



Curso Especializado Online en Big Data Analytics

PROYECTO FINAL

Presentado por:

Aura María Arango Toro

Carlos Pérez García

Jorge Eduardo Arias Morales

12 de abril de 2021

Tabla de contenido

0. Caso de uso – Sector asegurador	3
Descripción de la compañía	4
1. Proyecto de analítica predictiva	5
Principales líneas de actuación	5
2. Metodología de trabajo	6
Fases del proyecto	7
3. Organización de la empresa.....	8
4. Gestión del gobierno del dato	9
Herramientas data governance	10
5. Fuentes de datos	11
Fuentes de terceros	14
Protección de datos.....	14
Garantías de calidad	15
6. Arquitectura tecnológica	15
Fase 1: MVP.....	15
Fases siguientes	16
Arquitectura basada en cloud.....	18
Software requerido	19
Procesos de control	19
Ciencia de datos.....	20
Modelos analíticos	20
Métricas	24
7. Reporting	26
8. Otros casos de uso	28
9. Monetización.....	29
10. Rentabilidad	30
11. Anexos	33

CASO DE USO - SECTOR ASEGURADOR

El alumno forma parte de la dirección de la empresa **Verti**. Estáis buscando ideas para conseguir aumentar el número de clientes de la rama de seguros de coches. Surge la idea de monitorizar el estilo de conducción de los clientes potenciales y ofrecer condiciones ventajosas de contratación a aquellos que demuestren ser más prudentes y por tanto tengan menores probabilidades de tener un accidente. El ahorro de costes que supone para la empresa un escaso riesgo de accidente se le puede trasladar a esos clientes aplicándoles un descuento sobre la tarifa del seguro contratado.

En la actualidad, la inmensa mayoría de la población española lleva consigo un dispositivo móvil conectado a internet mientras conduce un coche. Si el conductor accede a instalar una aplicación en ese dispositivo y concede a la empresa el permiso de geolocalización del mismo, Verti dispondría de información muy valiosa sobre los hábitos de conducción del mismo y podría utilizarla para decidir si es rentable ofrecerle un descuento al conductor y la cuantía del mismo.

¿Qué pasos debe seguir el equipo de trabajo para lanzar esta iniciativa?

Propuesta principal: plantear el cálculo de la probabilidad que tiene cada conductor usuario de la aplicación de sufrir un accidente.

Big Data Analytics



Compañía de seguros online

DESCRIPCIÓN DE LA COMPAÑÍA

Verti es la compañía digital del Grupo MAPFRE. Una aseguradora nacida en enero de 2011, de capital 100% español y con el respaldo de un grupo líder en el mercado asegurador a nivel mundial. Pese a este breve tiempo, es la compañía de seguro directo que más rápido ha crecido en la historia de España, alcanzando cerca de 300.000 clientes.

Su principal comercialización es a través de internet, con productos sencillos y flexibles lo que se traduce en precios muy competitivos. Sin embargo deben adaptarse continuamente nuevos retos, como la personalización de la demanda y la agilidad de los servicios, que exige rapidez en la respuesta y una experiencia digital perfecta.

Los clientes Verti valoran de forma especial la experiencia digital y prefieren gestionar en autoservicio la relación con su aseguradora.

El análisis y tratamiento de los datos es, quizá, el reto más importante para el sector asegurador, porque a partir de ellos se pueden ofrecer operativas simplificadas y experiencias personalizadas, más relevantes para el cliente y que contribuyan a hacerle la vida un poco más fácil.

1. PROYECTO DE ANALÍTICA PREDICTIVA

Buscando ideas para conseguir aumentar el número de clientes de la rama de seguros de coches, surge la propuesta de monitorizar el estilo de conducción de los clientes potenciales y ofrecer condiciones ventajosas de contratación a aquellos que demuestren ser más prudentes y por tanto tengan menores probabilidades de tener un accidente. El ahorro de costes que supone para la empresa un escaso riesgo de accidente, puede ser trasladado a sus clientes, aplicándose un descuento sobre la tarifa del seguro contratado.



Principales líneas de actuación:

Esta idea además, se concibe como una transformación de los seguros en una nueva era de personalización y optimización a través de la tecnología, poniendo al cliente en el centro del negocio y obteniendo mayor y mejor conocimiento de lo que ocurre en la conducción, contribuyendo en el mejor de los casos a la disminución de accidentes.

Un proyecto ambicioso supone superar obstáculos en cada etapa del mismo.

Tener un conocimiento de los hábitos de conducción requiere superar ciertas barreras tecnológicas y emocionales, además de estimar un tiempo adecuado para obtener toda la información necesaria.

Entre las barreras tecnológicas destacamos la obtención de los datos de movimiento del coche, los cuales pueden ser monitorizados desde un dispositivo electrónico fabricado para tal efecto o como opción más ventajosa, desde el Smartphone del cliente/conductor a través de una app.

Superar la barrera emocional parte de conseguir que el cliente tenga instalada la app en su Smartphone, con los permisos adecuados para la captura de datos necesarios (acelerómetro, gps, giroscopio) y el compromiso de hacer uso de la misma cada vez que conduce.

Por último, disponer en el menor tiempo posible de un volumen de datos suficiente, que sirvan de prueba para elaborar los modelos predictivos y permitan poner en producción el proyecto.

2. METODOLOGÍA DE TRABAJO

Con el objetivo de poder adaptar la forma de trabajo a las condiciones del proyecto, consiguiendo **flexibilidad** e **inmediatez** además de autonomía por parte de los perfiles implicados, se hará uso de metodologías ágiles, adoptando Scrum como la más adecuada. Otras ventajas de trabajar con metodologías ágiles son:

- Mejora de la calidad del producto
- Mayor satisfacción del cliente
- Mayor motivación de los trabajadores
- Trabajo colaborativo
- Mejor control y capacidad de predicción
- Reducción de costes

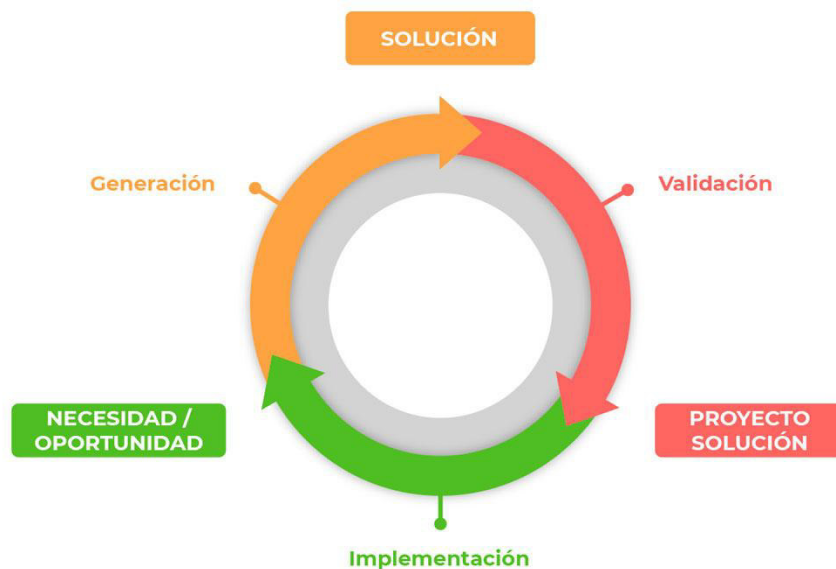
Los principales perfiles de la metodología Scrum son:

1. Product Owner, representa la voz del cliente, perfil que define y prioriza los objetivos del proyecto.
2. Scrum Master, se asegura de que todos los equipos de trabajo cumplen sus funciones para alcanzar los objetivos.
3. Scrum Team, es el equipo de trabajo encargado del desarrollo y entrega del proyecto.
4. Stakeholders, son los demás perfiles que dependen del proyecto para realizar su trabajo.

Partiendo de esta metodología de trabajo, desarrollaremos un **producto mínimo viable, MVP** con las características suficientes que permitan dar a conocer una primera fase del proyecto y proporcionen una retroalimentación suficiente para las fases futuras. En este proyecto, el **MVP** será la predicción mediante la clasificación de si ocurre o no un accidente, en función de los datos almacenados en una BBDD de siniestros de accidentes.

Los resultados obtenidos serán de gran interés, permitiendo además, detectar patrones recurrentes en los accidentes, ayudando a la compañía a tener una mejor visión de lo ocurrido y puntos a mejorar para la recolección de los datos en las fases siguientes del proyecto.

Esquema de procesos MVP:



Fases del Proyecto

Las fases estimadas para completar el proyecto las definimos a continuación:

Fase 1: MVP, análisis y preparación de los datos para elaborar un primer modelo predictivo con datos propios de la empresa recopilados en su BBDD.

Fase 2: Incorporación de más fuentes de datos externas, tanto de datos abiertos como las adquiridas a terceros para enriquecer el modelo.

Fase 3: Creación de una app como canal de comunicación con el cliente que sirva de fuente de recogida de datos.

Fase 4: Incorporación y procesado en batch de los datos recogidos por la app al modelo.

Fase 5: Monitorización de la conducción en tiempo real, a través de sensores recogidos en app y nuevos tipos de datos no estructurados como las imágenes.

Aunque desarrollemos en primera instancia la fase 1 con un MVP, definiremos toda la estructura necesaria para abordar el proyecto de Big Data.

