RESUMEN

SOFTWARE PARA LA APLICACIÓN DE CIENCIA
DE DATOS EN LA UNIÓN DE MEDIO ORIENTE
Y NORTE DE ÁFRICA. CASO
DE ESTUDIO EN IRAQ

por

Ceila Merari González Hernández

Asesor principal: Germán Harvey Alférez Salinas

RESUMEN DE TESIS DE MAESTRÍA

Universidad de Montemorelos

Facultad de Ingeniería y Tecnología

Título: SOFTWARE PARA LA APLICACIÓN DE CIENCIA DE DATOS EN LA UNIÓN DE MEDIO ORIENTE Y NORTE DE ÁFRICA. CASO DE ESTUDIO EN IRAQ

Investigador: Ceila Merari González Hernández

Asesor principal: Germán Harvey Alférez Salinas, Doctor en Informática

Fecha de terminación: Mayo de 2018

Problema

La zona geográfica con la mayor parte de la población mundial, conocida como Ventana 10/40, cuenta con una alta presencia de judaísmo, hinduismo, budismo e islamismo y, menormente, de cristianismo. En esta zona, la mayor parte de la población presenta diversas ideologías, intercambios culturales e ideas filosóficas y religiosas, lo cual la convierte en un desafío para la evangelización. La misión de la Iglesia Adventista del Séptimo Día consiste en predicar el evangelio en todas las regiones del mundo. Sin embargo, en esta zona existe un número muy reducido de feligreses y el porcentaje de crecimiento es inferior, comparado con el resto del mundo.

Método

Esta investigación basó su desarrollo en la metodología de ciencia de datos de la International Business Machines (IBM), compuesta por diez etapas iterativas que incluyen comprensión del problema, enfoque analítico, requerimiento de los datos, recolección de los datos, comprensión de los datos, preparación de los datos, modelamiento, evaluación, despliegue y retroalimentación.

Resultados

En esta investigación se aplicó aprendizaje automático, una subárea de la inteligencia artificial y pieza fundamental de la ciencia de datos, a datos abiertos obtenidos de la Global Database of Events, Language and Tone (GDELT), por medio de Big Query de Google. Esto con el objetivo de encontrar un modelo predictivo que ayude a descubrir necesidades de las personas en zonas específicas de Iraq. Este modelo es utilizado por un software creado en Python que, de acuerdo con cierta latitud y longitud introducida por el usuario, muestra la clasificación de la necesidad más apremiante del área. El modelo utilizado por el software fue el generado mediante Decision Trees, que obtuvo para el evento de refugiados un valor de precision de 0.76, recall y f1-score de 0.76, y para el evento de lucha con artillería, una precision de 0.74, recall y f1-score de 0.75, con una accuracy de 0.7629 para ambos eventos. El modelo resultante se obtuvo después de haber realizado en total 104 experimentos con los siguientes clasificadores: KNN, Decision Trees, Näive Bayes y Logistic Regression.

Conclusiones

Mediante un software basado en ciencia de datos, datos masivos, datos abiertos, aprendizaje automático y la metodología de ciencia de datos de IBM, es posible clasificar eventos sobre refugiados y lucha con artillería, basándose en un conjunto de datos sobre Iraq, país perteneciente a la Ventana 10/40. El modelo predictivo utilizado por el software que se generó mediante Decision Trees, presenta valores arriba de 0.74 para accuracy, precision, recall y f1-score. Además, los resultados coinciden con los ofrecidos en mapas oficiales.

Universidad de Montemorelos

Facultad de Ingeniería y Tecnología

SOFTWARE PARA LA APLICACIÓN DE CIENCIA DE DATOS EN LA UNIÓN DE MEDIO ORIENTE Y NORTE DE ÁFRICA. CASO DE ESTUDIO EN IRAQ

Tesis
presentada en cumplimiento parcial
de los requisitos para el grado de
Maestría en Ciencias Computacionales

por

Ceila Merari González Hernández

Mayo de 2018

SOFTWARE PARA LA APLICACIÓN DE CIENCIA DE DATOS EN LA UNIÓN DE MEDIO ORIENTE Y NORTE DE ÁFRICA. CASO DE ESTUDIO EN IRAQ

Proyecto

presentado en cumplimiento parcial de
los requisitos para el grado de
Maestría en Ciencias
Computacionales

por

Ceila Merari González Hernández

APROBADO POR LA COMISIÓN:

Germán Harvey Alférez Dr. Germán Harvey Alférez Salinas Asesor principal	 Dr. Abimael Lozano Reyes Asesor externo
 M.C. Alejandro Walterio García Mendoza Miembro	 Dr. Ramón Andrés Díaz Valladares Director de Posgrado e Investigación
 M.C. Saulo Hernández Osoria Miembro	1 de mayo de 2018 Fecha de aprobación

DEDICATORIA

Principalmente a Dios, porque él “es quien da la sabiduría; la ciencia y el conocimiento brotan de sus labios” (Proverbios 2:6, DHH).

A mis padres, Rocio Hernández y Fernando González que, aun en la distancia, me brindaron su apoyo.

A mis hermanos Jared y Yair, por su ánimo, comprensión e interés durante este proceso de investigación.

A mis amigos y amigas que, en la cercanía o lejanía, recibí de sus oraciones y palabras de ánimo.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RECONOCIMIENTOS

A mi padre celestial, porque Él ha guiado mi vida y me ha llevado por rumbos desconocidos, pero con propósitos fieles y firmes para honrar su nombre.

Al doctor Germán Harvey Alférez Salinas, mi asesor principal, por su orientación y dedicación durante esta investigación.

Al maestro Saulo Hernández Osoria, por su motivación y aprecio durante mi estancia en la Universidad de Montemorelos y por su apoyo como asesor secundario.

Al maestro Alejandro García Mendoza, director de la Facultad, por su apoyo como asesor secundario en esta investigación.

Al maestro Carlos Hernández Rentería, coordinador del posgrado, por su orientación al resolver mis dudas que surgían como estudiante.

Al maestro Daniel Gutiérrez Colorado, coordinador del posgrado, por su atención a los últimos detalles en el proceso de esta investigación.

Al pastor Jesús Fernández, capellán de la FIT, por su motivación e interés al preguntarme acerca de mi progreso en el trabajo de investigación.

Al doctor Jair del Valle, por su asesoría con el formato APA de esta investigación.

CAPÍTULO I

INTRODUCCIÓN

Planteamiento del problema

La zona conocida como Ventana 10/40, llamada así por ubicarse entre los paralelos 10 y 40 al norte del Ecuador, se considera como el área geográfica con la mayor parte de la población mundial. Esta zona sobresale por contar con una alta presencia de judaísmo, hinduismo, budismo e islamismo y una menor presencia de cristianismo (Gustin, 2006). La cosmovisión de las personas en estos lugares complica los intercambios culturales y de ideas tanto filosóficas como religiosas (Fakhoury, 2016. Ventana 10/40. Semana de misiones. Conferencia realizada por el instituto de Misiones de la Universidad de Montemorelos, Nuevo León, México). La misión de la Iglesia Adventista del Séptimo Día (IASD) es predicar el evangelio en todas las regiones del mundo. Sin embargo, de acuerdo con el reporte anual estadístico de la IASD, en la Ventana 10/40 existe un número muy reducido de feligreses y el porcentaje de crecimiento es inferior comparado con el resto del mundo (Seventh-day Adventist Church, 2014).

Fakhoury (2016) señala que la mayor parte de la población dentro de la Ventana 10/40 presenta diversas ideologías. También existe el desafío de las leyes religiosas e ideologías del Islam, que están en contra de otras religiones y creencias. Esto implica

que evangelizar mediante el mensaje bíblico en esa zona del mundo aun podría significar la muerte.

Una de las uniones de la IASD ubicada en la Ventana 10/40 es la Unión de Medio Oriente y Norte de África (MENA). Esta unión limita al oeste con Marruecos, al este con Irán, al norte con Turquía y al sur con Sudán. Los idiomas en esta región incluyen el árabe, el turco, el bereber, el persa, el francés y el inglés (Iglesia Adventista del Séptimo Día, 2016). Los países que conforman esta unión son: Argelia, Bahrein, Egipto, Irán, Iraq, Jordania, Kuwait, Líbano, Libia, Marruecos, Omán, Qatar, Arabia Saudita, Sudán, Túnez, Turquía, Emiratos Árabes Unidos, Sahara Occidental y Yemen (Seventh-day Adventist Church, 2014).

Declaración del problema

Hasta 2012, la estrategia que la IASD había usado para evangelizar en la Ventana 10/40 se basó en el establecimiento de cuatro industrias de alimentos, 33 hospitales, dos centros de medios, cuatro estaciones de radio y TV, 60 clínicas médicas y dentales, 23 asilos de ancianos, tres orfanatos, 12 casas publicadoras y 116 escuelas secundarias (Seventh-day Adventist Church, 2014).

Aunque se han realizado avances misionales en esta zona, la pregunta que surge es ¿cómo descubrir las necesidades de las personas en esta región del mundo con el fin de alcanzarlas mediante el método de Cristo? Este método consiste en resolver las necesidades actuales de las personas para luego invitarlas a seguirle (White, 1905). Una respuesta a esta pregunta se encuentra en la utilización de ciencia de datos para descubrir las necesidades de las personas en la Ventana 10/40. La

ciencia de datos se define como la ciencia que estudia datos para la obtención de conocimiento (Dhar, 2013).

En la Facultad de Ingeniería y Tecnología (FIT) de la Universidad de Montemorelos (UM) se han llevado a cabo estudios previos que muestran la eficacia de la aplicación de ciencia de datos en el evangelismo, por ejemplo, mediante el análisis de tweets para entender las necesidades de los habitantes de Nueva York (Alférez, 2016). Lo primordial en la ciencia de datos son, precisamente, los datos. No obstante, según el informe de Open Data Barometer Fourth Edition (Open Data Barometer, 2017), son pocos los países del medio oriente y norte de África con iniciativas de datos abiertos. Esto se debe a la poca participación de la sociedad civil y a la poca presión para que los gobiernos hagan públicos sus datos.

Iraq es uno de los países pertenecientes a la Ventana 10/40 y específicamente a la MENA. De acuerdo con la Seventh-Day Adventist Church (2014), este país no cuenta con ministros ordenados ni con iglesias establecidas. Asimismo, el número de miembros se ha reducido en los últimos dos años. De hecho, existen tres adventistas por cada millón de personas (Seventh-day Adventist Church, 2014). Igualmente, la Organización de las Naciones Unidas (2017) destaca un alto número de noticias en Iraq relacionadas con constantes desastres naturales, hechos de violencia, refugiados y masacres. Por las razones presentadas, se ha decidido enfocar el presente estudio en Iraq.

Aunque la administración de la MENA está interesada en conocer las necesidades actuales de los habitantes en Iraq, no se conocen todas las necesidades de regiones específicas en ese país por la falta de datos en algunas zonas. De acuerdo

con el informe de Open Data Barometer (2017), la mayoría de los datos de la región de MENA están bloqueados con restricciones legales o técnicas. Específicamente, Iraq es uno de los países que no se encuentra en el ranking regional respecto de la provisión de datos.

El aporte de esta investigación consiste en la aplicación de aprendizaje automático, una subárea de la inteligencia artificial y pieza fundamental de la ciencia de datos a datos abiertos obtenidos de la Global Database of Events, Language and Tone (GDELT) de 2012–2015 por medio de Big Query de Google. Esto con el fin de descubrir las necesidades de las personas en zonas específicas de Iraq. El modelo predictivo más exacto se ejecuta en un software que, de acuerdo con cierta longitud y latitud introducida por el usuario, muestra la clasificación de las necesidades más apremiantes del área. Esto es principalmente de interés para la clasificación de zonas sin datos disponibles.

GDELT es un sitio patrocinado por Google que contiene más de 200 millones de eventos geolocalizados con cobertura mundial respecto de noticias y acontecimientos importantes desde 1979 hasta el año actual (Leetaru y Schrodt, 2013). GDELT obtiene la información principalmente de las siguientes agencias de noticias: Lexis Nexis, The Agence France–Presse, Reuters, Associated Press y Xinhua. También utiliza como base conflict and mediation event observations (CAMEO), que es un código de contenido para las noticias electrónicas (Yonamine, 2012). El análisis de los datos proporcionados por esta organización mediante la aplicación de algoritmos de inteligencia artificial puede revelar ciertos comportamientos, características específicas o, posiblemente, predecir

acontecimientos futuros. Esto fue comprobado por Yonamine (2012), quien predijo en su investigación ciertos niveles de violencia en distritos de Afganistán.

Justificación

La IASD tiene como misión “hacer discípulos a todas las naciones instruyendo a las personas a servirle a Él (Jesús) como Señor y prepararlos para su pronto regreso” (Iglesia Adventista del Séptimo Día, 2013). Junto con la misión encomendada, Cristo estableció un método para acercarse a las personas. Él “trataba con los hombres como quien deseaba hacerles bien. Les mostraba simpatía, atendía sus necesidades y se ganaba su confianza. Entonces les decía: Sígueme” (White, 1905, p.143).

A nivel mundial, la IASD se organiza en 13 divisiones y tiene presencia en 206 países del mundo. El número de miembros por división se indica en la Tabla 1. De acuerdo con esta tabla se puede notar que una de las regiones con menor número (incluso disminución) de miembros se encuentra en la Unión de Medio Oriente y Norte de África.

De acuerdo con la Organización de las Naciones Unidas (2017), las principales características de la región de Medio Oriente son las siguientes:

Allí vive gran parte de la población mundial. La población total del mundo es de 7,244 millones, de los cuales África cuenta con 699 millones y Asia con 3,432 millones.

Otra característica consiste en que la mayoría de los pueblos más pobres del mundo se encuentran en esta zona.

Es necesario mencionar también que más del 90% de la población de la Ventana 10/40 no ha sido evangelizada por la IASD. Allí han nacido el judaísmo, el hinduismo, el budismo, el cristianismo y el islamismo, pero el cristianismo tiene poca presencia (Gustin, 2006).

Tabla 1

Feligreses por División, respecto al 2012 – 2014

No.	Divisiones en el mundo	Feligreses	% crecimiento
1	División Interamericana	3,686,255	1.4
2	División de África Meridional y Océano Índico	3,167,259	5
3	África Centro Oriental	2,856,708	1.3
4	División Sudamericana	2,263,194	4.6
5	División del Sudeste Asiático	1,510,326	-2.4
6	División de Asia Pacifico Sur	1,222,546	0.5
7	División Norteamericana	1,184,395	1.5
8	División de África Centro Occidental	769,607	-12.6
9	División de Asia Pacifico Norte	679,907	1
10	División del Pacífico Sur	420,962	4.6
11	División Inter Europea	178,199	0.2
12	División Euroasiática	116,013	-3.1
13	División Transeuropea	84,428	0.8
14	Unión de Medio Oriente y Norte de África	3,151	-1.9
15	Territorio de Israel	795	9.8

Asimismo, el crecimiento de feligreses de la IASD en estas zonas es menor comparado con el de otras partes del mundo, debido a que la cosmovisión de sus habitantes es totalmente opuesta a la cultura de occidente (Chuliá, 2016. Cosmovisión Islámica. Semana de misiones. Conferencia realizada por el instituto de Misiones de la Universidad de Montemorelos, Nuevo León, México). De hecho, los habitantes de la Ventana 10/40 consideran que los occidentales son liberales en sus prácticas, principalmente en la adoración a Dios. Esto conlleva a que ellos consideren a los cristianos de Occidente como “perdidos”.

Por medio de los datos es posible descubrir las necesidades específicas de las personas, con el fin de alcanzarlas con el método de Cristo (Alférez, 2016). La ciencia de datos junto con otras herramientas, tales como aprendizaje automático (Machine

learning, en inglés) y datos masivos (Big Data, en inglés) de GDELT, podrían brindar una solución para entender y clasificar las necesidades de las personas que viven en Iraq, que es el caso de estudio en esta investigación. Hasta donde se conoce, en la IASD a nivel mundial no existe un software que apoye como herramienta para clasificar las necesidades de las personas en regiones sin o con pocos datos, de la Ventana 10/40.

Objetivos

Los objetivos son importantes porque definen la orientación de la investigación, diseñando las metas a lograr durante el desarrollo de la misma. Es por ello que se describen a continuación.

Objetivo general

El objetivo general de este trabajo es construir un software para clasificar zonas de Iraq con respecto a eventos sobre refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres, utilizando ciencia de datos y aprendizaje automático.

Objetivos específicos

Los objetivos específicos de esta investigación son los siguientes:

1. Seleccionar por medio de consultas en Big Query los datos correspondientes en GDELT a eventos de refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres, respecto al país de Iraq entre los años 2012 y 2015.

2. Construir modelos predictivos mediante algoritmos de aprendizaje automático para clasificar zonas de Iraq respecto de los eventos de refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres con los resultados obtenidos mediante Big Query.

3. Evaluar los modelos predictivos de clasificación para seleccionar el modelo óptimo de acuerdo con una validación cruzada y una comparación de los resultados con mapas oficiales.

4. Crear un software que utilice el modelo predictivo elegido con los resultados óptimos para clasificar automáticamente zonas de Iraq.

Hipótesis

Un software basado en la aplicación de ciencia de datos con datos abiertos masivos permite descubrir las necesidades de las personas en Iraq, al clasificar automáticamente zonas en ese país.

Limitaciones

Existe un límite de consultas en GDELT. De acuerdo con el sitio Google Cloud Platform, GDELT establece 50 consultas simultáneas (Google Developers, 2017).

Las consultas en GDELT tienen un límite de datos para ser descargados. Para el tipo de archivo Comma Separated Values (CSV), el tamaño de fila y celda tiene como límite 10 megabytes (Google Developers, 2017).

La distancia entre la MENA, específicamente Iraq, y la Universidad de Montemorelos, lugar en donde se desarrolla esta investigación, no permite una validación en persona de los resultados obtenidos.

Delimitaciones

Para obtener una solución óptima al problema planteado, la investigación se delimitó como se describe a continuación:

El software es exclusivo para analizar eventos del país de Iraq que pertenece a la División de Medio Oriente y Norte de África de la IASD.

Después de la evaluación de los modelos predictivos con los algoritmos de clasificación KNN, Decision Trees, Naive Bayes y Logistic Regression, el software permite clasificar zonas en el país con respecto a eventos relacionados con refugiados y lucha con artillería.

Los datos fueron obtenidos del sitio de noticias de GDETL mediante Big Query en formato CSV para su manejo en el lenguaje de programación Python.

Se analizaron datos con respecto a un período de tres años, del año 2012 al año 2015. Se consideró este período de tiempo para comprender la situación actual del país. El uso de datos de años anteriores a estos dificultaría la generación de un modelo predictivo actual y relevante.

Se utilizó Python como lenguaje de programación y las librerías de Scikit Learn para elaborar modelos predictivos y de clasificación.

Se utilizaron los siguientes algoritmos de clasificación: K-Nearest Neighbors, Decision Trees, Näive Bayes y Logistic Regression.

Los modelos se evaluaron utilizando las siguientes medidas de evaluación: accuracy, recall, precision y f1-score. Además, los resultados se compararon con el mapa proporcionado por el Alto Comisionado de las Naciones Unidas (United Nations High Commissioner for Refugees, 2017) y por el Live Universal Awareness Map (2014).

Descripción de conceptos

A continuación, se describen términos no comunes presentes en esta investigación, con el objetivo de contextualizar al lector.

Actor. Se refiere a las entidades involucradas en los eventos (Global Database of Events, Language and Tone, GDELT, 2013 y Schrodt, 2012).

Aprendizaje automático. Machine learning en inglés, es una herramienta usada para automatizar la toma de decisiones basándose en patrones encontrados en datos disponibles (Murphree, 2016).

Datos masivos o Big Data. Se refiere al conjunto de datos que son mayores a la capacidad de procesamiento en bases de datos tradicionales (Mathur, Sihag, Bagaria y Rajawat, 2014).

Big Query. Es un servicio de Google que permite hacer análisis interactivo de datos masivos con el fin de almacenarlos y consultarlos (Camargo Vega, Camargo Ortega y Joyanes Aguilar, 2015).

CAMEO. Es la sigla en inglés de conflict and mediation event observations, es el código de contenido para noticias electrónicas (Yonamine, 2012).

Ciencia de datos. Es la ciencia que consiste en el estudio de datos para la obtención de conocimiento (Dhar, 2013).

Clase. Son los atributos de la variable dependiente (Harrington, 2012).

CSV. Es la sigla en inglés de comma separated values y es una extensión de archivos para valores separados por comas.

Dataset. Es un conjunto de datos que incluyen características similares o relación entre ellos (Moyano, Gibaja y Ventura, 2017).

Etiqueta o label. Es el nombre que se asigna a las divisiones o categorías del conjunto de datos (Hackeling, 2014).

Evento. De acuerdo con el manual de CAMEO, es una categoría respecto a algún acontecimiento (Schrodt, 2012).

Característica. Feature en inglés. Es un rasgo distintivo de una clase (Harrington, 2012).

GDELT. Es la sigla en inglés de Global Database of Events, Language, and Tone. Este es un proyecto patrocinado por Google que contiene más de 200 millones de eventos geolocalizados con cobertura mundial con respecto a noticias y acontecimientos importantes.

IBM. Son las siglas en inglés para International Business Machines.

Instancia. Son los registros a manera de filas en un conjunto de datos descritos mediante características (Harrington, 2012).

Inteligencia artificial. Es el estudio de cómo se adquiere, representa y almacena el conocimiento en una máquina (Nilsson, 2001).

MENA. Es The Middle East and North Africa, Unión de la Iglesia Adventista del Séptimo Día.

Open Data. Es utilización, reutilización y publicación de datos sin precio y sin uso de licencia (Murray, 2008).

Python. Es un lenguaje de programación recomendable y altamente utilizado en aprendizaje automático, debido a que presenta una sintaxis clara y la manipulación del texto es sencilla (Harrington, 2012).

Scikit Learn. Es un módulo de Python que integra diversos algoritmos de aprendizaje automático; es fácil de usar, presenta buen rendimiento y documentación (Pedregosa et al., 2011).

Software. Es el conjunto de programas, instrucciones y reglas informáticas para ejecutar ciertas tareas de una computadora (Real Academia Española, 2017).

CAPÍTULO II

MARCO TEÓRICO Y TRABAJO RELACIONADO

Marco teórico

Esta sección presenta la definición de los conceptos usados en la investigación: ciencia de datos, datos estructurados, no estructurados y semiestructuados, GDELT, BigQuery, aprendizaje automático, inteligencia artificial, algoritmos de clasificación, destacando K- Nearest Neighbors, Näive Bayes, Decision trees y Logistic Regression; aprendizaje supervisado y no supervisado, metodología de ciencia de datos de IBM, Python, incluyendo Scikit Learn, Numpy, Pandas y Anaconda; también se describen las métricas de evaluación para modelos de clasificación (accuracy, recall, precision y f1-score); asimismo, se presenta una sección de trabajo relacionado con ciencia de datos y aprendizaje automático en la Facultad de ingeniería y tecnología (FIT) de la Universidad de Montemorelos (UM).

Para sustentar la propuesta presentada y estructurar esta sección, se consultaron libros, artículos de revistas, sitios electrónicos y documentos de investigación, tales como tesis y publicaciones científicas. Dicha información fue extraída de las bases de datos EBSCO, Google Scholar, ProQuest, Institute of Electrical and Electronics Engineers (IEEE), Elsevier y Springer, por medio de la web.

En la Figura 1 se muestra el mapa conceptual de los términos presentados en esta sección, con el fin de facilitar su compresión por medio de su explicación y definición; estos son Ventana 10/40, Iraq, Metodología IBM, GDELT, BigQuery; también se describe Machine Learning y sus tipos de aprendizaje supervisado y no supervisado.

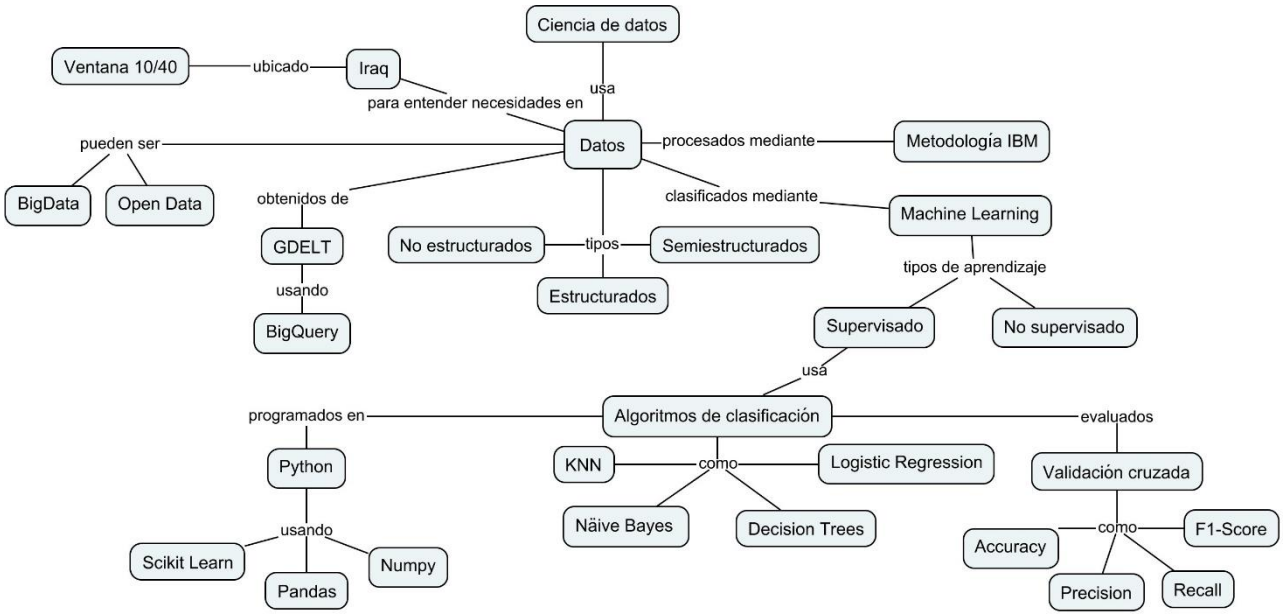

Figura 1. Mapa conceptual.

Datos

De acuerdo con la enciclopedia de términos de microcomputadoras (Christie y Christie, 1984), un dato es una unidad de información, que puede incluir números, letras, símbolos o hechos que representan un objeto, una idea, una condición, una situación o cualquier otro factor, que logran ser procesados o producidos por una computadora. Los datos que se obtienen pueden clasificarse, como se muestra en la Tabla 2.

Tabla 2

Ejemplos de tipos de datos (Maté Jiménez, 2014)

Datos estructurados	Datos semiestructurados	Datos no estructurados
Fichas de clientes	Correos electrónicos	Persona a persona: Comunicaciones en las redes sociales
Fecha de nacimiento	Parte estructurada: destinatario, receptores, tema	
Nombre	Parte no estructurada: cuerpo del mensaje	Persona a máquina: Dispositivos médicos, comercio electrónico, computadoras, móviles
Dirección		Máquina a máquina: sensores, dispositivos GPS, cámaras de seguridad
Transacciones en un mes		
Puntos de compra		

Datos estructurados

Los datos estructurados se refieren a cualquier conjunto de datos que se ajustan a un orden o esquema (Arasu y Garcia Molina, 2003).

Datos semiestructurados

Son datos que no están estrictamente tipificados o almacenados en sistemas de bases de datos convencionales (Abiteboul,1997); por ejemplo, los registros o *logs* de los servidores.

Datos no estructurados

Son datos que no tienen estructura alguna. Por ejemplo, mensajes de medios de comunicación, tales como correos electrónicos, imágenes y registros web (Digital Reasoning, 2018).

Ciencia de datos

Tradicionalmente se han manejado los datos por medio de bases de datos, las cuales han sido útiles en el lado administrativo, organizando datos para la obtención de informes o consultas (Pratt, Smatt y Wynn, 2017). Las bases de datos sirven para almacenar datos estructurados, representando sus entidades y sus interrelaciones (Camps Paré et al., 2005).

Aunque las bases de datos han sido una pieza fundamental en el desarrollo tecnológico, no es posible descubrir patrones ocultos en los datos, por medio de consultas en estas. Los motores de bases de datos solo dan informes de acuerdo con las peticiones del usuario relacionadas con el tipo y estructura de los datos almacenados. En cambio, la ciencia de datos consiste en obtener conocimiento de los datos, sean estos datos estructurados, semiestructurados o no estructurados (Dhar, 2013). Para esto se necesita realizar preguntas acertadas (Phethean, Simperl, Tiropaniss, Tinati y Hall, 2016). Esta ciencia se ha aplicado a diversas ramas del conocimiento, que van desde las ciencias sociales hasta las ciencias de la salud. También se apoya en otras disciplinas, tales como aprendizaje automático, inteligencia artificial, programación y estadística para la mejor comprensión de los datos (Phethean et al., 2016).

En este punto, es importante precisar la diferencia entre estadística y ciencia de datos. La estadística crea modelos con el propósito de aproximar valores, describiendo la estructura y comportamiento de los datos. En cambio, la ciencia de datos y el aprendizaje automático, que es uno de sus componentes fundamentales, se enfoca en la predicción usando valores desconocidos (Fawcett y Hardin, 2017).

