Análisis y aplicaciones del Big Data en bases de datos abiertas

Analysis and applications of Big Data in open databases

Christian Leal García

Resumen— En este proyecto vamos a tratar de mostrar las ventajas y aplicacions que nos ofrecen las técnicas y herramientas utilizadas en los procesos Big Data. Para ello, plasmaremos un marco teórico lógico donde definiremos términos, técnicas y herramientas relacionadas con el ámbito de Big Data para familiarizarnos con el entorno. Una vez realizada la parte teórica, seleccionaremos dos herramientas de procesado de datos, las cuales son RapidMiner y Tableau, para realizar un estudio comparativo entre estas con una serie de tests que nos ayudarán a seleccionar el programario que más encaje con las necesidades que se plantean en una situación ficticia que crearémos para definir los parámetros a comparar. Al finalizar el desarrollo de las pruebas, expondremos los resultados obtenidos en la comparativa de las herramientas de forma detallada y plasmaremos las conclusiones obtenidas al finalizar este proyecto en relación a la utilidad y creación de valor que nos ofrece Big Data en bases de datos abiertas hoy en día.

Palabras clave— Big Data, Mineria de datos, RapidMiner, Base de datos, Tableau, Hadoop, Valor empresarial, Analítica de datos, Open data.

Abstract—In this project we will try to show the advantages and applications offered by the techniques and tools used in Big Data processes. To do this, we will explain a logical theoretical framework where we will define terms, techniques and tools related to the Big Data field to familiarize ourselves with the environment. Once the theoretical part is completed, we will select two data processing tools, which are RapidMiner and Tableau, to carry out a comparative study between them with a series of tests that will help us to select the program that best suits the needs that arise in a fictitious situation that we will create to define the parameters to be compared. At the end of the development of the tests, we will present the results obtained in the comparison of the tools in detail and we will capture the conclusions obtained at the end of this project in relation to the utility and value creation offered by Big Data in open databases nowadays.

Index Terms— Big Data, Data Mining, RapidMiner, Database, Tableau, Hadoop, Business Value, Data Analytics, Open data.

1 Introducción

La información es un don del cual todos nos beneficiamos hoy en día, desde un detective buscando pruebas que incriminen a su acusado hasta una familia buscando las mejores ofertas por la web para irse de vacaciones.

En un mundo que vive en constante revolución tecnológica, almacenamos todo tipo de información en bases de datos virtuales, y estas bases necesitan de técnicas y herramientas de análisis para obtener la mayor eficiencia de búsqueda, rapidez y adecuación.

Todo este conjunto de datos masivos, procesos y análisis se denomina Big Data, una corriente creciente de infinitas posibilidades tanto en el mundo personal como en el empresarial, donde quien domina este ámbito tiene una cierta ventaja informativa contra todo opositor.

- *E-mail de contacto: Christian.Leal@e-campus.uab.cat*
- *Mención realizada: Tecnologies de la Informació.*
- *Trabajo tutorizado por: Ramón Musach Pi (Ingenieria de la Información y de las Comunicaciones)*
- *Curso 2019/20*

En este proyecto veremos diversas técnicas y herramientas analíticas que se utilizan sobre bases de datos públicas para extraer todo tipo de información deseada ajustando los parámetros para obtener el mejor resultado.

Tras estas pruebas, analizaremos las posibles aplicaciones que nos ofrece una búsqueda eficiente para añadir valor de conocimiento a cualquier tipo de dato.

2 Objetivos

2.1 Objetivos Generales

Mostrar las ventajas, aplicaciones y valor que nos proporciona un buen análisis de datos a gran escala con la utilización de herramientas óptimas para ello sobre bases de datos abiertas.

2.2 Objetivos Específicos

- Analizar y estudiar las herramientas más utilizadas en la analítica de Big Data.
- Observar distintas técnicas para el análisis de datos en Big Data.

- Creación de una situación ficticia a partir de bases de datos abiertas donde se detallarán los objetivos y parámetros a observar para la realización de una comparativa entre varias herramientas de procesado de datos.

- Detallar y comparar los resultados obtenidos tras los análisis realizados y seleccionar la herramienta más adecuada para nuestro caso práctico.

- Concretar y razonar las diversas aplicaciones que tienen estos análisis sobre bases de datos abiertas para añadir valor de conocimiento, eficiente y sofisticado, a cualquier tipo de búsqueda.

3 ESTADO DEL ARTE

En este apartado vamos a comentar varios artículos, nacionales e internacionales, que nos ayudaron a comprender el entorno tecnológico en el que se encuentra la tecnología Big Data actualmente para así poder enfocar nuestro proyecto de una forma más eficiente y coherente.

3.1 Investigaciones nacionales

El primer artículo a tratar está relacionado con las posibilidades que nos ofrece el uso del Big Data en las organizaciones redactado por el autor David López García. [1]

El objetivo que pretende alcanzar el autor es la explicación de términos relacionados con Big Data, tales como Data Mining o Cloud Computing, e intentar averiguar si esta tecnología perdurará en el futuro y como podrían las empresas obtener ventaja ante sus competidores con estos procesos.

Después de describir el marco teórico y detallar las diversas herramientas que son utilizadas en los procesos Big Data, el autor concluye su obra obteniendo como reflexión que la tecnología Big Data no solo es útil para obtener grandes cantidades de datos, sino que también para analizarlos.

También puntualiza al final que lo que hoy son grandes cantidades de datos, en un futuro se convertirán en ínfimas, es por eso que la tecnología Big Data debe seguir evolucionando de la mano de las organizaciones públicas y privadas.

3.2 Investigaciones internacionales

El segundo artículo a comentar está relacionado con el uso de Big Data en Business Intelligence y desarrollado por el autor Samuel Israel Goyzueta Rivera. [2]

El objetivo que pretende alcanzar el autor es la muestra y correlatividad de los términos Big Data, Business Analytics y Big Data Marketing ofreciéndonos los distintos procesos a seguir en el uso de estas tecnologías y las capacidades que debemos tener en cuenta como requisitos para poder aplicarlos.

