 108 datos, que eran los que no permitirán realizar la delimitación por comas, después de este proceso los datos en su totalidad son de 36,265. Por lo que se realizó el mismo proceso de delimitar por comas y por fecha.

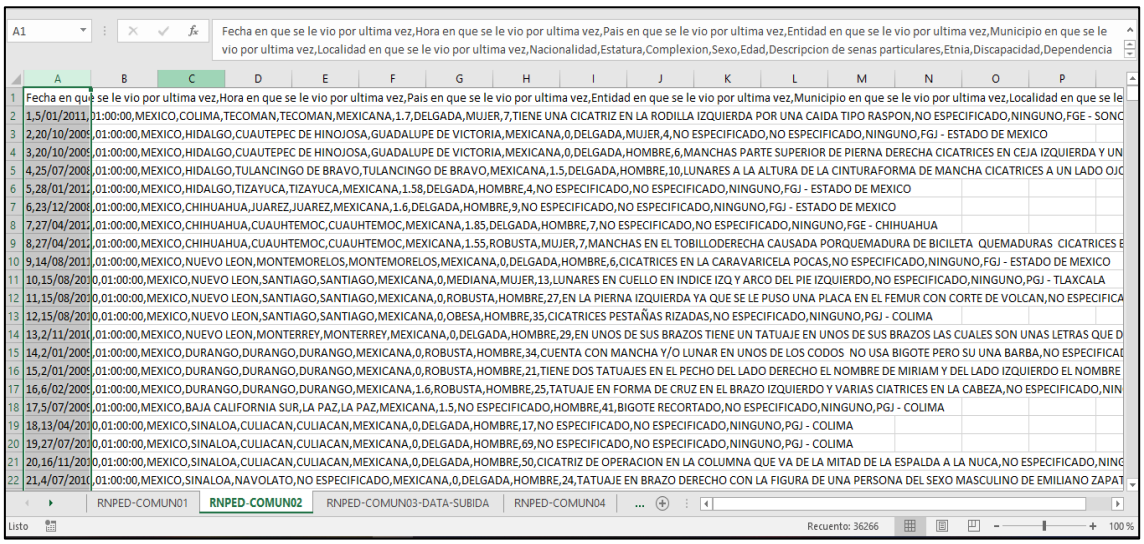


Figura 9. Pantalla de los datos sin delimitar (A)

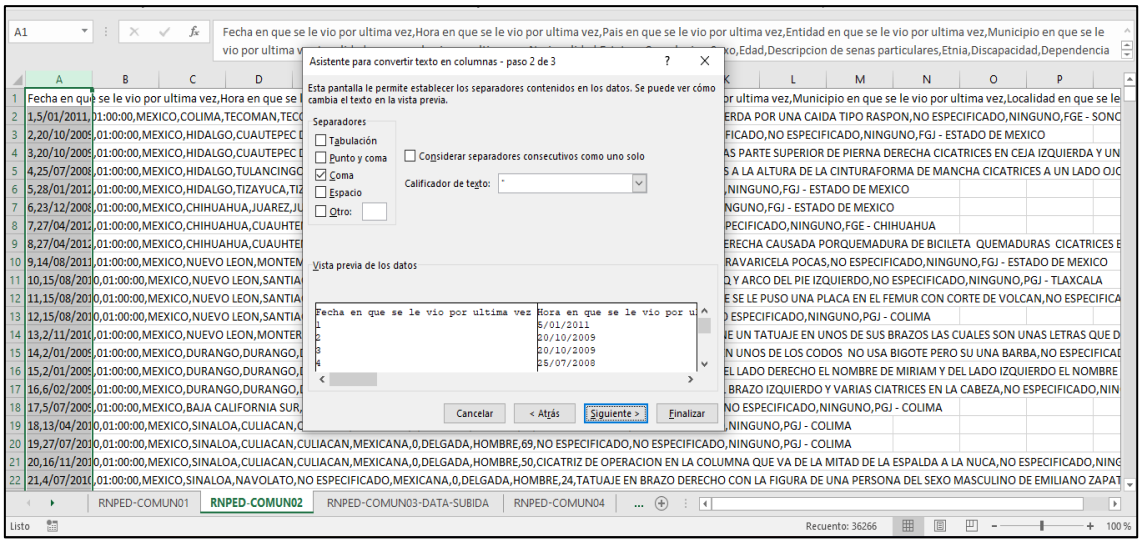


Figura 10. Pantalla de delimitado de datos (B)

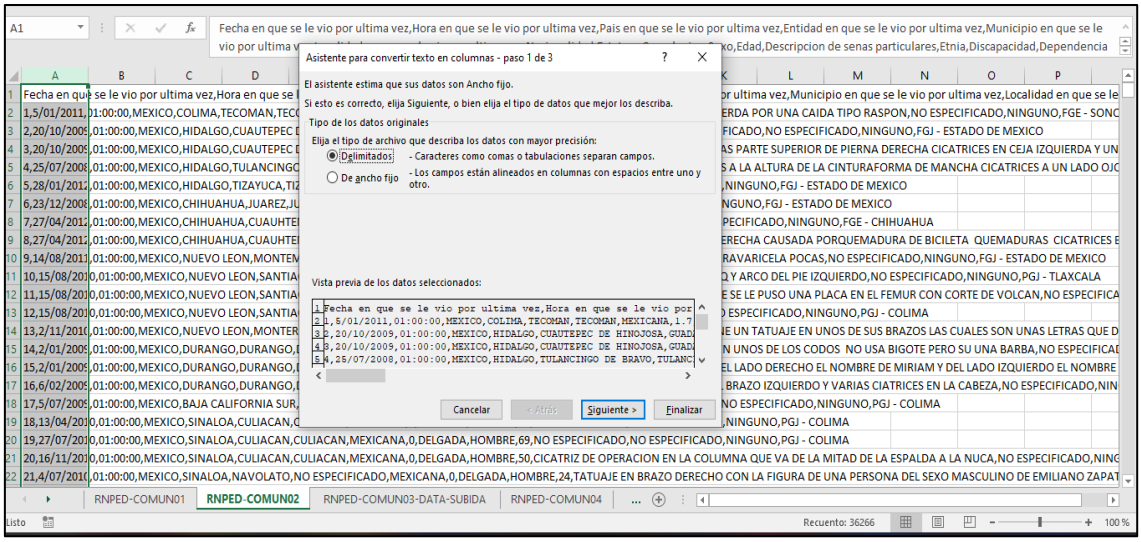


Figura 11. Pantalla de delimitado de datos por coma (B)

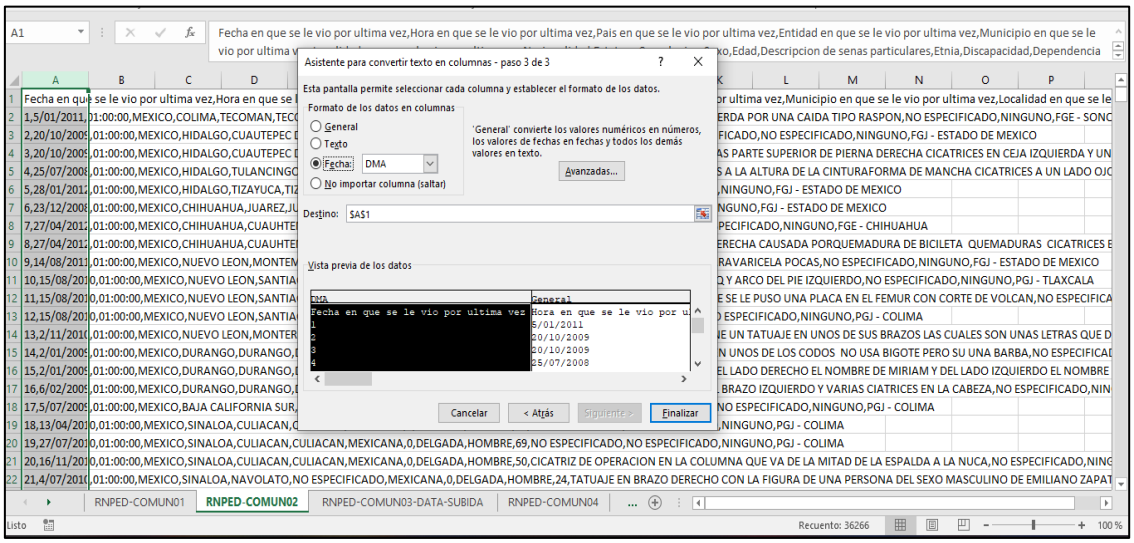


Figura 12. Pantalla de delimitado de datos por fecha (B)

En la figura 13 se muestra que el proceso de la delimitación de datos es correcto, por lo que la redistribución de datos en columnas es correcta como se muestra en la siguiente figura.

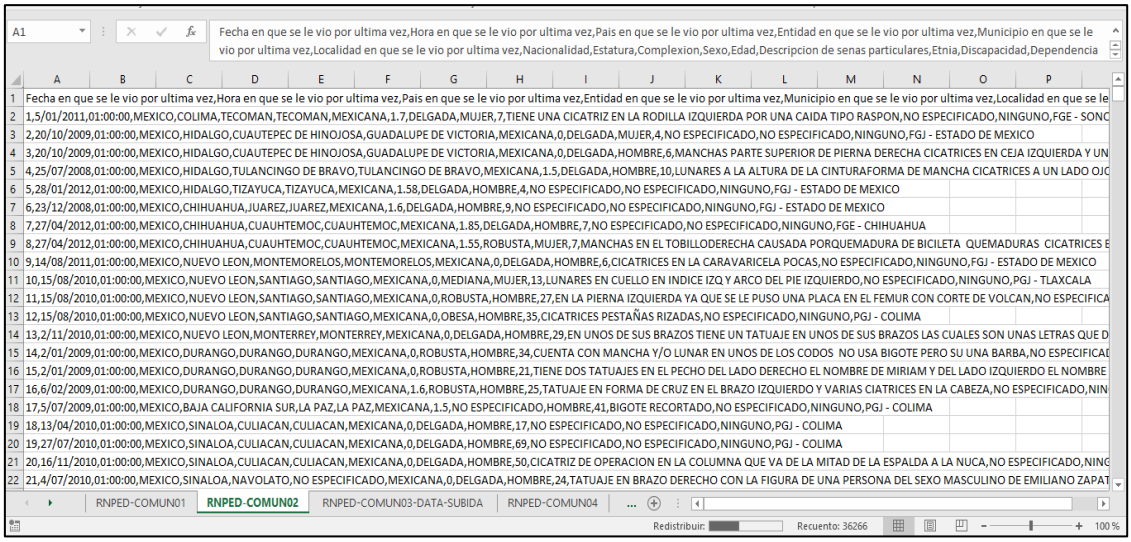


Figura 13. Pantalla de redistribución de datos

En la figura 14 se muestra la delimitación por columnas del total de datos que son en total 36,265 filas.

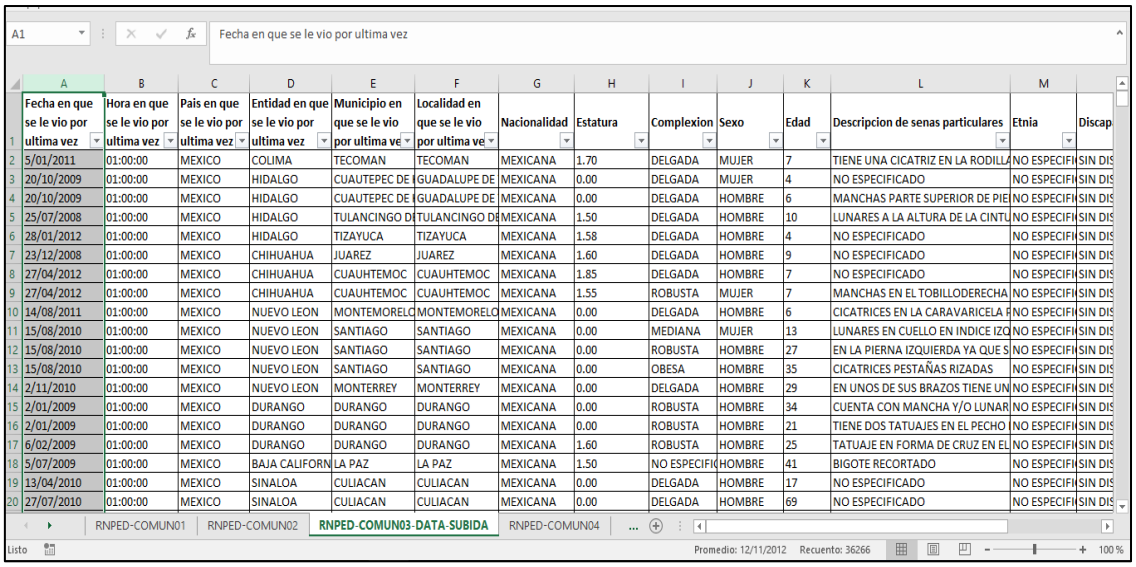


Figura 14. Pantalla de los datos separada por columnas

IV.2.2 Descripción de datos

En la tabla 9 se describen todos los campos que se ha obtenido después de la delimitación de datos de los cuales se realizó la descripción para su mayor comprensión de los mismos.

Tabla 9 *Descripción de los datos originales de los campos de tipo, formato y su volumetría respectiva*

Nombre	**Tipo**	**Formato**	**Volumetría**	**Descripción**
Fecha en la que se le vio por última vez	Fecha	DATE	-	En estos campos se muestra el día, mes y año en la cual se vio por última vez a la persona.
Hora en la que se le vio por última vez	Hora	TIME	-	Intervalo de tiempo en el cual a ocurrió el hecho.
País en que se le vio por última vez	Texto	VARCHAR	50	País donde ocurrió el hecho en este caso todos son del país de México.
Entidad en que se le vio por última vez	Texto	VARCHAR	100	Entidad donde ocurrió el suceso también puede ser llamado Estado por lo está en los 32 estados.
Municipio en que se le vio por última vez	Texto	VARCHAR	100	Municipio donde ocurrió el hecho.
Localidad que se le vio por última vez	Texto	VARCHAR	100	Localidad donde ocurrió el hecho.
Nacionalidad	Texto	VARCHAR	100	La nacionalidad en el campo varía, por lo que hay datos de otros países.
Estatura	Numérico	DECIMAL	(4,2)	Intervalo de estatura.
Complexión	Texto	VARCHAR	100	La complexión puede ser un aspecto influyente.
Sexo	Texto	VARCHAR	50	El sexo entre hombre y mujer.
Edad	Numérico	CHAR	3	Intervalo de años.

Nombre	Tipo	Formato	Volumetría	Descripción
Descripción de señas particulares	Texto	VARCHAR	3000	Descripción de alguna característica que ha tenido que ayudar a reconocer.
Etnia	Texto	VARCHAR	100	Característica en base a la etnia que en el país son variados, pero muchos sin especificar.
Discapacidad	Texto	VARCHAR	100	Si tuvo alguna discapacidad como síndrome de down o autismo.
Dependencia	Texto	VARCHAR	100	Lugar donde se recibió la denuncia correspondiente.

IV.2.3. Exploración de datos

Para el correcto análisis de datos se realizaron los siguientes procedimientos:

- Creación de la base de datos

```
CREATE DATABASE TESISFINAL;
USE TESISFINAL;
```

- Creación de la tabla “DATOS”

```
CREATE TABLE DATOS(
IDDESAPARECIDOS INT IDENTITY(1,1),
FECHA DATE,
HORA TIME,
PAIS VARCHAR(50),
ENTIDAD VARCHAR(100),
MUNICIPIO VARCHAR(100),
LOCALIDAD VARCHAR(100),
NACIONALIDAD VARCHAR(100),
ESTATURA DECIMAL(4,2),
COMPLEXION VARCHAR(100),
SEXO VARCHAR(50),
EDAD CHAR(3),
SENASPARTICULARES VARCHAR(3000),
ETNIA VARCHAR(100),
DISCAPACIDAD VARCHAR(100),
DEPENDENCIA VARCHAR(100),
PRIMARY KEY(IDDESAPARECIDOS)
);
```

Al tratar de subir los datos se tuvo el inconveniente entre los dispositivos que se están usando, realizando un análisis exhaustivo se llegó al punto en que el formato fecha no estaba correcto en un dispositivo por lo que se realizaron dos consultas para la inserción de datos; como se muestra en la siguiente imagen el conflicto en la subida de información.

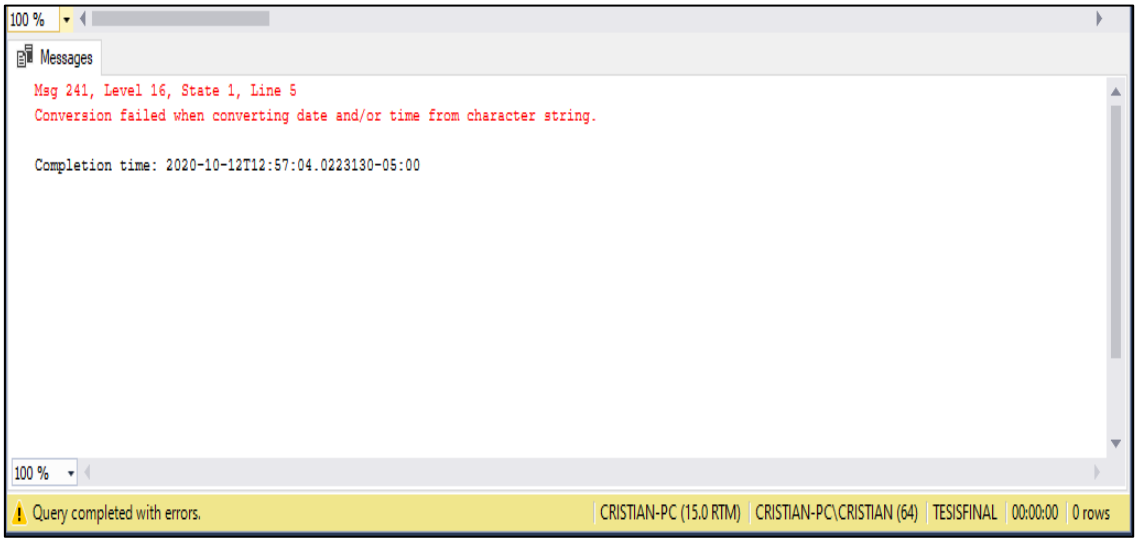


Figura 15. Pantalla de error en la subida de los datos a Microsoft SQL Server

- Consulta para insertar datos del equipo 1

```
----------INSERCIÓN DE DATOS----------

INSERT INTO DATOS VALUES(CONVERT(DATE,'2011-01-05'),'01:00:00','MEXICO','COLIMA','TECOMAN','TECOMAN','MEXICANA',1.7,'DELGADA','MUJER',7,'TIENE UNA CICATRIZ EN LA RODILLA IZQUIERDA POR UNA CAIDA TIPO RASPON','NO ESPECIFICADO','NINGUNO','FGE - SONORA');
INSERT INTO DATOS VALUES(CONVERT(DATE,'2009-10-20'),'01:00:00','MEXICO','HIDALGO','CUAUTEPEC DE HINOJOSA','GUADALUPE DE VICTORIA','MEXICANA',0,'DELGADA','MUJER',4,'NO ESPECIFICADO','NO ESPECIFICADO','NINGUNO','FGJ - ESTADO DE MEXICO');
.
.
.
Hasta los 36,265 datos
----------
```

- Tiempo en subir todos los datos (A)

  Tabla 10 *Tiempo de subida de los datos del equipo 1, mediante inserción de datos en SQL Server*

	Equipo 1
Tiempo	1 minuto con 31 segundos

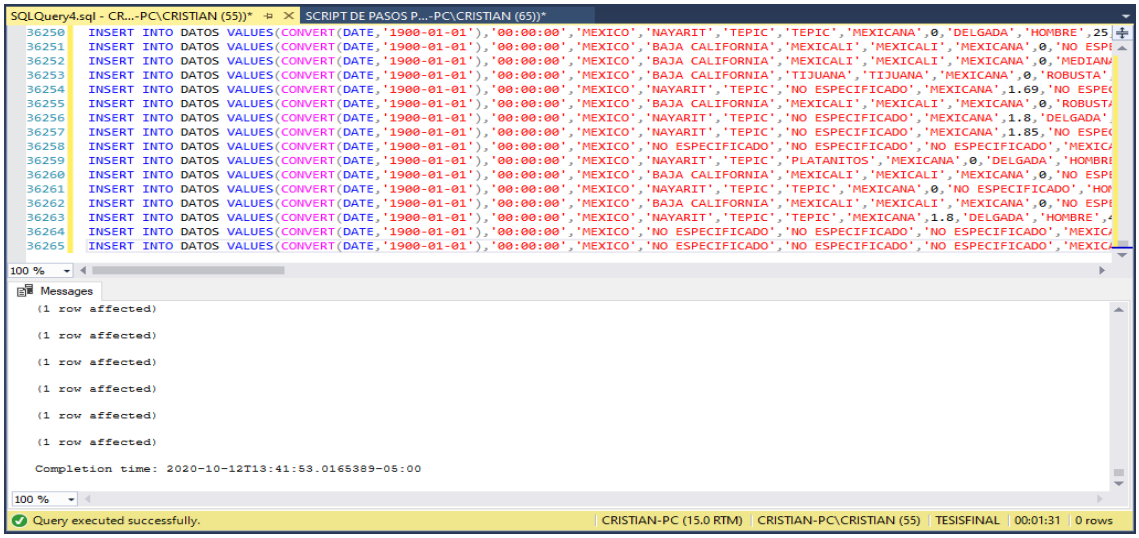


Figura 16. Pantalla de insert (insertar) en el equipo 1 a Microsoft SQL Server

- Consulta para insertar datos equipo 2

```
----------INSERCIÓN DE DATOS----------

INSERT INTO DATOS VALUES ('2007-05-15','12:00:00.0000000','MEXICO','HIDALGO','TULANCINGO DE BRAVO','TULANCINGO','MEXICANA',1.65,'DELGADA','HOMBRE',23,'VERRUGAS','NO ESPECIFICADO','SIN DISCAPACIDAD','FGE - CHIHUAHUA');
INSERT INTO DATOS VALUES ('2009-01-31','11:30:00.0000000','MEXICO','COAHUILA DE ZARAGOZA','TORREON','TORREON','MEXICANA',0,'DELGADA','HOMBRE',25,'VERRUGAS','NO ESPECIFICADO','SIN DISCAPACIDAD','FGJ - ESTADO DE MEXICO');
INSERT INTO DATOS VALUES ('2009-04-09','12:00:00.0000000','MEXICO','TAMAULIPAS','REYNOSA','REYNOSA','NO ESPECIFICADO',0,'DELGADA','MUJER',24,'VERRUGAS','NO ESPECIFICADO','SIN DISCAPACIDAD','FGE - PUEBLA');
.
.
.
Hasta los 36,265 datos
-----------------------------------------------------
```

- Tiempo en subir todos los datos (B)

Tabla 11 Tiempo de subida de los datos del equipo 2, mediante inserción de datos en SQL Server

	Equipo 2
Tiempo	20 segundos

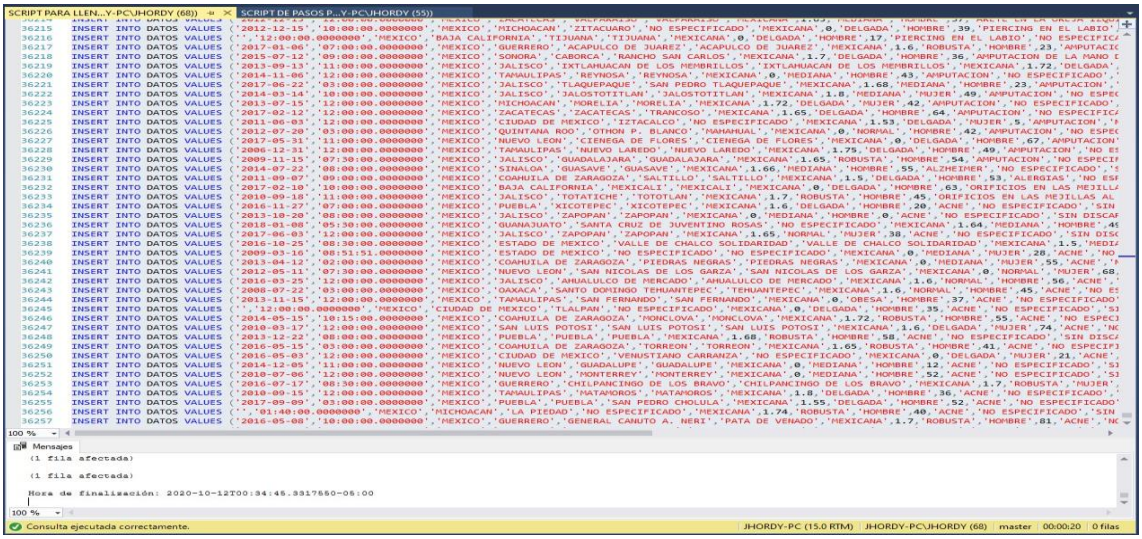
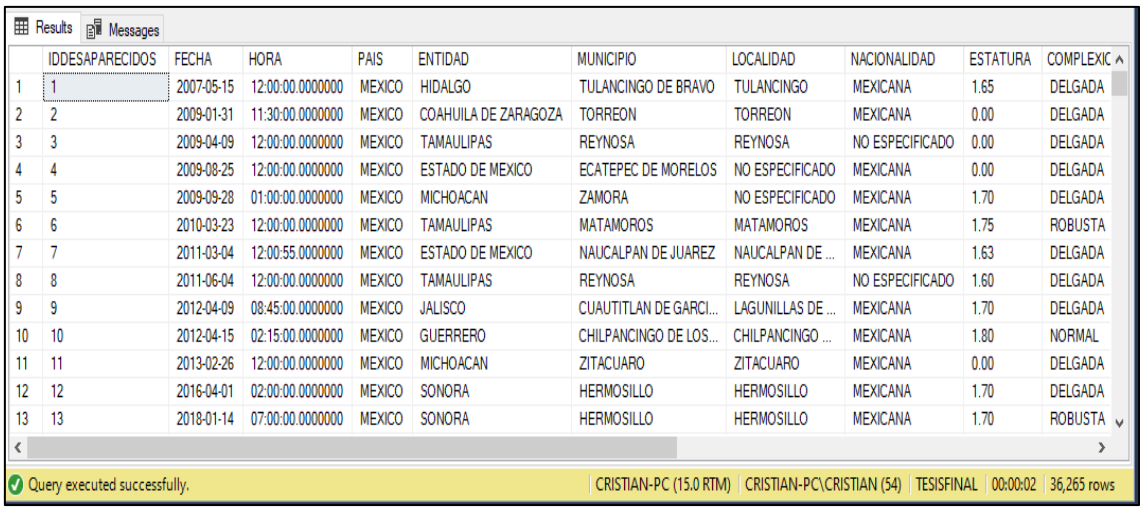


Figura 17. Pantalla de insert (insertar) en el equipo 2 a Microsoft SQL *Server*

```
----------VISTA DE DATOS----------
SELECT * FROM DATOS;
```



	IDDESAPARECIDOS	FECHA	HORA	PAIS	ENTIDAD	MUNICIPIO	LOCALIDAD	NACIONALIDAD	ESTATURA	COMPLEXIC
1	1	2007-05-15	12:00:00.0000000	MEXICO	HIDALGO	TULANCINGO DE BRAVO	TULANCINGO	MEXICANA	1.65	DELGADA
2	2	2009-01-31	11:30:00.0000000	MEXICO	COAHUILA DE ZARAGOZA	TORREON	TORREON	MEXICANA	0.00	DELGADA
3	3	2009-04-09	12:00:00.0000000	MEXICO	TAMAULIPAS	REYNOSA	REYNOSA	NO ESPECIFICADO	0.00	DELGADA
4	4	2009-08-25	12:00:00.0000000	MEXICO	ESTADO DE MEXICO	ECATEPEC DE MORELOS	NO ESPECIFICADO	MEXICANA	0.00	DELGADA
5	5	2009-09-28	01:00:00.0000000	MEXICO	MICHOACAN	ZAMORA	NO ESPECIFICADO	MEXICANA	1.70	DELGADA
6	6	2010-03-23	12:00:00.0000000	MEXICO	TAMAULIPAS	MATAMOROS	MATAMOROS	MEXICANA	1.75	ROBUSTA
7	7	2011-03-04	12:00:55.0000000	MEXICO	ESTADO DE MEXICO	NAUCALPAN DE JUAREZ	NAUCALPAN DE ...	MEXICANA	1.63	DELGADA
8	8	2011-06-04	12:00:00.0000000	MEXICO	TAMAULIPAS	REYNOSA	REYNOSA	NO ESPECIFICADO	1.60	DELGADA
9	9	2012-04-09	08:45:00.0000000	MEXICO	JALISCO	CUAUTITLAN DE GARCI...	LAGUNILLAS DE ...	MEXICANA	1.70	DELGADA
10	10	2012-04-15	02:15:00.0000000	MEXICO	GUERRERO	CHILPANCINGO DE LOS...	CHILPANCINGO ...	MEXICANA	1.80	NORMAL
11	11	2013-02-26	12:00:00.0000000	MEXICO	MICHOACAN	ZITACUARO	ZITACUARO	MEXICANA	0.00	DELGADA
12	12	2016-04-01	02:00:00.0000000	MEXICO	SONORA	HERMOSILLO	HERMOSILLO	MEXICANA	1.70	DELGADA
13	13	2018-01-14	07:00:00.0000000	MEXICO	SONORA	HERMOSILLO	HERMOSILLO	MEXICANA	1.70	ROBUSTA

Figura 18. Pantalla de la vista de los datos originales en Microsoft SQL Server

En la figura 18 se muestran los datos originales para poder familiarizarnos y así conocer cada aspecto de los datos, culminando este proceso se realizó una consulta de vista de datos.

```
----------CONTADOR DE TOTAL DE DATOS----------
SELECT COUNT(*)AS [CANTIDAD TOTAL] FROM DATOS;
```

En la figura 19 se muestra un contador del total de datos para tener a ciencia cierta la cantidad con la que se ha trabajado. El cual arrojó detalladamente al momento de realizar la delimitación de datos que son de 36,265 datos.

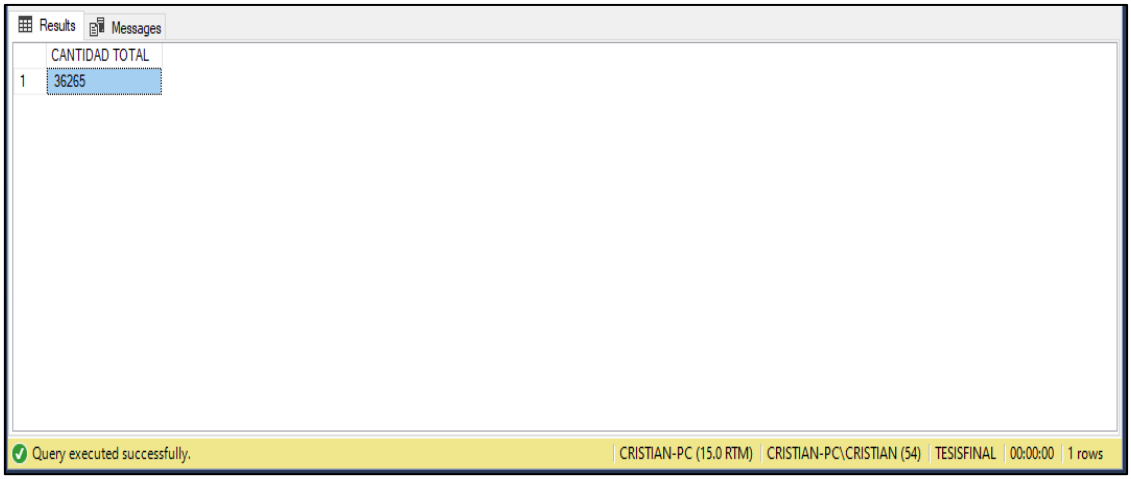


Figura 19. Pantalla del resultado de conteo total de datos dentro de Microsoft SQL Server

- **Fecha en que se le vio por última vez**

```
---------CANTIDAD DE DESAPARECIDOS POR AÑO----------
SELECT YEAR(FECHA) AS AÑO,COUNT(YEAR(FECHA))AS [CANTIDAD
DE DATOS] FROM DATOS
GROUP BY YEAR(FECHA)
ORDER BY YEAR(FECHA) ASC
```

En la tabla 12 se describen los resultados de la consulta para visualizar el conglomerado de datos por fecha.

Tabla 12 *Conteo total del campo de personas desaparecidas*

Año	Cantidad de datos	Año	Cantidad de datos
1900	338	1998	9
1968	1	1999	8
1971	1	2000	18
1972	1	2001	13
1976	1	2002	19
1977	2	2003	16
1978	1	2004	15
1979	3	2005	47

Año	Cantidad de datos	Año	Cantidad de datos
1980	1	2006	89
1984	1	2007	620
1985	2	2008	800
1987	2	2009	1372
1988	2	2010	3206
1989	4	2011	4064
1990	1	2012	3288
1991	2	2013	3650
1992	1	2014	3790
1993	3	2015	3272
1994	1	2016	4525
1995	2	2017	5426
1996	5	2018	1634
1997	9		

En la figura 20 se muestra un cuadro de gráfico que permitirá la visualización más detalla de la columna de personas desaparecidas por año.

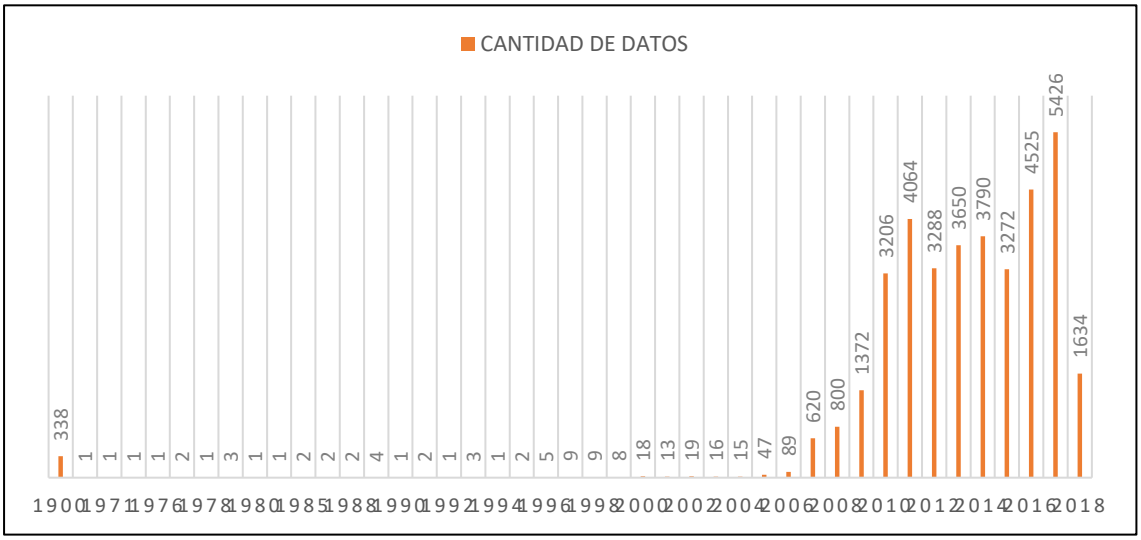
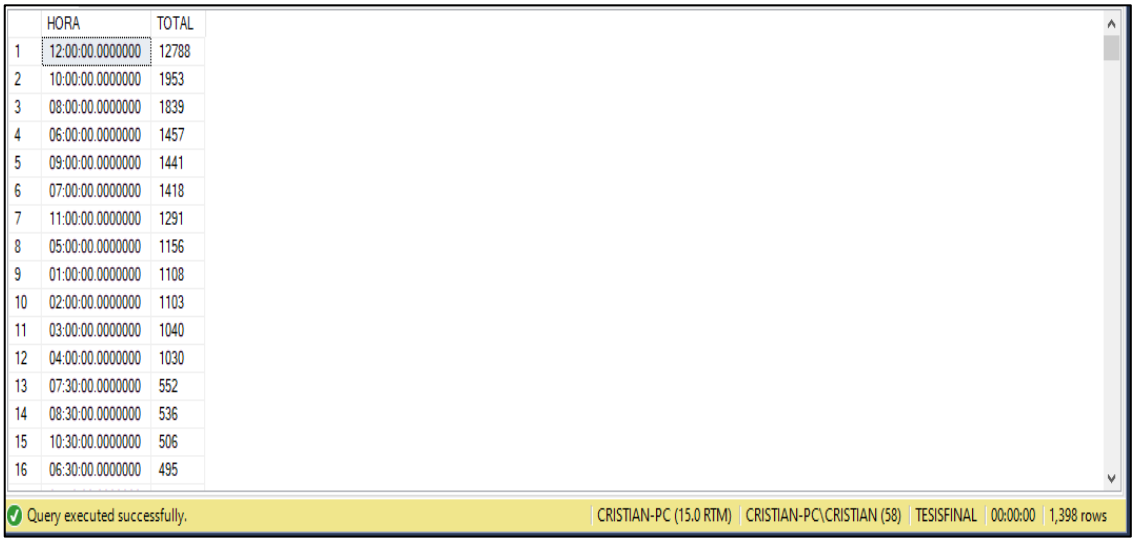


Figura 20. Pantalla de la cantidad de desaparecidos por año mediante gráfica lineal

- **Hora en que se le vio por última vez**

  En la figura 21 se realizará la consulta para visualizar la hora para ver los patrones de comportamiento

```
SELECT HORA, COUNT (*) AS TOTAL FROM DATOS
GROUP BY HORA
ORDER BY TOTAL DESC
```



	HORA	TOTAL
1	12:00:00.0000000	12788
2	10:00:00.0000000	1953
3	08:00:00.0000000	1839
4	06:00:00.0000000	1457
5	09:00:00.0000000	1441
6	07:00:00.0000000	1418
7	11:00:00.0000000	1291
8	05:00:00.0000000	1156
9	01:00:00.0000000	1108
10	02:00:00.0000000	1103
11	03:00:00.0000000	1040
12	04:00:00.0000000	1030
13	07:30:00.0000000	552
14	08:30:00.0000000	536
15	10:30:00.0000000	506
16	06:30:00.0000000	495

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (58) | TESISFINAL | 00:00:00 | 1,398 rows

Figura 21. Pantalla de agrupación del campo horas dentro de Microsoft SQL Server

En la Figura 22 se muestra un cuadro de gráfico que permitirá la visualización más detallada de la columna de personas desaparecidas por hora.

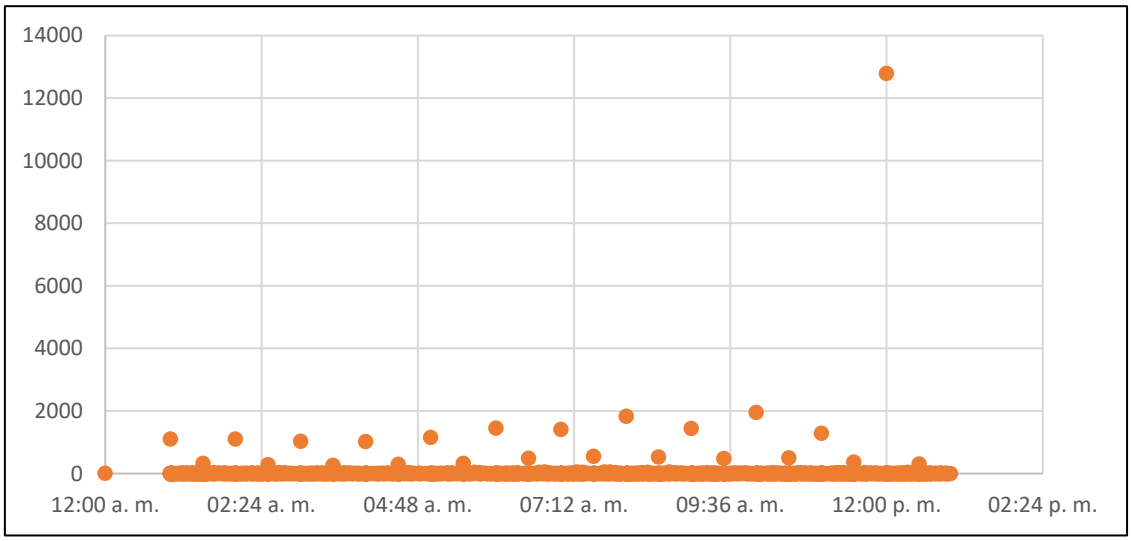


Figura 22. Pantalla de agrupación de las horas de los casos de desaparecidos mediante gráfica de dispersión

- **País en que se le vio por última vez**

  En la tabla 13 se describe la totalidad de país donde se vio al a persona por última vez en este caso todos son de México.

```
SELECT DISTINCT PAIS,COUNT(PAIS) AS TOTAL FROM DATOS
GROUP BY PAIS;
```

Tabla 13 Conteo total del campo país

PAÍS	TOTAL
MÉXICO	36265

- **Entidad en que se le vio por última vez**

  En la tabla 14 se describe la cantidad de estados del país de México con la totalidad de casos de personas desaparecidas.

```
SELECT DISTINCT ENTIDAD,COUNT(ENTIDAD) AS TOTAL FROM
DATOS
GROUP BY ENTIDAD
ORDER BY COUNT(ENTIDAD) DESC;
```

Tabla 14 *Conteo del total de los estados por la cantidad de desaparecidos*

ENTIDAD	TOTAL	ENTIDAD	TOTAL
TAMAULIPAS	5,990	DURANGO	420
ESTADO DE MÉXICO	3,890	QUERETARO	284
JALISCO	3,362	MORELOS	241
SINALOA	3,027	AGUASCALIENTES	223
NUEVO LEON	2,895	OAXACA	191
CHIHUAHUA	2,186	HIDALGO	173
SONORA	2,150	NAYARIT	145
PUEBLA	2,069	CHIAPAS	108
COAHUILA DE ZARAGOZA	1,753	YUCATAN	99
GUERRERO	1,482	SAN LUIS POTOSI	97
MICHOACAN	1,215	TABASCO	67
BAJA CALIFORNIA	1,024	QUINTANA ROO	61
CIUDAD DE MÉXICO	744	BAJA CALIFORNIA SUR	39
GUANAJUATO	615	CAMPECHE	35

ENTIDAD	TOTAL
COLIMA	593
VERACRUZ	524
ZACATECAS	510

ENTIDAD	TOTAL
NO ESPECIFICADO	29
TLAXCALA	24

En la figura 23 se muestra la gráfica de los estados en general donde se puede vizualizar que el Estado de taumlipas es el que tiene mas casos.

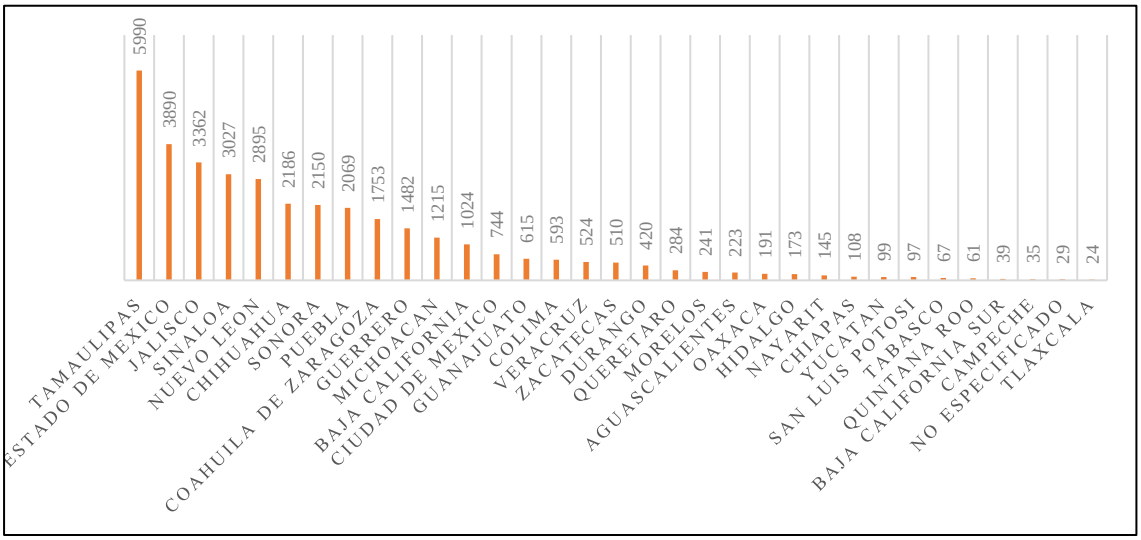
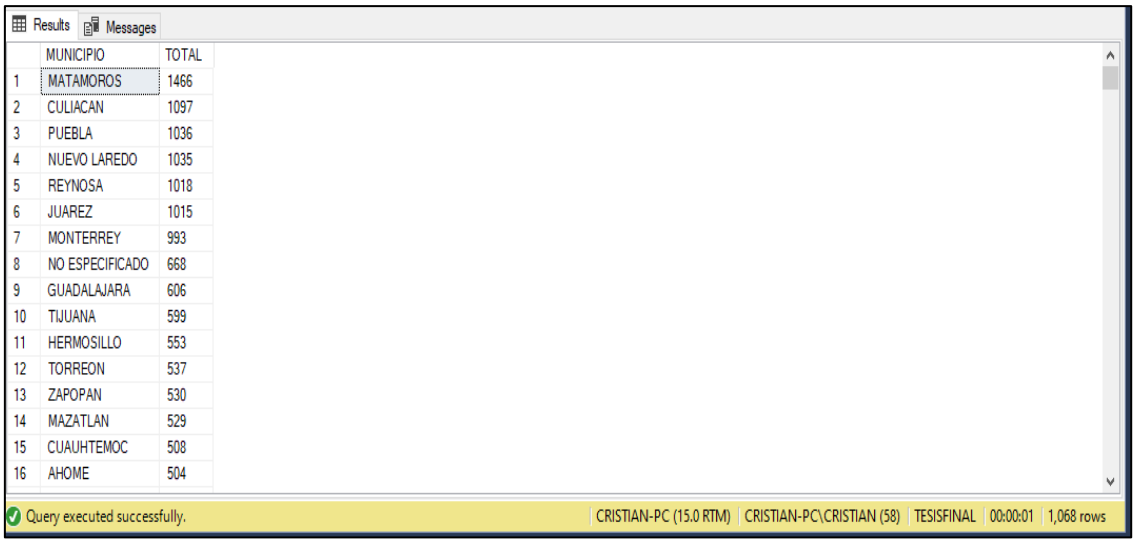


Figura 23. Pantalla de la cantidad de desaparecidos en los estados mediante gráfica de eje horizontal

- **Municipio en que se le vio por última vez**

  En la figura 24 se muestra la cantidad de municipios del país de México con la totalidad de casos de personas desaparecidas.

```
SELECT DISTINCT MUNICIPIO,COUNT(MUNICIPIO) AS TOTAL FROM
DATOS
GROUP BY MUNICIPIO
ORDER BY COUNT(MUNICIPIO) DESC;
```



Results Messages

	MUNICIPIO	TOTAL
1	MATAMOROS	1466
2	CULIACAN	1097
3	PUEBLA	1036
4	NUEVO LAREDO	1035
5	REYNOSA	1018
6	JUAREZ	1015
7	MONTERREY	993
8	NO ESPECIFICADO	668
9	GUADALAJARA	606
10	TIJUANA	599
11	HERMOSILLO	553
12	TORREON	537
13	ZAPOPAN	530
14	MAZATLAN	529
15	CUAUHTEMOC	508
16	AHOME	504

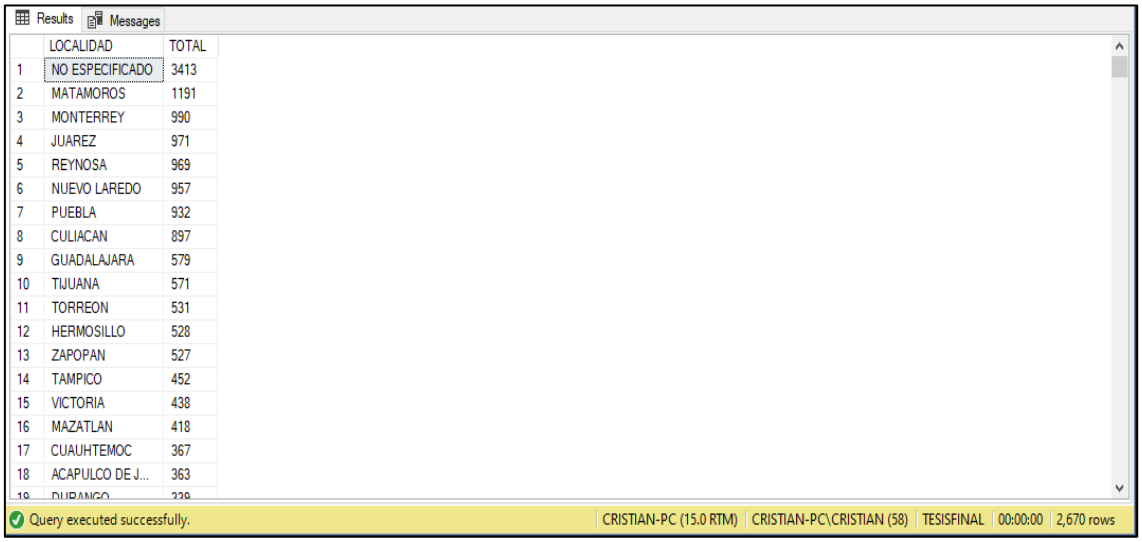
Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (58) | TESISFINAL | 00:00:01 | 1,068 rows

Figura 24. Pantalla del conteo total del campo de municipio dentro de Microsoft SQL Server

- **Localidad en que se le vio por última vez**

  En la figura 25 se muestra la cantidad de localidades del país de México con la totalidad de casos de personas desaparecidas.

```
SELECT DISTINCT LOCALIDAD,COUNT(LOCALIDAD) AS TOTAL FROM
DATOS
GROUP BY LOCALIDAD
ORDER BY COUNT(LOCALIDAD) DESC;
```



Results Messages

	LOCALIDAD	TOTAL
1	NO ESPECIFICADO	3413
2	MATAMOROS	1191
3	MONTERREY	990
4	JUAREZ	971
5	REYNOSA	969
6	NUEVO LAREDO	957
7	PUEBLA	932
8	CULIACAN	897
9	GUADALAJARA	579
10	TIJUANA	571
11	TORREON	531
12	HERMOSILLO	528
13	ZAPOPAN	527
14	TAMPICO	452
15	VICTORIA	438
16	MAZATLAN	418
17	CUAUHTEMOC	367
18	ACAPULCO DE J...	363
19	DURANGO	339

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (58) | TESISFINAL | 00:00:00 | 2,670 rows

Figura 25. Pantalla de conteo total del campo de localidad dentro de Microsoft SQL Server

- **Nacionalidad**

  En la tabla 15 se describen las diversas nacionalidades de los casos de personas desaparecidas.

```
SELECT DISTINCT NACIONALIDAD,COUNT(NACIONALIDAD) AS
TOTAL FROM DATOS
GROUP BY NACIONALIDAD
ORDER BY COUNT(NACIONALIDAD) DESC;
```

Tabla 15 *Conteo total del campo de nacionalidades*

NACIONALIDAD	TOTAL	NACIONALIDAD	TOTAL
MEXICANA	34,016	HAITIANA	2
NO ESPECIFICADO	2,040	FRANCESA	1
ESTADOUNIDENSE	124	DOMINICANA	1
HONDUREÑA	24	COSTARICENSE	1
GUATEMALTECO	12	CHINA	1
COLOMBIANA	9	ARGENTINA	1
SALVADOREÑA	8	HOLANDESA	1
VENEZOLANA	4	NICARAGUENSE	1
CUBANA	3	BRASILEÑA	1
ITALIANA	3	PARAGUAYA	1
CANADIENSE	3	PANAMEÑA	1
ESPAÑOLA	3	ERITREA	1
PERUANA	2	ALEMANA	1

En la figura 26 se muestra la gráfica de las nacionalidades de elas personas de casos de personas desaparecidas de las cuales la nacionalidad mexicana tiene el mayor porcentaje.