La ciencia de datos involucra a investigadores externos o a analistas conocidos como científicos de datos**,** cuya función principal consiste en diseñar modelos o algoritmos para ajustar los datos dependiendo del problema y plantear preguntas con respecto al tema de interés. Al analizar los datos, se identifican patrones y ciertas relaciones entre las variables. Por medio de estos datos es posible revelar ciertas tendencias y fenómenos (Phethean et al., 2016). Es así como diversas empresas, tales como LinkedIn, han usado las habilidades de los científicos de datos para encontrar patrones en los datos que han llevado al aumento de usuarios de la plataforma. Google también se apoya en los científicos de datos para refinar los algoritmos principales de búsqueda y publicación de anuncios. Netflix también aplicó ciencia de datos para mejorar su sistema de recomendación de películas (Davenport y Patil, 2012).

Proyecto GDELT

GDELT es un proyecto patrocinado por Google que contiene más de 200 millones de eventos geolocalizados con cobertura mundial respecto de noticias y acontecimientos importantes desde, 1979 (Leetaru y Schrodt, 2013). GDELT obtiene la información principalmente de las siguientes agencias de noticias: Lexis Nexis, The Agence France–Presse, Reuters, Associated Press y Xinhua. Utiliza como base CAMEO, el cual es un código de contenido para las noticias electrónicas (Yonamine,

2012). El análisis de los datos proporcionados por esta organización, mediante la aplicación de algoritmos de inteligencia artificial puede, revelar ciertos comportamientos, características específicas o posiblemente predecir acontecimientos futuros. Una muestra de esto fue en la predicción sobre niveles de violencia en distritos de Afganistán (Yonamine 2012).

Big Query

Big Query es un servicio de Google que permite hacer el análisis interactivo de datos masivos, con el fin de almacenar estos datos y consultarlos (Camargo Vega, Camargo Ortega y Joyanes Aguilar, 2015). Big Query tiene como características básicas la velocidad, la escalabilidad, la sencillez (lenguaje similar al lenguaje de consultas, SQL por sus siglas en inglés); permite el trabajo en grupo, la seguridad en el acceso y tiene una interfaz sencilla de uso. Google da la posibilidad de usar este servicio de manera gratuita, con un límite de 100 GB de datos almacenados y analizados en una base mensual, accediendo con una cuenta de Gmail (Vrbić, 2012).

Datos masivos

Se conoce como datos masivos o Big Data al conjunto de datos que son mayores a la capacidad de bases de datos tradicionales (Mathur et al., 2014). Para que los datos sean considerados como datos masivos, estos deben incluir cuatro características principales, que se engloban con el término 4V: volumen, velocidad, variedad y veracidad. Respecto del volumen, se refiere al conjunto de datos, que pueden ir de terabytes a petabytes. Es clave mencionar que estos datos van en incremento, de acuerdo con las actividades y funciones que cada empresa o sector

realice. La velocidad describe el tiempo en el que se crean, procesan y analizan los datos. Esto incluye el tiempo de espera desde que se crean los datos, su captura y la disponibilidad para usarlos. Actualmente, son tantos los datos que se generan que se complica la captura, el almacenamiento y su análisis. En la actualidad existe una mayor cantidad de datos no estructurados que estructurados. De hecho, estos datos se expanden rápidamente por diversas fuentes de datos disponibles, tales como imágenes de satélite, fotografías y video, redes sociales, datos móviles, sitios web, entre otros (Hurwitz, Nugent, Halper y Kaufman, 2013). La variedad tiene que ver con los tipos y fuentes de datos. En este aspecto, los datos se clasifican como datos estructurados, semiestructurados y no estructurados. Finalmente, la veracidad se refiere a la fiabilidad de los datos, enfatizando que la obtención de los datos debe ser de alta calidad (Schroeck, Shockley, Smart, Romero Morales y Tufano, 2012).

Las investigaciones con datos masivos se han desarrollado en áreas tales como la astronomía, la meteorología, la atención sanitaria, la salud pública, la política, los estudios de gobierno, la comercialización y las finanzas (Goes, 2014). Se pretende que estas investigaciones ayuden a los profesionales a ejecutar estrategias adecuadas con mayor precisión y eficiencia (Waller y Fawcett, 2013). Según Leung et al., (2016) se pueden obtener datos masivos principalmente de la Internet y los dispositivos móviles (Chan, 2013). Otras fuentes de datos masivos son las redes sociales (tales como Facebook, Google+, Twitter, entre otros), registros web, textos, transacciones comerciales, flujos de marketing o diversos medios sociales (Leung, Jiang, Zhang y Pazdor, 2016). Mediante la obtención de datos por estas fuentes, se puede llegar a conocer a las personas, sus actividades diarias y hasta sus tradiciones (Chan, 2013).

Open data

Open Data o datos abiertos es un término que surge para definir cómo los datos pueden ser utilizados y compartidos sin uso de licencia (Murray, 2008). Existen países, empresas y agencias gubernamentales que no están dispuestos a compartir el conjunto de información con el que cuentan (Shamsuddin y Hasan, 2015). Es por eso que Open Data extiende la posibilidad de acceder a la información, para usarla, modificarla y compartirla, sin ningún tipo de restricción. Los datos abiertos actualmente se aplican en áreas tales como logística, atención médica y servicios de seguros (Vostrovsky, Tyrychte y Ulman, 2015).

Aprendizaje automático

La inteligencia artificial tiene como objetivo diseñar y construir máquinas que puedan imitar el comportamiento inteligente de las personas. En este sentido, se espera que el computador no solo tenga como objetivo su uso en el trabajo o la investigación, sino que pueda ser útil como un intérprete de la realidad (Álvarez Munarriz, 1994).

Según Sánchez Montañés, Lago y González (2011), aprender se refiere a obtener conocimiento de algo a través del estudio, el entrenamiento o la práctica. También puede describirse como un proceso de estímulo del conocimiento. Aprender está relacionado con generalizar y saber reaccionar a casos nuevos: predecir/inferir.

Aprendizaje automático, que es una subárea de la inteligencia artificial, justamente permite que las máquinas aprendan mediante el conocimiento previo de comportamiento pasado (Phethean et al., 2016). Con este conocimiento, las máquinas aprenden (Turkson, Baagyere y Wenya, 2016). El aprendizaje automático es una

herramienta usada para automatizar la toma de decisiones, basándose en patrones encontrados en datos disponibles (Murphree, 2016). Las técnicas de aprendizaje automático pueden clasificarse en dos grupos, como se muestra en la Figura 2. Estas categorías se explican a continuación.

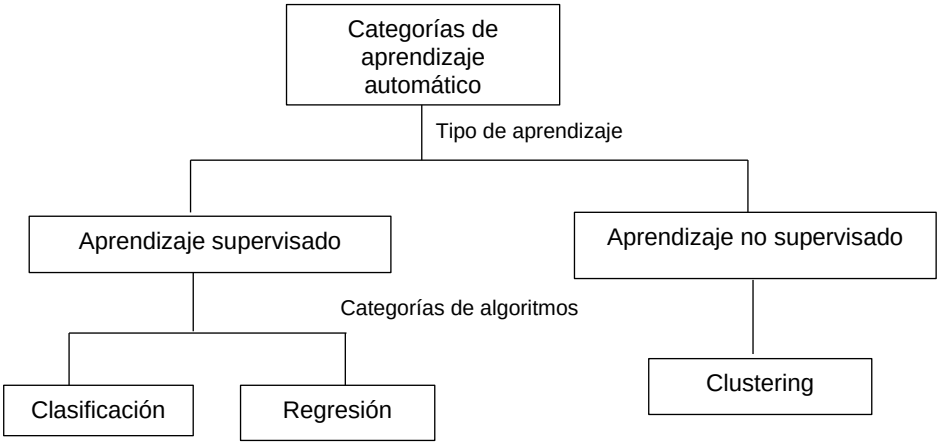

Figura 2. Tareas de aprendizaje automático (adaptado de Turkson et al., 2016).

Aprendizaje supervisado

El aprendizaje supervisado también es conocido como aprendizaje predictivo debido a que, a partir de datos conocidos o desconocidos, puede predecir el valor de una etiqueta estableciendo una relación entre esta y otra serie de atributos. (Moreno García, Miguel Quintales, García Peñalvo y Polo Martín, 2001). Para este tipo de aprendizaje, la computadora utiliza datos de entrenamiento y otros de prueba, con el fin de establecer reglas o procedimientos para nuevas clasificaciones (Learned Miller, 2014). Las categorías del lenguaje supervisado son clasificación y regresión. La diferencia entre estas radica en la variable dependiente. En la clasificación, la variable

dependiente toma valores discretos u ordenados. En la regresión, la variable dependiente toma valores continuos (Wei-Yin, 2011). El objetivo de este aprendizaje es usar los datos de entrenamiento para predecir valores en los datos de salida o pruebas (Hastie, Tibshirani y Friedman, 2009).

Aprendizaje no supervisado

El aprendizaje no supervisado, al no contar con supervisión intenta encontrar patrones en los datos (Hackeling, 2014). Explora datos sin que estos datos contengan una etiqueta pre asignada de salida relacionada con los datos de entrada. Este tipo de aprendizaje puede ser útil para descubrir correlaciones que no se conozcan previamente, agrupando datos con características similares (Murphree, 2016). En otras palabras, el aprendizaje no supervisado no tiene un objetivo o variable para predecir (Turkson et al., 2016).

Algoritmos de clasificación

Los algoritmos de clasificación sirven para desarrollar un modelo predictivo, usando características disponibles en los datos (García, Martínez, Núñez y Guzmán, 1998). En otras palabras, la clasificación consiste en que el sistema aprenda de la información disponible para generalizar la que no está disponible (Flores Pulido, 2012). Para realizar esta clasificación, es necesario dividir el conjuntos de datos en dos tipos: los que serán utilizados para el entrenamiento y los que serán utilizados para las pruebas (Advanced Tech Computing Group UTPL, 2008). A continuación, se describen los algoritmos de clasificación utilizados en este trabajo.

K-Nearest Neighbor (KNN)

El KNN es un algoritmo de clasificación, cuya función específica consiste en la suma de las distancias en un conjunto. Las instancias se asignan al centro más cercano (Kou, Wu, Li y Hong, 2016). La calidad de este algoritmo depende de la forma en que se calculen las distancias entre las diferentes instancias (Nguyen Cong, Rivero Pérez y Morell, 2015). La Figura 3 muestra un ejemplo de KNN que clasifica nuevas instancias dependiendo de la mayoría de votos con sus K vecinos más cercanos (Turkson et al., 2016). En este caso, se tienen tres clases: rojo, verde y azul.

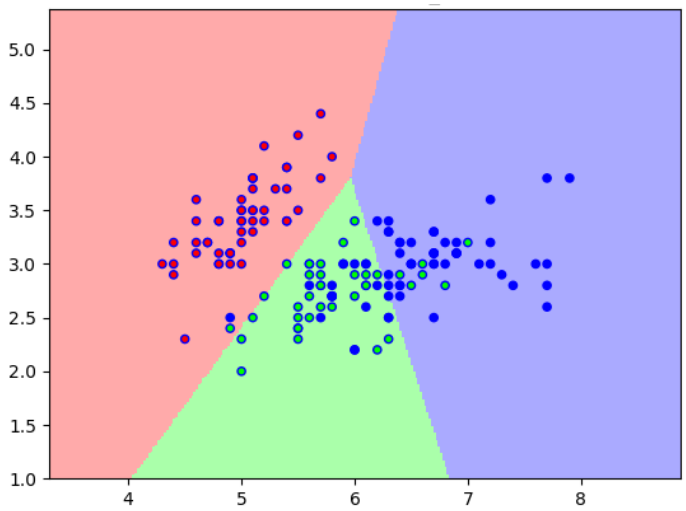

Figura 3. Algoritmo KNN (Scikit Learn, 2017b).

Este algoritmo consiste en las siguientes etapas, según Gómez, Prieto y Guzmán (2015): (a) adquirir muestras con las mismas características, (b) asignar aleatoriamente cada muestra a un espacio determinado, (c) calcular las distancias entre las muestras, para así lograr las respectivas agrupaciones; la más común es la

distancia euclidiana (Esqueda-Elizondo, 2002) y (d) asignar, mediante el algoritmo, la clasificación a la que pertenecen los puntos o muestras, (Moujahid, Inza y Larrañaga, 2008) una vez calculada la distancia entre los puntos.

Näive Bayes

El algoritmo de Näive Bayes es un subconjunto de la teoría bayesiana, la cual consiste en seleccionar la clase con mayor probabilidad, estableciendo las condiciones necesarias. Este algoritmo brinda una manera de estimar probabilidades desconocidas dependiendo de valores conocidos (Harrington, 2012).

Si la probabilidad de que una nueva muestra sea clasificada en la clase 1 es mayor que la probabilidad que la muestra sea clasificada en la clase 2 (p1(x, y) > p2 (x, y)), la nueva muestra se clasifica en la clase 1. Por el contrario, si la clase 2 es mayor que la clase 1 p2(x, y) > p1(x, y), la nueva muestra se clasifica en la clase 2. El resultado de esta clasificación se muestra en la Figura 4, que describe las dos clases posibles de la condición anterior, donde el color rojo representa la clase 1, correspondiente a círculos y el color azul representa la clase 2, correspondiente a triángulos (Harrington, 2012).

Decision Trees

De acuerdo con Kumar, Kumar y Saboo (2016), una de las ventajas que presenta el algoritmo de Decision Trees es que trabaja sobre el principio de elegir atributos con la entropía más baja. La entropía hace referencia a la distribución de elementos distintos y a la distancia entre ellos (Valiant y Valiant, 2017). Hay dos tipos de árboles de decisión: árboles de clasificación y árboles de regresión. En esta

investigación, se utilizaron árboles de decisión de tipo clasificación, debido a que se utilizan valores discretos para realizar las clasificaciones.

Este algoritmo clasifica las instancias por medio de las características de los valores. Cada nodo representa una instancia que se va a clasificar y cada rama representa un valor que el nodo puede adquirir (Kotsiantis, 2007).

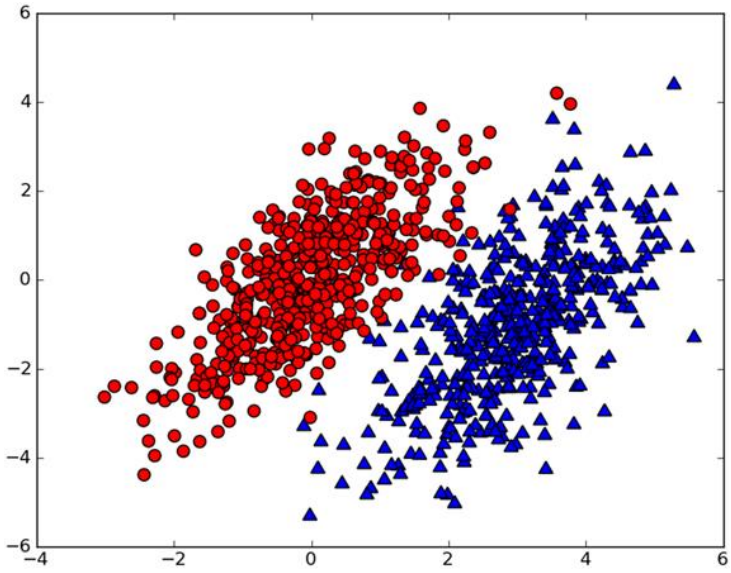

Figura 4. Algoritmo Näive Bayes (Harrington, 2012).

Los árboles de decisión son comúnmente obtenidos al dividir recursivamente el conjunto de entrenamiento basados en valores de las instancias para variables explicativas (Hackeling, 2014). La principal ventaja de este proceso consiste en la facilidad de interpretación de los resultados (Aluja, 2001). Mediante árboles de decisión se pueden manejar grandes cantidades de datos y los resultados de la clasificación son fáciles de deducir y explicar (He, Wang, Zhang y Cook, 2013).

La Figura 5 muestra una descripción gráfica de este algoritmo en el que se utilizó el conjunto de datos llamado “iris flor”. Este algoritmo clasifica los tipos de planta iris. El color azul clasifica iris setosa; el rojo, iris versicolor; y el verde, iris virginica. Esta clasificación se logra mediante el tamaño del pétalo y del sépalo, dos características, que son descritas por los ejes X y Y.

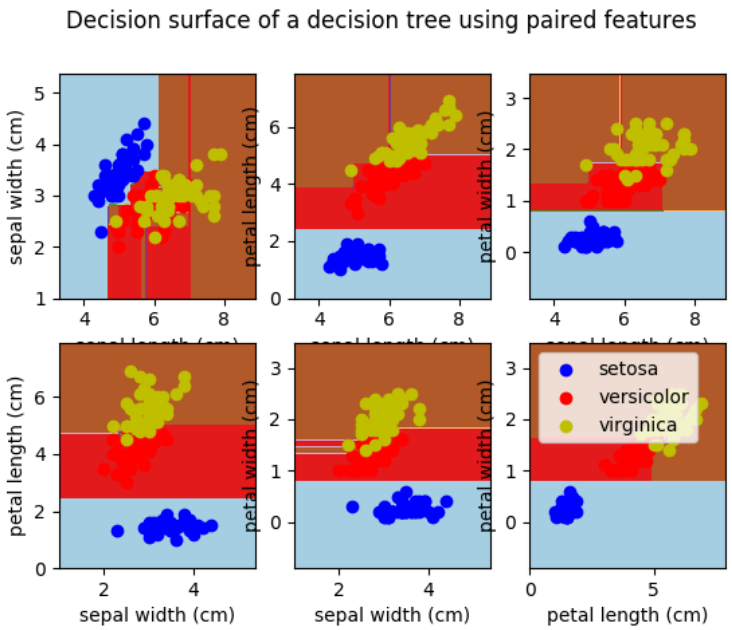

Figura 5. Algoritmo Decision Trees (Scikit Learn, 2017c).

Logistic Regression

La regresión logística usa como método la función logística o función sigmoide, que es una curva en forma de S que podría tomar valores reales y asignarlos a un valor entre 0 y 1, pero nunca exactamente en esos límites (Brownlee, 2016). También este algoritmo, al establecer los límites en los valores, hace uso de la función gradiente para establecer el mejor ajuste a los datos, resolviendo así los problemas de

optimización (Cárdenas Montes, 2014; Kaelo, Narayanan y Thuto, 2017). Como se muestra en la Figura 6, la línea verde es el límite entre los valores; los puntos rojos pertenecen a muestras falsas y los azules, a muestras verdaderas, donde el eje X es la variable dependiente y el eje Y, la (s) variable(s) independiente(s).

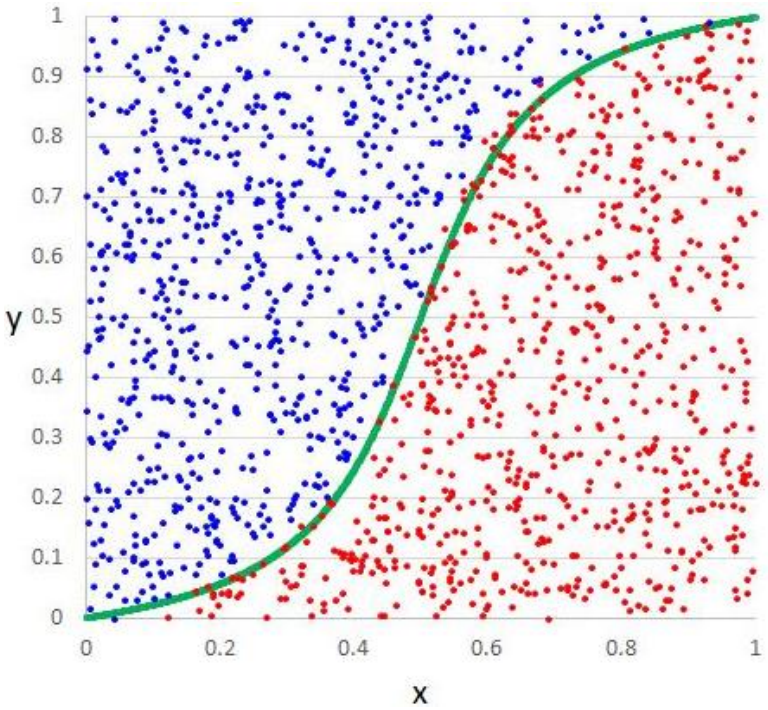

Figura 6. Algoritmo Logistic Regression (Technology of computing, 2016).

Métricas de evaluación

Luego de aplicar los algoritmos a los datos y obtener un modelo, es necesario evaluar los resultados para comprobar su veracidad. Para ello se describen las siguientes variables (Shazadi et al., 2014):

T_n = Verdadero negativo: datos negativos, predichos negativamente. Es decir, datos clasificados correctamente.

T_p = Verdadero positivo: datos positivos predichos positivamente. Esto es, datos clasificados correctamente.

F_n = Falso negativo: datos positivos predichos negativos. Clasificación incorrecta.

F_p = Falso positivo: datos negativos predichos positivos. Clasificación incorrecta.

Para valorar y seleccionar el modelo de clasificación en este estudio, se utilizaron las siguientes medidas de evaluación: accuracy, recall, precision y f1-score, las cuales se describen a continuación.

Accuracy

La accuracy se refiere a la exactitud, ya sea por fracción (por defecto) o por el conteo total de predicciones correctas. Si el total del conjunto de etiquetas predichas para una muestra coincide con el verdadero conjunto de etiquetas, entonces la exactitud se representa con el valor de 1,0; de lo contrario, es 0,0 (Scikit Learn, 2017a). Cuanto más cerca estén las medidas al valor conocido, más precisa será la medición (Helmenstine, 2014).

Como ejemplo, la accuracy radica en la frecuencia en un punto al azar que es distante al resultado esperado o blanco, que en la Figura 7 es el punto rojo.

Figura 7. Accuracy.

La accuracy puede entenderse como el número de predicciones correctas divididas entre el número total de predicciones hechas, multiplicadas por 100, para obtener el valor en porcentaje. No es suficiente considerar la accuracy como el único método de evaluación, debido a que los datos se comportan de manera distinta, por lo que es necesario apoyarse de otras medidas de evaluación para obtener predicciones sólidas (Brownlee, 2014).

Precision

La precision (en inglés) es una medida respecto de cuán cerca las medidas experimentales concuerdan entre sí (Helmenstine, 2014). En la Ecuación 1 se observa que la precision corresponde con el número de verdaderos positivos (T_p) sobre el número de verdaderos positivos más el número de falsos positivos (F_p) (Hackeling, 2014).

$$P = \frac{Tp}{Tp + Fp} \tag{1}$$

Una baja precision puede mostrar gran cantidad de falsos positivos (Brownlee, 2016). La Figura 8 muestra un ejemplo de precision, el cual consiste en la aproximación al punto rojo, considerado como blanco u objetivo.

Figura 8. Precision.

Recall

El recall o alcance se define como el número de verdaderos positivos (T_p) sobre el número de verdaderos positivos (T_p) más el número de falsos negativos (F_n) (Scikit-Learn, 2017d): esto se representa en la Ecuación 2. Por ejemplo, si un chico da una sorpresa de cumpleaños a su novia cada año en los últimos 10 años y la novia pregunta si este recuerda todas las sorpresas, el recall se refiere al número de eventos recordados correctamente respecto del total de eventos correctos (Choochaisri, 2016).

$$R = \frac{Tp}{Tp + Fn} \tag{2}$$

F1- score

El f1-score se define como la media armónica de precision y recall. También es conocida como F-measure. Se obtiene mediante la Ecuación 3 (Scikit-Learn, 2017d):

$$F1 = 2\frac{P \times R}{P + R} \tag{3}$$

Un modelo puede ser considerado perfecto cuando los puntajes de precision y recall alcanzan una puntuación de f1 en 1. A su vez, un modelo con perfecta precision y un puntaje de recall igual a 0, tendrá un f1 igual a 0 (Hackeling, G., 2014).

Python

Python es un lenguaje de programación de sintaxis clara. Python cuenta con una amplia comunidad de usuarios. Asimismo, este lenguaje permite programar mediante el paradigma de programación orientada a objetos o mediante el paradigma funcional. Las librerías de Python son de gran utilidad en la creación de sistemas basados en aprendizaje automático (Harrington, 2012).

De acuerdo con Panos y Christof (2013), Python es ideal en el análisis de datos, por las siguientes razones: (a) la facilidad al interconectar con otros sistemas, (b) la exportación de datos en distintos formatos y (c) los programas hechos en Python pueden interactuar vía web.

Herramientas Python

Python se apoya en módulos y librerías, herramientas que permiten ser usado con mayor facilidad en el aprendizaje automático. Algunas de estas librerías son las siguientes: Scikit Learn, Numpy, Pandas y Anaconda, para una mejor comprensión en la lectura del documento, se describen a continuación.

Scikit Learn

Scikit Learn es un módulo de Python que integra diversos algoritmos de aprendizaje automático. Es fácil de usar y tiene buen rendimiento y documentación (Pedregosa et al., 2011). Asimismo, permite evaluar conjuntos definidos de algoritmos para este tipo de aprendizaje.

Pandas

Pandas está diseñado para analizar datos estructurados. Es útil para el manejo y carga de datos en una hoja de cálculo de Excel, datos en series ordenadas y no ordenadas, o cualquier otra forma de conjuntos de datos estadísticos (McKinney, 2016).

Numpy

Abreviatura para Numerical Python. Esta biblioteca se utiliza en el desarrollo de operaciones con matrices, operaciones de álgebra lineal y generación de números aleatorios. Asimismo, usa herramientas para la conexión de C, C++ y Fortran con Python (McKinney, 2013).

Anaconda

Anaconda es una distribución de Python con herramientas para la ciencia computacional y para la ciencia de datos. Su instalación es gratuita para uso comercial y distribución (Johansson, 2016).

Metodología de ciencia de datos de IBM

La metodología de ciencia de datos de IBM se organiza en diez etapas iterativas para la comprensión y mejor análisis de datos, que abarca desde la comprensión del problema hasta la retroalimentación, estas etapas se describen en la Figura 9 para su comprensión. A continuación, se describe brevemente en qué consiste la metodología de ciencia de datos de IBM (Rollins, 2015):

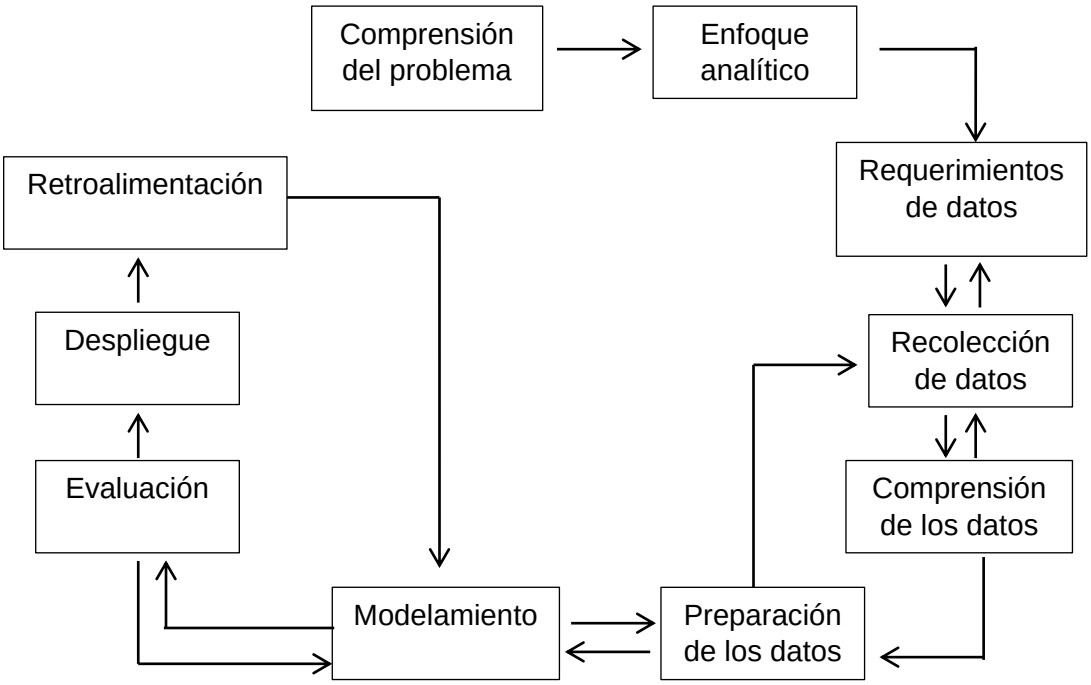

Figura 9. Metodología para ciencia de datos de IBM (Rollins, 2015).

1. Comprensión del problema: en este paso se identifica y comprende el problema o necesidad a resolver.

2. Enfoque analítico: al identificar el problema, se define el enfoque y se delimita el mismo. En este paso también se define el tipo de resultado que se espera obtener y las técnicas que ayudarán a alcanzarlo.

3. Requisitos de datos: en este paso se puntualizan los requisitos de los datos y métodos a utilizar para obtenerlos.

4. Recolección de datos: en este paso se obtienen los datos, considerando los requisitos definidos incluyendo el sitio de provisión de datos y el volumen para el análisis de los mismos.

5. Comprensión de los datos: luego de recopilar los datos, es necesario aplicar técnicas estadísticas u otras para comprenderlos y así evaluar la calidad de los mismos.

