machina .ai

a reboot company.

Desafíos de Kaggle

CONTENIDO

Los desafíos de Kaggle fortalecen al mundo del Data Science y Machine Learning 3

¿Quién usa Kaggle? 4

¿Cómo nació Kaggle? 5

¿Qué herramientas ofrece Kaggle? 6

Impacto de los rankings de Kaggle 8

Nuevos desafíos propuestos por Kaggle 9

Atracción de talento 11

Referencias 12

Los desafíos de Kaggle fortalecen al mundo del Data Science y Machine Learning

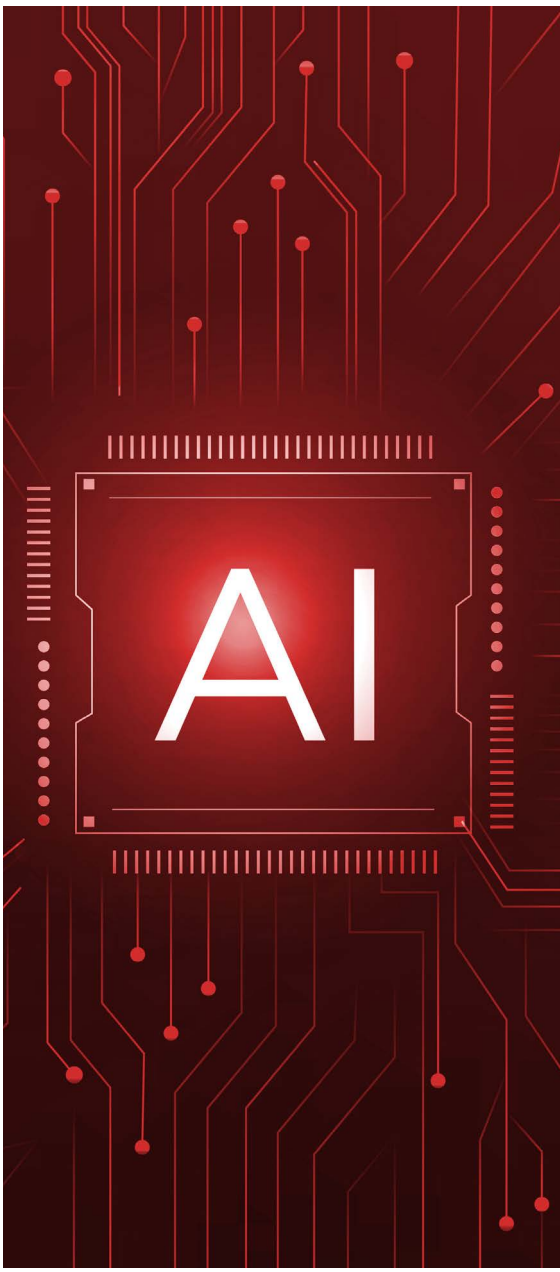
Los científicos de datos y machine learning engineers pueden beneficiarse enormemente, mejorar sus habilidades y colaborar con profesionales de diferentes partes del mundo cuando participan en los retos y desafíos que durante todo el año lanza la plataforma en línea y comunidad de expertos llamada ***Kaggle***.

Fundada en 2010 y adquirida por ***Google*** en 2017, ***Kaggle*** es ampliamente conocida porque convoca a los desarrolladores para participar en competencias, usando conjuntos de datos públicos, cursos y recursos educativos para profesionales y entusiastas de la ciencia de datos en todo el mundo.

En México, una de las compañías de ***inteligencia artificial*** y ciencia de datos que estimula la participación de sus ingenieros en ***Kaggle*** para mejorar sus habilidades, es ***Machina***, que ofrece soluciones y servicios en áreas como procesamiento del lenguaje natural, visión por computadora, analítica avanzada y aprendizaje automático.

Por ser una plataforma líder en su campo, utilizada por una amplia comunidad de científicos de datos y aprendizaje automático; conocer lo que ocurre dentro de la plataforma ***Kaggle*** es importante para entender el estado del arte en el campo donde trabaja ***Machina***.

