CIENCIA DE DATOS / IA

NÚMERO #4 ENERO 2022

The Data Scientist magazine

COMPUTACIÓN CUÁNTICA

DESCUBRE QUÉ ES LA COMPUTACIÓN CUÁNTICA. ANALIZAMOS SUS BASES FÍSICAS Y MATEMÁTICAS E INTRODUCIMOS ALGUNOS CONCEPTOS BÁSICOS QUE ES NECESARIO CONOCER

ROBOTS EN LA SOMBRA

ENTREVISTAMOS A IGNACIO G.R. GAVILÁN, AUTOR DEL LIBRO "ROBOTS EN LA SOMBRA" Y HABLAREMOS DE RPA, ROBOTS, CHATBOTS Y OTRAS FORMAS DE AUTOMATIZACIÓN COGNITIVA

DATOS SINTÉTICOS

BENEFICIOS DE LOS DATOS SINTÉTICOS PARA ORGANIZACIÓNES DATA-DRIVEN

Cursos/Tutoriales

Python / Pandas / R

¿Qué es MLOPS? II

Conoceremos las ventajas que ofrece MLOps en la estretegia de puesta en producción de modelos ML

Entrevistas

Ignacio G.R. Gavilán

Contenido 100% exclusivo

algoritmos empresas lenguajes formación entrevistas tendencias startups

En este número

01 **HA SIDO NOTICIA....**

OS TRAEMOS LAS NOTICIAS MÁS RELEVANTES EN EL ÁMBITO DE LA CIENCIA DE DATOS Y LA INTELIGENCIA ARTIFICIAL DURANTE LOS ÚLTIMOS MESES. **PAG. 6**

02 **DATOS SINTÉTICOS**

¿QUÉ SON LOS DATOS SINTÉTICOS, PARA QUÉ SIRVEN Y QUE BENEFICIOS APORTAN A LAS EMPREESAS DATA-DRIVEN? MARIO BRICIO, CO-FUNDADOR DE LA EMPRESA DEDOMENA.AI NOS RESPONDE ESTAS Y OTRAS CUESTIONES CON DETALLE. **PAG. 8**

03 **ENTREVISTAS. IGNACIO G.R. GAVILÁN**

ENTREVISTAMOS A IGNACIO G.R.GAVILÁN, GLOBAL DIRECTOR BIG DATA & IA USE CASES EN TELEFÓNICA Y UNO DE LOS PRINCIPALES REFERENTES EN INTELIGENCIA ARTIFICIAL Y CIENCIA DE DATOS DE NUESTRO PAÍS. **PAG. 16**

04 **COMPUTACIÓN CUÁNTICA**

¿QUIERES SABER QUÉ ES LA COMPUTACIÓN CUÁNTICA? ¿CUÁLES SON ALGUNOS PRINCIPIOS DE LA MECÁNICA CUÁNTICA QUE INTERVIENEN EN LA ARQUITECTURA DE UN ORDENADOR CUÁNTICO? ¿QUÉ ES UN QUBIT? ¿CUÁLES SON LOS USOS ACTUALMENTE DE LA COMPUTACIÓN CUÁNTICA? **PAG. 26**

05 **BENEFICIOS DE MLOPS**

ANALIZAMOS LOS PERFILES PROFESIONALES QUE PARTICIPAN EN LA CONFIGURACIÓN DE UN FRAMEWORK MLOPS Y RESALTAMOS ALGUNOS DE LOS BENEFICIOS QUE APORTA LA CORRECTA INTEGRACIÓN DE TALES PERFILES DENTRO DE LA ESTRATEGIA GLOBAL DE LAS EMPRESAS. **PAG. 32**

06 **PANDAS, UNA PODEROSA HERRAMIENTA PARA EL ANÁLISIS DE DATOS. PARTE 3**

CONTINUAMOS CON LA TERCERA ENTREGA DEL CURSO BÁSICO DE PANDAS A CARGO DE NUESTRO COMPAÑERO AMBROSIO NGUEMA. ESTA VEZ OS PRESENTAMOS UN JUPYTER NOTEBOOK PARA QUE PUEDAS REALIZAR TUS EJERCICIOS. **PAG. 37**

07 **CURSO DE R. MANIPULACIÓN DE DATOS CON TIDYVERSE: DPLYR Y TIDYR**

PRESENTAMOS LOS PRINCIPIOS BÁSICOS DEL TIDYVERSE DE R, QUE NOS PERMITE MANIPULAR LOS DATOS PARA DEJARLOS PREPARADOS PARA SU ANÁLISIS Y VISUALIZACIÓN. **PAG. 42**

08 **THE BLACK BOX. MEDIA, PROBABILIDAD Y JUEGOS DE AZAR**

¿SALE RENTABLE COMPRAR LOTERÍA EN NAVIDAD? ¿QUÉ PROBABILIDADES TENGO DE QUE ME TOQUE EL GORDO? ¿EXISTEN NÚMEROS CON MÁS PROBABILIDAD DE SALIR PREMIADOS? ¿QUIÉN ES REALMENTE EL PRINCIPAL GANADOR DEL SORTEO? **PAG. 52**

DEDOMENA

No hay inteligencia sin datos. No hay datos sin privacidad.

Reduce esfuerzos de desarrollo de IA hasta en un **80%**

Mejora la efectividad de tus modelos en un **30%**

Accede a un más de datos **90%**

Datos Sintéticos

Genera copias sintéticas estadísticamente idénticas a tus datos sin que contengan información identificable, garantizando la privacidad y el valor empresarial de los datos.

Herramientas de Anonimización

Ve más allá de las técnicas tradicionales de anonimización. Con Dedomena ya no tendrás que sacrificar valor de los datos por privacidad.

Soluciones IA de Negocio

Extrae el máximo valor de tus datos en cuestión de días, y no en meses o años. Ya tenemos los modelos que tu negocio necesita. Extraer valor de los datos nunca ha sido tan fácil.

Preciso. Fiable. Potente.

dedomena.ai

Editorial

Robots y Cuántica

Pedro Albarracín García
director de "The Data Scientist Magazine"
pedro.albarracin@thedatascientist.es

¿QUÉ SON LOS DATOS SINTÉTICOS? ¿QUÉ BENEFICIOS APORTAN A LAS EMPRESAS DATA-DRIVEN? ¿PODRÍAN LOS ROBOTS EN UN FUTURO PRÓXIMO SENTIRSE AMENAZADOS POR LAS PERSONAS? ¿SERÁN LAS RELACIONES ROBOTS-PERSONAS UNA RELACIÓN DE COLABORACIÓN O COMPETENCIA?

Nos hemos propuesto arrancar con fuerza este 2022 trayendo a The Data Scientist Magazine tres temas de mucha actualidad. Ya hemos hablado en algún número anterior sobre privacidad y los riesgos asociados a la recopilación y tratamiento masivo de los datos personales. Es un tema que preocupa y es por ello que han surgido diferentes iniciativas empresariales con soluciones muy interesantes para la anonimización de datos personales y la generación de datos sintéticos. ¿Datos sintéticos? Mario Bricio, co-fundador de la empresa **dedomena.ai** nos presenta en detalle qué son los datos sintéticos, qué problemas presentan los métodos tradicionales de anonimización, y cuáles son los beneficios de los datos sintéticos para las empresas data-driven.

¿Podrían los robots en un futuro próximo sentirse amenazados por las personas? "Robots en la sombra" es el último libro publicado por editorial Anaya y cuyo autor, Ignacio G.R. Gavilán cuenta con una dilatada experiencia en el campo de la inteligencia artificial, robotización de procesos, y un largo etc, es además, el responsable del área de relaciones humanos-robots en OdiseIA. Hablamos con Ignacio de robots, bots, RPA y muchos otros temas de actualidad.

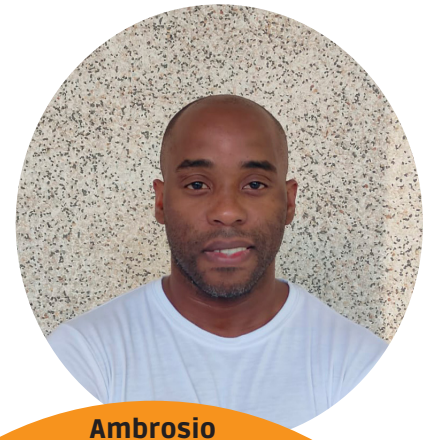
Por último, iniciamos una serie de artículos en el que trataremos en detalle todo lo que necesitas conocer sobre computación cuántica. En el primer artículo que publicamos este mes introducimos los conceptos básicos a modo de guía de iniciación. En próximos números conoceremos además, qué avances en el ámbito de la computación cuántica se están produciendo en España y qué organismos y compañías lideran el desarrollo en este apasionante campo.

Ya estamos trabajando en el número de febrero que dedicaremos por completo al área de People Analytics, con artículos y entrevistas de primer nivel.

Esperamos que los contenidos que hemos preparado estén a la altura y que este año 2022 podamos ver cumplidos nuestros anhelos y deseos tanto en lo personal como en lo profesional. Saludos.

Nuestro Equipo

Gema Fernández-Avilés Calderón es Licenciada en Administración y Dirección de Empresas y Doctora en Estadística. Actualmente es profesora titular de Estadística (Economía Aplicada) en la Universidad de Castilla-La Mancha y directora del Máster en Data Science & Business Analytics (con R software). Entre sus líneas de investigación destacan la Geoestadística Espacial y Espacio-Temporal, la Economía Aplicada, el Análisis Estadístico Multivariante y la Modelización de la Calidad del Aire. Entusiasta del Data Science y del Spatial Data Science. Coordinadora de la Revista y responsable de la sección "The Black Box"



Ambrosio Nguema es Ingeniero de telecomunicaciones y científico de datos, con amplia experiencia en el sector de Inteligencia Artificial (IA) y con una curtida trayectoria laboral en compañías como Ericsson, Jazztel, Indra, CornerJob, Jeff o Inditex, donde trabaja en estos momentos a través de la consultora Merlín Software como especialista NLP Senior, ayudando a que todos los productos de Inditex estén bien clasificados y categorizados, es también Autor del libro "Ciencia de Datos para adolescentes". Como parte del equipo de la redacción de la revista, Ambrosio lidera el área técnica en el ámbito del desarrollo Python y es responsable de la sección "Python para todos los públicos"

Bilal Laouah es graduado en Psicología del trabajo y de las organizaciones y máster en Big Data y Business Intelligence. Actualmente se dedica a la docencia en metodologías de la investigación cuantitativa y a la consultoría en formación y capacitación en áreas de IT. Participa en el equipo de la redacción liderando las áreas de innovación y control de proyectos.

The Data Scientist Magazine
ISSN 2792-3592
Alcalá de Guadaíra, Sevilla
revista@thedatascientist.es

Publicidad
publicidad@thedatascientist.es

NOTICIAS

CIENCIA DE DATOS / IA HA SIDO NOTICIA...

ESPAÑA COMPETIRÁ CON LOS LÍDERES DE LA COMPUTACIÓN CUÁNTICA MUNDIAL

El Gobierno de España aprobó una subvención de 22 millones de euros, ampliables hasta 60 durante el próximo trienio, para "impulsar la creación de un ecosistema de computación cuántica en España. El objetivo del proyecto "Quantum Spain" es "dar acceso a las empresas y al sector público para desarrollar un computador cuántico de altas prestaciones que se pondrá a disposición de la comunidad investigadora para el desarrollo de la Inteligencia Artificial, fortaleciendo el desarrollo tecnológico e industrial en España y la creación de empleo de alta cualificación". Según Carme Artigas, el proyecto tendrá su aplicación sobre problemas reales de sectores como química, finanzas, optimización de procesos de la cadena productiva y criptografía, entre otros.

ENLACE AL ECONOMISTA.ES

¿APRENDES IA O CON IA?

Mónica Villas, consultora de nuevas tecnologías, habla de la creciente demanda de formación en el ámbito de la inteligencia artificial y de los esfuerzos que muchos países están realizando en este sentido a través de sus estrategias nacionales de inteligencia artificial, tal y como se muestra en el informe publicado por la Universidad de Stanford "2021 AI Index Report". Por otro lado también pone manifiesto la necesidad de hacer uso del potencial de la inteligencia artifical como herramienta para mejorar el aprendizaje de cualquier materia. El reto a superar es llegar a la formación personalizada utilizando las técnicas de análisis del aprendizaje (learning analytics) a partir de de los datos obtenidos de los estudiantes.

ENLACE A EL ESPAÑOL

La Razón

La inteligencia artificial podría prevenir los suicidios en adolescentes. Mediante una encuesta de 20 preguntas, correctamente ponderadas, han permitido a la IA predecir los pensamientos suicidas con un 91% de aciertos, esto es: que de cada 10 encuestas analizadas, solo se equivoca en un adolescente a la hora de predecir si tiene pensamientos suicidas o no.

Enlace a La Razón

Libertad Digital

Cinco propuestas para saber más de inteligencia artificial. Obras de actualidad y clásicos reconocidos sobre el campo de la IA.

Enlace a Libertad Digital

La Vanguardia

La comunidad científica alerta sobre la ncesidad de regular el uso de la inteligencia artificial mediante tratados internacionales, un código de conducta para investigadores y una legislación completa debido "a que esta tecnología está empezando a tener un impacto realmente grande en el mundo real". Según declaraciones de el profesor Stuart Russell, fundador del Center for Human-Compatible Artificial Intelligence de la Universidad de California

Enlace a La Vanguardia

MINSAIT (INDRA) LANZA PLAIGROUND, UNA UNIDAD DE NEGOCIO ESPECIALIZADA EN INTELIGENCIA ARTIFICIAL

Minsait, empresa perteneciente a Indra, ha creado una nueva unidad de negocio, denominada Plaiground, para facilitar a las empresas el acceso a la inteligencia artificial a través de un entorno colaborativo. A través de esta unidad, la tecnológica española quiere ofrecer a sus clientes "la respuesta más adecuada" a los retos que se les plantee en este campo.
Minsait señala que uno de los claros ejemplos de la materialización del modelo de Plaiground es la reciente creación de AI Lab Granada, "uno de los centros de inteligencia artificial más avanzados de Europa", por parte de Minsait y la Universidad de Granada, con Google Cloud como socio tecnológico.
Este centro contará con más de 100 doctores en inteligencia artificial, 165 consultores, desarrolladores e investigadores y un ecosistema de startups y emprendedores, gracias al papel clave de impulso que ha jugado la Junta de Andalucía.

ENLACE A CINCO DIAS

GESTORES DE FLOTAS DE ROBOTS. LOS JEFES DE PERSONAL DEL FUTURO

Muchas de las tareas que se realizan en todas las profesiones tienen un grado potencial de automatización. Los dispositivos de inteligencia artificial están automatizando tareas que hace unos años sólo cabría imaginar que realizaran máquinas en películas de ciencia ficción.
La función principal de estos gestores de flotas de robots es la planificación y seguimiento de las actividades automatizadas que realizan los robots. Los conocimientos de robótica y programación, al menos por el momento, son indispensables para asumir esta responsabilidad

ENLACE A EL MUNDO

UN CENTENAR DE ENTIDADES EXIGE A LA UE UNA REGULACIÓN ESTRICTA DE LA INTELIGENCIA ARTIFICIAL

Diario Sur

La startup malagueña WeVoice premiada por usar la inteligencia artificial para mejorar la salud mental. La aplicación analiza un entramado de datos acumulados a base de detectar patrones de sueño y actividad, tiempo de uso de redes sociales y anomalías en el tecleo, el tono de la voz y otros hábitos. Aunque está pensada para trastornos leves, la plataforma WeVoice, puede servir como apoyo de tratamientos clínicos en casos más graves

Enlace a Diario Sur

La Vanguardia

El canal de Youtube "Netflix is a Joke" ha compartido la primera película de terror escrita enteramente por bots con el título "Mr Puzzles wants you to be less alive". Una inteligencia artificial escribió el guión después de visualizar más de 400.000 horas de cine de terror. El resultado es una mezcla de historias bajo una trama absurda sin ninguna originalidad.

Enlace a La Vanguardia

Diario.es

Tres hospitales usan inteligencia artificial para acelerar la detección de la Covid. Aúnan experiencia clínica en la lectura de la radiografía de tórax, de manera que aceleran la detección y mejoran la precisión diagnóstica de la covid con esta prueba.

Enlace a Diario.es

DATOS SINTÉTICOS

BENEFICIOS PARA LAS EMPRESAS DATA-DRIVEN

Autor: Mario Bricio, co-fundador y COO de Dedomena.ai

Estamos en la segunda década del siglo XXI y las organizaciones están empezando a asimilar los beneficios del análisis avanzado de grandes volúmenes de datos, lo que conocemos por el término Big Data, impulsado por un sinfín de aplicaciones de Inteligencia Artificial y Machine Learning, mientras, en paralelo, las consecuencias para la privacidad de las personas es un tema que cada vez preocupa más a clientes e instituciones. Pero, detengámonos a pensar en un asunto; lo que hace realmente poderoso al Big Data, con las herramientas y modelos de Machine Learning que lo explotan, por encima de las tecnologías e infraestructura necesaria para operarlos, es contar con grandes volúmenes de datos de calidad; que en la mayoría de los casos, corresponde con información sensible que hace existan barreras que bloquean, o en el mejor de los casos, retrasan, el desarrollo y adopción de este tipo de tecnologías.