6. Preparación de los datos: en este paso se ordenan, verifican y eliminan los datos que no son necesarios en los experimentos. También es posible combinar los datos para tener variables más útiles.

7. Modelado: el científico de datos realiza pruebas con diversos algoritmos para definir cuál es el mejor, de acuerdo con el enfoque analítico establecido. Si se desea generar un modelo predictivo, entonces es necesario usar datos para entrenamiento y otros para prueba.

8. Evaluación: en esta etapa, el científico de datos evalúa el modelo obtenido para comprender la calidad y verificar que se ha abarcado adecuadamente el problema planteado.

9. Despliegue: al obtener un modelo satisfactorio, se despliega el modelo en un ambiente de prueba. El desarrollo puede ser simple, como generar un informe con recomendaciones o más sofisticado, dependiendo de las necesidades planteadas.

10. Retroalimentación: al tener el modelo establecido, el científico de datos interactúa con la organización recibiendo información de acuerdo con el rendimiento del modelo. Dependiendo de la retroalimentación, el científico de datos concluye si debe refinar el modelo.

Ventana 10/40

En su artículo, Gustin (2006) describe a la Ventana 10/40 como una región que abarca el Norte de África, el Medio Oriente, Asia Central y Oriental. Se conoce con este nombre por ubicarse entre los paralelos 10 y 40 al norte del Ecuador. También

explica que esta área cuenta con características específicas: (a) allí vive la mayoría de la población del mundo, (b) las personas en esta zona son más receptivas a lo espiritual, (c) más del 80% de ellas tiene el estándar de vida más bajo y (d) el cristianismo brinda menos del 10 por ciento de sus fondos.

Uno de los hechos más preocupantes en esta zona consiste en que la mayoría de las personas no cuentan con alguna presencia cristiana entre ellos; también es preocupante que muchas de las personas de este lugar no han escuchado sobre el evangelio de manera significativa y, para poder acercarse a ellas, es necesario estar informado sobre una o más de las principales religiones y la cultura que destacan en ese lugar. La Figura 10 muestra un mapa de esta región.

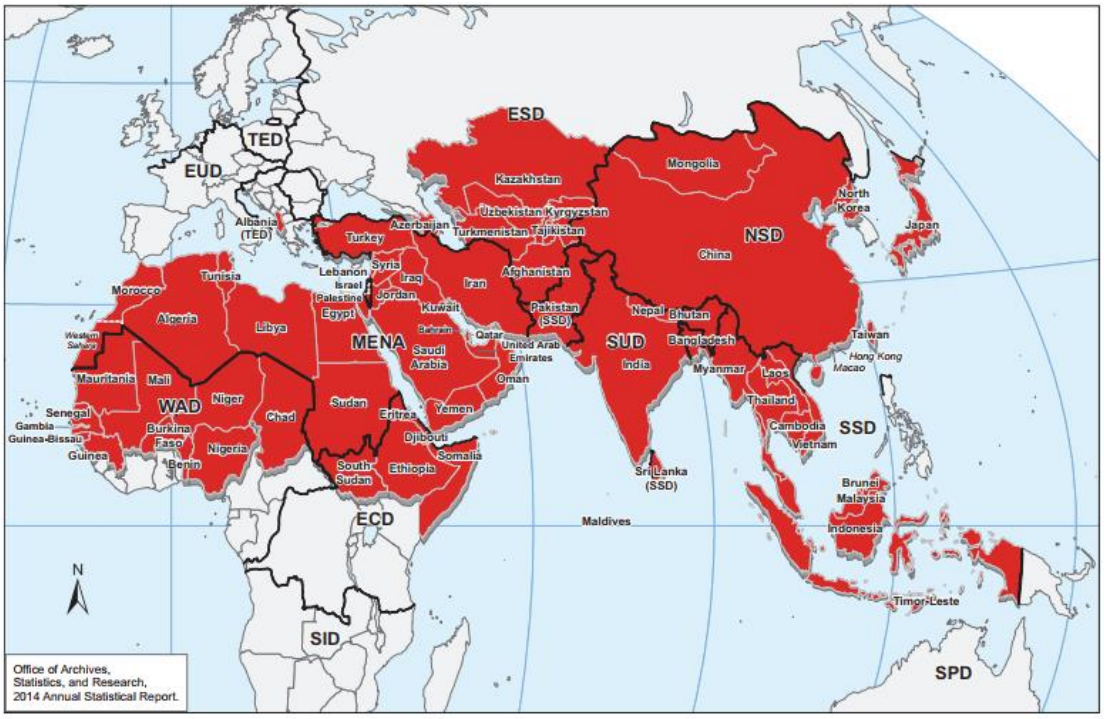

Figura 10. Ventana 10/40 (Seventh-day Adventist Church, 2014).

Iraq

País del suroeste de Asia, oficialmente conocido como República de Iraq, abarca mayormente la región noroeste de la cadena montañosa de Zagros, la parte oriental del desierto de Siria y la parte norte del desierto de Arabia. Este país cubre un área de 437.072 km^2 y, dentro de la categoría de países más grandes del mundo, ocupa el puesto 58. Su cultura se fundamenta tradicionalmente en el árabe (Ecured, 2017).

Trabajos relacionados

Se han desarrollado diversas investigaciones usando el sitio de noticias de GDELT. Por ejemplo, Qiao y Wang (2015) detectaron eventos de protestas en Nueva York y Hong Kong. En la evaluación de estos resultados, se utilizaron métricas como accuracy, el coeficiente de relación de Matthews y validación cruzada. También se crearon gráficas para analizar los resultados. El grado de precision de estos fue de 0.921.

Por otra parte, Su, Lan, Lin, Comfort, Joshi (2016) describen el análisis de datos del terremoto en Nepal ocurrido en 2015. Los datos fueron obtenidos de GDELT y los resultados muestran el aumento y disminución del interés de las personas respecto de contribuir por medio de donaciones. También explican el apoyo del gobierno y la eficacia de las organizaciones internacionales y locales. Asimismo, describen la importancia de un monitoreo cuidadoso y la pronta atención después de los desastres. Los datos fueron analizados por medio de gráficas y se obtuvieron conclusiones con respecto al comportamiento de los datos.

Bi, Gao, Wang y Cao (2015) usaron los datos de GDELT para estimar el grado de actividad e influencia de países respecto de la política, la economía, el comercio, la cultura, la milicia, entre otros. En su investigación, se destacan los países como

Estados Unidos, China y Rusia. Los datos fueron analizados por medio del cálculo de matrices, en las cuales los países involucrados fueron ordenados alfabéticamente y se les asignaron los valores obtenidos de GDELT para realizar los cálculos. Los resultados obtenidos de las matrices fueron evaluados por medio de gráficas.

Al-Naami, Seker y Khan (2014) usaron los datos de GDELT para entender y encontrar tendencias y comportamientos espaciales, sociales, perceptuales e internacionales. Los autores concluyeron que una base de datos tradicional no es suficiente para procesar y analizar los datos de GDELT. En esta investigación, el uso de los sistemas MapReduce y Hadoop, junto con la extensión Geographic Information System Querying Framework (GISQF) permitieron mejorar el análisis y el manejo de los datos.

Kumar, Benigni y Carley (2016) utilizaron los datos de GDETL para analizar por medio de gráficas la percepción de la población hacia los ataques cibernéticos en Estados Unidos. Al observarlos, concluyeron que los ataques cibernéticos han disminuido en relación con los cambios en la política cibernética del país. También se encontró que el sentimiento de la población hacia ese tipo de noticias es cada vez más negativo. Asimismo, se encontró que el aumento de ataques cibernéticos fue de entre un 50% y un 60% entre el 2013 y el 2016. La obtención de resultados fue sustentada con fórmulas bayesianas de probabilidad.

En otra investigación, Chen (2017) tomó datos de GDELT con respecto al discurso de negocios de Trump y analizó el efecto que este tuvo sobre la opinión pública, usando el algoritmo GCAM (sistema global de análisis de medición de contenidos). Se abarcaron diversas dimensiones de las medidas emocionales con base en la positividad y la negatividad. Se mostraron gráficas para conocer la media

sobre cuestiones respecto del impacto del comercio, los balances comerciales y los tipos de cambio. En esta investigación, se utilizó GDELT como medio efectivo para analizar y conocer las opiniones públicas, pasando mayormente de opiniones positivas a negativas. Para el análisis de estos datos, se mostraron gráficas que describen el comportamiento de estas opiniones, comparando las cuestiones entre sí con respecto a las emociones.

Uso de ciencia de datos en la FIT de la UM

En la Facultad de Ingeniería y Tecnología (FIT) de la Universidad de Montemorelos (UM) se han desarrollado investigaciones utilizando aprendizaje automático. Por ejemplo, se ha utilizado aprendizaje automático para la detección de melanomas malignos en muestras de nevus (Marín, Alférez, Córdova y González, 2015). Las nevus se describen como una alteración congénita en la pigmentación de la piel (Real Academia Española, 2017). Asimismo, se realizó un trabajo en colaboración con científicos de Loma Linda University acerca de la interpretación y validación de componentes geoquímicos mediante aprendizaje automático (Alférez, Rodríguez, Clausen y Pompe, 2015). También se construyó una herramienta basada en aprendizaje automático para mejorar el control del tráfico durante las horas pico en ciudades inteligentes, al permitir que el sistema de tráfico aprenda de forma automática acerca del flujo de vehículos. Los temporizadores de los semáforos cambian automáticamente de forma proactiva (antes que el problema de tráfico sea evidente), dependiendo del tráfico (Zavala y Alférez, 2015).

Asimismo, se aplicó ciencia de datos para corroborar la relación que existe entre diabetes y caries dental. En esta investigación se utilizaron datos de historias clínicas

de la Clínica Dental de la Universidad de Montemorelos (Alférez et al., 2016). Otra aportación interesante se dio al analizar, mediante ciencia de datos un conjunto de datos abiertos de la Secretaria de Salud en México respecto de las causas de muerte materna (Domínguez, 2017). En esta investigación se logró la clasificación de dos causas de muerte materna: eclampsia durante el parto y padecimiento de hemorragia.

Discusión

Los trabajos de investigación presentados en esta sección demuestran que los datos del sitio de noticias GDELT son apropiados para obtener información y resultados confiables. En algunas de las investigaciones presentadas en esta sección, se utilizaron técnicas estadísticas y matemáticas tales como las redes bayesianas. Otras investigaciones se apoyaron en el uso de gráficas o en la aplicación de cálculo de matrices con los datos obtenidos. No obstante, en las investigaciones mencionadas, no se utilizó inteligencia artificial ni algoritmos de clasificación. Por esta razón y siendo que GDELT es una plataforma confiable de datos usados por diversos autores en sus investigaciones, en este estudio se explota GDELT mediante la aplicación de ciencia de datos por medio de técnicas de aprendizaje automático.

CAPÍTULO III

METODOLOGÍA

Introducción

En la realización de la presente investigación se siguió la metodología de ciencia de datos de IBM (Rollins, 2015), introducida en el Capítulo II. Para facilitar la comprensión de los lectores, se hace referencia a las características de los datos seleccionados, su recopilación, la limpieza de los mismos, el análisis de las variables y su interpretación, al ser procesados con los algoritmos de aprendizaje supervisado como KNN, Näive Bayes, Decision Trees y Logistic Regression.

Metodología de ciencia de datos de IBM

La metodología de IBM está organizada en diez etapas que representan un proceso iterativo (Rollins, 2015). De manera frecuente, la ciencia de datos tiende a seguir un proceso general que incluye recopilación, limpieza, análisis y modelado de datos y permite su visualización y reporte (Phethean et al, 2016). En ciencia de datos, los datos obtenidos son procesados de manera distinta a las bases de datos tradicionales, por lo que se necesita aplicar una metodología para extraer de ellos la información o conocimiento necesario. De hecho, el uso de una metodología permite a los ingenieros desarrollar pruebas y así conducirse a modelos más eficaces (Murphree, 2016). A continuación, se describe la aplicación de la Metodología de ciencia de datos de IBM en esta investigación.

Comprensión del problema

En la comprensión del problema se describe que la Ventana 10/40 es una zona con muy pocos cristianos. Es por ello que, al conocer las tendencias, patrones o necesidades de esta zona por medio de ciencia de datos, se establecería un hito importante en la IASD para alcanzar a las personas con el mensaje de salvación. Como se describe en el Capítulo I, la Ventana 10/40 es una zona con culturas y creencias diferentes y hay lugares a los que no es posible tener un acercamiento directo. Esto impide obtener información sobre ellos respecto de sus necesidades o situaciones específicas. En esta investigación, se tomó el país de Iraq como caso de estudio, considerando que es uno de los países de la MENA con datos no disponibles, debido a restricciones legales.

Enfoque analítico

En la etapa del enfoque analítico, se aplicó aprendizaje automático para la comprensión y análisis de datos respecto de refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres. Estos eventos fueron seleccionados debido a su constante mención en noticias de la Organización de las Naciones Unidas (2017). En los experimentos, se utilizó Python junto con Scikit Learn para programar los siguientes algoritmos de aprendizaje automático: KNN, Decision Trees, Näive Bayes y Logistic Regression. Para el desarrollo de esta investigación, se optó por el aprendizaje supervisado, debido a que este predice datos con etiquetas desconocidas, a partir del aprendizaje de datos conocidos (Moreno García et al., 2001). En este caso, las clases son los eventos estudiados (ver Tabla 3).

Tabla 3

Descripción de eventos

Evento	Descripción
REF	Refugiados
073	Ayuda humanitaria
145	Protestas violentas
194	Lucha con artillería
202	Masacres

Requisitos de datos

En este paso de los requisitos de datos se realizó una consulta en Big Query para obtener los datos de GDELT con respecto a los eventos seleccionados en Iraq, que incluyen refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres. Los 42,027 datos obtenidos durante la consulta incluyen latitud y longitud del país de Iraq. Para realizar la consulta, se obtuvieron datos entre los años de 2012 y 2015 inclusive, siendo esto los años más próximos a la fecha actual; esto con el fin de obtener resultados relacionados con la situación reciente en ese país. Específicamente, en la Tabla 4 se muestra la descripción de variables usadas en la consulta de Big Query, donde el Actor se refiere a las entidades involucradas en los eventos (GDELT, 2013; Schrodt, 2012), que pueden referirse como Actor1 o Actor2. También se muestra la variable utilizada para asignar latitud y longitud en la consulta, la cual corresponde a Actor1Geo_Lat y Actor1Geo_Long. El Evento se refiere a una categoría respecto de algún acontecimiento, de acuerdo con el manual de CAMEO (Schrodt, 2012).

Tabla 4

Descripción de variables usadas en la consulta

Código	Descripción	Tipo
ActionGeo_CountryCode	País donde ocurrió el evento.	Cadena de caracteres
Actor1Geo_Lat	Latitud de la localización del actor en la noticia.	Flotante
Actor1Geo_Long	Longitud del actor.	Flotante
Actor1Type1Code	Tipo o rol del actor. Puede ser un rol específico como fuerzas de policía, gobierno, militares, refugiados u otro tipo de organización.	Cadena de caracteres
EventCode	Entidades relacionadas al evento	Cadena de caracteres
Year	Año del evento.	Entero

Recolección de datos

Los datos fueron obtenidos de GDELT mediante Big Query; usando consultas tipo SQL (Structured Query Language, lenguaje de consulta estructurada). Los resultados fueron descargados en formato CSV. En la Figura 11, se muestra la consola de administración de Big Query. Para acceder a este sitio, es necesario tener una cuenta de Google y crear un nuevo proyecto para almacenar las consultas. En la parte izquierda de la Figura 11 se muestra el menú principal donde se crea el nuevo proyecto para realizar las consultas.

Para la consulta en Big Query, se consideraron los límites del país de Iraq en términos de latitud y longitud. Específicamente, el código CAMEO permite la obtención de datos por medio de latitud y longitud. Con el fin de conocer la longitud y la latitud de

Iraq, se utilizó Google Maps debido a que esta herramienta ofrece un gran soporte en mapas e imágenes de satélite (Dincer y Uraz, 2013). Los límites utilizados en las consultas para el país de Iraq son los siguientes: (latitud > 29.12 < 37.29) y (longitud >39.22 <48.48). El ActionGeo_CountryCode corresponde a IZ. Estos fueron utilizados para la consulta de Big Query de Google.

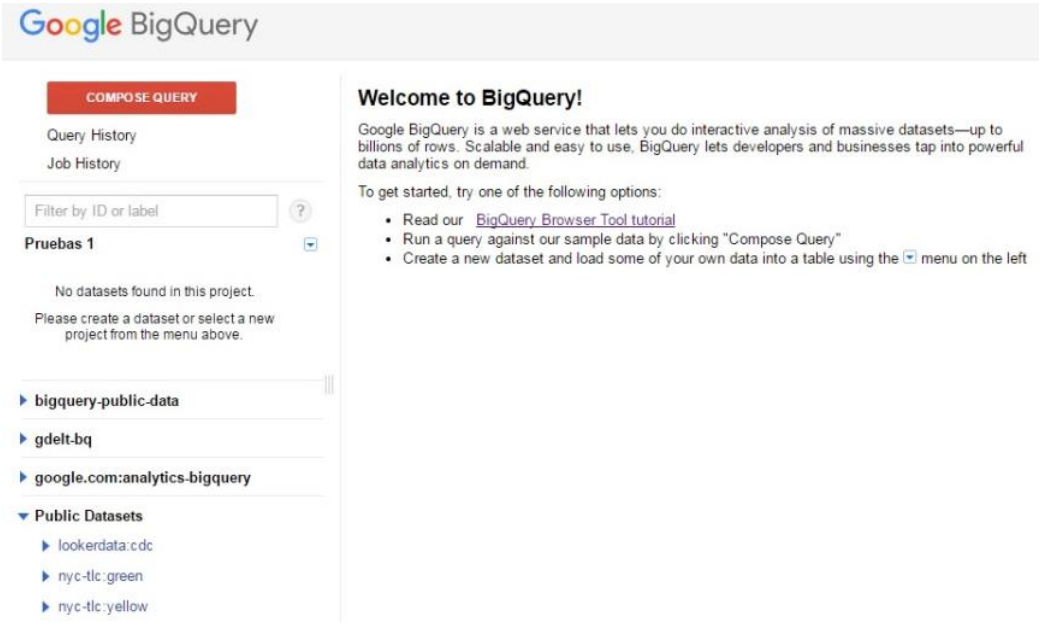

Figura 11. Consola de administración de Big Query.

La Figura 12 muestra los límites de latitud y longitud de Iraq, obtenidos mediante Google Maps, dados con el fin de obtener una visualización de esta zona y al final compararla e interactuar con los resultados de las consultas para los demás eventos. Este mapa puede ser consultado para su interacción en la siguiente liga: https://www.google.com/maps/d/edit?hl=es&mid=1YkzO6lzVeqJycs2n4vcIUUoeHHk&ll=34.543731422056055%2C42.85831259605618&z=6.

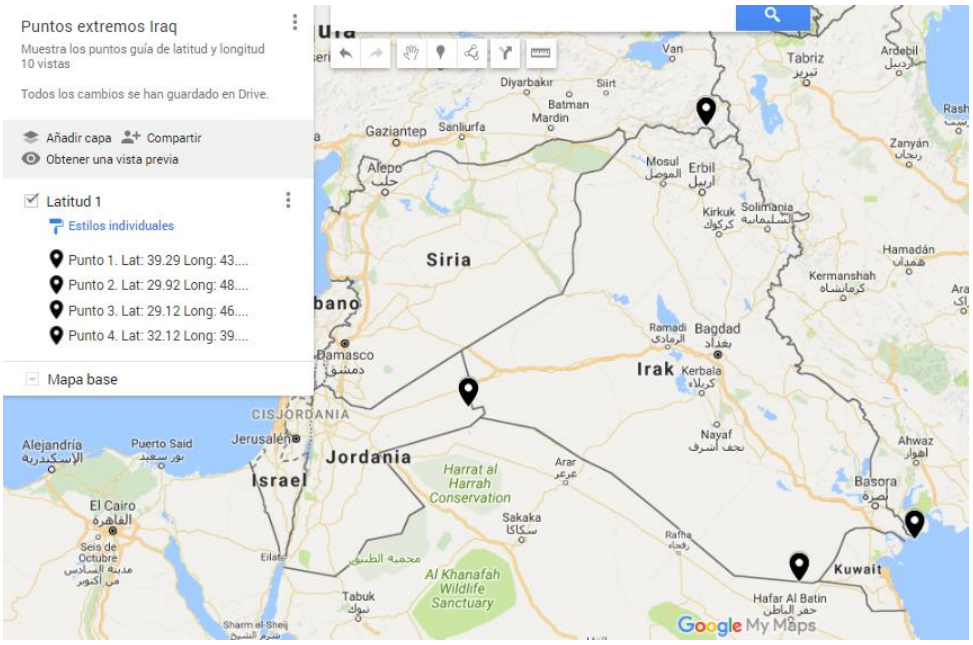

Figura 12. Latitud y longitud de Iraq, obtenido de Google Maps.

Para realizar las consultas, se utilizó el formato SQL en Big Query, manipulando los tipos de eventos y su símbolo con base en el código de GDELT–Data Format Codebook V 1.03 (GDELT, 2013), se obtuvieron los datos de acuerdo al tipo de evento y se describieron la cantidad de datos obtenidos en cada consulta del evento, así como también el total de datos, como se muestra en la Tabla 6.

La Figura 13 muestra el resultado de la consulta de la Tabla 5, que incluye el tiempo en el que se realizó la consulta y el tipo de formato para guardar los resultados. El Apéndice A describe el código SQL usado para obtener los datos de cada evento, utilizando Big Query de Google.

Tabla 5

Consulta sobre refugiados en Iraq

Código SQL	Descripción
SELECT Actor1Type1Code,Year, ActionGeo_CountryCode, Actor1Geo_Lat,Actor1Geo_Long, EventCode	Selecciona las variables a usar: tipo de evento, año, país, latitud, longitud y el código el evento
FROM [gdelt-bq:full.events]	Indica que los datos fueron obtenidos de GDELT
WHERE Actor1Type1Code="REF"	Especifica que la condición para esta consulta es obtener datos con el código "REF"
AND (Year> 2011 AND Year < 2016	Condición para especificar el período de tiempo
AND (Actor1Geo_Lat > 29.12 AND Actor1Geo_Lat < 37.29)	Condición para especificar el rango de latitud
AND (Actor1Geo_Long > 39.22 AND Actor1Geo_Long < 48.48)	Condición para especificar el rango de longitud
AND Actor1Geo_Lat IS NOT NULL AND Actor1Geo_Long IS NOT NULL	Condición donde especifica que no se incluyan latitudes y longitudes vacías que puedan existir en los datos de GDELT

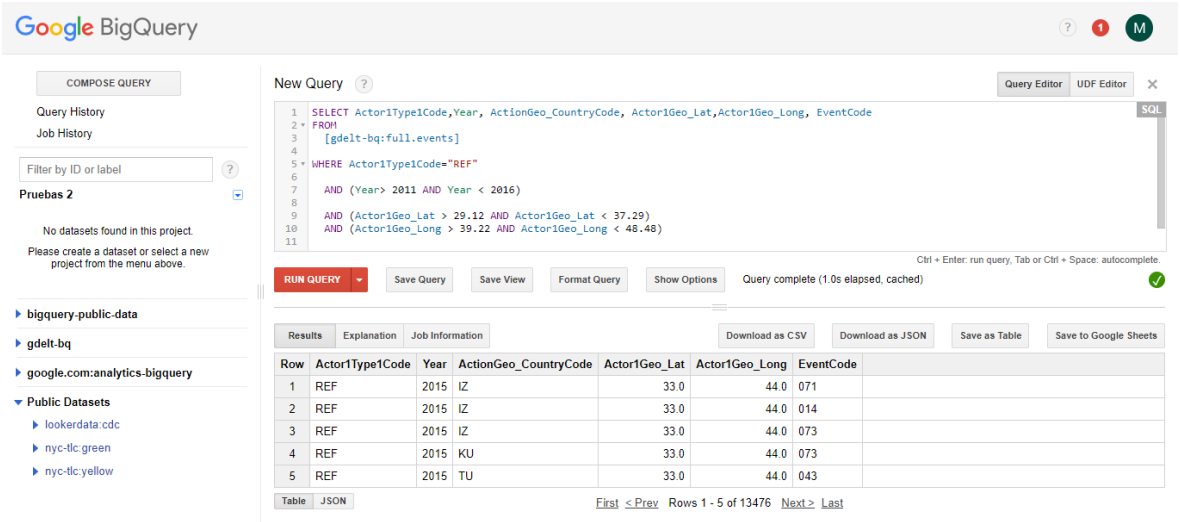

Figura 13. Resultado de consulta SQL sobre refugiados.

La Figura 14 muestra el archivo descargado de la consulta en la Tabla 5. El archivo está en formato CSV y contiene un total de 42,027 instancias. La Tabla 6 muestra la distribución de estos datos, abarcando el período de 2012 a 2015. El archivo CSV se puede descargar de la siguiente liga para su administración: https://drive.google.com/drive/folders/0B6LEG8jNfAY9MEJBTkZoYlRIZHc.

Tabla 6

Datos obtenidos en Big Query

Eventos	Descripción	Período 2012-2015
073	Ayuda humanitaria	10,414
145	Protestas violentas	3,068
194	Lucha con artillería	13,247
202	Masacres	1,822
REF	Refugiados	13,476
	Total datos	42,027

Comprensión de los datos

Para la comprensión y manejo inicial de los datos obtenidos de GDELT por medio de Big Query, se utilizó Microsoft Excel y Google Maps. Los mapas se basaron en los datos obtenidos mediante la consulta con Big Query. Los resultados de estos eventos se observan en las Figuras 15, 16, 17, 18 y 19. Algunos puntos exceden el límite establecido respecto de la latitud y a la longitud de Iraq.

	A	B	C	D	E	F	G
1	Actor1Type1Code	Year	ActionGeo_CountryCode	Actor1Geo_Lat	Actor1Geo_Long	EventCode	
2	REF	2015	IZ	33	44	71	
3	REF	2015	IZ	33	44	14	
4	REF	2015	IZ	33	44	73	
5	REF	2015	KU	33	44	73	
6	REF	2015	TU	33	44	43	
7	REF	2015	IZ	34.3482	45.3906	43	
8	REF	2015	IZ	36.8115	42.999	233	
9	REF	2015	IZ	36.8115	42.999	233	
10	REF	2015	IZ	33	44	12	
11	REF	2015	IZ	33	44	180	
12	REF	2015	IZ	34.3482	45.3906	43	
13	REF	2015	IZ	33	44	51	
14	REF	2015	IZ	33	44	51	
15	REF	2015	IZ	33.3558	43.7861	173	
16	REF	2015	IZ	33	44	20	
17	REF	2015	KU	33	44	173	
18	REF	2015	IZ	33	44	42	
19	REF	2015	AS	33	44	120	
20	REF	2015	IZ	33	44	141	
21	REF	2015	IZ	33	44	42	
22	REF	2015	IZ	33	44	42	
23	REF	2015	SW	33	44	120	
24	REF	2015	SW	33	44	233	
25	REF	2015	IZ	35.2523	45.5582	37	

Figura 14. Resultados de la consulta con respecto a refugiados.

Esto se debe a que estas distancias se obtuvieron de forma manual, utilizando los límites extremos en el mapa. En los mapas, los eventos seleccionados tienen mayor incidencia en las zonas norte y este del país. También sobresale que en Iraq el evento de ayuda humanitaria abarca la mayor parte del país, al igual que el evento de lucha con artillería. Aunque la simple visualización de los datos puede servir para encontrar patrones, es importante notar que hay zonas en donde no existen datos disponibles de acuerdo con los eventos establecidos. Es en estas zonas sin o con pocos datos en donde el uso de aprendizaje automático es de suma importancia para obtener clasificaciones que puedan ayudar a conocer las necesidades de las personas.