Este artículo nos muestra que las oportunidades de construir una cultura de innovación, la confianza de establecerla entre los consumidores, la escalabilidad y eficiencia de la plataforma con la que se trabaja, la organización y participación de los socios y una cultura de apoyo entre ellos, son la base imprescidible para implementar los procesos Big Data Marketing en una empresa.

Concluyendo el planteamiento inicial, la posibilidad de procesar grandes cantidades de información para lograr un mayor entendimiento de los gustos, deseos y necesidades del consumidor es una gran ventaja que no podemos dejar escapar si nos dedicamos al ámbito del marketing.

4 METODOLOGÍA

En el proyecto hemos elegido una metodología de trabajo tipo cascada para la realización de este de una forma versátil y cómoda y consta de las siguientes partes:

4.1 Identificación

Definir y especificar los límites del trabajo a realizar con consideraciones lógicas y coherentes hasta llegar al objetivo final de definir las diversas aplicaciones del Big Data.

4.2 Recopilar información

Reunir información relacionada con el Big Data. Centrándonos y profundizando en las diversas técnicas aplicables al análisis de datos y también en las diversas herramientas disponibles para la realización de estos.

4.3 Organización y análisis

Selección y estructuración de la información recogida previamente para el uso más eficiente de esta en nuestro favor.

4.4 Descripción

Detallar los tipos de técnicas y herramientas utilizadas para el análisis de datos, así como la elección de las bases de datos libres que existen, como por ejemplo las bbdd de las redes sociales.

- Posibles técnicas: Análisis descriptivo, de diagnóstico, predictivo, prescriptivo…

- Posibles herramientas: RapidMiner, Weka, Tableau, R, MapReduce….

4.5 Análisis

Selección de las herramientas deseadas para los análisis de datos y realización de estos en bases de datos libres utilizando diversas técnicas y eligiendo unos parámetros comparativos para obtener mejores resultados.

4.6 Comparativa

Pequeña comparativa sobre los datos obtenidos de los análisis entre las diversas herramientas. Para ello debemos realizar un análisis de requerimientos y así poder extraer la información más relevante para esta comparativa.

4.7 Razonamiento

Especificar, detallar y debatir las diversas aplicaciones que tienen el uso y análisis de Big Data en cualquier tipo de ámbito, como el individual, el corporativo o el empresarial.

5 MARCO TEÓRICO

En este apartado vamos a dar un pequeño repaso a todos los términos y tecnología involucrados en el proceso del análisis de Big Data para tener una visión más clara a la hora de la futura toma de decisiones y comprender el léxico utilizado en este trabajo:

5.1 Términos

- **Datos**: Son un el conjunto básico de hechos referentes a una persona, cosa o transacción. Un dato nos permite describir un objeto y dicho objeto podemos llamarlo entidad. Hay diferentes tipos de datos que se pueden tener en una base de datos, como caracteres, numéricos, imágenes, fechas, monedas, texto, bit, decimales y varchar.[3]

- **Tipos de datos:** Los tipos de datos se clasifican en estructurados, no estructurados y semi-estructurados:

  - Estructurados: Los datos estructurados tienen perfectamente definido la longitud, el formato y el tamaño de sus datos. Se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales.

  - No estructurados: Los datos no estructurados se caracterizan por no tener un formato específico. Se almacenan en múltiples formatos como documentos PDF o Word, correos electrónicos, ficheros multimedia de imagen, audio o video…

  - Semi-estructurados: Los datos estructurados son una mezcla de los dos anteriores no presenta una estructura perfectamente definida como los datos estructurados pero si presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones. [4]

- **Base de datos:** Una base de datos es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso. Actualmente la mayoría de las bases de datos están en formato digital y es por eso que se considera un componente electrónico más, por tanto se ha de tener en cuenta. [5]

  Hay programas denominados sistemas gestores de bases de datos (SGBD) que permiten almacenar y posteriormente acceder a los datos de forma rápida y estructurada. [6]

- **Big Data:** En definitiva, cuando hablamos de Big Data no nos referimos únicamente a los datos, sino sobre todo a la capacidad de poderlos explotar para extraer información y conocimiento de valor para nuestro negocio. [7] Una vez recogida y almacenada la información, se deben extraer indicadores que puedan ser útiles para tomar decisiones, incluso en tiempo real.

  Podríamos definir el Big Data con las cinco "Vs":

  - Volumen: Como hemos visto, la cantidad de datos se define "Big" no cuando supera un tamaño definido, sino cuando su almacenamiento, procesamiento y explotación empieza a ser un reto.

  - Velocidad: La segunda característica del Big Data está relacionada con el ritmo a los cuales los datos se están generando, que suele aumentar constantemente y que necesita una respuesta en tiempo real por parte de las empresas.

  - Variedad: El reto principal del Big Data reside en la gran diferencia de formatos distintos en los cuales encontramos los datos y que pueden ir desde texto sencillo, a imágenes, videos, hojas de cálculos y enteras bases de datos.

  - Veracidad: Los datos tienen que ser confiables y han que ser mantenidos limpios. Una gran cantidad de datos no tiene valor si son incorrectos y puede ser altamente perjudicial, sobre todo en la toma de decisión automatizada.

  - Valor: Los datos y su análisis tienen que generar un beneficio para las empresas. [8]

5.2 Técnicas

- **Data Mining:** Es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos de manera automática o semiautomática con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto. [9]

Básicamente, el datamining surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos a la obtención de conocimiento. [10]

Procesos a realizar para una minería de datos eficiente:

- Determinación de los objetivos: Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.

- Preprocesamiento de los datos: Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos.