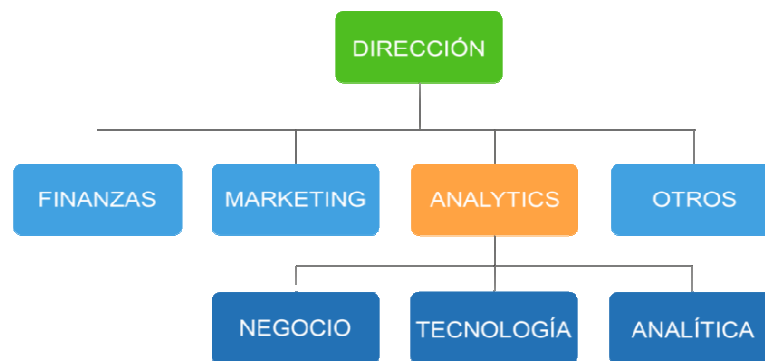
3. ORGANIZACIÓN EN LA EMPRESA

Para ejecutar el proyecto, la empresa debe contar con perfiles diversos y muy especializados en su campo. Algunos de ellos ya están trabajando en la compañía y son de suma importancia para aportar la visión de negocio, con un plan de formación específica pueden asumir los roles necesarios para contribuir al proyecto.

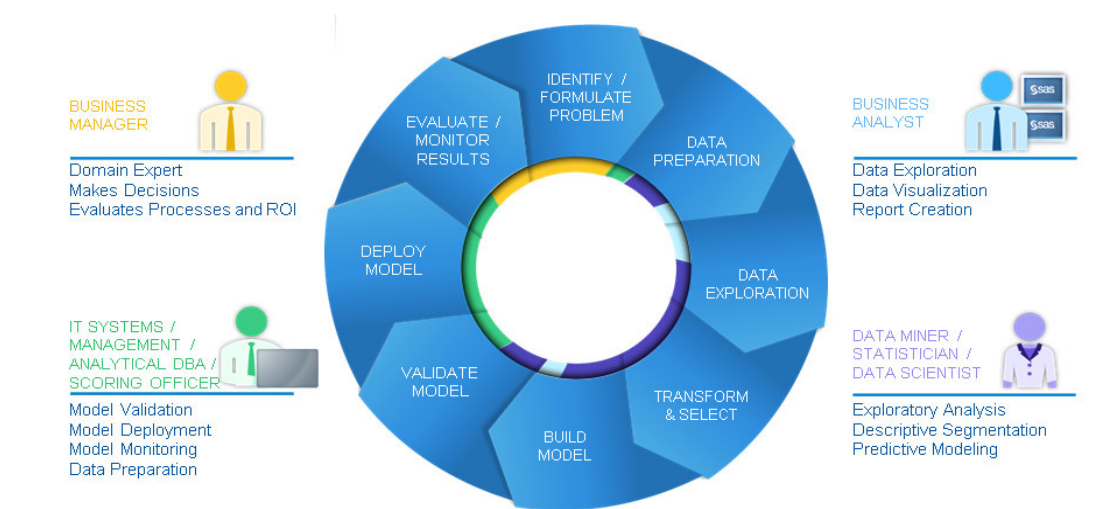
Entre los más relevantes destacamos:

Perfil	Funciones	Perfil Scrum
CDO (Chief Data Officer) Business Manager	Lidera la gestión global de los datos que tiene la empresa y debe obtener analíticas para la toma de decisiones. Es el director del resto de perfiles profesionales relacionados con el Big Data que haya en la empresa.	Scrum Master
Data Engineer IT Systems	Se encarga de ofrecer los datos de manera accesible a los data scientists. Tiene gran conocimiento en gestión de bases de datos, arquitecturas de big data, lenguajes de programación y sistemas de procesamiento de datos.	Scrum Team
Data Scientist Data Miner	Asume el rol de extraer conocimientos e información valiosa del análisis de los datos para que pueda ser tomada como información valiosa y relevante por el CDO.	Scrum Team
Data Governance	Perfil encargado de cuidar la calidad de los datos, garantizar la protección y gestionar el ciclo de vida de la información.	Product Owner
Analista de Datos Business Analyst	Se encarga de identificar las limitaciones y proporcionar recomendaciones a partir de los resultados obtenidos en los análisis.	Stakeholders

Al tratarse de una empresa 100% digital, la estructura jerárquica ya cuenta con una línea estratégica de análisis de datos, que facilita la sinergia entre los demás departamentos. Los perfiles dedicados al nuevo proyecto estarán guiados por las directrices marcadas en este departamento y deben contar con perfiles mixtos, capaces de abordar las tres áreas principales: Negocio, tecnología y analítica



Etapas del Proyecto Big data Analytics y roles implicados



4. GESTIÓN DEL GOBIERNO DEL DATO

La estrategia de gobierno del dato para llevar a cabo este caso de uso, debe desarrollarse en tres puntos estratégicos:

1. Alinear las partes involucradas en el modelo con el objetivo de poder obtener todos los datos necesarios que puedan plantear un problema que sea de aportación al desarrollo del modelo. Identificación de los stakeholders

(clientes y negocio) y partners (Departamentos internos involucrados en el desarrollo de la fuentes de datos) dentro y fuera de la empresa.

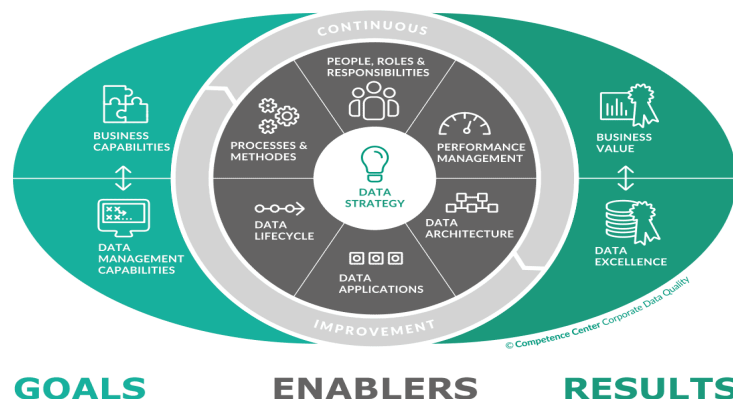
2. Evaluación de los activos Digitales: Debemos conocer de primera mano cómo están los activos digitales de la empresa, que capacidades de infraestructura dispone y qué políticas maneja el área de negocio para poder alinear la extracción de fuente de datos. ¿Qué arquitectura posee? ¿Qué capacidad de evolución se tiene por parte de negocio y el departamento de IT?.
3. Evaluación y análisis de la herramienta de Data Governance.

Herramienta de Data Governance

Una vez identificado el cuerpo del gobierno de datos, es muy importante que la herramienta cumpla con ciertos puntos fundamentales, como el que sea multitenencia, debe gestionar varias instancias y distintos roles de proyecto. Otra parte fundamental del Data Governance son las reglas de datos maestros que nos permitirá crear reglas de validación, de enriquecimiento, relación, correspondencia y consolidación del dato, al igual que identificar el umbral de confianza.

También es fundamental desarrollar adecuadamente y pensando en el futuro todos los siguientes puntos que deben estar integrados en la herramienta: Despliegue en Cloud, con Licencia o Open Source, que sea independiente del resto de tecnologías de la empresa, usabilidad de glosario de negocio y que pueda trabajar con plantillas, debe ser capaz de administrar los datos a nivel de políticas, términos de negocio, reglas de calidad de datos y tareas de datos maestros, definir los workflows de administración, diferentes perfiles de fuentes de datos con las herramientas que más se adecuen, como C3, QUERONA O DENODO, que tenga buena visualización de datos o que sea compatible con POWER BI O TABLEAU.

El modelo de data driven en Verti debe alinear a todas las partes implicadas en el proceso de beneficios estratégicos, y ayudar a la generación de nuevos clientes, así como a la captura de un mayor conocimiento de su cartera de clientes actuales, lo que debe permitir generar mejores condiciones para sus clientes y optimizar los costes de la compañía.



5. FUENTES DE DATOS.

Abordar un proyecto de estas características requiere la recopilación de diferentes fuentes de datos, que en su conjunto, ayudan a enriquecer el modelo predictivo y facilitar la toma de decisiones.

Para el proceso analítico se recopilan, entre otros, datos existentes en la empresa, datos abiertos, principalmente de entidades públicas recogidas en el portal **datos.gob.es** y, en una fase más avanzada del proyecto, datos recogidos por sensores durante la conducción.

Las fuentes de datos susceptibles a incorporar son:

Dataset	Organismo	Información relevante
Siniestros Verti	BBDD Verti	Datos de aseguradora con información de reclamaciones de accidentes
Parque de vehículos - Anuario - 2019	DGT	Clasificación vehículos más antiguos
Accidentes en función de la vía 2019	DGT	Clasificación vías peligrosas y accidentes comunes
Conductores y víctimas implicados 2019	DGT	Clasificación grupos de edad y perfil más accidentado
Características vehículos accidentados 2019	DGT	Fallos habituales en accidentes
Accidentes últimos 10 años en UK	Kaggle	Dataset con numerosas variables para elaborar un modelo predictivo
Aemet OpenData	AEMET	Valores climatológicos en tiempo real e históricos
Incidencias circulación	DGT	Estado de carreteras en tiempo real
Multas de tráfico	Datos.gob.es	Sanciones de tráfico recopiladas por ayuntamientos de Gijón y Madrid
Dataset estudio movimiento smartphone	Kaggle	Sensor de movimiento smartphone, para establecer patrones (Aceleración, parado...)

Actualmente podemos incorporar al análisis predictivo, datos históricos recogidos en la BBDD de la empresa, (tipología clientes, siniestralidad, tarifas, costes, etc.) Estos datos ayudarán a calcular con mayor precisión la póliza a contratar y predecir si el cliente puede sufrir un accidente.