La Figura 15 muestra un mapa realizado con los datos obtenidos de la consulta con Big Query respecto del evento de refugiados, siendo este el primer mapa obtenido de las

consultas realizadas, estos datos están representados con el color azul, para su fácil distinción. Este mapa está disponible en la siguiente liga para su consulta: https://www.google.com/maps/d/edit?hl=es&hl=es&authuser=0&authuser=0&mid=1uEeGUsxz38AdaYyYeHCJdWt5mG8&ll=33.137338087845784%2C43.81860000000006&z=6

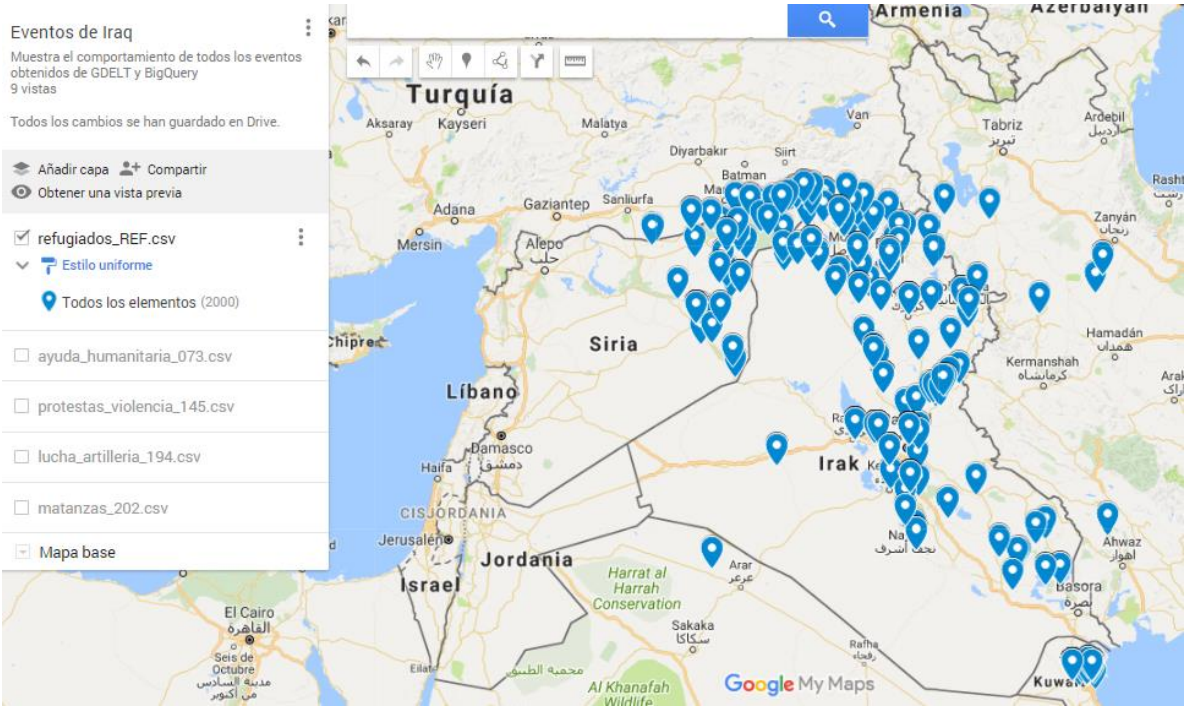

Figura 15. Mapa del evento REF, refugiados.

La Figura 16 muestra un mapa realizado con los datos obtenidos de la consulta con Big Query respecto del evento de ayuda humanitaria; a este resultado se le asignó el color violeta, el cual está disponible en la siguiente liga para su interacción: https://www.google.com/maps/d/edit?hl=es&hl=es&authuser=0&authuser=0&mid=1_qcVLrEUJ3jrDuzfD3LBu4qH4MU&ll=36.22191745381255%2C40.61688906249992&z=5.

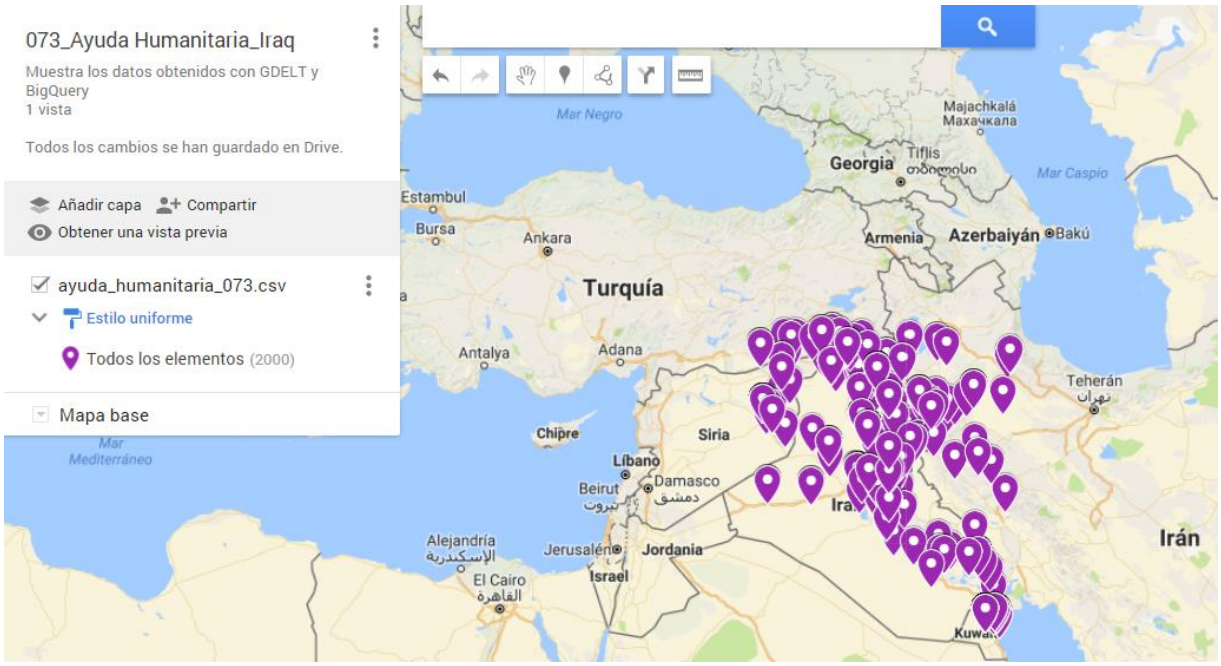

Figura 16. Mapa del evento 073, ayuda humanitaria.

La Figura 17 muestra un mapa realizado con los datos obtenidos de la consulta con Big Query respecto del evento de protestas violentas; a este resultado se le asignó el color naranja, el cual está disponible en la siguiente liga: https://www.google.com/maps/d/edit?hl=es&hl=es&authuser=0&authuser=0&mid=17Ph3fCJn-8De-HBsLEmoPuFjmnI&ll=34.414148847307146%2C43.05271503906249&z=6

La Figura 18 muestra un mapa realizado con los datos obtenidos de la consulta con Big Query respecto del evento de lucha con artillería, el cual está disponible en la siguiente liga para su interacción: https://www.google.com/maps/d/u/0/edit?hl=es&hl=es&mid=1uEeGUsxz38AdaYyYeHCJdWt5mG8&ll=32.75666932007098%2C43.818600000000006&z=6

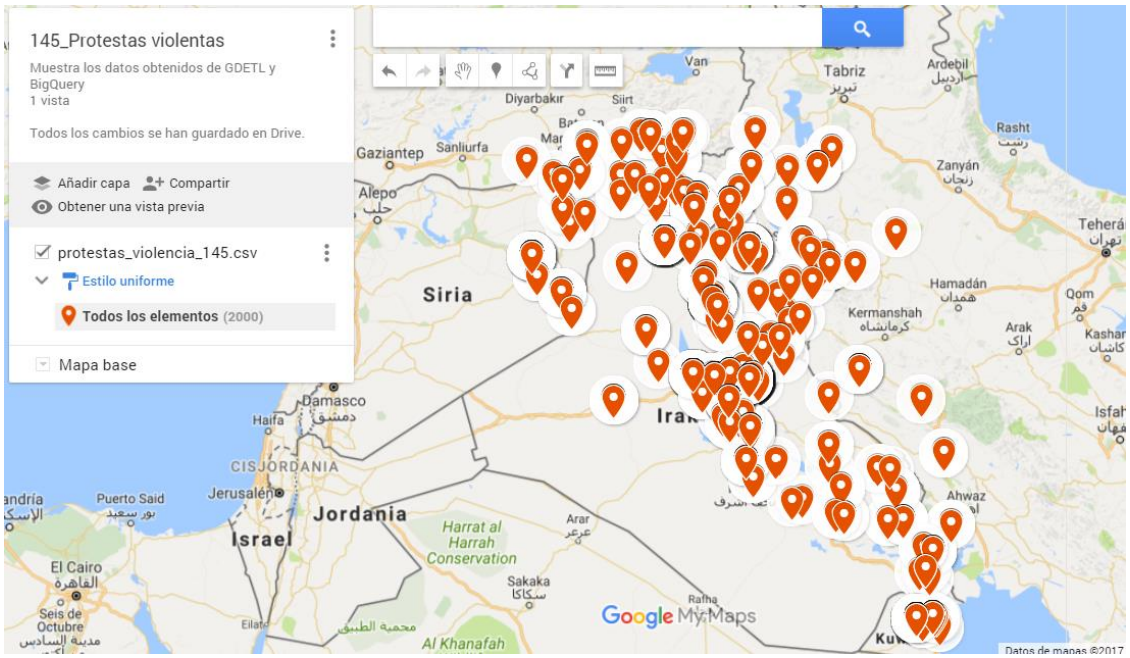

Figura 17. Mapa del evento 145, protestas violentas.

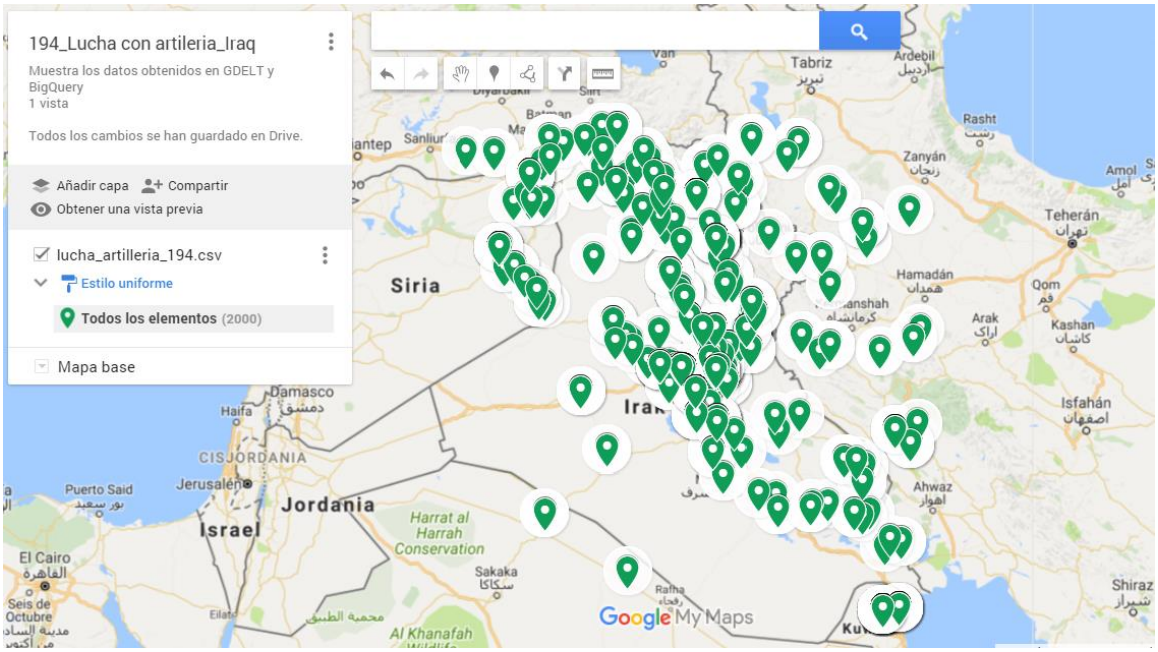

Figura 18. Mapa del evento 194, lucha con artillería.

La Figura 19 muestra un mapa realizado con los datos obtenidos de la consulta con Big Query respecto del evento de masacres, el cual está disponible en la siguiente liga: https://www.google.com/maps/d/edit?hl=es&hl=es&authuser=0&authuser=0&mid=1mjsmw-2KnWbpmLMbGOU88ijClpc&ll=34.36062796905575%2C43.213007812499995&z=6

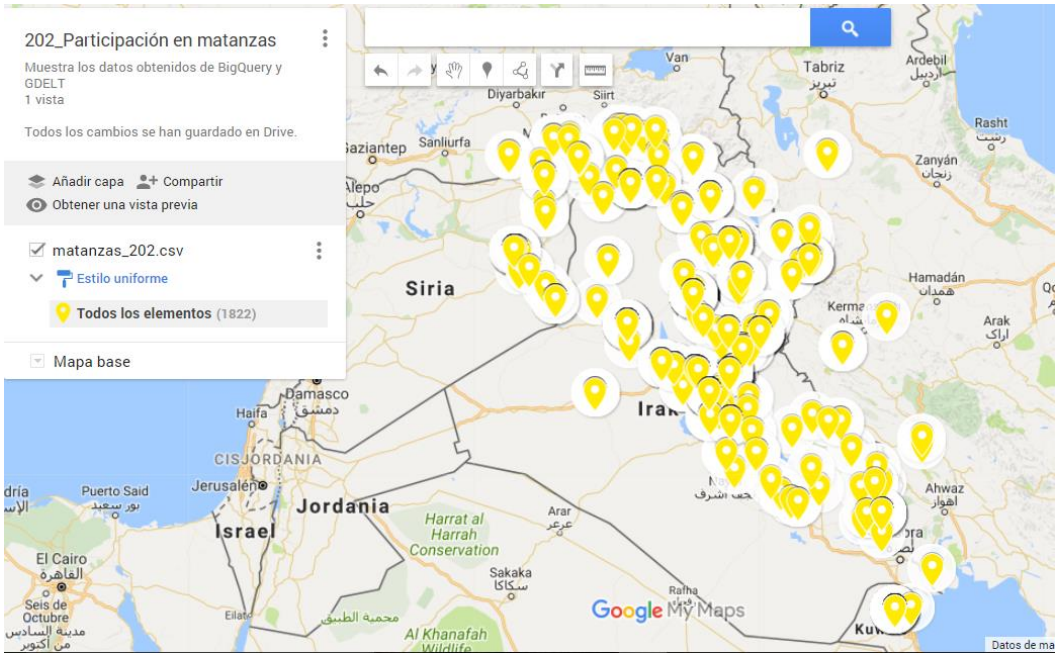

Figura 19. Mapa del evento 202, masacres.

Preparación de los datos

De los atributos usados en la consulta en Big Query se usaron las variables Latitud, Longitud, EventCode y Actor1Type1Code. Al evento de refugiados se le asignó el valor de 0 para su manejo en Python, debido a que los algoritmos de aprendizaje automático trabajan con datos de tipo numérico. Se realizaron consultas individuales por cada

evento y luego de obtener los resultados de los 5 eventos, se creó un nuevo archivo en formato CSV agrupando todos los eventos para aplicar las técnicas de aprendizaje automático a un solo conjunto de datos y así analizar el comportamiento de estos datos.

Modelado

En el paso del modelado, se generaron modelos predictivos mediante aprendizaje automático, utilizando los siguientes algoritmos de clasificación: KNN, Näive Bayes, Decision Trees y Logistic Regression, por formar parte de los mejores algoritmos para el análisis de datos mediante aprendizaje automático (Wu, et al., 2008). Para el modelado se utilizaron como clases los eventos de refugiados, la ayuda humanitaria, las protestas violentas y las masacres. Asimismo, se consideraron como características o features los valores de latitud y longitud. Los resultados de estos clasificadores se describen en el capítulo IV.

Evaluación

Evaluar los clasificadores es un tema dinámico e interesante porque el comportamiento y desempeño de ellos depende de la complejidad de los datos (Gómez, et al., 2015). Para la evaluación de los datos, se optó por graficar en mapas los datos obtenidos de la consulta y la clasificación de alguna nueva longitud y latitud a analizar. Los resultados fueron comparados con mapas oficiales como The United Nations High Commissioner for Refugees (UNHCR) y Live Universal Awareness Map (Liveumap). También, los modelos predictivos obtenidos por cada uno de los

algoritmos de clasificación fueron evaluados con el cálculo de la accuracy, la precision, el recall y el f1-Score. Para realizar las pruebas, se dividió el conjunto de datos en dos sub-grupos: 70% de las instancias fueron usadas para el entrenamiento y el 30% para la evaluación del modelo creado.

Despliegue

En el paso del despliegue se elaboró un software para clasificar zonas de Iraq relacionadas con los eventos que obtuvieron el modelo predictivo óptimo, los cuales corresponden a refugiados y lucha con artillería. El software se construyó en Python debido a que este es un lenguaje de programación recomendable para aprendizaje automático (Harrington, 2012). Para la construcción de la interfaz de usuario, se utilizó la biblioteca estándar TKinter (Marzal, Llorens y Gracia, 2003). El ambiente del software fue desplegado en Microsoft Windows de 64 bits, utilizando como herramientas de desarrollo la versión 2.7.12 de Python y Anaconda 4.2.0.

Diagrama de casos de uso

Los diagramas UML (lenguaje unificado de modelado) son útiles para comprender el comportamiento funcional del software, incluyendo las secuencias de actividades y los procesos (Chen, Jiang, Hong y Lin, 2018). A continuación, se describen de manera breve los casos de uso y los diagramas desarrollados en esta investigación.

1. Ejecución del clasificador: el científico de datos crea modelos de clasificación, utilizando KNN, Decision Trees, Näive Bayes y Logistic Regression.

2. Evaluación del clasificador: el científico de datos evalúa los modelos de clasificación, utilizando accuracy, recall, precision y f1-score.

3. Introducción de datos: el usuario final ingresa en la interfaz la latitud y la longitud del país de Iraq que desea clasificar.

4. Procesamiento del modelo: el algoritmo procesa los datos y los compara con el modelo establecido por el científico de datos.

5. Obtención de la clasificación: el usuario final y el científico de datos muestran la clasificación de la latitud y la longitud ingresada.

En la Figura 20 se muestra el diagrama de casos de uso.

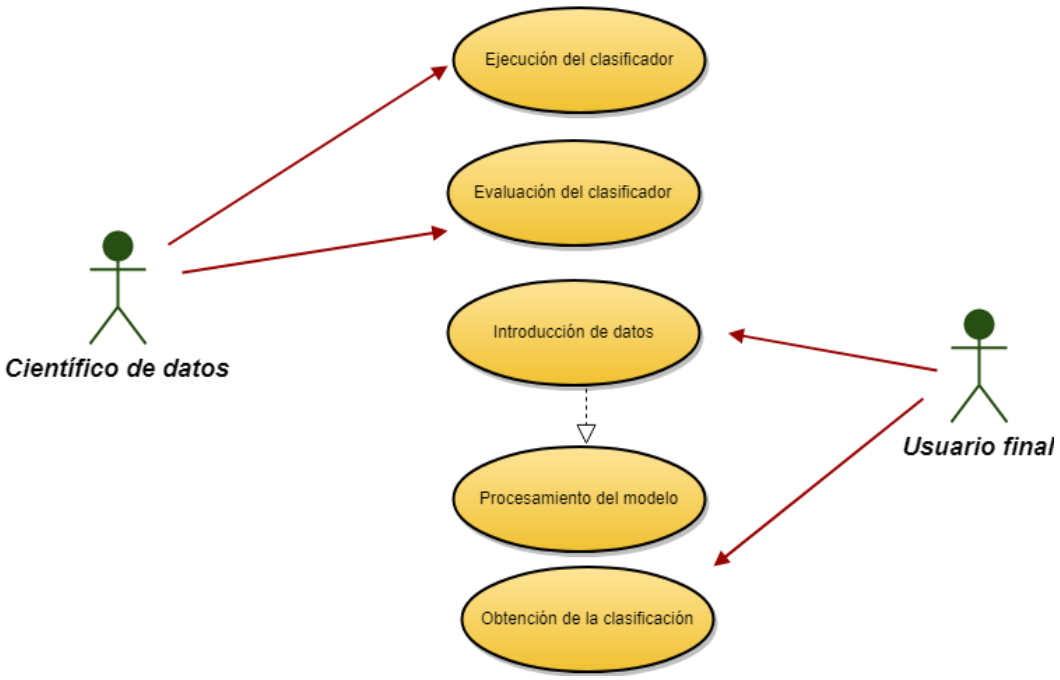

Figura 20. Diagrama de casos de uso.

En la Figura 21 se presenta el diagrama de secuencia UML del software. El usuario final inicia el proceso, al ejecutar la interfaz del software (Interfaz3.py). Enseguida, el software solicita que el usuario ingrese los valores de latitud y longitud a clasificar. Luego de ingresar los datos, el software evalúa los valores ingresados con el modelo establecido, mediante el algoritmo de clasificación. Acto seguido, el software muestra en un cuadro de diálogo la clasificación correspondiente a los valores ingresados. El usuario decide si continúa ingresando valores de latitud y longitud o si se sale del programa.

En la Figura 22 se presenta el prototipo de la interfaz de usuario para la clasificación de eventos. La interfaz muestra dos cuadros de texto en los que el usuario ingresa latitud y longitud. Para el ingreso de los datos, el software valida que los datos

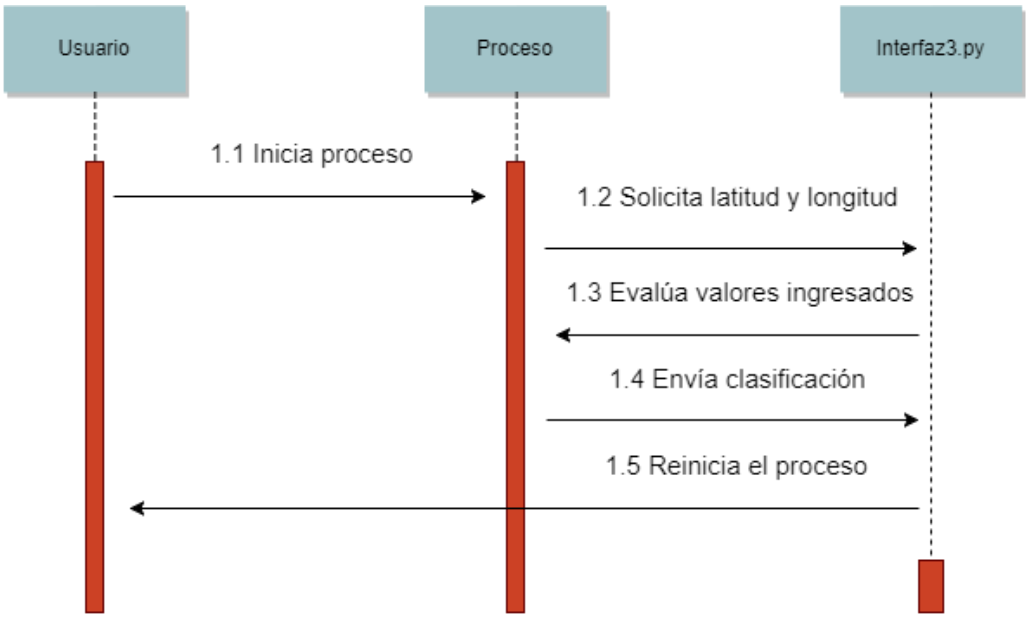

Figura 21. Diagrama de secuencias.

sean valores numéricos y que correspondan a Iraq para realizar la clasificación. Luego de que el usuario ingresa los valores de latitud y longitud, el software presenta el resultado de la clasificación, ya se de refugiados o de lucha con artillería tal como se muestra en las Figuras 23 y 24.

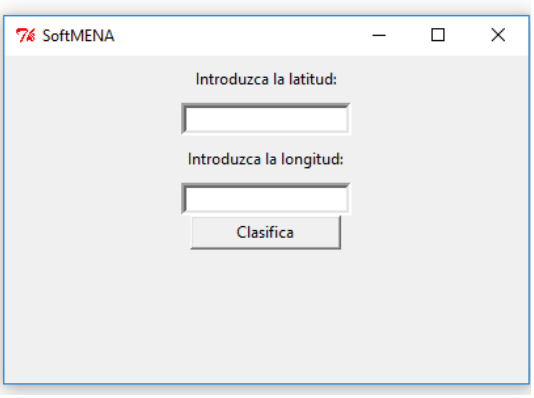

Figura 22. Interfaz de usuario.

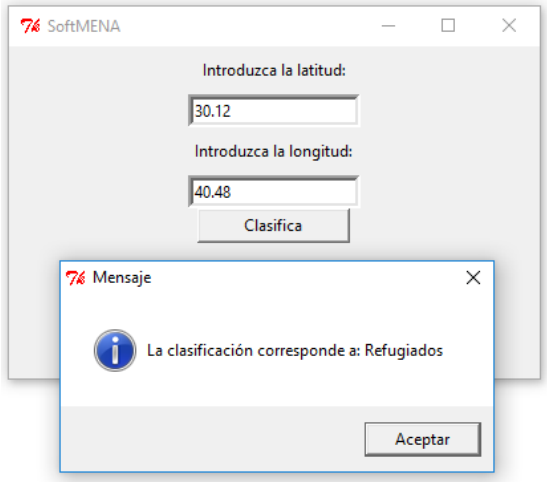

Figura 23. Resultado de la clasificación de refugiados.

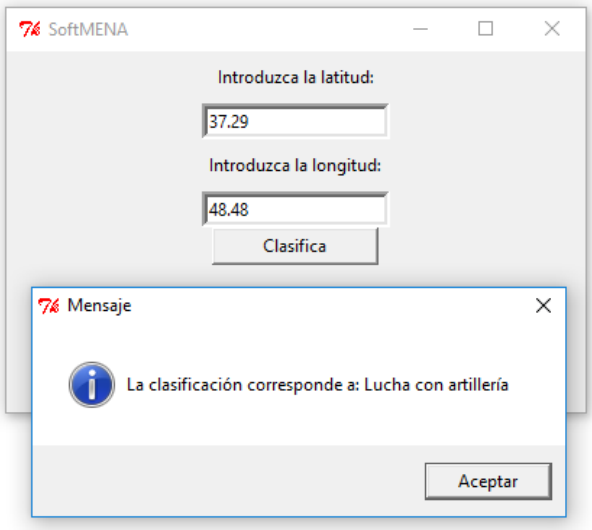

Figura 24. Resultado de la clasificación del evento lucha con artillería.

Retroalimentación

Los modelos predictivos generados por los algoritmos de clasificación se analizaron y validaron por la investigadora y el director del proyecto, doctor Harvey Alférez. Se realizaron varias iteraciones en la realización de los experimentos, con el fin de mejorar los modelos predictivos generados y seleccionar el modelo clasificador óptimo para ser ejecutado por el software. Es importante mencionar que, en 2016, la autora de este documento y el doctor Alférez se reunieron en el campus de la Universidad de Montemorelos con el pastor Rick McEdwards, presidente de la MENA, para hablar acerca del alcance de este proyecto. Aunque el pastor McEdwards quedó altamente interesado en el proyecto, no ha sido posible mantener una comunicación constante con él acerca de los avances realizados. Al terminar el presente proyecto, se enviarán los resultados obtenidos a los administradores de la MENA.

CAPÍTULO IV

PRESENTACIÓN Y ANÁLISIS DE RESULTADOS

Este capítulo presenta los resultados obtenidos con la metodología de ciencia de datos de IBM y al aplicar un conjunto de algoritmos de clasificación mediante aprendizaje automático a los datos de Iraq recuperados de GDELT mediante Big Query. Los experimentos realizados se usaron para encontrar el modelo predictivo más adecuado para clasificar eventos de refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres en el país de Iraq.

Infraestructura de cómputo para la construcción y evaluación de los modelos

Los experimentos se realizaron en un equipo portátil Lenovo con las siguientes características: procesador AMD A8-7410 APU con AMD Radeon R5 Graphics, Memoria RAM de 8.00 GB, y sistema operativo Windows 10 de 64 bits. Asimismo, se utilizó la versión 2.7.12 de Python, junto con la versión 4.2.0 de Anaconda, la cual contiene Pandas y Numpy.

El conjunto de datos de Iraq utilizado en los experimentos es un archivo que incluye el total de los datos obtenidos en la consulta con BigQuery sobre refugiados, ayuda humanitaria, protestas con violencia, lucha con artillería y masacres (42,027 instancias). Este conjunto de datos se puede descargar del siguiente link para su manejo https://drive.google.com/drive/folders/0B6LEG8jNfAY9MEJBTkZoYlRIZHc.

Descripción de los experimentos

Los experimentos consistieron en aplicar los algoritmos de KNN, Decision Trees, Näive Bayes y Logistic Regression al conjunto de datos obtenidos de Iraq, con el fin de elegir el óptimo para ser utilizado por el software clasificador de eventos. Estos cuatro clasificadores se eligieron, pues la literatura científica reconoce que son muy útiles en el aprendizaje automático (Wu et al., 2008) y también son insensibles a valores aislados o atípicos (Barabás y Kovács, 2009; Harrington, 2012; Iconiq Inc, 2016; Micenková, McWilliams y Assent, 2014).

Previo a la realización de las pruebas con los eventos, se realizó un cálculo de todas las posibles combinaciones de eventos disponibles con base en la Ecuación 4:

$$C(n,r) = {}^{n}C_r = {}_{n}C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} \qquad (4)$$

En esta ecuación, ***n*** es el número de entidades que se pueden elegir; en este caso, fueron los eventos (refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres); y ***r*** es la forma en que se pueden elegir estos eventos. Al aplicar la fórmula se obtuvieron 26 combinaciones de eventos, lo que equivale a 26 conjuntos de datos. Con estos conjuntos de datos se realizaron entrenamientos con cada algoritmo, llegando a un total de 104 pruebas (4 algoritmos de clasificación multiplicado por 26 conjuntos de datos). El total de experimentos realizados se muestra en el Apéndice B, al final de este documento y los archivos que incluyen el conjunto de

datos, están disponibles para su consulta y manejo en la siguiente liga: https://drive.google.com/drive/folders/0B6LEG8jNfAY9MEJBTkZoYlRIZHc. En las Tablas 7-9 se muestran las combinaciones obtenidas de los eventos y el número de instancias por combinación:

Tabla 7

Combinaciones con dos eventos

Evento	No. instancias
73, 145	13,482
73, 194	23,661
73, 202	12,236
73, REF	23,890
145, 194	16,315
145, 202	4,890
145, REF	16,544
194, 202	15,069
194, REF	26,723
202, REF	15,298

Tabla 8

Combinaciones con tres eventos

Evento	No. instancias
73, 145, 194	26,729
73, 145, 202	15,304
73, 145, REF	26,958
73, 194, 202	25,483
73, 194, REF	37,137
73, 202, REF	25,712
145, 194, 202	18,138
145, 194, REF	29,791
145, 202, REF	18,366
194, 202, REF	28,545

Tabla 9

Combinaciones con cuatro eventos

Evento	No. instancias
73, 145, 194, 202	28,551
73, 145, 194, REF	40,205
73, 145, 202, REF	28,780
73, 194, 202, REF	38,959
145, 194, 202, REF	31,613

En la combinación con los eventos 73, 145, 194, 202 y REF, se obtuvieron un total de 42, 027 instancias, para ser utilizadas mediante los algoritmos de clasificación. Los experimentos realizados se presentan a continuación:

1. En el primer experimento, se utilizó el algoritmo KNN, usando los conjuntos de datos con 2, 3, 4 y 5 eventos.

2. En el segundo experimento, se utilizó el algoritmo de Decision Trees, usando los conjuntos de datos correspondientes a 2, 3, 4 y 5 eventos.