Como ejemplo de sus alcances, para 2024 ***Kaggle*** planteó a su comunidad decenas de desafíos para abordar problemas tan diferentes entre sí como la detección rápida de fraudes con tarjetas de crédito; la identificación de noticias falsas o fake news; la catalogación más eficaz de películas de acuerdo con elecciones anteriores de los espectadores o el asesoramiento a principiantes que comienzan a invertir pequeñas cantidades de dinero en la bolsa de valores[1].



> "Actualmente ***Kaggle*** cuenta con más de 16 millones de afiliados[2] y participantes de 194 países que usan las bases de datos de uso libre y los entornos de programación online o Kernels de la plataforma[5]."

¿Quién usa Kaggle?

Actualmente la plataforma abierta ***Kaggle*** cuenta con más de 16 millones de afiliados[2] y participantes de 194 países que la consideran como su Airbnb porque pasan noches y fines de semana dentro de ella[3].

Por medio de retos y desafíos de participación abierta, además de recursos y herramientas gratuitas para los profesionales y entusiastas en estos campos, así como un ranking público de los solucionadores más eficaces, esta comunidad genera cada año cientos de soluciones reales para preguntas y problemas de usuarios de ***inteligencia artificial*** (IA), pero también con impacto económico y social.

Kaggle es el espacio de trabajo compartido de usuarios que van desde doctores en ciencias de la computación que realizan investigaciones de vanguardia hasta principiantes absolutos y autodidactas.

¿Cómo nació Kaggle?

El origen de ***Kaggle*** ocurrió en Melbourne, Australia, a partir de la idea de dos desarrolladores de software que deseaban estimular la competencia entre colegas para acelerar el desarrollo de soluciones basadas en el aprendizaje automatizado.

Estos dos especialistas, ***Anthony Goldbloom*** y ***Jeremy Howard***, construyeron la plataforma donde se llevarían a cabo las competencias que habían concebido y la pusieron en línea en abril de 2010. Su carta de presentación pública era ofrecer competencias de aprendizaje automático. Posteriormente

En la imagen: Jeremy Howard

En la imagen: Anthony Goldbloom

comenzaron a ofrecer otros servicios atractivos para los usuarios, como una plataforma de datos pública, un banco de trabajo basado en la nube para la ciencia de datos y educación en ***inteligencia artificial***.

Después de lanzar su emprendimiento, ***Anthony Goldbloom*** y ***Jeremy Howard*** incorporaron a ***Nicholas Gruen***, quien fue el presidente fundador. Como no existía otra plataforma con sus características, rápidamente creció su comunidad de usuarios y el valor de la empresa se elevó, en un solo año, a 25 millones de dólares.

En 2022, ***Kaggle*** mudó su sede central a Silicon Valley, recaudó 11 millones de dólares de inversionistas como ***Hal Varian*** (economista jefe de ***Google***), ***Max Levchin*** (de ***PayPal***), ***Index*** y ***Khosla Ventures.*** Finalmente fue adquirido por ***Google*** en marzo de 2017[3].

¿Qué herramientas ofrece Kaggle?

Para los desarrolladores, ***Kaggle*** proporciona recursos como data sets o conjuntos de datos públicos, tutoriales de aprendizaje automático y cuadernos de código que permiten aprender y practicar habilidades de ciencia de datos. Esto puede ayudar a principiantes y programadores ya experimentados a mejorar sus conocimientos y experiencia, haciéndolos más atractivos para posibles empleadores.