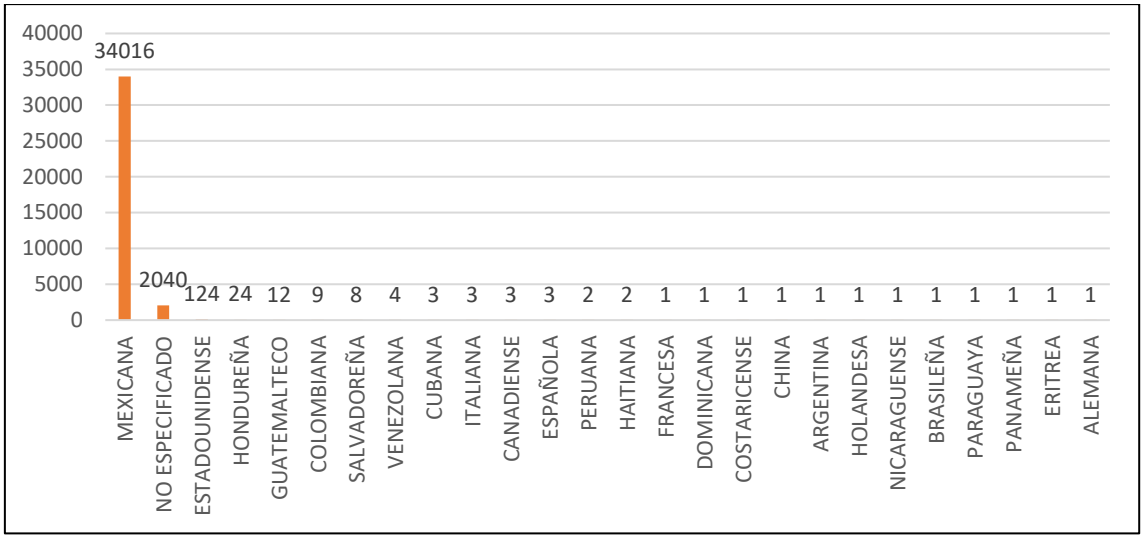


Figura 26. Pantalla del total de cantidad de nacionalidades mediante gráfica de barras

- **Estatura**

  En la figura 27 se muestra los patrones de comportamiento de la columna estatura con la cantidad.

```
SELECT DISTINCT ESTATURA,COUNT(ESTATURA)
AS TOTAL_ESTATURA FROM DATOS
GROUP BY ESTATURA
ORDER BY COUNT(ESTATURA) DESC;
```

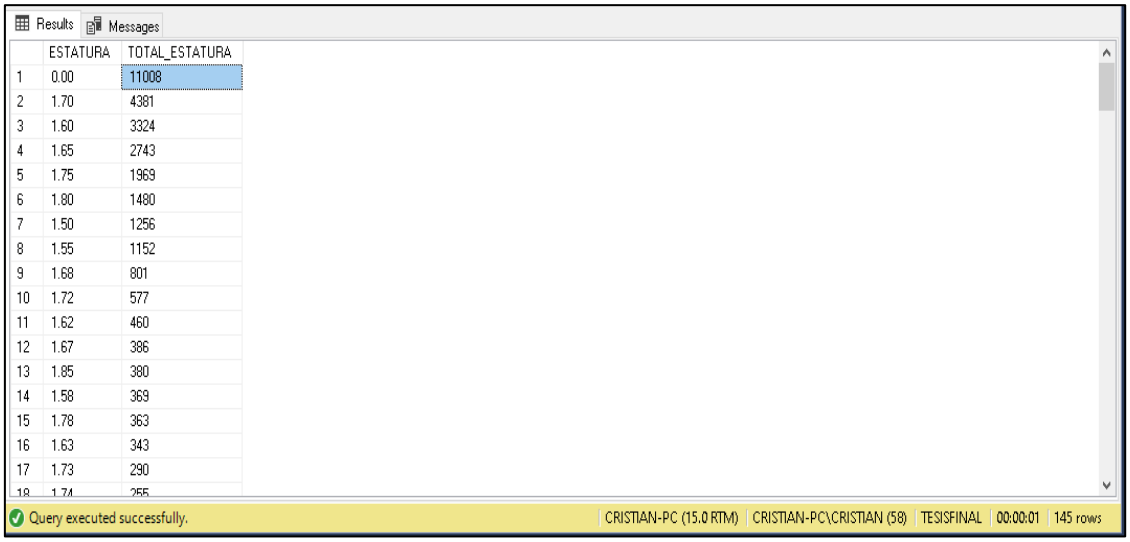


Figura 27. Pantalla de la cantidad total del campo estatura dentro de Microsoft SQL Server

- **Complexión**

  En la tabla 16 se describen los tipos de complexiones con la cantidad en total de cada una de ellas.

```
SELECT DISTINCT COMPLEXION,COUNT(COMPLEXION) AS TOTAL
FROM DATOS
GROUP BY COMPLEXION
ORDER BY COUNT(COMPLEXION) DESC;
```

Tabla 16 *Conteo del total del campo complexión*

COMPLEXION	TOTAL
DELGADA	13,349
NORMAL	11,295
ROBUSTA	6,208
MEDIANA	5,178
OBESA	235

En la figura 28 se muestra la gráfica de los tipos de complexión donde se tiene que la predominante es la delgada.

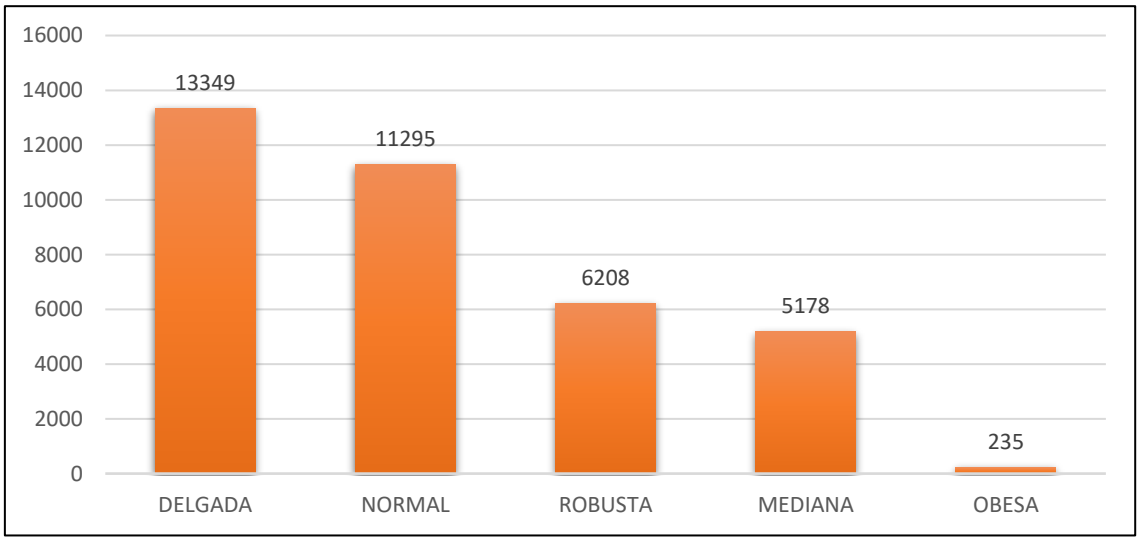


Figura 28. Pantalla de los tipos de complexiones mediante una gráfica de barras

- **Sexo**

  En la tabla 17 se describen los patrones de comportamiento entre el hombre y la mujer y su totalidad del mismo.

```
SELECT DISTINCT SEXO,COUNT(SEXO) AS TOTAL FROM DATOS
GROUP BY SEXO
ORDER BY COUNT(SEXO) DESC;
```

Tabla 17 Conteo del total del campo sexo

SEXO	TOTAL
HOMBRE	26,938
MUJER	9,327

En la figura 29 se muestra la gráfica donde se visualiza que el valor predominante es el del varón el cual tiene un mayor índice de casos de personas desaparecidas.

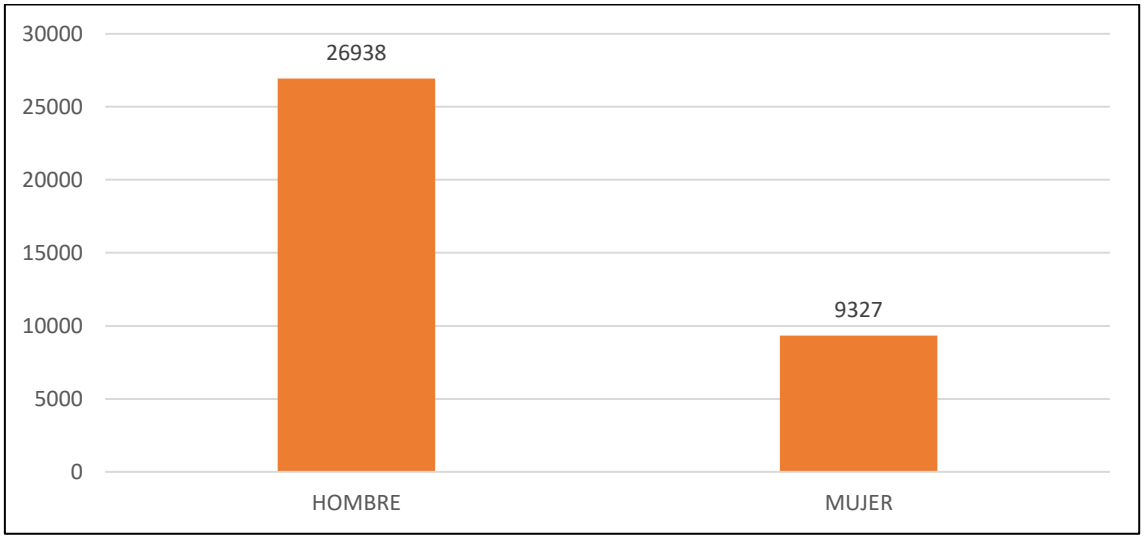
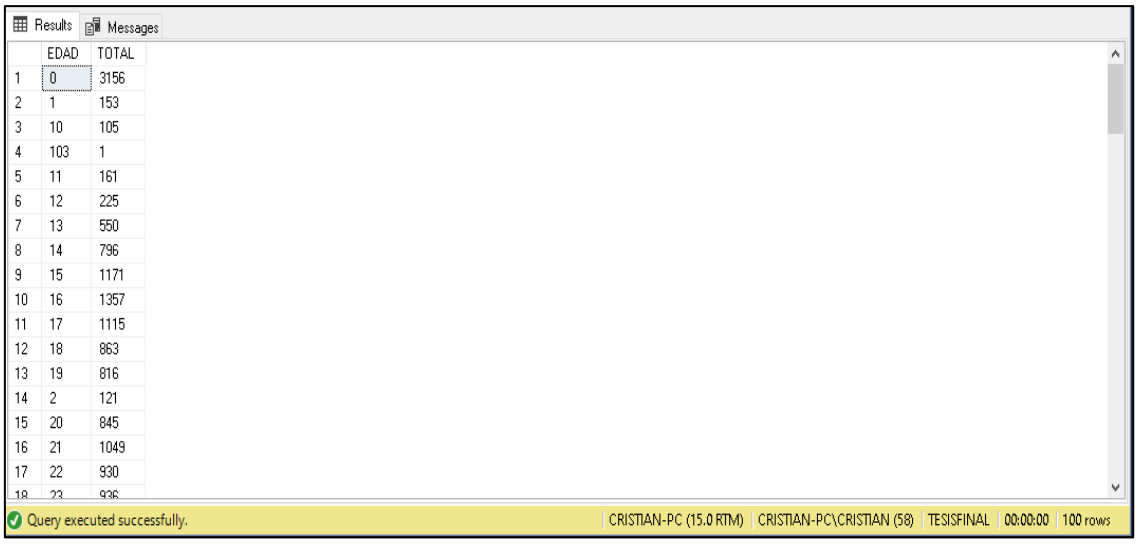


Figura 29. Pantalla de cantidad de datos del campo sexo mediante gráfica de barras

- **Edad**

  En la figura 30 se muestra la gráfica del total de patrones de comportamiento de la tabla edad por su total en cantidades.

```
SELECT DISTINCT EDAD,COUNT(EDAD) AS TOTAL FROM DATOS
GROUP BY EDAD
ORDER BY EDAD ASC;
```



Results Messages

	EDAD	TOTAL
1	0	3156
2	1	153
3	10	105
4	103	1
5	11	161
6	12	225
7	13	550
8	14	796
9	15	1171
10	16	1357
11	17	1115
12	18	863
13	19	816
14	2	121
15	20	845
16	21	1049
17	22	930
18	23	936

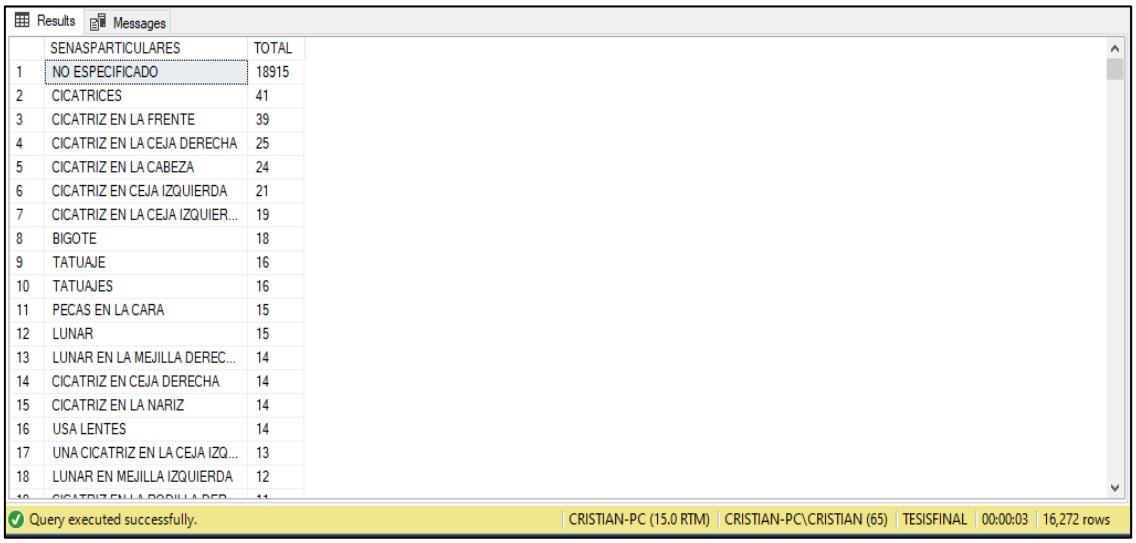
Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (58) | TESISFINAL | 00:00:00 | 100 rows

Figura 30. Pantalla del conteo total del campo de edad mediante Microsoft SQL Server

- **Descripción de señas particulares**

  En la figura 31 se muestra la agrupación de los patrones de comportamiento de las señas particulares.

```
SELECT DISTINCT
SENASPARTICULARES,COUNT(SENASPARTICULARES) AS TOTAL
FROM DATOS
GROUP BY SENASPARTICULARES
ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results Messages

	SENASPARTICULARES	TOTAL
1	NO ESPECIFICADO	18915
2	CICATRICES	41
3	CICATRIZ EN LA FRENTE	39
4	CICATRIZ EN LA CEJA DERECHA	25
5	CICATRIZ EN LA CABEZA	24
6	CICATRIZ EN CEJA IZQUIERDA	21
7	CICATRIZ EN LA CEJA IZQUIER...	19
8	BIGOTE	18
9	TATUAJE	16
10	TATUAJES	16
11	PECAS EN LA CARA	15
12	LUNAR	15
13	LUNAR EN LA MEJILLA DEREC...	14
14	CICATRIZ EN CEJA DERECHA	14
15	CICATRIZ EN LA NARIZ	14
16	USA LENTES	14
17	UNA CICATRIZ EN LA CEJA IZQ...	13
18	LUNAR EN MEJILLA IZQUIERDA	12

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (65) | TESISFINAL | 00:00:03 | 16,272 rows

Figura 31. Pantalla de conteo total del campo de señas particulares mediante Microsoft SQL Server

- **Etnia**

  En la tabla 18 se describen los diferentes tipos de etnias encontrados con su total de las mismas para un estudio más específico.

```
SELECT DISTINCT ETNIA,COUNT(ETNIA) AS TOTAL FROM DATOS
GROUP BY ETNIA
ORDER BY COUNT(ETNIA) DESC;
```

Tabla 18 *Conteo de las diferentes etnias*

ETNIA	**TOTAL**
NO ESPECIFICADO	36,133
MAYAS	27
NAHUAS	18
AMUZGOS OAXACA	13
CHATINOS	11
TLALPANECOS	11
OTOMIES	10
TARAHUMARAS - RARAMURI	9
CHICHIMECAS	6
CHINANTECOS	5
SERIS	4
MIXTECOS	4
MEXICANEROS	3
TOTONACAS	2
CHOLES	2
YAQUIS	2
HUICHOLES	2
PAMES	1
MAMES	1
HUASTECOS	1

En la figura 32 se muestran los diferentes tipos de etnias y cuál es el predominante para el estudio.

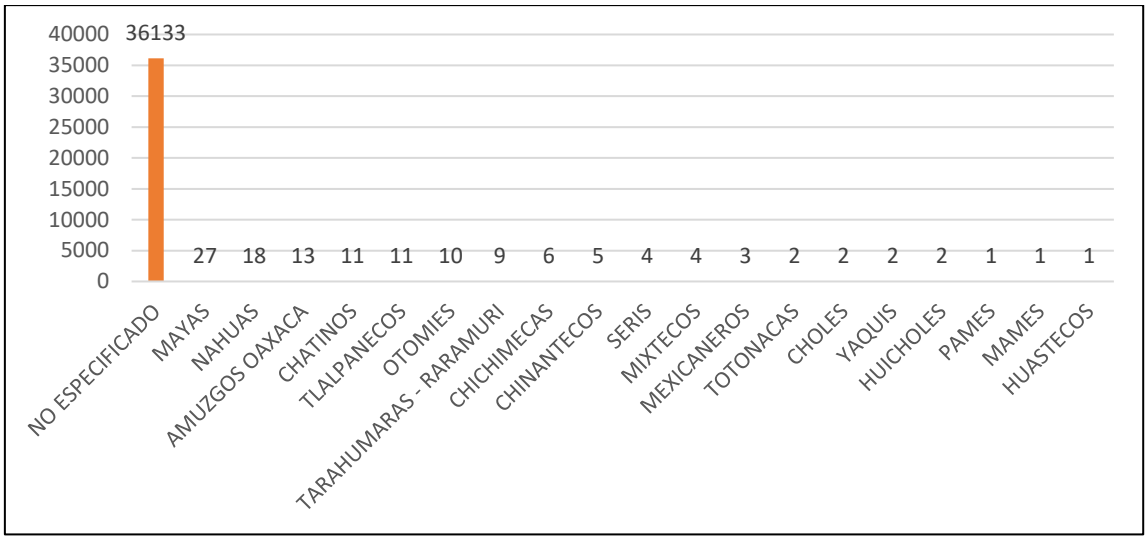


Figura 32. Pantalla de conteo total del campo de etnias mediante gráfica de barras

- **Discapacidad**

```
SELECT DISTINCT DISCAPACIDAD,COUNT(DISCAPACIDAD) AS
TOTAL FROM DATOS
GROUP BY DISCAPACIDAD
ORDER BY COUNT(DISCAPACIDAD) DESC;
```

En la tabla 19 se describe los diferentes tipos de discapacidades encontrados en los datos en su totalidad.

Tabla 19 *Conteo de los tipos de discapacidades*

DISCAPACIDAD	TOTAL
SIN DISCAPACIDAD (Ninguno)	35,181
NO ESPECIFICADO	1,069
SINDROME DE DOWN	8
AUTISMO	7

En la figura 33 se muestra los diferentes tipos de discapacidades y cuál es el valor dominante.

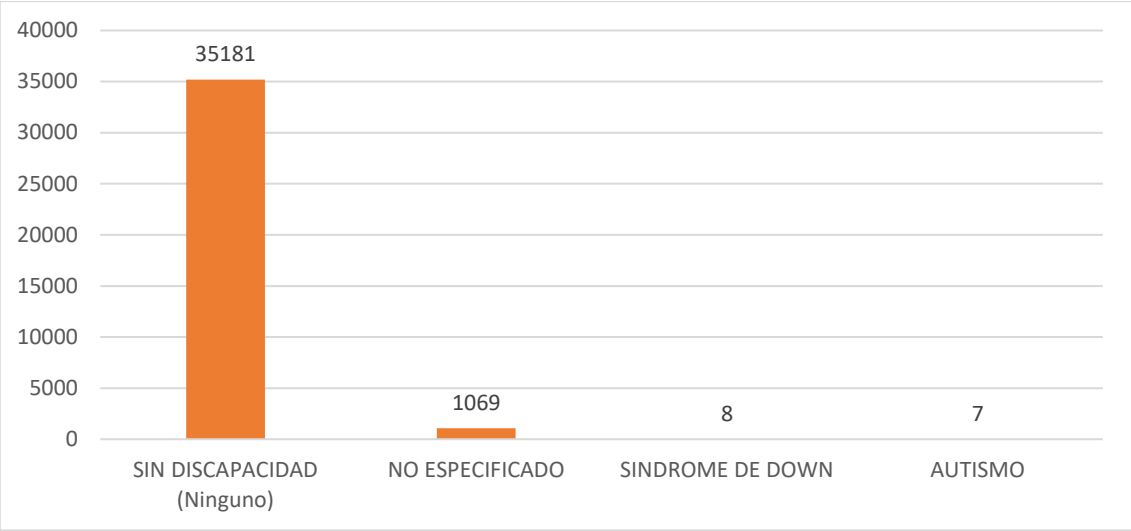


Figura 33. Conteo total del campo de discapacidad

- **Dependencia**

```
SELECT DISTINCT DEPENDENCIA,COUNT(DEPENDENCIA) AS TOTAL
FROM DATOS
GROUP BY DEPENDENCIA
ORDER BY COUNT(DEPENDENCIA) DESC;
```

En la tabla 20 se describen los diferentes tipos de dependencias encontradas.

Tabla 20 *Conteo de dependencias del país de México*

DEPENDENCIA	TOTAL	DEPENDENCIA	TOTAL
PGJ - TAMAULIPAS	5,978	FGE - DURANGO	406
FGJ - ESTADO DE MÉXICO	3,874	FGE - QUERETARO	361
FGE - JALISCO	3,377	FGE - MORELOS	240
FGE - SINALOA	3,025	FGE - AGUASCALIENTES	221
PGJ - NUEVO LEÓN	2,884	FGE - OAXACA	202
FGE - CHIHUAHUA	2,201	PGJ - HIDALGO	175
FGE - SONORA	2,146	FGE - NAYARIT	144
FGE - PUEBLA	2,078	FGE - CHIAPAS	107
FGE - COAHUILA	1,778	FGE - YUCATAN	98
FGE - GUERRERO	1,502	FGE - TABASCO	64
PGJ - MICHOACAN	1,203	FGE - QUINTANA ROO	60

DEPENDENCIA	TOTAL
PGJ - BAJA CALIFORNIA	1,012
PGJ - CIUDAD DE MÉXICO	742
PGJ - GUANAJUATO	615
PGJ - COLIMA	589
FGE - VERACRUZ	514
FGE - ZACATECAS	510

DEPENDENCIA	TOTAL
FGE - SAN LUIS POTOSI	53
PGJ - BAJA CALIFORNIA SUR	38
FGE - CAMPECHE	34
PGJ - TLAXCALA	21
NO ESPECIFICADO	2

En la figura 34 se muestran las dependencias y cuál es el que tuvo mayores casos de personas desaparecidas.

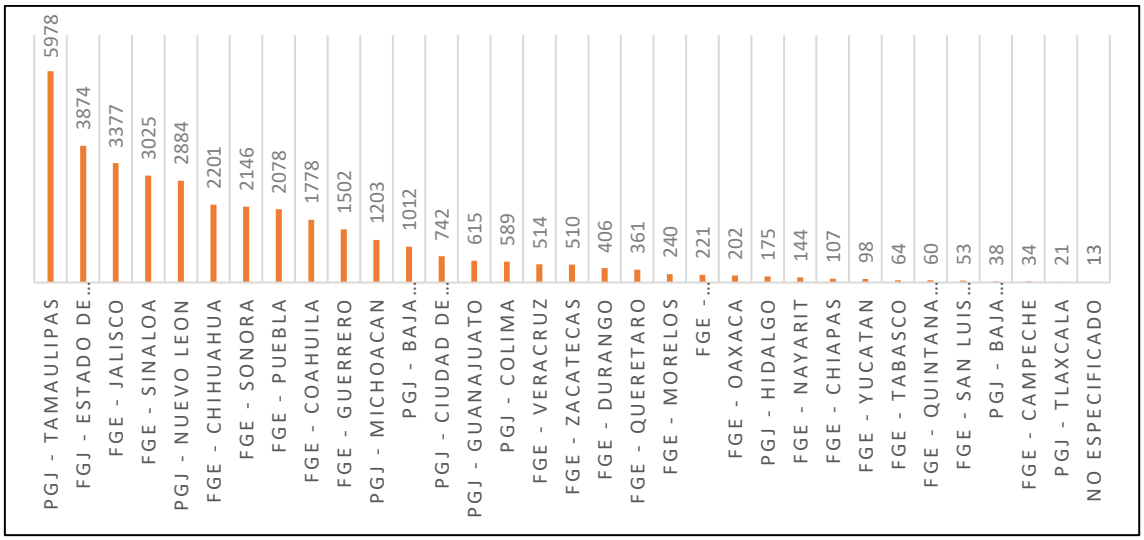


Figura 34. Pantalla de conteo de dependencias del país de México mediante una gráfica de barras

IV.3 Fase 3: Preparación de los datos

IV.3.1 Limpieza de datos

Para la depuración de datos se creará una nueva tabla para así no manipular o afectar a los datos originales.

```
CREATE TABLE DEPURACION(
IDDEPURACION INT IDENTITY(1,1) NOT NULL,
IDDESAPARECIDOS INT NOT NULL,
FECHA DATE NOT NULL,
HORA TIME NOT NULL,
PAIS VARCHAR(50),
```

```
ESTADO VARCHAR(100),
MUNICIPIO VARCHAR(100),
LOCALIDAD VARCHAR(100),
NACIONALIDAD VARCHAR(100),
COMPLEXION VARCHAR(100),
SEXO VARCHAR(50),
EDAD CHAR(3),
ESTATURA DECIMAL(4,2),
ETNIA VARCHAR(100),
SENASPARTICULARES VARCHAR(4000),
DEPENDENCIA VARCHAR(100),
PRIMARY KEY(IDDEPURACION)
);
```

- **HORA:** Se elaboró una consulta el cual se muestra el conglomerado de datos, asimismo arrojó que los datos son inconsistentes y no será beneficioso tomarlo en cuenta en el proyecto por esto se excluyó este campo.

```
SELECT HORA, COUNT (*) AS TOTAL
   FROM DATOS
GROUP BY HORA
ORDER BY TOTAL DESC
```

Tabla 21 *Depuración del campo horas*

HORA	TOTAL
12:00 p. m.	12,788
10:00 a. m.	1,953
08:00 a. m.	1,839
06:00 a. m.	1,457
09:00 a. m.	1,441
07:00 a. m.	1,418
11:00 a. m.	1,291
05:00 a. m.	1,156
01:00 a. m.	1,108
02:00 a. m.	1,103
03:00 a. m.	1,040
04:00 a. m.	1,030
07:30 a. m.	552
08:30 a. m.	536
10:30 a. m.	506
06:30 a. m.	495
09:30 a. m.	479
11:30 a. m.	370
05:30 a. m.	331

Estos datos se han depurado bajo el concepto de que solo existen registros de 12 horas, entre las 00:00:00 horas hasta las 12:00:00 horas por las cual hay pérdida de datos.

- **ESTATURA:** Se elaboró una consulta en el cual se tiene el conglomerado de datos, arrojando que es inconsistente y no será beneficioso tomarlo en cuenta en el proyecto por esto se excluyó este campo.

```
SELECT EDAD, ESTATURA, COUNT (*) AS TOTAL_ESTATURA
FROM DATOS
GROUP BY EDAD, ESTATURA
ORDER BY TOTAL_ESTATURA DESC
```

Tabla 22 *Depuración del campo estatura*

Edad	Estatura	Total_estatura
73	0	15
55	0	57
83	0	6
12	0	76
54	0	54
4	0	32
35	0	216
7	0	21
5	0	32
26	0	324
18	0	271
27	0	281
97	0	2
8	0	35
0	0	936
13	0	179
91	0	2
49	0	81
40	0	181
21	0	333

Edad	Estatura	Total_estatura
17	1	6
47	1	3
44	1	1
10	1	1
28	1	4
30	1	2
39	1	4
63	1	1
16	1	4
25	1	1
46	1	1
38	1	3
24	1	4
81	1	1
11	1	2
14	1	2
33	1	5
50	1	1
35	1	3
29	1	3

Según los especialistas de El Universal mx (2020) indicaron: “Que el mexicano promedio mide 1.64, mientras que las mujeres 1.58 de altura, añadieron que el promedio de los jóvenes es de 1.61 en mujeres y 1.67 en varones” (π. 2). Por la base teórica expuesta se decidió no contar con el campo de estatura puesto que abarcaba muchas inconsistencias.

Por lo que casi todo el documento no hay relación clara entre la edad y la estatura llevando a la conclusión que los datos en el campo de estatura no están bien planteados.

- **ETNIA:** El campo no cuenta con datos que aporten algo al estudio porque de los 36,265 datos, 132 son los que solo tienen alguna etnia y los otros 36,133 son sin especificar.

```
SELECT ETNIA, count (*) as TOTAL
   from DESAPARECIDOS
GROUP BY ETNIA
ORDER BY TOTAL DESC
```

Tabla 23 *Depuración del campo etnia*

ETNIA	**TOTAL**
NO ESPECIFICADO	36,133
MAYAS	27
NAHUAS	18
AMUZGOS OAXACA	13
CHATINOS	11
TLALPANECOS	11
OTOMIES	10
TARAHUMARAS - RARAMURI	9
CHICHIMECAS	6
CHINANTECOS	5
SERIS	4
MIXTECOS	4
MEXICANEROS	3
TOTONACAS	2
CHOLES	2
YAQUIS	2
HUICHOLES	2
PAMES	1
MAMES	1
HUASTECOS	1

En la figura 35 se muestran las etnias y cuál es la característica más dominante en los casos de personas desaparecidas.

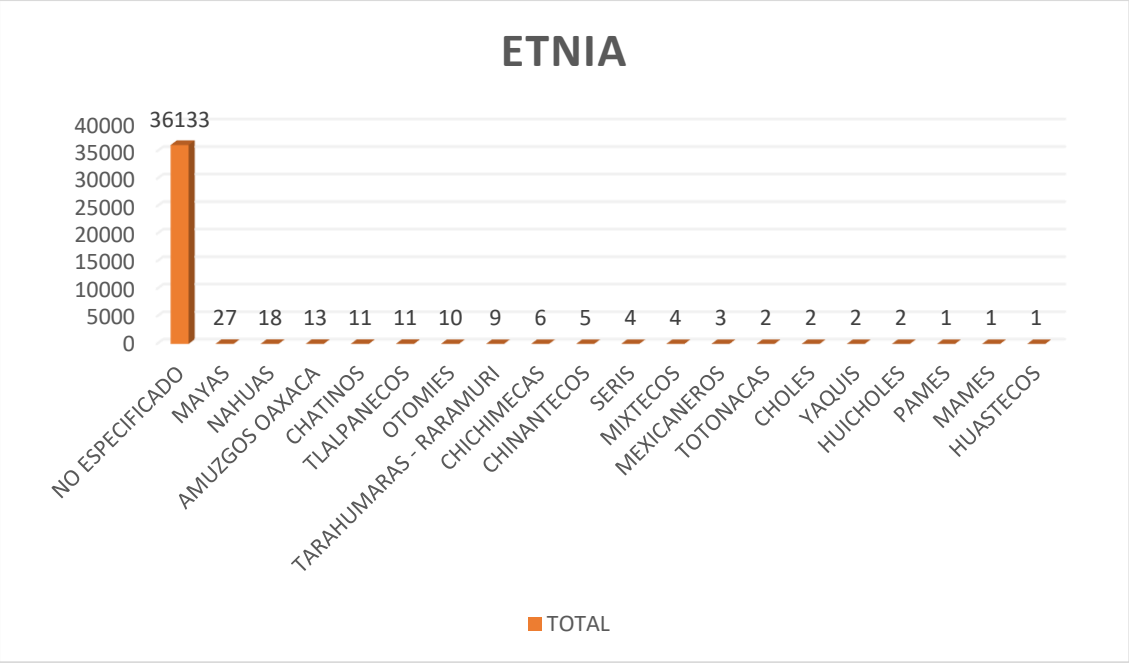


Figura 35. Pantalla de conteo de etnias del país de México mediante una gráfica de barras

En conclusión, los datos que se están depurando (HORA, ESTATURA Y ETNIA) son a coste de que salga un mejor estudio y estadísticas que ayuden al pronóstico y las técnicas que se han aplicado.

Se realiza un select (selección de los campos depurados) de la tabla DATOS.

```
SELECT HORA, ESTATURA, ETNIA FROM DATOS;
```

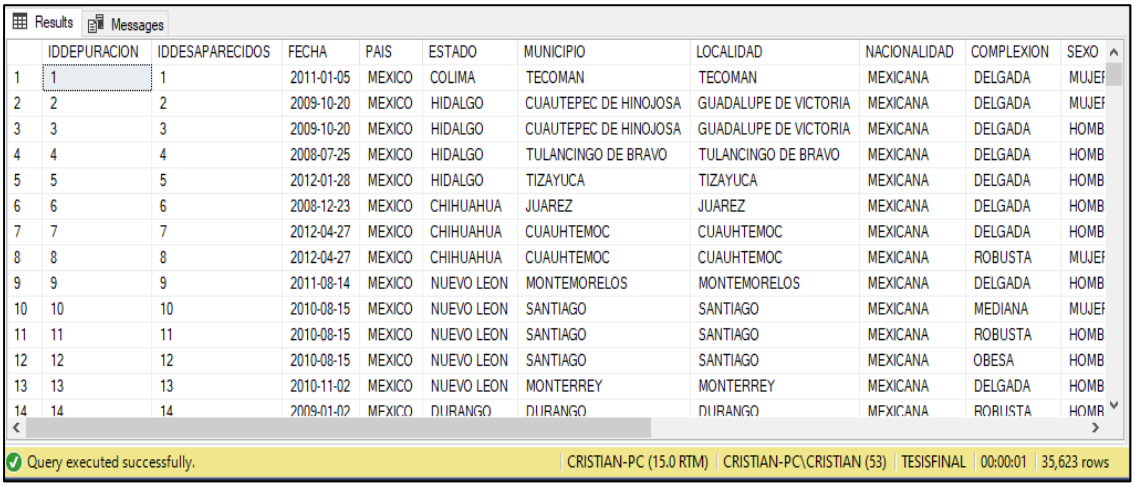
	HORA	ESTATURA	ETNIA
1	12:00:00.0000000	1.65	NO ESPECIFICADO
2	11:30:00.0000000	0.00	NO ESPECIFICADO
3	12:00:00.0000000	0.00	NO ESPECIFICADO
4	12:00:00.0000000	0.00	NO ESPECIFICADO
5	01:00:00.0000000	1.70	NO ESPECIFICADO
6	12:00:00.0000000	1.75	NO ESPECIFICADO
7	12:00:55.0000000	1.63	NO ESPECIFICADO
8	12:00:00.0000000	1.60	NO ESPECIFICADO
9	08:45:00.0000000	1.70	NO ESPECIFICADO
10	02:15:00.0000000	1.80	NO ESPECIFICADO
11	12:00:00.0000000	0.00	NO ESPECIFICADO
12	02:00:00.0000000	1.70	NO ESPECIFICADO
13	07:00:00.0000000	1.70	NO ESPECIFICADO
14	07:30:00.0000000	1.65	NO ESPECIFICADO
15	12:00:00.0000000	1.55	NO ESPECIFICADO
16	08:30:00.0000000	1.70	NO ESPECIFICADO
17	12:00:00.0000000	1.60	NO ESPECIFICADO
18	03:00:00.0000000	1.60	NO ESPECIFICADO
19	10:30:00.0000000	1.65	NO ESPECIFICADO
20	12:00:00.0000000	1.70	NO ESPECIFICADO
21	12:00:00.0000000	0.00	NO ESPECIFICADO
22	12:00:00.0000000	1.75	NO ESPECIFICADO
23	12:00:00.0000000	0.00	NO ESPECIFICADO
24	06:00:00.0000000	1.68	NO ESPECIFICADO

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (51) | TESISFINAL | 00:00:01 | 36,265 rows

Figura 36. Pantalla de los datos depurados (A) dentro de Microsoft SQL Server

- **Fecha y estado:** Después de analizar los datos se quitará los años menores al 2007 y los valores no especificados en el Estado.

```
INSERT DEPURACION
SELECT
   IDDESAPARECIDOS,FECHA,PAIS,ENTIDAD,MUNICIPIO,
   LOCAL
DAD,NACIONALIDAD,COMPLEXION,SEXO,EDAD,SENASPARTICULARE
S,DEPENDENCIA
FROM DATOS WHERE YEAR(FECHA) >=2007 AND ENTIDAD!='NO
ESPECIFICADO';
```



Results | Messages

	IDDEPURACION	IDDESAPARECIDOS	FECHA	PAIS	ESTADO	MUNICIPIO	LOCALIDAD	NACIONALIDAD	COMPLEXION	SEXO
1	1	1	2011-01-05	MEXICO	COLIMA	TECOMAN	TECOMAN	MEXICANA	DELGADA	MUJEI
2	2	2	2009-10-20	MEXICO	HIDALGO	CUAUTEPEC DE HINOJOSA	GUADALUPE DE VICTORIA	MEXICANA	DELGADA	MUJEI
3	3	3	2009-10-20	MEXICO	HIDALGO	CUAUTEPEC DE HINOJOSA	GUADALUPE DE VICTORIA	MEXICANA	DELGADA	HOMB
4	4	4	2008-07-25	MEXICO	HIDALGO	TULANCINGO DE BRAVO	TULANCINGO DE BRAVO	MEXICANA	DELGADA	HOMB
5	5	5	2012-01-28	MEXICO	HIDALGO	TIZAYUCA	TIZAYUCA	MEXICANA	DELGADA	HOMB
6	6	6	2008-12-23	MEXICO	CHIHUAHUA	JUAREZ	JUAREZ	MEXICANA	DELGADA	HOMB
7	7	7	2012-04-27	MEXICO	CHIHUAHUA	CUAUHTEMOC	CUAUHTEMOC	MEXICANA	DELGADA	HOMB
8	8	8	2012-04-27	MEXICO	CHIHUAHUA	CUAUHTEMOC	CUAUHTEMOC	MEXICANA	ROBUSTA	MUJEI
9	9	9	2011-08-14	MEXICO	NUEVO LEON	MONTEMORELOS	MONTEMORELOS	MEXICANA	DELGADA	HOMB
10	10	10	2010-08-15	MEXICO	NUEVO LEON	SANTIAGO	SANTIAGO	MEXICANA	MEDIANA	MUJEI
11	11	11	2010-08-15	MEXICO	NUEVO LEON	SANTIAGO	SANTIAGO	MEXICANA	ROBUSTA	HOMB
12	12	12	2010-08-15	MEXICO	NUEVO LEON	SANTIAGO	SANTIAGO	MEXICANA	OBESA	HOMB
13	13	13	2010-11-02	MEXICO	NUEVO LEON	MONTERREY	MONTERREY	MEXICANA	DELGADA	HOMB
14	14	14	2009-01-02	MEXICO	DURANGO	DURANGO	DURANGO	MEXICANA	ROBUSTA	HOMB

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 35,623 rows

Figura 37. Pantalla de los datos depurados (B) dentro de Microsoft SQL Server

Al término de la ejecución de la consulta se quedó con el total 24,476 datos con los que se trabajaron para la regresión y los modelos.

```
SELECT SUM(TOTAL)AS [CANTIDAD TOTAL] FROM
TOTALDESAPARECIDOS;
```

Resultados | Mensajes

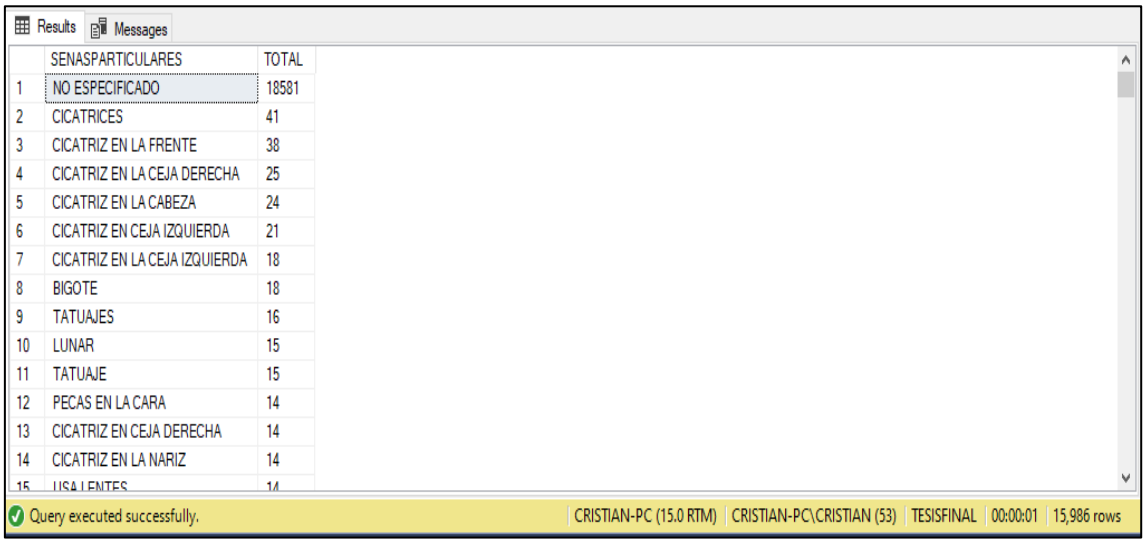
	CANTIDAD TOTAL
1	24476

Consulta ejecutada correctamente. | JHORDY-PC (15.0 RTM) | JHORDY-PC\JHORDY (64) | TESISFINAL | 00:00:00 | 1 filas

Figura 38 Pantalla de cantidad total de los datos depurados dentro de Microsoft SQL Server

- **Señas particulares:** Se analizó los patrones de comportamiento dentro de la columna de señas particulares encontrando en un inicio 15,986 valores.