3. En el tercer experimento, se utilizó el algoritmo de Näive Bayes, con los conjuntos de datos correspondientes a 2, 3, 4 y 5 eventos.

4. En el cuarto experimento, se utilizó el algoritmo de Logistic Regression, con los conjuntos de datos correspondientes a 2, 3, 4 y 5 eventos.

En la Figura 25 se muestra de manera gráfica el orden utilizado en estos experimentos.

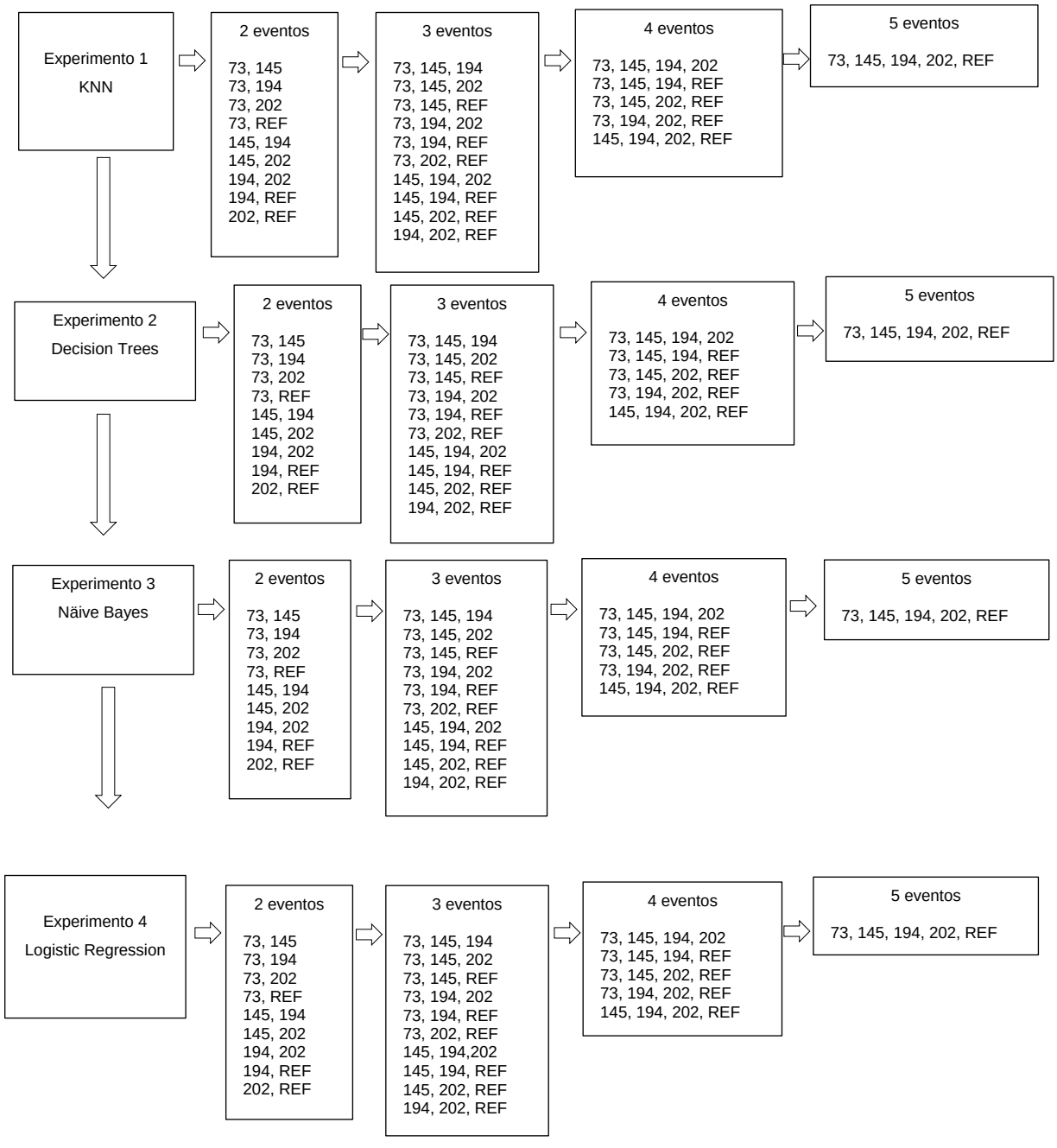

Figura 25. Secuencia de experimentos realizados.

Resultados

Los resultados obtenidos en los experimentos que se realizaron con los algoritmos KNN, Näive Bayes, Decision Trees y Logistic Regression se evaluaron en términos de accuracy, precision, recall y f1-score. Con respecto al accuracy, el valor esperado para aprobar el modelo de clasificación debía ser lo más cercano al valor 1 para obtener la mayor exactitud en la clasificación (Scikit Learn, 2017a). Para los valores de precision y recall también se consideraron apropiados los valores altos, cercanos a 1, para obtener una clasificación correcta. Un valor bajo en precision indica un número alto de falsos positivos (Brownlee, 2016). A su vez, un valor bajo en recall muestra un número alto de falsos negativos (Hackeling, 2014). El total de los experimentos realizados con los algoritmos están disponibles en el Apéndice B.

Para realizar los experimentos, el total de datos contenidos en cada conjunto de datos se dividió en 70% en la realización del entrenamiento y 30% en la realización de las pruebas. Los experimentos realizados con los eventos de refugiados y lucha con artillería obtuvieron los mejores resultados en términos de accuracy, precision, recall y f1-score. Para facilitar la labor del lector, a continuación, se describen los resultados de los experimentos con dos eventos que obtuvieron los mejores puntajes. Los demás resultados se encuentran en el Apéndice B.

Experimento 1 – algoritmo KNN

Al usar KNN, el mejor resultado se obtuvo con dos eventos o clases para el conjunto de datos de refugiados (evento 0, con 4,105 instancias) y lucha con artillería (evento 194, con 3,912 instancias) con un total de 8,017 instancias. La Tabla 10 describe el promedio de los resultados con respecto a precision, recall y f1-score.

Tabla 10

Resultados del algoritmo KNN

Evento	Precision	Recall	F1-Score
0	0.75	0.75	0.75
194	0.74	0.74	0.74

En este experimento se obtuvo una accuracy de 0.7545, que es alto y los resultados de las medidas de precision, recall y f1-score fueron altos. No obstante, se realizaron experimentos con los otros tres algoritmos de clasificación para observar y comparar los resultados obtenidos y así seleccionar el mejor clasificador.

Experimento 2 – algoritmo Decision Trees

En el segundo experimento, se aplicó el algoritmo de Decision Trees. El mejor resultado correspondió a la clasificación del evento de refugiados (evento 0, con 4,105 instancias) y al evento de lucha con artillería (evento 194, con 3,912 instancias). Se obtuvo una accuracy de 0.7629 y los valores en precision, recall y f1-score fueron entre 0.74 y 0.76, tal como se muestran en la Tabla 11.

Tabla 11

Resultados del algoritmo Decision Trees

Evento	Precision	Recall	F1-Score
0	0.76	0.76	0.76
194	0.74	0.75	0.75

Es notable que en este experimento se encuentra una similitud de resultados al compararlo con el experimento 1, debido a que el experimento 2 brinda un resultado favorable en términos de accuracy, precision, recall y f1-score. No obstante, se hicieron experimentos con los otros dos clasificadores.

Experimento 3 – algoritmo Näive Bayes

En el tercer experimento, se utilizó el algoritmo de Näive Bayes. El mejor resultado se obtuvo al aplicar este clasificador al evento de ayuda humanitaria (evento 73, con 3,124 instancias) y al evento de masacres (evento 202, con 513 instancias). Para estos eventos, la accuracy dio un valor alto de 0.8510. No obstante, los valores de precision, recall y f1-score en el evento de masacres fue de 0, tal como se describe en la Tabla 12.

Tabla 12

Resultados del algoritmo Näive Bayes

Evento	Precision	Recall	F1-Score
73	0.85	1.00	0.92
202	0.00	0.00	0.00

Por otra parte, en los experimentos realizados con este algoritmo para el evento de refugiados (evento 0, con 4,116 instancias) y en el evento protestas violentas (evento 145, con 915 instancias) se obtuvo una accuracy de 0.8145. No obstante, la precision en el caso del evento de protestas violentas fue de 0, tal como se muestra

en la Tabla 13. Esto demuestra que un valor alto en accuracy no es suficiente para la selección del algoritmo, así como se describe en la literatura (Brownlee, J., 2014). Se realizaron más experimentos con el algoritmo de Logistic Regression para observar el comportamiento de las métricas.

Tabla 13

Resultados del algoritmo Näive Bayes-protestas violentas y refugiados

Evento	Precision	Recall	F1-Score
0	0.82	1.00	0.90
145	0.00	0.00	0.00

Experimento 4 – algoritmo Logistic Regression

En el cuarto experimento, se utilizó el algoritmo de Logistic Regression. En este experimento, se obtuvieron resultados similares a los obtenidos mediante Näive Bayes, donde el evento de protestas violentas (evento 145, con 915 instancias) y masacres (evento 202, con 513 instancias) obtuvieron valores altos en accuracy, 0.8518. No obstante, el evento de masacres obtuvo valores bajos en precision, recall y f1-score (ver Tabla 14).

Asimismo, la Tabla 15 muestra los resultados para el evento de lucha con artillería (evento 194, con 4,022 instancias) y para el evento de masacres (evento 202, con 499 instancias). Con estos eventos se obtuvo una accuracy de 0.8896, pero el evento de masacres obtuvo el valor de 0 en precisión, recall y f1-score, por lo que no es un modelo óptimo para clasificación.

Tabla 14

Resultados del algoritmo Logistic Regression–ayuda humanitaria y masacres

Evento	Precision	Recall	F1-Score
73	0.86	1.00	0.92
202	0.00	0.00	0.00

Tabla 15

Resultados del algoritmo Logistic Regression–lucha con artillería y masacres

Evento	Precision	Recall	F1-Score
194	0.89	1.00	0.94
202	0.00	0.00	0.00

Discusión

De acuerdo con los experimentos presentados en este capítulo, sobresale el algoritmo Decision Trees para la clasificación de los eventos de refugiados (evento 0) y lucha con artillería (evento 194). Mediante la aplicación de este algoritmo, se obtuvo una accuracy de 0.7629. El evento de refugiados es clasificado con una precision de 0.76, un recall de 0.76 y un f1-score de 0.76. El evento de lucha con artillería (evento 194) es clasificado con una precision de 0.74, un recall de 0.75 y f1-score de 0.75. Por estos resultados, el modelo predictivo creado mediante Decision Trees fue seleccionado para la clasificación mediante el software. El código del software está disponible en el Apéndice C y se puede descargar en la siguiente liga: https://drive.google.com/drive/folders/0B6LEG8jNfAY9UklON0R3N0s1OFE.

Evaluación mediante geolocalización

Para demostrar que el modelo obtenido mediante el algoritmo Decision Trees es válido, además de las pruebas de validación cruzada, los resultados de clasificación con el modelo se compararon con resultados de geolocalización obtenidos en mapas de organismos oficiales.

En primer lugar, en la Figura 26, se muestra la distribución de zonas de refugiados (zonas en azul) del mapa proporcionado por el Alto Comisionado de las Naciones Unidas (UNHCR) para Iraq.

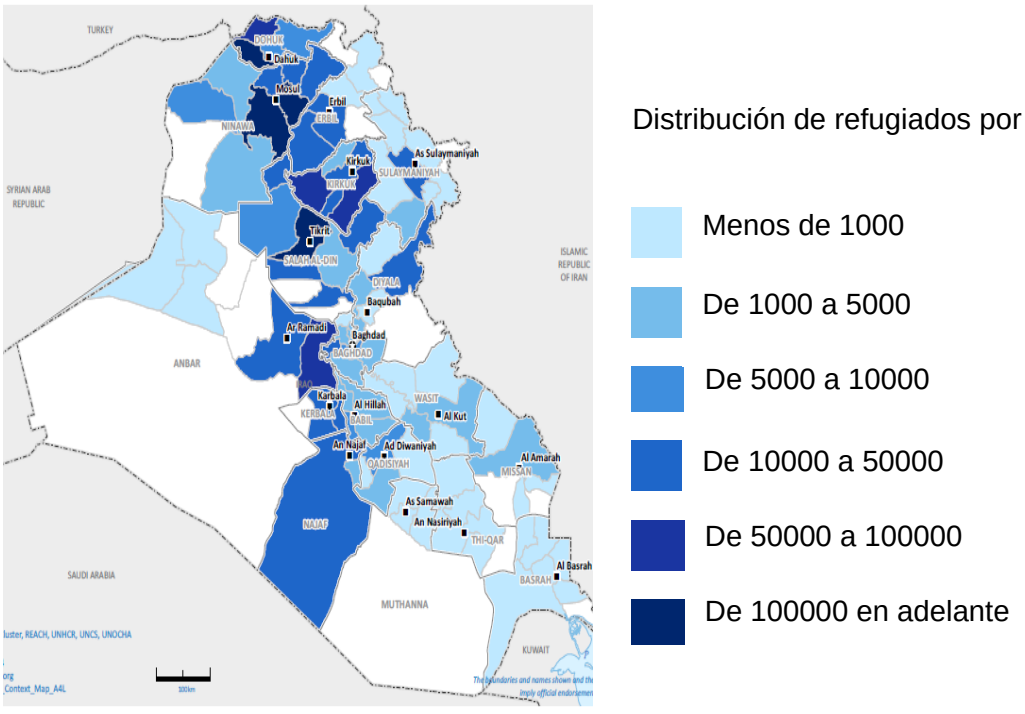

Figura 26. Mapa de refugiados (United Nations High Commissioner for Refugees, 2017).

Por otra parte, la Figura 27 muestra un mapa de los puntos seleccionados en la Figura 26 de las zonas clasificadas con mayor distribución de refugiados por distrito. Las zonas o puntos para su evaluación corresponden a: Duhok, Hilla, Suleimaniya, Saladino, Tal Afar, Ninaua, Kerbala y Erbil. Este mapa está disponible en la siguiente liga: https://drive.google.com/open?id=1Rj5HHvPL1HkO9Fm_Ee7vSrDzpVo&usp=sharing

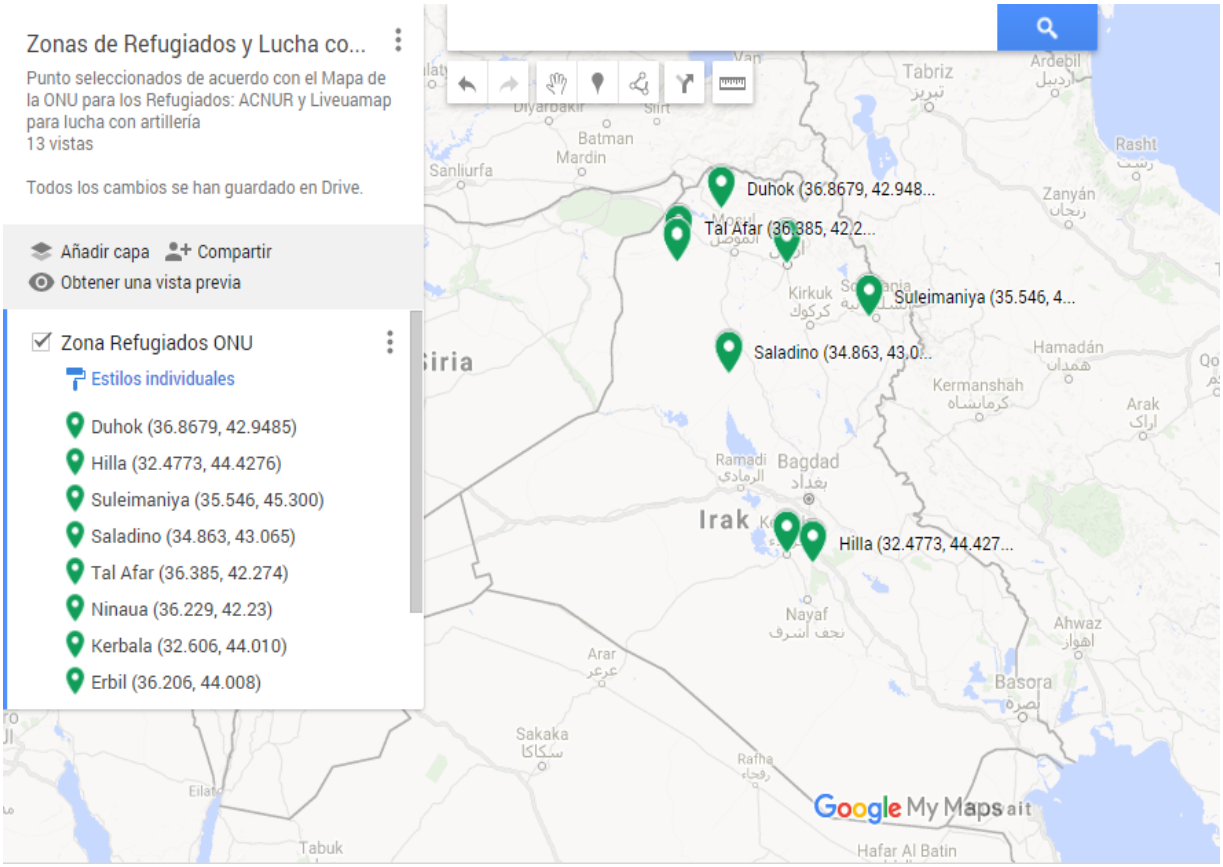

Figura 27. Zonas de refugiados con latitud y longitud con los datos del mapa UNHCR, 2017.

Con el fin de verificar si los resultados obtenidos con el software son similares a los presentados en los mapas de la Figura 26 y 27, se realizaron ocho experimentos usando la latitud y la longitud. Los resultados de estos experimentos se muestran en las Figuras 28-35. En estos experimentos, el software clasificó correctamente las zonas de refugiados al ser comparadas con el mapa oficial del Alto Comisionado de las Naciones Unidas (UNHCR) de la Figura 26. Esto confirma que el modelo obtenido con el algoritmo de Decision Trees es válido para la clasificación del evento de refugiados.

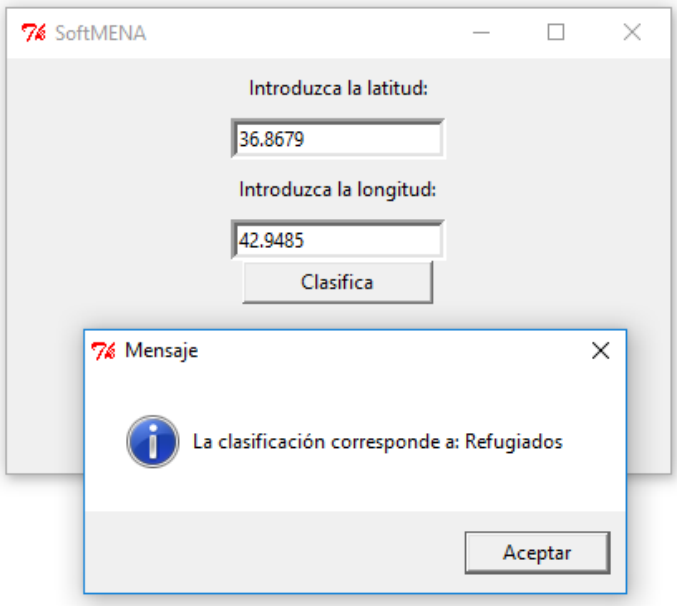

Figura 28. Clasificación Duhok, Iraq.

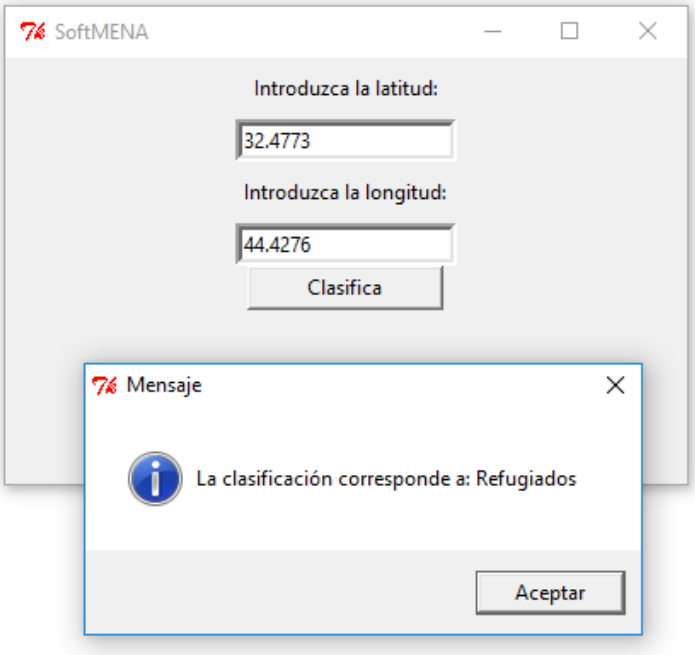

Figura 29. Clasificación Hila, Iraq.

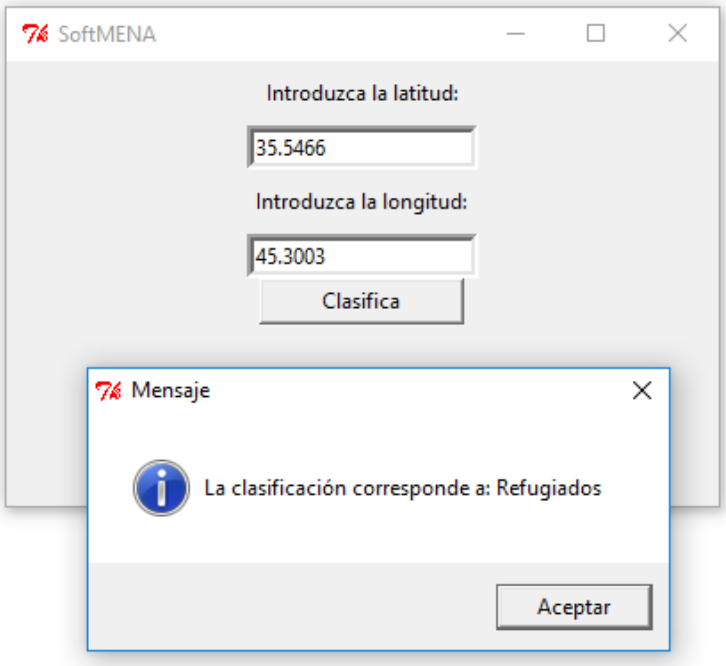

Figura 30. Clasificación Suleimaniya, Iraq.

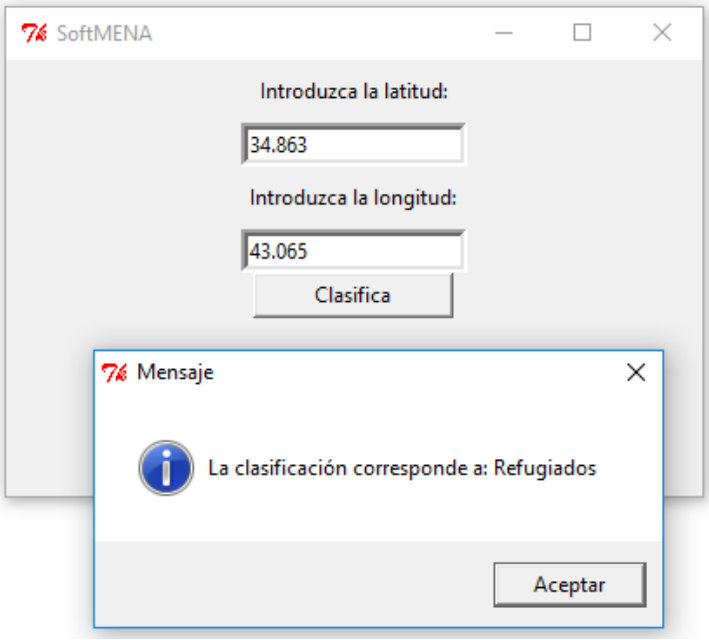

Figura 31. Clasificación Saladino, Iraq.

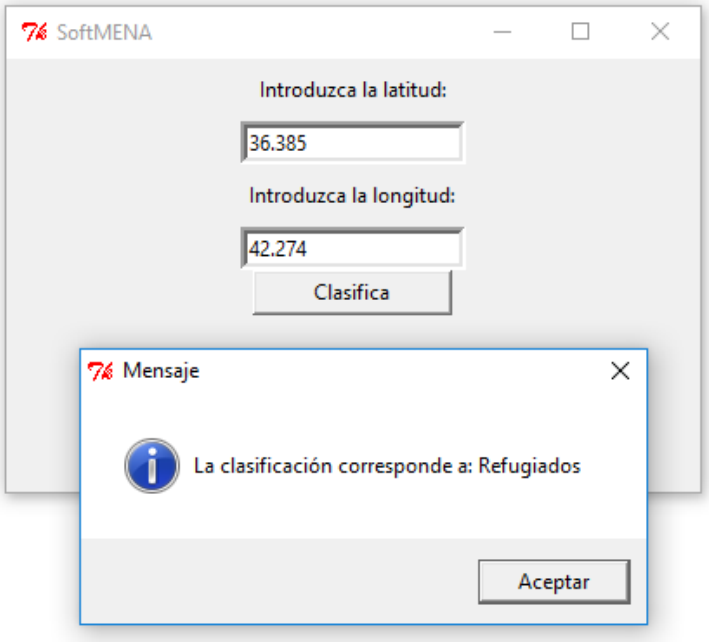

Figura 32. Clasificación Tal Afar, Iraq.

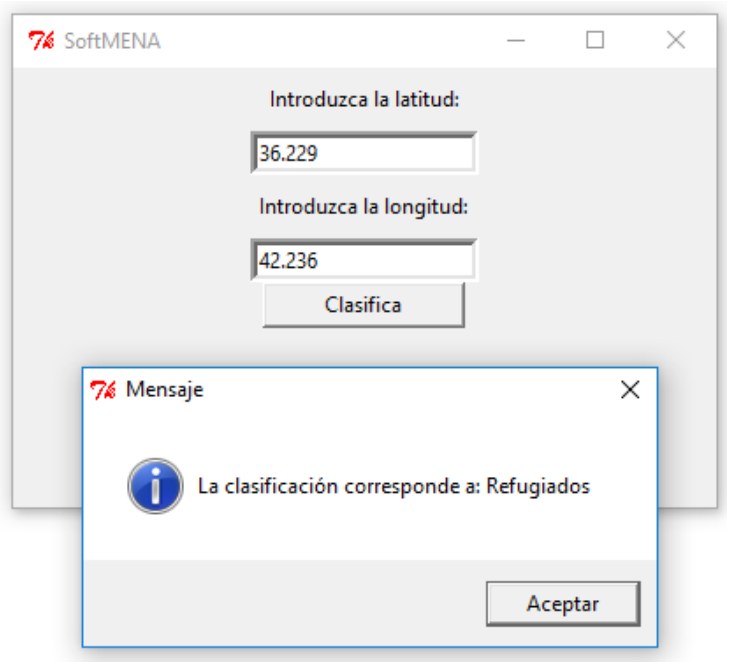

Figura 33. Clasificación Ninaua, Iraq.

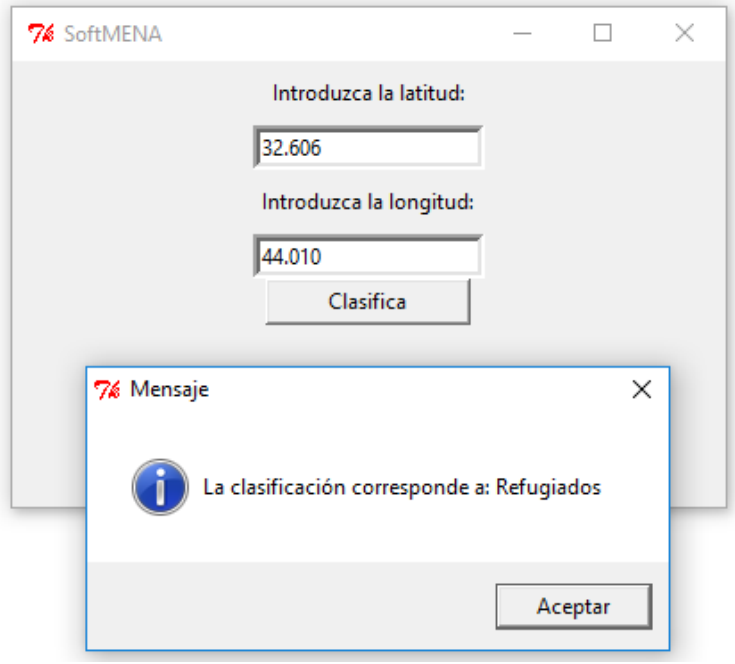

Figura 34. Clasificación Kerbala, Iraq.