Cuatro de las características clave de la plataforma y la comunidad ***Kaggle*** son:

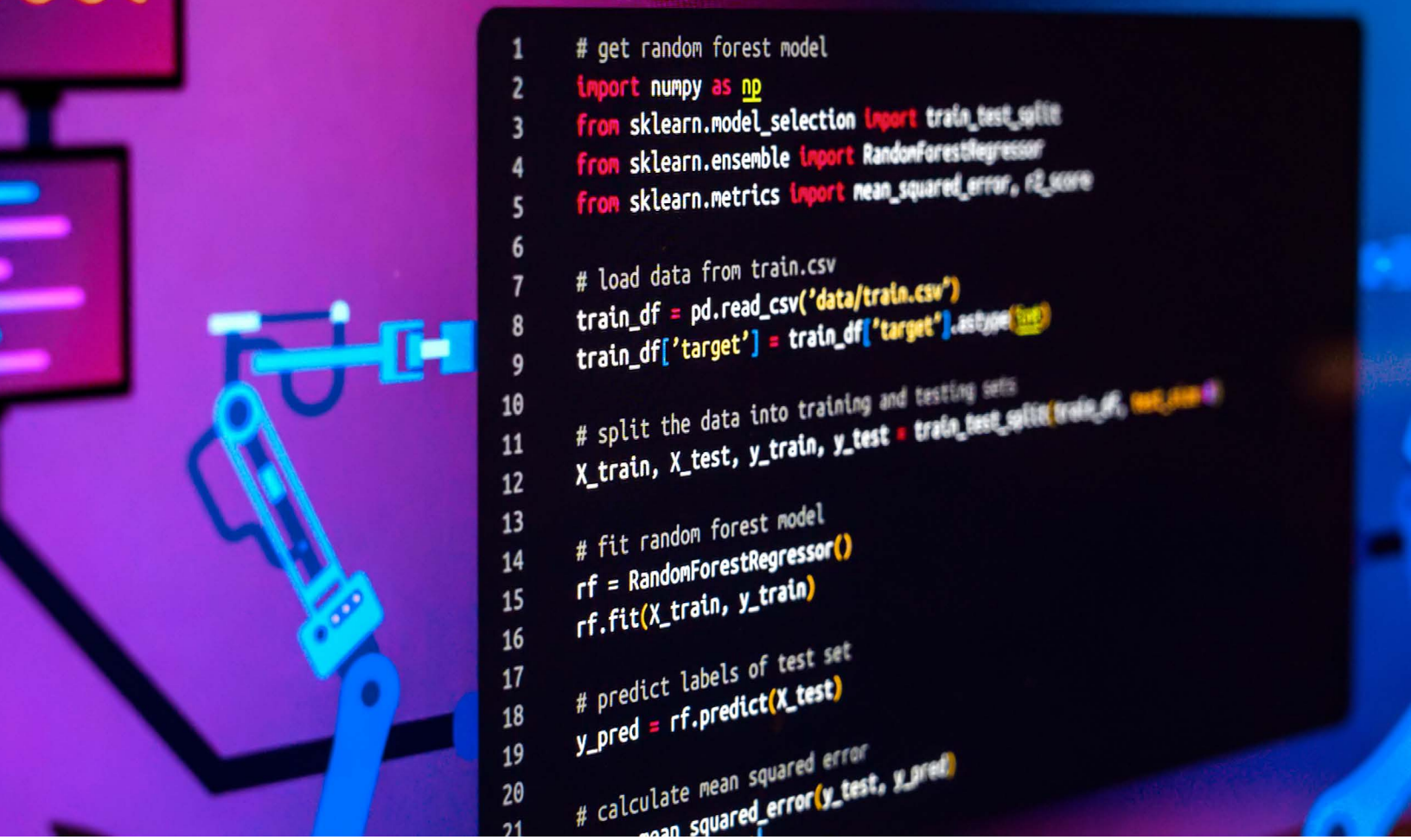
1) Los concursos de aprendizaje automático

Kaggle organiza regularmente competencias en las que los participantes ponen a prueba sus habilidades y conocimientos para resolver problemas de aprendizaje automático y análisis de datos. Estos concursos suelen ofrecer premios en efectivo y son una excelente manera de aplicar y mejorar las habilidades en aprendizaje automático. La mayor parte de los premios de estas competencias rondan entre los 15 mil y los 200 mil dólares para el ganador [4,5].