```
SELECT DISTINCT
SENASPARTICULARES,COUNT(SENASPARTICULARES) AS     TOTAL
FROM DATOS
GROUP BY SENASPARTICULARES
ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

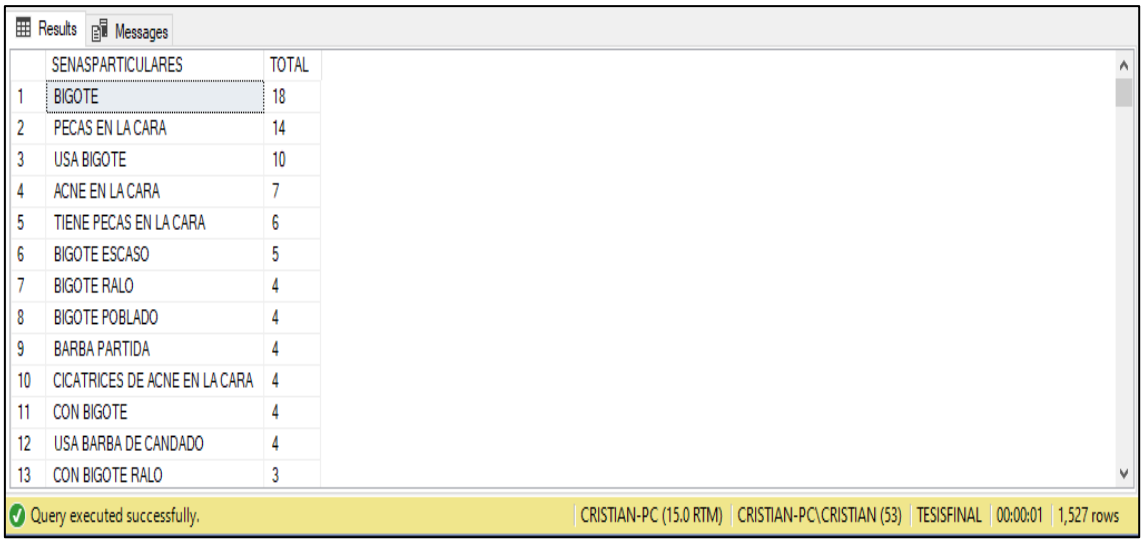
	SENASPARTICULARES	TOTAL
1	NO ESPECIFICADO	18581
2	CICATRICES	41
3	CICATRIZ EN LA FRENTE	38
4	CICATRIZ EN LA CEJA DERECHA	25
5	CICATRIZ EN LA CABEZA	24
6	CICATRIZ EN CEJA IZQUIERDA	21
7	CICATRIZ EN LA CEJA IZQUIERDA	18
8	BIGOTE	18
9	TATUAJES	16
10	LUNAR	15
11	TATUAJE	15
12	PECAS EN LA CARA	14
13	CICATRIZ EN CEJA DERECHA	14
14	CICATRIZ EN LA NARIZ	14
15	USA LENTES	14

Figura 39. Pantalla de agrupación del campo de señas particulares dentro de Microsoft SQL Server

Después del análisis de los datos se realizó la agrupación mediante ocho campos, estos mediante una consulta se agruparán simplificando el total de las señas particulares.

```
--Bigote, barba y cara

SELECT DISTINCT
SENASPARTICULARES,COUNT(SENASPARTICULARES)
AS TOTAL FROM DATOS
GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE
'%BIGOTE%' OR SENASPARTICULARES LIKE '%BARBA%' OR
SENASPARTICULARES LIKE '%CARA%'
ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

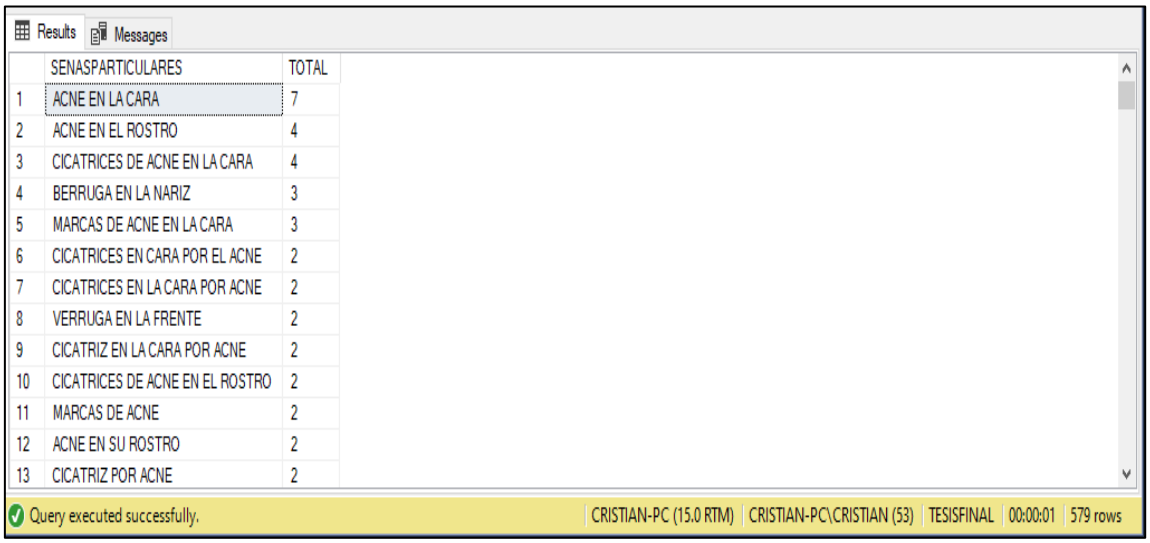
	SENASPARTICULARES	TOTAL
1	BIGOTE	18
2	PECAS EN LA CARA	14
3	USA BIGOTE	10
4	ACNE EN LA CARA	7
5	TIENE PECAS EN LA CARA	6
6	BIGOTE ESCASO	5
7	BIGOTE RALO	4
8	BIGOTE POBLADO	4
9	BARBA PARTIDA	4
10	CICATRICES DE ACNE EN LA CARA	4
11	CON BIGOTE	4
12	USA BARBA DE CANDADO	4
13	CON BIGOTE RALO	3

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 1,527 rows

Figura 40. Pantalla de agrupación por bigote, barba y cara usando una consulta mediante Microsoft SQL Server

```
--Acne y verruga

SELECT DISTINCT
SENASPARTICULARES,COUNT(SENASPARTICULARES)
AS TOTAL FROM DATOS
GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE
'%BERRUGA%' OR SENASPARTICULARES LIKE '%VERRUGA%' OR
SENASPARTICULARES LIKE '%ACNE%' ORDER BY
COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

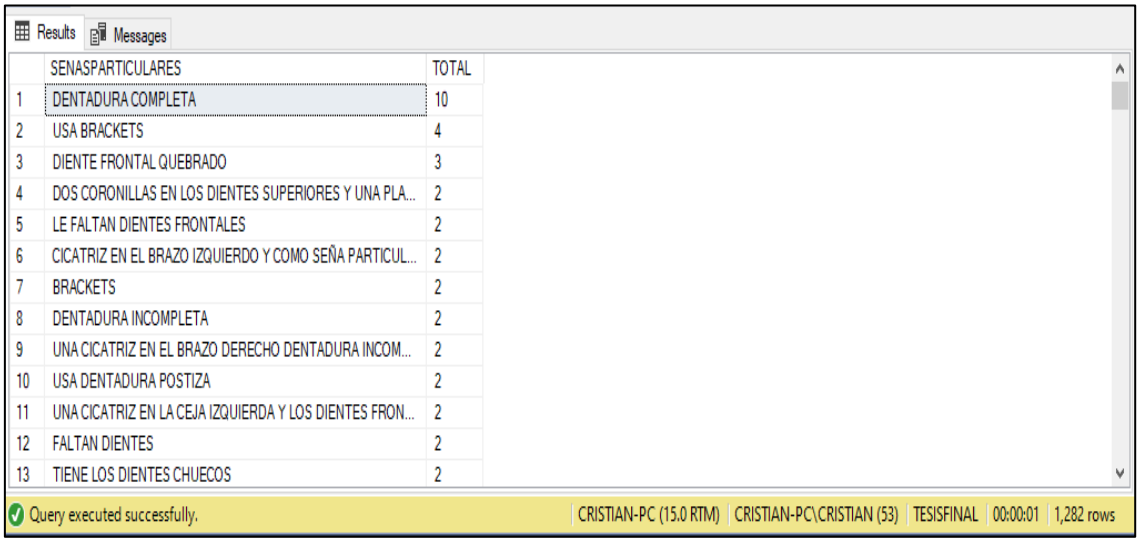
	SENASPARTICULARES	TOTAL
1	ACNE EN LA CARA	7
2	ACNE EN EL ROSTRO	4
3	CICATRICES DE ACNE EN LA CARA	4
4	BERRUGA EN LA NARIZ	3
5	MARCAS DE ACNE EN LA CARA	3
6	CICATRICES EN CARA POR EL ACNE	2
7	CICATRICES EN LA CARA POR ACNE	2
8	VERRUGA EN LA FRENTE	2
9	CICATRIZ EN LA CARA POR ACNE	2
10	CICATRICES DE ACNE EN EL ROSTRO	2
11	MARCAS DE ACNE	2
12	ACNE EN SU ROSTRO	2
13	CICATRIZ POR ACNE	2

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 579 rows

Figura 41. Pantalla de agrupación por acné y verruga, usando una consulta mediante Microsoft SQL Server

```
--Diente, muela y dentadura

SELECT DISTINCT SENASPARTICULARES, COUNT (SENASPARTICULARES) AS TOTAL FROM DATOS GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE '%DIENTE%' OR SENASPARTICULARES LIKE '%MUELA%' OR SENASPARTICULARES LIKE '%DENTADURA%' OR SENASPARTICULARES LIKE '%BRACKETS%' ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

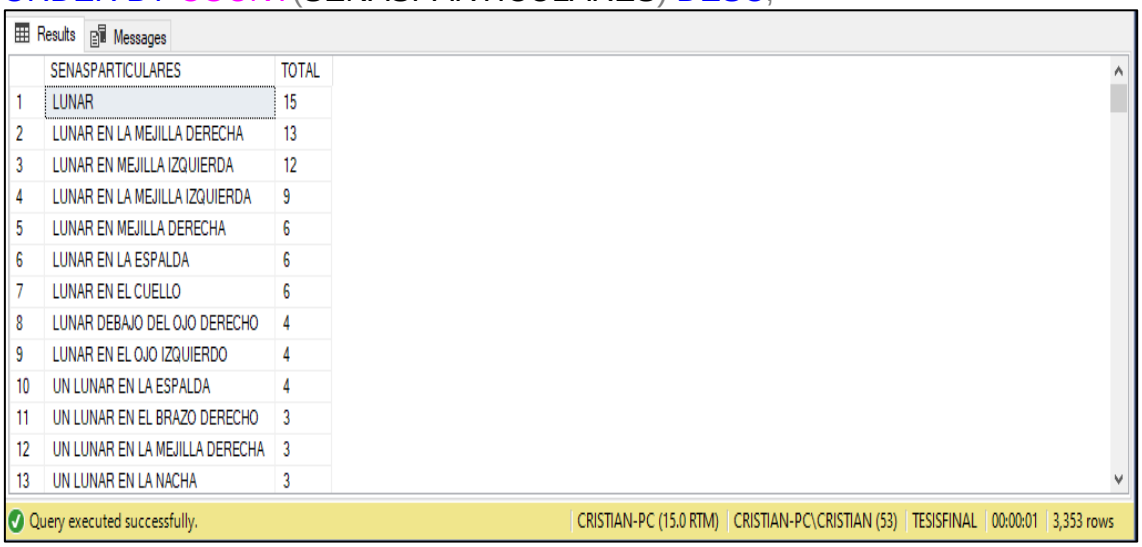
	SENASPARTICULARES	TOTAL
1	DENTADURA COMPLETA	10
2	USA BRACKETS	4
3	DIENTE FRONTAL QUEBRADO	3
4	DOS CORONILLAS EN LOS DIENTES SUPERIORES Y UNA PLA...	2
5	LE FALTAN DIENTES FRONTALES	2
6	CICATRIZ EN EL BRAZO IZQUIERDO Y COMO SEÑA PARTICUL...	2
7	BRACKETS	2
8	DENTADURA INCOMPLETA	2
9	UNA CICATRIZ EN EL BRAZO DERECHO DENTADURA INCOM...	2
10	USA DENTADURA POSTIZA	2
11	UNA CICATRIZ EN LA CEJA IZQUIERDA Y LOS DIENTES FRON...	2
12	FALTAN DIENTES	2
13	TIENE LOS DIENTES CHUECOS	2

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 1,282 rows

Figura 42. Pantalla de agrupación por diente, muela y dentadura, usando una consulta mediante Microsoft SQL Server

```
--Lunares

SELECT DISTINCT SENASPARTICULARES, COUNT (SENASPARTICULARES) AS TOTAL FROM DATOS GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE '%LUNAR%' OR SENASPARTICULARES LIKE '%LUNARES%'
ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

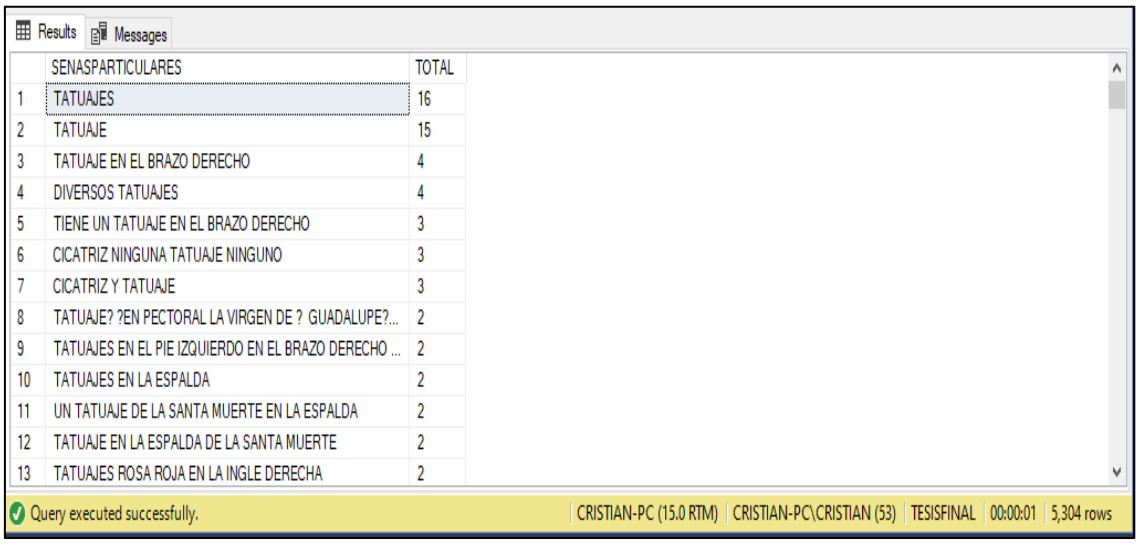
	SENASPARTICULARES	TOTAL
1	LUNAR	15
2	LUNAR EN LA MEJILLA DERECHA	13
3	LUNAR EN MEJILLA IZQUIERDA	12
4	LUNAR EN LA MEJILLA IZQUIERDA	9
5	LUNAR EN MEJILLA DERECHA	6
6	LUNAR EN LA ESPALDA	6
7	LUNAR EN EL CUELLO	6
8	LUNAR DEBAJO DEL OJO DERECHO	4
9	LUNAR EN EL OJO IZQUIERDO	4
10	UN LUNAR EN LA ESPALDA	4
11	UN LUNAR EN EL BRAZO DERECHO	3
12	UN LUNAR EN LA MEJILLA DERECHA	3
13	UN LUNAR EN LA NACHA	3

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 3,353 rows

Figura 43. Pantalla de agrupación por lunares, usando una consulta mediante Microsoft SQL Server

```
--Tatuaje o tatuajes

SELECT DISTINCT SENASPARTICULARES, COUNT (SENASPARTICULARES) AS TOTAL FROM DATOS GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE '%TATUAJE%' ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

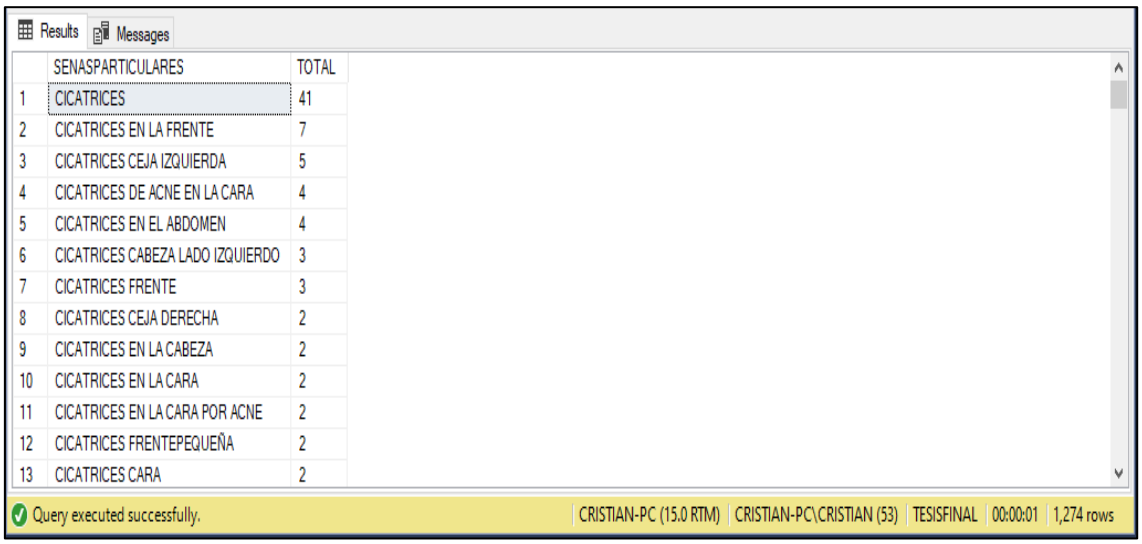
	SENASPARTICULARES	TOTAL
1	TATUAJES	16
2	TATUAJE	15
3	TATUAJE EN EL BRAZO DERECHO	4
4	DIVERSOS TATUAJES	4
5	TIENE UN TATUAJE EN EL BRAZO DERECHO	3
6	CICATRIZ NINGUNA TATUAJE NINGUNO	3
7	CICATRIZ Y TATUAJE	3
8	TATUAJE? ?EN PECTORAL LA VIRGEN DE ? GUADALUPE?...	2
9	TATUAJES EN EL PIE IZQUIERDO EN EL BRAZO DERECHO ...	2
10	TATUAJES EN LA ESPALDA	2
11	UN TATUAJE DE LA SANTA MUERTE EN LA ESPALDA	2
12	TATUAJE EN LA ESPALDA DE LA SANTA MUERTE	2
13	TATUAJES ROSA ROJA EN LA INGLE DERECHA	2

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 5,304 rows

Figura 44. Pantalla de agrupación por tatuajes, usando una consulta mediante Microsoft SQL Server

```
--Cicatriz o cicatrices

SELECT DISTINCT SENASPARTICULARES, COUNT (SENASPARTICULARES) AS TOTAL FROM DATOS GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE '%CICATRICES%' OR SENASPARTICULARES LIKE '%CICATRIS%' ORDER BY COUNT(SENASPARTICULARES) DESC;
```



Results | Messages

	SENASPARTICULARES	TOTAL
1	CICATRICES	41
2	CICATRICES EN LA FRENTE	7
3	CICATRICES CEJA IZQUIERDA	5
4	CICATRICES DE ACNE EN LA CARA	4
5	CICATRICES EN EL ABDOMEN	4
6	CICATRICES CABEZA LADO IZQUIERDO	3
7	CICATRICES FRENTE	3
8	CICATRICES CEJA DERECHA	2
9	CICATRICES EN LA CABEZA	2
10	CICATRICES EN LA CARA	2
11	CICATRICES EN LA CARA POR ACNE	2
12	CICATRICES FRENTEPEQUEÑA	2
13	CICATRICES CARA	2

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:01 | 1,274 rows

Figura 45. Pantalla de agrupación por cicatriz y cicatrices, usando una consulta mediante Microsoft SQL Server

```
--No especificado
SELECT DISTINCT SENASPARTICULARES, COUNT (SENASPARTICULARES) AS TOTAL FROM DATOS GROUP BY SENASPARTICULARES HAVING SENASPARTICULARES LIKE '%NO ESPECIFICADO%' ORDER BY COUNT(SENASPARTICULARES) DESC;
```

	SENASPARTICULARES	TOTAL
1	NO ESPECIFICADO	18581
2	NO ESPECIFICADOESTATURA 165 COMPLEXION DELGADA C...	1
3	USA PERFORACIONES CORPORALES NO ESPECIFICADOS	1
4	CICATRIZ EN LA CEJA IZQUIERDA TATUAJE DE LA VIRGEN D...	1
5	CUENTA CON 7 TATUAJES A LO ANCHO DE LA ESPALDA ALT...	1
6	LE FANTAN LOS DOS DIENTES UBICADOS EN LA PARTE SUP...	1
7	LUNAR NO ESPECIFICADOA VECES SE LE VA EL AVION	1

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (53) | TESISFINAL | 00:00:00 | 7 rows

Figura 46. Pantalla de agrupación por el valor no especificado usando una consulta mediante Microsoft SQL Server

Se elaboró el Query update (consulta de actualización), que permite agrupar por los 8 campos seleccionados

```
---Inicio del query para agrupar las señas particulares
DECLARE @CONT INT = 0;
DECLARE @ID INT = 1;
DECLARE @VALOR VARCHAR(5000)
DECLARE @BIGOTE INT = 0;
DECLARE @ACNEYVERRUGA INT = 0;
DECLARE @DENTADURA INT = 0;
DECLARE @LUNARES INT = 0;
DECLARE @TATUAJES INT = 0;
DECLARE @CICATRICES INT = 0;
DECLARE @NOESPECIFICADO INT = 0;
DECLARE @OTROS INT = 0;
WHILE @CONT < (SELECT COUNT(SENASPARTICULARES) FROM DEPURACION)
	BEGIN
		IF (SELECT SENASPARTICULARES FROM DEPURACION WHERE IDDEPURACION = @ID AND (SENASPARTICULARES LIKE '%BIGOTE%' OR SENASPARTICULARES LIKE '%BARBA%' OR SENASPARTICULARES LIKE '%CARA%')) = (SELECT SENASPARTICULARES FROM DEPURACION WHERE IDDEPURACION = @ID AND (SENASPARTICULARES
```

```
            LIKE '%BIGOTE%' OR SENASPARTICULARES LIKE
            '%BARBA%' OR SENASPARTICULARES LIKE
            '%CARA%'))
            BEGIN
                UPDATE DEPURACION SET
                SENASPARTICULARES = 'BIGOTE' WHERE
                IDDEPURACION = @ID; SET @BIGOTE =
                @BIGOTE+1
            END;
        ELSE IF(SELECT SENASPARTICULARES FROM
    DEPURACION WHERE ID
            DEPURACION = @ID AND (SENASPARTICULARES
            LIKE '%BERRUGA%' OR SENASPARTICULARES LIKE
            '%VERRUGA%' OR SENASPARTICULARES LIKE
            '%ACNE%')) = (SELECT SENASPARTICULARES FROM
            DEPURACION WHERE IDDEPURACION = @ID AND
            (SENASPARTICULARES LIKE '%BERRUGA%' OR
            SENASPARTICULARES LIKE '%VERRUGA%' OR
            SENASPARTICULARES LIKE '%ACNE%'))
            BEGIN
                UPDATE DEPURACION SET
                SENASPARTICULARES = 'ACNE Y VERRUGA'
                WHERE IDDEPURACION = @ID; SET
                @ACNEYVERRUGA = @ACNEYVERRUGA+1
            END;
    ELSE IF(SELECT SENASPARTICULARES FROM DEPURACION
        WHERE IDDEPURACION = @ID AND
        (SENASPARTICULARES LIKE '%DIENTE%' OR
        SENASPARTICULARES LIKE '%DIENTES%' OR
        SENASPARTICULARES LIKE '%MUELA%' OR
        SENASPARTICULARES LIKE '%DENTADURA%' OR
        SENASPARTICULARES LIKE '%BRACKETS%')) = (SELECT
        SENASPARTICULARES FROM DEPURACION WHERE
        IDDEPURACION = @ID AND (SENASPARTICULARES LIKE
        '%DIENTE%' OR SENASPARTICULARES LIKE
        '%DIENTES%' OR SENASPARTICULARES LIKE
        '%MUELA%' OR SENASPARTICULARES LIKE
        '%DENTADURA%' OR SENASPARTICULARES LIKE
        '%BRACKETS%'))
            BEGIN
                UPDATE DEPURACION SET
                SENASPARTICULARES = 'PROBLEMAS EN LA
                DENTADURA' WHERE IDDEPURACION = @ID;
                SET @DENTADURA = @DENTADURA+1
            END;
    ELSE IF(SELECT SENASPARTICULARES FROM DEPURACION
        WHERE IDDEPURACION = @ID AND
        (SENASPARTICULARES LIKE '%LUNAR%' OR
        SENASPARTICULARES LIKE '%LUNARES%')) = (SELECT
        SENASPARTICULARES FROM DEPURACION WHERE
```

```
                IDDEPURACION = @ID AND (SENASPARTICULARES LIKE
                '%LUNAR%'      OR      SENASPARTICULARES      LIKE
                '%LUNARES%'))
                        BEGIN
                          UPDATE              DEPURACION              SET
                          SENASPARTICULARES  =  'LUNARES'  WHERE
                          IDDEPURACION  =  @ID;  SET  @LUNARES  =
                          @LUNARES+1
                         END;
            ELSE IF(SELECT SENASPARTICULARES FROM DEPURACION
                WHERE       IDDEPURACION       =       @ID       AND
                (SENASPARTICULARES    LIKE    '%TATUAJE%'    OR
                SENASPARTICULARES LIKE '%TATUAJES%')) = (SELECT
                SENASPARTICULARES  FROM  DEPURACION  WHERE
                IDDEPURACION = @ID AND (SENASPARTICULARES LIKE
                '%TATUAJE%'      OR      SENASPARTICULARES      LIKE
                '%TATUAJES%'))
                        BEGIN
                          UPDATE              DEPURACION              SET
                          SENASPARTICULARES  =  'TATUAJES'  WHERE
                          IDDEPURACION  =  @ID;   SET  @TATUAJES  =
                          @TATUAJES+1
                         END;
            ELSE IF(SELECT SENASPARTICULARES FROM DEPURACION
                WHERE       IDDEPURACION       =       @ID       AND
                (SENASPARTICULARES    LIKE    '%CICATRICES%'    OR
                SENASPARTICULARES    LIKE    '%CICATRIS%'    OR
                SENASPARTICULARES LIKE '%CICATRIZ%')) = (SELECT
                SENASPARTICULARES  FROM  DEPURACION  WHERE
                IDDEPURACION = @ID AND (SENASPARTICULARES LIKE
                '%CICATRICES%'    OR    SENASPARTICULARES    LIKE
                '%CICATRIS%'      OR      SENASPARTICULARES      LIKE
                '%CICATRIZ%'))
                        BEGIN
                          UPDATE              DEPURACION              SET
                          SENASPARTICULARES  =  'CICATRICES'  WHERE
                          IDDEPURACION  =  @ID;  SET  @CICATRICES  =
                          @CICATRICES+1
                         END;
            ELSE IF(SELECT SENASPARTICULARES FROM DEPURACION
                WHERE       IDDEPURACION       =       @ID       AND
                SENASPARTICULARES LIKE 'NO ESPECIFICADO%') LIKE
                'NO ESPECIFICADO%'
                        BEGIN
                            SET          @NOESPECIFICADO          =
@NOESPECIFICADO+1
                        END;
                 ELSE
                        BEGIN
```

```
                    UPDATE DEPURACION SET
                    SENASPARTICULARES = 'OTRAS SENAS
                    PARTICULARES' WHERE IDDEPURACION =
                    @ID; SET @OTROS = @OTROS+1
                END;

            SET @ID = @ID+1;
            SET @CONT = @CONT+1
        END;
PRINT CONCAT('TERMINFO CONTADOR ',@CONT);
PRINT CONCAT('BIGOTE = ',@BIGOTE);
PRINT CONCAT('ACNE Y VERRUGA = ',@ACNEYVERRUGA);
PRINT CONCAT('DENTADURA = ',@DENTADURA);
PRINT CONCAT('LUNARES = ',@LUNARES);
PRINT CONCAT('TATUAJES = ',@TATUAJES);
PRINT CONCAT('CIACTRICES = ',@CICATRICES);
PRINT CONCAT('OTROS = ',@OTROS);
PRINT CONCAT('NO ESPECIFICADO = ',@NOESPECIFICADO);
GO

---Fin de la consulta para agrupar las señas particulares
```

Por el tipo de consulta elabora se comparó por el tiempo de demora

Tabla 24 *Tiempo de la consulta para el campo de señas particulares*

	Equipo 1	**Equipo 2**
Tiempo	13 minutos con 06 segundos	3 minutos con 15 segundos

En las siguientes figuras respectivamente se muestra el resultado de la consulta diseñada para el campo de señas particulares en esta se ve la descripción del equipo 1 y el equipo 2 dentro de la figura 47 y 48 detalladamente.

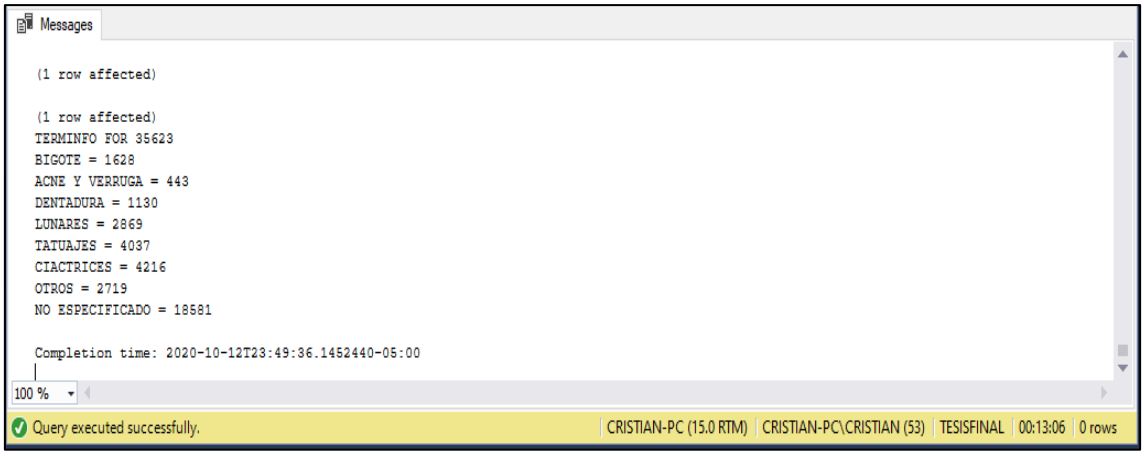


Figura 47. Pantalla del tiempo de ejecución de la consulta señas particulares – equipo 1

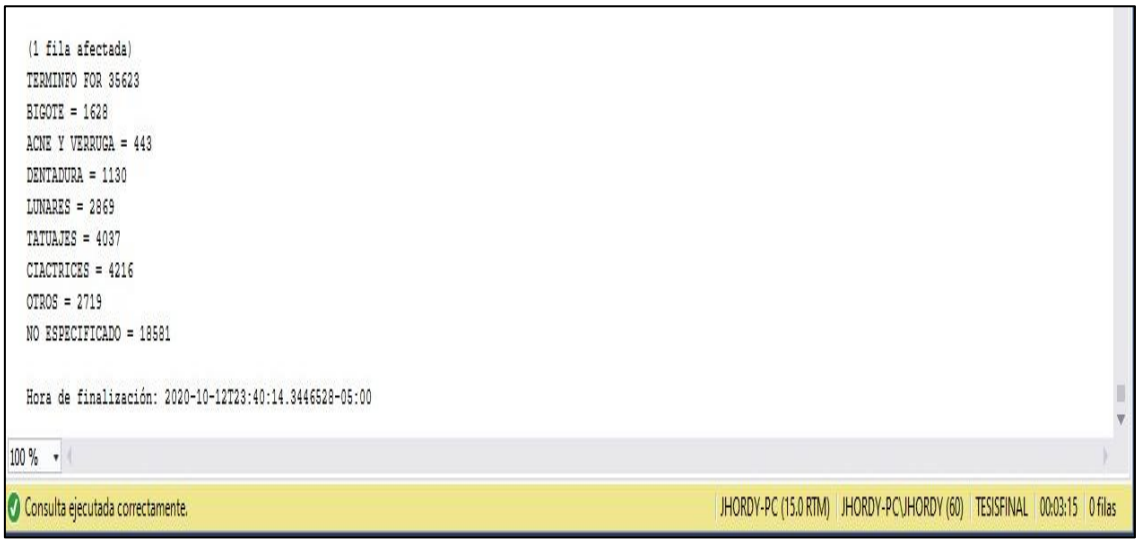


Figura 48. Pantalla del tiempo de ejecución de la consulta señas particulares – equipo 2

Consulta para visualizar los resultados después de ejecutarla anteriormente:

```
SELECT DISTINCT SENASPARTICULARES, COUNT
(SENASPARTICULARES)
AS TOTAL FROM DEPURACION
GROUP BY SENASPARTICULARES
ORDER BY COUNT(SENASPARTICULARES) DESC;
```

Tabla 25 *Agrupación final del campo señas particulares*

SENASPARTICULARES	**TOTAL**
NO ESPECIFICADO	18,581
CICATRICES	4,216
TATUAJES	4,037
LUNARES	2,869
OTRAS SENAS PARTICULARES	2,719
BIGOTE	1,628
PROBLEMAS EN LA DENTADURA	1,130
ACNE Y VERRUGA	443

En la figura 49 se muestra en una gráfica más visual el resultado que se obtuvo de la agrupación de los datos.

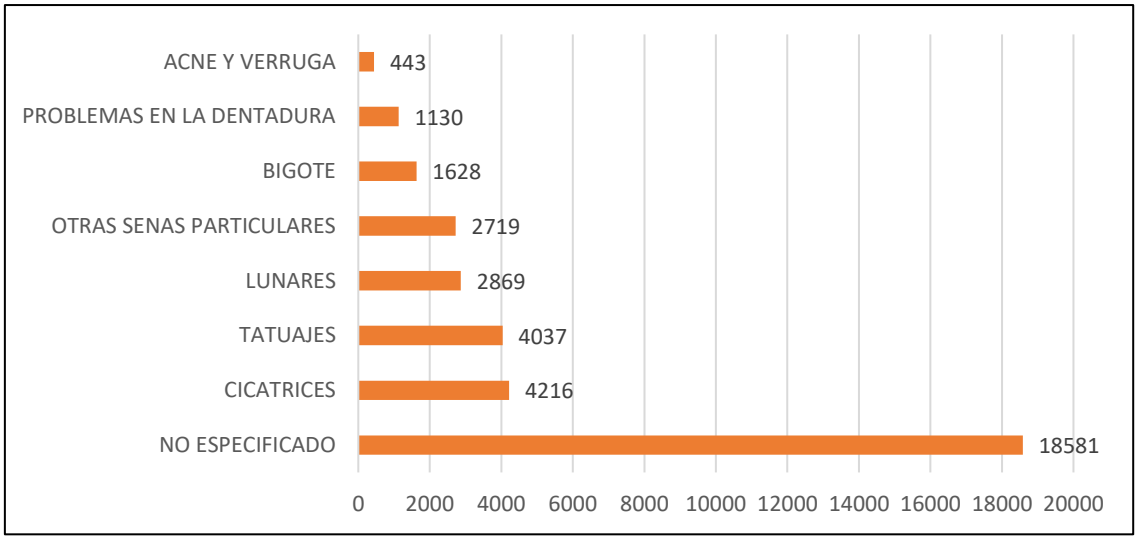


Figura 49. Pantalla de agrupación final del campo de las señas particulares mediante gráfica de barras

IV.3.2 Creación de indicadores

Creación de tablas para para asignación de claves foráneas

```
CREATE TABLE PAIS(
IDPAIS INT IDENTITY(1,1) NOT NULL,
NOMPAIS VARCHAR(50) NOT NULL,
PRIMARY KEY(IDPAIS));
INSERT PAIS
SELECT DISTINCT PAIS FROM DEPURACION;

SELECT * FROM PAIS;
```

Tabla 26 *Asignación del ID al país de México*

IDPAÍS	NOMPAÍS
1	MÉXICO

```
CREATE TABLE ESTADO(
IDESTADO INT IDENTITY(1,1) NOT NULL,
NOMESTADO VARCHAR(50) NOT NULL,
PRIMARY KEY(IDESTADO));
INSERT ESTADO
SELECT DISTINCT ESTADO FROM DEPURACION ORDER BY ESTADO;

SELECT * FROM ESTADO;
```

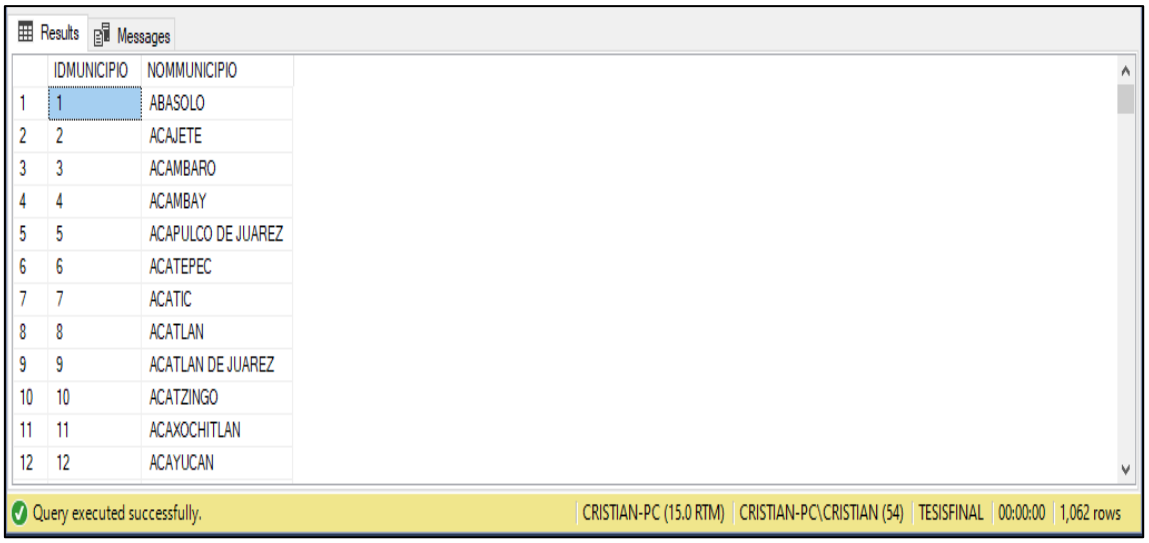
En la siguiente tabla 27 se muestran en detalle los nombres de los treinta y dos estados que conforman el país de México con sus ID que se le ha asignado para tener un mayor conocimiento.

Tabla 27 *Asignación del ID para los estados*

IDESTADO	NOMESTADO	IDESTADO	NOMESTADO
1	AGUASCALIENTES	17	MORELOS
2	BAJA CALIFORNIA	18	NAYARIT
3	BAJA CALIFORNIA SUR	19	NUEVO LEÓN
4	CAMPECHE	20	OAXACA
5	CHIAPAS	21	PUEBLA
6	CHIHUAHUA	22	QUERETARO
7	CIUDAD DE MÉXICO	23	QUINTANA ROO
8	COAHUILA DE ZARAGOZA	24	SAN LUIS POTOSI
9	COLIMA	25	SINALOA
10	DURANGO	26	SONORA
11	ESTADO DE MÉXICO	27	TABASCO
12	GUANAJUATO	28	TAMAULIPAS
13	GUERRERO	29	TLAXCALA
14	HIDALGO	30	VERACRUZ
15	JALISCO	31	YUCATAN
16	MICHOACAN	32	ZACATECAS

```
CREATE TABLE MUNICIPIO(
IDMUNICIPIO INT IDENTITY(1,1) NOT NULL,
NOMMUNICIPIO VARCHAR(50) NOT NULL,
PRIMARY KEY(IDMUNICIPIO));
INSERT MUNICIPIO
SELECT DISTINCT MUNICIPIO FROM DEPURACION ORDER BY
MUNICIPIO;

SELECT * FROM MUNICIPIO
```



Results Messages

	IDMUNICIPIO	NOMMUNICIPIO
1	1	ABASOLO
2	2	ACAJETE
3	3	ACAMBARO
4	4	ACAMBAY
5	5	ACAPULCO DE JUAREZ
6	6	ACATEPEC
7	7	ACATIC
8	8	ACATLAN
9	9	ACATLAN DE JUAREZ
10	10	ACATZINGO
11	11	ACAXOCHITLAN
12	12	ACAYUCAN

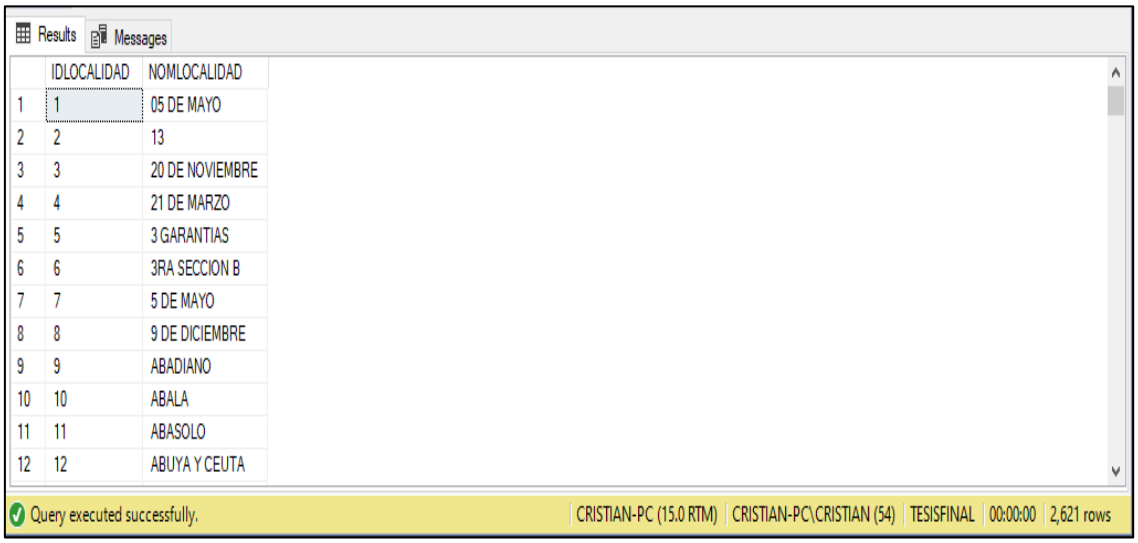
Query executed successfully. CRISTIAN-PC (15.0 RTM) CRISTIAN-PC\CRISTIAN (54) TESISFINAL 00:00:00 1,062 rows

Figura 50. Pantalla de asignación del ID para los municipios dentro de Microsoft SQL Server

```
CREATE TABLE LOCALIDAD(
IDLOCALIDAD INT IDENTITY(1,1) NOT NULL,
NOMLOCALIDAD VARCHAR(100) NOT NULL,
PRIMARY KEY(IDLOCALIDAD));
INSERT LOCALIDAD
SELECT DISTINCT LOCALIDAD FROM DEPURACION ORDER BY
LOCALIDAD;

SELECT * FROM LOCALIDAD
```

En la siguiente figura 51 se muestran las localidades con su ID asignado para su mejor control y manejo de los datos.



Results Messages

	IDLOCALIDAD	NOMLOCALIDAD
1	1	05 DE MAYO
2	2	13
3	3	20 DE NOVIEMBRE
4	4	21 DE MARZO
5	5	3 GARANTIAS
6	6	3RA SECCION B
7	7	5 DE MAYO
8	8	9 DE DICIEMBRE
9	9	ABADIANO
10	10	ABALA
11	11	ABASOLO
12	12	ABUYA Y CEUTA

Query executed successfully. CRISTIAN-PC (15.0 RTM) CRISTIAN-PC\CRISTIAN (54) TESISFINAL 00:00:00 2,621 rows

Figura 51. Pantalla de asignación del ID para las localidades mediante Microsoft SQL Server

```
CREATE TABLE NACIONALIDAD(
IDNACIONALIDAD INT IDENTITY(1,1) NOT NULL,
NOMNACIONALIDAD VARCHAR(100) NOT NULL,
PRIMARY KEY(IDNACIONALIDAD));
INSERT NACIONALIDAD
SELECT DISTINCT NACIONALIDAD FROM DEPURACION ORDER BY
NACIONALIDAD;

SELECT * FROM NACIONALIDAD;
```

En la tabla 28 se describe los diversos países que se encontraron en los datos de las personas desaparecidas deduciendo a que no solo son personas nacionales sino también extranjeras las cuales desaparecen, por esto se les asignó a los diversos países un ID para su mejor identificación.

Tabla 28 *Asignación del ID para las nacionalidades*

ID	NOMNACIONALIDAD
1	ALEMANA
2	ARGENTINA
3	BRASILEÑA
4	CANADIENSE
5	CHINA
6	COLOMBIANA
7	COSTARICENSE
8	CUBANA
9	DOMINICANA
10	ERITREA
11	ESPAÑOLA
12	ESTADOUNIDENSE
13	FRANCESA

ID	NOMNACIONALIDAD
14	GUATEMALTECO
15	HAITIANA
16	HOLANDESA
17	HONDUREÑA
18	ITALIANA
19	MEXICANA
20	NICARAGUENSE
21	NO ESPECIFICADO
22	PANAMEÑA
23	PARAGUAYA
24	PERUANA
25	SALVADOREÑA
26	VENEZOLANA

```
CREATE TABLE COMPLEXION(
IDCOMPLEXION INT IDENTITY(1,1) NOT NULL,
NOMCOMPLEXION VARCHAR(50) NOT NULL,
PRIMARY KEY(IDCOMPLEXION));
INSERT COMPLEXION
SELECT DISTINCT COMPLEXION FROM DEPURACION ORDER BY
COMPLEXION;
SELECT * FROM COMPLEXION;
```

En la tabla 29 se describen los diversos tipos de complexiones que se obtuvieron dentro de los datos de la RNPED las cuales se asignaron un ID para su identificación.

Tabla 29 *Asignación del ID para la complexión*

IDCOMPLEXION	NOMCOMPLEXION
1	DELGADA
2	MEDIANA
3	NO ESPECIFICADO
4	OBESA
5	ROBUSTA

```
CREATE TABLE SEXO(
IDSEXO INT IDENTITY(1,1) NOT NULL,
NOMSEXO VARCHAR(20) NOT NULL,
PRIMARY KEY(IDSEXO) );
INSERT SEXO
SELECT DISTINCT SEXO FROM DEPURACION;

SELECT * FROM SEXO;
```

En la tabla 30 se describe la designación de ID en este apartado solo se tiene los dos tipos hombre y mujer los cuales tienen el valor de 1 y 2 respectivamente esto ayudó a tener un mejor control y manejo de la información.

Tabla 30 *Asignación del ID para el campo sexo*

IDSEXO	NOMSEXO
1	HOMBRE
2	MUJER

```
CREATE TABLE SENASPARTICULARES(
IDSENASPARTICULARES INT IDENTITY(1,1) NOT NULL,
NOMSENAS VARCHAR(100) NOT NULL,
PRIMARY KEY(IDSENASPARTICULARES));
INSERT SENASPARTICULARES
SELECT DISTINCT SENASPARTICULARES FROM DEPURACION
ORDER BY SENASPARTICULARES;
SELECT * FROM SENASPARTICULARES;
```

Después de la elaboración de la consulta anteriormente detallada y explicada se concluyó a tener 8 tipos de señas particulares de las cuales estas fueron las que se trabajaron detectado esto se le asignó un ID a cada variable respectivamente.

Tabla 31 *Asignación del ID para las señas particulares*

ID	NOMSENAS	ID	NOMSENAS
1	ACNE Y VERRUGA	5	NO ESPECIFICADO
2	BIGOTE	6	OTRAS SENAS PARTICULARES
3	CICATRICES	7	PROBLEMAS EN LA DENTADURA
4	LUNARES	8	TATUAJES

```
CREATE TABLE DEPENDENCIA(
IDDEPENDENCIA INT IDENTITY(1,1) NOT NULL,
NOMDEPENDENCIA VARCHAR(100) NOT NULL,
PRIMARY KEY(IDDEPENDENCIA));
INSERT DEPENDENCIA
SELECT DISTINCT DEPENDENCIA FROM DEPURACION ORDER BY
DEPENDENCIA;

SELECT * FROM DEPENDENCIA;
```

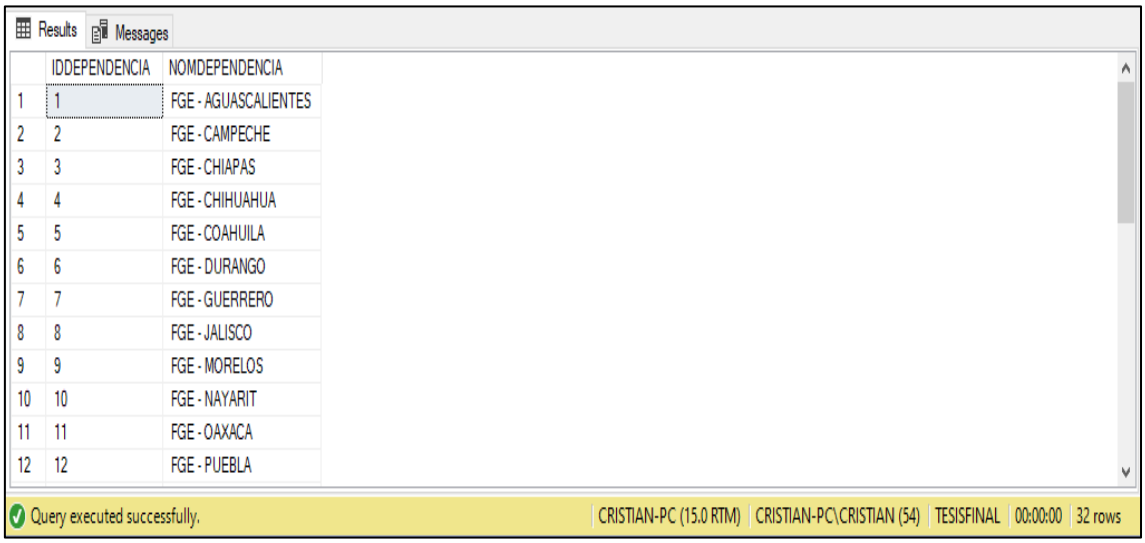


Figura 52. Pantalla de asignación del ID para las dependencias dentro de Microsoft SQL Server

```
--Creación de la tabla desaparecidos con datos relacionados

CREATE TABLE DESAPARECIDOS (
IDDESAPARECIDOS INT NOT NULL,
FECHA DATE NOT NULL,
IDPAIS INT NOT NULL,
IDESTADO INT NOT NULL,
IDMUNICIPIO INT NOT NULL,
IDLOCALIDAD INT NOT NULL,
IDNACIONALIDAD INT NOT NULL,
IDCOMPLEXION INT NOT NULL,
IDSEXO INT NOT NULL,
EDAD INT NOT NULL,
IDSENASPARTICULARES INT NOT NULL,
```

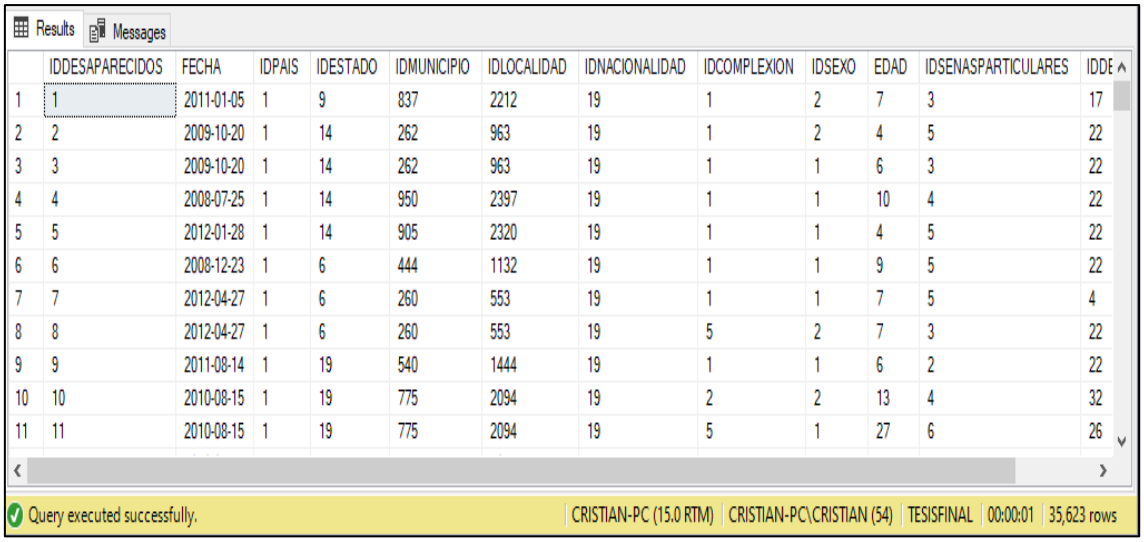
```
IDDEPENDENCIA INT NOT NULL,
FOREIGN KEY (IDPAIS) REFERENCES PAIS(IDPAIS),
FOREIGN KEY (IDESTADO) REFERENCES ESTADO(IDESTADO),
FOREIGN KEY (IDMUNICIPIO) REFERENCES MUNICIPIO(IDMUNICIPIO),
FOREIGN KEY (IDLOCALIDAD) REFERENCES LOCALIDAD(IDLOCALIDAD),
FOREIGN KEY (IDNACIONALIDAD) REFERENCES
NACIONALIDAD(IDNACIONALIDAD),
FOREIGN KEY (IDCOMPLEXION) REFERENCES
COMPLEXION(IDCOMPLEXION),
FOREIGN KEY (IDSEXO) REFERENCES SEXO(IDSEXO),
FOREIGN KEY (IDSENASPARTICULARES) REFERENCES
SENASPARTICULARES(IDSENASPARTICULARES),
FOREIGN KEY (IDDEPENDENCIA) REFERENCES
DEPENDENCIA(IDDEPENDENCIA),
PRIMARY KEY(IDDESAPARECIDOS));

--Inserción de datos en la tabla transaccional

INSERT DESAPARECIDOS
SELECT DEP.IDDESAPARECIDOS, DEP. FECHA, PA. IDPAIS, E.
IDESTADO, M. IDMUNICIPIO, L. IDLOCALIDAD, N. IDNACIONALIDAD, COM.
IDCOMPLEXION, SE. IDSEXO, DEP. EDAD, SP. IDSENASPARTICULARES,
DEPEN.IDDEPENDENCIA
FROM DEPURACION DEP
INNER JOIN PAIS PA ON PA.NOMPAIS=DEP.PAIS
INNER JOIN ESTADO E ON E.NOMESTADO=DEP.ESTADO
INNER JOIN MUNICIPIO M ON M.NOMMUNICIPIO=DEP.MUNICIPIO
INNER JOIN LOCALIDAD L ON L.NOMLOCALIDAD=DEP.LOCALIDAD
INNER JOIN NACIONALIDAD N ON
N.NOMNACIONALIDAD=DEP.NACIONALIDAD
INNER JOIN COMPLEXION COM ON COM.NOMCOMPLEXION=
DEP.COMPLEXION
INNER JOIN SEXO SE ON SE.NOMSEXO=DEP.SEXO
INNER JOIN SENASPARTICULARES SP ON
SP.NOMSENAS=DEP.SENASPARTICULARES
INNER JOIN DEPENDENCIA DEPEN ON
DEPEN.NOMDEPENDENCIA=DEP.DEPENDENCIA
```

En la figura 53 se muestra la tabla que se ha denominado como desaparecidos donde se tiene todos los ID anteriormente detallados esto permite tener un mejor control de los datos.