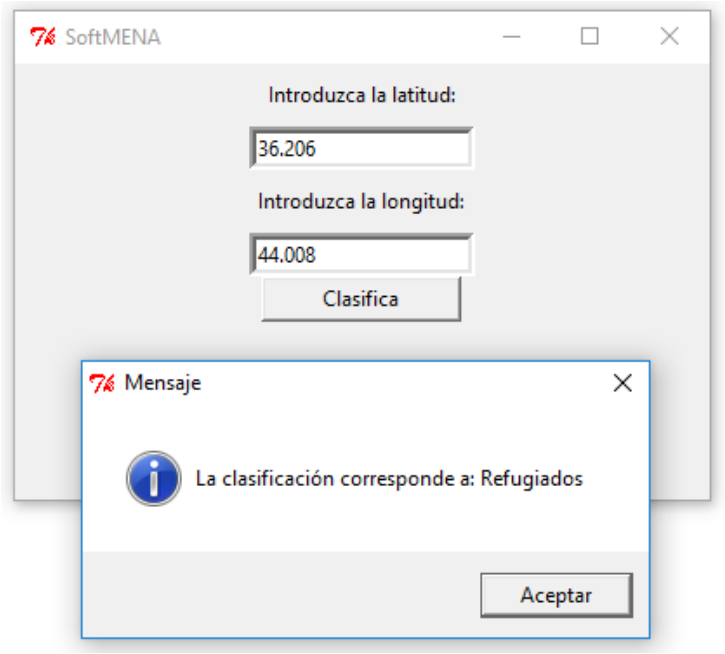

Figura 35. Clasificación Erbil, Iraq.

Por otra parte, en la Figura 36 se muestra un mapa proporcionado por Live Universal Awareness Map, que es un sitio de noticias interactivas por medio de mapas (Live Universal Awareness Map, 2014). Esta figura muestra la distribución de Iraq con respecto al evento de lucha con artillería. Los iconos muestran las noticias más sobresalientes. El color amarillo en las zonas representa las noticias recientes del día en el que se consultó el sitio y el color rojo las noticias de fechas anteriores, es necesario aclarar que este mapa se actualiza diariamente, siendo probable que al momento de su consulta no se observe de la misma manera las zonas y los iconos, como está en la figura de este documento.

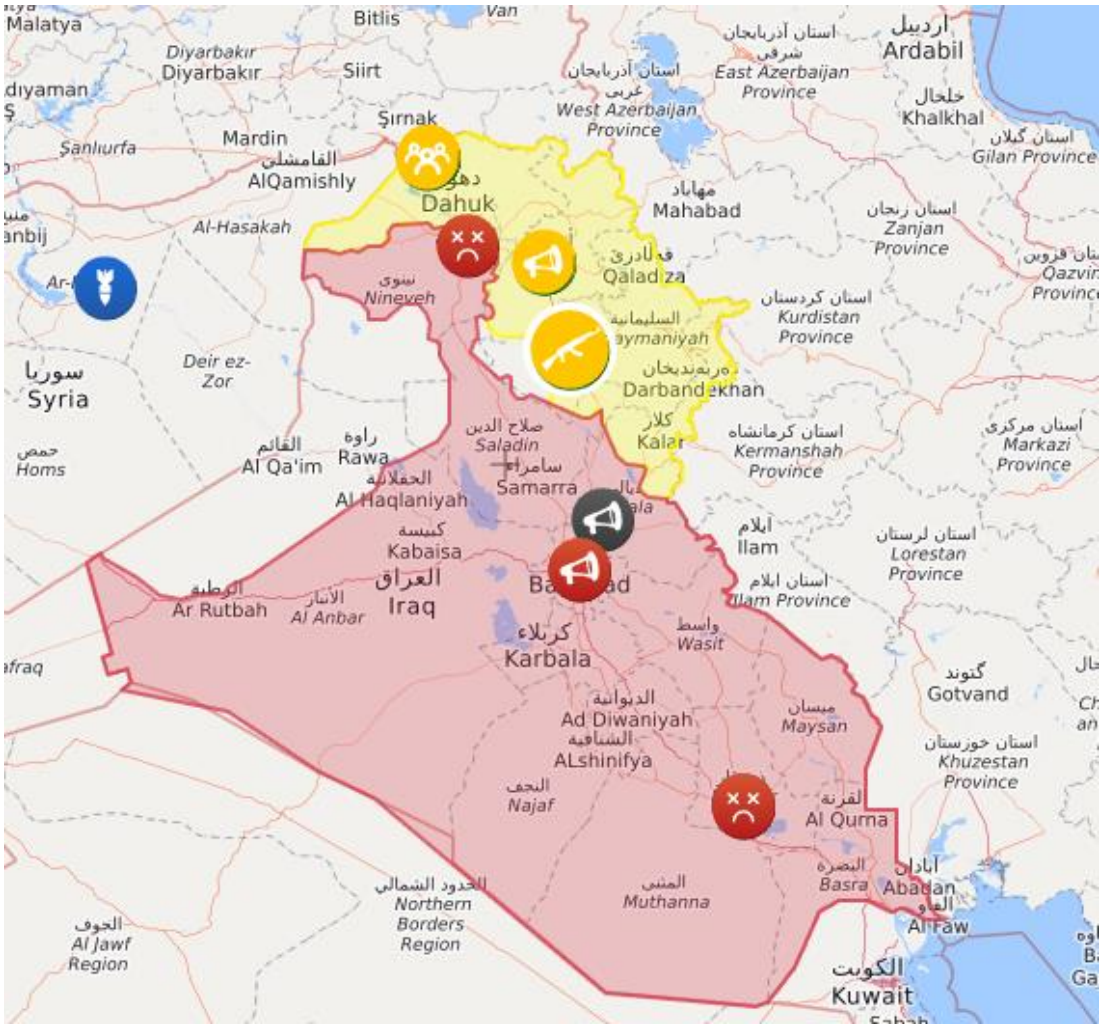

Figura 36. Mapa de lucha con artillería (Live Universal Awareness Map, 2014).

La Figura 37 muestra el mapa de algunas de las zonas clasificadas como lucha con artillería, basadas en el sitio de noticias Live Universal Awareness Map. Para la evaluación de estas zonas, se tomaron los valores de latitud y longitud basados en las zonas con noticias destacadas, que corresponden a: Ramadi, Bagdad, Kirkuk, Al-Hawija, Frontera Al-Kaim, Ambar y Rawa.

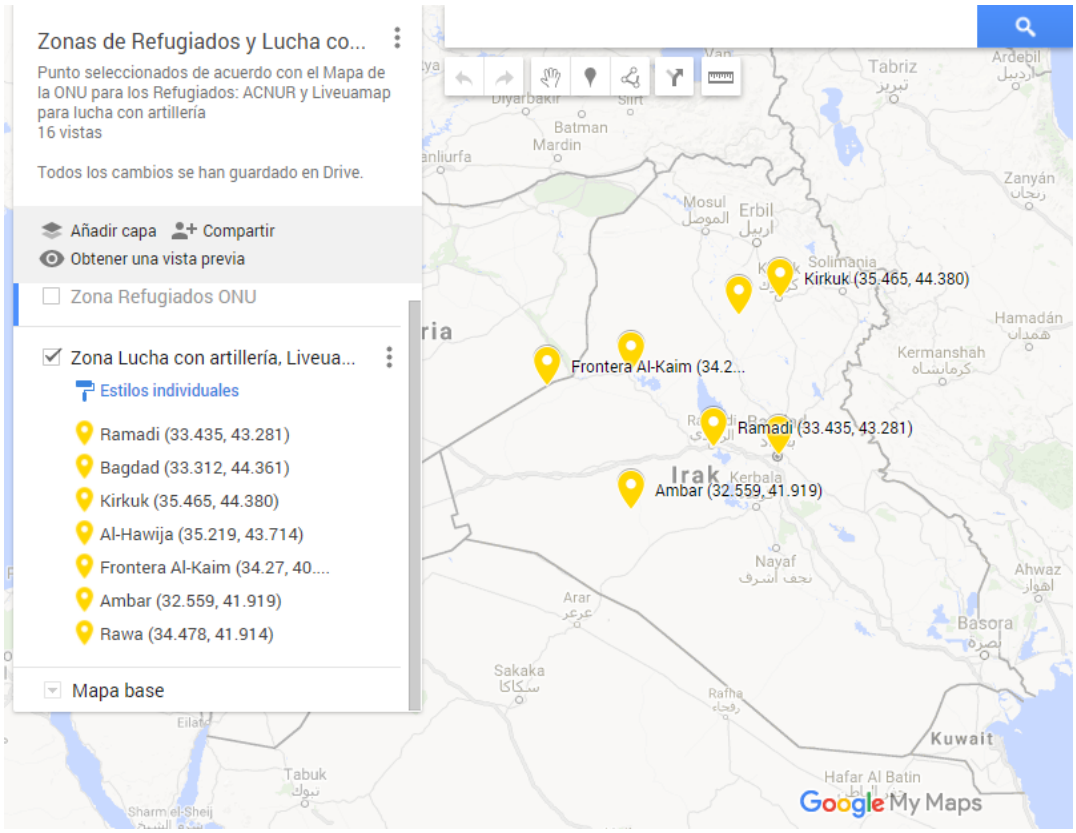

Figura 37. Mapa de lucha con artillería con latitud y longitud con los datos de Liveumap, 2014.

Estas zonas del mapa, relacionadas al evento lucha con artillería, fueron evaluadas con el software para comprobar que se obtiene la misma clasificación presentada en el mapa oficial de Live Universal Awareness Map. Los resultados de estos experimentos se observan en las Figuras 38 al 44. Estos resultados muestran que el software clasifica correctamente las zonas de lucha con artillería, al ser comparadas con el mapa proporcionado por Live Universal Awareness Map. El mapa está disponible para su consulta e interacción en la siguiente liga: https://drive.google.com/open?id=1Rj5HHvPL1HkO9Fm_Ee7vSrDzpVo&usp=sharing

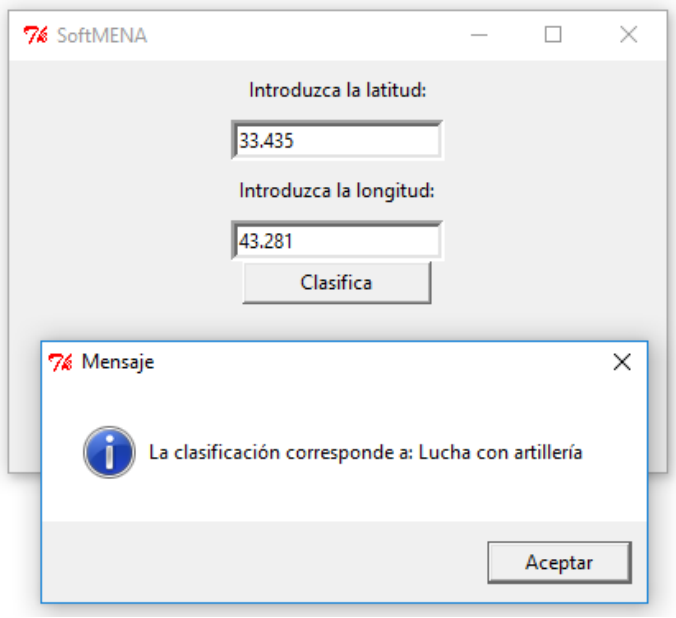

Figura 38. Clasificación Ramadi, Iraq.

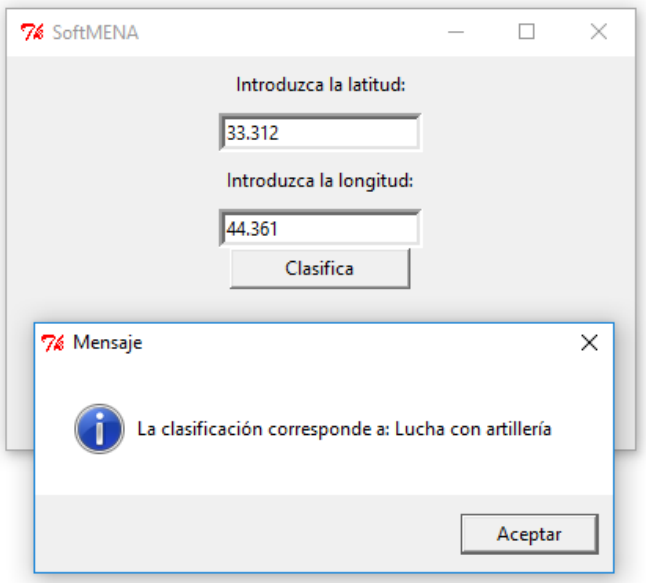

Figura 39. Clasificación Baghdad, Iraq.

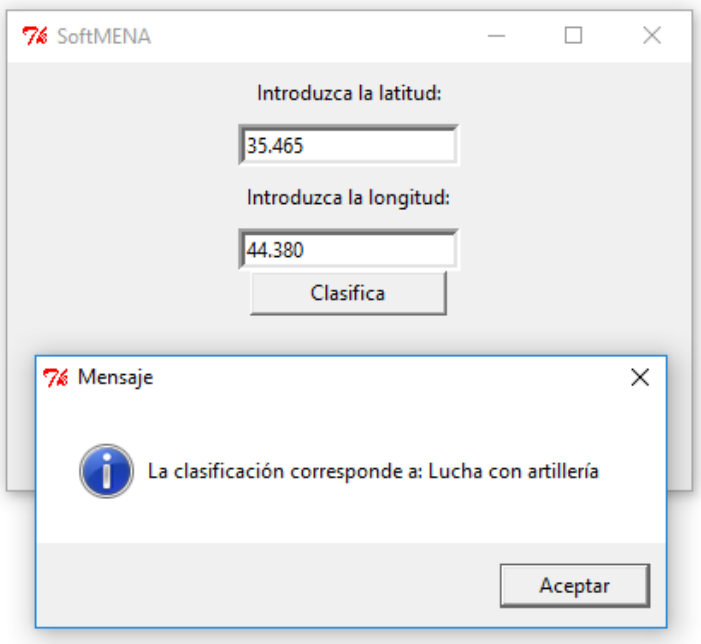

Figura 40. Clasificación Kirkuk, Iraq.

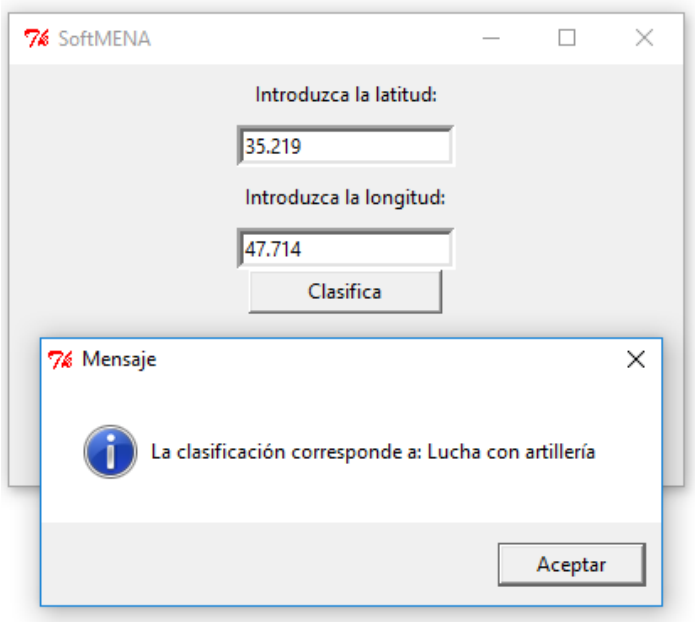

Figura 41. Clasificación Al-Hawija, Iraq.

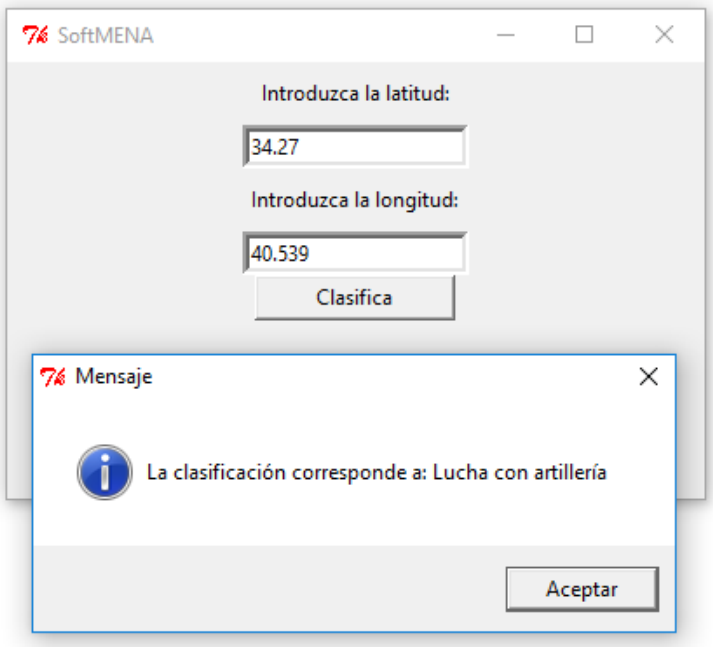

Figura 42. Clasificación Frontera Al-Kaim, Iraq.

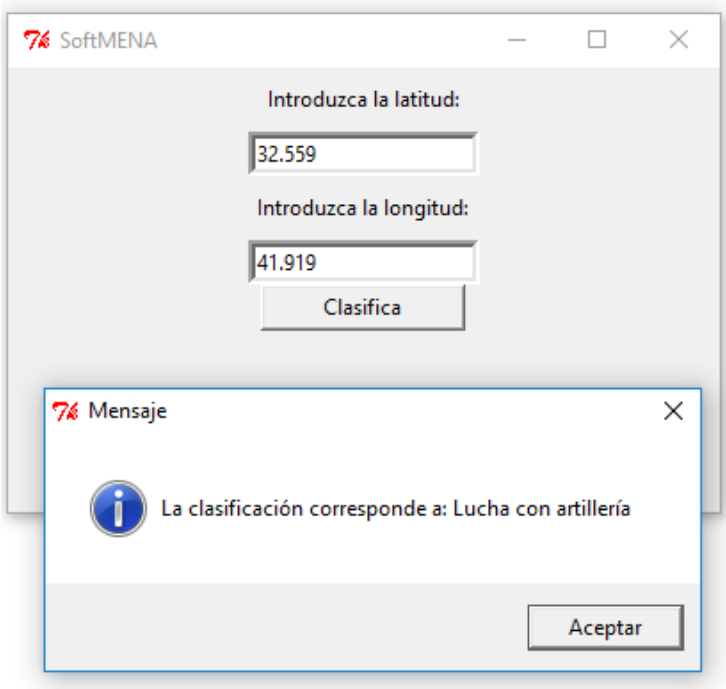

Figura 43. Clasificación Ambar, Iraq.

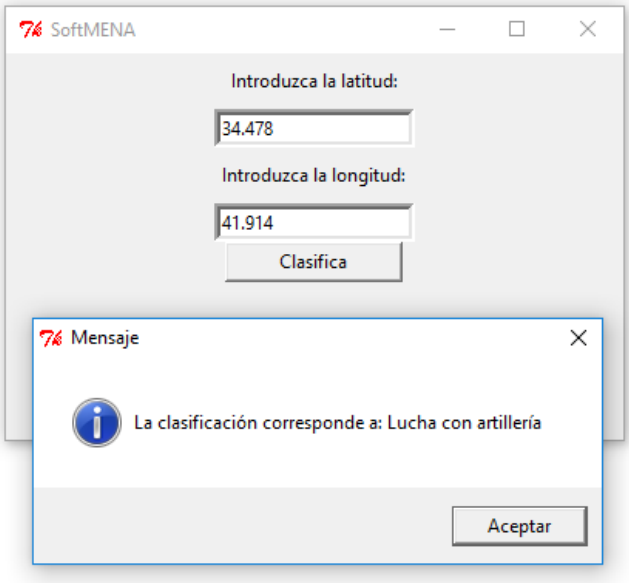

Figura 44. Clasificación Rawa, Iraq.

Experimentos en zonas sin datos disponibles

Los experimentos presentados en las secciones anteriores fueron realizados en zonas con clasificaciones disponibles para los eventos de refugiados y lucha con artillería. En esta sección, se presentan los resultados para zonas en Iraq sin datos disponibles para ser clasificadas con el software. Los puntos fueron seleccionados al azar en zonas sin datos disponibles para la clasificación. Los valores se tomaron del mapa proporcionado por el Alto Comisionado de las Naciones Unidas (UNHCR) y del mapa obtenido del sitio de noticias Live Universal Awareness Map. La Figura 45 muestra el mapa de los puntos clasificados por el software usando el algoritmo de Decision Trees. Los puntos verdes se refieren a la clasificación del evento de refugiados y los puntos amarillos a la clasificación para el evento de lucha con artillería. Este mapa está disponible para su consulta y análisis en la siguiente liga: https://drive.google.com/open?id=1FQbN1K_KF9_xNXS8HhftOiLCpTk&usp=sharing

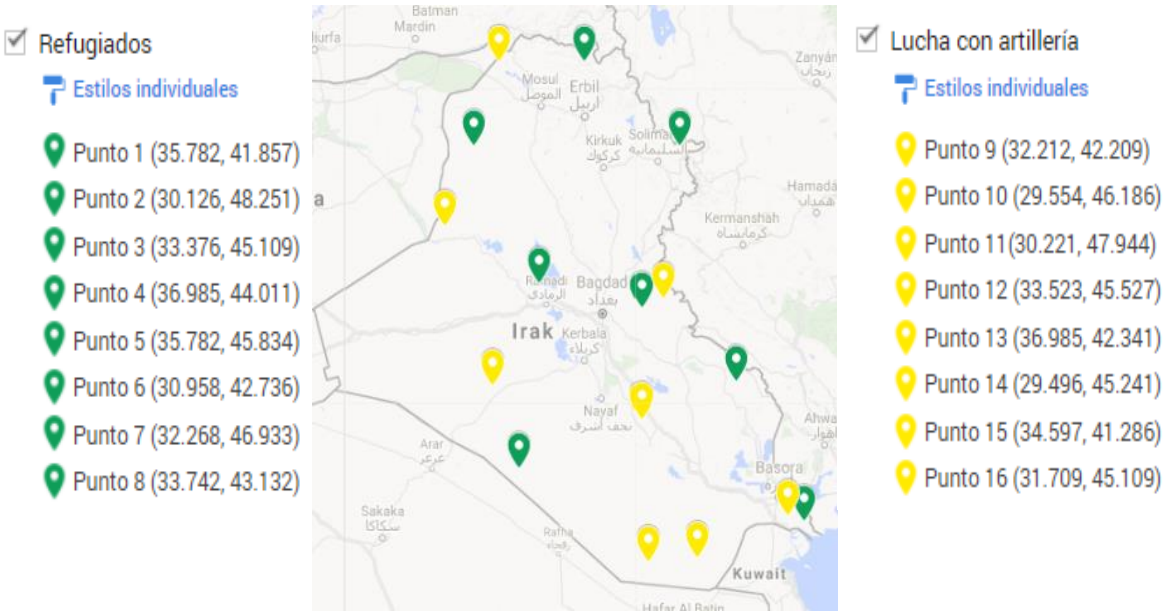

Figura 45. Puntos sin datos disponibles clasificados con el software.

CAPÍTULO V

DISCUSIÓN, CONCLUSIONES Y TRABAJO FUTURO

Discusión

Se realizaron 26 experimentos por algoritmo KNN, Decision Trees, Näive Bayes y Logistic Regression, haciendo un total de 104 experimentos. En estos experimentos, se observaron los resultados de validación cruzada en términos de accuracy, precision, recall y f1-score para los eventos de refugiados, ayuda humanitaria, protestas violentas, lucha con artillería y masacres.

El mejor clasificador para el país de Iraq es el basado en el algoritmo Decision Trees con los eventos de refugiados y lucha con artillería. El modelo generado con este clasificador es utilizado en un software que recibe la longitud y la latitud para clasificar la necesidad de los habitantes de alguna zona específica. Con base en el mapa proporcionado por el Alto Comisionado de las Naciones Unidas (UNHCR), para refugiados y en el mapa del sitio de noticias de Live Universal Awareness Map, para lucha con artillería, se seleccionaron las zonas ya clasificadas y se compararon con los resultados arrojados por el software. Los resultados presentaron la misma clasificación que la que se muestra en los mapas oficiales. Además, basados en los mapas oficiales, se tomaron valores de zonas sin clasificación, en los que el software proporcionó una clasificación con respecto a los eventos de refugiados y lucha con artillería. Los resultados obtenidos con el modelo predictivo de Decision Trees se

consideran confiables para los eventos de refugiados y lucha con artillería, debido a que la accuracy corresponde a 0.7629, donde el evento de refugiados obtuvo una precision de 0.76, un recall de 0.76 y un f1-score de 0.76 y el evento de lucha con artillería obtuvo una precision de 0.74, un recall de 0.75 y f1-score de 0.75.

Conclusiones

La Ventana 10/40 es la zona con menos cristianos en el mundo. La Iglesia Adventista del Séptimo Día no ha logrado una evangelización directa en los países de esta zona por diversos factores, principalmente por la diversidad de culturas y creencias, que son muy distintas a las de los cristianos de occidente. Es por ello que, como Iglesia Adventista del Séptimo Día y División de Medio Oriente y Norte de África, es necesario optar por técnicas novedosas que sirvan para conocer las necesidades de las personas, para luego acercase a ellas y cumplir así la misión encomendada de predicar el evangelio en todas las regiones del mundo.

Con el fin de descubrir las necesidades de las personas en la Ventana 10/40 y específicamente en Iraq, en esta investigación se construyó un software basado en los datos abiertos masivos de GDELT, a los cuales se les aplicó ciencia de datos y técnicas de aprendizaje automático, apoyados con la metodología de ciencia de datos de IBM.

El software construido utiliza un clasificador que usa el algoritmo de Decision Trees, con una accuracy de 0.7629. Al usar este clasificador, el evento de refugiados obtuvo un valor de precision de 0.76, un recall y un f1-score de 0.76. Para el evento de lucha con artillería, se obtuvo una precision de 0.74, un recall de 0.75 y un f1-score de 0.75. De esta manera, el software clasifica automáticamente las zonas de Iraq con

datos disponibles y no disponibles, al ingresar los valores de latitud y longitud. Los resultados fueron comparados con mapas oficiales. Uno de estos mapas fue tomado del Alto Comisionado de las Naciones Unidas (UNHCR) y muestra la distribución de refugiados en Iraq. El otro mapa fue proporcionado por el sitio de noticias Live Universal Awareness Map que muestra luchas con artillería en Iraq. Al comparar los resultados del clasificador con estos mapas, fueron positivos. El modelo es ejecutado cada vez que se introducen los valores de latitud y longitud en el software. Las contribuciones de esta investigación son las siguientes:

1. Se realizó un total de 104 experimentos con los algoritmos de KNN, Decision Trees, Näive Bayes y Logistic Regression. Dentro de estos experimentos, el mejor resultado se obtuvo con el algoritmo Decision Trees. Los resultados de este modelo se compararon con el mapa proporcionado por el Alto Comisionado de las Naciones Unidas (UNHCR), para el evento de refugiados y con el mapa proporcionado por el sitio de noticias Live Universal Awareness Map, para el evento de lucha con artillería. Los resultados con el software coincidieron en todos los casos con los mapas.

2. Se generó un modelo de clasificación para los eventos de refugiados y lucha con artillería en el país de Iraq que se puede utilizar aún en zonas en donde no existen datos disponibles.

3. Se creó un software que utiliza el modelo de clasificación de Decision Trees, que ofrece los mejores resultados para clasificar refugiados y lucha con artillería con respecto a validación cruzada: accuracy, precision, recall, y f1-score.

Trabajos futuros

Considerando que la ciencia de datos es relativamente nueva, a futuro se considera realizar lo siguiente:

1. Esta investigación se limitó a ciertos eventos del país de Iraq. Se espera incluir el análisis de nuevos eventos que sean de interés para los administradores de la MENA. Asimismo, se podrían incluir otros países que pertenezcan a esta zona.

2. La interfaz propuesta del software puede mejorarse, haciéndola dinámica a través de un mapa, el cual muestre al usuario la clasificación al dar clic sobre la zona o país de interés.

3. Podría ser de beneficio el análisis en tiempo real de los datos tomados de GDELT. Para este fin, se propone utilizar la librería Mlib de Apache Spark, que es una librería de aprendizaje automático para el procesamiento de datos masivos y utilizar la funcionalidad de streaming de Apache Spark, que facilita la construcción de aplicaciones escalables tolerantes a fallos que administran flujos de datos.

4. Extender el software de clasificación para que otras divisiones de la Iglesia Adventista del Séptimo Día puedan utilizarlo.