- Determinación del modelo: Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial. [11]

5.3 Herramientas

Hay dos tipos de herramientas para el análisis del Big Data, las privadas y las open source. Nosotros nos centraremos en las open source y realizaremos una descripción de las más conocidas y escogeremos dos para realizar un análisis de resultados y una comparativa. [7]

- **Weka:** La herramienta WEKA es una de las herramientas de código abierto de minería de datos más populares desarrollados en la Universidad de Waikato en Nueva Zelanda en 1992. Se trata de una herramienta basada en Java y se puede utilizar para implementar varios algoritmos de Machine Learning y minería de datos escritos en Java. [12]

  WEKA proporciona tanto, una GUI y CLI para realizar minería de datos y hace un buen trabajo de proporcionar apoyo a todas las tareas de minería de datos. WEKA soporta una variedad de formatos de datos como CSV, ARFF y binarios. También se centra en la representación textual de los datos en lugar de la visualización aunque proporciona soporte para mostrar algunos de visualización. [13]

- **Hadoop:** Es una infraestructura digital de desarrollo creada en código abierto bajo licencia Apache. Hadoop gestiona una gran cantidad de datos en petabytes, lo que hace posible que el funcionamiento de sus aplicaciones sea el adecuado. [14]

  Detecta los fallos y vuelve a ejecutar la instrucción, probando con otros caminos de nodos, para que los resultados que se obtienen no sean inconsistentes. Permite desarrollar tareas con datos de cantidad masiva, dividiéndolas en partes y distribuyéndolas a un conjunto de máquinas.

  El análisis realizado es en petabytes de datos, en entornos distribuidos formados por varias máquinas sencillas. En la actualidad, la tecnología avanza rápidamente, por esta razón trabajar de esta forma es una manera razonable, debido a la gran cantidad de información que se maneja y son utilizadas en grandes empresas como son Yahoo, Google, Twitter o Facebook. El núcleo de Hadoop está formado por dos componentes el sistema de ficheros distribuido HDFS7 y el modelo de procesamiento MapReduce.

- **Microsoft Azure:** Microsoft Azure, es una plataforma que se encuentra en la nube de Microsoft, actúa como SaaS teniendo la ventaja de nube híbrida, es decir, se puede tener todos los centros de datos y la nube pública al mismo tiempo, de esta manera, actúa de forma inmediata. Azure nos da la accesibilidad de crear servicios y aplicaciones con cualquier dispositivo.

  Otra característica a rescatar de esta herramienta, es que obtiene la libertad de crear e implementar donde quiera, utilizando las herramientas, las aplicaciones y los marcos que prefiera, así como también adaptarse a cualquier demanda, es decir, solo se paga por lo que se usa. [15]

- **Tableau:** Tableau es una poderosa herramienta de visualización de datos utilizados en la inteligencia de negocio y análisis de datos. Esta fué inventada en enero de 2003 y proporciona una visualización mejorada por completo y la posibilidad de obtener más conocimientos sobre los datos que estamos trabajando, también se puede utilizar para proporcionar predicciones más precisas.

  Este producto utiliza bases de datos relacionales, cubos OLAP, bases de datos en la nube y hojas de cálculo y luego genera un número de tipos de gráficos que se pueden mostrar en dashboards que se pueden compartir con seguridad sobre una Internet. [16]

  Tableau ha hecho posible explorar y presentar los datos de una manera mucho más simple y práctica. Trabajar en proyectos que usan Tableau consume menos tiempo y es fácil de manejar.

- **Oracle Big Data Appliance:** Oracle Big Data Appliance es una solución de infraestructura que servirá tanto a empresas que poseen grandes proyectos de datos como a organizaciones que aumentan sus necesidades a lo largo del tiempo.

  Se centra en la idea de que puede ampliar su actual arquitectura de información empresarial para incorporar datos grandes.

  Es un sistema que combina el hardware con una pila de datos en el software para ofrecer una solución completa y fácil de implementar, para adquirir y organizar gran cantidad de datos. Posee una configuración de rack completo con 18 servidores, con una capacidad total de almacenamiento de 648TB. [17]

- **R:** Es un lenguaje de programación y entorno de software para cálculo estadístico y gráficos. El lenguaje R es de los más usados por los estadistas y otros profesionales interesados en la minería de datos y las matemáticas financieras.

  Se parece más al lenguaje de las matemáticas que a otros lenguajes de programación, lo que puede ser un inconveniente para los programadores a la hora de elegir programar en R para temas de Big Data. [18]

  Lo que está claro es que si eliges usar R podrás disponer de una gran cantidad de librerías creadas por la comunidad de R y otras tantas herramientas de altísima calidad. [19]

- **MongoDB:** MongoDB es una base de datos orientada a documentos (guarda los datos en documentos, no en registros). Estos documentos son almacenados en BSON, que es una representación binaria de JSON.

  A pesar de que las bases de datos NoSQL no tienen una extensa variedad de uso, MongoDB tiene un ámbito de aplicación más amplio en diferentes tipos de proyectos. És especialmente útil en entornos que requieran escalabilidad y con sus opciones de replicación y sharding, podemos conseguir un sistema que escale horizontalmente sin demasiados problemas. [20]

- **RapidMiner:** RapidMiner es un programa informático, escrito en Java, para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico.

  Esta herramienta proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización. [21]

  También nos ofrece el 99% de una solución analítica avanzada a través de marcos basados en plantillas que aceleran la entrega y reducen los errores casi eliminando la necesidad de escribir código.

  RapidMiner proporciona una GUI para diseñar y ejecutar flujos de trabajo analíticos. Esos flujos de trabajo se denominan "Procesos" y consisten en múltiples "Operadores". Cada operador realiza una única tarea dentro del proceso, y la salida de cada operador forma la entrada de la siguiente.