Con la muestra inicial de clientes de la Fase 1 realizaremos una prueba piloto donde podremos recoger datos en directo con la App que estará enlazada al GPS de Google y que nos permitirá verificar tanto la capacidad de adaptación del cliente como los hábitos de conducción según las métricas existentes de GPS. La forma en la que llevaremos esta prueba a trabajo de campo será con la ayuda de una campaña de marketing y negocio, que nos permita premiar al cliente con puntos que puedan ser canjeados para próximos productos o la renovación de la misma póliza.

[Enlace al prototipo online](#)



Tabla de información recogida en la BBDD de la empresa

Variables	No Ocurrencia de Accidentes Siniestros (Y = 0)	Ocurrencia de Accidentes Siniestro (Y = 1)	Total
Edad (Años Asegurado)			
Género	Femenino		
	Masculino		
Experiencia Conducción (Años)			
Antigüedad del Vehículo (Años)			
Total kilómetros recorridos			
Porcentaje total de kilómetros recorridos en núcleos urbanos			
Porcentaje total de kilómetros sobre el límite obligatorio de velocidad			
Porcentaje de kilómetros recorridos de noche			
Total número de casos			

Otros datos no disponibles pero que pueden ser de gran utilidad, son las estadísticas de todos los accidentes de tráfico, independientemente de si presentan víctimas mortales, (como los recoge la DGT) y las causas por las cuales se producen los mismos. Es por este motivo que en una fase más avanzada del proyecto se quiera obtener datos de los sensores de movimiento, reconociendo patrones de conducción y otras conductas que proporcionen información más precisa.

Con el objetivo de aportar mayor eficacia a las predicciones, enriqueciendo el conjunto de datos y agilizando la obtención de los mismos, la empresa puede asumir una inversión económica accediendo a datos recopilados por otras compañías, como datos sin información personal pero sí que puedan mostrar ciertas métricas de desplazamientos o patrones de conducción o relacionados.

Entre las principales fuentes de información que pueden ser recopiladas por terceras empresas, destacamos los datos proporcionados de estudios estadísticos

demográficos y geográficos que permiten obtener conocimiento colectivo del sector automovilístico y que se distribuyen por ciertas asociaciones del sector. Para acceder a este tipo de información, existen asociaciones de seguros y empresas especializadas en repositorios de datos a las cuales puede adherirse Verti, destacando:

Fuentes de terceros

Nombre	Descripción
UNESPA	Unión Española de Entidades Aseguradoras y Reaseguradoras. Representa a cerca de 200 compañías que juntas abarcan el 98% del negocio en España.
ICEA	Investigación Cooperativa entre Entidades Aseguradoras y Fondos de Pensiones. Presta servicios de estudios del sector asegurador, siendo el organismo encargado de realizar y publicar todas las estadísticas sectoriales.
STATISTA	Plataforma de recolección de datos e indicadores de 170 sectores de más de 150 países.

Protección de datos

El tratamiento de todos los datos para este proyecto deben seguir las indicaciones del **Reglamento General de Protección de Datos (RGPD)**, regulación que establece los requisitos específicos para empresas y organizaciones sobre recogida, almacenamiento y gestión de los datos personales.

Las normas de protección de datos de la UE establecen que los datos deben tratarse de manera justa y lícita para un fin específico y legítimo y sólo deben tratarse los necesarios para alcanzar ese objetivo. La empresa debe cerciorarse de que se cumple una de las siguientes condiciones para el tratamiento de los datos personales:

- El interesado ha dado su **consentimiento**
- Los datos personales son necesarios para respetar una **obligación contractual** con el interesado
- Los datos personales son necesarios para cumplir una **obligación legal**
- Los datos personales son necesarios para proteger los **intereses vitales** del interesado
- Los datos personales se tratan para una **misión de interés público**
- Se actúa en **interés legítimo** de la empresa, siempre que en el tratamiento de los datos del interesado no se vean gravemente afectados los derechos y

libertades fundamentales de este; si los derechos de esa persona prevalecen sobre los intereses de la empresa, no se pueden tratar sus datos personales.

Además, Verti debe incorporar en el RGPD información relativa a:

- Autorizar el tratamiento de datos: consentimiento
- Proporcionar información transparente
- Normas específicas para menores
- Derecho de acceso y derecho a la portabilidad de los datos
- Derecho de rectificación y derecho de oposición
- Derecho de supresión (derecho al olvido)
- Decisiones automatizadas y elaboración de perfiles
- Violación de datos: proporcionar la notificación adecuada
- Responder a las solicitudes
- Evaluación de impacto
- Mantenimiento de registros

Garantías de calidad

Para garantizar la calidad de los datos y cumplir con la legislación, la empresa, dentro de su estrategia de gobierno del dato, ha de cumplir con tres premisas: **evaluar**, **dirigir** y **monitorizar** implicando en todo momento a personas, procesos y tecnología, estableciendo un marco de responsabilidad idóneo para la implementación de principios que afecten al uso de la información en aras a asegurar un uso más eficiente de la misma.

El sistema de Data Governance permitirá que este control sea efectivo y que los datos incorporados a los ficheros relacionales pasen por el filtro de limpieza que evitará subir datos con formatos incoherentes. Es importante identificar los sesgos de la información para direccionar correctamente el análisis.

6. ARQUITECTURA TECNOLÓGICA

Fase 1: MVP

En la primera fase del proyecto MVP, la empresa hará uso de software especializado tipo Open Source, el conjunto de datos con el que se trabajará provendrá de BBDD relacionales en su gran mayoría. El análisis y modelado predictivo precisará de ordenadores estándar con memoria RAM de al menos 8gb. Los programas más apropiados para estas tareas son Knime, R o Python. Para escalar el proyecto, se contratarán servicios cloud, cuya característica principal es el

uso bajo demanda, permitiendo a las empresas acceder a potentes recursos tecnológicos a un precio asumible y a medida de sus necesidades.

La elección del proveedor dependerá de diversas variables entre las que destacamos:

- I. Tipo de proyecto, algunos servicios cloud están mejor implementados con BBDD orientadas a grafos, tratamiento de imágenes y streaming.
- II. Experiencia de usuario y soporte, el departamento de IT requiere un entorno intuitivo y con numerosos recursos para consulta y documentación.
- III. Integración con el software usado en la empresa, para facilitar la agilidad en los proyectos y el uso transversal de sus implementaciones.

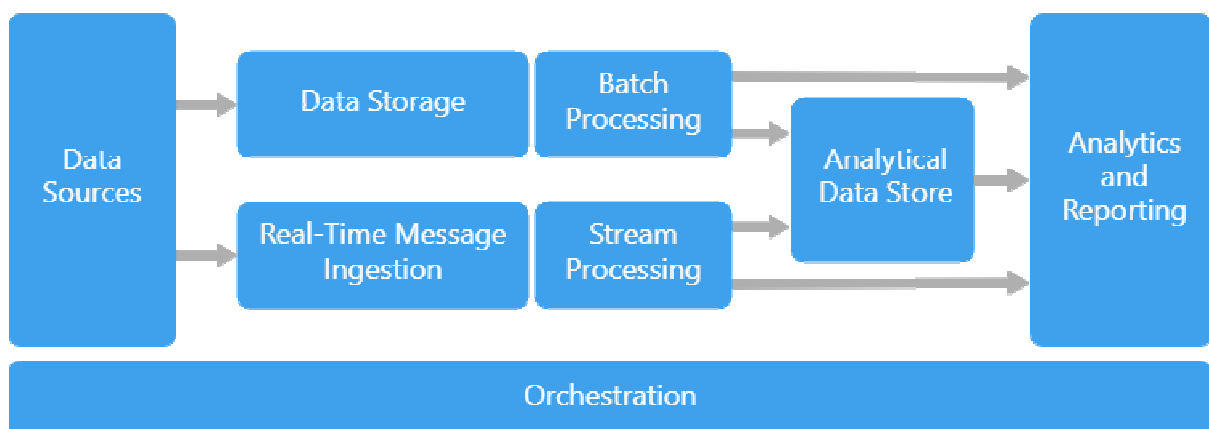
Entre los proveedores de servicios cloud más importantes se encuentran:



Con el fin de ampliar información en la búsqueda de servicios cloud, adjuntamos tabla comparativa en el apartado de anexos.

Fases siguientes

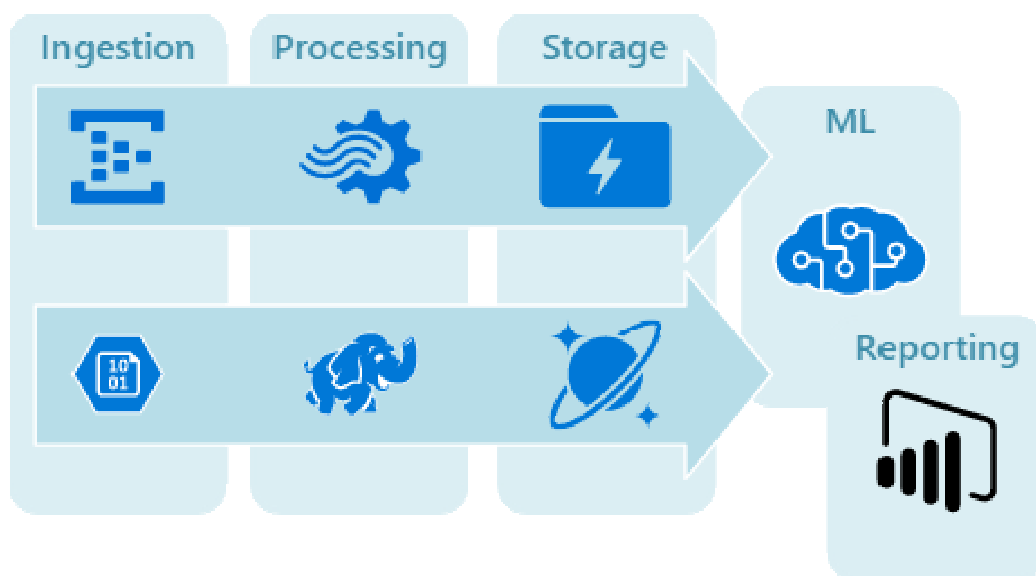
La arquitectura de alto nivel propuesta por Verti para el proyecto requerirá del diseño y construcción de una infraestructura que pueda abarcar todas las fases estimadas, soporte la ingesta y control de macrodatos, así como el procesamiento y el análisis de los mismos, los cuales serán capturados de diferentes fuentes de origen.