LOS DATOS POR SÍ SOLOS TIENEN MUY POCO VALOR, EL VERDADERO VALOR SE ENCUENTRA EN EL PROCESO DE REFINACIÓN LAS ORGANIZACIONES MÁS RICAS Y PODEROSAS SERÁN LAS QUE SEPAN ADQUIRIR Y GESTIONAR SUS DATOS DE FORMA ÓPTIMA LOS DATOS NO SON UN RECURSO FINITO

Todos estamos de acuerdo en que los datos son el nuevo petróleo de este siglo, pero al igual que el petróleo, los datos por si solos valen muy poco, el verdadero valor se encuentra en el proceso de refinación y los derivados que se pueden extraer, multiplicando exponencialmente su utilidad. No es casualidad que los países más ricos y poderosos del mundo sean los que mejor uso hacen de sus reservas de petróleo, lo mismo ocurrirá con los datos, las organizaciones más ricas y poderosas serán las que sepan adquirir y gestionar sus datos de

LA MAYORÍA DE LAS EMPRESAS, TANTO CLIENTES FINALES COMO PROVEEDORES DE TECNOLOGÍA, ESTÁN BLOQUEADAS O NO SON REALMENTE ÁGILES EN CUANTO A LA INNOVACIÓN Y LA GENERACIÓN DE MAYORES BENEFICIOS EN TORNO A LOS DATOS Y LA INTELIGENCIA ARTIFICIAL. LA PRIVACIDAD ES ESTRICTAMENTE NECESARIA, Y ASEGURARLA CONLLEVA DEMORAS EN EL ACCESO A DATOS Y PÉRDIDA DE CALIDAD DE LOS MISMOS. CADA VEZ SE NECESITAN MÁS DATOS, MÁS DIVERSOS, MÁS RÁPIDOS Y SOBRE TODO DE MAYOR CALIDAD. LOS DATOS SINTÉTICOS PROPORCIONAN ÉSTAS Y MUCHAS MÁS VENTAJAS A LAS EMPRESAS QUE ADOPTAN NUESTRA TECNOLOGÍA COMO CATALIZADOR DE SU ESTRATEGIA DATA-DRIVEN

manera óptima.

La gran diferencia entre el petróleo y los datos como motor de la economía, es que estos últimos no son un recurso finito, sino que cada vez se genera más y más cantidad a medida que avanza la revolución digital. Se podría decir que los datos son un recurso que se va regenerando con su propio uso. A medida que las empresas van transformando sus infraestructuras tecnológicas, modelos de negocio y la manera en que interactúan con sus clientes, se van generando nuevos puntos de información y recolección de datos que retroalimentan la cadena de valor.

Sin embargo, los datos tienen un importante hándicap que hasta ahora parecía no tener solución, y es que en la mayoría de los casos, se trata de información sensible y como es debido, sujeta a distintas regulaciones de protección de datos. De hecho, según Gartner (1), se estima que el 65% de la población mundial tendrá sus datos personales cubiertos por alguna de las nuevas leyes de privacidad de datos en los próximos años.

Siempre se ha pensado que la manera más sencilla y eficiente que tienen las empresas para acceder y compartir información de valor con colaboradores internos y externos de manera segura, era la anonimización de datos usando técnicas como privacidad diferencial, adicción de ruido, sustitución o tokenización. No obstante, se ha demostrado que estás técnicas no son útiles para anonimizar grandes volúmenes de datos, y mucho menos a la velocidad que se necesita para poder procesarlos en tiempo real.

¿Por qué los métodos tradicionales de anonimización no son la solución?

El Big Data es difícil de anonimizar, o para usar un término más técnico, es propenso a re-identificación. Se dice que un dato es propenso a ser re-identificado cuando, a través de inferencias, aislamientos, análisis o enlazamientos de la información en sí y/o con información pública o complementaria, se puede descubrir el sujeto al que pertenecen. Este es, sin lugar a duda, uno de los principales riesgos de la privacidad de la información, pero también es el menos conocido.

Lamentablemente, son muy comunes las filtraciones de datos o “data breach” por su término en inglés , lo cual es un signo claro de que los métodos de anonimización y la gestión de la privacidad de la información no están funcionando como se esperaban. Hoy en día, en empresas que poseen o procesan datos, con un elevado número de empleados , que no siempre son miembros de equipos internos de análisis y ciencia de datos, tienen acceso a grandes volúmenes de información privada y sensible. A pesar de los esfuerzos actuales por parte de los equipos de IT y gobierno del dato, este acceso es especialmente vulnerable a este tipo de situaciones.

En los últimos meses, hemos observado cómo los mismos clientes y usuarios son los que cada vez se preocupan más por la privacidad de sus datos personales. Según el estudio de CISCO (Cisco Consumer Privacy Study, 2021) (2), el 86% de los encuestados afirman que están preocupados por la seguridad de sus

DEDOMENA AI ES UNA START-UP DE INTELIGENCIA ARTIFICIAL QUE BRINDA SOLUCIONES DE ANONIMIZACIÓN, GENERACIÓN DE DATOS SINTÉTICOS Y EXTRACCIÓN DE VALOR DE LOS DATOS, SIENDO PIONEROS EN ESPAÑA Y SUR DE EUROPA EN EL DESARROLLO DE UNA PLATAFORMA COMPLETA DE INTELIGENCIA ARTIFICIAL USANDO LOS DATOS SINTÉTICOS COMO PARTE CENTRAL DE SU PROPUESTA DE VALOR.

datos, y el 50% de ellos ha cambiado o están dispuestos a cambiar de proveedor de servicios debido a las políticas de privacidad de la información de estas empresas.

Ahora bien, no son sólo los clientes los que están preocupados por la privacidad de la información. Las autoridades competentes y reguladores también son conscientes de la urgencia de endurecer las políticas de protección de datos para asegurar adecuadamente la privacidad de los individuos y las empresas: buscan que no sea posible utilizar fácilmente los datos de los clientes sin procesar o sin consentimiento, incluso dentro de las mismas organizaciones.

La Unión Europea presentó en 2018 el Reglamento General de Protección de Datos, más conocido como GDPR (General Data Protection Regulation) por sus siglas en Inglés, poniendo en práctica las reformas en protección de datos y privacidad que se venían debatiendo durante los últimos años. Este constituyó el primer paso hacia una definición más unificada de los derechos de privacidad ciudadana, escalando mundialmente después de su lanzamiento hacia otros países y regiones, siguiendo las mismas pautas y principios de los socios europeos. Países como Brasil, Canadá o China ya han creado su propia regulación en protección de datos, seguidos de otras tantas como la CCPA en California o la PDPB en India. Esta tendencia es un viaje sin retorno, ya que cada vez más países y regiones están endureciendo sus políticas de privacidad.

La mayoría de las técnicas actuales de anonimización de datos, en realidad, no son más que métodos de pseudo-anonimización. La misma GDPR define este término en su Artículo 4 (3), que dice textualmente: "Los datos personales que se hayan sometido a un proceso de pseudo-anonimización y que puedan atribuirse a una persona física mediante el uso de información adicional deben considerarse información sobre una persona física identificable". En este sentido, la regulación es bastante explícita, los datos resultantes de un proceso de pseudo-anonimización no son anónimos, y por lo tanto, deben regirse por los mismos principios que los datos personales. La información sensible, aunque haya pasado por un proceso de pseudo-anonimización, es fácil de ser identificada aplicando en muchos casos sólo ingeniería inversa; es por esto que los datos personales pseudo-anonimizados son un blanco fácil para los ataques a la privacidad.

En el otro extremo, técnicas más clásicas y conocidas de anonimización como la permutación, aleatorización y generalización, tienen el inconveniente de que "destruyen" los datos, o mejor dicho, el valor y la información que contienen debido a la agresividad de los métodos, sin por ello, garantizar al 100% la seguridad de la información.

La permutación es una técnica que consiste en alterar el orden de los registros de la base de datos para que no se correspondan idénticamente con la información original. Se trata de una modificación engañosa, ya que los datos pueden fácilmente permutarse de forma reversible si son cruzados con datos adicionales y por otro lado, se produce una gran pérdida de valor estadístico en cuanto a correlaciones, relaciones entre columnas, y por ende, poder predictivo.

LA MAYOR PARTE DE LOS DATOS SON INFORMACIÓN SENSIBLE Y POR LO TANTO SUJETOS A REGULACIONES DE PROTECCIÓN DE DATOS LOS MÉTODOS DE ANONIMIZACIÒN Y LA GESTIÓN DE LA PRIVACIDAD DE LA INFORMACION NO ESTÁN FUNCIONANDO COMO SE ESPERABA CADA VEZ MÁS CLIENTES SE PREOCUPAN POR LA PRIVACIDAD DE SUS DATOS PERSONALES

La aleatorización es otro enfoque bastante conocido, en el que las variables de los datos son modificadas de acuerdo a patrones aleatorios que son definidos previamente. La técnica más conocida es quizás la perturbación o adición de ruido, que consiste básicamente en añadir ruido sistemático al conjunto de datos. Por ejemplo, en un conjunto de datos que contiene las fechas en las que un paciente acudió al hospital, estos valores pueden ser ajustados aleatoriamente sumando o restando el mismo número de días a la fecha real de la visita. En este caso, al contrario que la permutación, algunas relaciones entre los datos y sus variables son preservados, aunque tampoco garantiza la privacidad de la información, ya que estos patrones pueden ser identificados y con ello identificar el dato original sensible.

La generalización, como su propio nombre indica, generaliza los datos diluyendo sus características. Se trata de convertir datos individuales en datos más genéricos o agregados, en los que no quepa sólo un individuo, sino un grupo de ellos, haciendo imposible identificar a alguno de los individuos originales por separado. El objetivo es convertir un dato específico en uno genérico, reduciendo la granularidad del mismo. Una de las técnicas de generalización más usadas es la conocida como K-anonymity. Al aplicar esta técnica, se debe elegir el parámetro "k" que define el balance entre privacidad y utilidad de los datos. No obstante, aún tomando un valor alto para "k", los problemas de privacidad permanecen tan pronto como la información sensible se vuelva homogénea. En tales casos, los datos se vuelven susceptibles a los llamados ataques de homogeneidad. Algunos autores proponen nuevas variantes de técnicas de generalización para evitar este tipo de riesgos, siendo las más conocidas l-diversity y t-closeness Sin embargo, incluso estas variantes, son insuficientes para garantizar la privacidad de la información en su conjunto.

Como consejo, creo que un auditor de algoritmos está guiado por la inquietud y la curiosidad de conocer cómo funciona el mundo y la voluntad de mejorarlo.

En general, el proceso de re-identificación es mucho más difícil con estas técnicas clásicas de anonimización que con técnicas de pseudo-anonimización. La reidentificación implica búsqueda de información adicional y complementaria a los datos que se pretenden re-identificar. Aún así, los ataques y casos de pérdidas de datos suceden con alarmante regularidad. A medida que se dispongan de más datos abiertos e interconectados, habilitados por las nuevas tecnologías y políticas, el riesgo de nuevos ataques a la privacidad aumentará exponencialmente si las organizaciones no toman medidas adecuadas, afectando negativamente su negocio y reputación.

Independientemente de los criterios que se tomen para prevenir la reidentificación de los datos personales y sensibles, siempre habrá un balance entre privacidad y utilidad de los datos. Con las técnicas actuales, los datos que se consideran completamente anonimizados no tendrán un gran valor estadístico, informacional y predictivo, y viceversa.

Las empresas orientadas al dato, por lo tanto, se encuentran con un importante dilema entre privacidad y utilidad de los datos, y en la práctica totalidad de los casos, esto lleva a disputas internas entre los responsables de asegurar la privacidad de la información y los encargados de extraer valor de ella. Es muy probable que la solución a todos estos problemas ya haya sido descubierta, quizás sea tan fácil como que las empresas pudieran crear sus propios datos: la respuesta está en los datos sintéticos.

¿Qué son los datos sintéticos y para qué sirven?

Los datos sintéticos son datos artificialmente generados a partir de los datos originales pero que mantienen las mismas características estadísticas, informacionales y predictivas. Mientras que los datos reales se recopilan en cada una de las interacciones físicas o digitales con las personas y a través de los

procesos internos, los datos sintéticos son generados con un algoritmo. Este algoritmo o modelo sintético es capaz de generar nuevos conjuntos de datos completamente nuevos y artificiales. En la creación de datos sintéticos está implícito un proceso de anonimización real, es decir, los datos sintéticos son un conjunto de datos 100% anónimos, ya que es imposible su re-identificación, a diferencia de las técnicas anteriormente comentadas en este artículo.

El término "datos sintéticos" no es nuevo, fue introducido por Donald B. Rubin, profesor de Estadística de la Universidad de Harvard, a comienzos de la década de los 90, cuando estaba ayudando al censo de los Estados Unidos a solventar problemas de recuento insuficiente, especialmente en zonas más pobres. Sin embargo, el auge cada vez mayor de la Inteligencia Artificial y el desarrollo de nuevos algoritmos de aprendizaje profundo o "deep learning" han acelerado el desarrollo y uso de nuevas herramientas de generación de datos sintéticos.

Los datos sintéticos satisfacen necesidades específicas y ciertas condiciones que no se pueden encontrar en los datos originales (reales), y esto es muy importante ya que abre la puerta a un océano de posibilidades para desarrolladores y científicos de datos. La adopción generalizada del uso de datos sintéticos por las organizaciones dotará a los desarrolladores con una poderosa herramienta para crear y disponer de la cantidad de datos que deseen, para así poder desarrollar y entrenar algoritmos de inteligencia artificial de forma rápida y asequible, nuevas aplicaciones y productos, probar entornos y nuevas funcionalidades, garantizando siempre la calidad y la privacidad de los datos.

Los datos sintéticos tendrán un gran impacto en la industria de la Inteligencia Artificial, ya que estos avances en los modelos generativos conducirán a un mejor rendimiento de los modelos de Machine Learning actuales y por venir. Hay quienes afirman que no es posible el desarrollo de Inteligencia Artificial eficiente sin el uso de datos sintéticos, y cada vez son más empresas las que están poniendo el foco en esta nueva tendencia. Según Gartner, en el año 2022, el 40% de los modelos de IA y ML estarán desarrollados y entrenados con este tipo de datos, y el porcentaje podría llegar al 60% para el año 2025 (4).

La adopción de los datos sintéticos como nuevo combustible para acelerar negocios orientados al dato dará como resultado mejores aplicaciones, productos y servicios, una completa comprensión del comportamiento de sus clientes y colaboradores, y por ende, mejores resultados para las empresas que ven el dato como parte central de la nueva economía digital. El cielo será el límite para la innovación en torno al dato, accediendo a conjuntos de datos que antes no era posible acceder.

Además, las empresas podrán extraer valor adicional y generar nuevas formas de monetizar sus activos de datos, ya que contarán con las capacidades necesarias para generar grandes volúmenes de datos anónimos con un alto valor industrial que podrán compartir o vender con terceros de forma segura. Las organizaciones tendrán total flexibilidad para intercambiar valor en forma de datos, sin necesidad de disponer de consentimiento expreso de sus clientes en muchos casos, ni poner en riesgo la información de los mismos.

Es por todo esto, que cada día escuchamos más el término "revolución de los datos sintéticos", y no se puede ser más preciso. Tener acceso a un gran volumen de datos sintéticos pondrá a las empresas que lo sepan aprovechar a la cabeza de la innovación, no sólo en su industria tradicional, sino en otras industrias antes inexploradas al acceder a nuevos datos e información sobre los que apalancar su crecimiento.

¿Qué beneficios ofrecen los datos sintéticos a las empresas data-driven?