  La funcionalidad RapidMiner se puede ampliar con complementos adicionales que están disponibles a través de RapidMiner Marketplace. RapidMiner Marketplace proporciona una plataforma para que los desarrolladores creen algoritmos de análisis de datos y los publiquen en la comunidad.

  La herramienta cuenta con dos componentes:

  - RapidMiner Studio: Versión standard para analistas que implementa todos los operadores de data mining, modelos predictivos, modelos descriptivos, transformación de datos, series de tiempo, etc.

  - RapidMiner Server: Versión Servidor de RapidMiner que nos permite trabajo colaborativo, escalable y concurrente múltiples usuarios, capacidad de delegar en bases de datos y otras mejoras de funcionalidad. [22]

6 MARCO PRÁCTICO

Debido a que el análisis de datos abarca muchos ámbitos y que el hecho de escoger una técnica y definir unos parámetros para minar bases de datos está sujeta directamente a la necesidad del cliente o entidad demandante, en este proyecto hemos decidido ambientar el desarrollo de las pruebas de la siguiente forma:

Pertenecemos a un equipo de marketing de cierta empresa que nos demanda un estudio concreto sobre los diversos accidentes mortales o graves en Catalunya para crear una campaña de concienciación en las carreteras de cara al año que viene. Con tal de realizar este nuevo proyecto de la forma más eficiente, hemos decidido realizar varios análisis con los softwares dedicados a la minería de datos: **RapidMiner** y **Tableau**.

Sabemos que una de las mejores opciones podría ser Hadoop dado que es una de las herramientas más utilizadas y que aporta una gran variedad de posibilidades y opciones en el entorno Big Data. Pero las herramientas que hemos seleccionado, rigiéndonos por nuestra situación fictícia, nos pueden aportar otro tipo de ventajas y características más interesantes en comparación

con Hadoop, por esto, hemos decidido no seleccionarlo como candidato para la realización de estas pruebas.
Una vez hayamos realizado los test, procederemos a comparar los resultados obtenidos por ambos programas y decidirnos por uno para realizar el proyecto.

6.1 Determinación de parámetros

Dada la naturaleza de nuestro proyecto hemos definido una serie de parámetros a observar y analizar para obtener un mejor resultado a la hora de crear la campaña:

- Año {2010-2018}
- Zona {Zona Urbana, Carretera}
- Vía {Todas las vías}
- Municipio {Todos los municipios}
- Comarca {Todas las comarcas}
- Provincia {Barcelona, Lleida, Girona, Tarragona}
- Número de Muertos {0-100}
- Número Heridos {0-100}
- Tipo de vehículo {Coche, Motocicleta, Ciclomotor}
- Climatología {Día soleado, Día lluvioso, Noche despejada, Noche lluviosa, Tarde soleada, Tarde lluviosa}
- Estado de la superficie {Seca, Mojada}
- Fecha y Hora {dd/mm/yyyy}
- Tipo de Accidente {En marcha, Atropello, Estacionado, Colisión frontal, lateral}

7 RESULTADOS

7.1 Resultado de las prestaciones

Cada herramienta de minería de datos y los datos de visualización tiene sus propias características especiales y también inconvenientes. Es por eso que, en este apartado del proyecto, vamos a definir una serie de características que debemos comparar para obtener el software que más se nos adecúe:

- **Usabilidad**: Esta característica determina la facilidad de uso de cada herramienta. Esto describe que la interfaz de usuario es comparativamente más fácil de usar.

- **Velocidad**: La velocidad es un factor importante que distingue entre las diferentes herramientas de minería de datos. Esta característica ayuda a entender cómo la configuración del sistema impacta el trabajo de una herramienta de minería de datos en particular.

- **Visualización**: La visualización es la característica más importante de una herramienta de minería de datos. Esta característica comparativa distingue a cada herramienta de minería de datos en base a diferentes opciones de visualización proporcionadas.

- **Algoritmos soportados:** Esta función clasifica las herramientas de minería de datos basados en la implementación del algoritmo con el apoyo de ellos y la elección de selección descriptor disponible.

- **Tamaño de conjunto de datos:** Datos pequeños o más grandes del soporte de conjuntos es otra de las características comparables entre diferentes herramientas de minería de datos.

- **Uso de memoria:** A medida que el uso de memoria afecta al rendimiento, el uso de memoria es otra característica importante para comparar herramientas de minería de datos.

- **Uso principal:** Cada herramienta de minería de datos tiene un uso particular que es una de las características comparables. Por ejemplo, tanto R y WEKA se pueden utilizar para implementar algoritmos de minería de datos, pero el uso principal de R está en el cálculo estadístico.

- **Tipo de interfaz soportado:** Desde el contexto de este estudio comparativo, el uso de la interfaz gráfica de usuario (GUI) o la línea de comandos de la interfaz (CLI) que diferencia a cada herramienta.

	RapidMiner	Tableau
Usabilidad	Fácil de usar	Muy fácil de usar
Velocidad	Requiere más memoria para operar	Funciona más rápido en cualquier máquina
Visualización	Varias opciones	Muchas opciones de visualización
Algoritmos soportados	Clasificación y Clustering	No implementa algoritmos
Tamaño de conjunto de datos	Soporta grandes y pequeños tamaños	Compatible con cualquier tamaño
Uso de memoria	Requiere más memoria	Menos memoria por lo tanto funciona más rápido
Uso principal	Minería de datos y Análisis predictivo	Business Intelligence
Tipo de interfaz soportado	GUI	GUI

7.2 Resultado de la prueba realizada

Gracias a la realización de las pruebas realizadas en las herramientas Big Data tales como RapidMiner y Tableau y del análisis detallado de las prestaciones que estas nos ofrecen, desde el equipo de marketing, hemos decidido escoger el software de código libre RapidMiner por los siguientes motivos:

- **Interface:** Poniendo en discusión la interface de estas dos herramientas, RapidMiner nos ofrece una versión más moderna y user-friendly a la hora de cargar los datos a procesar, brindándonos la posibilidad y asegurando la mayor cobertura de compatibilidad de formatos para este proceso. En cambio, con Tableau tuvimos varios problemas de compatibilidad a la hora de cargar los datos desde el mismo archivo csv, es por eso que realizamos varias modificaciones para que fuese posible la integración de estos.