2) Conjuntos de datos públicos

Kaggle proporciona una amplia variedad de conjuntos de datos públicos que los usuarios pueden explorar, analizar y utilizar para proyectos personales o educativos. Estos conjuntos de datos cubren una amplia gama de temas y fuentes de datos; por ejemplo, cifras y tendencias de suicidios en todo el mundo; estadísticas de ventas de autos en Estados Unidos desde 1974; cifras recientes de ventas en línea de Amazon o cambios en las preferencias de alojamientos a través de la plataforma Airbnb, además de todos los archivos públicos de datos estadísticos del gobierno de Estados Unidos[6].

"Las competencias organizadas por ***Kaggle*** suelen ofrecer premios en efectivo por resolver problemas y mejorar las habilidades en aprendizaje automático. La mayor parte de los premios de estas competencias rondan entre los 15 mil y los 200 mil dólares para el ganador[4,5]."



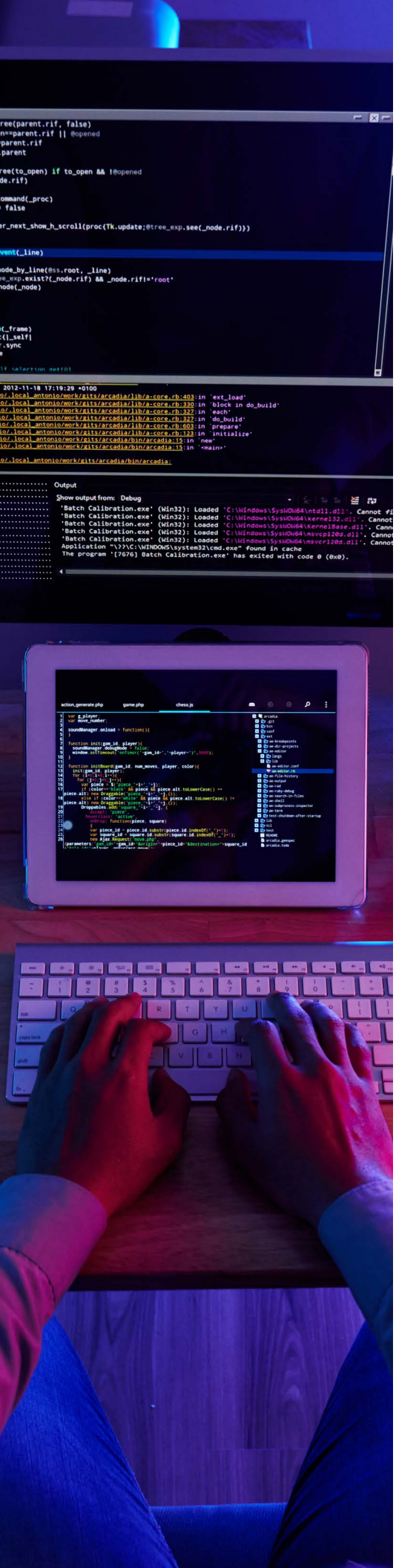
La plataforma tiene herramientas internas para clasificar las bases de datos en campos como entretenimiento, finanzas, ciencias sociales, biología, industria manufacturera, entre otros; también tiene pantallas para identificar las bases de datos que son tendencia en un momento determinado.

3) Entornos de programación Kernels

Los Kernels de ***Kaggle*** son entornos de programación en línea que permiten a los usuarios escribir y ejecutar código *Python* o *R* en la nube. Los Kernels son útiles para experimentar con datos, desarrollar modelos de aprendizaje automático y compartir análisis con la comunidad. Los Kernels ofrecen modelos de ciencia de datos para los afiliados a ***Kaggle*** o ***kagglers*** de 194 países con una variedad de antecedentes, desde ciencias de la computación hasta biología. Dentro de los Kernels las empresas e investigadores publican sus datos y problemas de science data y machine learning, para que los científicos de datos de todo el mundo compitan para producir los mejores modelos. Los concursos de ***Kaggle*** están abiertos a todos los científicos de datos registrados en el sitio[7].