Los beneficios de utilizar datos sintéticos para las organizaciones pueden llegar a ser realmente

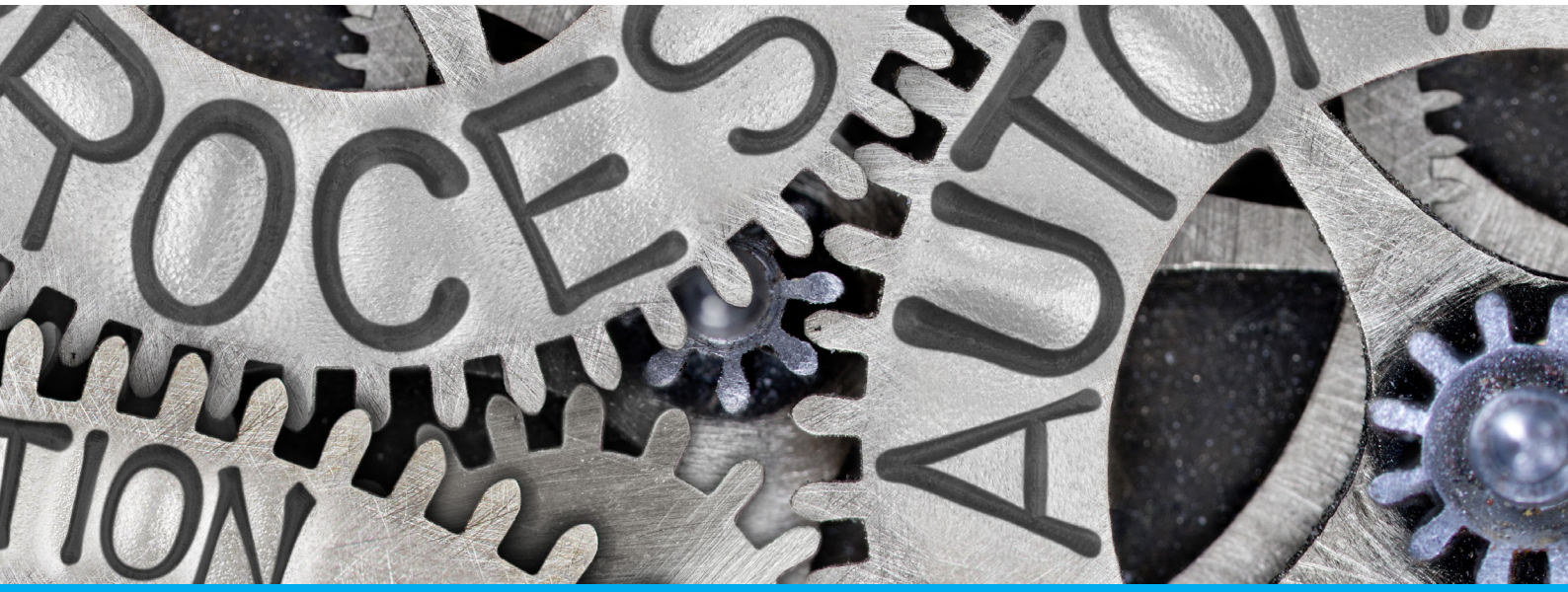
asombrosos, ya que pueden hacer factibles proyectos imposibles, acelerar significativamente las iniciativas de Inteligencia Artificial, mejorar sustancialmente los resultados de los algoritmos de Machine Learning, y como consecuencia de todo lo anterior, maximizar la monetización de su activo más preciado en la era digital, el dato.

El acceso a los datos es sin duda uno de los mayores desafíos a los que se enfrentan las empresas a la hora de desarrollar, implementar y aplicar Inteligencia Artificial y Machine Learning en sus procesos de negocio. Las organizaciones que quieran transformarse digitalmente necesitarán acceder y disponer de datos durante la mayor parte de su cadena de valor: se requieren datos para desarrollar, entrenar y validar modelos de Machine Learning, para probar aplicaciones y nuevos entornos, hacer demos o demostrar funcionalidades con nuevos clientes, o migraciones al Cloud, también para probar y evaluar tecnologías de IA desarrolladas por terceros.

Gracias a los datos sintéticos, las compañías se beneficiarán de tiempos de acceso al dato de hasta 10 veces más rápidos que con las técnicas actuales, beneficiándose por ende de un ahorro de costes de hasta un 80% en este tipo de actividades. Las limitaciones de usar los datos originales serán eliminadas gracias al uso de datos sintéticos, que además pueden ser generados a demanda sin necesidad de que ciertos eventos ocurran en la realidad, todo ello cumpliendo por definición con las más estrictas leyes de protección de datos.

Por otro lado, disponer de datos sintéticos de calidad ayuda considerablemente a extraer un valor superior de los mismos, ya que los modelos de Machine Learning entrenados con este tipo de datos artificiales pueden llegar a tener un desempeño de hasta un 30% mejor que los modelos entrenados sólo con los datos originales. Hasta ahora, los datos sintéticos han ganado una fuerte tracción en casos de uso como entrenamiento de vehículos autónomos, la atención médica digital o la detección de fraude. Reconocidas empresas tecnológicas como Uber o Google hacen uso de los datos sintéticos para el entrenamiento de sus vehículos autónomos, también Amazon utiliza este tipo de datos para entrenar su asistente virtual Alexa.

Por último y no menos importante, los datos sintéticos son la única solución hasta el momento viable para crear un entorno seguro donde compartir datos entre empresas y sectores. Aunque la cantidad de datos generados aumenta año tras año sin precedentes (se espera que la cantidad se duplique de 2022 a 2025), también lo hace la necesidad de que estos datos fluyan entre departamentos, empresas y fronteras, sin embargo el entorno regulatorio para que



esto sea posible está totalmente fragmentado. A día de hoy existen a nivel mundial más de 120 regulaciones diferentes en torno a la privacidad de la información, y este número continuará aumentando debido a la creciente preocupación de los clientes por su privacidad, lo que hace aún más necesario que nunca encontrar una solución única para superar este desafío.

LOS DATOS SINTÉTICOS SON DATOS ARTIFICIALMENTE GENERADOS A PARTIR DE LOS DATOS ORIGINALES PERO QUE MANTIENEN LAS MISMAS CARACTERÍSTICAS ESTADÍSTICAS, INFORMACIONALES Y PREDICTIVAS

LA ADOPCIÓN DE LOS DATOS SINTÉTICOS COMO NUEVO COMBUSTIBLE PARA HACER NEGOCIOS ORIENTADOS AL DATO DARÁ COMO RESULTADO MEJORES APLICACIONES, PRODUCTOS Y SERVICIOS

REFERENCIAS

1. HTTPS://WWW.GARTNER.COM/EN/NEWSROOM/PRESS-RELEASES/2020-09-14-GARTNER-SAYS-BY-2023--65--OF-THE-WORLD-S-POPULATION-W
2. HTTPS://WWW.CISCO.COM/C/DAM/EN_US/ABOUT/DOING_BUSINESS/TRUST-CENTER/DOCS/CISCO-CYBERSECURITY-SERIES-2021-CPS.PDF
3. HTTPS://GDPR-TEXT.COM/READ/ARTICLE-4/
4. HTTPS://BLOGS.GARTNER.COM/ANDREW_WHITE/2021/07/24/BY-2024-60-OF-THE-DATA-USED-FOR-THE-DEVELOPMENT-OF-AI-AND-ANALYTICS-PROJECTS-WILL-BE-SYNTHETICALLY-GENERATED/

ENTREVISTAS

IGNACIO G.R.GAVILÁN

"ROBOTS EN LA SOMBRA"

EL PAPEL DE ODISEIA ES MÁS QUE NECESARIO, ES IMPORTANTE DISPONER DE UNA VOZ NEUTRAL Y AUTORIZADA QUE ILUMINE EL USO ÉTICO Y RESPONSABLE DE LA IA Y QUE DISEMINE ESE CONOCIMIENTO ENTRE EL GRAN PÚBLICO

ES MUY IMPORTANTE SER CAPACES DE DISEÑAR UN PLAN DE IMPLANTACIÓN REALISTA Y DE VALOR Y CONSEGUIR ALINEAR A LA TOTALIDAD DE LA COMPAÑÍA EN TORNO A ÉL

Conocí a Ignacio al poco tiempo de incorporarme como socio en OdiseIA, el Obervatorio del Impacto Social y Ético de la IA. Me llamó la atención su trabajo en el área que dirige, "relaciones robots-humanos" y desde entonces se revisitado su participación en redes, charlas, etc. Siempre es un placer escuchar a Ignacio, tiene mucho que decir y aporta una visión muy didáctica sobre aspectos de la inteligencia artificial y la robótica quizás poco accesibles en algunos casos. Recientemente ha publicado su último trabajo titulado "Robots en la sombra" y aprovechando esa circunstancia hablamos con él de robots, RPA, bots y muchas otras cosas interesantes.

TDS: Ignacio, háblanos un poco de tu trayectoria profesional y de cómo nace tu interés por la inteligencia artificial

IG: Bueno, mi trayectoria profesional es ya bastante larga, lo cual es bueno por lo que tiene de experiencia y perspectiva, pero no tanto por lo que supone de edad (esto es una pequeña ironía).

Actualmente, y desde hace más de tres años trabajo bajo una firma propia, Reingeniería Digital en que intento aunar la visión tecnológica con la de negocio, especialmente bajo su faceta de procesos y transformación digital, y con especial foco en la automatización inteligente y la robotización de procesos. Bajo esa firma ofrezco servicios de asesoría a empresas,

IGNACIO G.R.GAVILÃN

MI VISIÓN RESPECTO A LA TECNOLOGÍA SIEMPRE ES Y SERÁ POSITIVA, A PESAR DE LOS RIESGOS LA TECNOLOGÍA, EN SU CONJUNTO, ES CLARA IMPULSORA DEL DESARROLLO Y BIENESTAR ECONÓMICO Y MATERIAL EN ESPAÑA ESTAMOS LEJOS DE APROVECHAR LAS TECNOLOGÍAS EN TODO SU POTENCIAL

formación y conferencias, además de mi labor como escritor.

Anteriormente estuve 25 años en Telefónica donde pasé por las áreas de Investigación y Desarrollo, Grandes Clientes y, finalmente, Operaciones y Red.

Toda mi carrera ha estado relacionado pues, en diferentes roles y perspectivas con la tecnología. Y durante toda mi carrera profesional he tenido mucho interés en la robótica y en la inteligencia artificial pero esta relación, como el propio desarrollo de la Inteligencia Artificial, ha tenido sus inviernos, épocas de mucha dedicación a la inteligencia artificial y épocas de estar muy al margen de la misma.

Hace ya muchos años cursé los estudios de doctorado en un programa de la UPM que se titulaba "Control de procesos e Inteligencia Artificial", aunque no pude traducirlo luego en una en una tesis doctoral por incompatibilidad absoluta a nivel horario y de exigencia con mi actividad profesional. Pero de aquella época conservo buena y muy abundante bibliografía y una afición e interés intactos hoy día.

Desde que me establecí bajo mi propia firma, hace ya más de tres años, la automatización inteligente, el uso de la inteligencia artificial en procesos y la robótica software han sido uno de mis focos de estudio e investigación y también de prestación de servicios profesionales.

Además, estoy adentrándome también en la investigación de la robótica social y la relación robots-personas que es un tema apasionante bajo todos los puntos de vista: tecnológico, psicológico, filosófico, ético. Y, especialmente en este apartado, y desde mi incorporación a OdiseIA hace ya más de año y medio estoy profundizando en su perspectiva ética y social mientras guardo para mi estudio particular la perspectiva tecnológica.

TDS: Como Director de Operaciones y del área de Robots-Personas en OdiseIA, el Observatorio del

Impacto Social y Ético de la IA. ¿Cómo definirías el papel de OdiseIA actualmente en el ámbito de la inteligencia artificial en nuestro país? Y ¿Cuál es el enfoque de los trabajos en el área que diriges?

IG: El papel de OdiseIA es más que necesario. Es casi imprescindible. En un momento como el actual en que hay tanto desarrollo tecnológico pero también tanta confusión en torno a la inteligencia artificial y con sus implicaciones no sólo de negocio sino también éticas y sociales, es muy importante disponer de una voz neutral y autorizada que ilumine el uso ético y responsable, que disemine ese conocimiento entre el gran público, que cree esa conciencia, que interactúe también con las administraciones y el tejido empresarial y que, incluso, pueda participar activamente en la elaboración de guías éticas, como es el caso de nuestro proyecto estrella, GuIA, en acciones de formación o en otros proyectos de acción ética y social.

En cuanto a mi actividad en la asociación realizo dos funciones. En la faceta de dirección de operaciones se trata de trabajo interno, intentando hacer lo más fluido y eficiente posible el funcionamiento de lo que denominamos áreas que son la columna vertebral del trabajo de OdiseIA, facilitar la interacción entre ellas y con el núcleo de gestión de la asociación. No es una tarea sencilla en un entorno de alto talento y creatividad pero también de colaboraciones desinteresadas y compatibilizadas con otras actividades.

Más interesante, quizá, visto externamente, es mi trabajo al frente del área Relaciones Robots-Personas donde trabajamos con las implicaciones sociales y éticas de los robots, especialmente en su relación con los humanos. Estamos enfocados en tres grupos de soluciones:

- **Robots sociales:** es decir, un tipo de robots orientados, específicamente, a su relación con las personas. Se trata de robots con frecuencia humanoides o zoomorfos con capacidades de interacción social, no solo mediante la voz y el lenguaje natural, sino también, mediante el

uso de la distancia social y el gesto, mediante la detección y expresión de emociones. Para aquellos lectores que no los conozcan estaríamos hablando de robots como Pepper, Nao o la famosísima Sophia.

- **Agentes conversacionales inteligentes:** nos referimos las versiones más avanzadas de robots conversacionales (chatbots, voicebots, altavoces inteligentes) donde, aunque con menos recursos que los robots sociales, por no disponer de un cuerpo físico, sí se pueden establecer vínculos emocionales con las personas, especialmente con el nuevo uso de sofisticados avatares realistas.
- Brain Computer Interface: es decir, las relaciones de humanos y máquinas pero ahora no mediante mecanismos sociales como el lenguaje o el gesto sino mediante interacción directa con el cerebro y sistema nervioso humanos mediante sensores y actuadores especializados.

Se trata de áreas desafiantes, de alto interés tecnológico pero con unas implicaciones sociales y filosóficas muy especiales, con mucho peso de las relaciones, de las emociones y de la propia concepción de la naturaleza humana.

Trabajamos en el estudio y divulgación multidisciplinar, incluyendo la visión técnica, la filosófica, la legal y la social pero, sobre todo, tengo mucho interés en transmitir y poner en práctica la capacidad de estas soluciones, de esos robots, para el bien, para el acompañamiento de ancianos, para la ayuda en casos de autismo, para el alivio del Parkinson, para la rehabilitación, para la ayuda en casos de demencia, etc.

Sin ignorar ni ocultar los riesgos, que existen, buscamos además una ética positiva y de acción, destacando las posibilidades y no sólo los riesgos y buscando, como gran aspiración, la participación en proyectos reales de implantación de soluciones de acción social positiva.

TDS: Recientemente has publicado tu último trabajo "Robots en la Sombra" de la editorial Anaya. Es un libro fascinante en el que hablas de robots software, chatbots, agentes o asistentes virtuales, etc., un "ejército" robótico intangible que está muy presente en nuestras vidas. ¿Ves un futuro en el que estas formas robóticas avanzadas puedan representar algún peligro para los seres humanos?

IG: Mi visión respecto a la tecnología siempre es y será positiva. No es que no haya riesgos, claro que los hay.

La automatización, y en concreto la automatización inteligente evidentemente elimina trabajo humano pero también crea nuevo trabajo por otro lado. Si sucede lo que en anteriores revoluciones tecnológicas, el impacto neto debería ser positivo, pero eso no quiere decir que, por el camino, no pueda haber situaciones de desempleo y sectores, empresas y sobre todo personas que puedan sufrir. Eso está ocurriendo y ocurrirá.

Y existen riesgos éticos, especialmente en el apartado de los robots conversacionales que pueden ser usados para espionaje o para un mal uso, intencionado o no, de las relaciones afectivas.

Pero estoy convencido de que la tecnología, en su conjunto, y en este caso la robótica software, es clara impulsora del desarrollo y bienestar económico y material que, en el fondo, constituye la base sobre la que construir, con más garantías, una sociedad más avanzada, justa e igualitaria. Creo que aunque algunas empresas o personas resulten perjudicadas es mucho mayor el beneficio conjunto, las mejoras en las vidas de la mayoría.

Dicho de otra manera, creo que hay riesgos y que hay y habrá perjuicios pero de alguna forma acotados, en momentos concretos, en casos concretos, en perfiles concretos, en personas concretas o en empresas concretas, pero que el conjunto es tremendamente

beneficioso para la sociedad y la mayor parte de empresas e instituciones.

TDS: La relación robots-personas ¿será una relación de colaboración o de competencia? ¿Podrían los robots en un futuro próximo sentirse amenazados por las personas?

IG: Quizá, más que hablar de competencia o colaboración, preferiría expresarlo como sustitución o colaboración.

Dicho eso, lo importante es entender que la relación entre robots y personas será, en el fondo, la que nosotros, los humanos, queramos que sea. Estamos al mando, absolutamente al mando, y no veo indicios creíbles de que esa situación se pueda revertir en un futuro previsible.

Dicho eso, creo que lo que parece razonable y previsible es que la relación sea de sustitución (sustitución de humanos por robots) en tareas rutinarias, peligrosas, desagradables y en las que los robots demuestren mayor eficacia y eficiencia que los humanos.

Y será de colaboración en aquellas tareas que por la necesidad de intuición, de creatividad, de liderazgo u otros ámbitos los humanos tengan mejor desempeño, ya por carencias cognitivas o mecánicas de los robots o en tareas que, simplemente, los humanos prefiramos reservarnos para nosotros.

En cualquier caso, el peso de unas u otras será cambiante en el tiempo, dependiendo del desarrollo de las capacidades de los robots y algoritmos y también de nuestras propias preferencias.