- **Muestra estadística:** Centrándonos ahora en la demostración de poderío estadístico de estas herramientas, RapidMiner es posiblemente una de las mejores de código libre que podemos escoger. El diseño al mostrarnos los datos es simplemente impecable, solo tenemos que procesar los datos de una forma genérica y este se encarga de mostrarnos las estadísticas prototípicas más interesantes para cada atributo y la capacidad de detallar más esta información con un solo clic, con lo que podríamos decir que está al alcance de cualquier usuario inexperto en este ámbito.

  Por otro lado, Tableau nos brinda también una amplia gama de posibilidades estadísticas pero con la contraposición de que el usuario que vaya a utilizarlo debe tener unos conocimientos más específicos para la realización de esta tarea.

- **Velocidad:** Entrando en el tema de velocidad de procesado de datos, RapidMiner necesitó 0.5 segundos para cargar el archivo .csv y 0.7 segundos para crear todo tipo de estadísticas y gráficos sin necesidad de peticionarlos. Por otra parte, Tableau requirió de 1.5 segundos en cargar los datos del archivo .csv, y una media de 1 segundo cada vez que se requiere la representación gráfica de los atributos de este, ya que dispone de muchas posibilidades de visualización y sería prácticamente imposible cargarlas todas.

- **Uso de memoria:** Debido a que RapidMiner es una herramienta dedicada exclusivamente a los procesos Big Data, el uso de memoria que necesita para cargar todos sus módulos y prototipos es más elevado que el uso de Tableau, que podría utilizarse solamente como visualizador de datos.

- **Representaciones gráficas:** Por último, pero no menos importante, analizando y comparando las representaciones gráficas que nos ofrecen estas dos herramientas, podemos afirmar que Tableau ofrece un gran abanico de posibilidades de mostrarnos gráficamente los datos y que es una de sus principales funciones en el mundo del Big Data.

  Pero volviendo a nuestro caso práctico, rompemos una lanza a favor de RapidMiner, que es capaz de representar gráficamente la información procesada de los datos de una forma muy eficiente y clara.

En el anexo de este proyecto se muestran distintas imagenes que representan el resultado obtenido de las pruebas realizadas en relación a la interface, la muestra estadística y las representaciones gráficas.

7.3 Resultado del caso práctico

Una vez elegida la herramienta RapidMiner para que nos ayude a la hora de realizar la campaña de concienciación en las carreteras de cara al año que viene, toca escoger la información procesada que nos será útil para esta.

Como hemos extraido la información previamente, pensando en enfatizar la campaña en mostrar los datos seleccionados podremos reducir el número de accidentes en las zonas con las condiciones más habituales de cara al año que viene.

Es por esto que, nuestra campaña de concienciación debería estar enfocada tanto en carreteras como en zonas urbanas de la provincia de Barcelona, insistiendo en que el peligro en nuestras vías no solo existe bajo condiciones climatológicas adversas, sino que también en días soleados.

Focalizando el grueso de estos accidentes entre semana y dejando entrever que los conductores somos el primer responsable de estos, ya que la mayoría de colisiones se realizan entre vehículos en marcha.

8 Conclusión

En el transcurso de este proyecto hemos realizado una pequeña introducción en el mundo del Big Data y la aplicación real y laboral que tienen este conjunto de técnicas y herramientas.

Como hemos observado, el hecho de que estemos en una época donde la globalización de la información es prácticamente total y en tiempo real, fuerza a las empresas y organizaciones en invertir en métodos de desarrollo e gestión de la información que va más allá del alcance de los seres humanos.

Dada esta necesidad, hemos podido analizar y estudiar diversas herramientas y distintas técnicas que nos ofrecen la posibilidad de gestionar estos procesos de una manera más sencilla.

Por este motivo, podemos decir que las diversas aplicaciones que nos ofrece el movimiento Big Data, fruto de la analítica avanzada de datos procesados, son la mejor forma de adquirir valor tanto empresarial como individual para seguir en la lucha tecnológica que estamos viviendo actualmente.

Una de las aplicaciones que hemos obtenido, al realizar las pruebas necesarias para nuestro caso práctico, es la obtención y creación de valor tanto económico como social al realizar una campaña de concienciación en las carreteras gracias al Big Data y a las bases de datos abiertas proporcionadas por el estado.

También hemos podido observar y comprobar de primera mano que la utilización de estas herramientas no nos garantiza la obtención de conocimiento útil para ofrecer valor a cualquier producto, servicio u organización, más bien debemos centrarnos en la elección de herramientas que congenien con la experiencia y el ámbito de trabajo de sus futuros usuarios, la finalidad con la que se adquiera esta misma y asegurarnos de un buen uso para crear valor en nuestra búsqueda de conocimiento.

Teniendo todos estos requisitos en cuenta, podemos afirmar con claridad que la gestión de grandes volúmenes de datos, en bases de datos abiertas, con las técnicas de Big Data óptimas para cada proceso, nos garantiza la creación de valor y nos aporta experiencia y una ventana de conocimiento de mercado en vista a un futuro que no para de evolucionar tecnológicamente.

De cara al futuro, podríamos centrar las líneas de investigación en mejorar los procesos internos, los algoritmos y los modelos prototipados de las herramientas del Big Data, para poder seguir trabajando con estas tecnologías de forma eficiente y global en bases de datos abiertas proporcionadas por todo tipo de personas u organizaciones, tanto públicos como privadas.