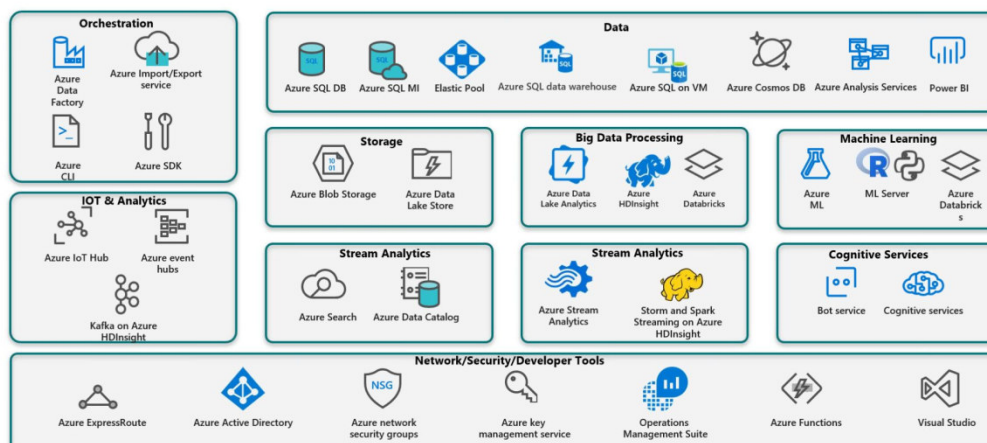
Las soluciones planteadas para el proyecto implican uno o varios de los tipos siguientes de cargas de trabajo:

- Procesamiento por lotes de orígenes de macrodatos en reposo.
- Exploración interactiva de macrodatos.
- Análisis predictivo y aprendizaje automático
- Procesamiento en tiempo real de macrodatos.
- Reporting o visualización.

Ejemplo de etapas en el procesamiento de datos en Azure.

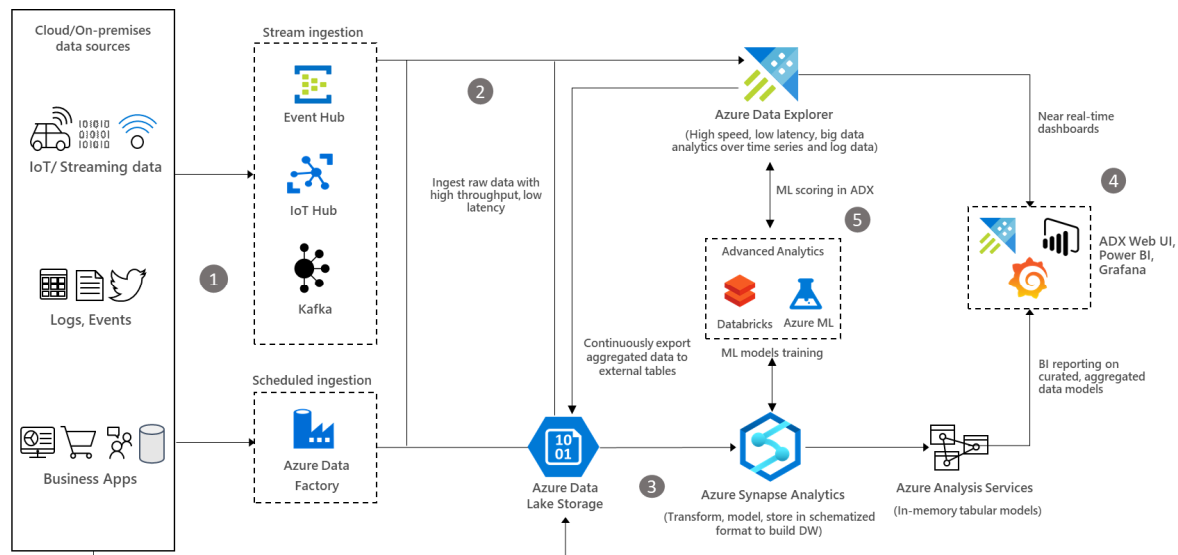


Ejemplo de herramientas disponibles para tratamiento de datos en Azure



Arquitectura basada en cloud

Una opción de arquitectura que abarque todas las fases del proyecto podría estar basada en **Azure Data Explorer** y **Azure Synapse Analytics** para el almacenamiento y análisis casi en tiempo real de datos con el siguiente esquema y componentes necesarios:



Componentes necesarios:

- Azure Event Hub: servicio de ingesta de datos en tiempo real y totalmente administrado simple, confiable y escalable.
- Azure IoT Hub: servicio administrado para habilitar la comunicación bidireccional entre los dispositivos de IoT y Azure.
- Kafka en HDInsight: servicio rentable y sencillo de nivel empresarial para el análisis de código abierto con Apache Kafka.
- Azure Data Explorer: servicio de análisis de datos rápido, totalmente administrado y muy escalable para el análisis en tiempo real de grandes volúmenes de datos que se transmiten en secuencias desde aplicaciones, sitios Web, dispositivos IoT, etc.
- Paneles de Azure Data Explorer: exporte de forma nativa las consultas de Kusto que se exploraron en la interfaz de usuario Web a los paneles optimizados.
- Azure Synapse Analytics: servicio de análisis que engloba el almacenamiento de datos empresariales y el análisis de macrodatos.

Software requerido

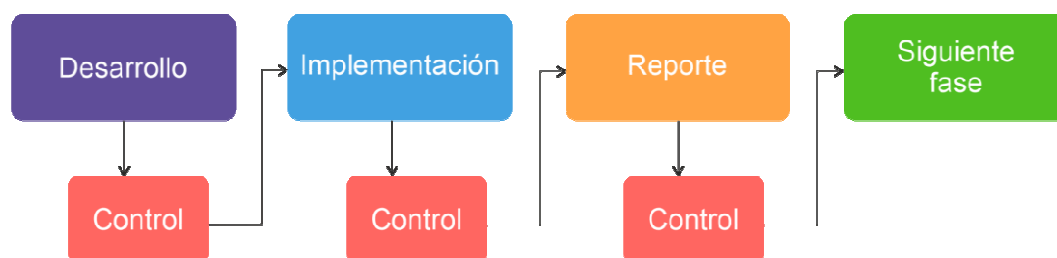
Dependiendo del entorno cloud con el que se trabaje, existe un gran ecosistema de tecnologías y software asociados que nos ofrecerán soluciones para la explotación, almacenamiento, procesamiento y análisis de los datos asociados con el proyecto de Verti para la predicción de siniestros con técnicas de machine learning.

Ejemplo de herramientas disponibles en el tratamiento del dato.



Procesos de control

Siguiendo los principios de metodologías ágiles, el proyecto será desarrollado por fases, implementando en cada una de ellas pruebas de control para minimizar riesgos. El proceso de cada fase es el siguiente:

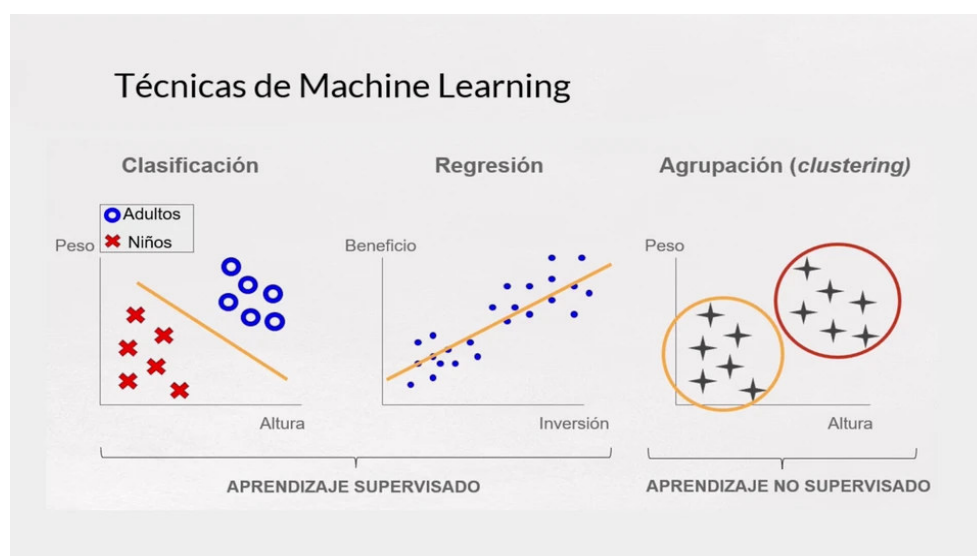


Una vez identificados los procesos a seguir, en cada fase haremos uso de la ciencia de datos para obtener la información más relevante y predictiva sobre los conductores/clientes generadores del dato, esta información la obtendremos a partir de tres importantes pasos:

- I. **Análisis exploratorio:** en un primer lugar se analizará el conjunto de datos para describir, resumir y visualizar la naturaleza de los mismos, aplicando técnicas básicas de estadística descriptiva. Gracias a este análisis podremos identificar la calidad, distribución y relación entre las variables implicadas, aportando insights muy valiosos para la empresa.
- II. **Corrección y transformación:** una vez conocidos los atributos de los datos a tratar, si es el caso, deben ser solventados los problemas más habituales, como la imputación de valores nulos y corrección de atípicos. «garbage in, garbage out» También pueden transformarse los datos para generar nuevas variables que aporten mayor visión al negocio o puedan mejorar el modelo predictivo.
- III. **Análisis predictivo:** con el conjunto de datos preparado adecuadamente, se puede proceder a crear un modelo analítico (algoritmo), que ayude a identificar patrones o relación entre las variables más relevantes y, gracias a ellas, predecir acciones futuras. Este proceso se conoce también como aprendizaje automático o machine learning.