Y no, los robots no se van a sentir amenazados porque, simplemente, no sienten y no parece que vayan a sentir en un futuro cercano, puede que nunca. Más que sentirse amenazados lo único que pudiera suceder sería que les impusiésemos unos objetivos que fuesen incompatibles con nuestros propios deseos, algo que no parece muy razonable, salvo error del que sí conviene tener salvaguarda (el famoso 'botón de stop').

TDS: Hablas en tu libro de tres cambios importantes que han ocurrido en los últimos años, desde el punto de vista tecnológico, el auge del cloud computing, el Big Data y las soluciones o formas de automatización cognitivas. ¿Crees que las empresas españolas son realmente conscientes del gran impacto que esta transformación tecnológica tendrá en sus negocios? ¿Están preparadas?

IG: Ahí tengo que reconocer que soy algo menos optimista. La verdad es que creo que en España estamos lejos de aprovechar las tecnologías en todo su potencial, que el nivel de conocimiento e implantación en el tejido empresarial está lejos de lo deseable.

Es cierto que, como país, estamos bastante bien situados en temas concretos como las comunicaciones de banda ancha o incluso la robótica industrial. Pero son situaciones no generales. Creo que nuestro desarrollo tecnológico es demasiado corto, que las empresas no conocen suficientemente las opciones que les da la tecnología y que nos las utilizan lo deseable. Y no hablo ni siquiera de las más avanzadas y brillantes, hablo de, por ejemplo, soluciones de Business Intelligence básico, de, por ejemplo, los visualizadores de datos o también, las inmensas posibilidades del cloud.

Y por si algo falta, tenemos una clara falta de talento en el campo tecnológico, de personas preparadas en este campo en el que hay una demanda no satisfecha.

Parece, eso sí, que la pandemia nos ha hecho despertar un poco a ese respecto. Confiemos en un mejor desarrollo futuro.

TDS: Hay un tema del que se oye hablar muy a menudo y que también abordas en tu libro, y es el de la necesidad de comprender por qué un algoritmo ha tomado una decisión en particular, es lo que se conoce como explicabilidad. Para algunos algoritmos esto es relativamente sencillo, para otros es casi imposible, hablamos de redes neuronales y Deep learning y los algoritmos black box. Explicabilidad o confianza ¿es ese el dilema?

IG: En cierto modo sí. Aparte de lo que cuento en el libro "Robots en la sombra", también en mi blog Blue Chip he publicado algunos artículos al respecto.

Por intentar resumir mi visión: mi posición es, por un lado, que los algoritmos de inteligencia artificial son perfectamente explicables, en el sentido de que, en un momento dado, son plenamente deterministas y se puede saber, ante una entrada concreta cuál va a ser su salida. El problema es que esa relación entrada-salida puede cambiar según aprenden y, sobre todo, que se explican en unos términos matemáticos que no son comprensibles por los humanos o, dicho de otra forma, que no es el tipo de explicación que estamos esperando.

Por otro lado que, los algoritmos no explicables (no explicables en el sentido humano), no son de 'caja negra' por malicia de los desarrolladores o de los propios algoritmos. Son de caja negra porque los humanos no hemos encontrado forma de aportar reglas concretas sobre cómo actuar y lo que le pedimos es que sean los algoritmos los que reconozcan los patrones de decisión adecuados. Y los algoritmos que mejor lo hacen, típicamente del Deep learning, son de caja negra. Y esos algoritmos hacen su labor y la hacen muy bien. Si ahora los sustituimos por otros algoritmos explicables o si los deformamos para hacerlos explicables, probablemente tengamos peores respuestas. Dicho de otra forma, con frecuencia la explicabilidad implica, al menos hoy en día, peor desempeño de los algoritmos, menos eficacia, menos eficiencia.

Por tanto, creo que no deberíamos exigir la explicabilidad de manera generalizada e indiscriminada sino cuando realmente tenga sentido.

¿Y cuándo tiene sentido exigir la explicabilidad? En el libro lo resumo en dos cosas: decisiones sin criterios claros, sin una lógica de decisión evidente (y me refiero con ello a que tampoco los humanos tenemos criterios claros de decisión, sino que también hay debate y opiniones) y decisiones que nos importan mucho.

Me refiero a una decisión de contratación o promoción en una empresa, a una decisión de concesión de préstamos o subvenciones, a una decisión de condena o exoneración de un acusado, etc En estos casos, deseamos, y es comprensible que así sea, saber por qué no nos han contratado, por qué no nos han concedido un préstamo o por qué nos han condenado. Deseamos entenderlo y deseamos poder protestar, argumentar en contra o presentar un recurso. Y para ello, necesitamos una decisión explicable.

Pero en la mayoría de los casos, donde prime mucho más la realización efectiva y eficiente de una tarea y no tenga las implicaciones anteriores, mejor no exigir la explicabilidad y contentarnos con que los algoritmos funcionen correctamente y de la forma más eficaz y eficiente. Sin más explicaciones.

TDS: ¿Cuál es el siguiente nivel de evolución de los chatbots y robots RPA en el futuro más inmediato?

IG: Hay una línea clara: más y más inteligencia. Los robots software hoy en día utilizan la inteligencia artificial sobre todo en la relación con el contexto, aplicaciones y documentos en el caso de RPA y los propios humanos en el caso de los chatbots. Así utilizan el lenguaje natural para comunicarse con personas de manera natural de forma escrita, el reconocimiento de voz para hacerlo de manera hablada, la visión artificial para obtener datos de imágenes en pantallas y documentos, el reconocimiento óptico de caracteres para obtener textos incrustados en imágenes o reconocimiento de texto en manuscritos.

Pero la lógica de la conversación en los chatbots y la lógica de proceso en RPA, normalmente está basada en reglas de negocio, especialmente en el caso de los segundos. Se trabaja en incorporación de más inteligencia no sólo en la relación con el exterior sino en su propia decisión, muy especialmente en el caso de los robots conversacionales.

También parece que en los próximos meses y años vamos a seguir observando avances en todo lo que

tiene que ver con el procesamiento del lenguaje natural y la voz, perfeccionando, probablemente mucho, esas capacidades que ya hoy día son buenas.

Creo que también es previsible la convergencia entre RPA y chatbots aunque aquí hay barreras de naturaleza no tanto técnicas, como históricas y de lógica de mercado.

Es muy interesante también, aunque no observo que esté alcanzando un gran éxito comercial, el concepto de trabajador digital, un robot que aúna características conversacionales con capacidades de interacción con back office y con unos conocimientos e inteligencia algo más avanzados, que le permiten cubrir la totalidad de tareas de un perfil laboral, de un puesto de trabajo.

Y, finalmente, y aunque no estoy convencido de su éxito en el corto plazo, tengo mucha curiosidad por ver cómo evoluciona el uso de avatares, con caras y aspecto humanos realista y con muy buenas capacidades gestuales y de expresión.

TDS: Hablando de inteligencia artificial y los principios éticos sobre los que debe construirse cualquier sistema inteligente, confianza, equidad, justicia, explicabilidad, etc. ¿Crees que estamos muy lejos de conseguir un consenso sobre cómo debemos desarrollar estos sistemas inteligentes con los que podamos establecer una relación de confianza?

IG: Creo que en algunos temas hay amplio consenso, al menos en Europa. Me refiero a temas como, por ejemplo, la privacidad o la equidad. Falta, eso sí, desarrollo normativo, pero, de nuevo en Europa el impulso es fuerte y vamos a ir presenciando acciones, normas y regulaciones muy concretas y creo que no en mucho tiempo.

En el mundo de los robots, sin embargo, creo que hay retos en que no existe un consenso porque está pendiente un paso previo, un debate de profundas raíces filosóficas y de concepción de la persona y su dignidad. ¿Tiene, por ejemplo, sentido la amistad entre robots y personas y, si pensamos que no, deberíamos tomar alguna precaución al respecto? ¿Y qué decir de los robots sexuales? ¿O qué precauciones hay que tomar, si es que hay que hacerlo, y me refiero a precauciones psicológicas y emocionales, en la relación de robots con colectivos vulnerables, ancianos, niños o personas con capacidades diferentes?

Ahí nos falta mucho camino. En cierto sentido, nos cogen por sorpresa esos robots tan avanzados y con capacidades relacionales y sociales convincentes. Creo que no nos ha dado tiempo a asimilarlo, a reflexionarlo, a debatirlo y mucho menos a consensuarlo.

TDS: En un post publicado recientemente en tu blog, hablas sobre transformación digital y la pereza de los directivos y Thomas M. Siebel en su libro “Digital Transformation: Survive and Thrive in a Era of Mass Extinction”, habla del papel del CEO transformador. ¿Entiendes que hay una necesidad real de directivos

LOS ROBOTS NO SE VAN A SENTIR AMENAZADOS PORQUE, SIMPLEMENTE, NO SIENTEN Y NO PARECE QUE VAYAN A SENTIR EN UN FUTURO CERCANO, PUEDE QUE NUNCA

LOS ALGORITMOS DE IA SON PERFECTAMENTE EXPLICABLES EN EL SENTIDO DE QUE, EN UN MOMENTO DADO, SON PLENAMENTE DETERMINISTAS Y SE PUEDE SABER, ANTE UNA ENTRADA CONCRETA CUÁL VA A SER SU SALIDA

comprometidos e implicados en dicha transformación, que en muchos casos puede resultar traumática para las compañías? ¿Cómo crees que puede combatirse esa "pereza"?

IG: Diría que los directivos, con independencia de que hayan tomado acciones o no, están inquietos y son conscientes, de una forma a veces abstracta, de que tienen que hacer algo en materia digital.

Y diría que, por supuesto hablando de manera general, no falta compromiso, en el sentido de compromiso con su empresa y compromiso con el hecho de tener que cambiar, de tener que evolucionar.

No, no es falta de compromiso sino pereza.

Y la pereza a que me refería es una pereza si se quiere, de naturaleza más intelectual, pero que es peligrosa a efectos estratégicos y operativos. Me refería a la pereza que les da el conocer y entender la tecnología, a la renuncia a hacerlo y preferir delegar en otros esa labor. Y a la comodidad de mantenerse lejos de ella y pensar que sólo con aportar una visión estratégica de alto nivel y alguna acción de comunicación también de alto nivel dentro de un programa de gestión del cambio o transformación cultural ya han hecho su labor.

Creo que nos es así. Creo que los directivos tienen que conocer el panorama tecnológico digital, entender los fundamentos de las tecnologías y las oportunidades y amenazas que les suponen. Si no, no van a saber qué elegir ni por qué y van a 'tocar de oídas' y en un terreno, el digital, en que, reconozcámoslo, hay mucho 'humo', muchos mensajes grandilocuentes, muchas expectativas infladas como dirían en Gartner, y muchos actores interesados en mantener esa confusión. ¿Cómo entonces va el directivo a tomar una decisión fundada, estratégica y responsable?

Querer acometer una transformación digital sin saber nada de tecnología digital es como querer entrar en un mercado nuevo sin saber nada de ese mercado, ni de marketing, ni de ventas.

Necesitamos una formación y divulgación serias y necesitamos un cambio de actitud por parte de muchos directivos. Pero no es fácil, nada fácil. Los directivos escuchan demasiadas voces, aparte de su propia voz interior, diciéndoles que no tienen por qué ocuparse de la tecnología, que eso no es labor suya y que la transformación digital va de cualquier cosa menos de tecnología digital.

TDS: ¿Podrías hablarnos de alguna tecnología o proyecto disruptivo que en tu opinión pueda resultar una revolución a corto o medio plazo?

IG: Bueno. Esa es la pregunta más difícil de todas, con diferencia, no sólo por aquello de que predecir es muy difícil, especialmente a futuro, sino porque, como se enseña cuando se habla de innovación, una de las características propias de las tecnologías disruptivas es, precisamente, que 'no se las ve venir', que al principio no llaman la atención.

CREO QUE NO DEBERÍAMOS EXIGIR LA EXPLICABILIDAD DE MANERA GENERALIZADA E INDISCRIMINADA, SINO CUANDO REALMENTE TENGA SENTIDO

QUERER ACOMETER UNA TRANSFORMACIÓN DIGITAL SIN SABER NADA DE TECNOLOGÍA DIGITAL ES COMO QUERER ENTRAR EN UN MERCADO NUEVO SIN SABER NADA DE ESE MERCADO, NI DE MARKETING NI DE VENTAS

Hay algunos campos, no digitales, en los que no me voy a arriesgar a apostar, porque los conozco menos, pero parece que pueden ser origen de grandes disrupciones: hablo de nanotecnología, de edición genética, bioingeniería o de nuevos materiales, por ejemplo.

En el campo digital, o casi mejor voy a decir de tecnologías de la información en este caso, la gran 'esperanza blanca', lo que puede ser un cambio de la noche al día, es la computación cuántica y no sólo por sí misma sino por lo que puede suponer como apalancamiento de otras tecnologías como la inteligencia artificial o el desafío para la ciberseguridad. Lo que no estoy seguro es de que lo veamos a corto o incluso medio plazo, aunque hay quien apuesto a que sí.

Pensando de una forma más cercana, y aunque no deje de tener su riesgo, señalaría dos explosiones pendientes y una muy posible. Las dos que creo pendientes son la de Internet de las cosas y la de la realidad aumentada, dos tipos de soluciones con un enorme potencial y variedad de aplicaciones pero que parece que no acaban de cumplir las expectativas, no tanto tecnológicas como de mercado y adopción.

Quizá el despliegue del 5G, cuando se produzca de forma masiva, sea el catalizador definitivo de Internet de las Cosas para el gran público y el ámbito doméstico. En cuanto a la realidad aumentada creo que, más bien, lo que falta es encontrar la o las 'killer app' que la convierta en masiva. ¿Será el metaverso?

Por la que yo apuesto como muy posible, pero casi de forma intuitiva, es la progresiva robotización de entornos de oficina y sobre todo del hogar. Creo que los diferentes dispositivos, aspiradoras, electrodomésticos, altavoces inteligentes, equipos domóticos, etc irán adquiriendo más y más caracteres robóticos, más uso de lenguaje natural y voz, más capacidades relacionales, más inteligencia, más interconexión y que poco a poco van a ir robotizando nuestros entornos de manera casi inadvertida pero firme.

Y, por supuesto, cabe seguir esperando mejoras, algunas seguramente impresionantes en el campo de la inteligencia artificial, quizá sobre todo en procesamiento de lenguaje natural y estaría muy atento a los desarrollos en el campo del BCI, y no me refiero sólo a Neuralink. A lo mejor hay sorpresas.

PUEDES ENCONTRAR A IGNACIO G.R.GAVILÁN EN:

PÁGINA OFICIAL: IGNACIOGAVILAN.COM

BLOG BLUE CHIP: IGNACIOGAVILAN.COM/BLUE-CHIP/

PERFIL EN LINKEDIN: LINKEDIN.COM/IN/IGRGAVILAN/

TWITTER: @IGRGAVILAN

CANAL YOUTUBE: YOUTUBE.COM/IGNACIOGRGAVILAN

COMPUTACIÓN CUÁNTICA

Autor: Pedro Albarracín García

La computación clásica, la que utilizamos a diario en nuestros ordenadores, funciona mediante el uso de circuitos electrónicos cuya base es lo que se conoce como lógica booleana. Los ordenadores están compuestos de múltiples circuitos electrónicos que contienen a su vez millones de transistores que funcionan a modo de interruptores (encendido o apagado). Estos transistores se utilizan para tratar con información o valores digitales, los cuales se encuentran en formato binario, es decir, información que puede representarse mediante los valores 0 y 1. La lógica o álgebra booleana utiliza variables y operaciones lógicas sobre dichas variables.

Nuestros ordenadores son capaces de procesar millones de instrucciones por segundo. Por ejemplo, un procesador Intel Core i9 10980XE en tareas de compresión de archivos con una aplicación como 7-Zip puede llegar a realizar 137.310 millones de operaciones por segundo (MIPS).

La unidad de información más pequeña que maneja un ordenador es el bit, que puede representar uno de los 2 estados posibles, 0 y 1, es decir, encendido o apagado o, dicho de otro modo, llega corriente o no llega corriente a la base. Como adelantamos al principio, el elemento que funciona a modo de interruptor y que verifica el paso de corriente es el transistor. El número de transistores aumenta con cada nueva generación de procesadores, los más modernos alcanzan más de 4.000 millones de esos transistores.

Mediante el uso combinado de transistores conseguimos realizar determinadas operaciones booleanas que se conocen como puertas lógicas, del tipo AND, OR, NOR, etc., y que conforman la base de los circuitos integrados y de la electrónica digital.

La capacidad de procesamiento en los ordenadores actuales no es infinita y ya se están alcanzado los límites físicos para la miniaturización de estos transistores; hoy nos movemos entre los 5 y 3 nanómetros de tamaño del transistor en los procesadores más modernos. Sin embargo, el problema que se presenta a determinados tamaños por debajo de los 3 nanómetros es que empiezan a producirse determinados comportamientos cuánticos.