Agradecimientos

Quiero agradecer la realización de este proyecto a Ramón Musach Pi ya que, como tutor del mismo, me ayudó en la toma de decisiones a la hora de enfocar el trabajo y siempre mostró un trato amable y seguro en las diversas entrevistas.

También quiero agrader el apoyo mostrado por parte de mi familia, que ha sido un pilar clave en la realización del proyecto, animándome cuando he tenido problemas de logística y proporcionándome ideas y posibles planteamientos a aplicar en este proyecto para fases futuras.

También agradecerle el esfuerzo a mi compañero de equipo, experto en Bases de Datos, que me dio una pequeña introducción sobre Big Data y los objetivos que esta tecnología desea alcanzar.

Por último, agradecerle la ayuda a mis compañeros de trabajo que me aconsejaron a la hora de realizar las pruebas prácticas.

Bibliografía

[1] David López García; (2018). “Análisis de las posibilidades de uso de Big Data en las organizaciones”. Perspectivas, Universidad de Cantabria UC, Santander.

[2] Samuel Israel Goyzueta Rivera; (2015). "Big Data Marketing: una aproximación". Perspectivas, Universidad Católica Boliviana "San Pablo", Unidad Académica Regional Cochabamba.

[3] Macrodatos. Wikipedia, la enciclopedia libre, 3 de octubre de 2019. Wikipedia, https://es.wikipedia.org/Macrodatos. Accedido 1 el de octubre de 2019.

[4] «Tipos de datos: estructurados, semiestructurados y no estructurados». Diego Calvo, 21 de noviembre de 2017, http://www.diegocalvo.es/tipos-de-datos-estructurados-semiestructurados-y-no-estructurados/. Accedido 20 de octubre de 2019.

[5] «Base de datos». Wikipedia, la enciclopedia libre, 20 de octubre de 2019. Wikipedia, https://es.wikipedia.org/w/index.php?title=Base_de_datos&oldid=120706258.

[6] «Bases de datos ¿qué son? ¿Qué tipos existen? Lo que necesitas saber cómo profesional». platzi.com, https://platzi.com/blog/bases-de-datos-que-son-que-tipos-existen/. Accedido 20 de octubre de 2019.

[7] ¿Qué es big data? | Oracle España. https://www.oracle.com/es/big-data/guide/what-is-big-data.html. Accedido 1 de octubre de 2019.

[8] Técnicas de Análisis de datos en Big Data. https://www.tecnologias-informacion.com/tecnicasbigdata.html. Accedido el 5 de octubre de 2019.

[9] Datamining (Minería de datos). https://www.sinnexus.com/business_intelligence/datamining.aspx. Accedido 20 de octubre de 2019.

[10] Las técnicas de análisis de datos en la era del ‘big data’. cognodata, 30 de octubre de 2018, https://www.cognodata.com/blog/tecnicas-analisis-datos-era-big-data. Accedido el 5 de octubre de 2019.

[11] Técnicas de Análisis de Big Data – Roc Reguant. https://rocreguant.com/tecnicas-de-analisis-de-big-data/306/. Accedido 20 de octubre de 2019.

[12] «Weka». SourceForge, https://sourceforge.net/projects/weka/. Accedido Accedido 25 de noviembre de 2019.

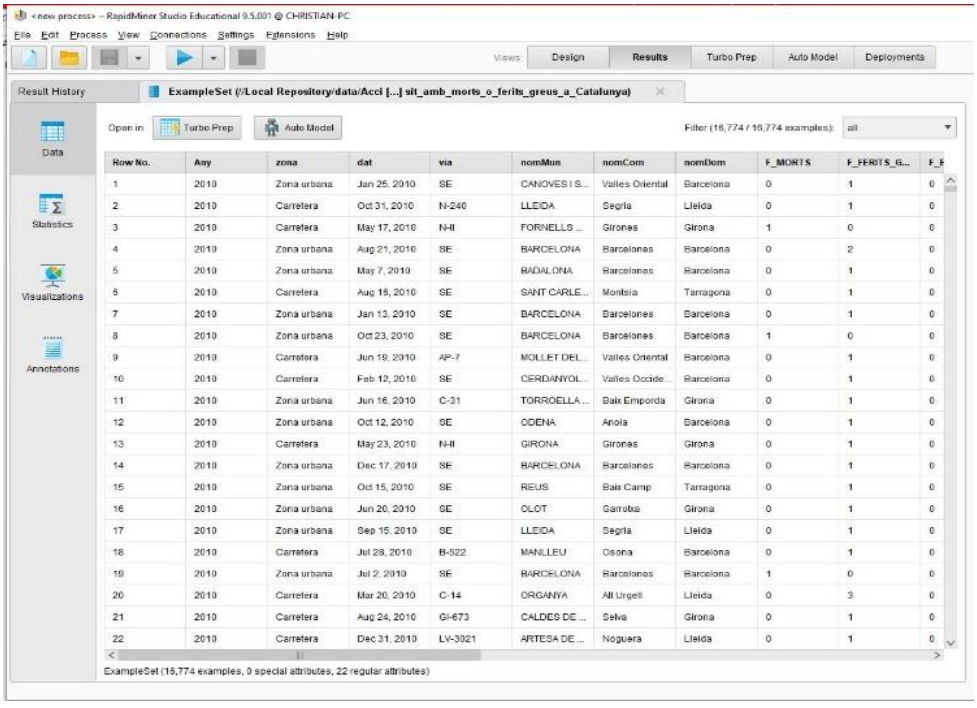
[13] «7 Herramientas Big Data para tu empresa - IIC». Instituto de Ingeniería del Conocimiento, 13 de octubre de 2016, http://www.iic.uam.es/innovacion/herramientas-big-data-para-empresa/. Accedido 20 de octubre de 2019.
[14] ¿Qué es agile analytics en Big Data? | Deusto Formación. https://www.deustoformacion.com/blog/gestion-empresas/que-es-agile-analytics-big-data. Accedido el 1 de octubre de 2019.
[15] 10 herramientas de Big Data imprescindibles para el análisis de datos. Blog de IEBSchool, 8 de marzo de 2019, https://www.iebschool.com/blog/mejores-herramientas-big-data/. Accedido el 5 de octubre de 2019.
[16] «4 departamentos dónde el Big Data Intelligence es rentable». Papeles de Inteligencia Competitiva, 17 de marzo de 2015, https://papelesdeinteligencia.com/departamentos-que-se-benefician-del-big-data-intelligence/. Accedido 25 de noviembre de 2019.
[17] «Versiones de prueba de los productos». Tableau Software, https://www.tableau.com/es-es/products/trial. Accedido 30 de diciembre de 2019.
[18] Dispositivo de Big Data | Oracle España . https://www.oracle.com/es/engineered-systems/big-data-appliance/. Accedido el 25 de noviembre de 2019.
[19] Analizar datos de twitter con R - Gender and Tech Resources. https: //gendersec.tacticaltech.org/wiki/index.php/Analizar_datos_de_twitter_con_R. Accedido 25 de octubre de 2019.
[20] «La Base de Datos Líder del Mercado Para Aplicaciones Modernas». MongoDB , https://www.mongodb.com/es. Accedido 25 de octubre de 2019.
[21] «RapidMiner». Wikipedia, 20 de octubre de 2019. Wikipedia, https://en.wikipedia.org/w/index.php?title=RapidMiner&oldid=921794576.
[22] «Qué es RAPID MINER?» prezi.com, https://prezi.com/grs49quva6m9/que-es-rapid-miner/. Accedido 20 de octubre de 2019.
[23] «Lightning Fast Data Science Platform for Teams | RapidMiner©». RapidMiner, https://rapidminer.com/. Accedido 25 de noviembre de 2019.