Modelos analíticos

En función de las características de cada proyecto y el tipo de dato, es necesario aplicar técnicas de aprendizaje automático diferentes y, en cada una de las técnicas, tenemos a disposición diversos modelos, algunos más avanzados y complejos que otros.



En el caso de uso de Verti, nos sometemos a técnicas de aprendizaje supervisado, concretamente a técnicas de clasificación. Es interesante aplicar varios modelos analíticos y comparar su efectividad para dotar de mayor precisión al proyecto.

El cálculo de la prima de un seguro se basa en múltiples parámetros muy complejos, unos se pueden estimar previamente a la firma del contrato: factores relativos al vehículo, al conductor y a la conducción (a priori); y otros posteriormente a la firma del contrato: historial de siniestralidad, multas, etcétera (a posteriori).

Este proyecto consistirá en el análisis de una base de datos de siniestros de accidente para predecir futuros, y además, conseguir agrupar los siniestros según tres criterios distintos:

- I. Según **la severidad del siniestro**: se crearán grupos basados en el coste económico de la reparación del siniestro. Dentro de cada uno de estos grupos, se volverá a realizar una separación en grupos, esta vez buscando la existencia de las ubicaciones, localizaciones y lugares de mayor ocurrencia predominantes en cada uno de los grupos de severidad.
- II. **Según las características y perfil del conductor**: se crearán grupos basados para obtener un perfil de conductor basado en las características de los asegurados en cada siniestro. A partir de la obtención de dichos perfiles, se crearán distintos grupos de perfil de conductor, analizando para intentar predecir qué grupos de asegurados son más propensos a sufrir accidentes.
- III. **Con datos procedentes de información Telemática e IoT**: se entrenará el modelo para intentar predecir las características y variables que tienen una relación directa asociada a los grupos de asegurados y perfil de conducción para predecir la siniestralidad y estimar el cálculo de la prima de los asegurados.

En fases más avanzadas del proyecto podremos recoger datos desestructurados en formato de imagen, y a través de técnicas de deep learning, podremos descifrar por medio de gestos, la variable estrés que puede afectar a la conducción y subir el nivel de accidentalidad.

Predicción con nuestro dataset

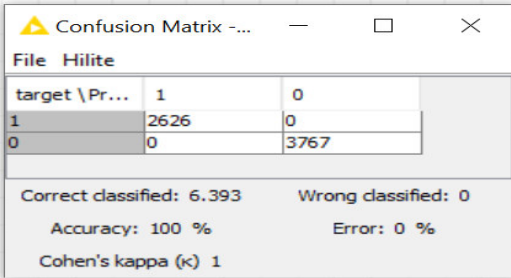
Predecir la ocurrencia de siniestros de accidentes en el seguro de automóviles es la base de la prima de cálculo, pero con el desarrollo de nuevos métodos de inteligencia artificial, la cuestión de elegir un modelo adecuado de predicción aún no se ha resuelto por completo. **Por eso estudios recientes, analizan los métodos propuestos de comparar el algoritmo XGBoost y la Regresión Logística**, para

establecer una comparativa con respecto a su desempeño predictivo en una muestra de conductores asegurados, incluyendo a su vez datos con información telemática.

En nuestro caso, una vez disponemos del dataset y antes de entrenar nuestro algoritmo tenemos que, seguir los pasos establecidos en la ciencia de datos, analizar, corregir, transformar y modelar. En este último punto, uno de los modelos utilizados es Xtreme Gradient Boosting, un algoritmo ensamblado caracterizado por ser un algoritmo interactivo que crea árboles de decisión que van corrigiendo los errores que tuvieron los anteriores, de manera que la predicción sea cada vez más precisa.

Los resultados obtenidos pueden ser analizados a través de diversas métricas, entre las que se encuentran las matrices de confusión.

Ejemplo de resultado obtenido de la matriz de confusión



Confusion Matrix -...		
File Hilite		
target \ Pr...	1	0
1	2626	0
0	0	3767

Correct classified: 6.393	Wrong classified: 0
Accuracy: 100 %	Error: 0 %
Cohen's kappa (κ) 1	

Tras analizar los resultados obtenidos en nuestras pruebas, se podría concluir que usando un volumen de 6.393 registros, nuestro modelo es capaz de detectar si habrá accidentes o no usando algunas características de los clientes. En concreto, el modelo detecta el 100% de los accidentes ocurridos. Esto nos puede llevar a dos conclusiones significativas: el modelo es muy bueno o existe un sobreajuste de los datos. Estos supuestos se analizarán más adelante.

Datos de entrada al algoritmo

Cualquiera de los algoritmos que queramos utilizar para la clasificación y predicción usarán como input el conjunto de datos también conocido como dataset, almacenado en la BBDD, este dataset deberá estar procesado o normalizado con los pasos que describimos en la ciencia de datos (análisis, corrección y transformación) para que el algoritmo pueda ser aplicado correctamente, ya que en algunos casos, los datos pueden contener texto o códigos que no tienen un significado numérico y el algoritmo no sabría interpretarlos.

La arquitectura de la solución propuesta propone una única plataforma de datos unificada para cumplir los requisitos del proyecto:

- Canalizaciones de datos relacionales tradicionales estructurados
- Transformaciones de macrodatos
- Ingesta de datos no estructurados y enriquecimiento con funciones basadas en inteligencia artificial
- Ingesta y procesamiento de flujos después de la arquitectura lambda
- Servicio de información para aplicaciones controladas por datos y visualización de datos enriquecida

Servicio Arquitectura Azure	Funcionalidades
Azure Data Factory	Ingesta de datos con Azure Data Factory
Azure Synapse Analytics	Implementación de un almacenamiento de datos con Azure Synapse Analytics
Azure Data Lake Storage Gen2	Procesamiento de datos a gran escala con Azure Data Lake Storage Gen2
Azure Cognitive Services	Rutas de aprendizaje y módulos de Cognitive Services
Azure Cosmos DB	Uso de datos NoSQL en Azure Cosmos DB
Azure Databricks	Realización de ingeniería de datos con Azure Databricks
Azure Event Hubs	Habilitación de mensajería confiable para aplicaciones de macrodatos con Azure Event Hubs
Azure Stream Analytics	Implementación de una solución de streaming de datos con Azure Stream Analytics
Power BI	Creación y uso de informes de análisis con Power BI

Información de salida del algoritmo

Cada algoritmo aplicado en el modelo nos permite conocer las variables que proporcionan mayor conocimiento o peso en el dataset, estableciendo por tanto una relación directa entre las variables y el resultado obtenido. En función del proyecto y el algoritmo podríamos obtener diez o quinientas variables predictivas, por lo tanto puede resultar difícil interpretarlas una a una, para ello se recurre a técnicas de evaluación como las matrices de confusión mencionadas anteriormente.

Si bien la preparación de los datos y el entrenamiento de aprendizaje del modelo son pasos clave en el proyecto, es igualmente importante medir el rendimiento del algoritmo. Lo bien que el algoritmo generaliza sobre datos no vistos es lo que define si éste es adecuado o no. Entre las técnicas de evaluación para modelos de clasificación, también conocidas como métricas, podríamos destacar:

- Matriz de confusión: compara los valores reales con los obtenidos por el algoritmo y da cuatro posibles respuestas, verdadero positivo, verdadero negativo, falso positivo y falso negativo.

		Resultado de la predicción		
		Positivo	Negativo	
Valor actual	Positivo	TP	FN	TP + FN
	Negativo	FP	TN	FP + TN

- Métrica de exactitud (accuracy): indica el número de elementos clasificados correctamente en comparación con el número total de artículos.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Métrica de exhaustividad o sensibilidad (recall): muestra la cantidad de verdaderos positivos que el modelo ha clasificado en función del número total de valores positivos.

$$\text{recall} = \frac{TP}{TP + FN}$$

- Métrica de precisión: representa el número de verdaderos positivos que son realmente positivos en comparación con el número total de valores positivos predichos.

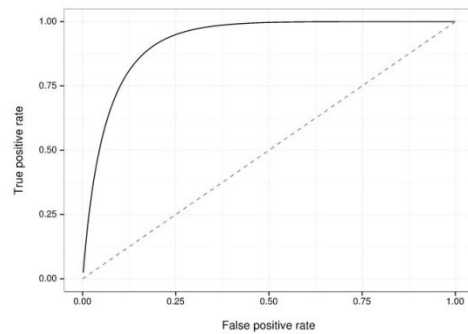
$$\text{precision} = \frac{TP}{TP + FP}$$

- Puntuación F1: es la combinación de las métricas de precisión y exhaustividad y sirve de compromiso entre ellas. La mejor puntuación F1 es igual a 1 y la peor

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

es igual a 0.