ORDENADOR CUÁNTICO

UN ORDENADOR CUÁNTICO ES CAPAZ DE APROVECHAR LAS PROPIEDADES QUE SE DESCRIBEN EN LA MECÁNICA CUÁNTICA PARA REALIZAR TAREAS DE COMPUTACIÓN

Computación Cuántica

Por lo general somos conscientes de los fenómenos físicos que ocurren a nivel macroscópico, es decir, a simple vista, sin ayuda del microscopio. Otros fenómenos ocurren bajo la influencia de partículas atómicas y subatómicas en lo que conocemos como principios de la mecánica cuántica.

Los principios de la computación cuántica se basan, por lo tanto, en las leyes que rigen la física cuántica, es decir, en las propiedades de la materia a escalas muy, muy pequeñas. Otra característica fundamental de la física cuántica y cuya importancia veremos más adelante es que es un sistema probabilista.

Utilizamos nuestros ordenadores para almacenar, procesar y presentar información y ya vimos al principio de nuestro artículo que esto es posible gracias al uso de transistores y el manejo de estados, 0 y 1, mediante bits. La física cuántica nos ha permitido descubrir que es posible utilizar nuevos mecanismos para procesar la información, diseñar ordenadores basados en una arquitectura diferente y sobre todo trabajar de una forma exponencialmente más rápida.

La computación cuántica se basa en el uso de QUBITS (quantum bits) como la unidad mínima de información, que a diferencia de los clásicos bits permiten representar más estados debido a una característica de la física cuántica denominada superposición, es decir, un qubit no sólo puede representar un estado 0 o 1, sino ambos a la vez, dicho de otro modo, un qubit puede tomar un rango continuo de valores que representen estados de superposición. Por ejemplo, un qubit podría estar con un 50% de probabilidades en estado 1 y un 50% de probabilidades en estado 0 al mismo tiempo, lo que se denomina "**superposición equivalente**".

Pero, ¿podemos utilizar transistores como qubits? No. Sólo algunos "objetos" del mundo micro, como fotones o electrones, entre otros, son capaces de manifestar determinadas propiedades, algunos efectos como el de la superposición cuántica únicamente existen a nivel micro y bajo determinadas circunstancias, como veremos más adelante cuando hablemos de la arquitectura de un ordenador cuántico.

Otra característica de los sistemas de computación

QUBIT (BIT CUÁNTICO)

UNIDAD MÍNIMA DE INFORMACIÓN EN UN SISTEMA CUÁNTICO. UN QUBIT PUEDE REPRESENTAR UN ESTADO 0 -1 Y AMBOS A LA VEZ EN UN RANGO CONTINUO DE VALORES QUE REPRESENTAN LA PROBABILIDAD DE QUE ESTÉ EN CUALQUIERA DE ESOS ESTADOS 0-1

QUTRIT (TRIT CUÁNTICO)

SISTEMAS QUE PUEDEN REPRESENTAR TRES ESTADOS, 0-1-2 O UNA SUPERSPOSICIÓN DE ESOS ESTADOS.

SUPERPOSICIÓN

SEGÚN LOS PRINCIPIOS DE LA MECÁNICA CUÁNTICA UN SISTEMA TIENDE A COLAPSAR O DEFINIR SU ESTADO SÓLO CUANDO ES OBSERVADO O MEDIDO. ANTES DE SU MEDICIÓN SE ENCUENTRA EN UN ESTADO INDETERMINADO, UNA VEZ ES OBSERVADO O MEDIDO PASA A UN ESTADO DE DEFINICIÓN.

ENTRELAZAMIENTO

EL ENTRELAZAMIENTO ES UN CASO ESPECIAL DE SUPERPOSICIÓN Y HACE REFERENCIA A LA FUERTE CORRELACIÓN QUE EXISTE ENTRE LA MEDICIÓN DE UN SISTEMA Y EL ESTADO DE OTRO SISTEMA SIN IMPORTAR SU DISTANCIA

cuántica son los niveles o diferentes estados que pueden representar los qubits. Un qubit es un sistema de dos niveles ya que puede representar los estados 0 y 1 o una superposición de ambos, otros sistemas más complejos están formados por QUTRITS, es decir, "trits cuánticos", o lo que es lo mismo, sistemas que pueden representar tres estados , 0, 1 y 2 o una superposición de esos estados.

No es fácil aumentar el número de qubits que forman parte de un procesador cuántico. En 2019 Google presentó el procesador cuántico Sycamore formado por 54 qubits, esto quiere decir que puede representar la cantidad de 1801439 8509481984 estados cuánticos, o lo que es lo mismo 2 elevado a 54. En 2021 IBM presentó el procesador cuántico Eagle formado por 127 qubits, y se prevé que durante el año 2022 IBM presente un ordenador cuántico de mas de 400 qubits.

El Ordenador Cuántico

Un ordenador cuántico es una máquina capaz de aprovechar las propiedades que se describen en la mecánica cuántica para realizar tareas de computación. Para describir un ordenador cuántico no debemos pensar en el paradigma clásico en términos de CPU, memoria RAM, disco duro o tarjeta gráfica.

Existen diferentes arquitecturas y diseños para un ordenador cuántico, aunque sí es cierto que estas arquitecturas requieren de un ordenador tradicional que hace las veces de sistema de control.

En cualquier proceso de computación cuántica, tras la medición, la salida resultante es información que debe ser procesada posteriormente por un ordenador clásico.

La complejidad de construir un ordenador cuántico se deriva de las extremas condiciones bajo las que deben trabajar sus componentes de forma que puedan mantener o preservar lo que se conoce como estado cuántico.

LA SUPERPOSICIÓN CUÁNTICA PRESENTA LA POSIBILIDAD DE QUE TODAS LAS ALTERNATIVAS POSIBLES ESTÁN OCURRIENDO AL MISMO TIEMPO CON UN GRADO DE PROBABILIDAD

ALGUNAS PREVISIONES

LA CONSULTORA DELOITTE FORMULÓ LAS SIGUIENTES PREVISIONES EN EL ÁMBITO DE LA COMPUTACIÓN CUÁNTICA:

- LOS ORDENADORES CUÁNTICOS NO REEMPLAZARÁN A LOS ORDENADORES TRADICIONALES, AL MENOS EN LAS PRÓXIMAS DÉCADAS
- EL MERCADO DE LOS ORDENADORES CUÁNTICOS DEL FUTURO TENDRÁN APROXIMADAMENTE LA MISMA ENVERGADURA QUE EL MERCADO DE LOS SUPERORDENADORES: UNOS 50.000 MILLONES DE DÓLARES AL AÑO
- ES POSIBLE QUE LOS PRIMEROS ORDENADORES CUÁNTICOS COMERCIALES DE USO GENERAL APAREZCAN A FINALES DE LA DÉCADA DE 2030

La construcción de un ordenador cuántico requiere [1]:

- Un qubit físico bien aislado del entorno y capaz de dirigirse y acoplarse a más de un qubit adicional de forma controlable
- Una arquitectura tolerante a fallos que soporte qubits lógicos fiables
- Puertas universales, inicialización y medición de qubits lógicos

Lo realmente complicado es conseguir que las partículas, es decir, eso que real y físicamente representan un qubit, lleguen a un estado cuántico y por lo tanto puedan expresar todas sus propiedades. Mantener un estado cuántico es difícil por su fragilidad implícita, por lo que se requiere un entorno totalmente aislado al vacío (incluso aislado del campo magnético terrestre) y a muy baja temperatura, lo más próximo posible al 0 absoluto (-273º C). Cualquier interferencia que se produzca, movimiento o colisión entre átomos, interacción con el entorno, romperá el estado cuántico y por lo tanto se producirán errores.

Usos de un ordenador cuántico

Puesto que la capacidad de cálculo de los ordenadores cuánticos para determinadas tareas es muy superior a un ordenador tradicional puede que te preguntes ¿puedo jugar al Fortnite con un ordenador cuántico o diseñar en 3D, navegar por Internet a gran velocidad? NO. Los usos actuales de la computación cuántica se centran en sectores donde existen problemas de cálculo intensivo, tales como: servicios financieros, industrias químicas, diseño de nuevos materiales, industria farmacéutica, criptografía, optimización de

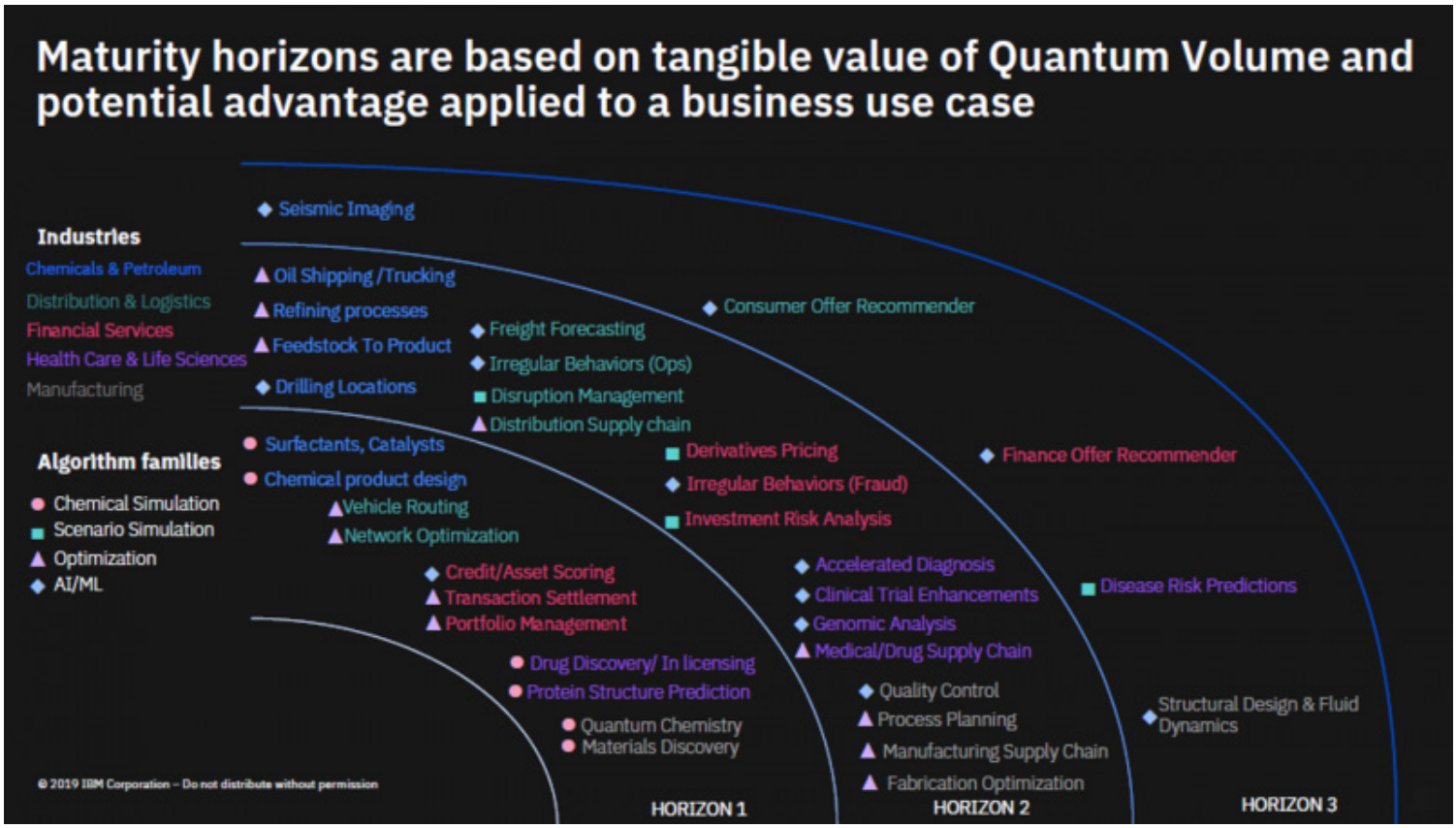


FIGURA 1: FUENTE IBM

sistemas logísticos o la inteligencia artificial.
IBM está probando sistemas cuánticos para entrenar y ejecutar algoritmos de aprendizaje automático con el fin de mejorar drásticamente tareas como la clasificación de datos. Esto podría permitirnos resolver problemas complejos más rápidamente, mejorando potencialmente aplicaciones como el diagnóstico de enfermedades, la detección de fraudes y la gestión eficiente de la energía. En la figura 1 se muestran las diferentes aplicaciones de la computación cuántica previstas por IBM para los próximos años según diversos horizontes de madurez. En el horizonte 1 se muestran las aplicaciones potenciales a corto plazo, en los próximos años, mientras que el horizonte 3 muestra las aplicaciones previstas para más allá de 15 años.

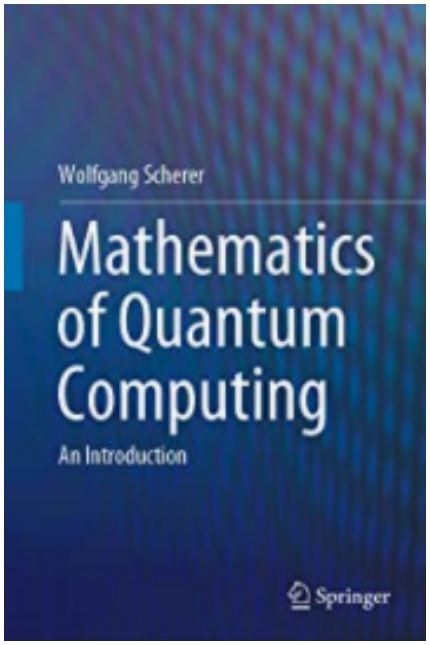
Referencias:

[1] Gambetta, J.M., Chow, J.M. & Steffen, M. Building logical qubits in a superconducting quantum computing system. npj Quantum Inf 3, 2 (2017). https://doi.org/10.1038/s41534-

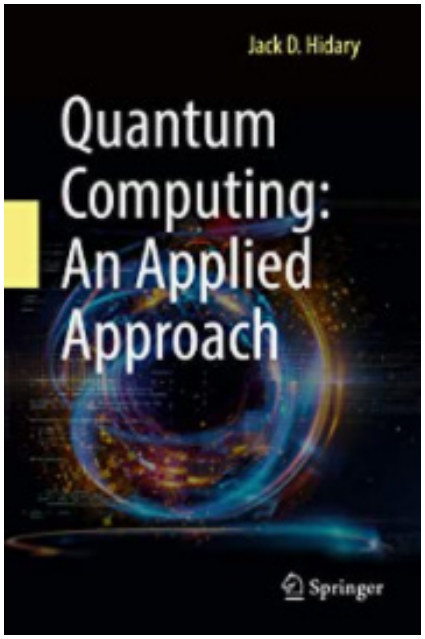
EN EL SIGUIENTE ARTÍCULO:

- LAS MATEMÁTICAS DE LA COMPUTACIÓN CUÁNTICA
- ¿QUÉ ES LA SUPREMACÍA CUÁNTICA?
- ¿QUÉ EMPRESAS LIDERAN EL DESARROLLO DE ORDENADORES CUÁNTICOS?
- PROGRAMACIÓN PARA COMPUTACIÓN CUÁNTICA

Lecturas Recomendadas



Autor: Wolfgang Scherer
Comprar en Amazon



Autor: Jack D. Hidary
Comprar en Amazon

¿QUIERES SABER QUÉ ES...

MLOPS (PARTE 2)

Autor: Pedro Albarracín García

Como analizamos en el artículo anterior hablamos de MLOps para referirnos a la metodología para la entrega y puesta en producción de modelos ML mediante flujos de trabajo sistematizados y eficientes que requieren la colaboración de diferentes perfiles y habilidades. Es precisamente este conjunto de perfiles y habilidades el que proporciona grandes beneficios en todo el ciclo de vida de desarrollo y despliegue de modelos ML. Analizamos a continuación algunos de esos beneficios.