```
SELECT * FROM DESAPARECIDOS;
```



Results | Messages

	IDDESAPARECIDOS	FECHA	IDPAIS	IDESTADO	IDMUNICIPIO	IDLOCALIDAD	IDNACIONALIDAD	IDCOMPLEXION	IDSEXO	EDAD	IDSENASPARTICULARES	IDDE
1	1	2011-01-05	1	9	837	2212	19	1	2	7	3	17
2	2	2009-10-20	1	14	262	963	19	1	2	4	5	22
3	3	2009-10-20	1	14	262	963	19	1	1	6	3	22
4	4	2008-07-25	1	14	950	2397	19	1	1	10	4	22
5	5	2012-01-28	1	14	905	2320	19	1	1	4	5	22
6	6	2008-12-23	1	6	444	1132	19	1	1	9	5	22
7	7	2012-04-27	1	6	260	553	19	1	1	7	5	4
8	8	2012-04-27	1	6	260	553	19	5	2	7	3	22
9	9	2011-08-14	1	19	540	1444	19	1	1	6	2	22
10	10	2010-08-15	1	19	775	2094	19	2	2	13	4	32
11	11	2010-08-15	1	19	775	2094	19	5	1	27	6	26

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:01 | 35,623 rows

Figura 53. Pantalla de la tabla desaparecidos dentro de Microsoft SQL Server

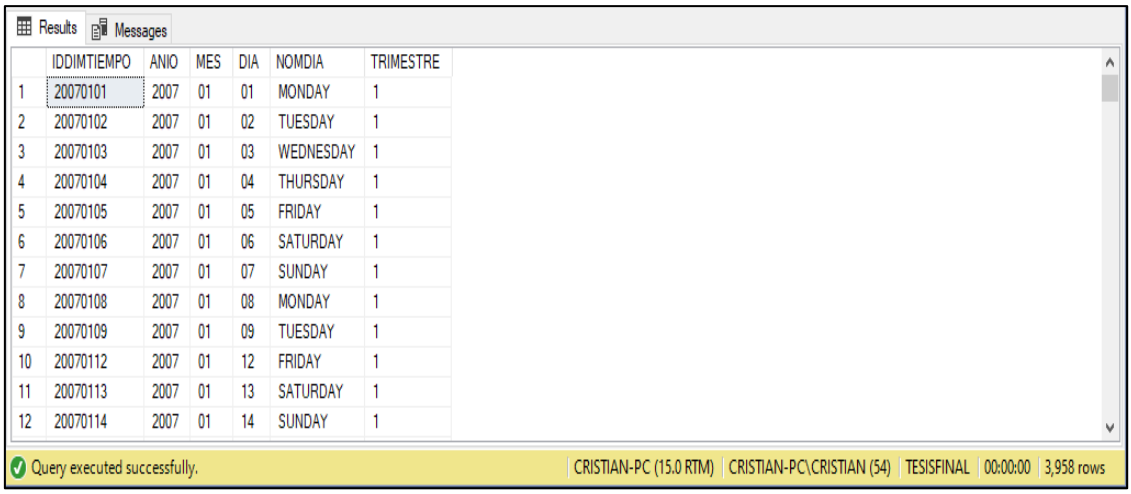
```
--Creación de las dimensiones

--Dimensión tiempo

CREATE TABLE DIM_TIEMPO(
IDDIMTIEMPO INT NOT NULL,
ANIO INT NOT NULL,
MES CHAR(2) NOT NULL,
DIA CHAR(2) NOT NULL,
NOMDIA VARCHAR(50) NOT NULL,
TRIMESTRE INT NOT NULL,
PRIMARY KEY(IDDIMTIEMPO));
INSERT DIM_TIEMPO
SELECT
        DISTINCT  YEAR(FECHA)*10000+MONTH(FECHA)*100+DAY(FECHA)
AS ID,
        YEAR(FECHA) AS ANIO,RIGHT(1000 + MONTH(FECHA), 2) AS
        MES,RIGHT(1000 + DAY(FECHA), 2) AS DIA,
        UPPER(DATENAME(WEEKDAY, FECHA)) AS NOMDIA,
        DATEPART(QUARTER,FECHA) AS TRIMESTRE
FROM DEPURACION
```

En la figura 54 se muestra la dimensión tiempo donde lo dividimos por el ID ya especificado el año, mes, día, el nombre del día y el trimestre.

```
SELECT * FROM DIM_TIEMPO
```



Results | Messages

	IDDIMTIEMPO	ANIO	MES	DIA	NOMDIA	TRIMESTRE
1	20070101	2007	01	01	MONDAY	1
2	20070102	2007	01	02	TUESDAY	1
3	20070103	2007	01	03	WEDNESDAY	1
4	20070104	2007	01	04	THURSDAY	1
5	20070105	2007	01	05	FRIDAY	1
6	20070106	2007	01	06	SATURDAY	1
7	20070107	2007	01	07	SUNDAY	1
8	20070108	2007	01	08	MONDAY	1
9	20070109	2007	01	09	TUESDAY	1
10	20070112	2007	01	12	FRIDAY	1
11	20070113	2007	01	13	SATURDAY	1
12	20070114	2007	01	14	SUNDAY	1

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:00 | 3,958 rows

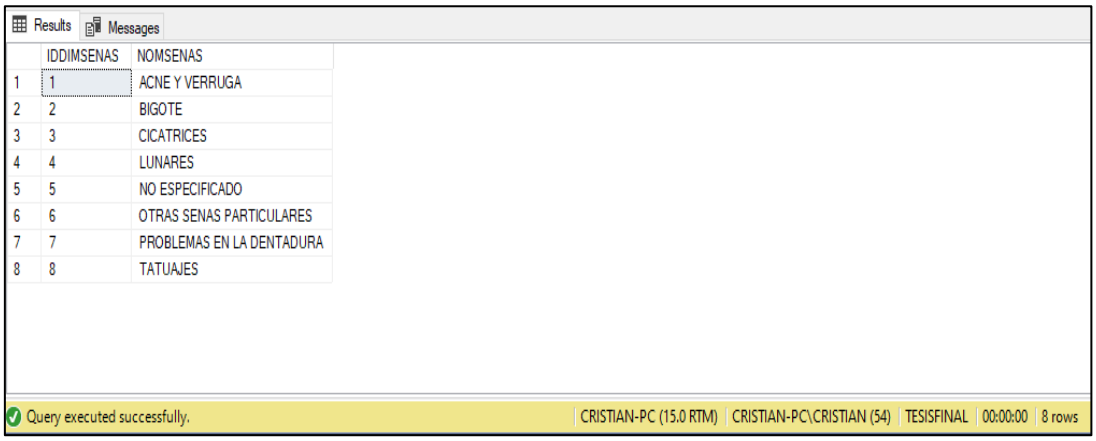
Figura 54. Pantalla de la dimensión tiempo dividida por las especificaciones dentro de Microsoft SQL Server

```
--Dimensión señas particulares

CREATE TABLE DIM_SENASPARTICULARES(
IDDIMSENAS INT NOT NULL,
NOMSENAS VARCHAR(100) NOT NULL,
PRIMARY KEY(IDDIMSENAS));
INSERT DIM_SENASPARTICULARES
SELECT DISTINCT  DE.IDSENASPARTICULARES,SENAS.NOMSENAS
 FROM DESAPARECIDOS DE
INNER JOIN SENASPARTICULARES SENAS ON
SENAS.IDSENASPARTICULARES = DE.IDSENASPARTICULARES

SELECT * FROM DIM_SENASPARTICULARES;
```

Dentro de la figura 55 donde se describen las señas particulares que anteriormente se dividieron en los ocho campos o id que se establecieron para su mayor comprensión y control de los mismos datos.



Results | Messages

	IDDIMSENAS	NOMSENAS
1	1	ACNE Y VERRUGA
2	2	BIGOTE
3	3	CICATRICES
4	4	LUNARES
5	5	NO ESPECIFICADO
6	6	OTRAS SENAS PARTICULARES
7	7	PROBLEMAS EN LA DENTADURA
8	8	TATUAJES

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:00 | 8 rows

Figura 55. Pantalla de la dimensión señas particulares dentro de Microsoft SQL Server

```
--Dimensión complexión

CREATE TABLE DIM COMPLEXION(
IDDIMCOMPLEXION INT NOT NULL,
VALOR INT NOT NULL,
NOMCOMPLEXION VARCHAR(50) NOT NULL,
PRIMARY KEY(IDDIMCOMPLEXION));
INSERT DIM_COMPLEXION
SELECT DISTINCT DESA.IDCOMPLEXION,'',COM.NOMCOMPLEXION
FROM DESAPARECIDOS DESA
INNER JOIN COMPLEXION COM ON COM.IDCOMPLEXION =
DESA.IDCOMPLEXION

SELECT * FROM DIM_COMPLEXION;
```

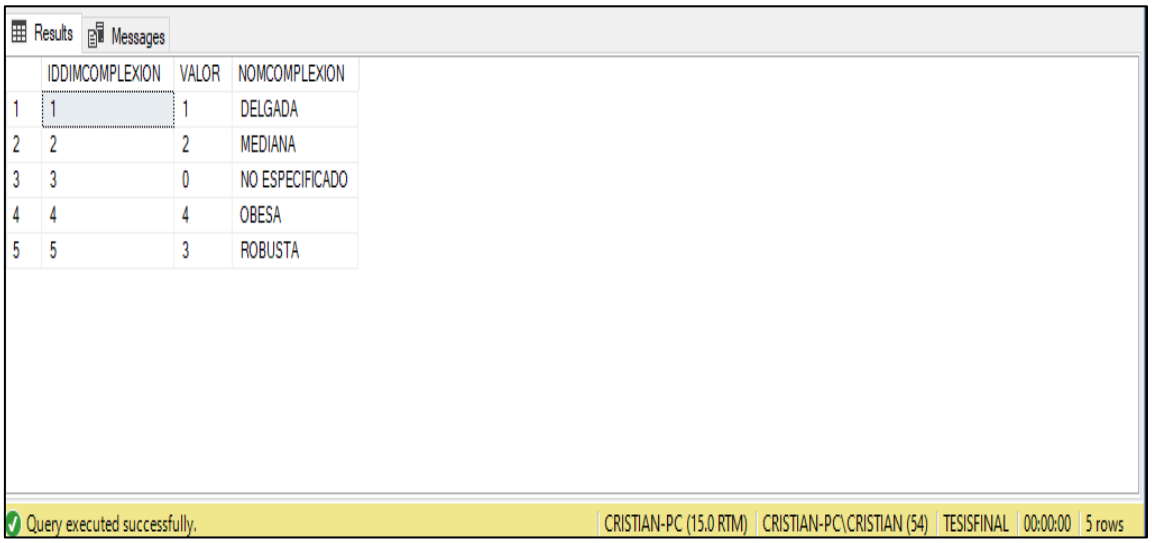


Figura 56. Pantalla de la dimensión complexión dentro de Microsoft SQL Server

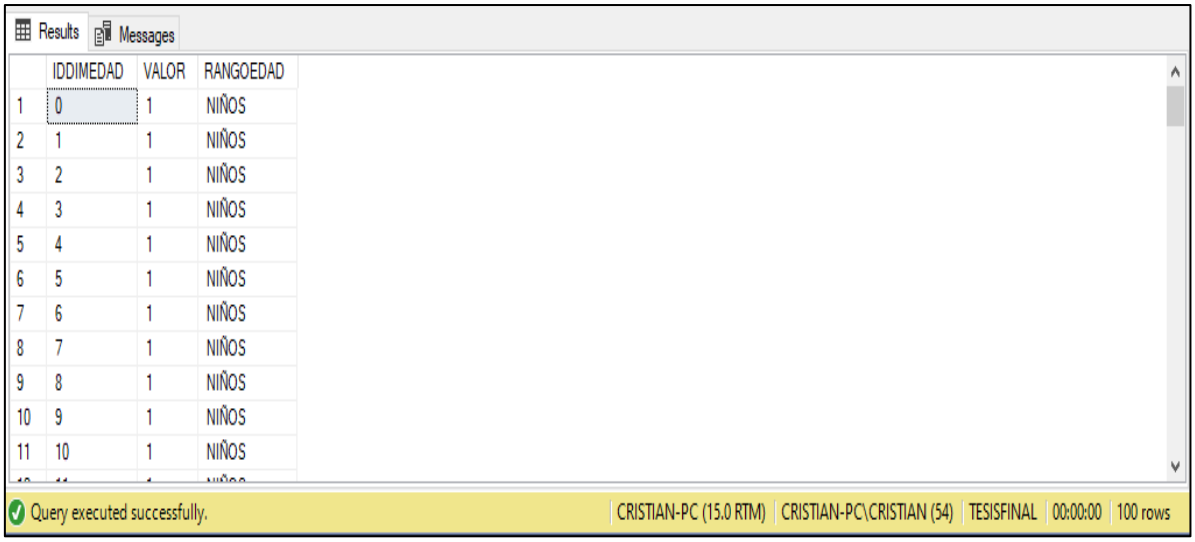
```
-- Dimensión edad

CREATE TABLE DIM_EDAD(
IDDIMEDAD INT NOT NULL,
VALOR INT NOT NULL,
RANGOEDAD VARCHAR(50) NOT NULL,
PRIMARY KEY(IDDIMEDAD));
INSERT DIM_EDAD
SELECT DISTINCT EDAD,'',
        CASE WHEN (EDAD BETWEEN 0 AND 11) THEN 'NIÑOS' ELSE
                CASE WHEN (EDAD BETWEEN 12 AND 17) THEN
'ADOLESCENTE' ELSE
                        CASE WHEN (EDAD BETWEEN 18 AND 29) THEN
'JOVENES' ELSE
                                CASE WHEN (EDAD BETWEEN 30 AND 59) THEN
                                'ADULTO' ELSE
```

```
                                    CASE WHEN (EDAD >= 60) THEN 'ADULTO
MAYOR'
                                    END
                                END
                            END
                        END
                    END RANGO
FROM DESAPARECIDOS
ORDER BY EDAD ASC

SELECT * FROM DIM_EDAD;
```

En la figura 57 se muestra la designación de los ID con sus valores para su mejor manejo dentro de estos se tiene intervalos de cada uno de ellos los cuales se dividen en niños, adolescentes, jóvenes, adultos y adultos mayores, esta consulta ayudó a tener un mejor manejo de este campo.



	IDDIMEDAD	VALOR	RANGOEDAD
1	0	1	NIÑOS
2	1	1	NIÑOS
3	2	1	NIÑOS
4	3	1	NIÑOS
5	4	1	NIÑOS
6	5	1	NIÑOS
7	6	1	NIÑOS
8	7	1	NIÑOS
9	8	1	NIÑOS
10	9	1	NIÑOS
11	10	1	NIÑOS

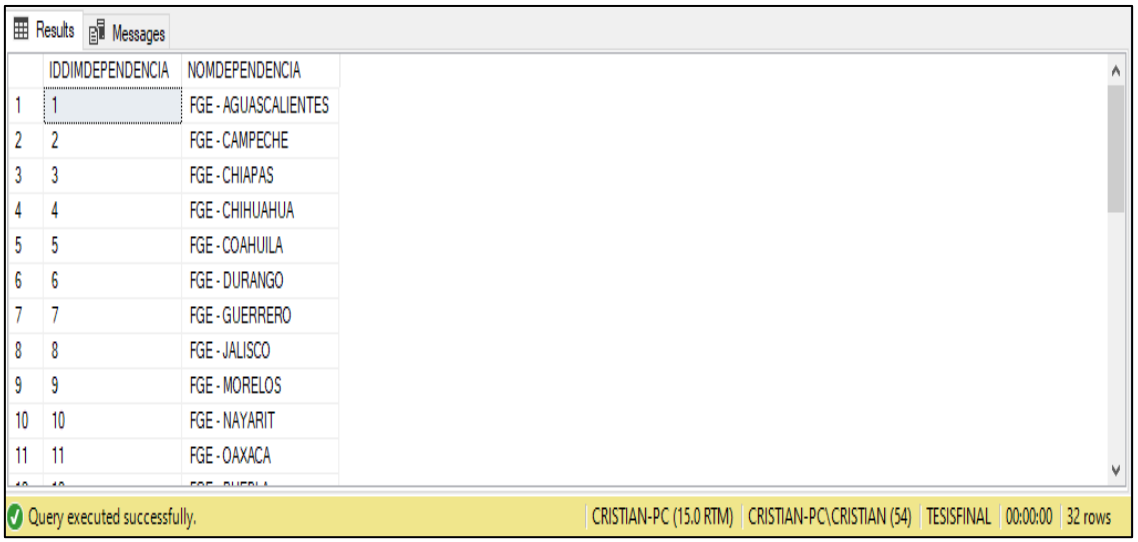
Figura 57. Pantalla de la dimensión edad dentro de Microsoft SQL Server

```
-- Dimensión dependencia

CREATE TABLE DIM_DEPENDENCIA(
IDDIMDEPENDENCIA INT NOT NULL,
NOMDEPENDENCIA VARCHAR(50) NOT NULL,
PRIMARY KEY(IDDIMDEPENDENCIA));
INSERT DIM_DEPENDENCIA
SELECT DISTINCT DESA.IDDEPENDENCIA,DEP.NOMDEPENDENCIA
FROM DESAPARECIDOS DESA
INNER JOIN DEPENDENCIA DEP ON DEP.IDDEPENDENCIA =
DESA.IDDEPENDENCIA

SELECT * FROM DIM_DEPENDENCIA;
```

En la figura 58 se muestra la designación de los ID para las diversas dependencias que se encontró a lo largo de los datos.



Results | Messages

	IDDIMDEPENDENCIA	NOMDEPENDENCIA
1	1	FGE - AGUASCALIENTES
2	2	FGE - CAMPECHE
3	3	FGE - CHIAPAS
4	4	FGE - CHIHUAHUA
5	5	FGE - COAHUILA
6	6	FGE - DURANGO
7	7	FGE - GUERRERO
8	8	FGE - JALISCO
9	9	FGE - MORELOS
10	10	FGE - NAYARIT
11	11	FGE - OAXACA

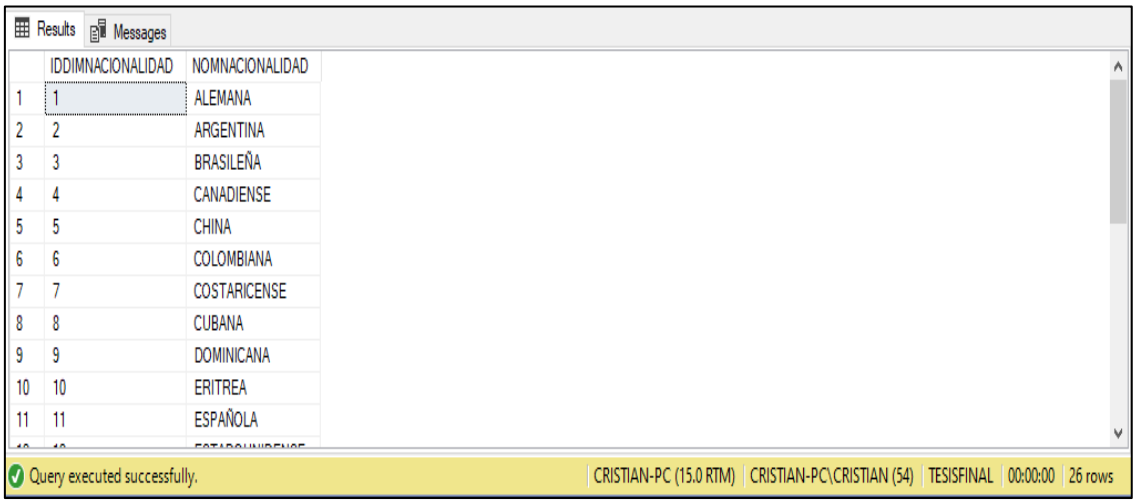
Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:00 | 32 rows

Figura 58. Pantalla de la dimensión dependencia dentro de Microsoft SQL Server

```
-- Dimensión nacionalidad

CREATE TABLE DIM_NACIONALIDAD(
IDDIMNACIONALIDAD INT NOT NULL,
NOMNACIONALIDAD VARCHAR(50) NOT NULL,
PRIMARY KEY(IDDIMNACIONALIDAD));
INSERT DIM_NACIONALIDAD
SELECT DISTINCT DESA.IDNACIONALIDAD,NAC.NOMNACIONALIDAD
FROM DESAPARECIDOS DESA
INNER JOIN NACIONALIDAD NAC ON NAC.IDNACIONALIDAD =
DESA.IDNACIONALIDAD

SELECT * FROM DIM_NACIONALIDAD;
```



Results | Messages

	IDDIMNACIONALIDAD	NOMNACIONALIDAD
1	1	ALEMANA
2	2	ARGENTINA
3	3	BRASILEÑA
4	4	CANADIENSE
5	5	CHINA
6	6	COLOMBIANA
7	7	COSTARICENSE
8	8	CUBANA
9	9	DOMINICANA
10	10	ERITREA
11	11	ESPAÑOLA

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:00 | 26 rows

Figura 59. Pantalla de la dimensión nacionalidad dentro de Microsoft SQL Server

```
-- Dimensión lugar

CREATE TABLE DIM_LUGAR(
IDDIMUBIGEO CHAR(11) NOT NULL,
NOMESTADO VARCHAR(200) NOT NULL,
NOMMUNICIPIO VARCHAR(200) NOT NULL,
NOMLOCALIDAD VARCHAR(200) NOT NULL,
PRIMARY KEY(IDDIMUBIGEO));
INSERT DIM_LUGAR
SELECT DISTINCT CONCAT (ESTA. IDESTADO, MUNI. IDMUNICIPIO, LOC.
IDLOCALIDAD) AS UBIGEO, ESTA. NOMESTADO, MUNI.NOMMUNICIPIO,
LOC.NOMLOCALIDAD FROM DESAPARECIDOS DESA
INNER JOIN ESTADO ESTA ON ESTA.IDESTADO = DESA.IDESTADO
INNER JOIN MUNICIPIO MUNI ON MUNI.IDMUNICIPIO = DESA.IDMUNICIPIO
INNER JOIN LOCALIDAD LOC ON LOC.IDLOCALIDAD = DESA.IDLOCALIDAD

SELECT * FROM DIM_LUGAR;
```

Results | Messages

	IDDIMUBIGEO	NOMESTADO	NOMMUNICIPIO	NOMLOCALIDAD
1	101491079	DURANGO	CANATLAN	J. GPE. AGUILERA
2	102891166	DURANGO	DURANGO	LA FERRERIA
3	102891502	DURANGO	DURANGO	NO ESPECIFICADO
4	102891652	DURANGO	DURANGO	POB. EL REFUGIO
5	102891708	DURANGO	DURANGO	PUEBLO NUEVO
6	102891803	DURANGO	DURANGO	SALTO P.N
7	102891898	DURANGO	DURANGO	SAN JOSE DE LA VINATA
8	102891967	DURANGO	DURANGO	SAN MIGUEL DE CRUCES
9	102892269	DURANGO	DURANGO	TEPEHUANES
10	102892465	DURANGO	DURANGO	VICTORIA DE DURANGO
11	10289608	DURANGO	DURANGO	DURABGO

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:00 | 3,538 rows

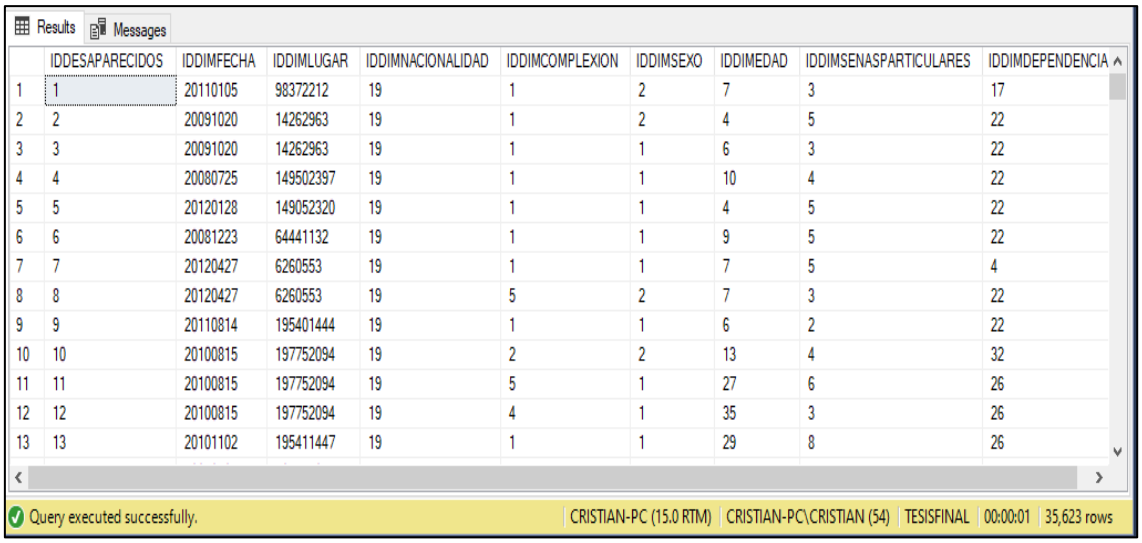
Figura 60. Pantalla de la dimensión lugar dentro de Microsoft SQL Server

```
-- Dimensión hechos

CREATE TABLE HECHOS DESAPARECIDOS(
IDDESAPARECIDOS INT NOT NULL,
IDDIMFECHA INT NOT NULL,
IDDIMLUGAR CHAR(11) NOT NULL,
IDDIMNACIONALIDAD INT NOT NULL,
IDDIMCOMPLEXION INT NOT NULL,
IDDIMSEXO INT NOT NULL,
IDDIMEDAD INT NOT NULL,
IDDIMSENASPARTICULARES INT NOT NULL,
IDDIMDEPENDENCIA INT NOT NULL
FOREIGN KEY (IDDESAPARECIDOS) REFERENCES
DESAPARECIDOS(IDDESAPARECIDOS),
FOREIGN KEY (IDDIMFECHA) REFERENCES DIM_TIEMPO(IDDIMTIEMPO),
```

```
FOREIGN KEY (IDDIMLUGAR) REFERENCES DIM_LUGAR(IDDIMUBIGEO),
FOREIGN KEY (IDDIMNACIONALIDAD) REFERENCES
DIM_NACIONALIDAD(IDDIMNACIONALIDAD),
FOREIGN KEY (IDDIMCOMPLEXION) REFERENCES
DIM_COMPLEXION(IDDIMCOMPLEXION),
FOREIGN KEY (IDDIMSEXO) REFERENCES DIM_SEXO(IDDIMSEXO),
FOREIGN KEY (IDDIMEDAD) REFERENCES DIM_EDAD(IDDIMEDAD),
FOREIGN KEY (IDDIMSENASPARTICULARES) REFERENCES
DIM_SENASPARTICULARES(IDDIMSENAS),
FOREIGN KEY (IDDIMDEPENDENCIA) REFERENCES
DIM_DEPENDENCIA(IDDIMDEPENDENCIA),
PRIMARY KEY(IDDESAPARECIDOS,
IDDIMFECHA,IDDIMLUGAR,IDDIMNACIONALIDAD,IDDIMCOMPLEXION,IDD
IMSEXO,IDDIMEDAD,IDDIMSENASPARTICULARES,IDDIMDEPENDENCIA));
INSERT HECHOS_DESAPARECIDOS
SELECT IDDESAPARECIDOS, FEC.IDDIMTIEMPO, LUG.IDDIMUBIGEO,
NAC.IDDIMNACIONALIDAD, COMP.IDDIMCOMPLEXION,
SEX.IDDIMSEXO,EDA. IDDIMEDAD, SP. IDDIMSENAS, DP.
IDDIMDEPENDENCIA
FROM DESAPARECIDOS DESA
INNER JOIN DIM_TIEMPO FEC ON FEC.IDDIMTIEMPO =
YEAR(DESA.FECHA)*10000+MONTH(DESA.FECHA)*100+DAY(DESA.FECH
A)
INNER JOIN ESTADO ESTA ON ESTA.IDESTADO = DESA.IDESTADO
INNER JOIN MUNICIPIO MUNI ON MUNI.IDMUNICIPIO =
DESA.IDMUNICIPIO
INNER JOIN LOCALIDAD LOC ON LOC.IDLOCALIDAD =
DESA.IDLOCALIDAD
INNER JOIN DIM_LUGAR LUG ON LUG.IDDIMUBIGEO =
CONCAT(ESTA.IDESTADO,MUNI.IDMUNICIPIO,LOC.IDLOCALIDAD)
INNER JOIN DIM_NACIONALIDAD NAC ON NAC.IDDIMNACIONALIDAD =
DESA.IDNACIONALIDAD
INNER JOIN DIM_COMPLEXION COMP ON COMP.IDDIMCOMPLEXION =
DESA.IDCOMPLEXION
INNER JOIN DIM_SEXO SEX ON SEX.IDDIMSEXO = DESA.IDSEXO
INNER JOIN DIM_EDAD EDA ON EDA.IDDIMEDAD = DESA.EDAD
INNER JOIN DIM_SENASPARTICULARES SP ON SP.IDDIMSENAS =
DESA.IDSENASPARTICULARES
INNER JOIN DIM_DEPENDENCIA DP ON DP.IDDIMDEPENDENCIA =
DESA.IDDEPENDENCIA

SELECT * FROM HECHOS_DESAPARECIDOS;
```



Results | Messages

	IDDESAPARECIDOS	IDDIMFECHA	IDDIMLUGAR	IDDIMNACIONALIDAD	IDDIMCOMPLEXION	IDDIMSEXO	IDDIMEDAD	IDDIMSENASPARTICULARES	IDDIMDEPENDENCIA
1	1	20110105	98372212	19	1	2	7	3	17
2	2	20091020	14262963	19	1	2	4	5	22
3	3	20091020	14262963	19	1	1	6	3	22
4	4	20080725	149502397	19	1	1	10	4	22
5	5	20120128	149052320	19	1	1	4	5	22
6	6	20081223	64441132	19	1	1	9	5	22
7	7	20120427	6260553	19	1	1	7	5	4
8	8	20120427	6260553	19	5	2	7	3	22
9	9	20110814	195401444	19	1	1	6	2	22
10	10	20100815	197752094	19	2	2	13	4	32
11	11	20100815	197752094	19	5	1	27	6	26
12	12	20100815	197752094	19	4	1	35	3	26
13	13	20101102	195411447	19	1	1	29	8	26

Query executed successfully. | CRISTIAN-PC (15.0 RTM) | CRISTIAN-PC\CRISTIAN (54) | TESISFINAL | 00:00:01 | 35,623 rows

Figura 61. Pantalla de la dimensión hechos dentro Microsoft SQL Server

En la figura 62 se muestra el modelo dimensional de la base de datos la cual se estructuró con lo anteriormente explicado.

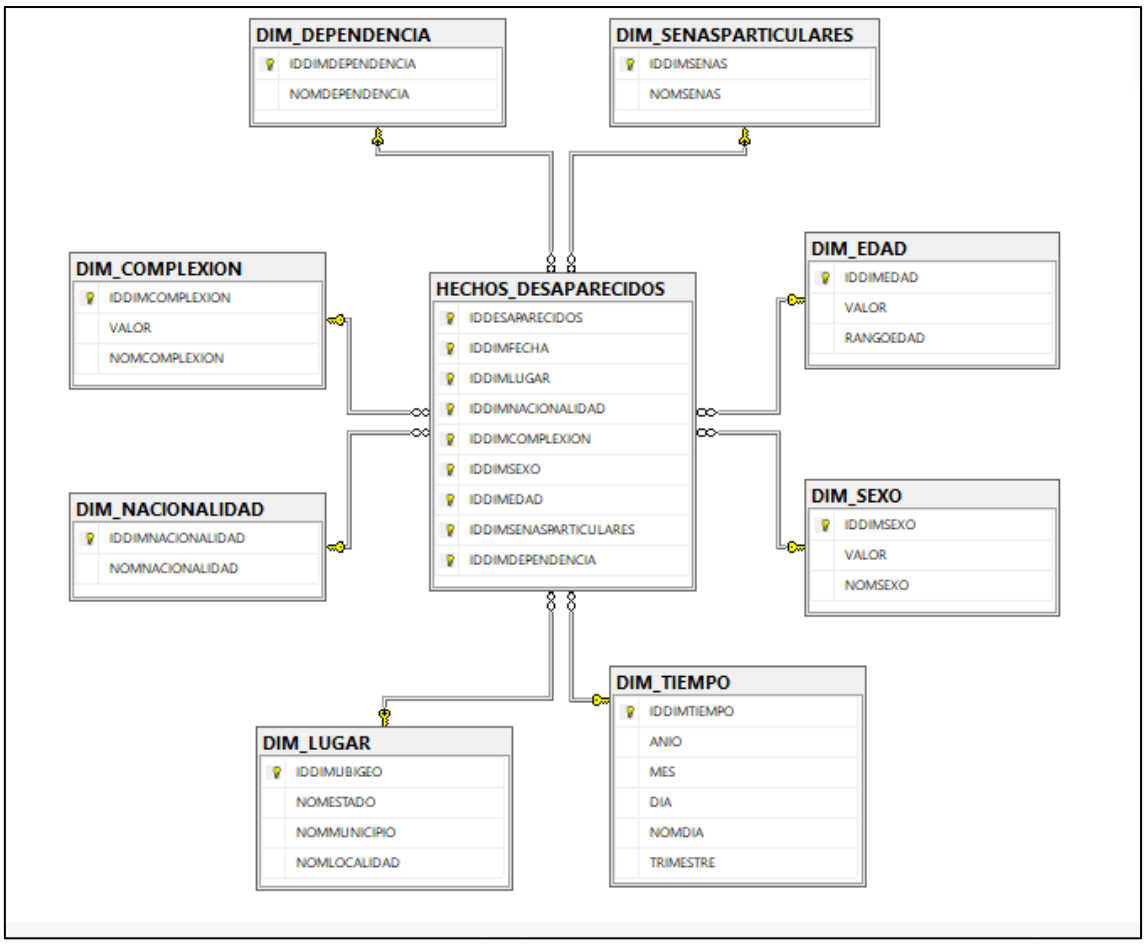


Figura 62. Pantalla del modelo dimensional generado desde Microsoft SQL Server

IV.3.3 Transformación de datos

En este apartado se definió los valores para aplicar la regresión logística binaria donde los valores asignados fueron HOMBRE=1 y MUJER=0.

```
SELECT
SEX.VALOR AS SEXO,COMPL.VALOR AS COMPLEXION,EDAD.VALOR
AS EDAD,
(CASE WHEN SP.NOMSENAS = 'ACNE Y VERRUGA' THEN 1 ELSE 0 END)
AS 'ACNE Y VERRUGA',
(CASE WHEN SP.NOMSENAS = 'BIGOTE' THEN 1 ELSE 0 END)
AS 'BIGOTE',
(CASE WHEN SP.NOMSENAS = 'CICATRICES' THEN 1 ELSE 0 END)
AS 'CICATRICES',
(CASE WHEN SP.NOMSENAS = 'LUNARES' THEN 1 ELSE 0 END)
AS 'LUNARES',
(CASE WHEN SP.NOMSENAS = 'TATUAJES' THEN 1 ELSE 0 END)
AS 'TATUAJES',
(CASE WHEN SP.NOMSENAS = 'PROBLEMAS EN LA DENTADURA' THEN 1
ELSE 0 END)
AS 'PROBLEMAS EN LA DENTADURA',
(CASE WHEN SP.NOMSENAS = 'OTRAS SENAS PARTICULARES' THEN 1
ELSE 0 END)
AS 'OTRAS SENAS PARTICULARES',
(CASE WHEN SP.NOMSENAS = 'NO ESPECIFICADO' THEN 1 ELSE 0 END)
AS 'NO ESPECIFICADO'
FROM HECHOS_DESAPARECIDOS HD
INNER JOIN DIM_TIEMPO TIE
ON TIE.IDDIMTIEMPO = HD.IDDIMFECHA
INNER JOIN DIM_NACIONALIDAD NAC
ON NAC.IDDIMNACIONALIDAD = HD.IDDIMNACIONALIDAD
INNER JOIN DIM_LUGAR LUG
ON LUG.IDDIMUBIGEO = HD.IDDIMLUGAR
INNER JOIN DIM_COMPLEXION COMPL
ON COMPL.IDDIMCOMPLEXION = HD.IDDIMCOMPLEXION
INNER JOIN DIM_SEXO SEX
ON SEX.IDDIMSEXO = HD.IDDIMSEXO
INNER JOIN DIM_EDAD EDAD
ON EDAD.IDDIMEDAD = HD.IDDIMEDAD
INNER JOIN DIM_SENASPARTICULARES SP
ON SP.IDDIMSENAS = HD.IDDIMSENASPARTICULARES
WHERE COMPL.NOMCOMPLEXION != 'NO ESPECIFICADO'
AND LUG.NOMESTADO = 'AGUASCALIENTES'
```

En este segundo apartado se definió los valores para aplicar la regresión logística binaria donde los valores asignados fueron HOMBRE=0 y MUJER=1.

```
SELECT
(CASE WHEN SEX.NOMSEXO = 'MUJER' THEN 1 ELSE 0 END) AS
'SEXO',COMPL.VALOR
AS COMPLEXION,EDAD.VALOR AS EDAD,
(CASE WHEN SP.NOMSENAS = 'ACNE Y VERRUGA' THEN 1 ELSE 0 END)
AS 'ACNE Y VERRUGA',
(CASE WHEN SP.NOMSENAS = 'BIGOTE' THEN 1 ELSE 0 END)
AS 'BIGOTE',
(CASE WHEN SP.NOMSENAS = 'CICATRICES' THEN 1 ELSE 0 END)
AS 'CICATRICES',
(CASE WHEN SP.NOMSENAS = 'LUNARES' THEN 1 ELSE 0 END)
AS 'LUNARES',
(CASE WHEN SP.NOMSENAS = 'TATUAJES' THEN 1 ELSE 0 END)
AS 'TATUAJES',
(CASE WHEN SP.NOMSENAS = 'PROBLEMAS EN LA DENTADURA' THEN 1
ELSE 0 END)
AS 'PROBLEMAS EN LA DENTADURA',
(CASE WHEN SP.NOMSENAS = 'OTRAS SENAS PARTICULARES' THEN 1
ELSE 0 END)
AS 'OTRAS SENAS PARTICULARES',
(CASE WHEN SP.NOMSENAS = 'NO ESPECIFICADO' THEN 1 ELSE 0 END)
AS 'NO ESPECIFICADO'
FROM HECHOS_DESAPARECIDOS HD
INNER JOIN DIM_TIEMPO TIE
ON TIE.IDDIMTIEMPO = HD.IDDIMFECHA
INNER JOIN DIM_NACIONALIDAD NAC
ON NAC.IDDIMNACIONALIDAD = HD.IDDIMNACIONALIDAD
INNER JOIN DIM_LUGAR LUG
ON LUG.IDDIMUBIGEO = HD.IDDIMLUGAR
INNER JOIN DIM_COMPLEXION COMPL
ON COMPL.IDDIMCOMPLEXION = HD.IDDIMCOMPLEXION
INNER JOIN DIM_SEXO  SEX
ON SEX.IDDIMSEXO = HD.IDDIMSEXO
INNER JOIN DIM_EDAD EDAD
ON EDAD.IDDIMEDAD = HD.IDDIMEDAD
INNER JOIN DIM_SENASPARTICULARES SP
ON SP.IDDIMSENAS = HD.IDDIMSENASPARTICULARES
WHERE COMPL.NOMCOMPLEXION != 'NO ESPECIFICADO'
AND LUG.NOMESTADO = 'BAJA CALIFORNIA'

---Query para asignar valores 0 a los campos que no tiene valor alguno

DECLARE @IDMIN INT,@IDMAX INT
SELECT @IDMIN=MIN(IDESTADO),@IDMAX=MAX(IDESTADO) FROM
ESTADO
DECLARE @NOMBREESTADO VARCHAR(100)
```

```
DECLARE @MESMIN INT,@MESMAX INT
SELECT @MESMIN=MIN(MES),@MESMAX=MAX(MES) FROM
PREDICCIONES
DECLARE @ANIOMIN INT,@ANIOMAX INT
SELECT @ANIOMIN=MIN(ANIO),@ANIOMAX=MAX(ANIO) FROM
PREDICCIONES
WHILE @IDMIN <= @IDMAX
BEGIN
	SET @NOMBREESTADO = (SELECT NOMESTADO FROM ESTADO
WHERE IDESTADO = @IDMIN)
	WHILE @ANIOMIN <= @ANIOMAX
		BEGIN
			WHILE @MESMIN <= @MESMAX
				BEGIN
					IF EXISTS (SELECT 1 FROM
					PREDICCIONES WHERE NOMESTADO =
					@NOMBREESTADO AND ANIO =
					@ANIOMIN AND MES=@MESMIN)
						BEGIN
							SET @MESMIN = @MESMIN+1

						END;
					ELSE
						BEGIN
							INSERT INTO PREDICCIONES
							VALUES
							(@NOMBREESTADO,@ANIOMI
							N,@MESMIN,0);
							SET @MESMIN = @MESMIN+1
						END;
				END;
			SET @MESMIN = (SELECT MIN(MES) FROM
PREDICCIONES)
			SET @ANIOMIN = @ANIOMIN+1
		END;
	SET @ANIOMIN = (SELECT MIN(ANIO) FROM PREDICCIONES)
	SET @IDMIN = @IDMIN+1
END
PRINT @MESMIN
PRINT @ANIOMIN
PRINT @IDMIN
SELECT NOMESTADO,CONCAT(ANIO,RIGHT(100+MES, 2)) AS [AÑO-
MES],TOTAL FROM PREDICCIONES ORDER BY NOMESTADO,ANIO,MES
ASC;

--Fin del query para asignar valores 0
```

Métodos de pronóstico

```
--Creación de tabla para el total

CREATE TABLE TOTALDESAPARECIDOS(
IDTOTALDESAPARECIDOS INT IDENTITY(1,1) NOT NULL,
ANIOMES INT NOT NULL,
TOTAL INT NOT NULL,
PRIMARY KEY (IDTOTALDESAPARECIDOS));


--Inserción de datos a la tabla

INSERT TOTALDESAPARECIDOS
SELECT CONCAT(ANIO,RIGHT(100+MES, 2)) AS [AÑO-MES],SUM(TOTAL)
AS TOTAL FROM PREDICCIONES
GROUP BY ANIO,MES,TOTAL
ORDER BY ANIO,MES,TOTAL ASC;


SELECT ANIOMES,SUM(TOTAL) AS TOTAL FROM TOTALDESAPARECIDOS
GROUP BY ANIOMES
ORDER BY ANIOMES ASC;

--Creación de tabla para la predicción por estado

CREATE TABLE PREDICCIONES(
IDPREDICCION INT IDENTITY(1,1) NOT NULL,
NOMESTADO VARCHAR(100) NOT NULL,
ANIO INT NOT NULL,
MES INT NOT NULL,
TOTAL INT NOT NULL,
PRIMARY KEY (IDPREDICCION));

--Inserción de datos a la tabla

INSERT PREDICCIONES
SELECT  LUG.NOMESTADO,TIE.ANIO,TIE.MES,COUNT(LUG.NOMESTADO)
AS TOTAL FROM HECHOS_DESAPARECIDOS HD
INNER JOIN DIM_LUGAR LUG
ON LUG.IDDIMUBIGEO = HD.IDDIMLUGAR
INNER JOIN DIM_TIEMPO TIE
ON TIE.IDDIMTIEMPO = HD.IDDIMFECHA
INNER JOIN DIM_COMPLEXION COMPL
ON COMPL.IDDIMCOMPLEXION = HD.IDDIMCOMPLEXION
WHERE COMPL.NOMCOMPLEXION!= 'NO ESPECIFICADO'
GROUP BY LUG.NOMESTADO,TIE.ANIO,TIE.MES
ORDER BY LUG. NOMESTADO, TIE.ANIO, TIE. MES, COUNT (LUG.
NOMESTADO) ASC;
```

```
SELECT * FROM ESTADO

SELECT CONCAT(ANIO,RIGHT(100+MES, 2)) AS [AÑO-MES],TOTAL
FROM PREDICCIONES WHERE NOMESTADO = 'AGUASCALIENTES'
ORDER BY NOMESTADO,ANIO,MES ASC;
```

IV.4 Fase 4: Modelado

IV.4.1 Selección de técnica de modelado

De acuerdo al presente estudio, se usó el entorno de desarrollo RStudio el cual cuenta con diversas técnicas y métodos que se ajustaron al mismo. La técnica que se acomodó más al fue la regresión logística binaria y usando las diversas series de tiempo como el modelo ARIMA, Holt-Winters y ETS, los cuales se compararon para seleccionar el mejor modelo para el pronóstico.

IV.4.2 Selección de datos de prueba

El modelado tendrá las siguientes observaciones, los datos serán de México el cual tendrá una variación entre los años 2007-2018. El país está dividido en 32 estados los cuales se estudiarán individualmente para lograr un mejor resultado y predicción, por lo que en alguno de estos estados los años varían en diverso intervalo de tiempo por lo que las predicciones mediante las series de tiempo pueden variar en los años a estudiar. El pronóstico dentro de las líneas de tiempo se dará por los 12 meses posteriores de acuerdo a que año termine el mismo.

Tabla 32 *Intervalo de años por los estados y el total de casos*

N°	Estado	Intervalo de años	Total de casos
0	País de México	2007/01 - 2018/04	24,476
1	Aguas Calientes	2007/01 - 2018/04	154
2	Baja California	2007/01 - 2018/01	575
3	Baja California Sur	2008/01 - 2016/05	30
4	Campeche	2008/01 - 2016/10	27
5	Chiapas	2011/01 - 2018/02	74
6	Chihuahua	2007/01 - 2018/03	1,489
7	Ciudad de México	2007/01 - 2017/12	509
8	Coahuila de Zaragoza	2007/01 - 2017/11	1,117
9	Colima	2008/01 - 2018/04	406
10	Durango	2007/01 - 2018/04	307

N°	Estado	Intervalo de años	Total, de casos
11	Estado de México	2007/01 – 2018/03	2,435
12	Guanajuato	2007/01 - 2018/03	440
13	Guerrero	2007/01 - 2018/04	933
14	Hidalgo	2017/01 - 2017/01	94
15	Jalisco	2007/01 - 2018/03	2,254
16	Michoacán	2007/01 - 2018/04	851
17	Morelos	2009/01 – 2018/03	159
18	Nayarit	2008/01 - 2018/04	83
19	Nuevo León	2007/01 - 2018/04	1,950
20	Oaxaca	2007/01 - 2015/04	138
21	Puebla	2007/01 - 2018/04	1,280
22	Querétaro	2007/01 - 2018/04	196
23	Quintana Roo	2007/01 - 2017/10	37
24	San Luis de Potosí	2010/01 - 2015/10	72
25	Sinaloa	2007/01 - 2018/04	2,008
26	Sonora	2007/01 - 2018/01	1,379
27	Tabasco	2007/01 - 2016/10	53
28	Tamaulipas	2007/01 - 2016/12	4,589
29	Tlaxcala	2009/01 - 2018/04	19
30	Veracruz	2007/01 - 201509	392
31	Yucatán	2007/01 - 2018/04	53
32	Zacatecas	2007/01 - 2018/04	373

IV.4.3 Obtención del modelo

A continuación, se detalla al país de México en general con la técnica de regresión logística binaria y las diversas series de tiempo usadas como el modelo ARIMA, Holt-Winters y ETS. Para los resultados ir al anexo 3 y 4 donde se encuentran detalladamente cada uno de ellos.

- **País México**

  A. Regresión logística binaria

  Con los siguientes comandos se obtendrá el modelo binomial. Asimismo, se recalca que la variable dependiente es el sexo masculino definido con 1 y femenino con 0.

  #Visualizar la cantidad en la variable dependiente

  Comando → table(PAISMEXICO$SEXO)

0	1
6375	18101

#Modelo binomial

Comando → MODELOPAISMEXICO= glm (SEXO~.,data= PAISMEXICO,family = binomial)

Se utiliza el comando summari para la visualización de la tabla binomial donde se muestra la desviación residual y otros valores a analizar.

#Tabla binomial

Comando → summary(MODELOPAISMEXICO)

Desviación Residual:

Tabla 33 *Regresión logística binaria - desviación residual*

Min	1Q	Median	3Q	Max
-2.3752	-1.0022	0.5593	0.7065	1.3952

#Tabla de Coeficientes

Comando → summary(MODELOPAISMEXICO)

Tabla 34 *Regresión logística binaria - tabla de coeficientes*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.005587	0.053561	-18.774	< 2e-16 ***
COMPLEXION	-0.003468	0.018089	-0.192	0.8479
EDAD	0.585308	0.014871	39.358	< 2e-16 ***
ACNEVERRUGA	0.318173	0.129203	2.463	0.0138 *
BIGOTE	0.841862	0.08139	10.344	< 2e-16 ***
CICATRICES	0.544793	0.047818	11.393	< 2e-16 ***
LUNARES	-0.064562	0.04879	-1.323	0.1857
TATUAJES	1.072715	0.056344	19.039	< 2e-16 ***
PROBLEMASDENTADURA	0.443868	0.08625	5.146	2.66e-07 ***
OTRASSENAS	0.347461	0.055621	6.247	4.19e-10 ***
NOESPECIFICADO	NA	NA	NA	NA

Seleccionando las variables que más influyen en el modelo final para el país de México quedaría de la siguiente forma:

#Modelo binomial 1

Comando → MODELOPAISMEXICO1=glm (SEXO~EDAD + ACNEVERRUGA + BIGOTE+ CICATRICES+ TATUAJES+ PROBLEMASDENTADURA+ OTRASSENAS,data =PAISMEXICO,family = binomial)

#Tabla binomial 1

Comando → summary(MODELOPAISMEXICO1)

Desviación Residual:

Tabla 35 Regresión logística binaria - desviación residual 01

Min	1Q	Median	3Q	Max
-2.3736	-0.9977	0.5601	0.6977	1.3684

#Tabla de Coeficientes

Tabla 36 *Regresión logística binaria - tabla de coeficientes 01*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.02309	0.04609	-22.196	< 2e-16 ***
EDAD	0.58457	0.01475	39.629	< 2e-16 ***
ACNEVERRUGA	0.33162	0.12875	2.576	0.01 *
BIGOTE	0.85565	0.08071	10.602	< 2e-16 ***
CICATRICES	0.55838	0.04665	11.969	< 2e-16 ***
TATUAJES	1.08674	0.05535	19.636	< 2e-16 ***
PROBLEMASDENTADURA	0.45794	0.08559	5.35	8.78e-08 ***
OTRASSENAS	0.36131	0.05463	6.614	3.73e-11 ***

LOGIT = -1.02309 + 0.58457EDAD+ 0.33162ACNEVERRUGA+ 0.85565BIGOTE+ 0.55838CICATRICES+ 1.08674TATUAJES+ 0.45794PROBLEMASDENTADURA+ 0.36131OTRASSENAS

Formula final con los ítems que más afectan al país de México:

$$P(X=1)=\frac{e^{-1.02309 + 0.58457EDAD+ 0.33162ACNEVERRUGA+ 0.85565BIGOTE + 0.55838CICATRICES+ 1.08674TATUAJES + 0.45794PROBLEMASDENTADURA+ 0.36131OTRASSENAS}}{1+e^{-1.02309 + 0.58457EDAD+ 0.33162ACNEVERRUGA+ 0.85565BIGOTE + 0.55838CICATRICES+ 1.08674TATUAJES + 0.45794PROBLEMASDENTADURA+ 0.36131OTRASSENAS}}$$

B. Modelo ARIMA

Dentro de este modelo se desarrolló dos pruebas como son la Dickey-Fuller y la de ruido blanco, con las cuales se prueba si existe la estacionalidad y la validez del modelo respectivamente. Partiendo con el modelo en primer lugar se tiene que definir la serie de tiempo, quiere decir que los datos tienen que ser TS (Serie de Tiempo)

Con estos comandos en primer lugar se selecciona la data a analizar, siguiente se realiza la serie de tiempo con (ts) a la vez se ingresa los años y el mes y la frecuencia de la data a analizar.