- Curva ROC (características operativas del receptor): es una representación gráfica que permite visualizar el equilibrio entre la tasa de verdaderos positivos y la de falsos positivos.



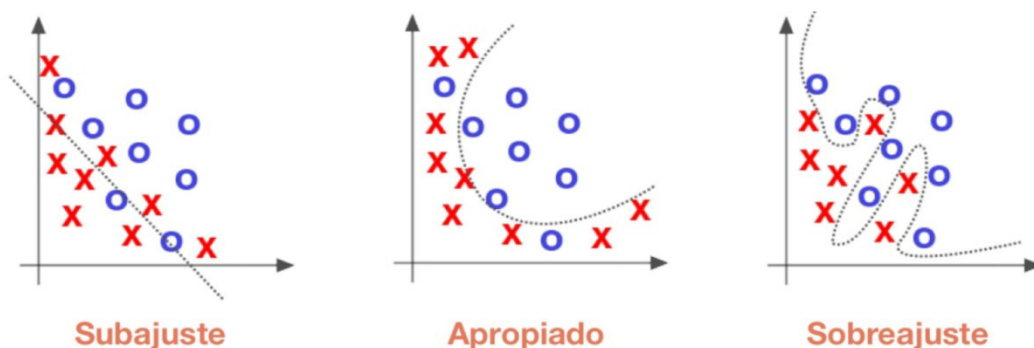
Las métricas pueden ayudar a la empresa a determinar qué algoritmo es el más adecuado para el proyecto y comprobar si este es capaz de generalizar sus predicciones sobre datos no vistos, además, permiten tomar decisiones y acciones para mejorar el poder predictivo antes de ponerlo en marcha para la producción sobre nuevos datos.

Evaluación y control

El proceso de evaluación y control debe continuar durante todo el ciclo, esto proporcionará la seguridad necesaria para obtener los mejores resultados a medida que pasa el tiempo, además los modelos pueden incurrir en **overfitting** (sobreajuste) o **underfitting** (subajuste) en función de los datos obtenidos.

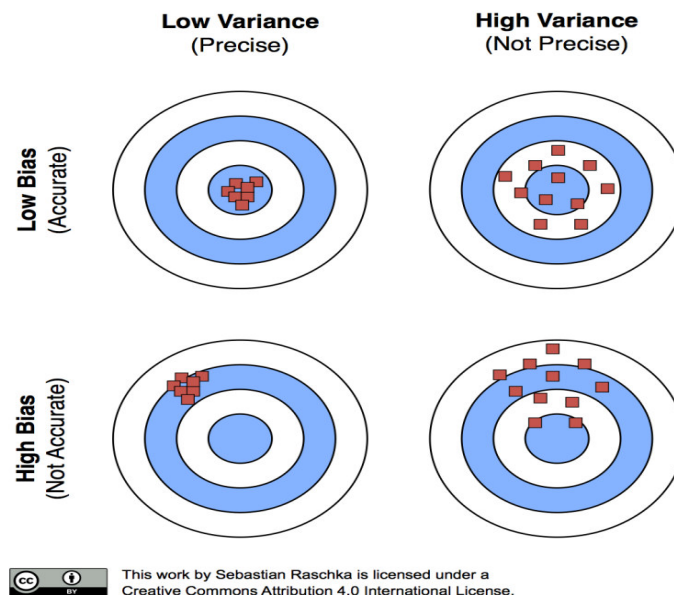
Cuando un modelo incurre en overfitting, indica que el algoritmo está considerando como válidos sólo los datos idénticos a los de entrenamiento, siendo incapaz de distinguir entradas buenas como fiables si se salen de los parámetros ya preestablecidos.

Ejemplo de sobreajuste y subajuste en un modelo de clasificación



Encontrar el punto medio apropiado es el objetivo, pero no siempre es tarea fácil.

Otros conceptos para conocer si el modelo está o no fallando son los términos de **bias** (sesgo) y **varianza**. Un alto sesgo o bias indica que el modelo sufre de underfitting o subajuste y una alta varianza indica que el modelo sufre de overfitting.



Monitorizar estos indicadores de forma recurrente ayudará a prevenir fallos, pueden establecerse frecuencias de horas, días, semanas o meses en función de la ingesta de datos y la fase en la que se encuentre el proyecto.

7. REPORTING

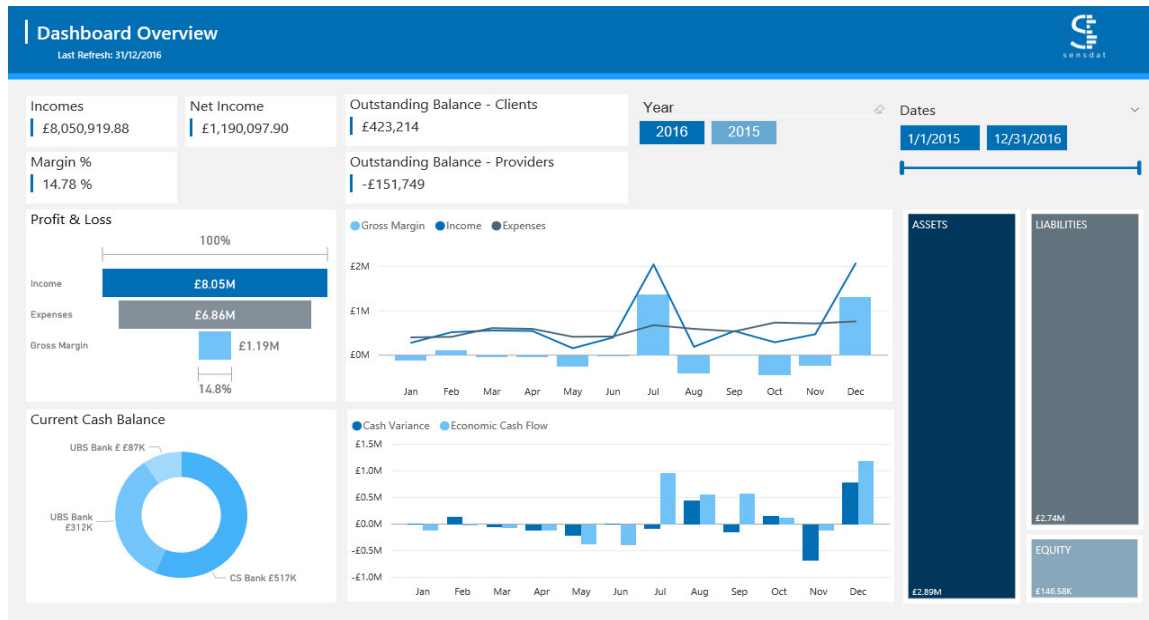
El reporting estará **soportado por las funcionalidades nativas de Azure Data Explorer para procesar los datos, agregarlos y analizarlos**. Permite obtener información detallada a gran velocidad, análisis casi en tiempo real con paneles de Azure Data Explorer, Power BI, Grafana y otras herramientas. El uso Azure Synapse Analytics para crear un almacenamiento de datos y combinarlo con los datos de Azure Data Explorer a fin de generar informes de BI en modelos de datos agregados y perfeccionados. El Explorador de datos de Azure es un servicio de exploración de datos altamente escalable y rápido para datos de telemetría y registro. Azure Data Explorer proporciona una aplicación web que permite ejecutar consultas y crear paneles. Los paneles están disponibles en la aplicación web independiente, la interfaz de usuario web. Azure Data Explorer también está integrado en otros servicios del panel, como Power BI y Grafana.

Los paneles de Azure Data Explorer ofrecen tres ventajas principales:

- Exportan de forma nativa las consultas de la interfaz de usuario Web a los paneles de Azure Data Explorer.

- Exploran los datos en la interfaz de usuario web.
- Mejor rendimiento de la representación del panel.

Ejemplo de dashboard generado con PowerBI



Azure Data Explorer proporciona la funcionalidad para conectarse a Power BI con varios métodos:

- Built-in native Power BI connector (Conector nativo de Power BI integrado)
- Query import from Azure Data Explorer into Power BI (Importación de consultas desde Azure Data Explorer a Power BI)
- SQL query
- Azure Data Explorer proporciona un conector para la conectividad con
 - Tableau
 - Qlick
 - Kibana
 - Grafana
 - Siense
 - Redhas

Ejemplo de reporting multiplataforma



8. OTROS CASOS DE USO

Nuevos modelos de negocio a través de los insights obtenidos en la exploración y análisis de los datos, pueden dar respuesta a inquietudes tales como:

¿Podemos mejorar la vida de los clientes con los datos que tenemos?

Gracias a la información procesada, podemos desarrollar un sistema de recomendación usando la técnica de análisis prescriptivo que nos permitirá hacer recomendaciones de seguridad en directo, informando al cliente de posibles situaciones de riesgo según el estado de su coche, los km que realiza, la velocidad media que utiliza y la posición geográfica donde transita habitualmente. Crear un sistema de recomendaciones donde el objetivo sea mejorar la calidad de vida del cliente ayudándole a prevenir accidentes en base a su conducta y entorno.

¿Podemos aumentar ingresos?

Con la misma técnica de análisis prescriptivo podemos sacar patrones que nos permitirán identificar variables que pueden indicar la oportunidad de ofrecer productos de valor añadido de verti. Como ejemplo podemos saber cuándo y cómo ha tenido un accidente el conductor por lo cual podremos

saber qué es lo que necesitaría para evitar que vuelva a ocurrir, si son accesorios de seguridad, mantenimiento del coche o simplemente hábitos de conducción. Se podrán diseñar acuerdos comerciales con marcas correspondientes y hacer que se obtengan acuerdos para ofrecer descuentos o beneficios que fidelicen al cliente y que nos permitan generar beneficio por venta de productos de valor añadido.