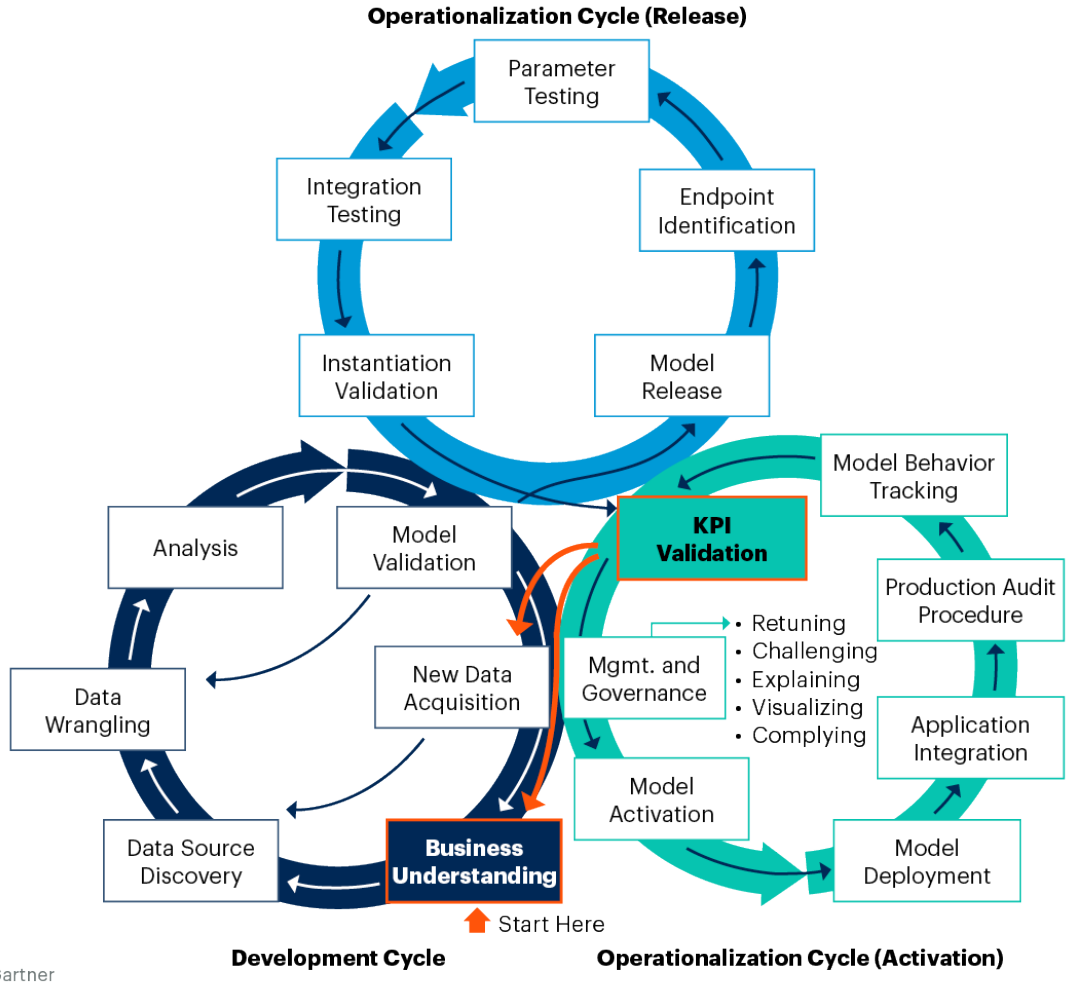
- Se acortan los tiempos de entrada en producción (time to market) de los modelos
- Más tiempo para desarrollo para los científicos de datos
- Aumenta la satisfacción de cliente
- Mejora la calidad de las predicciones

Para obtener esos beneficios es necesario implementar y automatizar:

- Integración continua
- Entrega continua
- Entrenamiento continuo

EL OBJETIVO DE MLOPS ES UNIFICAR EL DESARROLLO Y LAS OPERACIONES DEL SISTEMA DE ML

Gartner's MLOps Framework



Source: Gartner
725627_C

Figura 1: Framework MLOps de Gartner

Un sistema de desarrollo y producción de modelos ML no se diferencia en ese sentido de un sistema de software, por lo que en cierta medida es posible aplicar estrategias similares de desarrollo, pruebas o compilación.

Sin embargo la necesidad de probar y validar datos, la necesidad de entrenar y reentrenar los modelos, además de otras diferencias hace que MLOps represente un desafío para las organizaciones.

Uno de los principales desafíos precisamente es delimitar y coordinar los diferentes perfiles que intervienen en el desarrollo y puesta es producción del sistema ML y su correspondiente rol. En la figura 1 se muestra la propuesta de Gartner para un framework de MLOps, y es en ese framework donde deben encajar los siguientes perfiles:

- Expertos en el dominio
- Científicos de datos
- Ingenieros de datos
- Ingenieros de software
- DevOps
- Auditores
- Arquitectos ML

Expertos en el dominio

Con ellos comienza el ciclo del desarrollo. Expertos en la materia, es decir, conocedores del negocio, con los objetivos claramente definidos, KPI que quieren abordar así

como necesidades del negocio. Evalúan continuamente que el rendimiento de los modelos se alinean con las necesidades planteadas inicialmente mediante mecanismos de feedback.

Científicos de datos

Su función es la de desarrollar los modelos de acuerdo a las necesidades de negocio planteadas por los expertos. Entrega de modelos listos para despliegue en un entorno de producción y con datos también de producción. Evalúa la calidad de los modelos de acuerdo a los criterios establecidos por los expertos.

Ingenieros de datos

Optimiza la obtención y uso de los datos que se utilizarán en la fase de desarrollo. Trabajan en estrecha colaboración con los equipos de expertos para identificar los datos adecuados para el proyecto en cuestión y, posiblemente, también prepararlos para su uso. Trabajan también estrechamente con los científicos de datos para resolver cualquier problema de pipelines de datos que pueda hacer que un modelo no se comporte correctamente en producción.

Ingenieros de software

Se encargan de la integración de los modelos ML con otras aplicaciones de la compañía, aplicaciones web, apps para dispositivos móviles, etc. También se encargan de gestionar el control de versiones así como la integración con el circuito de soluciones CI/CD.

DevOps

Los equipos de DevOps se encargan de garantizar la seguridad, el rendimiento y la disponibilidad de los modelos de ML. En segundo lugar, son responsables de la gestión del pipeline de CI/CD.

El equipo de MLOps debe integrarse dentro de la estrategia global de DevOps de la compañía.

Auditores

Aseguran el cumplimiento de los requisitos internos y externos antes de la puesta en producción de los modelos ML.

Arquitectos ML

Aseguran entornos escalables para los modelos ML, desde el diseño, pasando por el desarrollo y la monitorización. Se encargan de la adopción de nuevas tecnologías que mejoren el rendimiento de los modelos en producción. Tienen un completo conocimiento de las necesidades de todos los actores que participan en el ciclo de vida del sistema ML.

Beneficios

Como mencionamos al principio del artículo la coordinación de perfiles MLOps proporcionan beneficios en todo el ciclo de vida de desarrollo y puesta en producción de modelos ML.

Time to Market

MLOps aporta automatización a los procesos de formación y reentrenamiento de modelos. También establece prácticas de integración continua y entrega continua (CI/CD) para el despliegue y la actualización de los pipelines de aprendizaje automático. Como resultado, las soluciones basadas en ML entran en producción más rápidamente.

Más tiempo para desarrollo

Con MLOps, un entorno de producción es el área de responsabilidad de los profesionales de operaciones, mientras que los científicos de datos pueden centrarse en sus principales tareas de desarrollo.

Mejora la experiencia de usuario

Gracias a las prácticas de MLOps, tales como la formación continua y la monitorización de los modelos

es posible actualizar y mantener el sistema cuando es necesario, lo que mejora la satisfacción del cliente.

Mejora la calidad de las predicciones

MLOps se encarga de la validación de los datos y del modelo, de la evaluación de su rendimiento en producción y del reentrenamiento con nuevos conjuntos de datos. Todo esto garantiza que pueda confiar en los resultados producidos por su algoritmo en la toma de decisiones importantes.

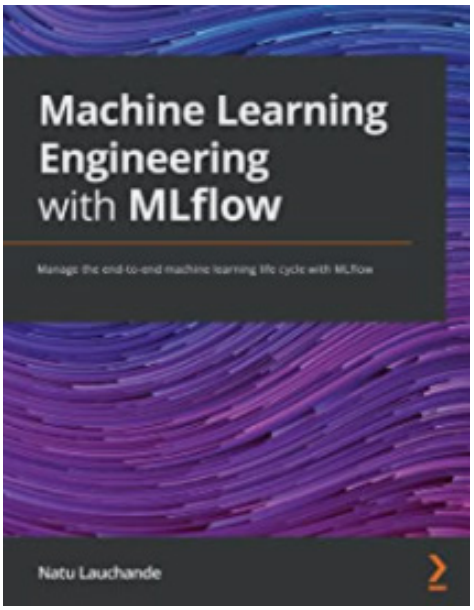
Conclusiones

- Son muchos los profesionales expertos que participan de la estrategia MLOps de la compañía y todos juegan un papel importante en la puesta en producción y mantenimiento de los modelos ML.
- Un correcta estrategia MLOps redunda en un mejor rendimiento de los modelos y esto favorece una mayor confianza de los sistemas ML.
- Las empresas deben hacer esfuerzos para superar los retos que derivan de la implantación de MLOps en el ciclo de vida de un sistema ML.

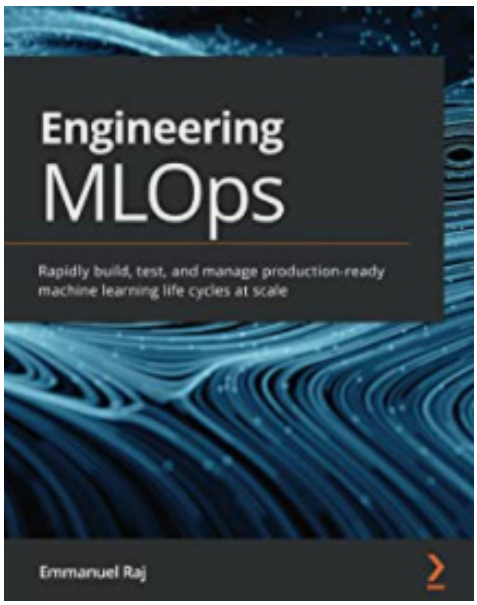
EN EL SIGUIENTE ARTÍCULO:

- ¿QUÉ ES MLFLOW?
- ¿CÓMO ENCAJA EN EL CICLO DE VIDA DE DESARROLLO Y DESPLIEGUE DE MODELOS ML?

Lecturas Recomendadas



Autor: Natu Lauchande
Comprar en Amazon



Autor: Emmanuel Raj
Comprar en Amazon

DEDOMENA

No hay inteligencia sin datos. No hay datos sin privacidad.

Reduce esfuerzos de desarrollo de IA hasta en un **80%**

Mejora la efectividad de tus modelos en un **30%**

Accede a un más de datos **90%**

Datos Sintéticos

Genera copias sintéticas estadísticamente idénticas a tus datos sin que contengan información identificable, garantizando la privacidad y el valor empresarial de los datos.

Herramientas de Anonimización

Ve más allá de las técnicas tradicionales de anonimización. Con Dedomena ya no tendrás que sacrificar valor de los datos por privacidad.

Soluciones IA de Negocio

Extrae el máximo valor de tus datos en cuestión de días, y no en meses o años. Ya tenemos los modelos que tu negocio necesita. Extraer valor de los datos nunca ha sido tan fácil.

Preciso. Fiable. Potente.

dedomena.ai

LIBRERÍAS

NIVEL: BÁSICO

ENTREGA III

PANDAS

UNA PODEROSA HERRAMIENTA PARA EL ANÁLISIS DE DATOS

Autor: Ambrosio Nguema

NUESTRO COMPAÑERO AMBROSIO NGUEMA, QUE DIRIGE ESTA SECCIÓN NOS OFRECE UN JUPYTER NOTEBOOK DONDE CONTINUA EXPLICANDO EL ANÁLISIS BÁSICO DE UN DATASET CON PANDAS

AL FINAL DEL ARTÍCULO TE PROPONEMOS ALGUNAS IDEAS POR SI TE ATREVES CON ELLAS

PUEDES OBTENER EL DATASET DEL ARTÍCULO EN:

HTTPS://WWW.KAGGLE.COM/KAGGLE/SF-SALARIES?SELECT=SALARIES.CSV

Salaries Employees

Como siempre importamos las librerías necesarias

y creamos un dataframe a partir del archivo .csv

Utilizamos la propiedad shape para conocer la forma del dataframe

y la función head() para ver las primeras filas

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
```

In [3]:
```
data = pd.read_csv("Salaries.csv")
```

In [4]:
```
print(data.shape)
data.head()
```

```
(148654, 13)
```

Out[4]:

	Id	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	1	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN
1	2	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN
2	3	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91	335279.91	2011	NaN	San Francisco	NaN
3	4	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61	332343.61	2011	NaN	San Francisco	NaN
4	5	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.19	326373.19	2011	NaN	San Francisco	NaN

In [5]:
```
data.drop('Id', axis=1, inplace=True)
```

In [6]:
```
data.dtypes
```

Out[6]:
```
EmployeeName        object
JobTitle            object
BasePay             object
OvertimePay         object
OtherPay            object
Benefits            object
TotalPay           float64
TotalPayBenefits   float64
Year                 int64
Notes              float64
Agency              object
Status              object
dtype: object
```

In [14]:
```
# Identfificamos de esta otra manera las columnas categoricas, lo hacemos así por si luego queremos trabajar solo con ellas

categoricas = [var for var in data.columns if data[var].dtype == 'O']
data[categoricas] = data[categoricas].astype('O')
data[categoricas].head()
```

Out[14]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	Agency	Status
0	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	San Francisco	NaN
1	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	San Francisco	NaN
2	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	San Francisco	NaN
3	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	San Francisco	NaN
4	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	San Francisco	NaN

In [13]:
```
# Identfificamos de esta otra manera las columnas numericas, lo hacemos así por si luego queremos trabajar solo con ellas

numericas = [var for var in data.columns if data[var].dtype != 'O']
data[numericas] = data[numericas].astype('float')
data[numericas].head()
```

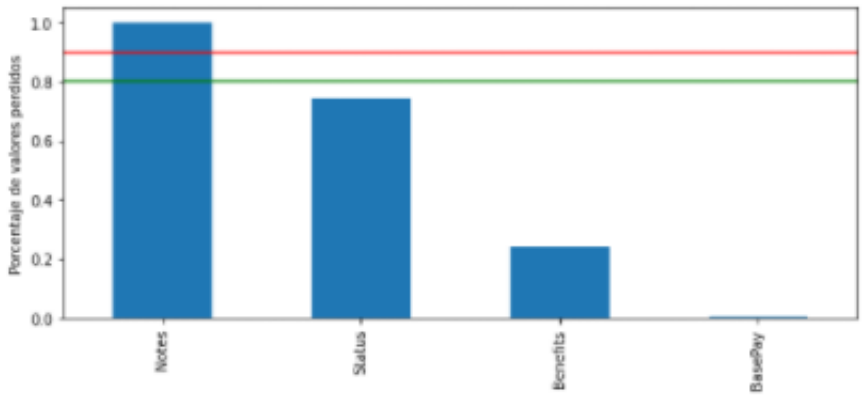
Out[13]:

	TotalPay	TotalPayBenefits	Year	Notes
0	567595.43	567595.43	2011.0	NaN
1	538909.28	538909.28	2011.0	NaN
2	335279.91	335279.91	2011.0	NaN
3	332343.61	332343.61	2011.0	NaN
4	326373.19	326373.19	2011.0	NaN

In [16]:
```
# Hacemos una lista de las variables que contienen valores perdidos.