ANEXO

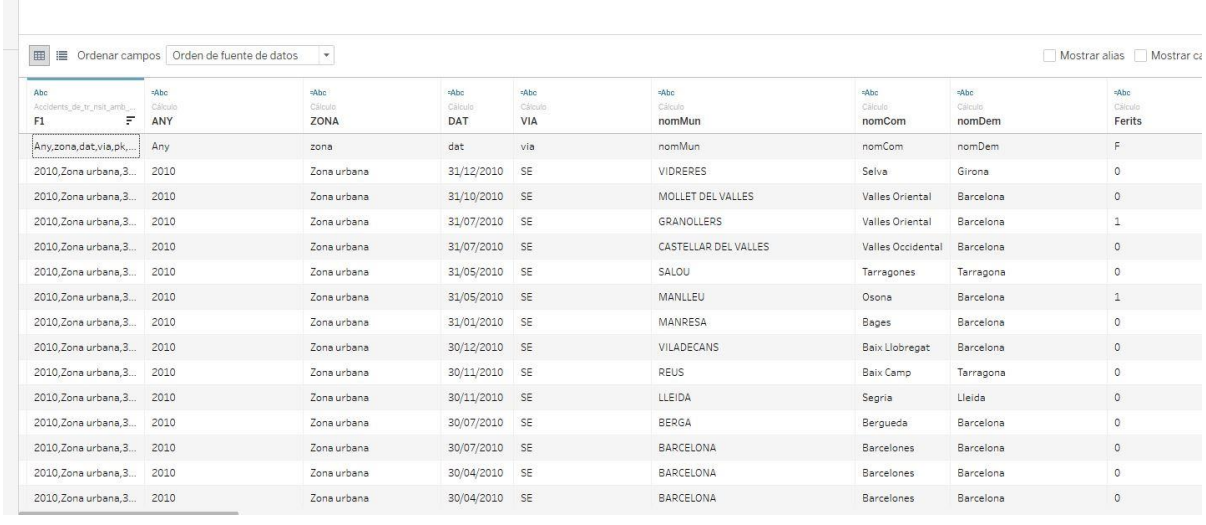
Aquí se muestran las diversas imágenes fruto de los resultados obtenidos en las pruebas comparativas entre RapidMiner y Tableau.