5. Presentar los resultados obtenidos en diferentes niveles de la administración de la Iglesia Adventista del Séptimo Día, con el fin de buscar nuevos escenarios en los cuáles aplicar ciencia de datos.

6. Publicar un artículo en una revista científica o de la Iglesia Adventista del Séptimo Día, en el que se muestre la importancia del uso de ciencia de datos para entender las necesidades de las personas y para que se expliquen los resultados obtenidos en esta investigación.

APÉNDICE A

CONSULTAS REALIZADAS EN BIG QUERY

CONSULTAS REALIZADAS EN BIG QUERY

Refugiados

```
SELECT Actor1Type1Code,Year, ActionGeo_CountryCode,
Actor1Geo_Lat,Actor1Geo_Long, EventCode

FROM
  [gdelt-bq:full.events]

WHERE Actor1Type1Code="REF"

AND (Year> 2011 AND Year < 2016)

AND (Actor1Geo_Lat > 29.12 AND Actor1Geo_Lat < 37.29)
AND (Actor1Geo_Long > 39.22 AND Actor1Geo_Long < 48.48)

AND Actor1Geo_Lat IS NOT NULL AND Actor1Geo_Long IS NOT NULL
```

Ayuda humanitaria

```
SELECT ActionGeo_CountryCode, Year,Actor1Geo_Lat,Actor1Geo_Long,
EventCode,

FROM
  [gdelt-bq:full.events]

WHERE EventCode="073"

AND (Year> 2011 AND Year < 2016)

AND (Actor1Geo_Lat > 29.12 AND Actor1Geo_Lat < 37.29)
AND (Actor1Geo_Long > 39.22 AND Actor1Geo_Long < 48.48)

AND Actor1Geo_Lat IS NOT NULL AND Actor1Geo_Long IS NOT NULL
```

Protestas violentas

```
SELECT ActionGeo_CountryCode, Year,Actor1Geo_Lat,Actor1Geo_Long,
EventCode,

FROM
  [gdelt-bq:full.events]

WHERE EventCode="145"

AND (Year> 2011 AND Year < 2016)

AND (Actor1Geo_Lat > 29.12 AND Actor1Geo_Lat < 37.29)
AND (Actor1Geo_Long > 39.22 AND Actor1Geo_Long < 48.48)

 AND Actor1Geo_Lat IS NOT NULL AND Actor1Geo_Long IS NOT NULL
```

Lucha con artillería

```
SELECT ActionGeo_CountryCode, Year,Actor1Geo_Lat,Actor1Geo_Long,
EventCode,

FROM
  [gdelt-bq:full.events]

WHERE EventCode="194"

AND (Year> 2011 AND Year < 2016)

AND (Actor1Geo_Lat > 29.12 AND Actor1Geo_Lat < 37.29)
AND (Actor1Geo_Long > 39.22 AND Actor1Geo_Long < 48.48)

AND Actor1Geo_Lat IS NOT NULL AND Actor1Geo_Long IS NOT NULL
```

Masacres

```
SELECT ActionGeo_CountryCode, Year,Actor1Geo_Lat,Actor1Geo_Long,
EventCode,

FROM
  [gdelt-bq:full.events]

WHERE EventCode="202"

AND (Year> 2011 AND Year < 2016)

AND (Actor1Geo_Lat > 29.12 AND Actor1Geo_Lat < 37.29)
AND (Actor1Geo_Long > 39.22 AND Actor1Geo_Long < 48.48)

AND Actor1Geo_Lat IS NOT NULL AND Actor1Geo_Long IS NOT NULL
```

APÉNDICE B

RESULTADOS DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

RESULTADOS DE ALGORITMOS DE
APRENDIZAJE AUTOMÁTICO

Este apéndice presenta el total de los experimentos realizados con los algoritmos KNN, Decision Trees, Näive Bayes y Logistic Regression, basándose en las combinaciones posibles para cada evento. Cada imagen muestra los resultados obtenidos en las métricas de evaluación: precision, recall y f1-score, avg/total y support indican el promedio obtenido de los resultados.

KNN -2 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.728749443703
[[ 73  73  73 ..., 145 145 145]
 [ 73  73  73 ..., 145 145 145]]
0.728306551298
             precision    recall  f1-score   support

         73       0.84      0.81      0.82      3140
        145       0.41      0.46      0.43       905

avg / total       0.74      0.73      0.73      4045
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.728749443703
[[ 73  73  73 ..., 145 145 145]
 [ 73  73  73 ..., 145 145 145]]
0.728306551298
             precision    recall  f1-score   support

         73       0.84      0.81      0.82      3140
        145       0.41      0.46      0.43       905

avg / total       0.74      0.73      0.73      4045
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.842186989212
[[ 73  73  73 ..., 202 202 202]
 [ 73  73  73 ...,  73  73  73]]
0.841187687279
             precision    recall  f1-score   support

         73       0.86      0.97      0.91      3122
        202       0.39      0.11      0.17       549

avg / total       0.79      0.84      0.80      3671
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.590288823776
[[73 73 73 ...,  0  0  0]
 [ 0  0  0 ...,  0  0 73]]
0.583089158644
             precision    recall  f1-score   support

          0       0.59      0.85      0.69      4000
         73       0.56      0.25      0.34      3167

avg / total       0.58      0.58      0.54      7167
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.831688630095
[[145 145 145 ..., 194 194 194]
 [194 194 145 ..., 145 194 194]]
0.831052093973
             precision    recall  f1-score   support

        145       0.74      0.12      0.21       896
        194       0.83      0.99      0.91      3999

avg / total       0.82      0.83      0.78      4895
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.644376278119
[[145 145 145 ..., 202 202 202]
 [145 145 145 ..., 202 145 145]]
0.631220177232
             precision    recall  f1-score   support

        145       0.73      0.65      0.69       911
        202       0.51      0.60      0.55       556

avg / total       0.64      0.63      0.64      1467
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.833474371373
[[145 145 145 ...,   0   0   0]
 [145 145 145 ...,   0   0 145]]
0.832594681708
             precision    recall  f1-score   support

          0       0.85      0.96      0.90      4072
        145       0.58      0.26      0.36       892

avg / total       0.80      0.83      0.81      4964
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.869268033712
[[194 194 194 ..., 202 202 202]
 [194 194 194 ..., 202 194 194]]
0.872373368724
             precision    recall  f1-score   support

        194       0.89      0.97      0.93      3984
        202       0.40      0.15      0.22       537

avg / total       0.84      0.87      0.85      4521
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.754593421397
[[194 194 194 ...,   0   0   0]
 [194   0 194 ...,   0   0 194]]
0.746912810278
             precision    recall  f1-score   support

          0       0.75      0.75      0.75      4105
        194       0.74      0.74      0.74      3912

avg / total       0.75      0.75      0.75      8017
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.886063537717
[[202 202 202 ...,   0   0   0]
 [202 202   0 ...,   0   0   0]]
0.888235294118
             precision    recall  f1-score   support

          0       0.90      0.98      0.94      4062
        202       0.54      0.20      0.29       528

avg / total       0.86      0.89      0.86      4590
```

KNN -3 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.711251960272
[[ 73  73  73 ..., 202 202 202]
 [ 73  73  73 ..., 202  73 145]]
0.698170731707
             precision    recall  f1-score   support

         73       0.70      0.97      0.82      3095
        145       0.65      0.14      0.23       967
        202       0.65      0.11      0.18       530

avg / total       0.68      0.70      0.62      4592
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.711251960272
[[ 73  73  73 ..., 202 202 202]
 [ 73  73  73 ..., 202  73 145]]
0.698170731707
             precision    recall  f1-score   support

         73       0.70      0.97      0.82      3095
        145       0.65      0.14      0.23       967
        202       0.65      0.11      0.18       530

avg / total       0.68      0.70      0.62      4592
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.475851324282
[[73 73 73 ...,  0  0  0]
 [73 73 73 ...,  0  0 73]]
0.470944609298
             precision    recall  f1-score   support

          0       0.61      0.32      0.42      4064
         73       0.41      0.77      0.53      3101
        145       0.75      0.13      0.22       923

avg / total       0.55      0.47      0.44      8088
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.622140250363
[[ 73  73  73 ..., 202 202 202]
 [ 73  73  73 ..., 194  73 194]]
0.618966644866
             precision    recall  f1-score   support

         73       0.57      0.55      0.56      3126
        194       0.65      0.75      0.70      3989
        202       0.49      0.08      0.14       530

avg / total       0.61      0.62      0.60      7645
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.455825726365
[[ 73  73  73 ...,   0   0   0]
 [  0   0   0 ..., 194 194 194]]
0.448393466164
             precision    recall  f1-score   support

          0       0.42      0.86      0.57      4052
         73       0.34      0.11      0.17      3116
        194       0.63      0.29      0.40      3974

avg / total       0.47      0.45      0.39     11142
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.560672059739
[[73 73 73 ...,  0  0  0]
 [ 0  0  0 ...,  0  0 73]]
0.558076225045
             precision    recall  f1-score   support

          0       0.56      0.90      0.69      4056
         73       0.57      0.19      0.29      3118
        202       0.52      0.09      0.16       540

avg / total       0.56      0.56      0.49      7714
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.680707945085
[[145 145 145 ..., 202 202 202]
 [145 145 145 ..., 194 145 145]]
0.678610804851
             precision    recall  f1-score   support

        145       0.30      0.48      0.37       908
        194       0.82      0.80      0.81      3996
        202       0.61      0.10      0.18       538

avg / total       0.71      0.68      0.67      5442
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.538014836696
[[145 145 145 ...,   0   0   0]
 [145 145 145 ...,   0   0 145]]
0.538599239203
             precision    recall  f1-score   support

          0       0.69      0.74      0.71      4084
        145       0.16      0.45      0.24       929
        194       0.70      0.35      0.47      3925

avg / total       0.64      0.54      0.56      8938
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.762713710117
[[145 145 145 ...,   0   0   0]
 [145 145 145 ...,   0   0 145]]
0.762976406534
             precision    recall  f1-score   support

          0       0.79      0.96      0.87      4105
        145       0.54      0.25      0.34       893
        202       0.45      0.05      0.10       512

avg / total       0.71      0.76      0.71      5510
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.709756524785
[[194 194 194 ...,   0   0   0]
 [194   0 194 ..., 194 194 194]]
0.704694068192
             precision    recall  f1-score   support

          0       0.72      0.74      0.73      4078
        194       0.69      0.75      0.72      3947
        202       0.62      0.08      0.14       539

avg / total       0.70      0.70      0.69      8564
```

KNN – 4 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.488984624006
[[ 73  73  73 ..., 202 202 202]
 [145 145 145 ..., 202  73 145]]
0.482138687836
             precision    recall  f1-score   support

         73       0.55      0.35      0.43      3084
        145       0.15      0.38      0.22       925
        194       0.65      0.66      0.66      4027
        202       0.22      0.12      0.15       530

avg / total       0.54      0.48      0.49      8566
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.444621315757
[[ 73  73  73 ...,   0   0   0]
 [ 73  73  73 ...,  73  73 194]]
0.438318686785
             precision    recall  f1-score   support

          0       0.56      0.19      0.28      4003
         73       0.32      0.61      0.42      3142
        145       0.39      0.22      0.28       913
        194       0.57      0.60      0.59      4004

avg / total       0.49      0.44      0.42     12062
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.440722724114
[[73 73 73 ...,  0  0  0]
 [73 73 73 ...,  0  0 73]]
0.43525596479
             precision    recall  f1-score   support

          0       0.51      0.42      0.46      4073
         73       0.39      0.60      0.47      3088
        145       0.37      0.16      0.22       943
        202       0.45      0.12      0.19       530

avg / total       0.45      0.44      0.42      8634
```

```
C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
redicted model accuracy: 0.458071305732
[ 73  73  73 ...,   0   0   0]
[ 73  73  73 ..., 194 194 194]]
.452173169062
             precision    recall  f1-score   support

          0       0.53      0.21      0.31      4036
         73       0.33      0.65      0.44      3107
        194       0.60      0.59      0.60      4013
        202       0.49      0.05      0.09       532

vg / total       0.50      0.45      0.43     11688
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.649321481669
[[145 145 145 ...,   0   0   0]
 [145 145 145 ..., 194 194 145]]
0.644559257697
             precision    recall  f1-score   support

          0       0.66      0.76      0.70      4094
        145       0.43      0.21      0.28       918
        194       0.66      0.71      0.68      3930
        202       0.59      0.04      0.08       542

avg / total       0.63      0.64      0.62      9484
```

KNN- 5 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python KNN.py
Predicted model accuracy: 0.40821376734
[[ 73  73  73 ...,   0   0   0]
 [  0   0   0 ...,   0   0 194]]
0.40130065826
             precision    recall  f1-score   support

          0       0.37      0.73      0.49      4058
         73       0.36      0.27      0.31      3090
        145       0.37      0.20      0.26       920
        194       0.57      0.26      0.36      3988
        202       0.51      0.05      0.09       553

avg / total       0.44      0.40      0.37     12609
```

Decision Trees -2 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.806556890669
0.800494437577
             precision    recall  f1-score   support

         73       0.81      0.97      0.88      3140
        145       0.67      0.22      0.33       905

avg / total       0.78      0.80      0.76      4045
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.719073580998
0.713339907029
             precision    recall  f1-score   support

         73       0.69      0.66      0.67      3165
        194       0.73      0.76      0.75      3934

avg / total       0.71      0.71      0.71      7099
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[202]
Predicted model accuracy: 0.868257600523
0.86298011441
             precision    recall  f1-score   support

         73       0.87      0.99      0.92      3122
        202       0.68      0.16      0.26       549

avg / total       0.84      0.86      0.83      3671
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[0]
Predicted model accuracy: 0.618794474676
0.60583228687
             precision    recall  f1-score   support

          0       0.61      0.85      0.71      4000
         73       0.61      0.30      0.40      3167

avg / total       0.61      0.61      0.57      7167
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.836714679743
0.834320735444
             precision    recall  f1-score   support

        145       0.63      0.22      0.33       896
        194       0.85      0.97      0.91      3999

avg / total       0.81      0.83      0.80      4895
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.848222920696
0.845286059629
             precision    recall  f1-score   support

          0       0.86      0.98      0.91      4072
        145       0.69      0.25      0.37       892

avg / total       0.83      0.85      0.81      4964
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.889375539186
0.888520238885
             precision    recall  f1-score   support

        194       0.90      0.99      0.94      3984
        202       0.61      0.17      0.27       537

avg / total       0.86      0.89      0.86      4521
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.762938292856
0.752276412623
             precision    recall  f1-score   support

          0       0.76      0.76      0.76      4105
        194       0.74      0.75      0.75      3912

avg / total       0.75      0.75      0.75      8017
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[0]
Predicted model accuracy: 0.895738004968
0.899346405229
             precision    recall  f1-score   support

          0       0.90      0.99      0.95      4062
        202       0.79      0.17      0.28       528

avg / total       0.89      0.90      0.87      4590
```

Decision Trees-3 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.650342324816
0.643471754583
             precision    recall  f1-score   support

         73       0.60      0.66      0.63      3087
        145       0.52      0.22      0.31       911
        194       0.69      0.73      0.71      4021

avg / total       0.64      0.64      0.63      8019
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.72288290643
0.706881533101
             precision    recall  f1-score   support

         73       0.72      0.95      0.82      3095
        145       0.60      0.23      0.34       967
        202       0.57      0.13      0.21       530

avg / total       0.68      0.71      0.65      4592
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.562022405223
0.556256181998
             precision    recall  f1-score   support

          0       0.60      0.72      0.65      4064
         73       0.48      0.44      0.46      3101
        145       0.59      0.23      0.33       923

avg / total       0.55      0.56      0.54      8088
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[202]
Predicted model accuracy: 0.671898913001
0.667102681491
             precision    recall  f1-score   support

         73       0.63      0.66      0.64      3126
        194       0.70      0.75      0.73      3989
        202       0.50      0.08      0.15       530

avg / total       0.66      0.67      0.65      7645
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[73]
Predicted model accuracy: 0.566497024531
0.558517321845
             precision    recall  f1-score   support

          0       0.54      0.68      0.60      4052
         73       0.48      0.18      0.27      3116
        194       0.60      0.73      0.66      3974

avg / total       0.54      0.56      0.53     11142
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[0]
Predicted model accuracy: 0.581207218419
0.576873217527
             precision    recall  f1-score   support

          0       0.58      0.85      0.69      4056
         73       0.57      0.30      0.39      3118
        202       0.62      0.11      0.19       540

avg / total       0.58      0.58      0.53      7714
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.760158791421
0.753031973539
             precision    recall  f1-score   support

        145       0.50      0.21      0.30       908
        194       0.78      0.96      0.86      3996
        202       0.57      0.13      0.22       538

avg / total       0.71      0.75      0.70      5442
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[0]
Predicted model accuracy: 0.69826457655
0.690087267845
             precision    recall  f1-score   support

          0       0.70      0.76      0.73      4084
        145       0.58      0.19      0.29       929
        194       0.69      0.73      0.71      3925

avg / total       0.68      0.69      0.68      8938
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[0]
Predicted model accuracy: 0.774311227268
0.770780399274
             precision    recall  f1-score   support

          0       0.78      0.97      0.87      4105
        145       0.61      0.23      0.33       893
        202       0.58      0.10      0.17       512

avg / total       0.74      0.77      0.72      5510
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.718970047294
0.711816907987
             precision    recall  f1-score   support

          0       0.72      0.76      0.74      4078
        194       0.71      0.75      0.73      3947
        202       0.70      0.11      0.19       539

avg / total       0.71      0.71      0.70      8564
```

Decision Trees-4 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[145]
Predicted model accuracy: 0.612728100592
0.605766985758
             precision    recall  f1-score   support

         73       0.56      0.66      0.61      3084
        145       0.47      0.22      0.30       925
        194       0.66      0.72      0.69      4027
        202       0.45      0.09      0.16       530

avg / total       0.59      0.61      0.58      8566
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[73]
Predicted model accuracy: 0.53182439995
0.524208257337
             precision    recall  f1-score   support

          0       0.51      0.68      0.58      4003
         73       0.45      0.19      0.27      3142
        145       0.46      0.21      0.29       913
        194       0.57      0.70      0.63      4004

avg / total       0.51      0.52      0.49     12062
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[73]
Predicted model accuracy: 0.531897150799
0.523164234422
             precision    recall  f1-score   support

          0       0.57      0.71      0.63      4073
         73       0.45      0.44      0.44      3088
        145       0.52      0.21      0.30       943
        202       0.59      0.12      0.20       530

avg / total       0.52      0.52      0.50      8634
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[0]
Predicted model accuracy: 0.542416386458
0.534052019165
             precision    recall  f1-score   support

          0       0.52      0.67      0.58      4036
         73       0.45      0.19      0.27      3107
        194       0.58      0.72      0.64      4013
        202       0.54      0.10      0.17       532

avg / total       0.52      0.53      0.50     11688
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.661152057698
0.655314213412
             precision    recall  f1-score   support

          0       0.67      0.76      0.71      4094
        145       0.49      0.20      0.28       918
        194       0.65      0.74      0.69      3930
        202       0.57      0.08      0.14       542

avg / total       0.64      0.66      0.63      9484
```

Decision Trees-5 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python DTpruebas2singraf
icos.py
[194]
Predicted model accuracy: 0.511385537869
0.504639543183
             precision    recall  f1-score   support

          0       0.49      0.68      0.57      4058
         73       0.43      0.19      0.26      3090
        145       0.43      0.20      0.27       920
        194       0.55      0.70      0.61      3988
        202       0.49      0.09      0.16       553

avg / total       0.49      0.50      0.47     12609
```

Näive Bayes-2 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.772437323839
0.776266996292
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.78      1.00      0.87      3140
        145       0.00      0.00      0.00       905

avg / total       0.60      0.78      0.68      4045
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.596044123241
0.595576841809
             precision    recall  f1-score   support

         73       0.60      0.28      0.38      3165
        194       0.59      0.85      0.70      3934

avg / total       0.60      0.60      0.56      7099
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.851095129127
0.85044946881
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.85      1.00      0.92      3122
        202       0.00      0.00      0.00       549

avg / total       0.72      0.85      0.78      3671
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.561364587694
0.552671968746
             precision    recall  f1-score   support

          0       0.56      0.91      0.70      4000
         73       0.47      0.10      0.16      3167

avg / total       0.52      0.55      0.46      7167
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[145 145]
Predicted model accuracy: 0.781060373889
0.780183861083
             precision    recall  f1-score   support

        145       0.19      0.06      0.09       896
        194       0.82      0.94      0.87      3999

avg / total       0.70      0.78      0.73      4895
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[202 145]
Predicted model accuracy: 0.630470347648
0.626448534424
             precision    recall  f1-score   support

        145       0.63      0.98      0.77       911
        202       0.59      0.05      0.09       556

avg / total       0.61      0.63      0.51      1467
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.814555125725
0.820306204674
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.82      1.00      0.90      4072
        145       0.00      0.00      0.00       892

avg / total       0.67      0.82      0.74      4964
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[202 202]
Predicted model accuracy: 0.878757714513
0.880557398806
             precision    recall  f1-score   support

        194       0.88      1.00      0.94      3984
        202       0.00      0.00      0.00       537

avg / total       0.78      0.88      0.83      4521
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.572091456798
0.572159161781
             precision    recall  f1-score   support

          0       0.65      0.35      0.46      4105
        194       0.54      0.80      0.65      3912

avg / total       0.60      0.57      0.55      8017
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.880899463982
0.884967320261
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.88      1.00      0.94      4062
        202       0.00      0.00      0.00       528

avg / total       0.78      0.88      0.83      4590
```

Näive Bayes-3 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.527292453889
0.530240678389
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.50      0.28      0.36      3087
        145       0.00      0.00      0.00       911
        194       0.54      0.84      0.66      4021

avg / total       0.46      0.53      0.47      8019
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.680475692629
0.67399825784
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.67      1.00      0.81      3095
        145       0.00      0.00      0.00       967
        202       0.00      0.00      0.00       530

avg / total       0.45      0.67      0.54      4592
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.497700126122
0.49727992087
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.50      0.91      0.65      4064
         73       0.42      0.10      0.16      3101
        145       0.00      0.00      0.00       923

avg / total       0.42      0.50      0.39      8088
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.553310049837
0.556180510137
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.55      0.28      0.37      3126
        194       0.56      0.85      0.67      3989
        202       0.00      0.00      0.00       530

avg / total       0.52      0.56      0.50      7645
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.40978000377
0.404774726261
             precision    recall  f1-score   support

          0       0.46      0.26      0.33      4052
         73       0.34      0.09      0.15      3116
        194       0.40      0.80      0.53      3974

avg / total       0.40      0.40      0.35     11142
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.521546359676
0.520611874514
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.53      0.91      0.67      4056
         73       0.45      0.10      0.16      3118
        202       0.00      0.00      0.00       540

avg / total       0.46      0.52      0.42      7714
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[202 145]
Predicted model accuracy: 0.728896730441
0.7335538405
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

        145       0.35      0.01      0.02       908
        194       0.73      1.00      0.85      3996
        202       0.00      0.00      0.00       538

avg / total       0.60      0.73      0.62      5442
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.51344365748
0.515663459387
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.59      0.35      0.44      4084
        145       0.00      0.00      0.00       929
        194       0.49      0.81      0.61      3925

avg / total       0.49      0.52      0.47      8938
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.733747141457
0.74500907441
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.75      1.00      0.85      4105
        145       0.00      0.00      0.00       893
        202       0.00      0.00      0.00       512

avg / total       0.56      0.75      0.64      5510
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.535750569277
0.535497431107
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.62      0.35      0.44      4078
        194       0.51      0.80      0.62      3947
        202       0.00      0.00      0.00       539

avg / total       0.53      0.54      0.50      8564
```

Näive Bayes-4 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.494658680957
0.499649778193
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.49      0.28      0.35      3084
        145       0.00      0.00      0.00       925
        194       0.50      0.85      0.63      4027
        202       0.00      0.00      0.00       530

avg / total       0.41      0.50      0.42      8566
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[73 73]
Predicted model accuracy: 0.494658680957
0.499649778193
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.49      0.28      0.35      3084
        145       0.00      0.00      0.00       925
        194       0.50      0.85      0.63      4027
        202       0.00      0.00      0.00       530

avg / total       0.41      0.50      0.42      8566
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.466296038916
0.466643502432
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.47      0.92      0.62      4073
         73       0.41      0.10      0.15      3088
        145       0.00      0.00      0.00       943
        202       0.00      0.00      0.00       530

avg / total       0.37      0.47      0.35      8634
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESpruebas
.py
[0 0]
Predicted model accuracy: 0.38966605919
0.390571526352
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision a
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.44      0.27      0.33      4036
         73       0.33      0.10      0.15      3107
        194       0.38      0.80      0.52      4013
        202       0.00      0.00      0.00       532

avg / total       0.37      0.39      0.33     11688
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESprueba
.py
[0 0]
Predicted model accuracy: 0.483598519596
0.479228173766
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.56      0.34      0.42      4094
        145       0.00      0.00      0.00       918
        194       0.45      0.80      0.58      3930
        202       0.00      0.00      0.00       542

avg / total       0.43      0.48      0.42      9484
```

Näive Bayes-5 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>python NAIVEBAYESprueba
.py
[0 0]
Predicted model accuracy: 0.36138672758
0.357125862479
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMetricWarning: Precision
nd F-score are ill-defined and being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.40      0.25      0.31      4058
         73       0.32      0.10      0.15      3090
        145       0.00      0.00      0.00       920
        194       0.35      0.80      0.49      3988
        202       0.00      0.00      0.00       553

avg / total       0.32      0.36      0.29     12609
```

Logistic Regression-2 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[73]
0.772805933251
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.77      1.00      0.87      3126
        145       0.00      0.00      0.00       919

avg / total       0.60      0.77      0.67      4045
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.571066347373
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.00      0.00      0.00      3045
        194       0.57      1.00      0.73      4054

avg / total       0.33      0.57      0.42      7099
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[73]
0.858894034323
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.86      1.00      0.92      3153
        202       0.00      0.00      0.00       518

avg / total       0.74      0.86      0.79      3671
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.565369052602
             precision    recall  f1-score   support

          0       0.57      0.93      0.71      4067
         73       0.49      0.09      0.15      3100

avg / total       0.54      0.57      0.47      7167
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.814096016343
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

        145       0.00      0.00      0.00       910
        194       0.81      1.00      0.90      3985

avg / total       0.66      0.81      0.73      4895
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[145]
0.629175187457
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

        145       0.63      1.00      0.77       923
        202       0.00      0.00      0.00       544

avg / total       0.40      0.63      0.49      1467
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.811643835616
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.81      1.00      0.90      4029
        145       0.00      0.00      0.00       935

avg / total       0.66      0.81      0.73      4964
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.889626188896
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

        194       0.89      1.00      0.94      4022
        202       0.00      0.00      0.00       499

avg / total       0.79      0.89      0.84      4521
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.491330921791
             precision    recall  f1-score   support

          0       0.49      0.89      0.63      3989
        194       0.47      0.10      0.17      4028

avg / total       0.48      0.49      0.40      8017
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.884095860566
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.88      1.00      0.94      4058
        202       0.00      0.00      0.00       532

avg / total       0.78      0.88      0.83      4590
```

Logistic Regression-3 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.489836637985
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.00      0.00      0.00      3177
        145       0.00      0.00      0.00       914
        194       0.49      1.00      0.66      3928

avg / total       0.24      0.49      0.32      8019
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[73]
0.678571428571
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.68      1.00      0.81      3116
        145       0.00      0.00      0.00       924
        202       0.00      0.00      0.00       552

avg / total       0.46      0.68      0.55      4592
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.492457962413
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.50      0.93      0.65      4014
         73       0.43      0.08      0.14      3134
        145       0.00      0.00      0.00       940

avg / total       0.41      0.49      0.37      8088
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.528842380641
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.00      0.00      0.00      3049
        194       0.53      1.00      0.69      4043
        202       0.00      0.00      0.00       553

avg / total       0.28      0.53      0.37      7645
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.351732184527
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.36      0.89      0.51      3988
         73       0.00      0.00      0.00      3140
        194       0.28      0.09      0.14      4014

avg / total       0.23      0.35      0.23     11142
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.523074928701
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.53      0.92      0.67      4086
         73       0.45      0.08      0.14      3102
        202       0.00      0.00      0.00       526

avg / total       0.46      0.52      0.41      7714
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.727306137449
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

        145       0.00      0.00      0.00       932
        194       0.73      1.00      0.84      3958
        202       0.00      0.00      0.00       552

avg / total       0.53      0.73      0.61      5442
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.441597672857
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.44      0.88      0.59      4013
        145       0.00      0.00      0.00       920
        194       0.42      0.11      0.17      4005

avg / total       0.39      0.44      0.34      8938
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.724863883848
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.72      1.00      0.84      3994
        145       0.00      0.00      0.00       968
        202       0.00      0.00      0.00       548

avg / total       0.53      0.72      0.61      5510
```

Logistic Regression-4 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[194]
0.466845668924
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