# Determinar el porcentaje de valores perdidos (expresados como decimales)
# Y mostramos el resultado ordenado por% de datos perdidos

variables_nulas = [var for var in data.columns if data[var].isnull().sum() > 0]
data[variables_nulas].isnull().mean().sort_values(ascending=False)
```

Out[16]:
```
Notes       1.000000
Status      0.743572
Benefits    0.243243
BasePay     0.004070
dtype: float64
```

In [17]:
```
# Ploteamos
# Nuestro conjunto de datos contiene algunas variables con una gran proporción de valores perdidos
# Esto significa que para entrenar un modelo de aprendizaje automático con este conjunto de datos,
# necesitaríamos imputar los datos faltantes en estas variables. Hay varias técnicas que mostraré en otros capítulos


data[variables_nulas].isnull().mean().sort_values(
    ascending=False).plot.bar(figsize=(10, 4))
plt.ylabel('Porcentaje de valores perdidos')
plt.axhline(y=0.90, color='r', linestyle='-')
plt.axhline(y=0.80, color='g', linestyle='-')

plt.show()
```



In [10]:
```
# Ahora podemos determinar qué variables, de aquellas con datos faltantes,
# son numéricos y cuáles son categóricos

cat_na = [var for var in categoricas if var in variables_nulas]
num_na = [var for var in numericas if var in variables_nulas]

print('Number of categorical variables with na: ', len(cat_na))
print('Number of numerical variables with na: ', len(num_na))
```

```
Number of categorical variables with na:  3
Number of numerical variables with na:  1
```

In [11]:
```
# Ahora nos podemos centrar en la variable TotalPay encontrar las 5 observaciones con el valor más grande
# Encontramos las 5 observaciones con el valor más alto, la paga total

data.nlargest(5,"TotalPay")
```

Out[11]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43	567595.43	2011	NaN	San Francisco	NaN
1	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28	538909.28	2011	NaN	San Francisco	NaN
110531	David Shinn	Deputy Chief 3	129150.01	0.0	342802.63	38780.04	471952.64	510732.68	2014	NaN	San Francisco	PT
110532	Amy P Hart	Asst Med Examiner	318835.49	10712.95	60563.54	89540.23	390111.98	479652.21	2014	NaN	San Francisco	FT
36159	Gary Altenberg	Lieutenant, Fire Suppression	128808.87	220909.48	13126.31	44430.12	362844.66	407274.78	2012	NaN	San Francisco	NaN

In [12]:
```
# Veamos cómo haríamos para encontrar las 5 observaciones con el valor más pequeño
data.nsmallest(5,"TotalPay")
```

Out[12]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
148653	Joe Lopez	Counselor, Log Cabin Ranch	0.00	0.00	-618.13	0.00	-618.13	-618.13	2014	NaN	San Francisco	PT
36156	PAULETTE ADAMS	STATIONARY ENGINEER, WATER TREATMENT PLANT	0.0	0.0	0.0	NaN	0.00	0.00	2011	NaN	San Francisco	NaN
36157	KAUKAB MOHSIN	TRANSIT OPERATOR	0.0	0.0	0.0	NaN	0.00	0.00	2011	NaN	San Francisco	NaN
36158	JOSEPHINE MCCREARY	MANAGER IV	0.0	0.0	0.0	NaN	0.00	0.00	2011	NaN	San Francisco	NaN
70877	Roland Baylon	Deputy Court Clerk II	0.0	0.0	0.0	3728.05	0.00	3728.05	2012	NaN	San Francisco	NaN

In [13]:
```
# Decidimos por ejemplo establecer como thereshold el valor superior de la mediana de TotalPay

mediana = data["TotalPay"].median()
len(data[data['TotalPay']>mediana])
```

Out[13]: 74327

In [14]:
```
median_employees = data[data['TotalPay']>mediana]
```

In [17]:
```
# Creamos un groupby y así podemos visualizar que puestos de trabajo son los que tienen mejores salarios. TOP 10

median_employees.groupby(['JobTitle','TotalPay']).count().reset_index().sort_values(by = 'TotalPay', ascending=False)[['JobTitle
```

Out[17]:

	JobTitle	TotalPay
22190	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	567595.43
5934	CAPTAIN III (POLICE DEPARTMENT)	538909.28
11212	Deputy Chief 3	471952.64
3535	Asst Med Examiner	390111.98
25263	Lieutenant, Fire Suppression	362844.66
6866	Captain 3	350403.41
5512	Battalion Chief, Fire Suppress	347102.32
3534	Asst Med Examiner	344187.46
7775	Chief Investment Officer	339653.70
7849	Chief of Police	339282.07

In [28]:
```
# Y si agregamos el simbolo del dolar a la columna TotalPay?

median_employees['TotalPay'] = median_employees['TotalPay'].apply(lambda x: str(x) + '$')
```

La columna TotalPay seguira siendo float, o se ha convertido a otro type??

In [19]:
```
median_employees.head()
```

Out[19]:

	EmployeeName	JobTitle	BasePay	OvertimePay	OtherPay	Benefits	TotalPay	TotalPayBenefits	Year	Notes	Agency	Status
0	NATHANIEL FORD	GENERAL MANAGER-METROPOLITAN TRANSIT AUTHORITY	167411.18	0.0	400184.25	NaN	567595.43$	567595.43	2011	NaN	San Francisco	NaN
1	GARY JIMENEZ	CAPTAIN III (POLICE DEPARTMENT)	155966.02	245131.88	137811.38	NaN	538909.28$	538909.28	2011	NaN	San Francisco	NaN
2	ALBERT PARDINI	CAPTAIN III (POLICE DEPARTMENT)	212739.13	106088.18	16452.6	NaN	335279.91$	335279.91	2011	NaN	San Francisco	NaN
3	CHRISTOPHER CHONG	WIRE ROPE CABLE MAINTENANCE MECHANIC	77916.0	56120.71	198306.9	NaN	332343.61$	332343.61	2011	NaN	San Francisco	NaN
4	PATRICK GARDNER	DEPUTY CHIEF OF DEPARTMENT,(FIRE DEPARTMENT)	134401.6	9737.0	182234.59	NaN	326373.19$	326373.19	2011	NaN	San Francisco	NaN

AQUÍ TIENES ALGUNAS IDEAS, ¿TE ATREVES?

- **OBTENER TODOS LOS EMPLEADOS CUYO APELLIDO EMPIEZE POR LA LETRA "A"**
- **¿QUÉ PUESTOS TIENEN PEOR SALARIO Y EN BASE A QUE CONDICION?**
- **SACAR UNA LABEL, PERO ¿CON QUÉ COLUMNA?, ¿QUÉ METRICA ASIGNARIAS 1 Y 0 Y QUÉ CONCLUSIONES SACARÍAS AL RESPECTO?**
- **¿TE ATREVES CON UN BOXPLOT O UN HISTOGRAMA?, Y EN ELLOS ¿QUÉ QUIERES EXPLICAR SOBRE LA DATA?**

Enlaces de descarga:

Descargar dataset desde Kaggle

Jupyter Notebook del artículo

CURSO R

MANIPULACIÓN DE DATOS CON EL TIDYVERSE: DPLYR Y TIDYR

Autores: Aurora González Vidal y Antonio Maurandi López

Este artículo presenta los principios básicos del tidyverse de R, que nos permiten manipular los datos para dejarlos preparados para su análisis y visualización. El artículo muestra el uso de las librerías dplyr y tidyr para seleccionar, filtrar y realizar operaciones básicas por grupos en un dataframe, cambiar el formato de las tablas y las columnas del mismo entre otras, así como el uso de pipes para facilitar la lectura y reutilización del código.

Introducción al tidyverse

El universo tidy o tidyverse está compuesto por una serie de librerías de R diseñados para una manipulación eficiente de los datos de cara a su análisis. Todas estas librerías comparten una filosofía de diseño, gramática y estructura de datos. Todos ellos han sido creados por el popular programador de R Hadley Wickham en colaboración con otros desarrolladores.

Resulta interesante leer el manifiesto del tidyverse(1), donde se definen los principios para proveer una interfaz uniforme a través de la cual las librerías de R puedan combinarse de una forma natural y una vez se consige manejar uno con cierta destreza, el resto resulte más sencillo dada su cohesión. En resumen, las librerías del tidyverse cumplen las siguientes características básicas

- Reutilizan estructuras de datos existentes,
- Las funciones simples las componen usando pipes (tuberías),
- Respetan y acogen la programación funcional, ya que R lo es,
- Diseñan los pasos para humanos.

Las librerías del tidyverse que introduciremos brevemente son dplyr y tidyr, pero hay más:

Ggplot2: probablemente el más conocido de esta lista, es un sistema organizado de visualización de datos con elementos bien definidos y que consigue resultados profesionales.

Forcats: librería para trabajar con variables categóricas que recoge una serie de herramientas para resolver los prolemas relacionados con los factores.

Tibble: Las tibbles son dataframes con características ampliadas, modernizados y adaptados al universo tidy

TIDYVERSE ESTÁ COMPUESTO POR UNA SERIE DE LIBRERÍAS DE R DISEÑADAS PARA UNA MANIPULACIÓN EFICIENTE DE LOS DATOS DE CARA A SU ANÁLISIS LA MANIPULACIÓN DE DATOS ES EL PROCESO POR EL CUAL LOS DATOS CRUDOS O RAW DATA SE TRANSFORMAN EN DATOS LIMPIOS Y ORDENADOS PREPARADOS PARA SU ANÁLISIS UNA TUBERÍA O PIPE CONSISTE EN UNA CADENA DE PROCESOS CONECTADOS

Readr: para importar y exportar los conjuntos de datos

Stringr: reune una serie de funciones para que trabajar con cadenas de datos sea más intuitivo.

Purrr: para hacer bucles tratándolos como funciones naturales.

Todos estas librerías se pueden instalar con una sola expresión (ver figura 1):

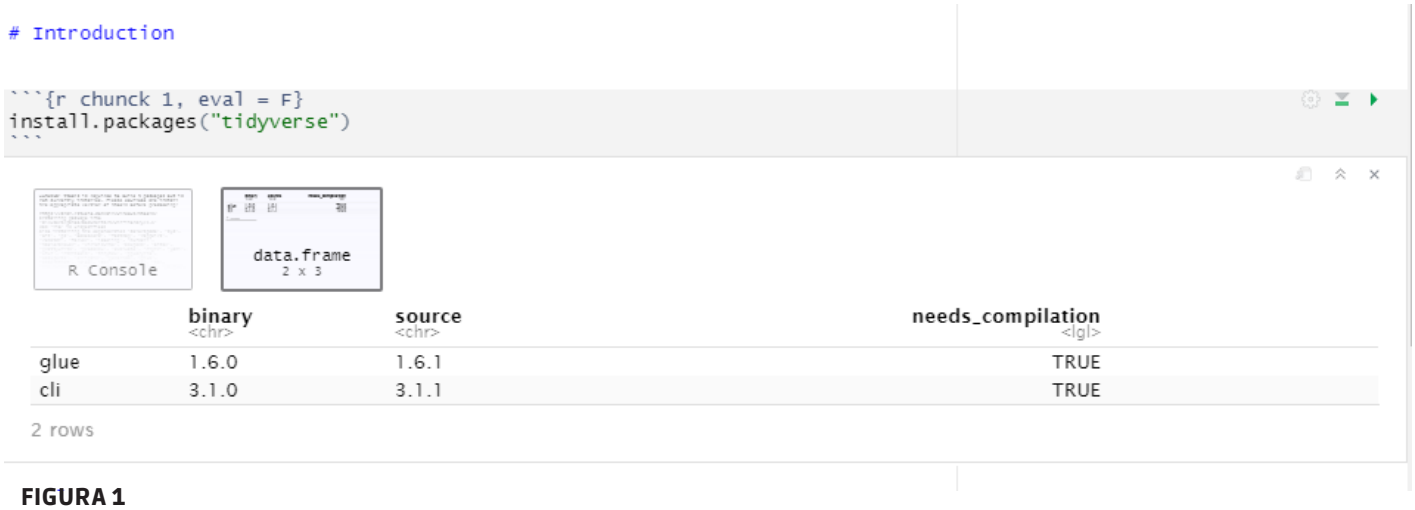


FIGURA 1

La manipulación de datos es el proceso a través del cual los datos crudos del mundo real, que presentan valores faltantes, formatos incorrectos, nombres de variables protegidos, etc, se transforman, de forma hábil, en datos limpios y ordenados preparados para su análisis. En lo que sigue veremos cómo utilizar los plaquetes Dplyr y Tidyr para este propósito y, además, cómo componer funciones simples mediante pipes para crear funciones más complejas.

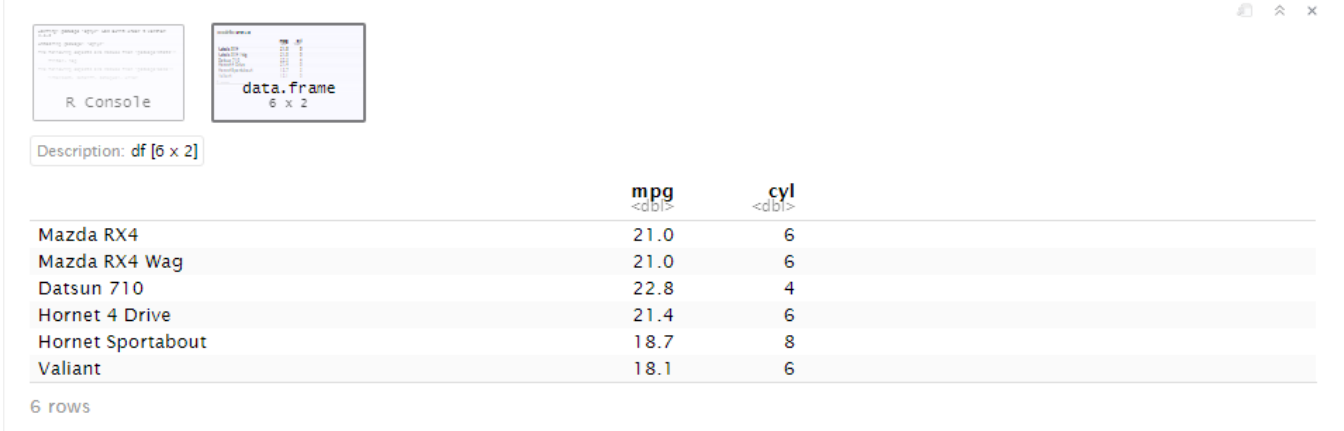
DPLYR

La librería dplyr contiene una colección de funciones para realizar operaciones de manipulación de datos comunes como: filtrar por fila, seleccionar columnas específicas, reordenar filas, añadir nuevas filas y agregar datos. Además, también contiene funciones que sirven para realizar una tarea que se denomina: split-apply-combine. Esto consiste en aplicar funciones por grupos como indica su denominación: split para cortar por grupos, apply para aplicar la función y combine para

combinar los grupos para dotarlos de una estructura.

Veamos con código, las funciones más sencillas aplicadas al conjunto de datos mtcars. Select sirve para filtrar los datos por columnas utilizando sus nombres y tipo. Por defecto pondremos el nombre del dataframe seguido de los nombres de las columnas que se quieren seleccionar (ver figura 2)

```
library("dplyr")
head( select( mtcars, mpg, cyl ) )
```



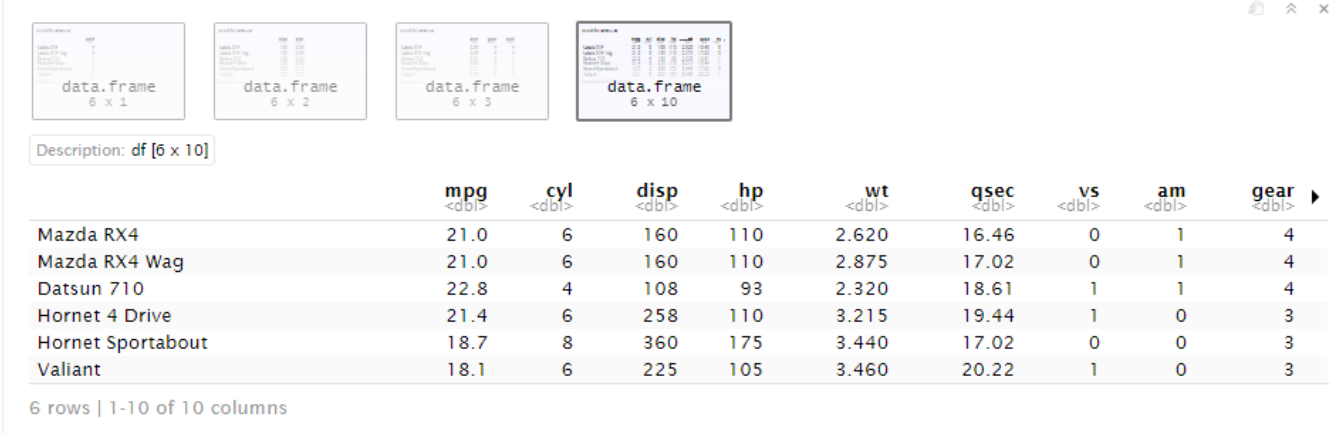
	mpg <dbl>	cyl <dbl>
Mazda RX4	21.0	6
Mazda RX4 Wag	21.0	6
Datsun 710	22.8	4
Hornet 4 Drive	21.4	6
Hornet Sportabout	18.7	8
Valiant	18.1	6

6 rows

FIGURA 2

Asimismo, hay muchas otras formas de seleccionar columnas, por ejemplo la última, las que empiezan por "d", las que contienen una "r", o eliminar columnas añadiendo un signo menos como se aprecia en el código. Para ver todas las posibilidades que hay con select, se puede utilizar la ayuda "select_helpers". (ver figura 3)

```
head(select(mtcars, last_col()))
head(select(mtcars, starts_with("d")))
head(select(mtcars, contains("r")))
head(select(mtcars, -drat))
?select_helpers
```



	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>	gear <dbl>
Mazda RX4	21.0	6	160	110	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160	110	2.875	17.02	0	1	4
Datsun 710	22.8	4	108	93	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258	110	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360	175	3.440	17.02	0	0	3
Valiant	18.1	6	225	105	3.460	20.22	1	0	3

6 rows | 1-10 of 10 columns

FIGURA 3

Filter sirve para seleccionar filas de acuerdo a condiciones que se cumplan por columnas. (ver figura 4)

Arrange sirve para ordenar las filas de acuerdo la condición indicada, por ejemplo por cilindrada (cyl) y por peso (wt). (ver figura 5)

```
filter(mtcars, cyl == 6)
```

Description: df [7 x 11]

	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1

7 rows | 1-10 of 11 columns

FIGURA 4

```
arrange(mtcars, cyl, wt)
```

Description: df [32 x 11]

	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0

1-10 of 32 rows | 1-10 of 11 columns Previous 1 2 3 4 Next

FIGURA 5

También podemos crear una nueva columna que sea el resultado de transformar otra, por ejemplo, podemos convertir el Horse Power (hp) a caballos de vapor como se muestra en la figura 6.

La función summarise agrupa los valores en una tabla de acuerdo a la función que le indiquemos. Para realizar el conteo de frecuencias se puede utilizar la función "n()" como se aprecia en el ejemplo. Habitualmente se utiliza junto a group_by(). (ver figura 7)