```
#Definir la serie de tiempo(ts)
```

Comando → `CANTPAISMEXICO.ts <- ts(CANTPAISMEXICO, start = c(2007,1), frequency = 12)`

Con este comando se indica la cantidad de valores por los años establecidos y meses.

```
#Vista de los datos seleccionados por los años definidos
```

Comando → `CANTPAISMEXICO.ts`

Tabla 37 *Modelo ARIMA - serie de tiempo*

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	57	28	28	29	48	45	33	47	26	37	34	32
2008	72	31	35	45	42	49	58	56	51	65	45	37
2009	85	78	84	58	72	93	85	81	105	97	84	54
2010	116	123	215	196	218	232	276	236	162	179	193	157
2011	218	287	331	336	243	280	234	251	247	191	152	184
2012	216	181	207	196	235	211	207	188	204	190	155	165
2013	166	142	202	234	233	198	200	225	267	251	217	200
2014	225	222	257	239	248	246	212	246	216	199	167	168
2015	172	178	234	168	207	178	144	192	190	192	193	142
2016	207	210	244	309	288	299	310	266	236	230	221	234
2017	295	276	282	276	309	283	297	260	261	321	252	241
2018	313	320	292	156								

Con estos comandos se tiene la descripción del año en el que empieza y termina la serie de tiempo (ts).

#Descripcion del año en el que empieza y termina la (ts)

Comando → ts_info(CANTPAISMEXICO.ts)

La CANTPAISMEXICO.ts es un objeto de serie de tiempo con 1 variable y 136 observaciones

Frecuencia: 12
Fecha de inicio: 2007 / 1
Fecha fin: 2018 / 4

#Gráficas de los datos

Comando → autoplot(CANTPAISMEXICO.ts, col="blue") + ggtitle("PAÍS DE MÉXICO") + ylab("Total de Desaparecidos") + xlab("Total de Años")

Figura 63. Pantalla del modelo ARIMA - datos Iniciales

B.1 Prueba de dickey-fuller

Para probar que el modelo es válido es decir, si es estacionario se tiene que cumplir con una regla la cual es:

No es estacionario > 0.5
Es estacionario < 0.5

#Prueba de Dickey Fuller

Comando → adf.test(CANTPAISMEXICO.ts,alternative = "stationary")

#Resultado

```
        Augmented Dickey-Fuller Test

data:  CANTPAISMEXICO.ts
Dickey-Fuller = -2.4454, Lag order = 5, p-value = 0.3912
alternative hypothesis: stationary
```

Figura 64. Pantalla del modelo ARIMA – resultado de la prueba de Dickey Fuller

En la figura 64, se obtuvo como resultado de 0.3912. En este sentido, la prueba de Dickey Fuller resulto ser menor a valor permitido de 0.5, concluyendo que es estacionario y es válido.

#Vista De La Diferencia

Comando → autoplot(DIFERENCIAPAISMEXICO) + ggtitle("DIFERENCIA PAÍS DE MÉXICO") + ylab("Total de Desaparecidos") + xlab("Total de Años")

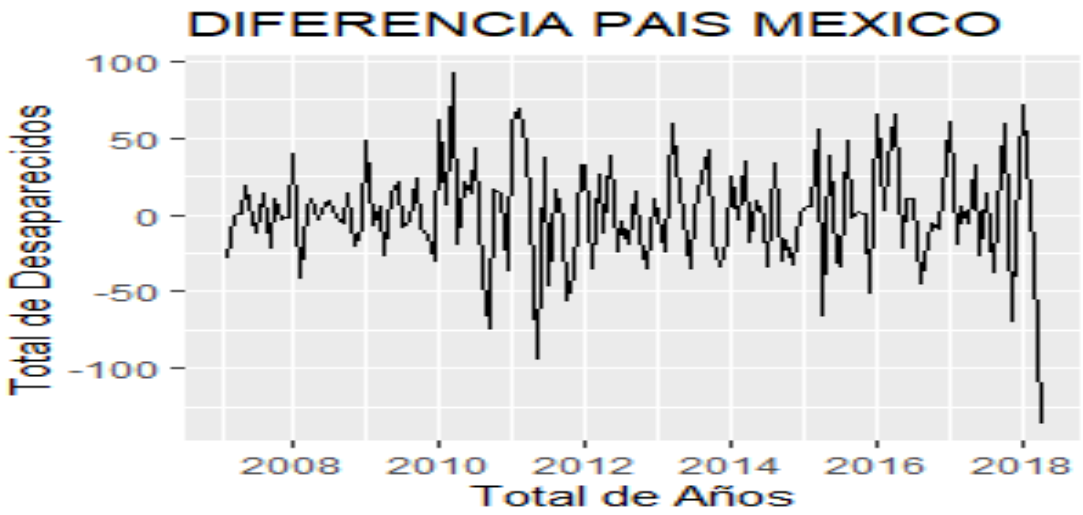


Figura 65. Pantalla del modelo ARIMA - diferencia

#Para obtener el mejor modelo ARIMA, se ejecutó:

Comando→MODELOPAISMEXICO <-auto.arima (CANTPAISMEXICO.ts, trace = T)

#Resultado

Best model: ARIMA (1,1,1)(1,0,0)

#Vizualizar el modelo final

Comando → print(MODELOPAISMEXICO)

```
Series: CANTPAISMEXICO.ts
ARIMA(1,1,1)(1,0,0)[12]

Coefficients:
          ar1      ma1     sar1
      0.5451  -0.7846  0.1990
s.e.  0.1907   0.1394  0.0925

sigma^2 estimated as 1091:  log likelihood=-662.51
AIC=1333.02   AICc=1333.33   BIC=1344.65
```

Figura 66. Pantalla del modelo ARIMA – librería auto.arima

En la figura 66, se obtuvo los resultados del mejor modelo ARIMA seleccionado mediante la librería auto.arima el cual nos arrojó que el modelo final para el pronóstico es el de (1,1,1)(1,0,0).

#Vizualizar el modelo seleccionado

Comando → checkresiduals**(**MODELOPAISMEXICO**)**

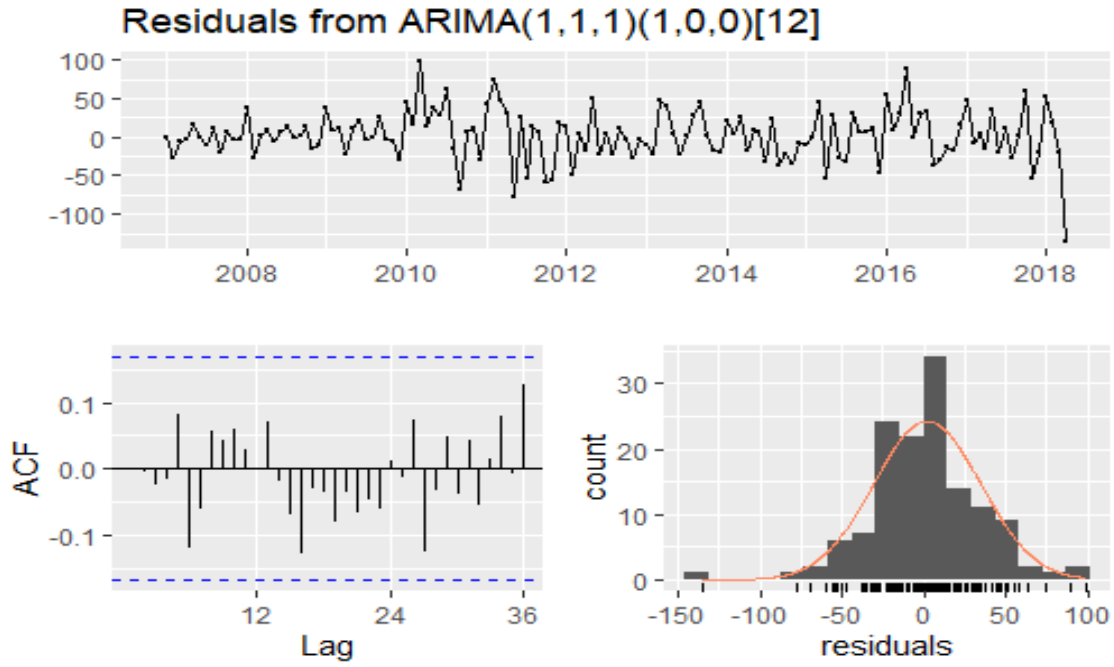


Figura 67. Pantalla del modelo ARIMA - gráfico de residuos

B.2 Prueba de ruido blanco

Para probar que existe ruido blanco (o es válido el modelo) tiene que cumplir con unas características que son:

Ruido blanco > 0.5
No hay ruido blanco < 0.5

Esta prueba dentro de RStudio se llama Ljung Box Test, tiene que existir el ruido blanco para realizar el pronóstico.

#Prueba de Ljun-Box

Comando → Box.test(residuals(MODELOPAISMEXICO), type = "Ljung-Box")

```
        Box-Ljung test

data:  residuals(MODELOPAISMEXICO)
X-squared = 0.0015702, df = 1, p-value = 0.9684
```

Figura 68. Pantalla del modelo ARIMA – prueba de ruido blanco

En la figura 68 se muestra que el modelo es mayor a 0.5 quiere decir que tiene ruido blanco por ende el modelo es válido.

#Vizualizar gráficamente la prueba Ljung Box Test

Comando → tsdiag (MODELOPAISMEXICO)

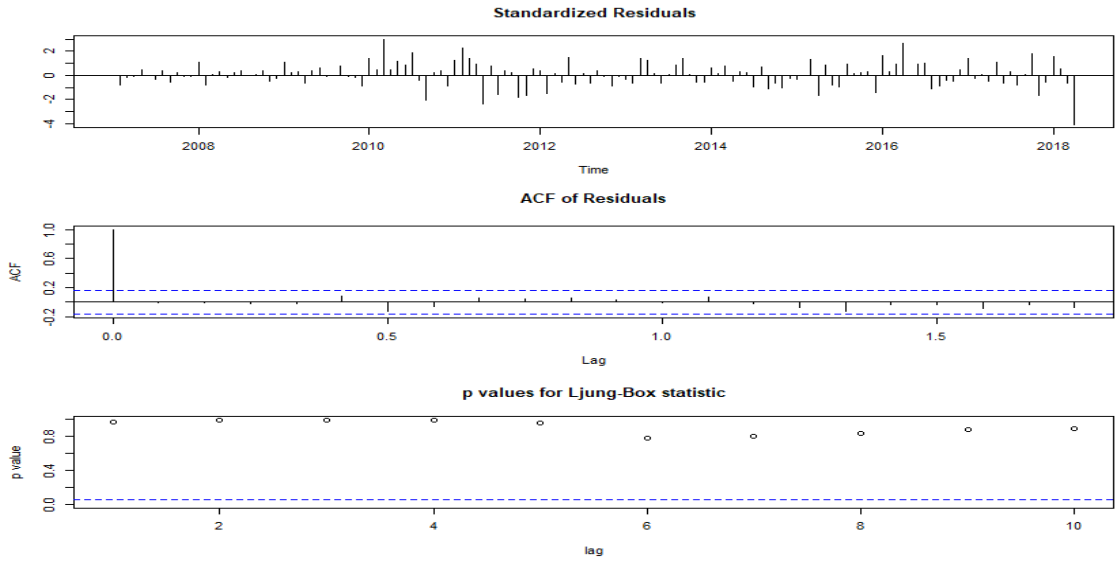


Figura 69 Pantalla del modelo ARIMA – prueba de Test-LjungBox

En este sentido y teniendo estos resultados se procede a generar el error para visualizar si esta media es igual a cero

#Generar el error

Comando → error = residuals(MODELOPAISMEXICO)plot(error)

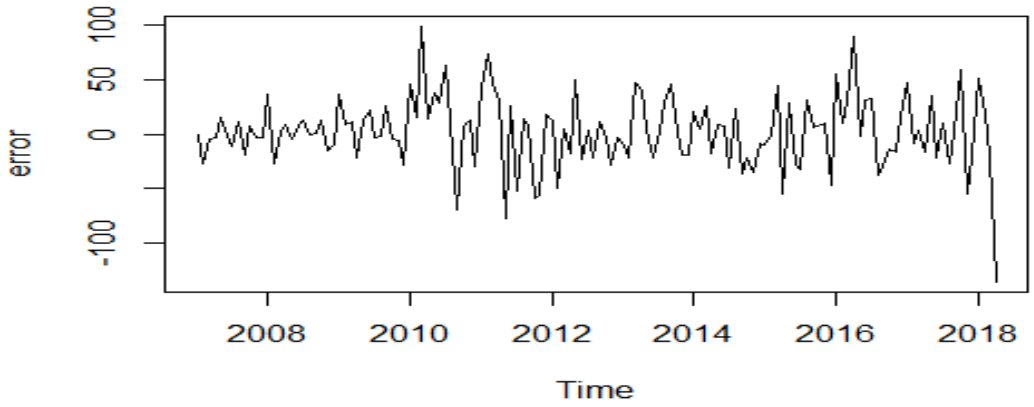


Figura 70. Pantalla del modelo ARIMA - medir el error

B.3 Pronóstico

#Gráfico del pronóstico

Comando → autoplot(PREDICCIONPAISMEXICO) + ggtitle("PRONÓSTICO PAÍS DE MÉXICO") + ylab("Total de Desaparecidos") +xlab("Total de Años")

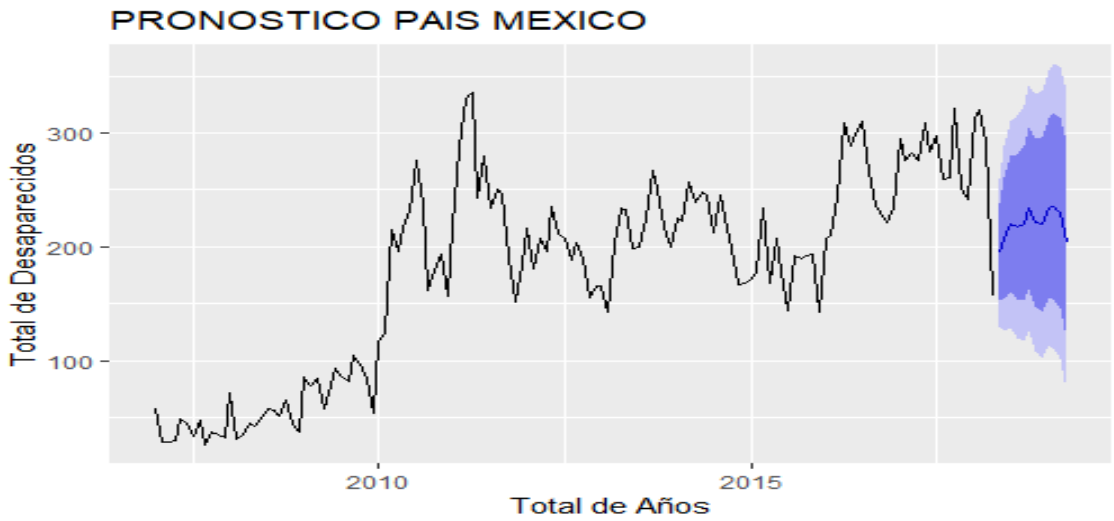


Figura 71. Pantalla del modelo ARIMA - pronóstico

#Ajuste entre el pronóstico y los datos originales

Comando → autoplot(PREDICCIONPAISMEXICO)+ autolayer(fitted (PREDICCIONPAISMEXICO), series="PRONÓSTICO")+ ggtitle("AJUSTE ENTRE EL PRONÓSTICO Y LA SERIE DE TIEMPO") + ylab("Total de Desaparecidos") +xlab("Total de Años")

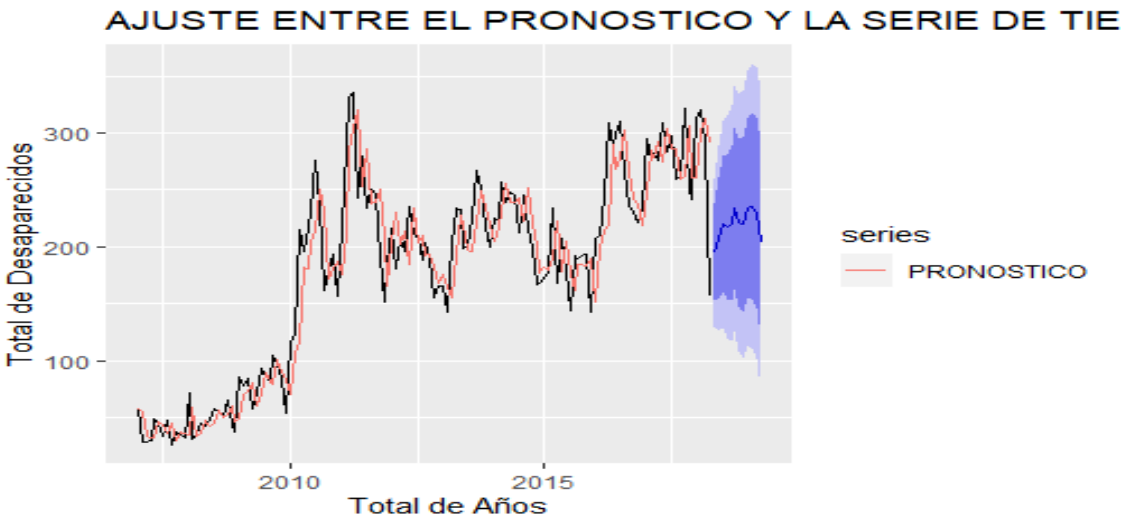


Figura 72. Pantalla del modelo ARIMA - ajuste entre el pronóstico y los datos originales

#Visualizar el resumen total del modelo

Comando →print(summary(PREDICCIONPAISMEXICO))

```
Forecast method: ARIMA(1,1,1)(1,0,0)[12]

Model Information:
Series: CANTPAISMEXICO.ts
ARIMA(1,1,1)(1,0,0)[12]

Coefficients:
         ar1      ma1    sar1
      0.5451  -0.7846  0.1990
s.e.  0.1907   0.1394  0.0925

sigma^2 estimated as 1091:  log likelihood=-662.51
AIC=1333.02   AICc=1333.33   BIC=1344.65

Error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 1.993372 32.53905 24.20511 -1.633303 16.26545 0.4854331 -0.003360735
```

Figura 73. Pantalla del modelo ARIMA – resumen del modelo

#Visualización de los resultados del pronóstico

Comando → RESULTADOSPAISMEXICO **<-**as.data.frame **(**PREDICCIONPAISMEXICO**)**

#Resultado

Comando → RESULTADOSPAISMEXICO

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
May 2018	194.8838	152.5563	237.2114	130.14939	259.6183
Jun 2018	207.3230	154.1478	260.4983	125.99852	288.6476
Jul 2018	219.7104	160.2266	279.1942	128.73774	310.6830
Aug 2018	217.5791	153.5676	281.5907	119.68194	315.4763
Sep 2018	220.6305	152.9425	288.3185	117.11068	324.1503
Oct 2018	234.1273	163.2242	305.0305	125.69029	342.5644
Nov 2018	221.2413	147.4003	295.0824	108.31123	334.1714
Dec 2018	219.5138	142.9186	296.1091	102.37146	336.6562
Jan 2019	234.0961	154.8794	313.3129	112.94452	355.2477
Feb 2019	235.6266	153.8918	317.3614	110.62407	360.6291
Mar 2019	230.1284	145.9612	314.2956	101.40583	358.8510
Apr 2019	203.1004	116.5746	289.6262	70.77064	335.4302

Figura 74. Pantalla del modelo ARIMA – *Lo80 y Lo95*

Lo80: Límite Inferior con 80% de confianza
Hi80: Limite Superior con 80% de confianza
Lo95: Límite Inferior con 95% de confianza
Hi95: Limite Superior con 95% de confianza

Como se visualiza en los resultados el mes con mayor cantidad de posibles desaparecidos es el de mes de febrero del 2019 con 238.1284 posibles casos y el menor mes en mayo del 2018 con 194.8838 de posibles casos.

C. Modelo Holt-Winters

Para comenzar este modelo se utilizó los siguientes comandos:

#Ajuste entre modelo Holt-Winters y los datos originales

Este comando se usó con la serie de tiempo ya definida en el modelo anterior (ARIMA).

Comando → MODELOPAISMEXICO **=** HoltWinters**(**CANTPAISMEXICO.ts, seasonal **=** "additive"**)**

#Gráfico de la relación entre el modelo y los datos

Comando → plot(MODELOPAISMEXICO)

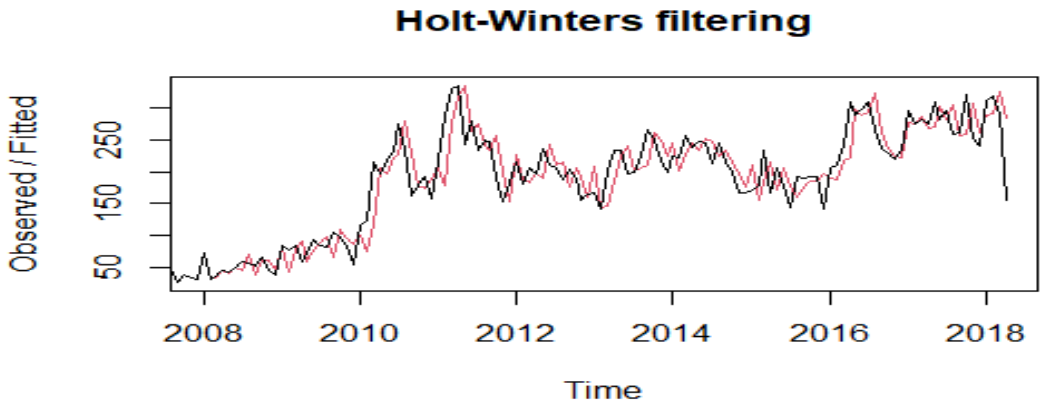


Figura 75. Pantalla del modelo Holt-Winters gráfico del modelo

Leyenda:

- Línea negra: Data original
- Línea roja: Modelo Holt-Winters

C.1 Datos Fitted

#Tabla fitted

Comando → plot(fitted(MODELOPAISMEXICO))

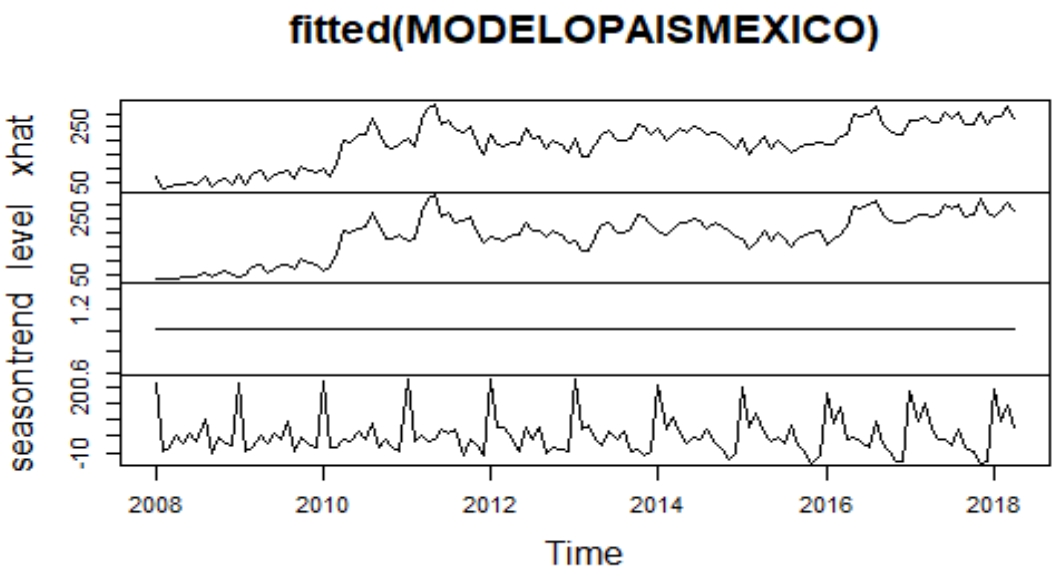


Figura 76. Pantalla del modelo Holt-Winters gráfico fitted

#Información de los datos Fitted

Comando → fitted(MODELOPAISMEXICO)

Tabla 38 *Modelo Holt-Winters datos fitted*

Month	xhat	level	trend	season
Jan 2008	69.17254	35.54482	1.016608	32.6111111
Feb 2008	30.17966	38.96861	1.016608	-9.8055556
Mar 2008	34.478	40.68362	1.016608	-7.2222222
Apr 2008	43.73068	42.14463	1.016608	0.5694444
May 2008	41.20293	44.24188	1.016608	-4.0555556
Jun 2008	49.23146	45.93708	1.016608	2.2777778
Jul 2008	45.05101	46.75663	1.016608	-2.7222222
Aug 2008	70.34184	58.79745	1.016608	10.5277778
Sep 2008	37.73175	47.60403	1.016608	-10.8888889
Oct 2008	60.08605	59.91666	1.016608	-0.8472222
Nov 2008	61.86952	65.1168	1.016608	-4.2638889
Dec 2008	46.60747	51.77142	1.016608	-6.1805556
Jan 2009	78.45714	44.60863	1.016608	32.8319008
Feb 2009	42.47066	51.19555	1.016608	-9.7414972
Mar 2009	76.29548	82.46033	1.016608	-7.1814608
Apr 2009	91.72141	90.03624	1.016608	0.6685624
May 2009	59.36716	62.34387	1.016608	-3.9933145
Jun 2009	77.39185	74.11554	1.016608	2.2597033
Jul 2009	87.72581	88.42026	1.016608	-1.7110655
Aug 2009	97.5407	87.11624	1.016608	9.4078563
Sep 2009	65.2146	74.05079	1.016608	-9.8528014
Oct 2009	109.49211	108.939	1.016608	-0.4635031
Nov 2009	94.75578	99.32036	1.016608	-5.5811911
Dec 2009	85.26578	91.17996	1.016608	-6.9307808
Jan 2010	99.93763	65.5782	1.016608	33.3428176
Feb 2010	74.31915	80.26963	1.016608	-6.9670926
Mar 2010	117.16783	122.73105	1.016608	-6.5798322
Apr 2010	206.08976	207.03782	1.016608	-1.9646656
May 2010	197.4742	199.46443	1.016608	-3.0068453
Jun 2010	222.45096	217.95584	1.016608	3.478508
Jul 2010	226.19479	227.1021	1.016608	-1.9239175
Aug 2010	279.65359	270.52075	1.016608	8.1162309
Sep 2010	228.6431	234.37255	1.016608	-6.7460506
Oct 2010	178.22967	178.65204	1.016608	-1.4389832
Nov 2010	174.92	180.32448	1.016608	-6.4210852
Dec 2010	188.37798	196.73363	1.016608	-9.3722543
Jan 2011	206.65005	171.03635	1.016608	34.5970913
Feb 2011	179.56672	181.71582	1.016608	-3.1657162
Mar 2011	276.27284	274.19657	1.016608	1.0596583
Apr 2011	320.06962	321.80556	1.016608	-2.752552
May 2011	335.9972	336.38462	1.016608	-1.4040322
Jun 2011	263.46814	258.22736	1.016608	4.224171
Jul 2011	276.30036	273.31849	1.016608	1.9652575

Month	xhat	level	trend	season
Aug 2011	244.0464	238.32237	1.016608	4.707422
Sep 2011	234.32553	245.25898	1.016608	-11.9500581
Oct 2011	256.70387	257.06609	1.016608	-1.3788299
Nov 2011	198.15256	202.14521	1.016608	-5.0092596
Dec 2011	153.06361	163.86949	1.016608	-11.8224894
Jan 2012	227.72402	191.22403	1.016608	35.4833826
Feb 2012	188.49942	182.2593	1.016608	5.2235031
Mar 2012	183.24101	176.89123	1.016608	5.3331773
Apr 2012	197.64325	198.13523	1.016608	-1.5085849
May 2012	190.10348	197.75284	1.016608	-8.6659706
Jun 2012	243.52416	236.99244	1.016608	5.5151059
Jul 2012	209.99809	210.31936	1.016608	-1.3378808
Aug 2012	215.05055	208.78352	1.016608	5.2504126
Sep 2012	176.82672	186.77045	1.016608	-10.9603376
Oct 2012	205.42835	210.92123	1.016608	-6.509495
Nov 2012	191.2062	198.8028	1.016608	-8.6132077
Dec 2012	160.60486	168.99499	1.016608	-9.4067373
Jan 2013	209.33792	173.75343	1.016608	34.5678808
Feb 2013	143.52847	137.87397	1.016608	4.6378908
Mar 2013	145.79438	137.58931	1.016608	7.1884622
Apr 2013	185.8367	186.457	1.016608	-1.6369027
May 2013	224.3343	228.47779	1.016608	-5.160104
Jun 2013	240.86399	236.87201	1.016608	2.9753689
Jul-13	200.84065	201.39604	1.016608	-1.5719949
Aug 2013	205.85166	201.69695	1.016608	3.1380974
Sep 2013	211.19382	219.01565	1.016608	-8.8384381
Oct 2013	260.84562	267.54327	1.016608	-7.7142593
Nov 2013	249.75388	260.17774	1.016608	-11.4404672
Dec 2013	225.26216	233.30908	1.016608	-9.0635307
Jan 2014	245.01889	212.81856	1.016608	31.1837216
Feb 2014	202.32708	196.79194	1.016608	4.5185359
Mar 2014	227.15127	214.55723	1.016608	11.5774308
Apr 2014	244.12645	240.98578	1.016608	2.1240591
May 2014	234.17114	237.63795	1.016608	-4.4834191
Jun 2014	251.07269	250.42786	1.016608	-0.371782
Jul 2014	246.50477	247.1258	1.016608	-1.6376396
Aug 2014	224.41647	218.76651	1.016608	4.6333476
Sep 2014	234.69437	238.15842	1.016608	-4.4806609
Oct 2014	215.79297	223.25944	1.016608	-8.4830812
Nov 2014	196.99769	209.97923	1.016608	-13.9981426
Dec 2014	175.43749	185.45707	1.016608	-11.0361951
Jan 2015	210.77882	180.14172	1.016608	29.6204922
Feb 2015	155.21504	148.14369	1.016608	6.0547492
Mar 2015	183.4833	168.55844	1.016608	13.9082497
Apr 2015	215.32318	212.58282	1.016608	1.7237464
May 2015	170.92354	173.31049	1.016608	-3.4035549
Jun 2015	205.28977	205.04106	1.016608	-0.7678967

Month	xhat	level	trend	season
Jul 2015	179.50889	182.82432	1.016608	-4.332038
Aug 2015	160.94553	153.61017	1.016608	6.3187562
Sep 2015	176.14138	181.06523	1.016608	-5.9404616
Oct 2015	185.10268	193.88048	1.016608	-9.7944055
Nov 2015	185.44519	200.76917	1.016608	-16.340594
Dec 2015	197.61726	208.21763	1.016608	-11.6169714
Jan 2016	189.49301	161.88406	1.016608	26.5923427
Feb 2016	186.65596	177.80538	1.016608	7.8339743
Mar 2016	217.5657	198.69611	1.016608	17.8529835
Apr 2016	221.26275	222.21776	1.016608	-1.9716124
May 2016	298.36031	297.93013	1.016608	-0.5864266
Jun 2016	288.24412	290.12641	1.016608	-2.8988927
Jul 2016	294.21188	300.30011	1.016608	-7.1048457
Aug 2016	324.5184	314.75806	1.016608	8.7437287
Sep 2016	262.11292	265.95459	1.016608	-4.8582735
Oct 2016	236.50056	244.73976	1.016608	-9.2558099
Nov 2016	225.48802	240.22207	1.016608	-15.7506558
Dec 2016	222.47438	237.41777	1.016608	-15.9599964
Jan 2017	277.22283	248.2468	1.016608	27.9594236
Feb 2017	275.0716	264.39813	1.016608	9.6568571
Mar 2017	287.13893	266.20514	1.016608	19.9171773
Apr 2017	268.74288	262.84669	1.016608	4.8795885
May 2017	269.66286	270.0417	1.016608	-1.3954399
Jun 2017	303.50589	304.54827	1.016608	-2.058991
Jul 2017	283.25166	288.10704	1.016608	-5.8719872
Aug 2017	306.01917	300.8284	1.016608	4.1741609
Sep 2017	256.78548	262.66624	1.016608	-6.897372
Oct 2017	258.5241	267.27092	1.016608	-9.7634239
Nov 2017	306.39235	321.47685	1.016608	-16.1011152
Dec 2017	262.14275	276.18613	1.016608	-15.059987
Jan 2018	289.56691	259.2027	1.016608	29.347602
Feb 2018	290.91521	280.16925	1.016608	9.7293538
Mar 2018	326.47991	305.94741	1.016608	19.5158901
Apr 2018	284.07218	277.60929	1.016608	5.4462802

#Gráfico fitted

En la figura 77 se tiene mediante grafico los datos anteriormente mostrados para su visualización grafica.

Comando → autoplot(fitted(MODELOPAISMEXICO))+ ylab ("Fitted -Total de Desaparecidos")

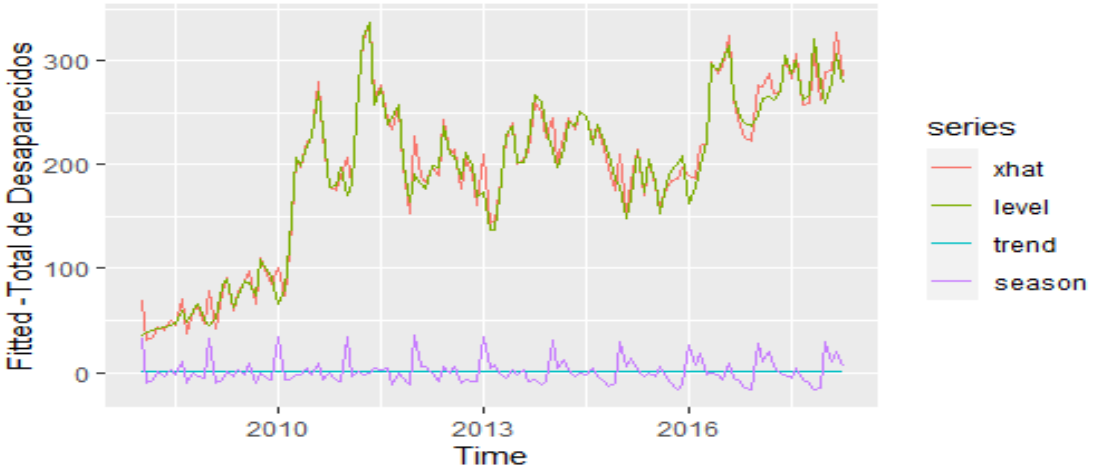


Figura 77. Pantalla del modelo Holt-Winters ajuste del gráfico fitted

C.2 Prueba de ruido blanco

Para probar que existe ruido blanco (o es válido el modelo) tiene que cumplir con unas características que son:

Ruido blanco > 0.5
No hay ruido blanco < 0.5

Esta prueba dentro de RStudio se llama Ljung Box Test, tiene que existir el ruido blanco para realizar el pronóstico.

#Prueba de Ljun-Box

Comando → Box.test(residuals(MODELOPAISMEXICO), type = "Ljung-Box")

```
        Box-Ljung test

data:  residuals(MODELOPAISMEXICO)
X-squared = 0.38152, df = 1, p-value = 0.5368
```

Figura 78. Pantalla del modelo Holt-Winters ajuste del gráfico fitted

En la figura 78 se muestra que el modelo es mayor a 0.5 quiere decir que tiene ruido blanco por ende el modelo es válido.

C.3 Pronóstico

#Código para pronosticar por 1 año

Comando → PREDICCIONPAISMEXICO **<-** predict **(**MODELOPAISMEXICO, n.ahead**=**12, prediction.interval **=** T, level **=** 0.95**)**

#Gráfico del pronóstico

Comando → plot**(**MODELOPAISMEXICO, PREDICCIONPAISMEXICO**)**

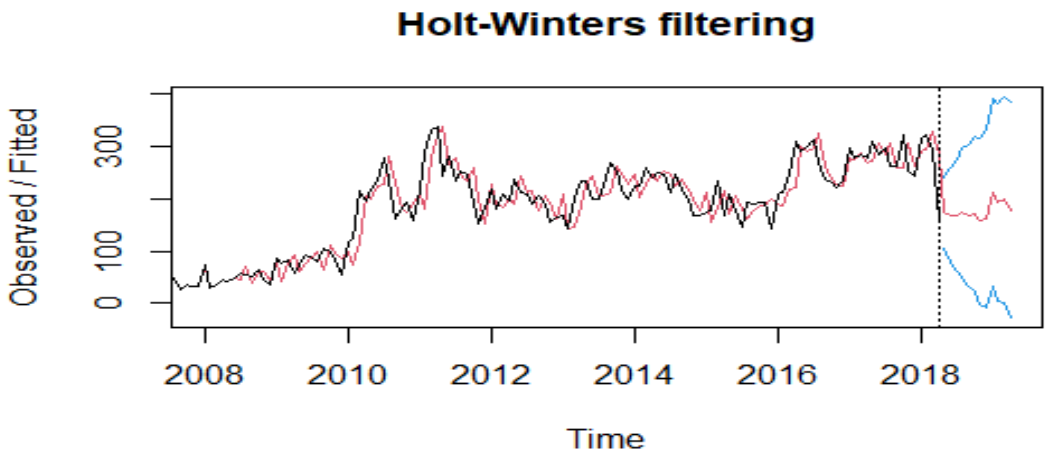


Figura 79. Pantalla del modelo Holt-Winters pronóstico

#Visualizar el resumen total del modelo

Comando → summary **(**forecast**(**MODELOPAISMEXICO**))**

```
Error measures:
                    ME     RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 0.07565164 34.69519 25.83165 -1.498269 14.72901 0.5180534 0.05480457

Forecasts:
         Point Forecast       Lo 80    Hi 80      Lo 95    Hi 95
May 2018       172.2836 127.6396517 216.9275 104.006575 240.5606
Jun 2018       167.9636 109.3318588 226.5954  78.294064 257.6332
Jul 2018       167.8421  97.9686971 237.7155  60.979954 274.7042
Aug 2018       174.2377  94.6959509 253.7795  52.589069 295.8864
Sep 2018       168.1055  79.9493314 256.2616  33.282299 302.9286
Oct 2018       170.8055  74.8049511 266.8061  23.985326 317.6257
Nov 2018       156.3585  53.1077273 259.6092  -1.549905 314.2668
Dec 2018       161.0126  50.9884072 271.0367  -7.254868 329.2800
Jan 2019       209.9176  93.5134633 326.3217  31.892838 387.9424
Feb 2019       191.7573  69.3051421 314.2094   4.482894 379.0317
Mar 2019       197.5968  69.3816278 325.8120   1.508612 393.6850
Apr 2019       177.2354  43.5053085 310.9655 -27.287125 381.7579
```

Figura 80. Pantalla del modelo Holt-Winters Lo80 y Lo95

Lo80: Límite Inferior con 80% de confianza
Hi80: Limite Superior con 80% de confianza
Lo95: Límite Inferior con 95% de confianza
Hi95: Limite Superior con 95% de confianza

Como se visualiza en los resultados el mes con mayor cantidad de posibles desaparecidos es el de mes de abril del 2019 con 177.2354 posibles casos y el menor mes en septiembre del 2018 con 168.1055 de posibles casos.

#Visualizar el pronóstico en gráfico final

Comando → autoplot**(**forecast**(**MODELOPAISMEXICO**))**

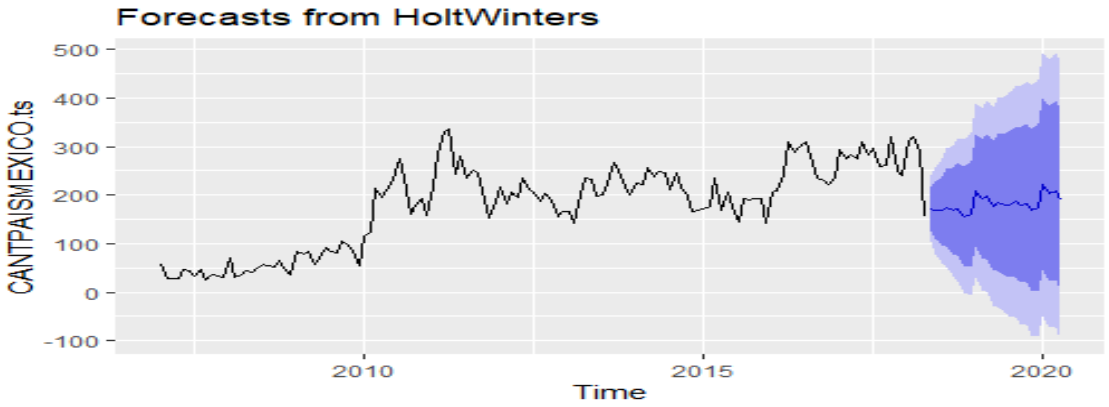


Figura 81. Pantalla del modelo Holt-Winters gráfico del pronóstico

D. Modelo ETS

Para comenzar este modelo se usó de los códigos hasta el plot del diseño, por lo que es la misma sintaxis.

D.1 Prueba de ruido blanco

Para probar que existe ruido blanco (o es válido el modelo) tiene que cumplir con unas características que son:

Ruido blanco > 0.5
No hay ruido blanco < 0.5

Esta prueba dentro de RStudio se llama Ljung Box Test, tiene que existir el ruido blanco para realizar el pronóstico.

#Prueba de Ljun-Box

Comando → Box.test**(**residuals**(**ETSPAISMEXICO**)**, type **=** "Ljung-Box"**)**

```
        Box-Ljung test

data:  residuals(ETSPAISMEXICO)
X-squared = 0.12546, df = 1, p-value = 0.7232
```

Figura 82. Pantalla del modelo Holt-Winters prueba de ruido blanco

En la figura 68 se muestra que el modelo es mayor a 0.5 quiere decir que tiene ruido blanco por ende el modelo es válido.

D.2 Pronóstico

#Código para pronosticar por 1 año

Comando → PREDICCIONPAISMEXICO **<-** forecast **(**ETSPAISMEXICO, 12**)**

#Gráfico del pronóstico

Comando → plot**(**PREDICCIONPAISMEXICO**)**

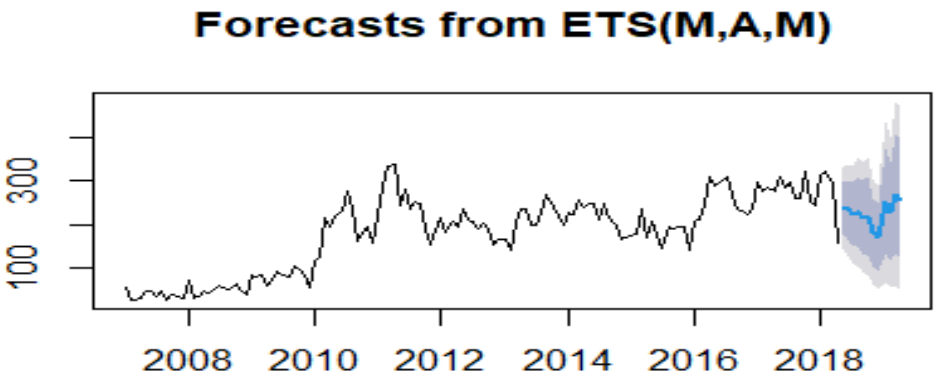


Figura 83. Pantalla del modelo ETS - gráfico del pronóstico

#Ajuste entre el pronóstico y los datos originales

Comando → autoplot**(**PREDICCIONPAISMEXICO**)+**autolayer**(**fitted **(**PREDICCIONPAISMEXICO**)**, series **=** "PRONÓSTICO"**)**

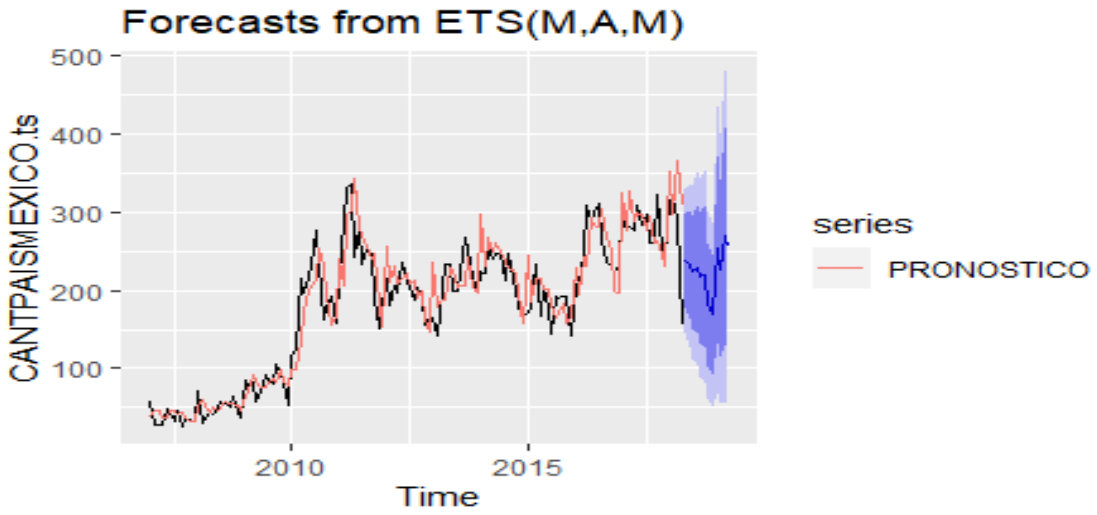


Figura 84. Pantalla del modelo ETS - ajuste entre el pronóstico y los datos originales

#Visualizar la información completa de los resultados obtenidos

Comando → print(summary(PREDICCIONPAISMEXICO))

```
Forecast method: ETS(M,A,M)

Model Information:
ETS(M,A,M)

Call:
 ets(y = CANTPAISMEXICO.ts)

  Smoothing parameters:
    alpha = 0.5611
    beta  = 0.0013
    gamma = 1e-04

  Initial states:
    l = 31.9074
    b = 3.1551
    s = 0.734 0.801 0.9712 0.9726 1.0384 1.0287
           1.0852 1.1215 1.0658 1.131 0.9659 1.0847

  sigma:  0.1977

     AIC     AICc      BIC
1614.212 1619.398 1663.727

Error measures:
                    ME    RMSE      MAE       MPE     MAPE      MASE      ACF1
Training set -3.601086 32.3174 22.81646 -5.192176 15.50703 0.4575838 0.1060333
```

Figura 85. Pantalla del modelo ETS – resumen del modelo

Con los siguientes comandos se tiene la visualización final del pronóstico en intervalos de confianza.

#Visualización del pronóstico final

Comando → RESULTADOSPAISMEXICO **<-**as.data.frame **(**PREDICCIONPAISMEXICO**)**

#Resultado

Comando → RESULTADOSPAISMEXICO