Para cumplir estos nuevos casos de uso deben cumplirse premisas importantes como la de recoger datos actualizados diariamente, con previa autorización del asegurado para poder identificar variables de tiempo, espacio, velocidad, (geográficas y demográficas). analizarlos, definirlos y alimentar el modelo para que pueda prescribir al cliente lo que queremos que haga en el momento más adecuado para la motivación a la acción. **BOTTOM – UP**

Se pueden implementar muchos otros casos de uso basados en soluciones big data que aporten una perspectiva muy útil para desarrollar nuevos proyectos en Verti, basadas en:

- Reducir costes de riesgos y abandono.
- Escoger y retener a los mejores clientes, en función de los perfiles idóneos.
- Aumentar su capacidad de innovación para la creación de nuevos productos y servicios que garanticen la satisfacción del cliente.

9. MONETIZACIÓN

La viabilidad del proyecto deber ir ligada a un beneficio para Verti, en primer lugar, la implementación del modelo y su uso ayudará a captar más cantidad de clientes gracias a precios más competitivos, pero además, proporciona un conocimiento extra en el sector asegurador, que son transformados en nuevos modelos de negocio.

Una vez obtenidos los datos que nos permiten definir, en base a criterios objetivos, los perfiles de clientes, estamos en capacidad de desarrollar nuevos modelos de recomendaciones en función de los hábitos de conducción o preferencias que podamos relacionar con accesorios para el vehículo y el conductor.

Otra vía interesante de monetización es la venta de información a terceras empresas, dicha información debe cumplir los requisitos del Reglamento General de Protección de Datos, existen diversas técnicas para cumplirlo, entre las más comunes la de anonimizar los datos, permitiendo compartir resultados obtenidos

tras los análisis, datos que asocien hábitos a perfiles de conductores, estadísticas y segregaciones según tipo de coche, zona geográfica etc.

También realizaremos acuerdos comerciales con marcas relacionadas que nos paguen por los anuncios y la vinculación a sus tiendas virtuales

10. Rentabilidad

Los costes principales en los que incurrirá el proyecto se detallan en la tabla de ejemplo. Entre los principales conceptos reflejados destacan:

- **Costes directos:** Hardware, Software y Management (equipo técnico).
- **Costes indirectos:** Soporte y promoción interna implantación proyecto (equipos transversales soporte).

La infraestructura de la solución constará de almacenamiento de datos, servidores, redes y herramientas de monitorización. Todos los costes deberán ser proporcionales al tamaño de la plataforma y volumen de ingesta y tratamiento de datos.

Igualmente otro coste muy significativo de construir una solución de análisis de Big Data son los recursos humanos. La solución dependiendo de la complejidad, requiere un conocimiento real del negocio e involucra a numerosos especialistas. La mayoría del equipo deben ser ingenieros con experiencia en Big Data, que es un recurso bastante escaso y caro. Una lista parcial de los expertos que requiere el proyecto serán: desarrolladores de ETL, expertos en infraestructura en la nube, desarrolladores de Java / Python, administradores de bases de datos (DBA), analistas de datos, desarrolladores, etc.

En nuestro caso hemos optado por ir a una solución global de mercado con un partner como Microsoft y aportar por una solución en Azure en la que se incluyan todos los componentes necesarios de las soluciones que proporciona el Microsoft en cuanto a (Almacenamiento, Contenedores Procesos, Bases de Datos, AI y Machine Learning, Seguridad, Analítica y Reporting, etc.)

Estimación de costes y ROI Proyecto Verti

COSTE PROYECTO BIG DATA (PROYECTO VERTI)

COSTES DIRECTOS		Año 1 (Pilot)		Año 2	Año 3	TCO LIFE CYCLE
Hardware		Coste	50%	30%	20%	
Servidores		18,000,00 €	9,000,00 €	5,400,00 €	3,600,00 €	18,000,00 €
Network Componentes & Redes		6,000,00 €	3,000,00 €	1,800,00 €	1,200,00 €	6,000,00 €
Comunicaciones		5,000,00 €	2,500,00 €	1,500,00 €	1,000,00 €	5,000,00 €
Total Hardware Cost			14,500,00 €	8,700,00 €	5,800,00 €	29,000,00 €
Software		Coste	50%	30%	20%	
Licencias y Software Microsoft Azure						
Data Factory		1.664,40 €	832,20 €	499,32 €	332,88 €	1.664,40 €
Azure Data Lake Storage Gen1		33.811,00 €	16.905,50 €	10.143,30 €	6.762,20 €	33.811,00 €
Azure Synapse Analytics		11.897,65 €	5.948,82 €	3.569,29 €	2.379,53 €	11.897,65 €
Azure Cognitive Services		3.262,35 €	1.631,18 €	978,71 €	652,47 €	3.262,35 €
Azure Cosmos DB		3.123,42 €	1.561,71 €	937,03 €	624,68 €	3.123,42 €
Azure Databricks		2.982,11 €	1.491,06 €	894,63 €	596,42 €	2.982,11 €
Event Hubs		765,79 €	382,90 €	229,74 €	153,16 €	765,79 €
Azure Stream Analytics		478,24 €	239,12 €	143,47 €	95,65 €	478,24 €
Power BI Embedded		4.328,90 €	2.164,45 €	1.298,67 €	865,78 €	4.328,90 €
Mantenimiento Software (h/xCosteh)	1.000,0	65,00 €	32.500,00 €	19.500,00 €	13.000,00 €	65.000,00 €
Desarrollo Software (h/xCosteh)	1.000,0	75,00 €	37.500,00 €	22.500,00 €	15.000,00 €	75.000,00 €
Total Software Cost			101.156,93 €	60.694,16 €	40.462,77 €	202.313,87 €
Management		FTEs	Coste salarial em	30%	50%	20%
Arquitectura / Infraestructura	1,00		50.000,00 €	15.000,00 €	25.000,00 €	10.000,00 €
Comunicaciones	0,50		45.000,00 €	13.500,00 €	22.500,00 €	9.000,00 €
Seguridad	0,50		45.000,00 €	13.500,00 €	22.500,00 €	9.000,00 €
Reporting	1,00		40.000,00 €	12.000,00 €	20.000,00 €	8.000,00 €
Support Staff	0,50		30.000,00 €	9.000,00 €	15.000,00 €	6.000,00 €
Total Management Cost				63.000,00 €	105.000,00 €	42.000,00 €
180.000,00 €						
Implementación Data Science		FTEs	Coste salarial em	30%	50%	20%
Chief Data Officer	1,00		100.000,00 €	30.000,00 €	50.000,00 €	20.000,00 €
Data Engineer	1,00		75.000,00 €	22.500,00 €	37.500,00 €	15.000,00 €
Data Scientist	1,00		65.000,00 €	19.500,00 €	32.500,00 €	13.000,00 €
Data Governance	1,00		65.000,00 €	19.500,00 €	32.500,00 €	13.000,00 €
Business Analyst	1,00		45.000,00 €	13.500,00 €	22.500,00 €	9.000,00 €
Total Implementation Cost				105.000,00 €	175.000,00 €	70.000,00 €
350.000,00 €						
TOTAL COSTES DIRECTOS				283.656,93 €	349.394,16 €	158.262,77 €
761.313,87 €						
COSTES INDIRECTOS		Año 1 (Pilot)		Año 2	Año 3	TCO LIFE CYCLE
Implantación Proyecto		Coste empresa	40%	20%	20%	
Program Management	0,50	20.000,00 €	8.000,00 €	4.000,00 €	4.000,00 €	16.000,00 €
Account Management	0,50	20.000,00 €	8.000,00 €	4.000,00 €	4.000,00 €	16.000,00 €
Business Operations	0,50	20.000,00 €	8.000,00 €	4.000,00 €	4.000,00 €	16.000,00 €
Marketing Promotions	0,50	20.000,00 €	8.000,00 €	4.000,00 €	4.000,00 €	16.000,00 €
Total Programmatic Cost			32.000,00 €	16.000,00 €	16.000,00 €	64.000,00 €
TOTAL COSTES INDIRECTOS			32.000,00 €	16.000,00 €	16.000,00 €	64.000,00 €
TOTAL COSTES			315.656,93 €	365.394,16 €	174.262,77 €	855.313,87 €
TOTAL COSTE / AÑO						285.104,62 €
POTENCIAL EFICIENCIA (ROI)		Año 1 (Pilot)		Año 2	Año 3	TCO LIFE CYCLE
Mejora Siniestralidad			4,4%	5,3%	4,4%	
Mejora Ratio Siniestralidad		74,34%	71,07%	70,40%	71,07%	
Prima media seguros		491,00 €	512,60 €	539,77 €	563,52 €	
Coste Siniestralidad			364,30 €	380,00 €	400,49 €	
Margen Técnico			148,30 €	159,77 €	163,03 €	
Polizas Impacto proyecto Big Data			1.837,00	2.393,00	1.675,00	
TOTAL POTENCIAL EFICIENCIA			272.429,41 €	382.335,89 €	273.079,13 €	927.844,43 €
POTENCIAL ROI %			86%	105%	157%	108%
POTENCIAL ROI / AÑO						€ 309.281,48

Aumento de ingresos y reducción de costes

En nuestro supuesto estimamos que las palancas sobre las que el Proyecto tendrá un impacto de mejora y que pueden generar incremento de ingresos, o reducción de costes estarán relacionadas con el propio negocio del seguro de auto.