         73       0.00      0.00      0.00      3113
        145       0.00      0.00      0.00       902
        194       0.47      1.00      0.64      3999
        202       0.00      0.00      0.00       552

avg / total       0.22      0.47      0.30      8566
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.333609683303
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.35      0.36      0.35      4046
         73       0.00      0.00      0.00      3182
        145       0.00      0.00      0.00       942
        194       0.33      0.66      0.44      3892

avg / total       0.22      0.33      0.26     12062
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.471623812833
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.47      0.93      0.63      4093
         73       0.44      0.09      0.15      3036
        145       0.00      0.00      0.00       941
        202       0.00      0.00      0.00       564

avg / total       0.38      0.47      0.35      8634
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.338124572211
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.35      0.88      0.50      4019
         73       0.00      0.00      0.00      3107
        194       0.28      0.10      0.15      4002
        202       0.00      0.00      0.00       560

avg / total       0.22      0.34      0.22     11688
```

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.428511176719
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.43      0.37      0.40      4050
        145       0.00      0.00      0.00       931
        194       0.43      0.66      0.52      3931
        202       0.00      0.00      0.00       572

avg / total       0.36      0.43      0.38      9484
```

Logistic Regression-5 eventos

```
(C:\Program Files\Anaconda2) C:\Users\JAMEFER\Documents\Maestria\Merari Compu\Algoritmos\Iraq 2>
python LogistRegresion.py
[0]
0.331588547863
C:\Program Files\Anaconda2\lib\site-packages\sklearn\metrics\classification.py:1113: UndefinedMe
tricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predic
ted samples.
  'precision', 'predicted', average, warn_for)
             precision    recall  f1-score   support

          0       0.35      0.35      0.35      4116
         73       0.00      0.00      0.00      3124
        145       0.00      0.00      0.00       915
        194       0.32      0.70      0.44      3941
        202       0.00      0.00      0.00       513

avg / total       0.21      0.33      0.25     12609
```

APÉNDICE C

CÓDIGO DE INTERFAZ DE USUARIO

CÓDIGO DE INTERFAZ DE USUARIO

```
#Importaciones necesarias.
#-*- coding: 850 -*-
from Tkinter import *
import tkMessageBox
import pickle
from sklearn import model_selection
from sklearn.externals import joblib
from sklearn.metrics import accuracy_score
import string

#Función para realizar la clasificación
def Clasificacion():
  global lat,lon
  #valida que los datos ingresados sean numéricos
  try:
    lat = float(caja1.get())
  except ValueError:
    tkMessageBox.showwarning("Cuidado","Introduzca latitud correcta")

  try:
    lon = float(caja2.get())
  except ValueError:
    tkMessageBox.showwarning("Cuidado","Introduzca longitud correcta")
    return False

  #valida que los datos ingresados correspondan a la latitud y longitud de Iraq
  if lat > 29.12 and lat < 37.29 and lon > 39.22 and lon < 48.48:

    #Carga el modelo para la clasificación en formato .pkl
    filename = 'model_DT.pkl'
    clf=joblib.load (filename)
    a=clf.predict([[lat,lon]])

    if a == [0]:
      tkMessageBox.showinfo("Mensaje","La clasificación corresponde a: Refugiados")
    if a == [194]:
      tkMessageBox.showinfo("Mensaje","La clasificación corresponde a: Lucha con
artillería")
  else:
    tkMessageBox.showinfo("Mensaje","Valores incorrectos")
```

```
  caja1.delete(0,20)
  caja2.delete(0,20)

#Creación del GUI (Interfaz gráfica de usuario)
gui = Tk()
#Titulo del GUI
gui.title("SoftMENA")
#Dimensiones (ancho,alto,posición x,posición y)
gui.geometry("400x250+400+150")

#Creación de una etiqueta
var1 = StringVar()
var1.set("Introduzca la latitud:")
label1 = Label(gui,textvariable=var1,height = 2)
label1.pack()

#Creación de una caja de texto para latitud
lat1=StringVar()
caja1=Entry(gui,bd=4,textvariable=lat1)
caja1.pack()

#Creación de otra etiqueta
var2 = StringVar()
var2.set("Introduzca la longitud:")
label2 = Label(gui,textvariable=var2,height = 2)
label2.pack()

#Creación de otra caja de texto para longitud
long2=StringVar()
caja2=Entry(gui,bd=4,textvariable=long2)
caja2.pack()

#Botón para clasificar
boton1 = Button(gui, text = "Clasifica", command = Clasificacion,width=15)
boton1.pack()

#Cargar el GUI
gui.mainloop()
```

APÉNDICE D

CÓDIGOS DE CLASIFICADORES

CÓDIGOS DE CLASIFICADORES

Algoritmo KNN

```
#Importar librerías
import numpy as np
import pandas as pd
import csv
from sklearn import neighbors, datasets
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import cross_validation

#Leer dataset
df = pd.read_csv("Iraq-con194_REF.csv")

#Columnas dataset
X= np.array(df.drop(["3"],1))
y= np.array(df["3"])

#Valor - variables
X_train, X_test, y_train, y_test = cross_validation.train_test_split(X,y, test_size= 0.3,
random_state=0)

#Creación/ división de datasets
X_train.shape, y_train.shape
X_test.shape, y_test.shape

# Crea una instancia del clasificador KNN, luego ajusta los datos de entrenamiento
clf = neighbors.KNeighborsClassifier().fit(X_train, y_train)
clf.score(X_train, y_train)

#Hace predicciones de clase para las muestras en X
Z = clf.predict(X)

#Compara las clases entre etiquetas predichas y las reales
accuracy=clf.score(X,y)
print ("Predicted model accuracy: "+ str(accuracy))

#Agrega una fila de clases pronosticadas a la matriz, para facilitar la comparación
A = np.vstack([y, Z])
```

```
print(A)
clf.fit(X_train, y_train)
accuracy=clf.score(X_train,y_train)

##Reporte de clasificación##
y_pred = clf.predict(X_test)

#Muestra accuracy
#print(accuracy_score(y_test, y_pred))
print classification_report(y_test, y_pred)
```

Algoritmo Näive Bayes

```
#Importa librerías
import csv
from sklearn.naive_bayes import GaussianNB
import pandas as pd
import numpy as np
from sklearn.metrics import classification_report #classification report
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import cross_validation

#Leer dataset
df = pd.read_csv("Iraq-con073_202.csv")

#Columnas dataset
X= np.array(df.drop(["3"],1))
y= np.array(df["3"])

#Variables
X_train, X_test, y_train, y_test = cross_validation.train_test_split(X,y, test_size= 0.3,
random_state=0)

#Creación/ división de datasets
X_train.shape, y_train.shape
X_test.shape, y_test.shape

#Crea un clasificador Gaussiano
model = GaussianNB()
```

```
#Entrena el modelo usando los conjuntos de entrenamiento
model.fit(X_train, y_train)
model.score(X_train, y_train)
model.fit(X_train, y_train)

#Predicciones
accuracy=model.score(X,y)
print ("Predicted model accuracy: "+ str(accuracy))

##Reporte de clasificación##
y_pred = model.predict(X_test)
#print(accuracy_score(y_test, y_pred))
print classification_report(y_test, y_pred)
```

Algoritmo Decision Trees

```
#Importa librerías
import csv
import numpy as np
import pandas as pd
from sklearn import tree
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import cross_validation

import pickle
from sklearn import model_selection

#Leer dataset
df = pd.read_csv("Iraq-con194_REF.csv")

#Columnas dataset
X= np.array(df.drop(["3"],1))
y= np.array(df["3"])

#Valor - variables
X_train, X_test, y_train, y_test = cross_validation.train_test_split(X,y, test_size= 0.3,
random_state=0)
```

```
#Creación/ división de datasets
X_train.shape, y_train.shape
X_test.shape, y_test.shape

#Descripción del clasificador
clf = tree.DecisionTreeClassifier().fit(X_train, y_train)
clf.score(X_train, y_train)
clf.fit(X_train, y_train) #ajusta el modelo de acuerdo con los datos de entrenamiento

#Guarda el modelo
filename = 'model_DT.txt'
pickle.dump(clf, open(filename, 'wb')) #wb --> escritura

#Predicciones - precisión
accuracy=clf.score(X,y)
print ("Predicted model accuracy: "+ str(accuracy))

##Reporte de clasificación##
y_pred = clf.predict(X_test)
#print(accuracy_score(y_test, y_pred))
print classification_report(y_test, y_pred)
```

Algoritmo Logistic Regression

```
#Importa librerías
import csv
import numpy as np
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
from sklearn import cross_validation

#Lee el conjunto de datos
df0 = pd.read_csv("Iraq-con073_202.csv")
df = df0.sample(frac=1)

X = np.array(df.drop(["3"],1))
y = np.array(df["3"])

#Valor - variables
X_train, X_test, y_train, y_test = cross_validation.train_test_split(X,y, test_size=0.3,
random_state=0)

#Crea una instancia y adapta los datos
LB = LogisticRegression().fit(X_train, y_train)
y_pred = LB.predict(X_test)

#Crea una instancia de Logistic Regresion y ajusta los datos
LB.fit(X, y)

##Muestra accuracy##
print(accuracy_score(y_test, y_pred))
print classification_report(y_test, y_pred)
```

REFERENCIAS

Abiteboul, S. (1997). Querying semi-structured data. En *ICDT '97 Proceedings of the 6th International Conference on Database Theory.* Londres: Springer-Verlag

Advanced Tech Computing Group UTPL. (2008, 14 de abril). *Clasificación supervisada y no supervisada.* Recuperado de https://advancedtech.wordpress.com/2008/04/14/clasificacion-supervisada-y-no-supervisada/

Alférez, G. H. (2016). Tweeting in New York city: Data science can teach us to sympathize. *Adventist Review*, *193*(2), 47-49.

Alférez, G. H., Jiménez, J., Hernández-Navarro, H., González, M., Domínguez, R., Briones, A., y Hernández Villalvazo, H. (2016). *Aplicación de ciencia de datos en las historias clínicas de una clínica ubicada en el noreste de México para corroborar la relación entre caries dental y diabetes*. Ponencia presentada en el 2° Congreso de Investigación Universitaria de la División Interamericana (CIDIA 2016), Montemorelos, Nuevo León, México.

Alférez, G. H., Rodríguez, J., Clausen, B. y Pompe, L. (2015). Interpreting the geochemistry of Southern California granitic rocks using machine learning. En *Proceedings on the International Conference on Artificial Intelligence (ICAI)* (pp. 592-598). Atenas: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Al-Naami, K. M., Seker, S. y Khan, L. (2014). GISQF: An efficient spatial query processing system. En *CLOUD '14 Proceedings of the 2014 IEEE International Conference on Cloud Computing* (pp. 681-688). Washington: IEEE Computer Society. doi:10.1109/CLOUD.2014.96

Aluja, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial. *Qüestiió, 25*(3), 479-498.

Álvarez Munarriz, L. (1994). *Fundamentos de inteligencia artificial.* Murcia: Ediciones de la Universidad de Murcia.

Arasu, A. y García Molina, H. (2003, junio). *Extracting structures data from web pages*. Conferencia presentada en SIGMOD '03 Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. New York, NY. doi:10.1145/872757.872799

Barabás, P. y Kovács, L. (2009). *Usability of summation hack in bayes classification*. Recuperado de http://conf.uni-obuda.hu/cinti2008/15_cinti2008_submission.pdf

Bi, S., Gao, J., Wang, Y. y Cao, Y. (2015). *A contrast of the degree of activity among the three major powers, USA, China, and Russia: Insights from media reports*. Presentada en International Conference on Behavioral, Economic and Socio-Cultural Computing, Nanjing, China. doi: 10.1109/BESC.2015.7365955

Brownlee, J. (2014). *Classification accuracy is not enough: more performance measures you can use.* Recuperado de http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

Brownlee, J. (2016). *Logistic regression for machine learning.* Recuperado de http://machinelearningmastery.com/logistic-regression-for-machine-learning/

Camargo Vega, J. J., Camargo Ortega, J. F. y Joyanes Aguilar, L. (2015). Conociendo Big Data. *Facultad de Ingeniería, 24*(38), 63-77. Recuperado de http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-11292015000100006&lng=en&tlng=es.

Camps Paré, R., Casillas Santillán, L., Costal Costa, D., Gibert Ginestá, M., Martín Escofet, C. y Pérez Mora, O. (2005). *Bases de datos*. Recuperado de http://www.uoc.edu/masters/oficiales/img/913.pdf

Cárdenas-Montes, M. (2014). *Regresión lineal*. Recuperado de http://wwwae.ciemat.es/~cardenas/docs/lessons/RegresionLineal.pdf. Fecha de consulta: 2 noviembre 2017.

Chan, J. O. (2013). An architecture for big data analytics. *Communications of the IIMA, 13*(2). Recuperado de http://scholarworks.lib.csusb.edu/ciima/vol13/iss2/1/

Chen, H., Jiang, J., Hong, Z. y Lin, L. (2018). Decomposition of UML activity diagrams. *Software: Practice & Experience*, 48(1), 105-122. doi:10.1002/spe.2519

Chen, K. (2017). Análisis de Big Data: el efecto Trump sobre el discurso comercial. *BBVA Research*. Recuperado de https://www.bbvaresearch.com/wp-content/uploads/2017/03/170322_US_TradeNarratives_esp.pdf

Choochaisri, S., (2016, 5 de junio). *What is the best way to understand the terms "precision" and "recall"?* Recuperado de https://www.quora.com/What-is-the-best-way-to-understand-the-terms-precision-and-recall

Christie, L. y Christie, J. (1984), *Enciclopedia de términos de microcomputadoras,* México: Prentice-Hall.

Davenport, T. H. y Patil, D. J. (2012, octubre). Data scientist: The sexiest job of the 21st century. *Harvard Business Review, 90*(10), 70-76.

Dhar, V. (2013). Data Science and prediction. *Communications of the Acm*, 56(12), 64-73. doi:10.1145/2500499

Digital Reasoning. (2018). *Unstructured data: A big deal in big data*. Recuperado de http://www.digitalreasoning.com/resources/Holistic-Analytics.pdf

Dincer, A., y Uraz, B. (2013). *Google Maps JavaScript API Cookbook*. Birmingham: Packt Publishing. Recuperado de http://www.luciopanasci.it/Ebooks/Google%20Maps%20JavaScript%20API%20Cookbook.pdf

Domínguez, R. (2017). *Aplicación de ciencia de datos para la creación de software predictivo de morbimortalidad maternal en México* (Tesis de maestría). Universidad de Montemorelos, Nuevo León, México.

Ecured. (2017). *Iraq*. Recuperado de https://www.ecured.cu/Irak Consultado el 14 noviembre 2017.

Esqueda-Elizondo, J. J. (2002, noviembre). *Fundamentos de procesamiento de imágenes.* Trabajo presentado en el *CONATEC*, Instituto Tecnológico de Ciudad Madero, Tamaulipas, México. Recuperado de goo.gl/k9C14U

Fawcett, T., y Hardin, D. (2017, agosto). *Machine learning vs. statistics: The Texas death match of data science*. Recuperado de http://www.kdnuggets.com/2017/08/machine-learning-vs-statistics.html

Flores Pulido, L. (18 de octubre de 2012). Técnicas de clasificación supervisada [Mensaje en un blog]. Recuperado de https://aicitel.files.wordpress.com/2012/10/ic_clase5.pdf

García, A., Martínez, G., Núñez, G. y Guzmán, A. (1998). *Clasificación supervisada: inducción de árboles de decisión, algoritmo k-d*. Recuperado de http://www.cic.ipn.mx/aguzman/papers/109%20Algoritmo%20KD.%20Clasificacion%20supervisada.pdf

Global Database of Events, Language and Tone. (2013). *Data format codebook V 1.03*. Recuperado de http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf

Goes, B. P. (2014). Big Data and IS research. *MIS Quarterly, 38*(3), 3-5.

Gómez, D., Prieto, F. y Guzmán, M. (2015). Nearest neighbors by adaptive simulated annealing. *IEEE Latin America Transactions, 13*(7), 2398-2404. doi: 10.1109/TLA.2015.7273804

Google Developers. (2017). *Quotas and limits*. Recuperado de https://cloud.google.com/bigquery/quotas

Gustin, P. (2006). *La Ventana 10/40: nuevas oportunidades para la misión*. Recuperado de http://circle.adventist.org/files/CD2008/CD2/dialogue/articles/12_2_gustin_s.htm

Hackeling, G. (2014). *Mastering machine learning with scikit-learn*. Recuperado de http://pdf.th7.cn/down/files/1603/Mastering%20Machine%20Learning%20with%20scikit-learn.pdf

Harrington, P. (2012). *Machine learning in action*. Recuperado de https://www.manning.com/books/machine-learning-in-action

Hastie, T., Tibshirani, R. y Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. New York: Springer.

He, S., Wang, G., Zhang, M. y Cook, D. (2013). Multivariate process monitoring and fault identification using multiple decision tree classifiers. *International Journal of Production Research*, *51*(11), 3355-3371. doi: 10.1080/00207543.2013.774474

Helmenstine, T. (2014, 11 de junio). *What is the difference between accuracy and precision?* Recuperado de https://sciencenotes.org/what-is-the-difference-between-accuracy-and-precision/

Hurwitz J., Nugent, A., Halper, F. y Kaufman, M. (2013). *Unstructured data in a big data*. Recuperado de http://www.dummies.com/programming/big-data/engineering/ unstructured-data-in-a-big-data-environment/

Iconiq Inc. (2016, 29 de junio). *Top 10 machine learning algorithms*. Recuperado de https://www.dezyre.com/article/top-10-machine-learning-algorithms/202

Iglesia Adventista del Séptimo Día. (2013). *División Sudamericana*. Recuperado de http://www.adventistas.org/es/misionglobal/2013/11/14/que-es-mision-global/

Iglesia Adventista del Séptimo Día. (2016). *Iglesia mundial*. Recuperado de https://www.adventist.org/es/iglesia-mundial/

Johansson, R., (2016). *Introduction to scientific computing in Python*. Recuperado de http://forum.myquant.cn/uploads/default/original/1X/d87be13e0e121369e22b59cb6d74d295f7e8a04e.pdf

Kaelo, P., Narayanan, S. y Thuto, M. V. (2017). A modified quadratic hybridization of Polak-Ribiere-Polyak and Fletcher-Reeves conjugate gradient method for unconstrained optimization problems. *An International Journal of Optimization and Control: Theories & Applications (IJOCTA), 7*(2), 177-185. doi:10.11121/ijocta.01.2017.00339

Kotsiantis, S. B. (2007). Supervised Machine Learning: A review of classification techniques. *Informatica, 31,* 249-268. doi: 10.1.1.95.9683

Kou, Y., Wu, Q., Li, X. y Hong, S. (2016, junio). *Fast clustering based on state Learning Machine*. Conferencia presentada en 12th World Congress on intelligent control and automation (WCICA). doi: 10.1109/WCICA.2016.7578369

Kumar, S., Benigni, M. y Carley, K. (2016, septiembre). *The impact of US cyber policies on cyber-attacks trend*. Conferencia presentada en Intelligence and Security Informatics (ISI). doi: 10.1109/ISI.2016.7745464

Kumar, V. B., Kumar, S. S. y Saboo, V. (2016, septiembre). *Dermatological disease detection using image processing and machine learning*. Conferencia presentada en el Third International Conference on Artificial Intelligence and Pattern Recognition, Lodz, Poland. doi:10.1109/ICAIPR.2016.7585217

Learned Miller, E. G. (2014). *Introduction to supervised learning*. Recuperado de https://people.cs.umass.edu/~elm/Teaching/Docs/supervised2014a.pdf

Leetaru, K. y Schrodt, P. (2013). GDELT: *Global Data on Events, Location and Tone*, 1979-2012. Recuperado de http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf

Leung, C., Jiang, F., Zhang, H. y Pazdor, A. (2016, agosto). *A data science model for big data analytucs of frequent patterns*. Conferencia presentada en la 14th Intl Conf on Pervasive Intelligence and Computing. Auckland, New Zealand. doi:10.1109/DASC-PICom-DataCom-CyberSciTec.2016.148

Live Universal Awareness Map. (2014). *Map data, LiveuaMap Open Street Map*. Recuperado de https://iraq.liveuamap.com/

Marín, C., Alférez G. H., Córdova J. y González V. (2015, junio). *Detection of melanoma through image recognition and artificial neural networks*. Conferencia presentada en la World Congress on Medical Physics and Biomedical Engineering, Toronto, Canadá. doi: 10.1007/978-3-319-19387-8_204

Marzal, A., Llorens, D. y Gracia, I. (2003). *Aprender a programar con Python: una experiencia docente*. Recuperado de http://www.academia.edu/16383969/Aprender_a_programar_con_Python_una_experiencia_docente

Maté Jiménez, C. (2014). Big data. Un nuevo paradigma de análisis de datos. *Anales de Mecánica y Electricidad,* 10-16. Recuperado de https://www.iit.comillas.edu/docs/IIT-14-153A.pdf

Mathur, A., Sihag, A, Bagaria, E. y Rajawat, S. (2014, marzo). *A new perspective to data processing: Big Data*. Conferencia presentada en International Conference on Computing for Sustainable Global Development, New Delhi, India. doi: 10.1109/IndiaCom.2014.6828111

McKinney, W. (2013). *Python for data analysis: Data wrangling with Pandas, Numpy, and IPython*. Recuperado de http://opencarts.org/sachlaptrinh/pdf/28232.pdf

McKinney, W. y PyData development team. (2016). *Pandas: powerful Python data analysis tollkit*. Recuperado de http://pandas.pydata.org/pandas-docs/stable/pandas.pdf

Micenková, B., McWilliams, B. y Assent, I. (2014, agosto). *Learning outlier ensembles: The best of both worlds – Supervised and unsupervised*. Conferencia presentada en ODD2 ’14, New York, USA. Recuperado de https://www.inf.ethz.ch/personal/mcbrian/pdfs/odd.pdf

Moreno García, M. N., Miguel Quintales, L. A., García Peñalvo, F. J. y Polo Martín, M. J. (2001, noviembre). *Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requisitos de software*. Conferencia realizada en ADIS 2001 (Apoyo a la decisión en ingeniería del software), Almagro, Ciudad Real, España. Recuperado de goo.gl/Sgptm8

Moujahid, A., Inza, I. y Larrañaga, P. (2008). *Clasificadores K-NN*. Recuperado de http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/t9knn.pdf

Moyano, J., Gibaja, E. y Ventura, S. (2017). MLDA: A too for analyzing multi-label dataset. *Elsevier Science Publishers B. V., 121*(C), 1-3. doi: 10.1016/j.knosys.2017.01.018

Murphree, J. (2016, septiembre). *Machine learning anomaly detection in large systems*. Conferencia presentada en IEEE Autotestcon, Anaheim, CA, USA. doi:10.1109/AUTEST.2016.7589589

Murray, P. (2008). Open data in science. *Serials Review, 34*(1), 52-64. doi**:**10.1080/00987913.2008.10765152

Nguyen Cong, B., Rivero Pérez, J. L. y Morell, C. (2015). Aprendizaje supervisado de funciones de distancia: estado del arte. *Revista Cubana de Ciencias Informáticas*, *9*(2), 14-28.

Nilsson, N. (2001). *Inteligencia artificial: una nueva síntesis*. Madrid. McGraw-Hill.

Open Data Barometer. (2017). *Middle East and North Africa*. Recuperado de http://opendatabarometer.org/4thedition/regional-snapshot/middle-east-north-africa/

Organización de las Naciones Unidas. (2017). *Noticias por regiones: Oriente Medio.* Recuperado de http://www.un.org/spanish/News/region.asp?regioncode=ME

Panos, L. y Christof, E. (2013). Embedded analytics and statistics for Big Data, *IEEE Software*, *30*(6), 33-39. doi:10.1109/MS.2013.125

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit – learn: Machine learning in python. The *Journal of Machine Learning Research, 12,* 2825-2830.

Phethean, C., Simperl, E., Tiropaniss, T., Tinati, R. y Hall, W. (2016). The role of data science in web science. *IEEE Intelligent Systems, 31*(3), 102-107. doi:10.1109/MIS.2016.54

Pratt, R. E., Smatt, C. T. y Wynn, D. E. (2017). Tourism through Travel Club: A Database project. *Information Systems Education Journal*, *15*(2), 40-47.

Qiao, F. y Wang, H. (2015, octubre). *Computational approach to detecting and predicting occupy protest events*. Trabajo presentado en la International Conference on Identification, Information, and Knowledge in the Internet of Thing, Beijing, China. doi 10.1109/IIKI.2015.28.

Real Academia Española. (2017). Diccionario de la lengua española. Madrid: Autor.

Rollins, J. (2015). Foundational methodology for data science. *IBM Analytics IBM*. Recuperado de https://public.dhe.ibm.com/common/ssi/ecm/im/en/imw14824usen/IMW14824USEN.PDF

Sánchez Montañés, M., Lago, L. y González, A. (2011). *Métodos avanzados en aprendizaje artificial: teoría y aplicaciones a problemas de predicción* [diapositivas de PowerPoint]. Recuperado de http://arantxa.ii.uam.es/~msanchez/docencia/maaa/transparencias/Intro_1112.pdf

Schrodt, P. (2012). *Conflict and mediation event observations event and actor codebook (CAMEO*). Recuperado de http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

Schroeck, M., Shockley, R., Smart, J., Romero Morales, D. y Tufano, P. (2012) Analytics: el uso de big data en el mundo real. Cómo las empresas más innovadoras extraen valor de datos inciertos. *IBM Global Business Services*. Recuperado de https://www-05.ibm.com/services/es/gbs/consulting/pdf/El_uso_de_Big_Data_en_el_mundo_real.pdf

Scikit-Learn. (2017a). *Accuracy score*. Recuperado de goo.gl/ZumKR5

Scikit-Learn. (2017b). N*earest centroid classification*. Recuperado de http://scikit-learn.org/stable/auto_examples/neighbors/plot_nearest_centroid.html#sphx-glr-auto-examples-neighbors-plot-nearest-centroid-py

Scikit-Learn. (2017c). *Plot the decision Surface of a decision tree on the iris dataset*. Recuperado de http://scikit-learn.org/stable/auto_examples/tree/plot_iris.html#sphx-glr-auto-examples-tree-plot-iris-py

Scikit-Learn. (2017d). *Precision-recall*. Recuperado de goo.gl/bA34rL

Seventh-day Adventist Church. (2014). *Annual statistical report: 150th report of the General Conference of Seventh-day Adventists for 2012 and 2013.* Silver Spring, MD: ASTR.

Shamsuddin, S. y Hasan, S. (2015, octubre). *Data science vs big data @ UTM big data centre.* Trabajo presentado en la International Conference on Science in Information Technology (ICSITech). doi: 10.1109/ICSITech.2015.7407766

Shazadi, K., Petrovski, S., Roten, A., Miller, H., Huggins, R. M., Brodie, M. J.,...Sills, G. J. (2014). Validation of a multigenic model to predict seizure control in newly treated epilepsy. *Epilepsy Research*, *108*(10), 1797-1805. doi:10.1016/j.eplepsyres.2014.08.022

Su, Y., Lan, Z., Lin, Y., Comfort, L. y Joshi, J. (2016, noviembre). *Tracking disaster response and relief efforts following the 2015 Nepal earthquake*. Conferencia presentada en la 2nd International Conference on Collaboration and Internet Computing (CIC), Pittsburgh, PA, USA. doi: 10.1109/CIC.2016.075

Technology of computing (2016). *A short introduction - logistic regression algorithm*. Recuperado de https://helloacm.com/a-short-introduction-logistic-regression-algorithm/

Turkson, R., Baagyere, E. y Wenya, G. (2016, septiembre). *A machine learning approach for predicting bank credit worthiness*. Conferencia presentada en la International Conference on Artificial Intelligence and Pattern Recognition (AIPR), Lodz, Poland. doi:10.1109/ICAIPR.2016.7585216

United Nations High Commissioner for Refugees. (2017). *Iraq: CCCM - IDP Population in temporary settlements by district.* Recuperado de http://www.refworld.org/docid/5977453c4.html.

Valiant, G. y Valiant, P. (2017). Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM, 64*(6), 1-41. doi:10.1145/3125643

Vostrovsky, V., Tyrychr, J. y Ulman, M. (2015). Potential of open data in the agricultural eGovernment. *Agriso on-line Papers in Economics and Informatics, 7(*2), 103-113.

Vrbić, R. (2012). Data mining and cloud computing. *Journal of Information Technology and Applications, 2*(2), 75-87. doi:10.7251/JIT1202075V

Waller, M. y Fawcett, E., (2013). Click here for a data scientist: Big Data, predictive analytics, and theory development in the era of a maker movement supply chain. *Journal of Business Logistics, 34*(4), 249-252. doi:10.1111/jbl.12024

Wei-Yin, L. (2011). Classification and regression trees. *Wires, Data Mining and Knowledge Discovery*, *1*(1), 14-23. doi:10.1002/widm.8

White, E. (1905). *El ministerio de curación*. Mountain View, CA: Pacific Press.

Wu, X., Kumar, V., Quinlan, R., Ghosh J., Yang, Q., Motoda, H., … Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1-37. doi:10.1007/s10115-007-0114-2

Yonamine, J. (2012). *Predicting future levels of violence in Afghanistan districts using GDELT*. Recuperado de http://data.gdeltproject.org/documentation/Predicting-Future-Levels-of-Violence-in-Afghanistan-Districts-using-GDELT.pdf

Zavala, B. y Alférez, G. (2015, julio). Proactive control of traffic in smart cities. Conferencia presentada en el congreso de Inteligencia Artificial, Montemorelos, Nuevo León, México.