```
Forecasts:
         Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
May 2018       237.6702 177.45521 297.8852 145.57931 329.7611
Jun 2018       232.8375 165.04699 300.6281 129.16086 336.5142
Jul 2018       223.4020 150.94311 295.8609 112.58569 334.2183
Aug 2018       228.2407 147.38656 309.0949 104.58495 351.8965
Sep 2018       216.3372 133.77034 298.9041  90.06208 342.6123
Oct 2018       218.5567 129.58491 307.5286  82.48607 354.6274
Nov 2018       182.3817 103.79306 260.9704  62.19072 302.5728
Dec 2018       169.0323  92.39868 245.6660  51.83128 286.2334
Jan 2019       252.6659 132.72945 372.6024  69.23892 436.0929
Feb 2019       227.4999 114.88480 340.1150  55.26997 399.7298
Mar 2019       269.3702 130.78571 407.9546  57.42353 481.3168
Apr 2019       256.6379 119.80381 393.4719  47.36825 465.9075
```

Figura 86. Pantalla del modelo ETS – Lo80 y Lo95

Lo80: Límite Inferior con 80% de confianza
Hi80: Limite Superior con 80% de confianza
Lo95: Límite Inferior con 95% de confianza
Hi95: Limite Superior con 95% de confianza

Como se visualiza en los resultados el mes con mayor cantidad de posibles desaparecidos es el de mes de marzo del 2019 con 269.3702 posibles casos y el menor mes en diciembre del 2018 con 168.0323 de posibles casos.

IV.5 Fase 5: Evaluación del modelo

IV.5.1 Regresión logística binaria

Los resultados finales de la regresión aplicados en todos los estados y el país de México en general dieron como resultado lo siguiente:

Tabla 39 *Resultados PseudoR² Cox y Snell - Nagelkerke*

N°	Estado	Pseudo R^2 Cox y Snell	Pseudo R^2 Nagelkerke
1	Tlaxcala	0.4898199	1.000000
2	Baja California Sur	0.3486215	0.4648287

N°	Estado	Pseudo R^2 Cox y Snell	Pseudo R^2 Nagelkerke
3	Tabasco	0.33218444	0.45236202
4	San Luis de Potosí	0.31210513	0.49806273
5	Quintana Roo	0.30759289	0.47467264
6	Oaxaca	0.2858749	0.4132273
7	Yucatán	0.2175738	0.3400114
8	Campeche	0.1965937	0.3189035
9	Morelos	0.17401603	0.24993177
10	Nayarit	0.15866139	0.2213468
11	Ciudad de México	0.1525289	0.2269652
12	Hidalgo	0.13847172	0.20625161
13	Colima	0.13654981	0.20716334
14	Durango	0.12746364	0.18926061
15	Michoacán	0.1229156	0.1815299
16	Estado de México	0.1184347	0.1777085
17	Aguas Calientes	0.11599257	0.171198928
18	Querétaro	0.11147275	0.16041936
19	Coahuila de Zaragoza	0.11055004	0.16142987
20	Chiapas	0.10951749	0.15558307
21	Zacatecas	0.10495713	0.14966409
22	Guanajuato	0.10436554	0.14976604
23	Sinaloa	0.10427131	0.1551677
24	País de México	0.1002332	0.14687577
25	Nuevo León	0.1002332	0.14687577
26	Tamaulipas	0.1002332	0.14687577
27	Guerrero	0.09761455	0.14969423
28	Baja California	0.09585671	0.13801648
29	Veracruz	0.0924964	0.13557235
30	Sonora	0.09242729	0.13738677
31	Chihuahua	0.09035949	0.13492474
32	Jalisco	0.07844032	0.11592364
33	Puebla	0.06827002	0.10465141

IV.5.2 Modelo ARIMA

En la tabla 40 se muestra el resultado final del modelo ARIMA basándose en la medición del resultado del error absoluto medio o más conocido como el MAE.

Tabla 40 *Resultado del error absoluto medio (MAE) del modelo ARIMA*

N°	Nombre	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
0	País de México	1.9934	32.5391	**24.2051**	-1.6333	16.2655	0.4854	-0.0034
28	Tamaulipas	0.1422	14.6094	**10.4301**	Inf	Inf	0.4509	-0.0105
19	Nuevo León	0.1294	8.1617	**5.4395**	Inf	Inf	0.4747	-0.0432

N°	Nombre	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
11	Estado de México	0.8642	8.0739	**5.0160**	-14.0423	41.3270	0.5876	-0.0168
25	Sinaloa	-0.0557	6.2946	**4.3903**	Inf	Inf	0.6481	0.2511
15	Jalisco	0.0203	6.0390	**4.3413**	Inf	Inf	0.5791	0.0204
26	Sonora	-0.0780	5.7020	**3.9951**	Inf	Inf	0.7324	-0.0230
8	Coahuila de Zaragoza	-0.0202	5.7366	**3.9335**	Inf	Inf	0.6360	-0.0121
6	Chihuahua	0.4902	4.7295	**3.6288**	-15.3530	42.8492	0.8100	0.0442
2	Baja California	0.0075	7.4954	**3.5628**	NaN	Inf	0.5309	-0.0531
21	Puebla	-0.0001	3.8462	**2.7014**	Inf	Inf	0.5658	0.0028
7	Ciudad de México	0.0234	3.8156	**2.4935**	Inf	Inf	0.5810	0.0147
16	Michoacán	0.2116	3.3367	**2.4296**	Inf	Inf	0.5684	-0.0126
30	Veracruz	0.0095	3.8779	**2.3524**	Inf	Inf	0.4029	-0.1678
13	Guerrero	0.2080	3.4226	**2.3132**	NaN	Inf	0.5432	-0.0117
12	Guanajuato	-0.0079	2.5133	**1.9036**	Inf	Inf	0.8019	-0.0094
20	Oaxaca	-0.0437	3.5701	**1.7746**	Inf	Inf	0.9241	0.0970
9	Colima	-0.0097	2.6334	**1.7056**	Inf	Inf	0.5618	0.0346
32	Zacatecas	0.2782	2.2422	**1.5731**	Inf	Inf	0.7171	-0.0200
10	Durango	0.0074	2.2308	**1.4513**	NaN	Inf	0.5503	0.0080
24	San Luis de Potosí	0.0143	2.5213	**1.1286**	Inf	Inf	0.6964	-0.1685
1	Aguas Calientes	0.1318	1.5701	**1.0397**	NaN	Inf	0.8211	-0.0085
22	Querétaro	0.1304	1.4054	**1.0241**	Inf	Inf	0.7839	-0.0120
17	Morelos	0.1170	1.3347	**0.8963**	Inf	Inf	0.6622	-0.0252
14	Hidalgo	0.0114	1.2281	**0.8341**	Inf	Inf	0.7577	0.0438
5	Chiapas	-0.0034	1.2303	**0.8250**	Inf	Inf	0.9539	-0.1089
18	Nayarit	0.0081	1.9281	**0.5726**	Inf	Inf	0.6478	-0.1215
3	Baja California Sur	0.1065	0.7960	**0.4093**	Inf	Inf	0.9106	-0.0117
27	Tabasco	0.0759	0.7910	**0.4054**	Inf	Inf	0.8426	-0.0262
31	Yucatán	-0.0146	0.5334	**0.3964**	Inf	Inf	0.9102	-0.0112
23	Quintana Roo	0.0674	0.7629	**0.3523**	NaN	Inf	0.8484	0.0176
4	Campeche	0.0353	0.5803	**0.2935**	NaN	Inf	1.0219	0.0118
29	Tlaxcala	0.0000	0.4409	**0.2878**	Inf	Inf	0.8721	-0.0956

IV.5.3 Modelo Holt-Winters

En la tabla 41 se muestra el resultado final del modelo Holt-Winters basándose en la medición del resultado del error absoluto medio o más conocido como el MAE.

Tabla 41 *Resultado del error absoluto medio (MAE) del modelo Holt-Winters*

N°	Nombre	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
0	País de México	0.0757	34.6952	**25.8317**	-1.4983	14.7290	0.5181	0.0548
28	Tamaulipas	-0.1216	16.5641	**12.2144**	Inf	Inf	0.5281	-0.0492
19	Nuevo León	0.2006	9.1952	**6.3657**	NaN	Inf	0.5555	0.0451
11	Estado de México	0.9288	8.5538	**5.3710**	-6.5975	37.3463	0.6292	-0.0127
15	Jalisco	-0.4136	6.3558	**4.7475**	NaN	Inf	0.6333	0.0468
25	Sinaloa	0.5672	6.5296	**4.5768**	-24.8417	51.3214	0.6756	0.1580
8	Coahuila de Zaragoza	-0.1241	6.2991	**4.5437**	Inf	Inf	0.7346	0.3241
26	Sonora	0.2955	6.0530	**4.2089**	NaN	Inf	0.7716	0.0448
2	Baja California	-0.0965	9.9980	**4.1057**	NaN	Inf	0.6118	0.0098
6	Chihuahua	0.1372	5.2318	**3.9677**	-17.8009	43.5068	0.8857	0.0818
7	Ciudad de México	0.0408	4.3991	**2.7731**	NaN	Inf	0.6461	0.1051
16	Michoacán	0.1341	3.5733	**2.7344**	Inf	Inf	0.6397	0.0534
21	Puebla	0.4200	4.0380	**2.7179**	NaN	Inf	0.5693	-0.0466
13	Guerrero	0.0108	3.6902	**2.6713**	NaN	Inf	0.6274	0.0446
30	Veracruz	-0.0050	4.1675	**2.6618**	NaN	Inf	0.4559	0.0217
12	Guanajuato	0.2804	2.6742	**1.8935**	NaN	Inf	0.7976	0.1370
20	Oaxaca	-0.4828	3.9323	**1.8410**	NaN	Inf	0.9586	0.0462
9	Colima	0.3592	2.8144	**1.8192**	NaN	Inf	0.5993	0.1987
32	Zacatecas	-0.2141	2.2772	**1.7520**	NaN	Inf	0.7987	0.0269
10	Durango	-0.1963	2.4808	**1.6536**	NaN	Inf	0.6270	-0.0305
1	Aguas Calientes	0.1842	1.7454	**1.1746**	NaN	Inf	0.9277	0.0557
22	Querétaro	0.1555	1.5052	**1.1498**	NaN	Inf	0.8801	0.0066
24	San Luis de Potosí	-0.1529	1.6270	**1.1335**	NaN	Inf	0.6994	0.2240
17	Morelos	0.0039	1.4807	**1.0413**	NaN	Inf	0.7693	-0.0418
14	Hidalgo	-0.1148	1.4599	**0.9839**	NaN	Inf	0.8937	0.0100
18	Nayarit	-0.0234	2.0376	**0.7995**	NaN	Inf	0.9044	0.1221
5	Chiapas	0.2029	1.3453	**0.7158**	NaN	Inf	0.8277	-0.0454
27	Tabasco	0.0312	0.9124	**0.5477**	NaN	Inf	1.1385	0.0456
3	Baja California Sur	0.1035	0.8707	**0.4835**	NaN	Inf	1.0758	-0.0158
23	Quintana Roo	-0.0092	0.9122	**0.4591**	NaN	Inf	1.1055	0.3683
31	Yucatán	0.0589	0.5576	**0.3914**	NaN	Inf	0.8987	0.1530
4	Campeche	0.0612	0.5871	**0.3687**	NaN	Inf	1.2835	-0.0243
29	Tlaxcala	0.0479	0.4864	**0.2986**	NaN	Inf	0.9048	-0.1239

IV.5.4 Modelo ETS

En la tabla 42 se muestra el resultado final del modelo ETS basándose en la medición del resultado del error absoluto medio o más conocido como el MAE.

Tabla 42 *Resultado del error absoluto medio (MAE) del modelo ETS*

N°	Nombre	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
0	País de México	-3.6011	32.3174	**22.8165**	-5.1922	15.5070	0.4576	0.1060
28	Tamaulipas	-0.0163	15.5480	**10.9623**	Inf	Inf	0.4739	-0.0522
19	Nuevo León	0.0485	8.5810	**5.4841**	Inf	Inf	0.4786	0.0054
11	Estado de México	0.3172	8.3720	**5.1761**	-21.7131	42.8238	0.6063	0.0806
25	Sinaloa	0.7708	6.3672	**4.2660**	Inf	Inf	0.6297	0.1852
15	Jalisco	0.0488	6.0389	**4.3548**	Inf	Inf	0.5810	0.0220
26	Sonora	0.6464	5.8623	**3.9593**	Inf	Inf	0.7259	0.0110
8	Coahuila de Zaragoza	-0.0488	5.9894	**4.0299**	Inf	Inf	0.6516	0.2269
6	Chihuahua	-0.1760	4.4175	**3.3406**	-25.3659	44.4592	0.7457	0.0814
2	Baja California	0.0123	9.3380	**3.4108**	NaN	Inf	0.5083	0.0303
21	Puebla	0.3985	3.8865	**2.5446**	NaN	Inf	0.5330	-0.0347
7	Ciudad de México	0.0461	4.1217	**2.4792**	NaN	Inf	0.5777	0.1008
16	Michoacán	0.2015	3.3693	**2.4521**	Inf	Inf	0.5737	0.0509
30	Veracruz	0.0096	3.8173	**2.2503**	Inf	Inf	0.3854	0.0074
13	Guerrero	0.1285	3.4852	**2.3460**	Inf	Inf	0.5510	0.0764
12	Guanajuato	0.1730	2.5225	**1.8066**	Inf	Inf	0.7610	0.2035
20	Oaxaca	-0.0814	3.8188	**1.7827**	Inf	Inf	0.9283	0.0107
9	Colima	0.3117	2.6206	**1.5952**	NaN	Inf	0.5255	0.1904
32	Zacatecas	0.0785	2.1673	**1.5463**	Inf	Inf	0.7049	0.0273
10	Durango	0.0109	2.2467	**1.4356**	Inf	Inf	0.5444	-0.0394
24	San Luis de Potosí	-0.0080	2.0640	**1.2567**	Inf	Inf	0.7754	0.2190
1	Aguas Calientes	0.1347	1.5979	**1.0554**	Inf	Inf	0.8336	0.0509
22	Querétaro	0.1380	1.4054	**1.0261**	Inf	Inf	0.7854	-0.0094
17	Morelos	0.1181	1.3346	**0.8997**	Inf	Inf	0.6647	-0.0187
14	Hidalgo	0.0185	1.2279	**0.8431**	Inf	Inf	0.7658	0.0534
5	Chiapas	0.1691	1.1758	**0.6873**	Inf	Inf	0.7946	-0.0609
18	Nayarit	0.0081	1.9147	**0.5690**	NaN	Inf	0.6437	-0.0039
3	Baja California Sur	-0.0122	0.7704	**0.4634**	NaN	Inf	1.0310	-0.0001
27	Tabasco	0.0179	0.8076	**0.4142**	Inf	Inf	0.8609	0.0593
31	Yucatán	0.0669	0.5361	**0.3890**	NaN	Inf	0.8933	0.2548

N°	Nombre	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
23	Quintana Roo	0.0096	0.8358	**0.3626**	Inf	Inf	0.8732	0.1453
4	Campeche	0.0672	0.5660	**0.3311**	NaN	Inf	1.1526	-0.0150
29	Tlaxcala	0.0000	0.4410	**0.2878**	Inf	Inf	0.8720	-0.0956

IV.6 Fase 6: Implementación del modelo

Culminando el estudio y revisando los resultados se procedió a escoger los mejores modelos de pronóstico para su proyección en la plataforma escogida que en este caso es la plataforma de Power BI. Asimismo, monitorear las acciones para detectar áreas de oportunidad o incluso nuevos problemas (Espinoza, 2019, p. 4).

IV.7 Pruebas de hipótesis

IV.7.1 Hipótesis específica 1

$HE1_0$: La complexión, el rango de edad, la presencia de acné y verrugas, los bigotes, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares no están relacionados con los casos de desaparición de varones.

$HE1_1$: La complexión, el rango de edad, la presencia de acné y verrugas, los bigotes, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares están relacionados con los casos de desaparición de varones.

Con los resultados obtenidos mediante técnica de regresión logística binaria, la cual fue aplicada a los 32 estados y el país México en general, por lo que se concluyó que se rechaza la $HE1_0$ y se acepta la $HE1_1$ la cual indica si existe la relación entre los ítems edad, acné, verruga, bigote, cicatrices, tatuajes, problemas con la dentadura y otras señas las cuales predomina en los varones en los casos de desaparición (ver anexo 4).

IV.7.2 Hipótesis específica 2

$HE2_0$: El uso del modelo ARIMA mejoró la precisión del pronóstico de casos de personas desaparecidas no fue de al menos del 95%

$HE2_1$: El uso del modelo ARIMA mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 95%

Con los resultados obtenidos mediante la investigación con el uso del modelo ARIMA la cual fue aplicada a cada uno de los 32 estados y el país de México en total, se llegó a la conclusión de que se rechaza la $HE2_0$ y se acepta la $HE2_1$ la cual indicó que la precisión del pronóstico fue de por lo menos del 95% (ver anexo 3).

IV.7.3 Hipótesis específica 3

$HE3_0$: El uso del modelo Holt-Winters mejoró la precisión del pronóstico de casos de personas desaparecidas no fue de al menos del 85%.

$HE3_1$: El uso del modelo Holt-Winters mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 85%.

Con los resultados obtenidos dentro la investigación y el modelo Holt-Winters se llegó a la conclusión de que se rechaza la $HE3_0$ y se acepta la $HE3_1$ la cual indica la precisión del pronóstico fue de por lo menos del 85% (ver anexo 3).

IV.7.4 Hipótesis específica 4

$HE4_1$: El uso del modelo ETS mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 90%

$HE4_0$: El uso del modelo ETS mejoró la precisión del pronóstico de casos de personas desaparecidas no fue de al menos del 90%

Con los resultados obtenidos dentro de la investigación y modelo ETS, se llegó a la conclusión de que se rechaza la $HE4_0$ y se acepta la $HE4_1$ la cual indica la precisión del pronóstico fue de por lo menos del 90% (ver anexo 3).

IV.7.5 Hipótesis general

HG_1: El uso de la regresión logística binaria y modelos de series de tiempo mejoró la precisión del pronóstico de casos de personas desaparecidas.

HG_0: El uso de la regresión logística binaria y modelos de series de tiempo no mejoró la precisión del pronóstico de casos de personas desaparecidas.

Con los resultados obtenidos mediante la técnica de regresión logística binaria y las series de tiempo, se concluyó que se rechaza la $HE1_0$ y se acepta la $HE1_1$ la cual indica que usando estas técnicas y métodos se puede mejorar el pronóstico para los casos de personas desaparecidas (ver anexo 3-4).

IV.7.6 Resumen

A continuación, se detallará las hipótesis aceptadas en la investigación:

Tabla 43 *Resumen general de las hipótesis*

Hipótesis	Termino	Detalle de hipótesis	Resultado
HG	HG_1	El uso de la regresión logística binaria y modelos de series de tiempo mejoró la precisión del pronóstico de casos de personas desaparecidas.	Aceptada
HE1	$HE1_1$	La complexión, el rango de edad, la presencia de acné y verrugas, los bigotes, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares están relacionados con los casos de desaparición de varones.	Aceptada
HE2	$HE2_1$	El uso del modelo ARIMA mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 95%.	Aceptada
HE3	$HE3_1$	El uso del modelo Holt-Winters mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 85%.	Aceptada

Hipótesis	Termino	Detalle de hipótesis	Resultado
HE4	$HE4_1$	El uso del modelo ETS mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 90%.	Aceptada

En la tabla 43 se observa que los resultados de esta investigación fueron muy favorables, debido a que se llegó a aceptar todas las hipótesis planteadas. Se aceptó la hipótesis general y las hipótesis específicas.

V. DISCUSIÓN

Este estudio se basa en cómo prevenir o disminuir el índice de personas desaparecidas, usando la regresión logística binaria y modelos de series de tiempo ARIMA, Holt-Winters y ETS. En este sentido, se trabajó con un total de 36,265 datos los cuales fueron depurados mediante la metodología CRISP-DM, logrando determinar la precisión del pronóstico de casos de personas desaparecidas.

Los resultados de la regresión logística binaria para el Estado de Jalisco fueron: (a) Pseudo R^2 Cox y Sneel: 7.84% y (b) Pseudo R^2 Nagalkerke: 11.59%. Estos resultados fueron menores a los presentados por Delgado et al. (2014) quienes evaluaron la regresión logística binaria para el rendimiento académico alcanzando una estimación del 80.5%, debido a que tuvieron dieciocho variables independientes (de las cuales tuvieron cinco en el modelo final), mientras que en esta investigación se tuvo nueve variables (de las cuales quedaron cuatro en el modelo final).

Con el modelo de serie de tiempo ARIMA para el Estado de Campeche se obtuvo una precisión del 99.70% para los casos de personas desaparecidas, lo que fue mayor al resultado del estudio de Sosa et al. (2019) en el que obtuvieron una precisión del 95.00% para el monitoreo de parámetros ambientales. Asimismo, se obtuvo un mayor resultado que el logrado en el estudio de Sánchez et al. (2014), quienes obtuvieron una precisión del 85.00% usando el modelo ARIMA enfocado en la predicción de la producción de leche.

Si bien se obtuvo una mayor precisión en los resultados de la investigación, se resalta que estos resultados no son coherentes con la cantidad de 35 registros utilizada para el modelo de ARIMA, ya que en el estudio de Sosa et al. (2019) obtuvieron 5,889 registros de parámetros ambientales y en el estudio de Sánchez et al. (2014) obtuvieron 132 registros de producción de leche (132 meses). Las cantidades de registros mayores de los estudios citados debieron contribuir a una mejor precisión. También, esto exige revisar con mayor profundidad los 35 registros que conformaron la muestra para el Estado de Campeche.

Los resultados de la serie de tiempo ARIMA para el Estado de Campeche a través de la prueba de ruido blanco indican que el modelo es válido si es mayor a 0.5 y si es menor a 0.5 no es válido. En este sentido, al obtener un 0.9022 (90.22%) en el estudio se indica que existe ruido blanco y el modelo es válido, lo que fue mayor al resultado del estudio de Palomino et al. (2019) en el que obtuvieron a través de la prueba un 0.05 (5%) para la predicción de variables ambientales (temperatura, humedad y velocidad del aire). Por lo tanto, los resultados de Palomino et al. (2019) fueron menores debido a que solo trabajaron con tres variables, mientras que en esta investigación se manejaron nueve de las cuales quedaron seis (complexión, edad, cicatrices, lunares, tatuajes y otras señas) para el modelo final.

Con el modelo de serie de tiempo Holt-Winters para el Estado de Campeche se obtuvo una precisión del 99.63% y un error del 0.37% para los casos de personas desaparecidas, lo que fue mayor al resultado del estudio de Roque et al. (2016) en el que obtuvieron una precisión del 84.42% para su modelo final enfocados en el pronóstico de ventas. Asimismo, se obtuvo un mayor resultado que el logrado en el estudio de Mariño et al. (2021) quienes obtuvieron un error aproximado del 3% usando el modelo Holt-Winters enfocado en el pronóstico de la demanda eléctrica del sector de explotación de minas.

Si bien se obtuvo una mayor precisión en los resultados de la investigación, donde se resalta que se trabajó con un intervalo de tiempo entre Enero 2008 - Octubre 2016 (106 meses) para el estado de Campeche, mientras que en el estudio de Roque et al. (2016) se tuvo un intervalo entre agosto 2011 y junio 2014 (35 meses) para la predicción de ventas y en el estudio de Mariño et al. (2021), quienes tuvieron un intervalo de tiempo entre enero 2010 y junio 2018 (102 meses) para el análisis de la demanda comercial de energía. Las cantidades de meses no son contrastables a los registros, por lo tanto, esto exige revisar con mayor profundidad los 35 registros que conforman la muestra para el Estado de Campeche.

Con el modelo de serie de tiempo Holt-Winters para el Estado de Tamaulipas se obtuvo una precisión del 87.78% y error del 12.22% para los casos de personas desaparecidas, lo que fue mayor al resultado del estudio de Roque et al. (2016) en el que obtuvieron una precisión del 84.42% y error del 15.58% para la predicción de venta. Si bien se obtuvo una mayor precisión en los resultados de la investigación, cabe resaltar que estos resultados son inferiores a la cantidad de la muestra del Estado de Tamaulipas con 4,589 datos, mientras que en el estudio de Roque et al. (2016) trabajaron con 4,6122 datos de registros de ventas.

Con el modelo de serie de tiempo ETS para el Estado de Campeche se obtuvo una precisión del 99.66% para los casos de personas desaparecidas, lo que fue mayor al resultado del estudio de Roque et al. (2016) en el que obtuvieron una precisión del 90.51% para el pronóstico de ventas. Si bien se obtuvo una mayor precisión en el resultado de la investigación, cabe resaltar que estos son mayores en el tiempo de pronóstico de los casos de personas desaparecidas, ya que en el estudio se tuvo como objetivo pronosticar los casos por un año (12 meses) mientras que en el estudio de Roque et al. (2016) realizaron un pronóstico de tres meses (90 días).

Con el modelo de serie de tiempo ETS para el Estado de Campeche se obtuvo un total de 16 modelos finales para los casos de personas desaparecidas, lo que fue menor al resultado del estudio de Del Pilar et al. (2018) en el que obtuvieron 17 modelos finales para los resultados de tendencia en las metodologías para el pronóstico mandante la serie de tiempo ETS. En este sentido, el estudio es menor a los presentados por Del Pilar et al. (2018) por la diferencia de un modelo, donde se resalta que en este estudio se tiene a la serie de tiempo ETS como modelo dominante ante los otros modelos (ARIMA y Holt-Winters).

Mediante los modelos de serie de tiempo orientados a pronosticar los casos de personas desaparecidas se tomaron los 32 estados y al país de México en general, lo que fue comparado con los resultados del estudio de Mammadova et al. (2020) quienes evaluaron una estrategia para pronosticar la población utilizando series temporales difusas. Asimismo, se hace la comparación de

estudios en el aspecto de los países puesto que Mammadova et al. (2020) se enfocaron en pronosticar el número de habitantes de Azerbaiyán, mientras que este estudio tuvo como objetivo el pronóstico de casos de personas desaparecidas en México.

Mediante los casos de personas desaparecidas se tienen diversas características o señas particulares. En este sentido, se tiene el estudio de Mammadova et al. (2020), quienes tomaron en cuenta algunas características como la: población económicamente activa, población de diferentes grupos de edad, muertes y nacimientos. Al respecto, los ítems usados en este estudio tienen una semejanza en cuanto a: población total, estados, señas particulares, rango de edad y sexo.

Las características más comunes en un caso hipotético de desaparición de personas muestran los siguientes rasgos: acné, verrugas, bigote, cicatrices, tatuajes, problemas con la dentadura y otras señas particulares. En este sentido, Rodríguez et al. (2020) proporcionaron datos de análisis actualizados sobre casos de feminicidios teniendo como objetivo demostrar la existencia de estrategias sobre la desaparición de personas, lo que sirve de apoyo a lo presentado en capítulos anteriores, lo cual es definir qué características afectan más a cada Estado y al país en general.

En el Estado de Tamaulipas se tiene un intervalo de 9 años entre el 2007 al 2016 con un total de 4,589 casos de personas desaparecidas, lo que fue mayor al resultado del estudio de Rodríguez et al. (2020) quienes en su estudio lo realizaron en un intervalo de 6 años desde el 2008 hasta el 2013 sobre las estrategias del Estado Mexicano para minimizar los feminicidios. Asimismo, Roque (2016) estudió un intervalo de 3 años (desde el 2011 hasta el 2014) con una totalidad de 46,122 datos en el pronóstico de ventas. En este sentido el estudio es mayor en la cantidad de años de datos teniendo una diferencia de 3 años (36 meses) respecto al estudio de Rodríguez et al. (2020) y 6 años (72 meses) en el estudio de Roque (2016).

En este estudio se obtuvo un total de 36,265 datos procesándolos mediante la metodología CRISP-DM, lo que fue comparado al estudio de

Barbulescu et al. (2021) quienes estudiaron una gran cantidad de datos y a la vez realizaron pronósticos de la carga diaria, a comparación de esta investigación que procesó casos de personas desaparecidas. En ese sentido, se asemeja al estudio de Severiukhina et al. (2020) quienes estudiaron los pronósticos a gran escala de la difusión de información, utilizando la métrica de MAE para la microescala. Asimismo, se precisa que este estudio está bajo la premisa del método de medición de error (MAE), la cual se desarrolló en los modelos de pronóstico y series de tiempo.

Si bien se obtuvo un error menor al 10% mediante la métrica de error MAE, en los estados y el país de México en general en base a las series de tiempo y modelo de regresión logística binaria, estos resultados fueron similares al objetivo de Severiukhina et al. (2020), en el que tuvieron como objetivo un error menor al 10% mediante métricas en el pronóstico a gran escala de la difusión de información. En este sentido, se tiene una precisión de pronóstico semejante entre ambos estudios.

La serie de tiempo ARIMA para el país de México se obtuvo mediante la librería auto.arima el modelo final (1,1,1)(1,0,0) para los casos de personas desaparecidas. Estos resultados fueron comparados con los resultados del estudio de Sosa et al. (2019), en el que obtuvieron como modelo final para la serie de tiempo ARIMA (5,1,0) (0,0,1) para el monitoreo de parámetros ambientales. Debido a lo detallado y teniendo como punto de comparación en la cantidad de valores autorregresivos, diferencia y media móvil. En este sentido se tiene la diferencia de autorregresivos mientras que en el estudio de Sosa et al. (2019) obtuvieron cinco autorregresivos, teniendo una diferencia de cuatro autorregresivos.

Con los modelos de series de tiempo trabajados para el país de México se manejó un total de 24,476 datos en el estudio, desarrollando en cada modelo la confianza al 80% y 95% en el entorno y lenguaje de programación R, obteniendo un resultado mayor al estudio de Severiukhina et al. (2020), quienes en su estudio manejaron 5,889 datos para la predicción a gran escala de la difusión. Asimismo, se obtuvo un mayor resultado que el logrado en el estudio de Severiukhina et al. (2020), quienes en su estudio solo se enfocaron en aplicar

la confianza al 95%, lo que hizo que se tenga una menor aproximación. En este sentido, al aplicar la confianza al 80% y 95% estos permiten realizar una comparación más amplia, mediante los porcentajes, ya sea por los estados o por el país de México para los casos de personas desaparecidas.

VI. CONCLUSIONES

Las conclusiones de la investigación fueron las siguientes:

1. La técnica de regresión logística binaria para el pronóstico de casos de personas desaparecidas tuvo buenos resultados, ya que se obtuvieron las características que más afectan en general al país de México las cuales fueron: rango de edad (niños, adolescentes, jóvenes, adultos y adultos mayores), acné, verrugas, bigote, cicatrices, tatuajes, problemas con la dentadura y otras señas particulares. Por otro lado, las características excluyentes fueron: complexión (delgada, normal, robusta, mediana y obesa) y los lunares, lo que se obtuvo mediante la fórmula aplicada.

2. Considerando a la variable dependiente “sexo-masculino” y las variables independientes como el resto de los “ítems” para cada uno de los estados, se obtuvo lo siguiente:
   2.1 El ítem complexión (delgada, normal, robusta, mediana y obesa) tiene una mayor probabilidad que afecte a las personas en los siguientes estados: Baja California Sur, Campeche, Oaxaca, Tlaxcala y Yucatán.
   2.2 El ítem acné y verruga tienen una mayor probabilidad que afecte a las personas en los siguientes estados: Oaxaca, Tlaxcala, Yucatán, Tamaulipas y San Luis de Potosí.
   2.3 El ítem bigote tiene una mayor probabilidad que afecte a las personas en los siguientes estados: Oaxaca, Tlaxcala, Yucatán, Tamaulipas, Baja California, Chihuahua, Ciudad de México, Coahuila de Zaragoza, Estado de México, Guanajuato, Jalisco, Nuevo León, Puebla, Sinaloa y Veracruz.
   2.4 El ítem cicatrices tiene una mayor probabilidad que afecte a las personas en los siguientes estados: Oaxaca, Tlaxcala, Yucatán, Tamaulipas, Baja California, Chihuahua, Ciudad de México, Coahuila de Zaragoza, Guanajuato, Jalisco, Nuevo León, Puebla, Sinaloa, Veracruz, Campeche, Aguas Calientes, Colima, Michoacán y Sonora.
   2.5 El ítem lunar tiene una mayor probabilidad que afecte a las personas en los siguientes estados: Oaxaca, Tlaxcala, Yucatán, Tamaulipas, Coahuila de Zaragoza, Campeche, Sonora y Estado de México.

2.6 El ítem problemas con la dentadura tiene una mayor probabilidad que afecte a las personas en los siguientes estados: Oaxaca, Tlaxcala, Yucatán, Tamaulipas, Michoacán, Guerrero y Chiapas.

2.7 El ítem otras señas tiene una mayor probabilidad que afecte a las personas en los siguientes estados: Oaxaca, Tlaxcala, Yucatán, Tamaulipas, Michoacán, Guerrero, Chiapas, Campeche, Nuevo León, Veracruz y Durango.

2.8 El ítem tatuajes tiene una mayor probabilidad que afecte a las personas en la mayor cantidad de estados, pero con una menor probabilidad en los siguientes estados: San Luis de Potosí, Baja California Sur, Hidalgo, Nayarit, Quintana Roo y Tabasco.

2.9 El ítem edad (niños, adolescentes, jóvenes, adultos y adultos mayores) tiene una mayor probabilidad que afecte a las personas en la mayor cantidad de estados, pero con una menor probabilidad en los siguientes estados: Chiapas y San Luis de Potosí.

3. La serie de tiempo ARIMA se ajustó mejor a los estados Tamaulipas, Nuevo León, Estado de México, Sinaloa y Jalisco, consiguiendo errores de 10.4301%, 5.4395%, 5.0160%, 4.3903% y 4.3413%, respectivamente, mediante la métrica MAE y una precisión del 89.5700%, 94.5605%, 94.9840%, 95.6098% y 95.6587%, respectivamente.

4. La serie de tiempo ARIMA enfocada para la predicción de casos de personas desaparecidas cumple con su objetivo de mostrar modelos de pronósticos finales. En este sentido, se tiene un total de 17 estados que tienen una mejor precisión en el pronóstico, tomando como ejemplo al Estado de Tabasco con una precisión del 99.59% a comparación de las otras series de tiempo con un índice del 99.45% en el modelo Holt-Winters y 99.58% en el modelo ETS.

5. La serie de tiempo ARIMA enfocada para la predicción de casos de personas desaparecidas cumplen con pronosticar buenos modelos y usando la librería auto.arima se optimiza el proceso de búsqueda de modelos finales. Tomando al país de México se adoptó el modelo (1,1,1)(1,0,0), ya que este es el mejor que se ajusta para el pronóstico final.

6. Con los modelos de series de tiempo ARIMA, Holt-Winters y ETS para el Estado de Campeche, los resultados fueron los siguientes: (a) para ARIMA: precisión del 99.7065% y error del 0.2935% (modelo final), (b) para Holt-Winters: 99.6313% de precisión y error del 0.3687% y (c) para ETS: 99.6689% de precisión y error del 0.3311%. Cumpliendo con el objetivo de determinar la precisión del pronóstico de casos de personas desaparecidas resaltando el Modelo ARIMA como modelo final para este estado.

7. El modelo de serie de tiempo Holt-Winters se ajustó mejor a los estados de: Tamaulipas, Nuevo León, Estado de México, Jalisco y Sinaloa**;** consiguiendo un 12.2144%, 6.3657%, 5.3710%, 4.7475%, 4.5768% respectivamente, mediante la métrica MAE y una precisión del 87.7856%, 93.6343%, 94.6290%, 95.2525% y 95.4232%, respectivamente.
8. La serie de tiempo Holt-Winters enfocada para la predicción de casos de personas desaparecidas no cumple con su objetivo de mostrar modelos de pronósticos con mayor precisión a comparación de las otras series de tiempo (ARIMA y ETS). En este sentido, no hubo estados para los modelos finales.