Estimamos que se producirá una mejora en el Ratio de Siniestralidad, lo que implica una eficiencia en los costes y mejora el margen técnico de rentabilidad. Asociado a la mejora de la información de la clasificación de clientes y estimación de potencial de los siniestros en la cartera de asegurados, podremos realizar una mejor estimación en el cálculo de la prima con un incremento proporcional al riesgo de siniestralidad obtenido.

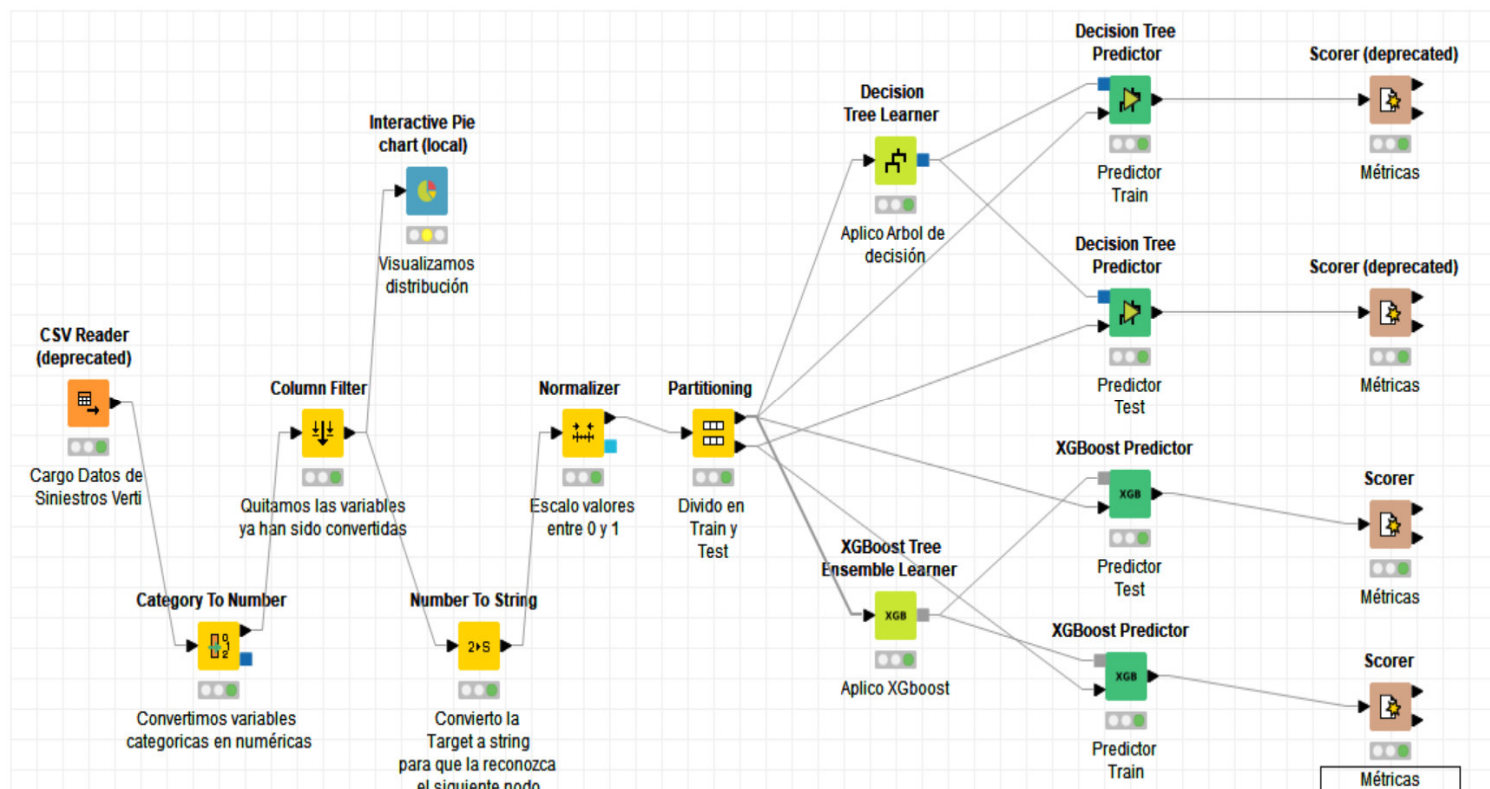


11. ANEXOS:

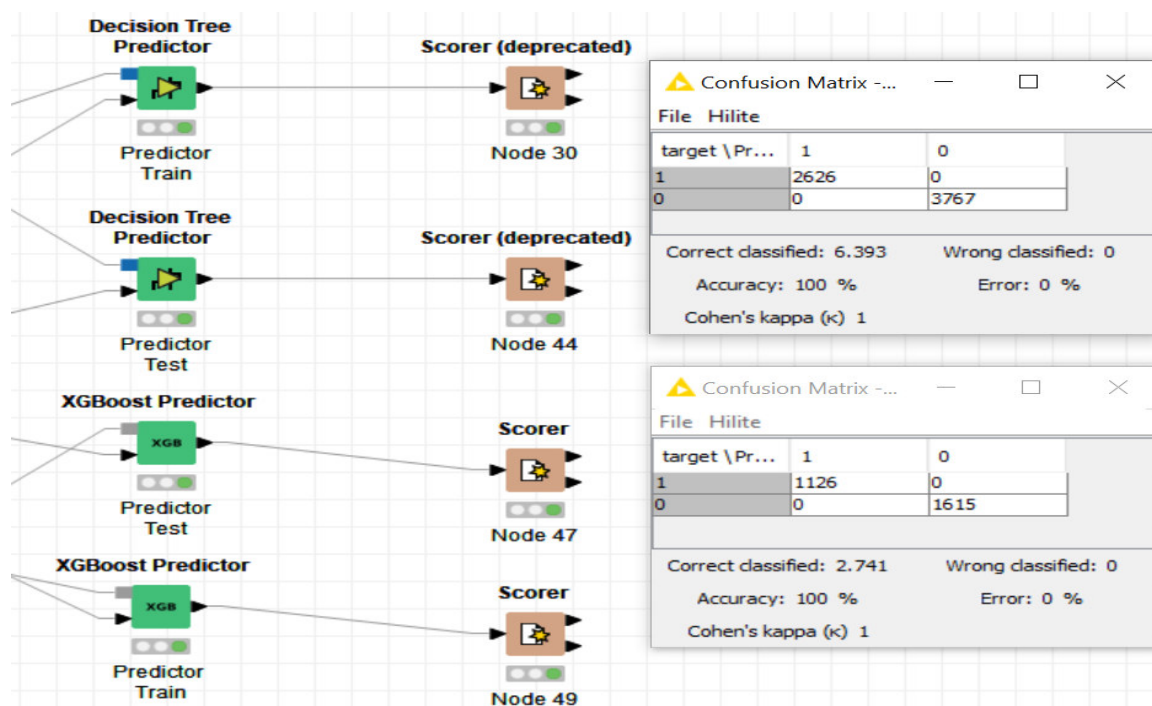
1- Notebook de Google Colaboratory con transformación de datos en R:

https://colab.research.google.com/drive/1aMt6s34y3FWC2dVJp0uTBkliz_r5GKxu

2- Flujo de trabajo en KNIME



3- Matrices de confusión generadas con algoritmos Decision Tree y XGBoost



4. Tabla comparativa de servicios cloud para desarrollos Big Data

CARACTERÍSTICAS	AWS	AZURE	GCP
Análisis de grandes cantidades de datos basada en Hadoop y/o Apache Spark	Amazon EMR	Azure Databricks HDInsight	Cloud Dataproc
Ingesta y el procesamiento de streams de datos	Amazon Kinesis	Stream Analytics Data Lake Store Análisis con Azure Data Lake	Pub/Sub
Streaming de datos	Kinesis Data Firehose Kinesis Data Streams	Event Hubs	
Servicio gestionado para almacenar datos empresariales y consultarlos mediante SQL estándar	Amazon Athena	Análisis con Azure Data Lake	BigQuery
Orquestación de workflows de datos	Data Pipeline AWS Glue	Data Factory Data Catalog	Cloud Composer
Visualización	QuickSight	PowerBI	Data Studio
Machine Learning & AI			
Servicio gestionado para Machine Learning	SageMaker	Azure Machine Learning Studio Servicio Azure Machine Learning	Cloud Machine Learning Engine
Reconocimiento de voz e interfaz de conversación	Amazon Lex	Bing Speech API Speaker Recognition API	Dialogflow Enterprise Edition
Texto a voz	Amazon Polly	Bing Speech API	Text-To-Speech
Visión	Amazon Rekognition	Computer Vision Face API Emotions API	Cloud Vision
Procesamiento de lenguaje natural	Amazon Comprehend	Language Understanding Intelligent Service (LUIS)	Cloud Natural Language
Traducción	Amazon Translate	Translator Text	Cloud Translation
Video	Amazon Rekognition Video	Video API	Cloud Video Intelligence
Servicios de asistente personal	Alexa Skills Kits	Azure AI Bot Framework	Actions on Google
Servicio ML automatizado		Cloud AutoML	
IoT y otros			
Conectar y supervisar dispositivos IoT	AWS IoT Core	Azure IoT Hub	Cloud IoT Core
Administración de dispositivos de forma remota	AWS IoT Device Management	Azure IoT Hub Device Management	
Detección de eventos y respuesta	AWS IoT Events	Event Grid	
Suite empresarial integral de comunicaciones, correo, documentos	WorkMail WorkDocs Chime	Office 365	Gsuite

5. Esquema visual de servicios en AWS, AZURE y GCP

Big Data Pipelines on AWS, Microsoft Azure, and GCP

scgupta.link/big-data-pipeline