4) Foros y discusiones

Kaggle cuenta con foros activos donde los usuarios pueden hacer preguntas, compartir conocimientos, discutir técnicas y colaborar en proyectos. Estos foros son una excelente manera de obtener ayuda y orientación de otros miembros de la comunidad.



Impacto de los rankings de Kaggle

Para los desarrolladores de software en ***data science*** y machine learning, los rankings de ***Kaggle*** pueden tener un impacto positivo para conseguir nuevas oportunidades laborales; es por esto que compañías como ***Machina*** lo consideran como una herramienta de preselección de talento laboral. Una buena posición en rankings se puede sumar a la cantidad de experiencia práctica y las habilidades técnicas de cada programador para abrirse un espacio favorable en un mundo laboral cada vez más competitivo[8,9].

Obtener una alta posición en los rankings de ***Kaggle*** demuestra habilidades sólidas en la resolución de problemas de ciencia de datos y aprendizaje automático y permite que la solución del reto sea presentada a la industria como un resultado tangible de las capacidades del programador. Además, competir y ganar alguno de estos retos otorga visibilidad ante los empleadores y los propios colegas porque los rankings de ***Kaggle*** son públicos y pueden aumentar la presencia pública de un desarrollador; lo que le proporciona más oportunidades de colaboración, networking y reconocimiento en la industria.

Adicionalmente, un buen desempeño en competiciones de ***Kaggle*** puede agregar credibilidad profesional a un desarrollador de software, lo que puede ser valorado por empleadores actuales y futuros. Otro valor agregado de competir y ganar alguno de los desafíos o retos de ***Kaggle***, es que en los últimos años han aparecido algunas empresas de software y reclutadores que buscan activamente desarrolladores con un historial destacado en ***Kaggle*** para puestos relacionados con la ciencia de datos y el aprendizaje automático. Tener un buen ranking en ***Kaggle*** puede aumentar las oportunidades de ser contactado para roles interesantes y bien remunerados.