9. El modelo de serie de tiempo ETS se ajustó mejor a los estados: Tamaulipas, Nuevo León, Estado de México, Jalisco y Sinaloa**;** consiguiendo un 10.9623%, 5.4841%, 5.1761%, 4.3548%, 4.2660% respectivamente, mediante la métrica MAE y una precisión del: 89.0377%, 94.5159%, 94.8239%, 95.6452% y 95.7340%, respectivamente.

10. La serie de tiempo ETS enfocada para la predicción de casos de personas desaparecidas cumplió con su objetivo de mostrar modelos de pronósticos finales. En este sentido, se tuvo un total de 16 estados que tuvieron mejor precisión en el pronóstico, tomando como ejemplo al Estado de Chiapas con una precisión del 99.31% a comparación de las otras series de tiempo con un índice del 99.28% en el modelo Holt-Winters y 99.17% en el modelo ARIMA.

11. La metodología CRISP-DM obtuvo una puntuación alta, ya que se tuvo un mejor alcance en la comprensión del problema por su orientación a la minería de datos e incluir: comprensión de datos, modelado de los mismos y su implementación, de manera detallada. Esta metodología brindó una guía para

proyectar todos los aspectos de los objetivos que se plantearon en este estudio enfocado en pronosticar casos de personas desaparecidas.

VII. RECOMENDACIONES

Las recomendaciones para futuras investigaciones fueron las siguientes:

1. Estimar los resultados obtenidos con otros modelos de series de tiempos como son: Naive (modelo que se enfoca en los valores observados), Elman y Jhordan. Así se podrá comparar la precisión obtenida en este estudio con los modelos propuestos. También se sugiere evaluar los resultados con otras técnicas de pronóstico como: las redes neuronales y las combinaciones de las regresiones con las redes neuronales (Mariño, 2021, p. 18).
2. Implementar este sistema en otras plataformas interactivas como el Tableau o Dundas BI para que junto al Power BI se pueda llegar a una cantidad mayor de usuarios.
3. Recabar bases de datos con más países para ampliar la información, tomando en cuenta a Argentina (3,446), Guatemala (3,154), Perú (3,006) y Colombia (1,260), ya que estos países tienen mayores índices de casos de desapariciones en América Latina (Agudo, 2016, π. 3).
4. Aplicar los pasos desarrollados con la metodología CRISP-DM en esta investigación como guía para el desarrollo de modelos de pronóstico de casos de desapariciones de personas en países con características similares.
5. Ampliar la investigación con diversos ítems como: hora, estatura, etnia o alguna discapacidad, dentro de la base de datos, para así conseguir más características y señas que permitan relacionarlas para lograr un mejor análisis o pronóstico a desarrollar.
6. Realizar un análisis completo de las características disponibles en la base de datos como: complexión (delgada, normal, robusta, mediana y obesa), edad (niños, adolescentes, jóvenes, adultos y adultos mayores), acné, verruga, bigote, cicatrices, lunares, tatuajes, problemas con la dentadura y otras señas particulares, con el fin de clasificar a las personas desaparecidas en grupos que sirvan para el mejor análisis de las autoridades pertinentes con la técnica estadística del análisis de conglomerados.

7. Agregar otros criterios de análisis a la base de datos, tales como: las coordenadas geográficas de la desaparición y la demora en la ubicación de personas desaparecidas, para mejorar la clasificación de los grupos de personas con el análisis de conglomerados.

REFERENCIAS

AGUDO, A [en línea]. *¿Ha visto a este joven?,* 2016 [consulta: 19 de setiembre de 2020]. Disponible en: https://bit.ly/3FUgtiE

AQUINO, A. A., MOLERO, G. y ROJANO, R. Hacia un nuevo proceso de minería de datos centrado en el usuario. *Instituto tecnológico de Celaya* [en línea]. 2015, **114**(1), pp.272-291 [consulta: 19 de setiembre de 2020]. ISSN 1405-1249. Disponible en: https://bit.ly/34tGBzq

ARANA, V., CABRERA, F., PEREZ, J., DORTA, P., HERNANDEZ, S. y JIMENEZ, E. Challenges of an Autonomous Wildfire Geolocation System Based On Synthetic Vision Technology [Desafíos de un sistema autónomo de geolocalización de incendios forestales basado en la tecnología de visión sintética]. *Sensors* [en línea]. 2018, **18**(1), pp.1-16 [fecha consulta 10 de octubre de 2020]. doi:10.3390/s18113631. Disponible en: https://bit.ly/3nIElw8

ARANEDA ASTUDILLO, A. y GATICA LEIVA, C. *Sistema de difusión de información dependiente de la geolocalización para el ambiente universitario* [en línea]. Tesis de pregrado. Universidad del Bio-bio, Concepción-Chile, 2013 [consultado: 04 de diciembre de 2019]. Disponible en: https://bit.ly/2ZQo26d

ARIAS. J., L. *Proyecto de Tesis: Guía para la elaboración*. Arequipa-Perú: Arias Gonzales, 2020 [consultado: 17 de setiembre de 2020]. ISBN: 978-612-00-5416-1 (Online). Disponible en: https://bit.ly/2WzHTVi

BANDA, H., GARZA, R. Aplicación teórica del método Holt-Winter's al problema de credit scroring de las instituciones de micro finanzas. Revista Universidad Autónoma de Querétaro [en línea]. 2014, **15**(2), pp.5-21. [fecha consulta 27 de octubre de 2020]. ISSN 1810-9993. DOI: https://bit.ly/3rq2GJk. Recuperado de https://bit.ly/3NhfqMb

BARBULESCU, C., KILYENI, S., CHIS, V., CRACIUN, M. y SIMO, A. *Daily load curve forecasting. comparative analysis: Conventional vs. unconventional Methods [Predicción de la curva de carga diaria. Análisis comparativo: Métodos convencionales vs. no convencionales].* 8vo Taller Internacional de Aplicaciones de Soft Computing SOFA 2018. Conference

held in Arad, Romania, 2021 [consultado: 17 de setiembre de 2020]. Disponible en: https://bit.ly/3aqUd2p

BAZÁN, W. Fundamentos para pronosticar una serie de tiempo estacionaria con información de su propio pasado. *Revista industrial* [en línea]. 2020, **23**(1), pp.207-228. [fecha consulta 27 de octubre de 2020]. ISSN 1810-9993. DOI: https://bit.ly/3rq2GJk. Recuperado de https://bit.ly/3avpjGg

BLACONÁ, M., MAGNANO, L., y ANDEOZZI. Características de los modelos de espacio de estado de innovaciones, con aplicaciones. *Ciencias Económicas y Estadística* [en línea]. 2012, **1**(1), pp.207-228. [fecha consulta 27 de octubre de 2020]. ISSN 1810-9993. DOI: https://bit.ly/3rq2GJk. Recuperado de https://bit.ly/37ILPfC

BBC [en línea]. *La "Epidemia Silente" de la desaparición de mujeres en Perú (y como ha impactado el coronavirus)*, 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bbc.in/2K8Pc3C

BELTRÁN, G. La geolocalización Social. *Polígonos Revista de Geografía* [en línea]. 2015, **27**(1), pp.97-118. [fecha consulta 24 de octubre de 2019]. ISBN 978-84-686-5448-5. Recuperado de: https://bit.ly/3mzqRkJ

BRITOS P. V. (2008). P*rocesos de explotación de información basados en sistemas inteligentes* [en línea]. Tesis doctoral. Universidad Nacional de la Plata, Buenos Aires-Argentina, 2008 [consultado: 19 de setiembre de 2020]. Disponible en: https://bit.ly/34tGBzq.

CABELLO, L., M. Geolocalización a través de direcciones IP. *Revista de Derecho UNED* [en línea]. 2017, **20**(1), pp.283- 301 [fecha consulta 11 de octubre de 2019]. Disponible en: https://bit.ly/3rbhc7H

CABRERA, G. y DE LEON, A. Markovian modeling to identify and to forecast the dynamics of the industrial production index from 1980 to 2018 [Modelización Markoviana para identificar y pronosticar la dinámica del índice de producción industrial d 1980 a 2018]. *EconoQuantum* [en línea]. 2018, **16**(2), pp.23-41 [fecha consulta 24 de octubre de 2020]. Disponible en: https://bit.ly/37Gk6cV

CASTELLANO, J. A. y LÓPEZ, Y. Protocolo de actuación policial en la búsqueda de personas desaparecidas, uso de unidades caninas. *Revista Profesional de Investigación, Docencia y Recursos Didácticos* [en línea]. 2017, **79**(1), pp.565-602 [fecha consulta 13 de octubre de 2019]. Recuperado de https://bit.ly/38lQJvn

CÓRDOVA, D. SANTA MARIA, F. Aplicación del método autor regresivo integrado de medias móviles para el análisis de series de casos de Covid-19 en Perú. *Revista Facultad Humana Universidad Ricardo Palma* [en línea]. 2021, **21**(1), pp. 65-74 [fecha consulta 13 de octubre de 2019]. Recuperado de https://bit.ly/3FO9jMX

COLEGIO DE INGENIEROS DEL PERÚ [en línea]. *Código de Ética de Colegio de Ingenieros del Perú,* 2020 [consulta: 22 de diciembre de 2020]. Disponible en: https://bit.ly/38rVR1d

CORRILLO, M. F. y GUITIÉRREZ, Q. M. Estudio de Localización de un Proyecto. *Ventana científica* [en línea]. 2016, **7**(11), pp.29-33 [fecha consulta 25 de octubre de 2019]. ISSN 2305 – 6010. Recuperado de https://bit.ly/2KJB7JS

DELGADO, R., GUITIÉRREZ, M. y DUEÑAS, J. (2014). Predicción del rendimiento académico en la asignatura de Matemática Básica en la UNALM utilizando la técnica de Regresión Logística Binaria. *Anales Científicos* [en línea]. 2014, **75**(2), pp.310-315 [fecha consulta 27 de octubre de 2020]. ISSN 2519-7398. DOI: https://bit.ly/34yrJjr. Recuperado de https://bit.ly/3rhWG5l

DEL PILAR, A., y VERGEL, M. *Metodologías para el pronóstico de series de tiempo* [en línea]. Tesis de maestría. Universidad Javeriana, Bogotá-Colombia, 2018 [consultado: 04 de diciembre de 2019]. Disponible en: https://bit.ly/3FOdBUz

DÍAZ, F. Sistemas de Geolocalización y Monitoreo de pacientes médicos en alto riesgo. *Revista Pensamiento AMERICANO* [en línea]. 2011, **4**(7), pp.23-27 [fecha consulta 02 de diciembre de 2019]. ISSN: 2027-2448. Disponible en https://bit.ly/2J7z3Ln

DUNDAS BI [en línea]. *Acerca de Dundas BI,* 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/34vqOjJ

EL UNIVERSAL MX [en línea]. *El mexicano promedio mide 1.64 metros y pesa 74 kilos: estudio,* 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/34fHmMw

ESPINOZA, J. J. Implementation of the CRISP-DM methodology for geographical segmentation using a public database [Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública]. *Ingeniería Investigación y Tecnología* [en línea]. 2019. **21**(1), pp.1-17 [fecha consulta 15 de diciembre de 2020]. ISSN 2594-0732. Disponible en: https://bit.ly/3rnuJJt

FABARA, C. P., MALDONADO, D. A., SORIA, M. S. y TOVAR, A. F. Predicción de la Generación para un Sistema Fotovoltaico mediante la aplicación de técnicas de Minería de Datos. *Revista Técnica "energía"* [en línea]. 2019, **16**(1), pp.69-77. [fecha consulta 15 de diciembre de 2020]. ISSN 1390-5074. Disponible en: https://bit.ly/2LQQuAR

FRIGOLÉ, J. Violencia, Riesgo e incertidumbre: La desaparición Forzada en México a Través de la voz de las madres. *Revista de Antropología* [en línea]. 2019, **74**(2), pp.1-19. [fecha consulta 18 de setiembre de 2020]. ISSN: 2659-6881. Disponible en: https://bit.ly/2Jdk2b6

FUNDACIÓN JUSTICIA [en línea]. *Exigimos continúe la búsqueda de personas desaparecidas durante la pandemia COVID 19*, 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/2LI6R2J

HURTADO, F. Segmentación de clientes usando algoritmo de clustering K-Mean [en línea].2005, **20**(2), 123-130. [consulta: 22 de diciembre de 2020]. Disponible en: https://bit.ly/3nOqHr3

GARCÍA, A. y CUNJAMA, E. D. (2020). La desaparición forzada de personas en México. Apuntes para un análisis crítico criminológico. *Cotidiano - Revista de La Realidad Mexicana* [en línea]. 2020, **351**(219), pp.39–52. [fecha consulta 18 de setiembre de 2020]. Disponible en: https://bit.ly/3aCfh5U

GROS, B. y FORÉS, A. (2013). El uso de geolocalización en educación secundaria para la mejora del aprendizaje situado: Análisis de dos estudios de caso. *Revista Latinoamericana de Tecnología Educativa* [en línea]. 2013, **12**(2), pp.41-53. [fecha consulta 14 de octubre de 2020]. ISSN 1695288X. Disponible en: https://bit.ly/3asRzsO

GOBIERNO DE MÉXICO [en línea]. *Registro Nacional de Datos de Personas Extraviadas o Desaparecidas (RNPED),* 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/36Rcmo3

GUZMÁN, J., M. *Notificación electrónica para los actos procesales de comunicación en una entidad supervisora de servicios de salud* [en línea]. Tesis de pregrado. Universidad Cesar Vallejo, Lima-Perú, 2017 [consultado: 04 de diciembre de 2019]. Obtenido de https://bit.ly/2WvaLxU

HERNÁNDEZ, C., L. y DUEÑAS, M., X. Hacia una metodología de gestión del conocimiento basada en minería de datos. *Revista Universidad Garcilaso de la Vega* [en línea]. 2009, **1**(1), pp.79-95 [fecha consulta 19 de diciembre de 2020]. Disponible en: https://bit.ly/38tcaLg

HERNÁNDEZ, R., FERNÁNDEZ, C. y BAPTISTA, P. Metodología de la Investigación (6ª ed.). México DF: McGraw-Hill. 2014, [fecha consulta 19 de diciembre de 2020]. ISBN: 978-1-4562-2396-0

HERNÁNDEZ, R. y MENDÓZA, C. (2018). Metodología de la investigación, las rutas cuantitativa cualitativa y mixta. Ciudad de México, México: Mc Graw Hill. 2014, [fecha consulta 19 de diciembre de 2020]. Disponible en: https://bit.ly/2Kfwho9

LI, X., KANG, Y. y LI, Y. Forecasting with time series imaging [Pronóstico con imágenes de series temporales]. *Scopus* [en línea]. 2020, **160**(1), [fecha consulta 17 de setiembre de 2020]. DOI: 10.1016/j.eswa.2020.113680. Disponible en: https://bit.ly/38qJoum

LEGARRETAETXEBARRIA, A. *Sistema de localización y seguimiento de personas en interiores mediante cámara PTZ basado en las tecnologías Kinect y Ubinese.* Tesis de Grado, Universidad del País Vasco, Lejona,

España, 2011. [consultado: 04 de diciembre de 2019]. Disponible de https://bit.ly/2WBTZ0m

LÓPEZ, P. y FACHELLI, S. *Análisis de regresión logística, Metodología de la investigación Social Cuantitativa.* Bellaterra (Cerdanyola del Vallès): Depósito Digital de documentos, Universidad Autónoma de Barcelona. 1ª edición, 2016. Edición digital: Disponible en: https://bit.ly/34vFJu9

MAMANI, Z., DEL PINO, L. y CORTEZ, A. Minería de datos distribuida usando clustering k-means en la predictibilidad del proceso petitorio en una organización pública. Industrial Data [en línea].2017, **20**(2), 123-130. [consulta: 22 de diciembre de 2020]. Disponible en: https://bit.ly/3nOqHr3

MAMMADOVA, M. y JABRAYILOVA, Z. Development of Methodological and Functional Principles of the Intelligent Demographic Forecasting System [Desarrollo de los principios metodológicos y funcionales de pronóstico demográfico inteligente]. Bakú, Azerbiayán: Springer, 2020. Disponible en: https://bit.ly/37yhFsG

MARIÑO, M., ARANGO, A., LOTERO, L., y JIMÉNEZ. Modelos de series temporales para pronóstico de la demanda eléctrica del sector de explotación de minas y canteras en Colombia. *Revista EIA* [en línea]. 2021, **18**(35), pp. 1-23 [consulta: 27 de setiembre de 2020]. Disponible en: https://bit.ly/3MeOlcJ

MENACHO, C. Comparación de los métodos de series de tiempo y redes neuronales. *Anales Científicos* [en línea]. 2013, **75**(2), pp.245-252 [consulta: 27 de setiembre de 2020]. Disponible en: https://bit.ly/3paQuKy

MICROSOFT [en línea]. *SQL Server 2019: Requisitos de hardware y software,* 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/2GNCDJ8

MICROSOFT EXCEL [en línea]. *Recursos de Microsoft 365 y Office,* 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/30Q3StI

MICROSOFT POWER BI [en línea]. Get Power BI Desktop*,* 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/36VK2RP

MOINE J. M. *Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo* [en línea]. Tesis doctoral. Universidad Nacional de la Plata, Buenos Aires-Argentina, 2013 [consultado: 19 de setiembre de 2020]. Disponible en: https://bit.ly/34tGBzq.

MUNTANÉ, J. Introducción a la investigación básica. *RAPD ONLINE* [en línea]. 2019, **33**(3), pp.221-227 [consulta: 18 de setiembre de 2020]. Disponible en: https://bit.ly/3axNP9p

MUN, J. *Modelación de riesgos, Aplicación de la simulación de Monte Carlo, análisis de opciones reales, pronóstico estocástico, optimización de portafolio, análisis de datos, inteligencia de negocios y modelación de decisiones.* California, United States. IIPER Press. 2016. [fecha consulta 27 de octubre de 2020]. ISSN 1810-9993. DOI: https://bit.ly/3rq2GJk. Recuperado de https://bit.ly/3avpjGg

ÑAUPAS, H., MEJÍA, E., NOVOA, E. y VILLAGÓMEZ, A. Metodología de la investigación Cuantitativa-Cualitativa y Redacción de la Tesis (4ª ed.) Ediciones de la U, Bogotá, Educación. 2014, [fecha consulta 29 de noviembre de 2019]. ISBN: 978-958-762-188-4

ÑAUPAS, H., VALDIVIA, M., R., PALACIOS, J., J., ROMERO, H., E. Metodología de la investigación Cuantitativa-Cualitativa y Redacción de la Tesis (5ª ed.) Ediciones de la U, Bogotá, Educación. 2018, [fecha consulta 29 de noviembre de 2019]. ISBN: 978-958-762-876-0

PALOMINO, J., TORRES, O., ANGULO, Y. Dispositivo basado en modelo ARIMA para predicción de variables ambientales (temperatura, humedad, velocidad del aire) en el área agrícola del departamento del meta. *Gestión y Organización de negocios* [en línea]. 2020, **7**(2), pp. 1-12 [consulta: 23 de setiembre de 2020]. Disponible en: https://bit.ly/3a1HkO7

PÉREZ, J. La Regresión Logística Como Modelo De Predicción Del Riesgo Crediticio en Las Organizaciones De La Economía Social Y Solidaria. *Revista Ciencia Administrativa* [en línea]. 2017, **2**(1), pp. 232-243 [consulta: 23 de setiembre de 2020]. Disponible en: https://bit.ly/3swOeS0

POWER DATA [en línea]. *Procesos ETL: Definición, Características, Beneficios y Retos*, 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/3mpZg5w

REAL ACADEMIA ESPAÑOLA [en línea]. *Precisión*, 2020a [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/3MxOnfF

REAL ACADEMIA ESPAÑOLA [en línea]. *Efecto*, 2020b [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/3lw7I4Y

REAL ACADEMIA ESPAÑOLA [en línea]. *Pronóstico*, 2020c [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/3NHiamD

REAL ACADEMIA ESPAÑOLA [en línea]. *Impacto*, 2020d [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/3yURfiG

RAMOS, L., F. El uso de las licencias libres en los datos públicos abiertos. *Revista Española de Documentación científica* [en línea]. 2016, **40**(3), pp.1-17 [consulta: 23 de setiembre de 2020]. Disponible en: https://bit.ly/2KJvlrE

REYES, J., ESCOBAR, C., DUARTE, J., RAMÍREZ. Una aplicación del modelo de regresión logística en la predicción del rendimiento estudiantil. *Revista de Estudios Pedagógicos XXXII.* [en línea]. 2007, **2**(1), pp.101-120 [consulta: 23 de setiembre de 2020]. Disponible en: https://bit.ly/3Mj7lH4

RODRÍGUEZ, J., J., ESCOBAR, N., E., B. y RAMÍREZ, N., G. Estrategias del Estado mexicano para minimizar los feminicidios. *Revista Estudios Feministas* [en línea]. 2020, **28**(1), pp.1-12 [consulta: 18 de setiembre de 2020]. DOI: 10.1590/1806-9584-2020v28n157811 Disponible en: https://bit.ly/3rfN62N

ROQUE, I., L. *Análisis comparativo de técnicas de minería de datos para la predicción de ventas* [en línea]. Tesis de pregrado. Universidad de Señor de Sipán, 2016 [consultado: 04 de noviembre de 2020]. Disponible en: https://bit.ly/2INWzgp

RSTUDIO [en línea]. *Documentación técnica de R-Studio*, 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/2FfSfog

SÁNCHEZ, F., A. Fundamentos Epistémicos de la Investigación Cualitativa y Cuantitativa: Consenso y Disensos. *Revista digital de Investigación* [en línea]. 2018, **13**(1), pp.102-122 [consulta: 24 de setiembre de 2020]. Disponible en: https://bit.ly/3rkh3i0

SÁNCHEZ, L., CABANAS, G, A., TORRES, V. Utilización de modelos ARIMA para la predicción de la producción de leche. Estudio de caso en la UBPC "Maniabo", Las Tunas. *Revista Cubana de Ciencia Agrícola.* [en línea]. 2014, **48**(3), pp. 2013-2018 [consulta: 24 de setiembre de 2020]. Disponible en: https://bit.ly/3Niwa61

SCHULZ, Ch. y SALAZAR, M. La búsqueda de la verdad: Una necesidad para tejer el futuro. La búsqueda de personas desaparecidas desde una perspectiva de los derechos humanos. *El cotidiano 209* [en línea]. 2018, **1**(1), pp.103-109 [consulta: 24 de setiembre de 2020]. Disponible en: https://bit.ly/3mBsQFj

SEVERIUKHINA, O., KESAREV, S., BOCHENINA, K., BOUKHANOVSKY, A., LEES, M. y SLOOT, P. Large-Scale forecasting of information spreading [Predicción a gran escala de difusión de la información]. *Journal of Big Data* [en línea]. 2020, **7**(1), pp.1-17 [consulta: 17 de setiembre de 2020]. Disponible en: https://bit.ly/3aw7J4D

SIFUENTES, O. (2018). Modelos predictivos de la deserción estudiantil en una universidad privada peruana. *Revista industrial Data* [en línea]. 2020, **21**(2), pp.47-52 [consulta: 18 de setiembre de 2020]. Disponible en: https://bit.ly/38ilUHY

SOSA, J., ROQUE, E. y PALOMINO, R. Modelo de pronóstico ARIMA en el monitoreo de parámetros ambientales. *Ciencia y Desarrollo-Universidad Alas Peruanas* [en línea]. 2019, **21**(2), pp.49-59 [consulta: 01 de noviembre de 2020]. Disponible en: https://bit.ly/3nxWyMC

TABLEAU [en línea]. *¿Qué es Tableau?*, 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://tabsoft.co/3pbS0ft

TIMARÁN, S. R., HERNÁNDEZ, I., CAICEDO, S. J., HIDALGO, A. y ALVARADO, J. C. *El proceso de descubrimiento de conocimiento en bases*

de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional, 2016. Bogotá: Ediciones Universidad Cooperativa de Colombia, 63-86. Disponible en: https://bit.ly/3p3npAC

UNIVERSIDAD CÉSAR VALLEJO [en línea]. *RESOLUCIÓN DE CONSEJO UNIVERSITARIO N° 0262-2020/UCV*, 2020 [consulta: 22 de diciembre de 2020]. Disponible en: https://bit.ly/2M4cLvn

UNODC [en línea]. Plan de Investigación para el delito de Desaparición Forzada de Personas Bogotá, Colombia, Scripto, 2010 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/3h4hkRw

VANDEPUT, N. [en línea]. *KPI de Pronóstico: RMSE, MAE, MAPE y segó*, 2020 [consulta: 04 de diciembre de 2020]. Disponible en: https://bit.ly/2Wv12I9

VIALARDI, C., CHUE, J., PECHE, J., ALVARADO, G., VINATEA, B., ESTRELLA, J. y ORTIGOSA, Á. A data mining approach to guide students through the enrollment process based on academic performance [Un enfoque de minería de datos para guiar a los estudiantes a través del proceso de inscripción basado en el rendimiento académico]. *User modeling and user-adapted interaction*. [en línea]. 2011, Vol. **21** (1). 217-248. [fecha de consulta 17 de diciembre del 2020]. DOI 10.1007/s11257-011-9098-4. Disponible en https://bit.ly/3my409n

YAVUZ, Ö., KARAHOCA, A. y KARAHOCA, D. A data mining approach for desire and intention to participate in virtual communities [Un enfoque de minería de datos para el deseo y la intención de participar en comunidades virtuales]. *International Journal of Electrical and Computer Engineering(IJECE)* [en línea]. Turkey, 2019, **9**(5), pp.3714-3719 [consulta: 11 de noviembre de 2019]. ISSN:2088-8708. DOI: 10.11591/ijece.v9i5.pp3714-3719. Disponible en: https://bit.ly/37xQQ7Y

ZITA, A [en línea]. Exactitud y precisión, 2014. [consulta: 01 de noviembre de 2019]. Disponible en: https://bit.ly/3h7jiAB

ANEXOS

Anexo 1: Matriz de operacionalización de la variable

En esta tabla se muestra el proceso de operacionalización en el que se puede apreciar la transformación de la variable teórica o constructos, en dimensiones y estas en indicadores e índices (Ñaupas et al*.,* 2014, p. 191).

Tabla 44 *Matriz de operacionalización*

Variable	Definición Conceptual	Definición Operacional	Dimensión	Indicadores	Instrumentos
Efecto del uso del sistema geolocalizado en la precisión de pronóstico de casos de personas desaparecidas	Castellano et al*.* (2017) concluyeron que hay muchos casos de personas desaparecidas en los que el resultado de la búsqueda, por falta de medios y en otros por falta de organización y conocimiento de los recursos, los resultados no han sido los esperados (p. 565). Gros et al*.* (2013) añadieron que el sistema es un novedoso enfoque que busca desarrollar modelos de propuestas para la implementación de proyectos de aprendizaje basados en la geolocalización del contenido con la voluntad de servir a la comunidad (p. 42).	Un sistema que permitirá dar un pronóstico mediante técnicas que ayuden en los resultados, ya que también con estos mismos se podrá reconocer los rasgos de las personas que pueden ser propensas a algunos casos de secuestro o desaparición.	Precisión (Zita, 2018, π, 8-9; Legarretaetxebarria, 2011, p. 49)	$PseudoR^2$ (López et al*.,* 2016) MAE (Error Absoluto Medio) (Vandeput, 2019, π. 13)	Hoja de tabulación (Gobierno de México, 2020)

Anexo 2: Matriz de consistencia

Tabla 45 *Matriz de consistencia*

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLE	INDICADORES	MÉTODO
GENERAL	**GENERAL**	**GENERAL**			
¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el uso de la regresión logística binaria y modelos de series de tiempo?	Determinar la precisión del pronóstico de casos de personas desaparecidas con la regresión logística binaria y los modelos de series de tiempo.	El uso de la regresión logística binaria y modelos de series de tiempo mejoró la precisión del pronóstico de casos de personas desaparecidas.		-	**TIPO DE INVESTIGACIÓN** Investigación Aplicada
ESPECIFICO	**ESPECIFICO**	**ESPECIFICO**		**INDICADORES**	**DISEÑO DE INVESTIGACIÓN**
¿Cuáles son las relaciones entre el rango de edad, la presencia de acné y verrugas, los bigotes, la complexión, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares con los casos de desaparición de varones?	Determinar las relaciones entre el rango de edad, la presencia de acné y verrugas, los bigotes, la complexión, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares con los casos de desaparición de varones.	La complexión, el rango de edad, la presencia de acné y verrugas, los bigotes, las cicatrices, los lunares, los tatuajes, los problemas con la dentadura y otras señas particulares están relacionados con los casos de desaparición de varones (Delgado et al., 2014; Pérez, 2017; Reyes et al., 2007)	Efecto del uso del sistema geolocalizado en la precisión de pronóstico de casos de personas desaparecidas	$PseudoR^2$ López et al. (2016)	Transversal Correlacional Causal Longitudinal de Tendencia **ENFOQUE DE INVESTIGACIÓN** Cuantitativo
¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el modelo ARIMA?	Determinar la precisión del pronóstico de casos de personas desaparecidas con el modelo ARIMA.	El uso del modelo ARIMA mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 95% (Sosa et al., 2019; Sánchez et al., 2014; Córdova et al., 2021; Palomino et al., 2020)			**POBLACIÓN** Base de datos obtenida de RNPED del país de los estados unidos mexicanos (Gobierno de México, 2020)
¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el modelo Holt-Winters?	Determinar la precisión del pronóstico de casos de personas desaparecidas con el modelo Holt-Winters.	El uso del modelo Holt-Winters mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 85% (Roque, 2016; Mariño et al., 2021; Banda et al., 2014)		MAE (Error Absoluto Medio) Vandeput ,2019, π. 13)	
¿Cuánto fue la mejora en la precisión del pronóstico de casos de personas desaparecidas con el modelo ETS?	Determinar la precisión del pronóstico de casos de personas desaparecidas con el modelo ETS.	El uso del modelo ETS mejoró la precisión del pronóstico de casos de personas desaparecidas por lo menos del 90% (Roque, 2016; Del Pilar et al., 2018)			**MUESTRA** Son 36157 datos (Gobierno de México, 2020)

Anexo 3: Conglomerado de resultados los modelos de series de tiempo

En el capítulo cuatro de resultados se detalló al país de México, usando la regresión logística binaria ya detallada y los métodos ARIMA, Holt-Winters y ETS. A continuación, se detallará el conglomerado de los resultados de los estados en general. Con los resultados finales el modelo que resultó el más dominante fue el modelo ARIMA con 17 modelos finales y ETS en segundo lugar con 16 modelos. Por lo que se puede concluir que aplicando ya sea el modelo ARIMA o ETS los resultados serán confíales.

Tabla 46 *Resultado de la precisión y selección de los modelos finales*

N°	Nombre	Precisión			Error			Modelo Final
		ARIMA	Holt-Winters	ETS	ARIMA	Holt-Winters	ETS	
		Precisión	Precisión	Precisión	Error	Error	Error	
0	País de México	75.7949	74.1684	77.1835	24.2051	25.8316	22.8165	ETS con un 77.1835% de precisión
1	Aguas Calientes	98.9603	98.8254	98.9446	1.0397	1.1746	1.0554	ARIMA con un 98.9603% de precisión
2	Baja California	96.4372	95.8943	96.5892	3.5628	4.1057	3.4108	ETS con un 96.5892% de precisión
3	Baja California Sur	99.5907	99.5165	99.5366	0.4093	0.4835	0.4634	ARIMA con un 99.5907% de precisión
4	Campeche	99.7065	99.6313	99.6689	0.2935	0.3687	0.3311	ARIMA con un 99.7065% de precisión
5	Chiapas	99.1750	99.2842	99.3127	0.8250	0.7158	0.6873	ETS con un 99.3127% de precisión
6	Chihuahua	96.3712	96.0323	96.6594	3.6288	3.9677	3.3406	ETS con un 96.6594% de precisión
7	Ciudad de México	97.5065	97.2269	97.5208	2.4935	2.7731	2.4792	ETS con un 97.5208% de precisión
8	Coahuila de Zaragoza	96.0665	95.4563	95.9701	3.9335	4.5437	4.0299	ARIMA con un 96.0665% de precisión
9	Colima	98.2944	98.1808	98.4048	1.7056	1.8192	1.5952	ETS con un 98.4048% de precisión
10	Durango	98.5487	98.3464	98.5644	1.4513	1.6536	1.4356	ETS con un 98.5644% de precisión
11	Estado de México	94.9840	94.6290	94.8239	5.016	5.371	5.1761	ARIMA con un 94.984% de precisión
12	Guanajuato	98.0964	98.1065	98.1934	1.9036	1.8935	1.8066	ETS con un 98.1934% de precisión

N°	Nombre	Precisión			Error			Modelo Final
		ARIMA	Holt-Winters	ETS	ARIMA	Holt-Winters	ETS	
		Precisión	Precisión	Precisión	Error	Error	Error	
13	Guerrero	97.6868	97.3287	97.6540	2.3132	2.6713	2.346	ARIMA con un 97.6868% de precisión
14	Hidalgo	99.1659	99.0161	99.1569	0.8341	0.9839	0.8431	ARIMA con un 99.1659% de precisión
15	Jalisco	95.6587	95.2525	95.6452	4.3413	4.7475	4.3548	ARIMA con un 95.6587% de precisión
16	Michoacán	97.5704	97.2656	97.5479	2.4296	2.7344	2.4521	ARIMA con un 97.5704% de precisión
17	Morelos	99.1037	98.9587	99.1003	0.8963	1.0413	0.8997	ARIMA con un 99.1037% de precisión
18	Nayarit	99.4274	99.2005	99.4310	0.5726	0.7995	0.569	ETS con un 99.431% de precisión
19	Nuevo León	94.5605	93.6343	94.5159	5.4395	6.3657	5.4841	ARIMA con un 94.5605% de precisión
20	Oaxaca	98.2254	98.159	98.2173	1.7746	1.841	1.7827	ARIMA con un 98.2254% de precisión
21	Puebla	97.2986	97.2821	97.4554	2.7014	2.7179	2.5446	ETS con un 97.4554% de precisión
22	Querétaro	98.9759	98.8502	98.9739	1.0241	1.1498	1.0261	ARIMA con un 98.9759% de precisión
23	Quintana Roo	99.6477	99.5409	99.6374	0.3523	0.4591	0.3626	ARIMA con un 99.6477% de precisión
24	San Luis de Potosí	98.8714	98.8665	98.7433	1.1286	1.1335	1.2567	ARIMA con un 98.8714% de precisión
25	Sinaloa	95.6098	95.4232	95.7340	4.3902	4.5768	4.266	ETS con un 95.734% de precisión
26	Sonora	96.0049	95.7911	96.0407	3.9951	4.2089	3.9593	ETS con un 96.0407% de precisión
27	Tabasco	99.5946	99.4523	99.5858	0.4054	0.5477	0.4142	ARIMA con un 99.5946% de precisión
28	Tamaulipas	89.5699	87.7856	89.0377	10.4301	12.2144	10.9623	ARIMA con un 89.5699% de precisión
29	Tlaxcala	99.7122	99.7014	99.7122	0.2878	0.2986	0.2878	ETS con un 99.7122% de precisión
30	Veracruz	97.6476	97.3382	97.7497	2.3524	2.6618	2.2503	ETS con un 97.7497% de precisión
31	Yucatán	99.6036	99.6086	99.6110	0.3964	0.3914	0.3890	ETS con un 99.611% de precisión
32	Zacatecas	98.4269	98.2480	98.4537	1.5731	1.7520	1.5463	ETS con un 98.4537% de precisión

Anexo 4: Conglomerado de resultados la regresión logística binaria

En esta tabla se muestra los resultados finales de la regresión logística binaria la cual dio por resultado que el rango de edad es la que afecta más a la población las cicatrices y los lunares y las que menos afectan fueron la complexión, acné, verruga, lunares, problemas con la dentadura y otras señas.