```
head( mutate(mtcars, cv = hp * 0.9863) )
```

Description: df [6 x 12]

disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>	gear <dbl>	carb <dbl>	cv <dbl>
160	110	3.90	2.620	16.46	0	1	4	4	108.4930
160	110	3.90	2.875	17.02	0	1	4	4	108.4930
108	93	3.85	2.320	18.61	1	1	4	1	91.7259
258	110	3.08	3.215	19.44	1	0	3	1	108.4930
360	175	3.15	3.440	17.02	0	0	3	2	172.6025
225	105	2.76	3.460	20.22	1	0	3	1	103.5615

6 rows | 4-13 of 12 columns

FIGURA 6

```
summarise(group_by(mtcars, gear), mean = mean(disp), freq = n())
```

A tibble: 3 x 3

gear <dbl>	mean <dbl>	freq <int>
3	326.3000	15
4	123.0167	12
5	202.4800	5

3 rows

FIGURA 7

PIPES

En informática, una tubería o pipe consiste en una cadena de procesos conectados de forma tal que la salida de cada elemento de la cadena es la entrada del próximo. Permiten la comunicación y sincronización entre procesos.

Para nosotros, el pipe será una función nueva que se define con el operador %>%. Esta función puede parecer extraña y farragosa, pero se necesitaba una función definida por >, que implica que es una cadena o sucesión de órdenes y en R, para que un símbolo defina una función debe estar escrito entre porcentajes.

El pipe viene de la librería magrittr de Stefan Milton Bache. Las librerías del tidyverse cargan %>% automáticamente, por lo que usualmente no habrá que cargar **magrittr** de forma explícita.

Vamos a encadenar sucesivamente funciones de la librería dplyr para combinarlas. Vemos que este tipo de programación es más gramatical, puede ser leída como sigue: del conjunto de datos mtcars filtramos las filas cuyo "cyl=6", de ese conjunto de datos resultante seleccionamos las columnas que no sean drat y am y, a continuación, las que contienen le letra "r". Del conjunto de datos resultante se muestra solo la cabecera. (ver figura 8)

```
mtcars %>%
  filter(cyl==6)%>%
  select(-drat,-am )%>%
  select(contains("r"))%>%
  head
```

Description: df [6 x 2]

	gear <dbl>	carb <dbl>
Mazda RX4	4	4
Mazda RX4 Wag	4	4
Hornet 4 Drive	3	1
Valiant	3	1
Merc 280	4	4
Merc 280C	4	4

6 rows

FIGURA 8

La principal diferencia entre usar pipes y no es que con las funciones tradicionales tenemos que leer de dentro a fuera, mientras que con pipes se van aplicando de una forma secuencial, que se asemeja al lenguaje humano.

Podemos también modificar lo que anteriormente habíamos hecho con la función summarise y group_by para realizarlo con pipes como sigue (ver figura 9)

```
mtcars %>%
  group_by(gear) %>%
  summarise(mean = mean(disp), n = n())
```

A tibble: 3 x 3

gear <dbl>	mean <dbl>	n <int>
3	326.3000	15
4	123.0167	12
5	202.4800	5

3 rows

FIGURA 9

TIDYR

El objetivo de la librería Tidyr es crear tablas de datos limpias, es decir, aquellas donde cada columna es una variable, cada fila una observación y cada celda contiene un único valor [1]. Las funciones de la librería que repasaremos son pivot_longer(), pivot_wider(), separate() y unite().

- **Formatos de tabla wide (ancha) y long (larga)**

Wide y Long son términos que se utilizan para describir dos formas de presentar los datos en una tabla o en un dataframe.

- **Wide**: Cada variable de los datos se presenta en una sola columna
- **Long**: Una columna contiene todos los valores y otra columna contiene los nombres de todas las variables.

La función pivot_longer() colapsa los valores de todas las columnas seleccionadas en una sola, es decir, nos transporta de una tabla ancha a una larga. La dimensión del conjunto de datos mtcars es de 32 filas y 11 columnas. (ver figura 10)

```
library("tidyr")

mtcars %>%
  dim

mtcars %>%
  head
```

Description: df [6 x 11]

	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0

6 rows | 1-10 of 11 columns

FIGURA 10

Sin embargo, al hacer que 3 columnas colapsen con la función pivot_longer() observamos que su dimensión ha cambiado, siendo ahora de 96 filas y 10 columnas, donde las 2 últimas contienen los nombres de las variables colapsadas y los valores de éstas respectivamente. (ver figura 11)

```
mtcars %>% pivot_longer(c(mpg, cyl, disp)) %>%
  dim

mtcars %>% pivot_longer(c(mpg, cyl, disp)) %>%
  head


```

A tibble: 6 x 10

hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>	gear <dbl>	carb <dbl>	name <chr>	value <dbl>
110	3.9	2.620	16.46	0	1	4	4	mpg	21
110	3.9	2.620	16.46	0	1	4	4	cyl	6
110	3.9	2.620	16.46	0	1	4	4	disp	160
110	3.9	2.875	17.02	0	1	4	4	mpg	21
110	3.9	2.875	17.02	0	1	4	4	cyl	6
110	3.9	2.875	17.02	0	1	4	4	disp	160

6 rows

FIGURA 11

De forma alternativa, con la función pivot_wider() se puede volver al estado anterior, es decir, pasar del formato long al formato wide. (ver figura 12)

Siguiendo con las transformaciones de las columnas, para pegar múltiples columnas en una sola utilizamos la función unite(). El separador por defecto es la barra baja "_". (ver figura 13)

```
df<- mtcars %>% pivot_longer(c(mpg, cyl, disp))
pivot_wider(df, names_from = name, values_from = value)
```

A tibble: 32 x 11

hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs <dbl>	am <dbl>	gear <dbl>	carb <dbl>	mpg <dbl>	cyl <dbl>
110	3.90	2.620	16.46	0	1	4	4	21.0	6
110	3.90	2.875	17.02	0	1	4	4	21.0	6
93	3.85	2.320	18.61	1	1	4	1	22.8	4
110	3.08	3.215	19.44	1	0	3	1	21.4	6
175	3.15	3.440	17.02	0	0	3	2	18.7	8
105	2.76	3.460	20.22	1	0	3	1	18.1	6
245	3.21	3.570	15.84	0	0	3	4	14.3	8
62	3.69	3.190	20.00	1	0	4	2	24.4	4
95	3.92	3.150	22.90	1	0	4	2	22.8	4
123	3.92	3.440	18.30	1	0	4	4	19.2	6

1-10 of 32 rows | 1-10 of 11 columns Previous 1 2 3 4 Next

FIGURA 12

```
mtcars %>%
  unite("vs:am", c("vs","am"),sep = ":")
```

Description: df [32 x 10]

	mpg <dbl>	cyl <dbl>	disp <dbl>	hp <dbl>	drat <dbl>	wt <dbl>	qsec <dbl>	vs:am <chr>	gear <dbl>
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0:1	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0:1	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1:1	4
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1:0	3
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0:0	3
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1:0	3
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0:0	3
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1:0	4
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1:0	4
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1:0	4

1-10 of 32 rows | 1-10 of 10 columns Previous 1 2 3 4 Next

FIGURA 13

Por último, para convertir una única columna en múltiples separando los caracteres que componen a la inicial se puede utilizar la función separate(). (ver figura 14)

```
df <- read.table("http://gauss.inf.um.es/datos/longlat.txt", sep=";", head=T)
df$Location <- gsub("[()]", "", df$Location)
separate(df, col = Location, into = c("lat","long"), sep = ",")
```

Description: df [6 x 5]

Date <chr>	Time <int>	Accident.Type <chr>	lat <chr>	long <chr>
07/01/2012	1630	PD	39.26699	-76.560642
07/02/2012	1229	PD	39.000549	-76.399312
07/02/2012	1229	PD	39.00058	-76.399267
07/02/2012	445	PI	39.26367	-76.56648
07/02/2012	802	PD	39.240862	-76.599017
07/02/2012	832	PD	39.27022	-76.63926

6 rows

FIGURA 14

CONCLUSIÓN

Hemos realizado una breve introducción al universo tidy de R, o tidyverse. Hemos visto cómo usar pipes para combinar funciones y la librería dplyr y tidyr para manipular datos. Como siguientes pasos dentro del tidyverse, los 2 paquetes básicos que se recomiendan son readr y ggplot2. Para profundizar más en temas de tabulación, visualización y reporte de datos, se recomienda el curso online autónomo tabularCola (2).

Enlace descarga script R

REFERENCIAS

[1] Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.".

[2] González-Vidal, Aurora, Maurandi-López, Antonio, & Del Río, Laura. (2019, April 4). tabularCola - Tabulación de datos con R: Curso on line autónomo. Zenodo. https://doi.org/10.5281/zenodo.2628892

ENLACES A LA UNIVERSIDAD DE MURCIA

WEB: HTTPS://GAUSS.INF.UM.ES/UMUR/

TWITTER: HTTPS://TWITTER.COM/UMUR_MURCIA

CANAL DE YOUTUBE: HTTPS://WWW.YOUTUBE.COM/CHANNEL/UC_L88_FH-EA5QZV-BVPCYKQ

Aurora González Vidal

Presidenta de la Asociación de Usuarios de R Murcia (UMUR) e investigadora postdoctoral del Dept. de Ingeniería de la Información y las Comunicaciones, Fac. de Informática, Universidad de Murcia.

E-mail: aurora.gonzalez2@um.es

Perfil RRSS: https://www.researchgate.net/profile/

Antonio Maurandi López

Vocal de la Asociación de Usuarios de R Murcia (UMUR) y profesor del Dept. Didáctica de las Ciencias Matemáticas y Sociales, Fac. de Educación, Universidad de Murcia.

E-mail: amaurandi@um.es

Perfil RRSS: https://amaurandi.github.io/


Aquí tu publicidad
Contacta en publicidad@thedatascientist.es
50% descuento 3 primeros meses

THE BLACK BOX ES UNA SECCIÓN QUE TIENE POR OBJETIVO EXPLICAR CONCEPTOS CIENTÍFICOS RELACIONADOS CON LA CIENCIA DE DATOS Y LA INTELIGENCIA ARTIFICIAL

ES UNA SECCIÓN EMINENTEMENTE DIVULGATIVA Y TRATA DE ILUSTRAR Y EXPLICAR DESDE CERO CONCEPTOS QUE SE MEZCLAN, SE RELACIONAN Y SE OCULTAN, LA MAYORÍA DE LAS VECES, BAJO EL TÉRMINO "ALGORITMO".

ABRIMOS THE BLACK BOX Y DEJAMOS TODOS SUS SECRETOS AL DESCUBIERTO.

PROBABILIDAD Y JUEGOS DE AZAR

Autora: Gema Fernández-Avilés Calderón

Como cada año, al llegar la Navidad, la ilusión de muchas personas brilla radiantemente, y no solo por lo entrañable y emotiva que pueden resultar estas fechas, sino porque cada fin de año, desde el 18 de diciembre de 1812, se celebra el sorteo de la Lotería de Navidad, con el que la esperanza de ser premiados con "El Gordo" hace que nos gastemos unos ahorrillos en los Décimos y Participaciones de la Lotería, pero… ¿sale rentable comprar Lotería de Navidad?

Pues como pasa la mayoría de las veces la respuesta a esta pregunta es **DEPENDE**, porque la lotería toca, lo vemos en la televisión cada 22 de diciembre, pero **¿qué probabilidad tengo de que me toque el gordo en la Lotería de Navidad?** Para responder a esta pregunta hagamos un simple cálculo probabilístico. En el sorteo de Lotería de Navidad se juegan 100.000 números de 5 cifras, que van desde el 00000 hasta el 99.999, por tanto, si sólo compro de un número (una participación, un décimo, una serie…) la probabilidad de que toque es de 1/100.000, lo que es igual a 0,01 por MIL. Es decir, **MUY, MUY, MUY BAJA. ¿Y cual sería la probabilidad de recibir un premio cualquiera de la Lotería de Navidad si sólo juego de un número?** Pues bien, dijimos antes que, en el sorteo de Lotería de Navidad hay 100.000 números que participan en el juego, de los cuales tan solo resultarán premiados 14.272. Por tanto, la probabilidad de ser ganador de cualquier premio (incluyendo la pedrea) es de 14.272/100.00, poco más del 14%.

TODOS LOS NÚMEROS QUE PARTICIPAN EN EL SORTEO TIENEN LA MISMA PROBABILIDAD DE SALIR PREMIADOS PROBABILIDAD ES SINÓNIMO DE ALEATORIEDAD, AZAR NO EXISTEN NÚMEROS MÁS AFORTUNADOS, NI NÚMEROS BONITOS O FEOS

Hasta ahora estoy suponiendo que sólo juego de un número, pero pare obvio pensar que **si juego más números tendré mas posibilidades de que me toque El Gordo**. Y efectivamente, así es. Sí en vez de comprar un único número compras 5, por ejemplo, la probabilidad de que te toque el gordo es de 5/100.000. Y si compras 50.000 números distintos, tendrías una probabilidad de que te el gordo del 50%. Pero ojo, para tener una probabilidad del 50% de ganar El Gordo "SOLO" tienes que gastarte como mínimo 1.000.000 de euros (suponiendo que compras 50.000 números distintos en forma de décimos x 20 euros). **¿Y qué pasaría si voy a comprar a la Administración X porque siempre toca? ¿Tendría más probabilidad de ser agraciado con El Gordo?** Pues no, no te pasaría nada en especial. Para empezar SIEMPRE y PROBABILIDAD no pueden ir en la misma frase, porque PROBABILIDAD es sinónimo de ALEATORIEDAD/ AZAR y estos términos son opuestos a SIEMPRE. Si la administración X, por ejemplo, la famosa Dª Manolita en Madrid, es la que más números vende, por supuesto es la que tiene la mayor probabilidad que te "venda" El Gordo, pero si yo solo compro de un número, mi probabilidad será 1/100.000 siguiendo el razonamiento anterior.

Una cuestión muy recurrente son los "**números bonitos**", los cuales no existen, lo que sí existen son los números preferidos, pero **NO MÁS AFORTUNADOS**. No existen números ni bonitos ni feos. El 00000 tiene la misma probabilidad de ser el premiado con el Gordo que el día que nacen tus hijos, el día que te casas, el día de la erupción del volcán de La Palma, el día de la declaración del Estado de Alarma, el día que tal...

Otro aspecto importante para tener en cuenta es la tributación a la Agencia Estatal de Administración Tributaria de un décimo premiado por El Gordo para no llevarse sorpresas desagradables. Una persona que lleve un décimo y le "toque" El Gordo, es premiada con 400.000 euros. Los primeros 40.000 euros están exentos de tributar. De los otros 360.000 euros, Hacienda grava el 20%, es decir, se queda con 72.000 euros. Luego, la persona premiada con un décimo de El Gordo de Navidad recibirá 328.000 euros, que ya no tendrán que tributar en el Impuesto sobre la Renta de las Personas Físicas, ya te lo dan descontado, pero sí en el Impuesto sobre Patrimonio.

Para finalizar estas reflexiones sobre el sorteo de Lotería de Navidad y probabilidad, la pregunta clave: ¿Quién es el principal ganador del sorteo? La respuesta está clara: Loterías y Apuestas del Estado, una empresa pública que como empresa tiene sus beneficios. Hagamos unos pequeños números para ver el volumen de dinero que mueve este sorteo de Lotería de Navidad. Merece la pena.

Para empezar, de cada número se imprimen 10 Décimos, lo que se denomina SERIE y cada Serie se imprime 172 veces. Por tanto:
POR CADA NÚMERO Loterías y Apuestas del Estado pone a la venta 10x172 =1.720 décimos, lo que le proporciona un ingreso de 1.720 décimos x 20 euros = 34.400 euros.

COMO SE JUEGAN 100.000 números, el ingreso asciende a 100.000 números x 34.400 euros = 34.400 millones de euros, de los cuales (i) reparte en premios el 70% de la emisión =34.400 millones € x 0.7= 2.408 millones de euros y (ii) y obtiene un ingreso directo por el resto, el 30% de la emisión 34.400 millones de euros x 0.3= 1.032 millones de euros y (iii) un ingreso indirecto del 20% de los premios mayores de 40.000 euros (que están exentos de tributar a Hacienda como se explicó anteriormente).

LOTERÍAS
Y APUESTAS DEL ESTADO

¿Sabías que...?

En el año 2022, según consta en los Presupuestos Generales del Estado, el Gobierno prevé ganar los 1.800 millones de euros, los cuales se incorporarán a los recursos del Estado para prestar servicios públicos: Sanidad, Educación, Infraestructuras...

Por tanto, el principal ganador es la Banca, Loterías y Apuestas del Estado, y el que nunca pierde y, además, gana, es aquella persona que no compra lotería y recibe mejoras públicas vía presupuestos.


Aquí tu publicidad
Contacta en publicidad@thedatascientist.es
50% descuento 3 primeros meses