INTERFACE

RapidMiner

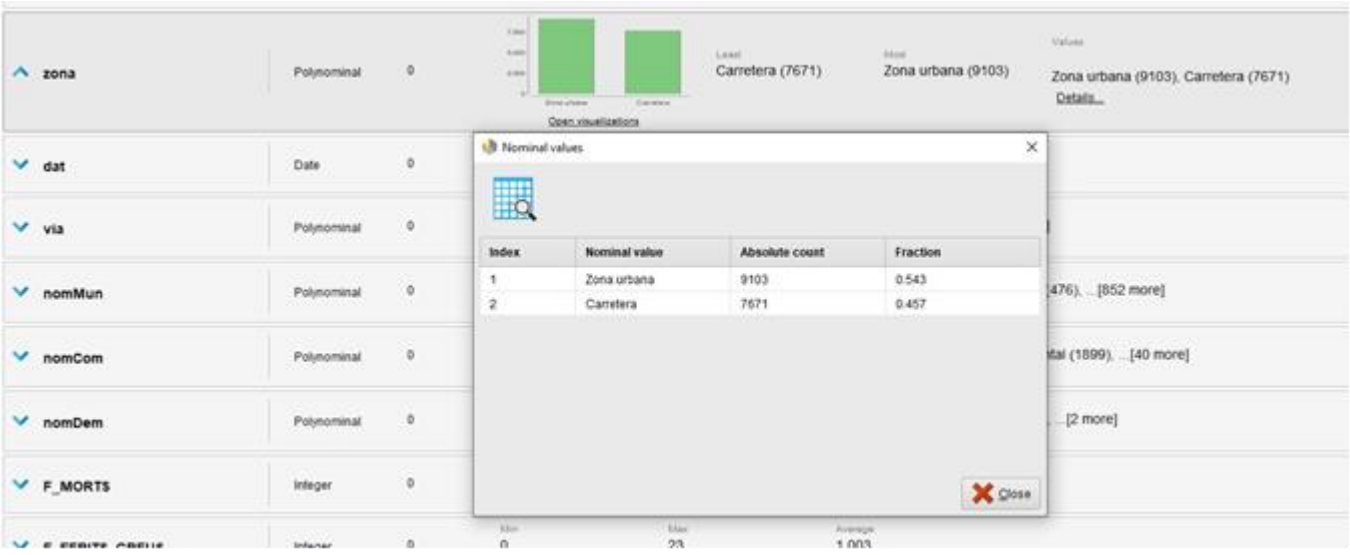


Tableau



MUESTRA ESTADÍSTICA

RapidMiner



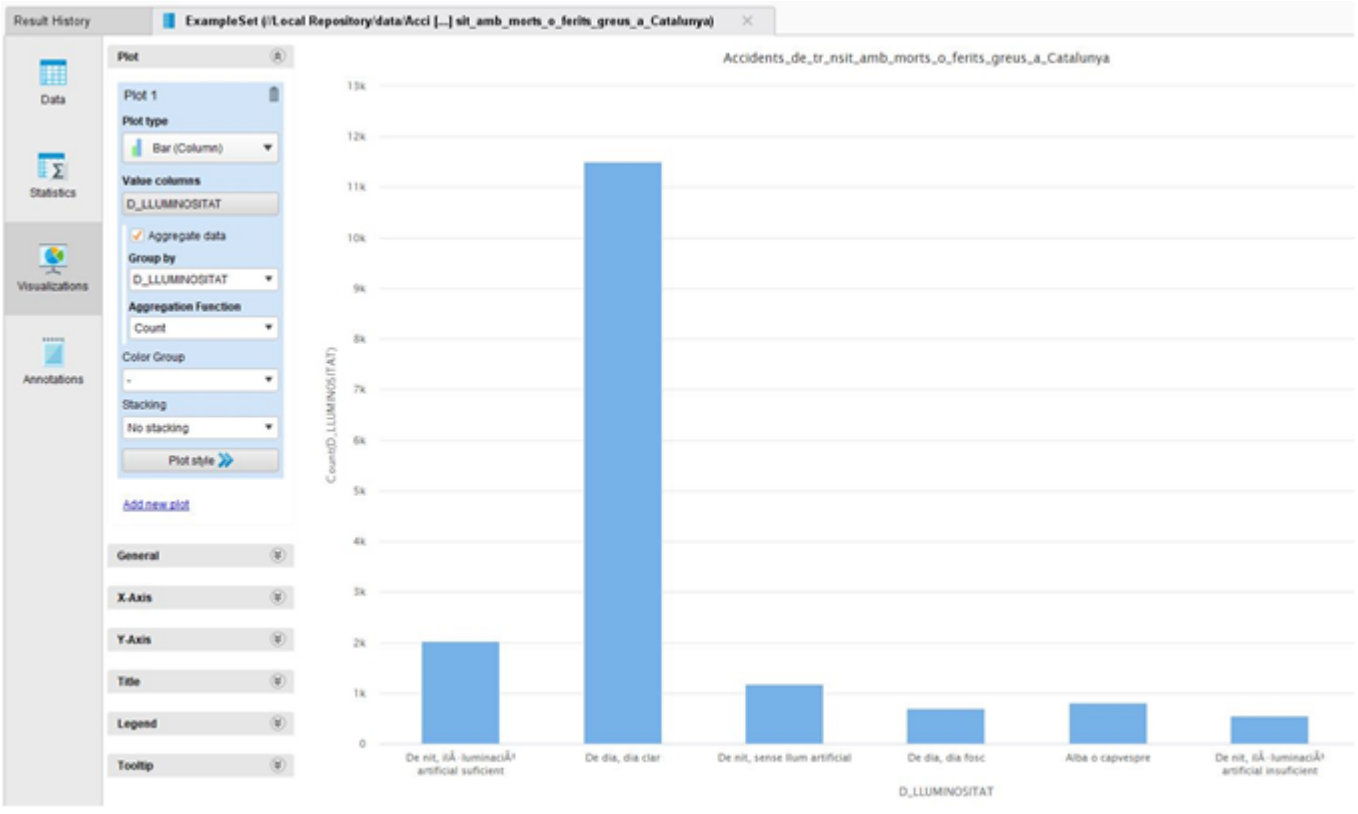
Tableau

ESTADÍSTICA

ANY	DAT	VIA	nomMun	nomCom	nomDem
2010	01/01/2010	BV-2241	PIERA	Anoia	Barcelona
		SE	ANGLES	Selva	Girona
			HOSPITALET DE LLOBREG..	Barcelones	Barcelona
			SANT FELIU DE BUIXALLEU	Selva	Girona
	01/02/2010	C-53	"FULIOLA	LA"	Urgell
		SE	AITONA	Segria	Lleida
			BARCELONA	Barcelones	Barcelona
			OLOT	Garrotxa	Girona
	01/03/2010	C-55	MANRESA	Bages	Barcelona
		C-58	MONTCADA I REIXAC	Valles Occidental	Barcelona
		N-420	PRADELL DE LA TEIXETA	Priorat	Tarragona
		SE	BARCELONA	Barcelones	Barcelona
			CERDANYOLA DEL VALLES	Valles Occidental	Barcelona
			SANT BOI DE LLOBREGAT	Baix Llobregat	Barcelona
			TORREDEMBARRA	Tarragones	Tarragona
	01/04/2010	AP-7	LLINARS DEL VALLES	Valles Oriental	Barcelona
		C-44	VANDELLOS I L'HOSPITAL..	Baix Camp	Tarragona
		SE	BARCELONA	Barcelones	Barcelona

REPRESENTACIONES GRÁFICAS

RapidMiner



Tableau