Tabla 47 *Resultado final de la regresión logística binaria*

N°	Nombre	Regresión Logística Binaria								
		Complexión	Edad	Acné, Verruga	Bigote	Cicatrices	Lunares	Tatuajes	Problemas Dentadura	Otras Señas
0	País de México		R	R	R	R		R	R	R
1	Aguas Calientes		R			R		R		
2	Baja California		R		R	R		R		
3	Baja California Sur	R	R							
4	Campeche	R	R			R	R	R		R
5	Chiapas								R	R
6	Chihuahua		R		R	R		R		
7	Ciudad de México		R		R	R		R		
8	Coahuila de Zaragoza		R		R	R	R	R		
9	Colima		R			R		R		
10	Durango		R					R		R
11	Estado de México		R		R			R		
12	Guanajuato		R		R	R		R		
13	Guerrero		R					R	R	R
14	Hidalgo		R							
15	Jalisco		R		R	R		R		
16	Michoacán		R			R		R	R	R
17	Morelos		R							
18	Nayarit		R							
19	Nuevo León		R		R	R		R		R
20	Oaxaca	R	R	R	R	R	R	R	R	R
21	Puebla		R		R	R		R		
22	Querétaro		R					R		
23	Quintana Roo		R							
24	San Luis de Potosí			R						

N°	Nombre	Regresión Logística Binaria								
		Complexión	Edad	Acné, Verruga	Bigote	Cicatrices	Lunares	Tatuajes	Problemas Dentadura	Otras Señas
25	**Sinaloa**		R		R	R		R		
26	**Sonora**		R			R	R	R		
27	**Tabasco**		R							
28	**Tamaulipas**		R	R	R	R	R	R	R	R
29	**Tlaxcala**	R	R	R	R	R	R	R	R	R
30	**Veracruz**		R		R	R		R		R
31	**Yucatán**	R	R	R	R	R	R	R	R	R
32	**Zacatecas**		R					R		

Tabla 48 *Leyenda del cuadro final de la regresión logística binaria*

Relacionado	R

Anexo 5: Diseño de la base de datos

En la figura 87 se muestra el modelo dimensional de la base de datos, esta fue obtenida después de realizar la mayor parte de los procesos de la metodología que se desarrolló en la investigación, la cual fue CRISP-DM.

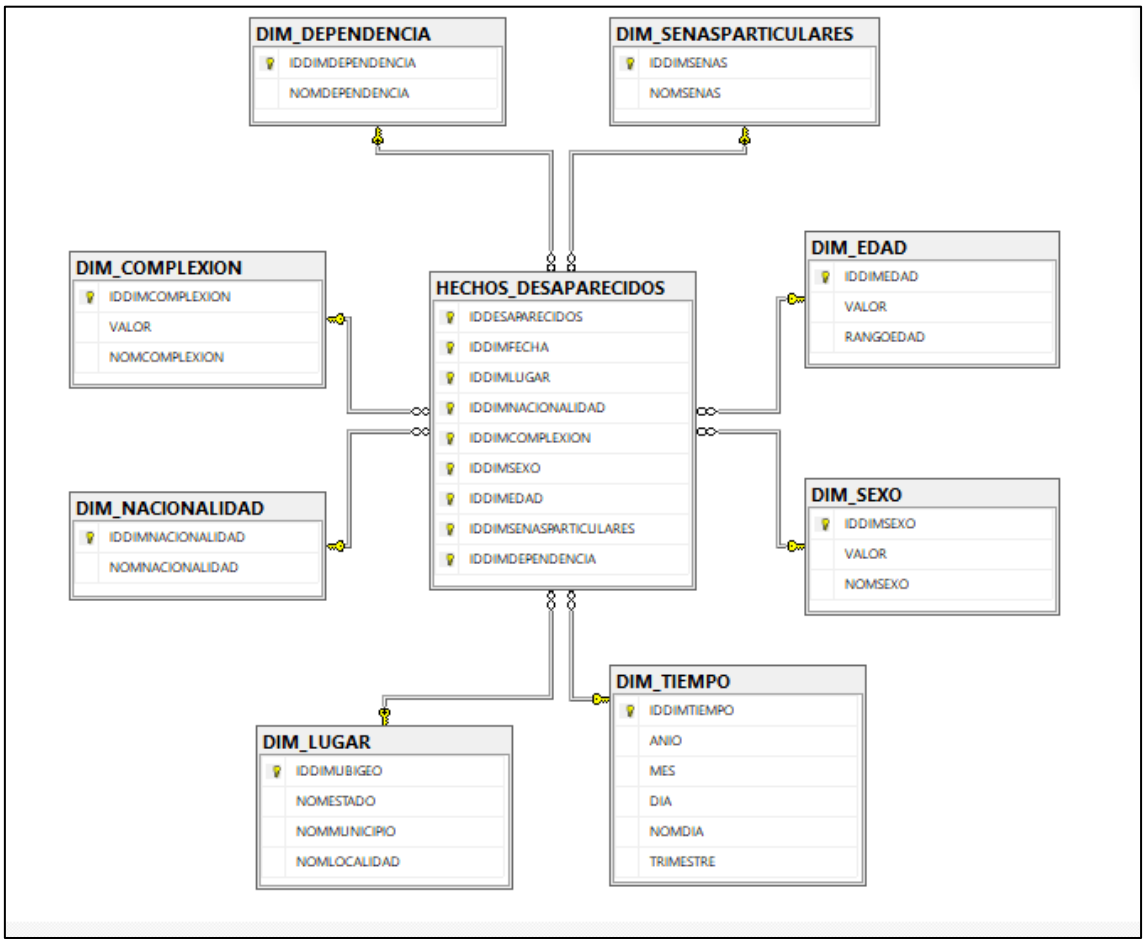


Figura 87. Pantalla del diagrama de las dimensiones

Anexo 6: Descripción de las dimensiones

En la tabla 49 se describe la descripción de todas las dimensiones estas se dividen en 09 tablas donde la última viene tabla de hechos, en la tabla 01 se describe el nombre de la dependencia con su respectiva ID para su identificación oportuna, en la tabla 02 se describen la seña particular y su ID respectivo, en la tabla 03 se tiene la edad donde se tiene el valor, el rango de edad ya detallado anteriormente y su ID respectivo, en tabla 04 se tiene el tipo del sexo de las personas, en tabla 05 se tiene al tiempo donde se divide en año, mes día, nombre del día y el trimestre, en la tabla 06 el lugar se subdivide con el nombre del Estado, el nombre del municipio y el nombre de la localidad, en la tabla 07 la nacionalidad se subdivide en el nombre de la nacionalidad por lo que ya se detalló anteriormente se tiene más nacionalidades no solo la mexicana, en la tabla 08 se tiene el tipo de complexión y en los hechos se indica los ID de cada tabla para su respectiva unión de cada una.

Tabla 49 *Descripción de todas las dimensiones*

TABLA 01	**"DIM_DEPENDENCIA"**			
Descripción de la Tabla	Esta dimensión aloja a las dependencias de los estados del país de México			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMDEPENDENCIA	INT	-	PK	Llave primaria de la tabla
NOMDEPENDENCIA	VARCHAR	100	-	Nombre de la dependencia

TABLA 02	**"DIM_SENASPARTICULARES"**			
Descripción de la Tabla	Esta dimensión aloja las diversas señas dentro de los datos			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMSENAS	INT	-	PK	Llave primaria de la tabla
NOMSENAS	VARCHAR	100	-	Clasificación de las señas

TABLA 03	**"DIM_EDAD"**			
Descripción de la Tabla	Esta dimensión aloja el rango de edad			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMEDAD	INT	-	PK	Llave primaria de la tabla
VALOR	INT	-	-	Valor calificador
RANGOEDAD	VARCHAR	50		Rango de edad

TABLA 04	**"DIM_SEXO"**			
Descripción de la Tabla	Esta dimensión aloja el sexo de las personas			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMSEXO	INT	-	PK	Llave primaria de la tabla
VALOR	INT	-	-	Valor calificador
NOMSEXO	VARCHAR	50	-	Descripción del sexo de las personas

TABLA 05	"DIM_TIEMPO"			
Descripción de la Tabla	Esta dimensión aloja los tiempos en diversos formatos			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMTIEMPO	INT	-	PK	Llave primaria de la tabla
ANIO	INT	-	-	Formato en años
MES	CHAR	2	-	Formato en meses
DIA	CHAR	2	-	Formato en días
NOMDIA	VARCHAR	50	-	Formato en nombre de días
TRIMESTRE	INT	-	-	Formato en trimestre

TABLA 06	"DIM_LUGAR"			
Descripción de la Tabla	Esta dimensión aloja el Estado, municipio y la localidad dentro del país			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMDEPENDENCIA	INT	-	PK	Llave primaria de la tabla
NOMESTADO	VARCHAR	200	-	Nombre del Estado
NOMMUNICIPIO	VARCHAR	200	-	Nombre del municipio
NOMLOCALIDAD	VARCHAR	200	-	Nombre de la localidad

TABLA 07	"DIM_NACIONALIDAD"			
Descripción de la Tabla	Esta dimensión aloja a las la nacionalidad de las personas desaparecidas			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMDEPENDENCIA	INT	-	PK	Llave primaria de la tabla
NOMNACIONALIDAD	VARCHAR	50	-	Nombre de la nacionalidad

TABLA 08	"DIM_COMPLEXION"			
Descripción de la Tabla	Esta dimensión aloja a las dependencias de los estados del país de México			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDIMDEPENDENCIA	INT	-	PK	Llave primaria de la tabla
VALOR	INT	-	-	Valor calificador
NOMCOMPEXION	VARCHAR	50		Clasificación de la complexión

TABLA 09	"HECHOS_DESAPARECIDOS"			
Descripción de la Tabla	Esta dimensión aloja a las dependencias de los estados del país de México			
Campos	**Tipo de Datos**	**Tamaño**	**(PK) o (FK)**	**Descripción**
IDDESAPARECIDOS	INT	-	PK	Llave primaria de la dimensión desaparecidos
IDDIMFECHA	INT	-	FK	Llave foránea de la dimensión fecha
IDDIMLUGAR	INT	-	FK	Llave foránea de la dimensión lugar
IDDIMNACIONALIDAD	INT	-	FK	Llave foránea de la dimensión nacionalidad
IDDIMCOMPLEXION	INT	-	FK	Llave foránea de la dimensión complexión
IDDIMSEXO	INT	-	FK	Llave foránea de la dimensión sexo
IDDIMEDAD	INT	-	FK	Llave foránea de la dimensión edad
IDDIMSENASPARTICULARES	INT	-	FK	Llave foránea de la dimensión señas particulares
IDDIMDEPENDENCIA	INT	-	FK	Llave foránea de la dimensión dependencia

Anexo 7: Arquitectura tecnológica

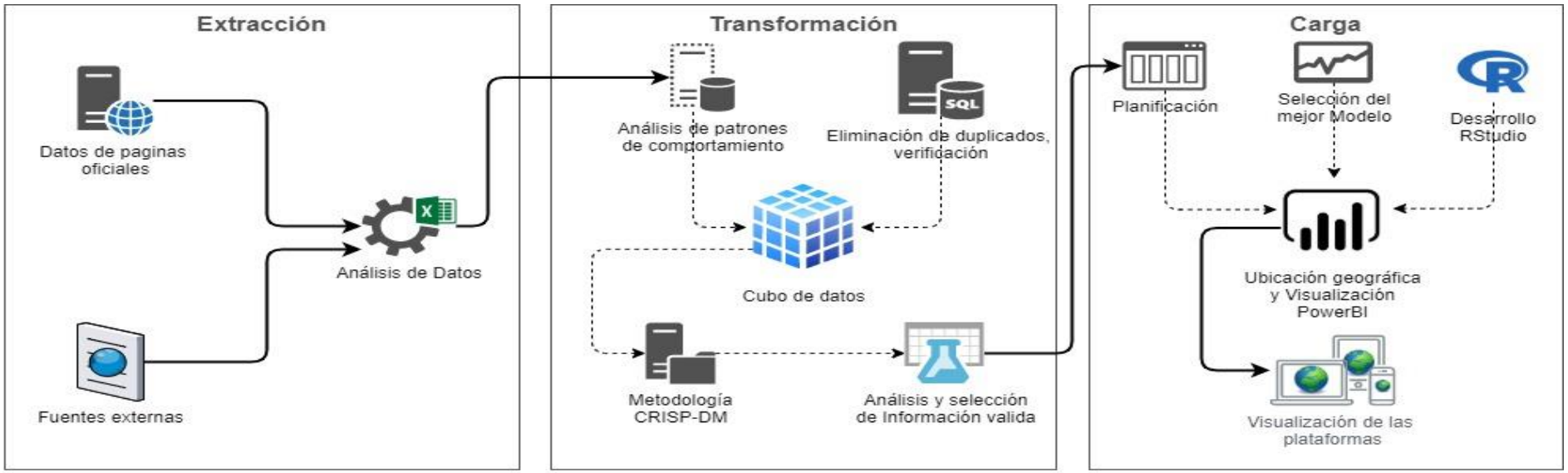


Figura 88. Pantalla de la arquitectura tecnológica

En el presente trabajo de investigación se realizó el siguiente diagrama el cual fue realizado con la herramienta drwa.io. Este benefició para el desarrollo de la arquitectura tecnológica el cual se dividió en tres fases; la primera es la etapa de extracción de información de nuestra fuente (RNPED), en la segunda se tiene la transformación donde se realiza un análisis más profundo con los patrones de comportamiento, eliminación de datos duplicados y verificación de datos que se usaran para el análisis, bajo la metodología seleccionada (CRISP-DM), por último en la etapa carga, se muestra la información y predicciones mediante los modelos estudiados que culminando con la comparación respectiva serán mostradas en la plataforma de Power BI la cual fue escogida mediante una clasificación de plataformas, desarrolladas en el entorno de RStudio el cual ha permitido realizar la técnica de regresión binaria y los modelos de series de tiempo(ARIMA, Holt-Winters, ETS).

Anexo 8: Metodología CRISP-DM

La metodología CRISP-DM se divide en seis fases las cuales de acuerdo a Espinoza (2019) son las siguientes:

- Fase 1: Comprensión del problema o negocio “Esta es la etapa más importante, ya que, si no se tiene una correcta comprensión del problema, o negocio, de nada servirán las etapas siguientes” (Espinoza, 2019, p. 3). Dentro de esta fase se subdivide en sub-fases las cuales son:
  - Fase 1-1: Identificación del problema “Consiste en entender y delimitar la problemática, así como identificar los requisitos, supuestos, restricciones y beneficios del proyecto” (Espinoza, 2019, p. 3).
  - Fase 1-2: Determinación de Objetivos “Puntualiza las metas a lograr al proponer una solución basada en un modelo de minería de datos” (Espinoza, 2019, p. 3).
  - Fase 1-3: Evaluación de la situación actual “Específica el Estado actual antes de implementar la solución de minería de datos propuesta, a fin de tener un punto de comparación que permita medir el grado de éxito del proyecto” (Espinoza, 2019, p. 3).
- Fase 2: Comprensión de los datos: En esta fase se incluye las actividades principales las cuales se subdivide en sub-fases las cuales son:
  - Fase 2-1: Recolección de datos “Consiste en obtener los datos a utilizar en el proyecto identificando las fuentes, técnicas empleadas en su recolección los problemas encontrados en su obtención y la forma como se resolvieron los mismos” (Espinoza, 2019, p. 3).
  - Fase 2-2: Descripción de datos “Identifica el tipo, formato, volumetría y significado de cada dato” (Espinoza, 2019, p. 3).
  - Fase 2-3: Exploración de datos “Radica en aplicar pruebas estadísticas básicas que permitan conocer las propiedades de los datos a fin de entenderlos lo mejor posible” (Espinoza, 2019, p. 3).

- Fase 3: Preparación de los datos “Esta es la etapa que consume más tiempo en el proyecto y es donde se seleccionan los datos que se transforman de acuerdo con los resultados de la etapa anterior a fin de utilizarlos en la etapa de modelado” (Espinoza, 2019, p. 3). Dentro de esta se subdivide en sub-fases las cuales son:
  - Fase 3-1: Limpieza de datos “Se aplican diferentes técnicas para la respectiva depuración” (Espinoza, 2019, p. 3).
  - Fase 3-2: Creación de indicadores “Genera indicadores que potencian la capacidad predictiva de los datos a partir de los datos existentes y ayudan a detectar comportamientos interesantes para modelar” (Espinoza, 2019, p. 3).
  - Fase 3-3: Transformación de datos “Cambia el formato o estructura de ciertos datos sin modificar su significado, a fin de aplicarles alguna técnica particular en etapa de modelado” (Espinoza, 2019, p. 3).
- Fase 4: Modelado “En esta etapa se obtiene propiamente el modelo de minería de datos” (Espinoza, 2019, p. 3). Dentro de esta se subdivide en sub-fases las cuales son:
  - Fase 4-1: Selección de Técnica de modelado “Elige la técnica apropiada de acuerdo con el problema a resolver, los datos disponibles, las herramientas de minería de datos disponibles, así como el dominio de la técnica elegida” (Espinoza, 2019, p. 3).
  - Fase 4-2: Selección de datos de prueba: “En algunos tipos de modelos se requiere dividir la muestra en datos de entrenamiento de validación” (Espinoza, 2019, p. 3).
  - Fase 4-3: Obtención del modelo: “Genera el mejor modelo mediante un proceso iterativo de modificación de parámetros del mismo” (Espinoza, 2019, p. 3).

- Fase 5: Evaluación del Modelo, en esta etapa se determina la calidad del modelo con base en el análisis de ciertas métricas estadísticas del mismo, comparando los resultados con resultados previos, o bien, analizando los resultados con apoyo de expertos en el dominio del problema; de acuerdo con los resultados de esta etapa se determina seguir con la última fase de la metodología, regresar a alguna de las etapas anteriores o incluso partir de cero con un nuevo proyecto (Espinoza, 2019, p. 4).
- Fase 6: Implementación del Modelo, esta etapa explota, mediante acciones concretas, el conocimiento adquirido mediante el modelo; aquí también es importante documentar los resultados de manera clara para el usuario final y asegurarse de que todas las etapas de la metodología se documenten debidamente para hacer una revisión del proyecto a fin de obtener lecciones aprendidas durante el proceso. Asimismo, monitorear las acciones para detectar áreas de oportunidad o incluso nuevos problemas (Espinoza, 2019, p. 4).

Anexo 9: Metodología KDD

De acuerdo a Timarán et al. (2016), la metodología KDD se divide en cinco etapas:

- Etapa 1: Selección en esta etapa, una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento, la selección de los datos varía de acuerdo con los objetivos del negocio (Timarán et al., 2016, p. 65).
- Etapa 2: Procesamiento/Limpieza: “En esta se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos, se selecciona estrategias para el manejo de datos desconocidos” (Timarán et al., 2016, p. 65).
- Etapa 3: Transformación/reducción “En esta etapa se buscan características útiles para representar los datos dependiendo de la meta del proceso” (Timarán et al., 2016, p. 66).
- Etapa 4: Minería de datos (data mining) “El objetivo de esta fase es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento” (Timarán et al., 2016, p. 66).
- Etapa 5: Interpretación/evaluación “Se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones” (Timarán et al., 2016, p. 65).

Anexo 10: Metodología SEMMA

La metodología SEMMA se divide en cinco etapas de acuerdo a Hernández et al. (2009) son las siguientes:

- Etapa 1: Muestreo “Dentro de esta fase se realiza la extracción de una muestra de los datos que permita representar las características comunes de la población para posteriormente comenzar el análisis de los mismos” (Hernández, 2009, p. 83).
- Etapa 2: Exploración “A través de esta fase usando técnicas estadísticas permite realizar un seguimiento a los mismos logrando detectar, identificar y posteriormente eliminar datos que presenten anomalías o deficiencias en las fases siguientes hacia el descubrimiento de información” (Hernández, 2009, p. 83).
- Etapa 3: Modificación “Se realiza una selección y transformación de los datos de acuerdo a las variables seleccionadas para el proceso de minado” (Hernández, 2009, p. 84).
- Etapa 4: Modelado “En este punto se hace uso de herramientas de software que permitan la utilización de técnicas y métodos propios de la minería de datos” (Hernández, 2009, p. 84).
- Etapa 5: Evaluación “Dentro de esta fase se evalúa los resultados obtenidos dentro de la etapa anterior para verificar el éxito del proyecto” (Hernández, 2009, p. 84).

Anexo 11: Metodología Catalyst

La metodología Catalyst está compuesta por dos modelos, de acuerdo a Aquino et al. (2015) citando a Britos (2008) son las siguientes:

- Modelo 1: Circunstancias del negocio
  - Modelo 1-1: Dato “El proyecto comienza con un conjunto de datos con el objetivo de explorarlos para encontrar patrones de interés” (Aquino et al., 2015, p. 280). Modelo 1-2: Oportunidad es “donde el proyecto inicia como un problema u oportunidad de negocio que debe ser explorada” (Aquino et al., 2015, p. 281). Modelo 1-3: Prospectiva es “El objetivo del proyecto, descubrir donde la minería de datos puede ofrecer un valor a la organización” (Aquino et al., 2015, p. 281). Modelo01-4: Definido en esta fase “el proyecto comienza con la premisa de crear las especificaciones del modelo de minería de datos con un propósito específico” (Aquino et al., 2015, p. 281). Modelo 1-5: Estratégico es donde se “comienza con una estrategia de análisis para dar soporte a un escenario planificado por la organización” (Aquino et al., 2015, p. 281).
- Modelo 2: Explotación de información este modelo de acuerdo a Aquino et al. (2015) citando a Moine (2013) comprende de los siguientes pasos:
  - Modelo 2-1: Preparación de los datos el “cual incluye una serie de actividades que permiten comprobar la calidad de los datos a utilizar” (Aquino et al., 2015, p. 281). Modelo02-2: Selección de herramientas y modelado inicial. Modelo 2-3: Refinar el modelo seleccionado. Modelo02-4: Implementar el modelo. Modelo 2-5: Comunicación de resultados “Donde se presenta el resultado obtenido al público interesado y los responsables de la toma de decisiones” (Aquino et al., 2015, p. 281).

Anexo 12: Selección de la metodología de desarrollo

Al terminar de analizar las cuatro metodologías, las cuales estas se basan en el data mining se procedió a realizar una clasificación de la cual se quedaría con la que tenga mayor puntaje basándonos en diversos criterios como son: en la fase de comprensión del problema en la cual la metodología CRISP-DM tiene un mayor alcance y mayor detalle, cada metodología tiene diversas fases pero entre sí tienen varias relaciones por lo que al ver exhaustivamente y calificando cada fase, el puntaje más alto se llevó la metodología CRISP-DM con 28 puntos por lo que está más detallada en sus fases y en la actualidad es la metodología que se trabaja más para minería de datos.

Tabla 50 *Comparación de metodologías para el desarrollo*

	CRISP-DM	**KDD**	**SEMMA**	**Catalyst**
Compresión del Problema	**4**	3	3	4
Está orientada a la minería de datos	**5**	5	4	4
Esta detalladamente definido	**5**	4	4	3
Comprensión de datos	**5**	4	4	4
Modelado de datos	**4**	4	4	4
Implementación	**5**	4	4	4
Total	**28**	24	23	23

Tabla 51 *Leyenda de la clasificación de las metodologías*

Muy Bueno	5
Bueno	4
Regular	3
Malo	2
Muy Malo	1

Anexo 13: Selección de la plataforma de desarrollo

A continuación, se visualiza una tabla de las comparaciones entre las plataformas en donde se puede realizar el front end (visual) del sistema en donde se tiene, la plataforma Power BI como dominante entre Tableau y Dundas BI por diversas razones una de las cuales son que su implementación de las series de tiempo con el R es más sencillo a comparación de las otras, la visualización, la posición en el mercado y Power BI es una herramienta de Microsoft y por ende es un poco más conocida, a la vez tiene una amplia gama de librerías y extensiones que permite tener un dashboard (tablero) profesional y excelentemente visual para el usuario final. Culminando con la clasificación con los diversos aspectos encontrados se encuentra en la parte inferior una leyenda del cómo se ha calificada cada plataforma dando como ganadora la ya recomendada Power BI con 33 puntos.

Tabla 52 *Comparación de plataformas para el desarrollo*

	Power BI	Tableau	Dundas BI
Series de Tiempo	**5**	4	4
Visualización	**5**	5	4
Posición	**5**	4	3
Librerías	**5**	4	4
Extensiones	**5**	3	3
Licencia	**4**	3	3
Plataformas	**4**	4	4
Total	**33**	27	25

Tabla 53 *Leyenda de la clasificación de las plataformas*

Muy Bueno	5
Bueno	4
Regular	3
Malo	2
Muy Malo	1

Anexo 14: Información del equipo 1

En la figura 89 se hace la descripción básica del equipo 1, desde la edición del Windows, el tipo de procesador, la memoria RAM, el tipo de sistema y lo que se necesitó para llevar a cabo sin ningún inconveniente el proyecto.

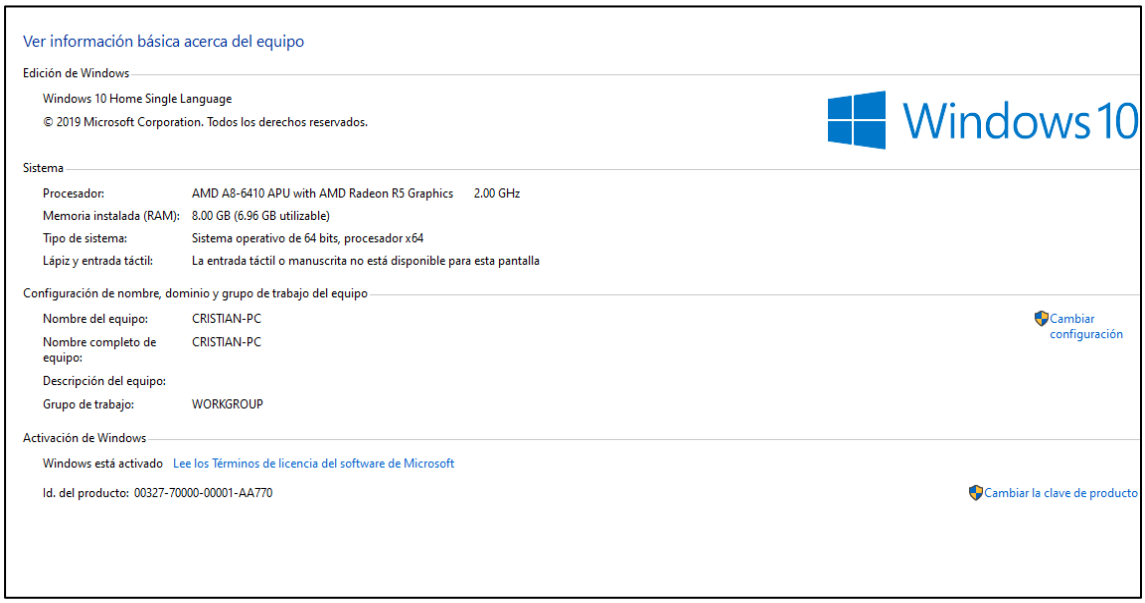


Figura 89. Pantalla de las características del equipo 1

Figura 90. Pantalla de la velocidad de internet del equipo 1

En la figura 90 se muestra la velocidad del internet tanto como la descarga y la carga.

Anexo 15: Información del equipo 2

En la figura 91 se hace la descripción básica del equipo 2, desde la edición del Windows, el tipo de procesador, la memoria RAM, el tipo de sistema y lo que se necesitó para llevar a cabo sin ningún inconveniente el proyecto.

Figura 91. Pantalla de las características del equipo 2

Figura 92. *Pantalla de la velocidad de internet del equipo 2*

En la figura 92 se muestra la velocidad del internet tanto como la descarga y la carga.

Anexo 16: Aspectos técnicos

A continuación, se presentó las librerías que se usaron dentro del entorno de desarrollo RStudio

Tabla 54 *Descripción de las librerías usadas en RStudio*

Librerías	Descripción	Instalación	Cargar
Readxl	Es un paquete diseñado para hacer una sola tarea: importar hojas de Excel a R. Esto hace que sea un paquete ligero y eficiente.	install.packages("readxl")	library(readxl)
DescTools	Una colección de diversas funciones estadísticas básicas y envoltorios de conveniencia para describir datos de manera eficiente	install.packages("DescTools")	library(DescTools)
Forecast	Librería para el pronóstico de las series de tiempo(TS)	library(forecast)	install.packages("forecast")
Tidyverse	Librería para realizar transformaciones en general del conjunto de datos	library(tidyverse)	install.packages("tidyverse")
Lubridate	Librería para realizar transformaciones en un campo fecha	library(lubridate)	install.packages("lubridate")
TSstudio	Proporciona un conjunto de herramientas para el análisis descriptivo y predictivo de datos de series temporales. Incluye funciones para la visualización interactiva de objetos de series temporales y también funciones de utilidad para la automatización del pronóstico de series temporales	library(TSstudio)	install.packages("TSstudio")
Polynom	Un conjunto de funciones para implementar una clase para univariantes manipulaciones polinómicas.	library(polynom)	install.packages("polynom")
Tseries	Análisis de series temporales y financiación computacional	library(tseries)	install.packages("tseries")

Anexo 17: Manual de usuario

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

MANUAL DE USUARIO

SISTEMA GEOLOCALIZADO DE PRONÓSTICO DE CASOS DE PERSONAS DESAPARECIDAS

Índice

Índice de Figuras

¿POR QUÉ POWER BI?

Se ha desarrollado sistemas geolocalizado en la solución de **Microsoft Power BI**, las ventajas de esta herramienta en relación con otras es una interfaz de usuario más amigable, la existencia previa de sistemas de almacenamiento y repositorio de datos, la adaptabilidad a dispositivos móviles y la facilidad de conectar a otros productos de Microsoft a través de los enlaces, a través de la plataforma **Office 365** (Microsoft Power BI, 2020).

Power BI es una colección de servicios de software, aplicaciones y conectores que funcionan conjuntamente para convertir orígenes de datos sin relación entre sí en información coherente, interactiva y atractiva visualmente. Tanto si se trata de una sencilla hoja de cálculo de Excel como de una colección de almacenes de datos híbridos locales o basados en la nube, Power BI le permite conectar fácilmente los orígenes de datos, visualizar (o descubrir) lo más importante y compartirlo con quien quiera (Microsoft Power BI, 2020).

Power BI consta de una aplicación de escritorio de Windows denominada **Power BI Desktop**, un servicio SaaS (software como servicio) en línea denominado **servicio Power BI** y aplicaciones móviles de Power BI disponibles para teléfonos y tabletas Windows, así como para dispositivos iOS y Android. (Microsoft Power BI, 2020).

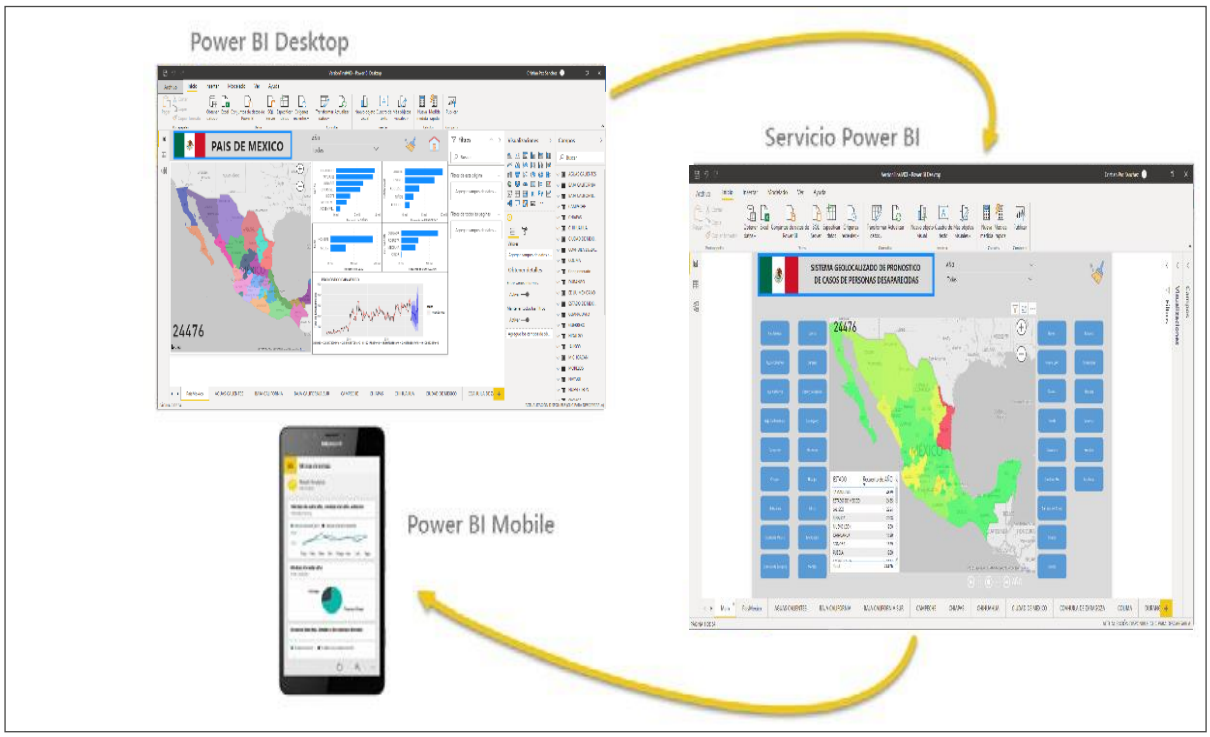


Figura 1 Plataformas Power BI

ANTES DE EMPEZAR

Antes de empezar a revisar el contenido de la Aplicación, se deben tener en cuenta los siguientes puntos: 12

Licencias de usuario

Para poder acceder a la aplicación se debe contar con una licencia de usuario o cuenta gratuita, la cual es provista por la misma página de **Power BI**. Antes de ingresar verifique que dispone de un usuario puede registrarse con su correo institucional totalmente gratis; la cantidad de usuarios que dispone la institución es limitada; ya que solo se puede obtener 1 usuario por cada correo.

Descargar aplicación de escritorio (opcional)

Tal como se ha señalado, Microsoft Power BI cuenta con una aplicación de escritorio que permite al usuario trabajar sobre copias locales de los reportes disponibles en el Portal. Para descargar la aplicación de escritorio siga los siguientes pasos:

1. Ingrese en su navegador el sitio web https://bit.ly/2KJde57
2. Seleccione la opción Productos>Power BI Desktop
3. Seleccione “Descarga gratuita”; de forma automática se descargó el instalador de la aplicación
4. Ejecute el instalador y siga las instrucciones que ahí se indican. Una vez terminada la instalación reinicie el equipo.

Una vez haya reiniciado el equipo, abra la aplicación de Power BI. Debería ver una pantalla como esta

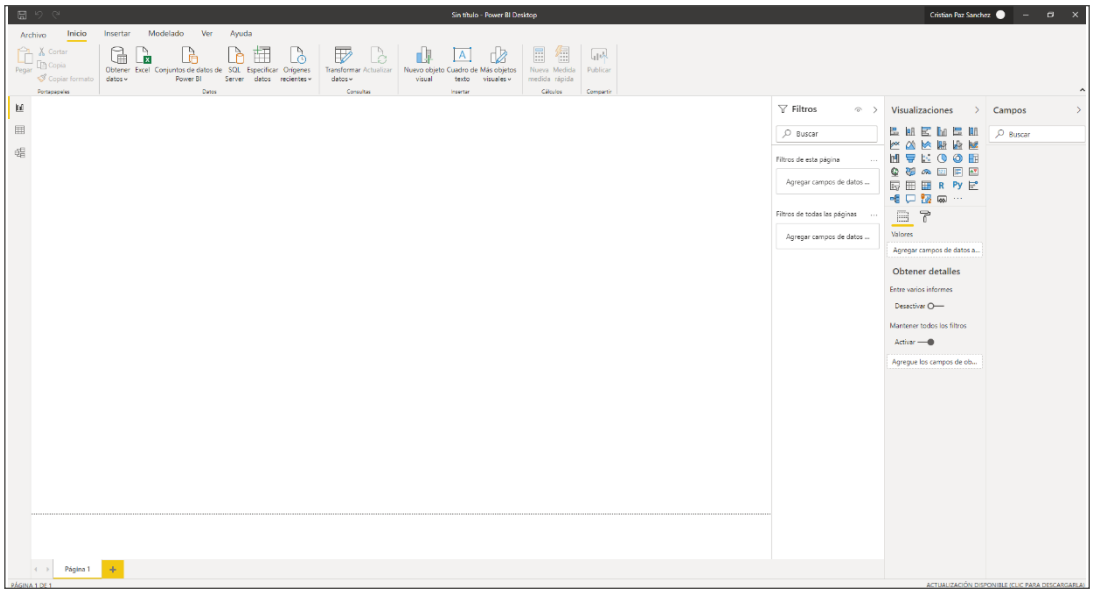


Figura 2 Área de trabajo Power BI

Automáticamente se desplegará una pantalla para ingresar el usuario (su dirección de correo electrónico institucional) y su contraseña; una vez ingresadas sus credenciales, se desplegará una pantalla como la siguiente:

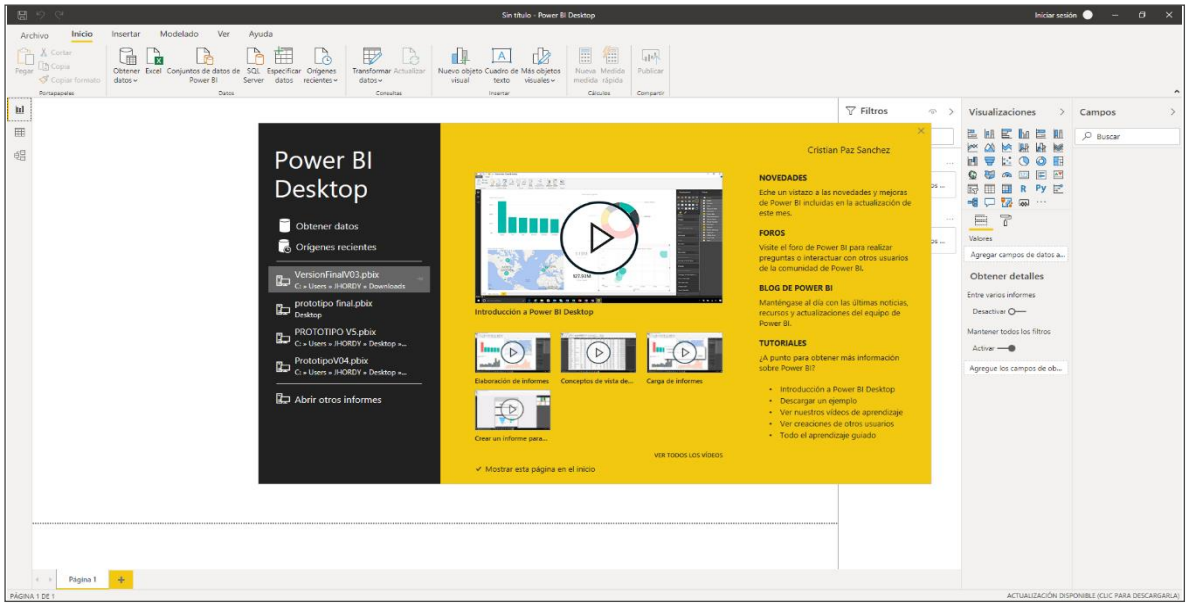


Figura 3 Selección de informe Power BI

¡Recuerde!: Para poder acceder a la aplicación, no es necesario descargar la aplicación de escritorio; sin embargo, para poder visualizar su contenido, debe crear su usuario ya que la aplicación no está disponible de forma pública.

ACCESO A LA APLICACIÓN

1. Ingrese al siguiente link **(Tener en cuenta que tiene que tener su cuenta para poder visualizar la información**):

   https://bit.ly/2WyLkeS

2. En la ventana, presione el botón "Sign-In":

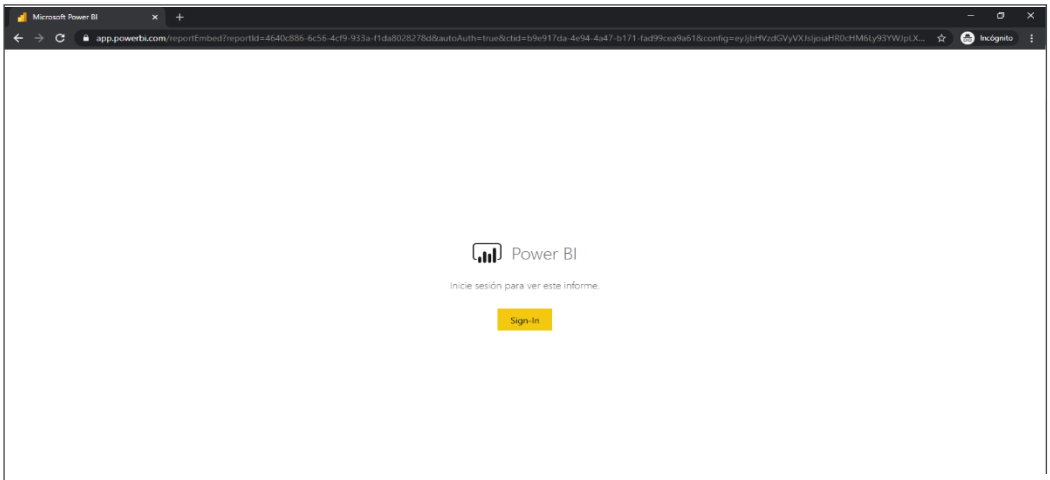


Figura 4 Inicio de sesión Power BI

3. En la siguiente pantalla tiene 2 opciones una para poder iniciar sesión y la otra para poder crear su cuenta:

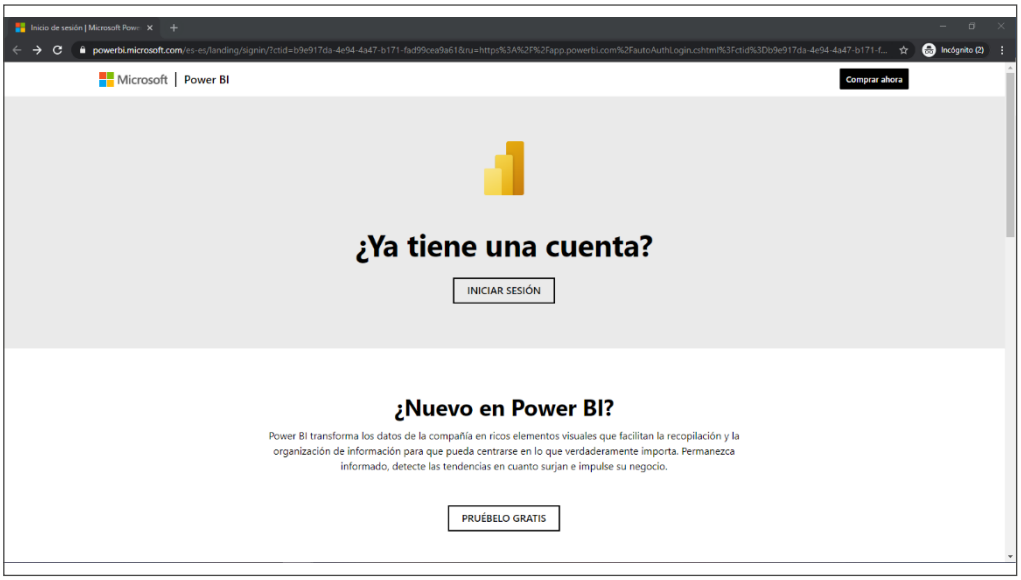


Figura 5 Creación o inicio de sesión en Power BI

4. En la siguiente pantalla se pedirá acceder con nuestra cuenta institucional con la cual se podrá acceder a la información de la aplicación:

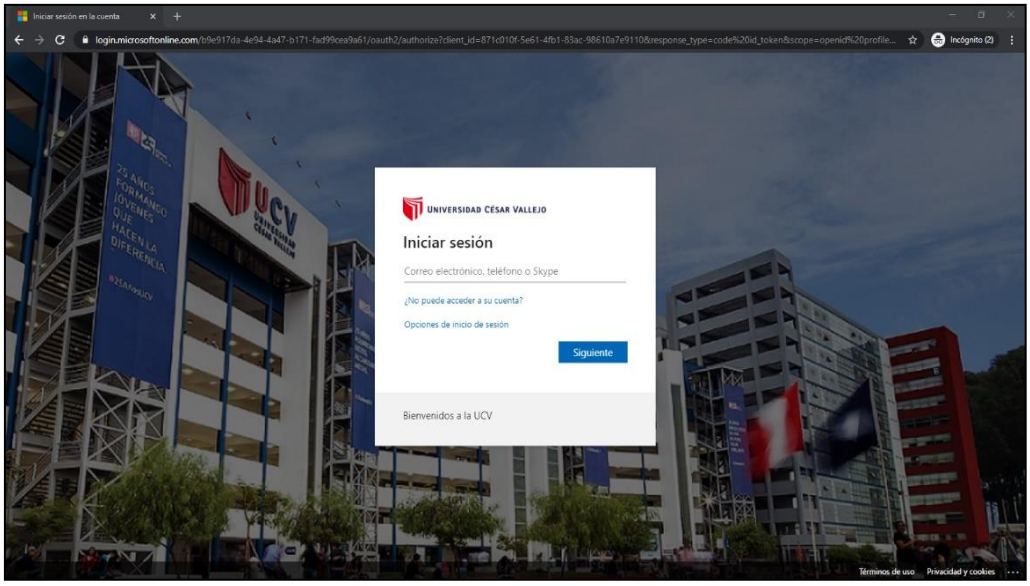


Figura 6 Inicio de sesión con correo institucional

5. Una vez haber iniciado sesión se podrá acceder a la aplicación para poder visualizar el dashboard:

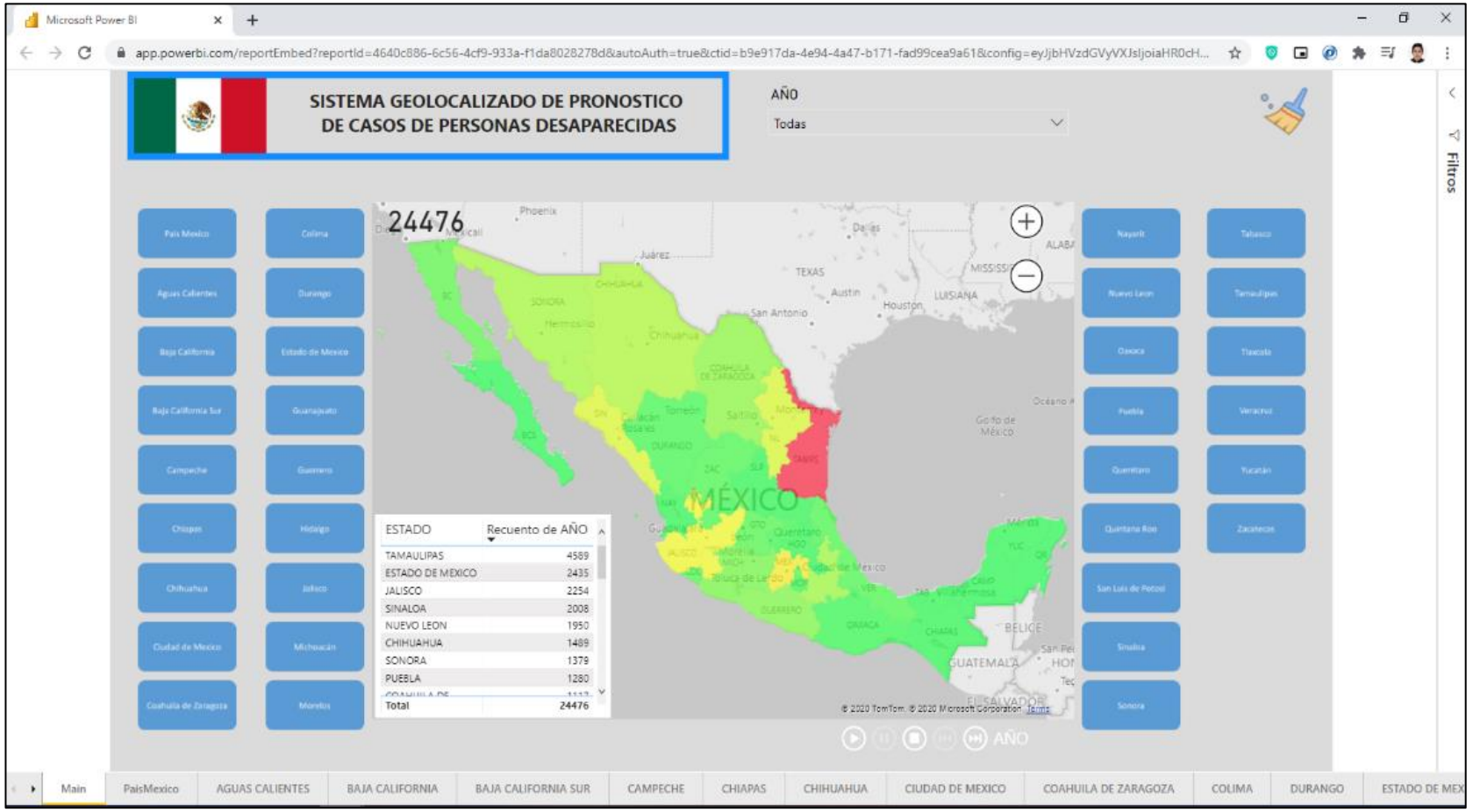


Figura 7 Sistema de geolocalización Power BI

Las opciones y funciones a las que podrá acceder son las siguientes:

1. Estados disponibles para visualización de información y pronósticos

   ✓ País México
   ✓ Aguas Calientes
   ✓ Baja California
   ✓ Baja California Sur
   ✓ Campeche
   ✓ Chiapas
   ✓ Chihuahua
   ✓ Ciudad de México
   ✓ Coahuila de Zaragoza
   ✓ Colima
   ✓ Durango
   ✓ Estado de México
   ✓ Guanajuato
   ✓ Guerrero
   ✓ Hidalgo
   ✓ Jalisco
   ✓ Michoacán
   ✓ Morelos
   ✓ Nayarit
   ✓ Nuevo león
   ✓ Oaxaca
   ✓ Puebla
   ✓ Querétaro
   ✓ Quintana Roo
   ✓ San Luis de Potosí
   ✓ Sinaloa
   ✓ Sonora
   ✓ Tabasco
   ✓ Tamaulipas
   ✓ Tlaxcala
   ✓ Veracruz
   ✓ Yucatán
   ✓ Zacatecas

2. Descripción específica por cada Estado:

   ✓ Señas particulares
   ✓ Rango de Edad
   ✓ Sexo
   ✓ Complexión
   ✓ Pronóstico por Estado
   ✓ Recuento por meses

ACCESO A LOS INFORMES

Para consultar alguno de los informes, siga los siguientes pasos:

1. Seleccione el botón **"Áreas de trabajo"**
2. Seleccione el área de trabajo que desea consultar:

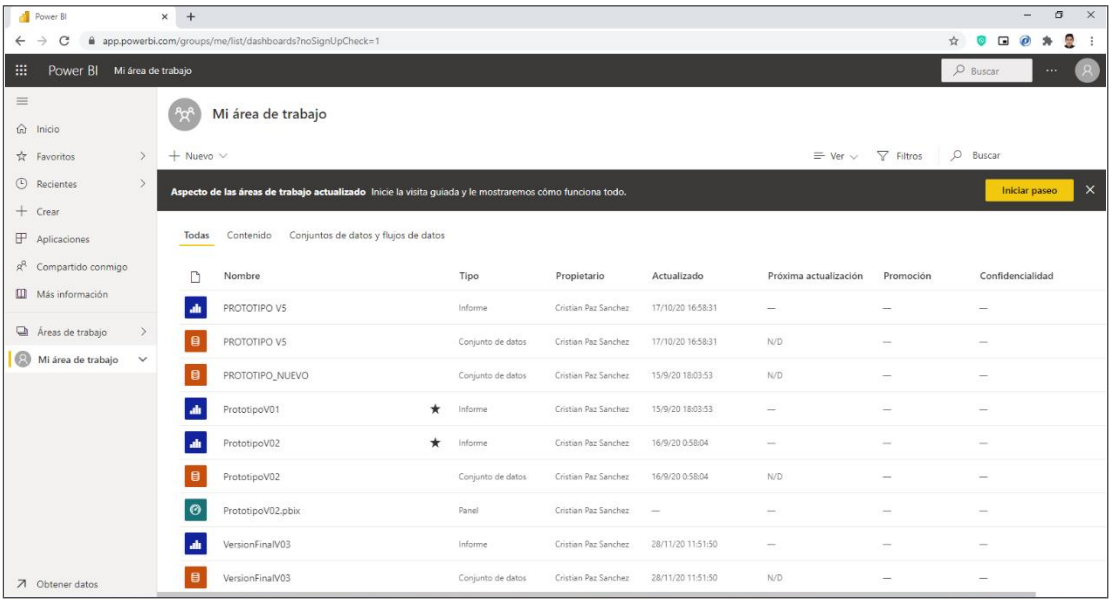


Figura 8 Selección del área de trabajo Power BI

3. Una vez ingresado al área de trabajo, seleccione la pestaña "Contenido"; se desplegará una lista con los informes disponibles:

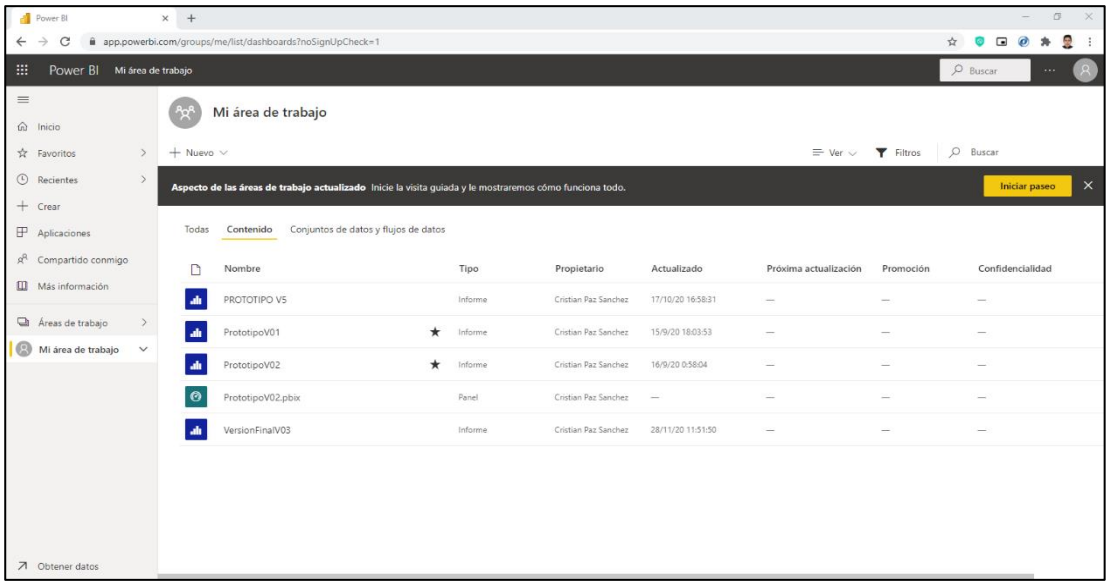


Figura 2 Trabajos disponibles en Power BI

4. Seleccione el informe de su interés. En este caso y a modo de ejemplo, se ha seleccionado el área “Pronósticos Final”.

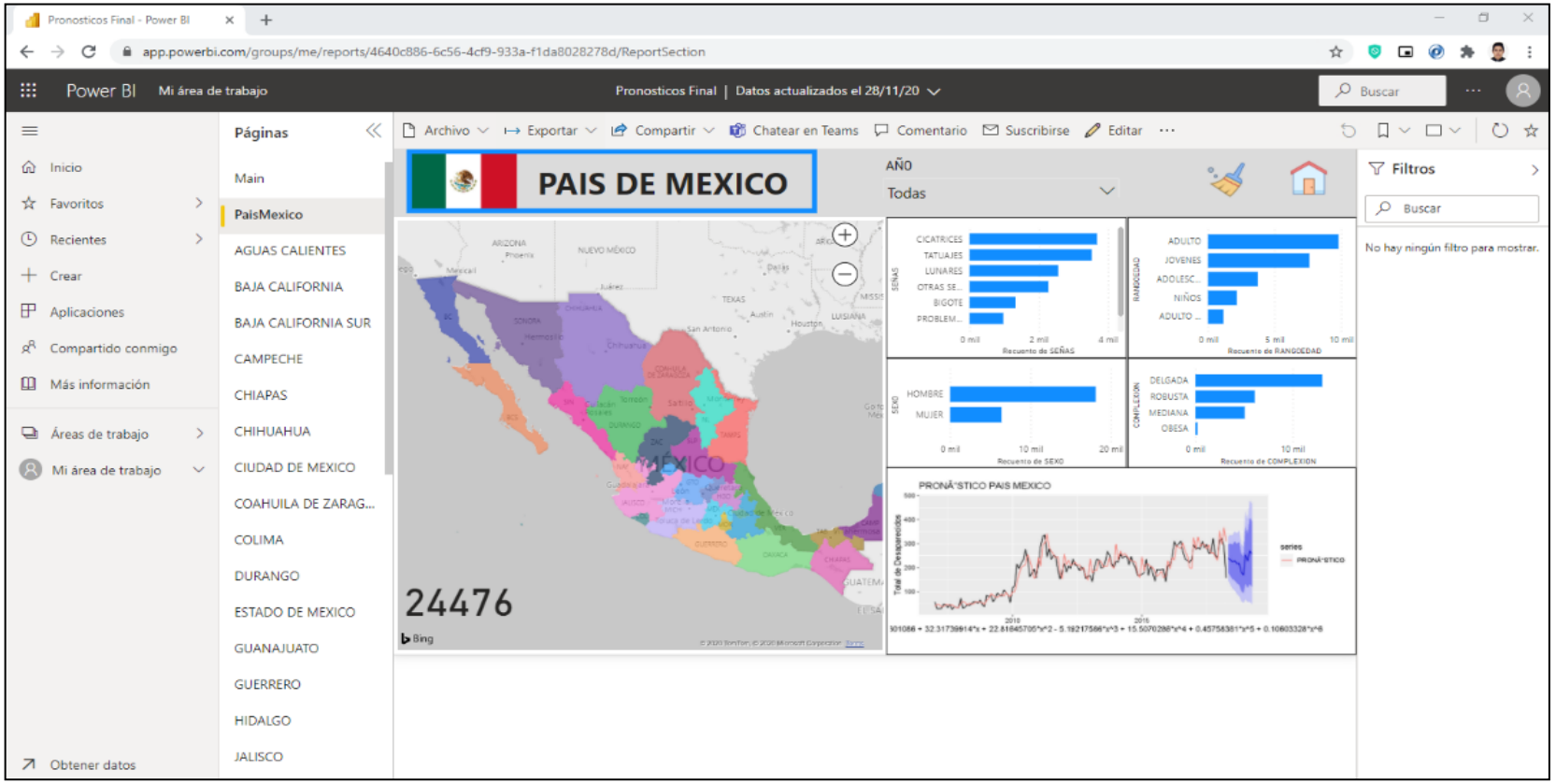


Figura 10 Interfaz dentro de Power BI

¡Tenga presente! Dentro de un informe, cualquier representación de datos **(gráfico, tabla, filtro, número, entre otros)**, se considera una visualización; de esta manera, un reporte puede contener una o muchas visualizaciones como se aprecia en la pantalla anterior.

QUÉ HACER EN LA APLICACÍON

Navegar por los estados

De acuerdo con la naturaleza de cada informe, usted puede realizar las siguientes acciones:

- Filtrar los períodos de interés: para ello, debe seleccionar -en los informes que dispongan de dicha función- en la parte superior del informe los años correspondientes.

- Cambiar los atributos de una visualización: puede cambiar los datos que una visualización muestra al seleccionar atributos de otra visualización.

Figura 11 Sistema geolocalizado con mapa de calor Power BI

- Si se selecciona un Estado como ejemplo “País México”, cambiarán los datos mostrados en la visualización “País de México más detallado”:

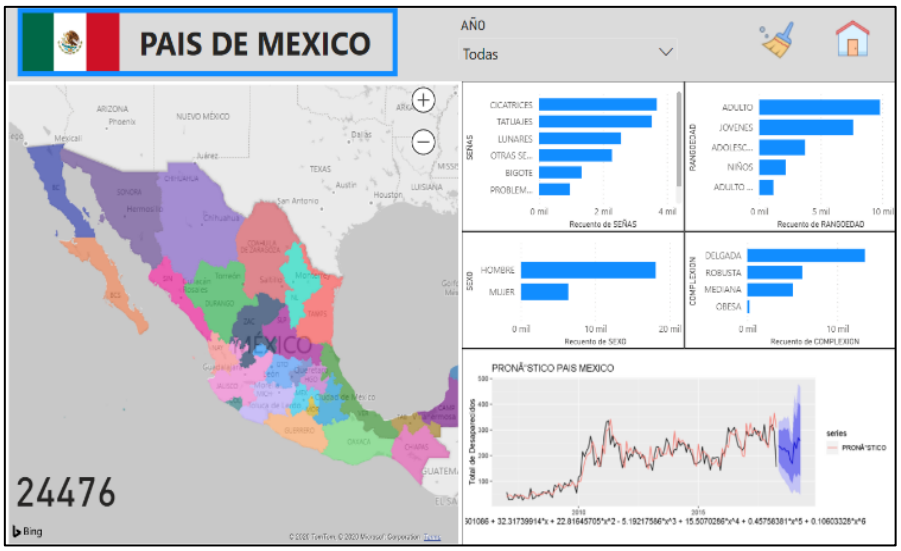


Figura 12 Datos estadísticos

- Pestañas del informe: de forma similar a una planilla Excel, los informes de Power BI cuentan con “pestañas”. En los informes que cuenten con más de una pestaña, puede acceder a más información seleccionado la pestaña de su interés:

Figura 14 Gráfico por escala de colores de los estados

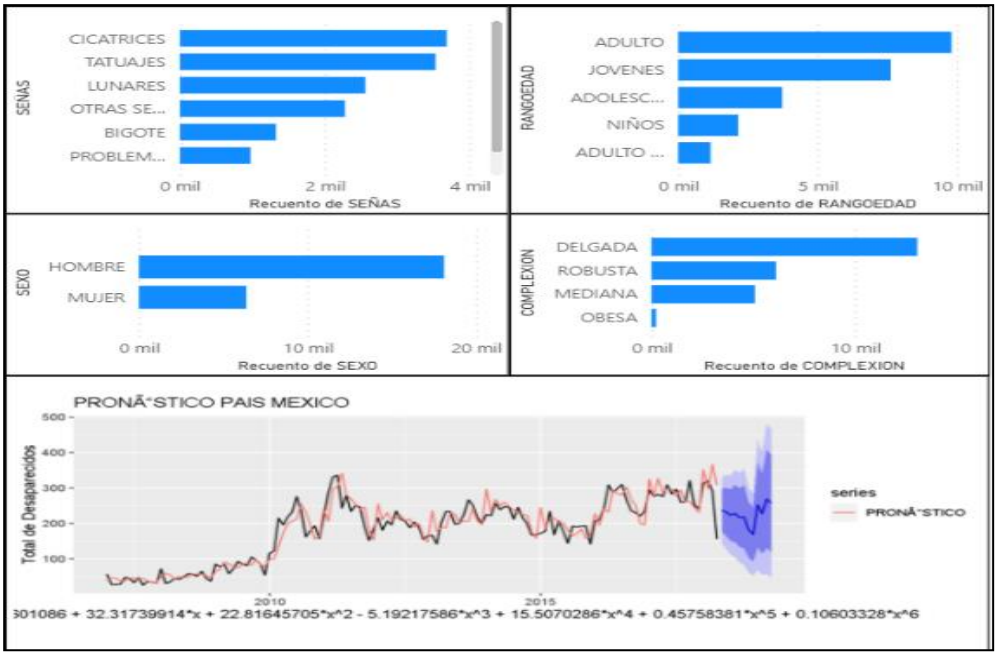


Figura 13 Visualización general de las características y el pronóstico

- Se tiene que resaltar el uso de librerías adicionales para el uso del mapa de calor como se ve a continuación:

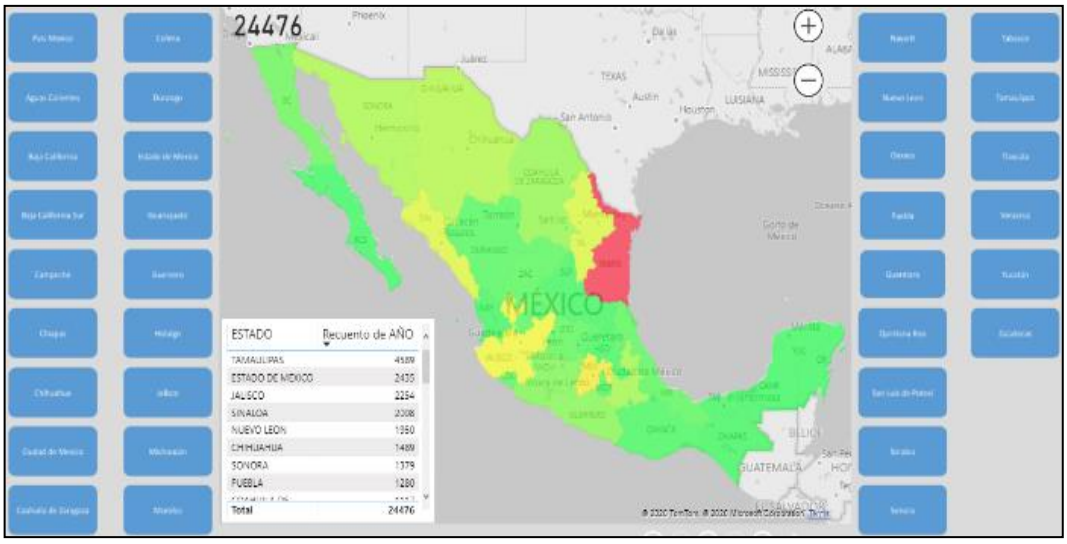


Figura 15 Mapa de calor Power BI 01

- En las siguientes imágenes se ve, el cambio de color que se da en cada Estado y como en alguno se va poniendo de color rojo lo que indica que ese Estado está en riesgo y tienen a comparación de otros estados más casos de personas desaparecidas, por lo que esto permitirá visualizar esos aspectos de una manera interactiva y didáctica

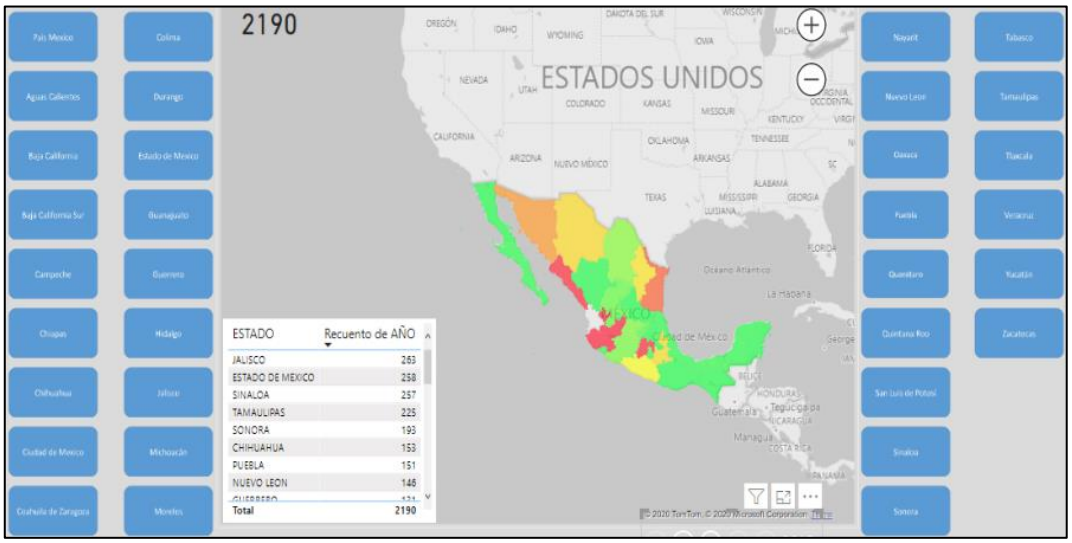


Figura 16 Mapa de calor Power BI 02

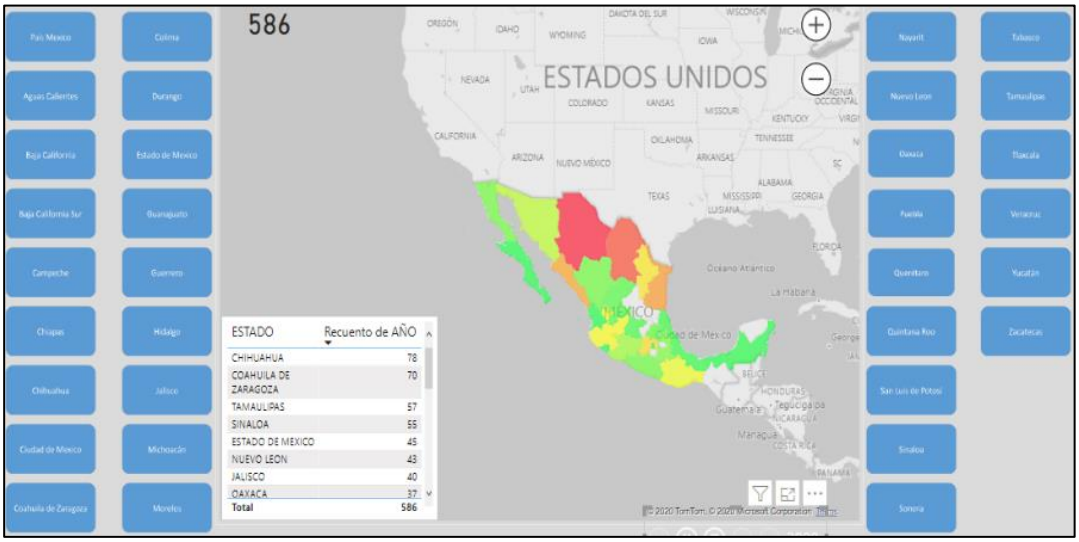


Figura 17 Mapa de calor Power BI 03

- Para ello se tiene que realizar lo siguiente ingresar al siguiente enlace para poder descargar la librería.

  https://bit.ly/3h6hDv2

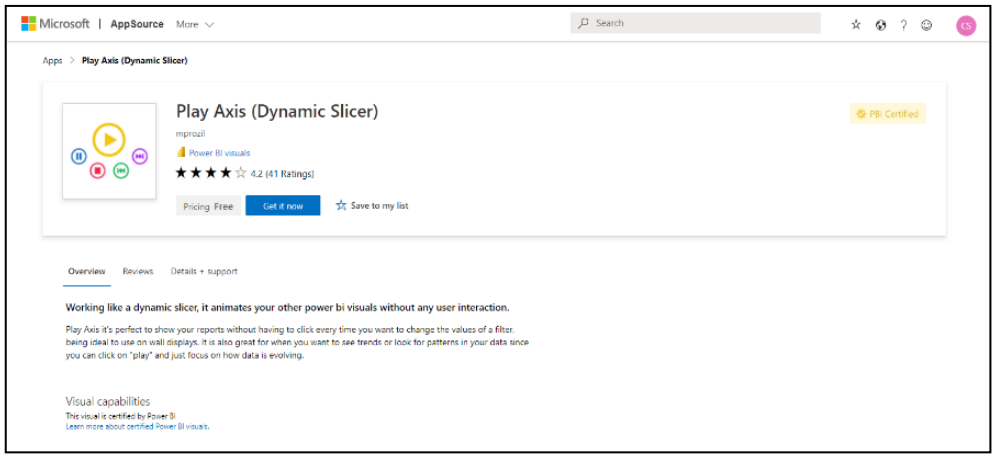


Figura 18 Descarga librería Play Axis (Dynamic Slicer)

- Una vez descargado la librería se importa (sube) desde el Power BI de la siguiente manera.

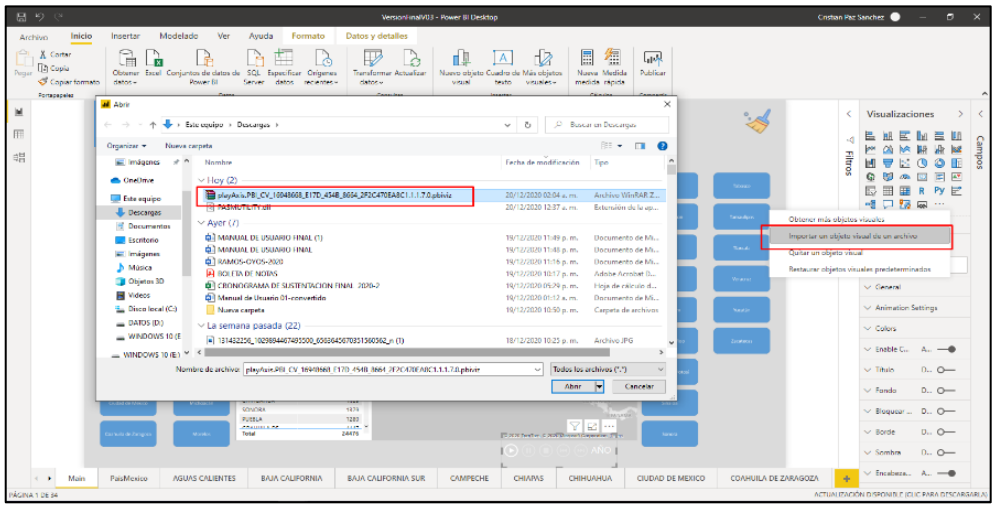


Figura 19 Importar y selecciona librería Play Axis

- Con esto se puede hacer uso de la librería y en nuestro panel de visualizaciones aparecerá un icono como este para poder arrástralo a nuestra área de trabajo del Power BI.

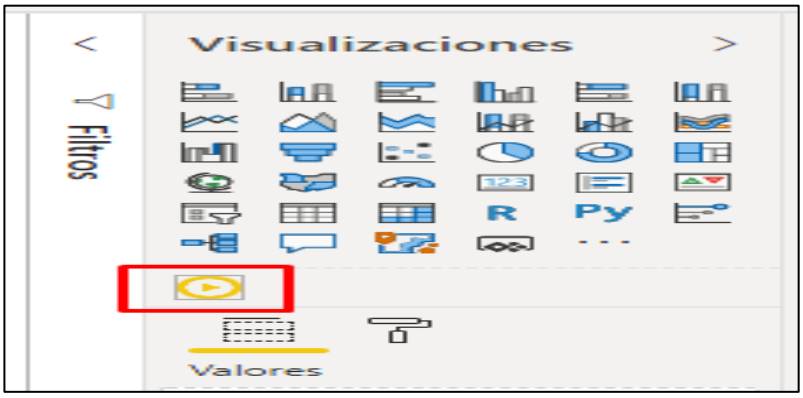


Figura 20 Librería importada

¿MÁS DUDAS?

En caso de necesitar más información y tutoriales sobre el uso de Power BI, visite el sitio web de Microsoft:

https://bit.ly/3muq1Ws

Referencias

Microsoft Power BI. (2020). Power BI. Recuperado de https://powerbi.microsoft.com/es-es/

ANEXOS

Anexo 1. Qué es el Business Intelligence

La inteligencia de negocios o Business Intelligence es un conjunto de elementos, metodologías, aplicaciones y procesos que permiten transformar los datos en información y luego en conocimiento. Algunas de las principales ventajas de esta disciplina son las siguientes:

- Intuitivo: las herramientas de BI son fáciles de utilizar con poco entrenamiento de usuarios; en ese sentido, están diseñadas para que cualquier persona con conocimientos de acceso a internet y uso básico de computadores pueda utilizar estos sistemas.
- Confiable: no es necesario esperar semanas o meses para contar con información de calidad; la velocidad en el procesamiento de datos, implementadas en arquitecturas de sistemas de información robustos permite contar con información en tiempo real o con fechas de corte establecidas por los administradores.
- Accesible: en los últimos 5 años se ha producido una disminución de costos que ha permitido el acceso a estas herramientas a costos abordables para empresas que deseen contar con sistemas de reportaría sofisticados. El mercado dispone de múltiples herramientas de acuerdo con la complejidad de cada organización, o de acuerdo a los niveles de arquitectura de sistemas presentes en ellas.

Debido a su reciente presencia de manera formal y a la naturaleza del contexto en el que se enmarca, el business intelligence ha sufrido una evolución que ha resultado positiva para el usuario final de los datos, tal como lo presenta la siguiente gráfica:

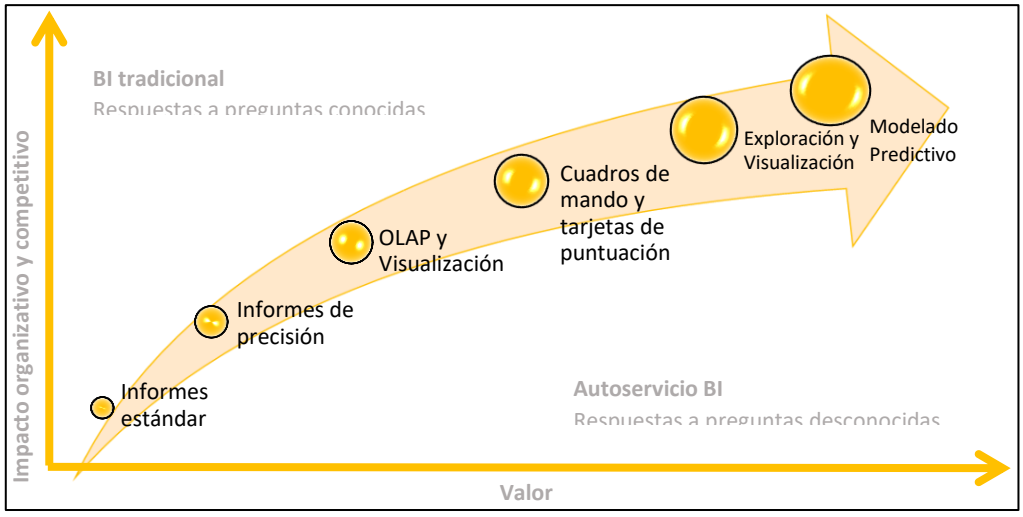


Figura 21 Inteligencia de Negocios en Power BI

UCV
UNIVERSIDAD
CÉSAR VALLEJO

Power BI