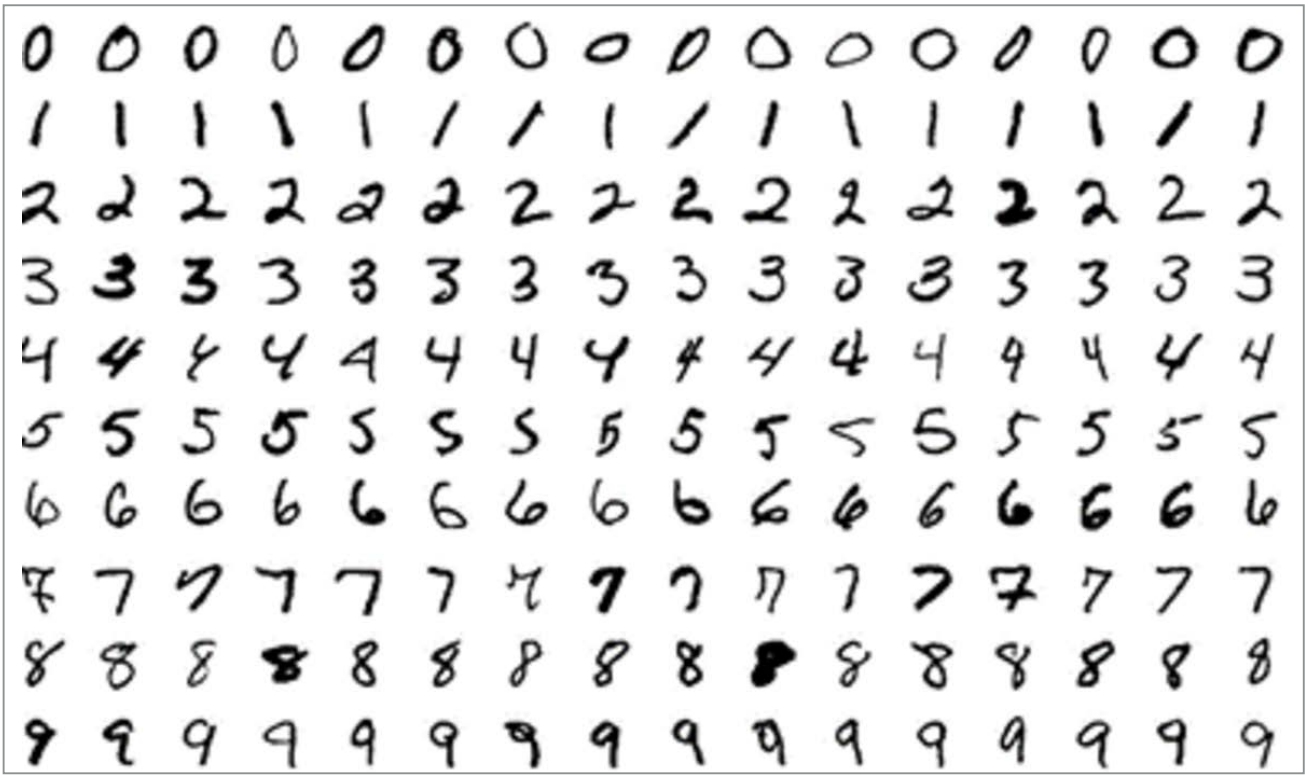
Nuevos desafíos propuestos por Kaggle

Para estimular el aprendizaje y desarrollo de nuevas habilidades en su comunidad, ***Kaggle*** ha propuesto algunos retos que van desde fáciles hasta avanzados para abordar en 2024. El objetivo es que, al resolver estos desafíos los desarrolladores de software puedan obtener experiencia en varios aspectos, desde el preprocesamiento de datos y el análisis exploratorio de datos hasta la implementación de modelos de aprendizaje automático.

Algunos de los retos más visibles son los siguientes:

El primer reto de ***Kaggle***, aparentemente fácil, pide crear un modelo de reconocimiento visual capaz de clasificar números escritos a mano mediante el conjunto de datos ***MNIST***. Este proyecto es una introducción fundamental para la clasificación de imágenes y, a menudo, se considera un punto de partida para aquellos que se inician en el aprendizaje profundo. Como insumo, ***Kaggle*** ofrece un banco de datos ***MNIST*** que contiene imágenes de números escritos a mano del 0 al 9 en escala de grises; para que los programas trabajen en su identificación[10].

Un segundo reto, de dificultad más elevada, pide crear un modelo de aprendizaje automático para segmentar a los clientes en función de su comportamiento de compra anterior, de modo que cuando



En la imagen: **Conjunto de datos MNIST**

el mismo cliente vuelva, ese sistema pueda recomendar cosas pasadas para aumentar las ventas. De esta manera, al utilizar la segmentación, las organizaciones pueden orientar el marketing y ofrecer servicios personalizados a todos los clientes. Dado que se trata de un tipo de problema de aprendizaje no supervisado, no se requerirán etiquetas para tales tareas y se pueden usar conjuntos de datos que contengan datos de transacciones de clientes, conjuntos de datos minoristas en línea o cualquier conjunto de datos relacionado con el comercio electrónico, como Amazon, ***Flipkart***[11].

Entre los desafíos de complejidad media, para 2024, está el desarrollo de herramientas para detectar noticias falsas o fake news. Para ese proyecto, se tiene que desarrollar un modelo de aprendizaje automático que ayude a encontrar la diferencia entre artículos de noticias reales y falsas recopilados de diferentes aplicaciones de redes sociales utilizando técnicas de procesamiento de lenguaje natural; este proyecto implica el preprocesamiento de texto, la extracción de características y la clasificación. Para trabajar se sugiere usar tecnologías como bibliotecas de procesamiento de lenguaje natural como *NLTK* o *spaCy* y algoritmos de aprendizaje automático como *Naive Bayes* o modelos de aprendizaje profundo y se ofrece un banco de datos llamado "Conjunto de datos de noticias falsas" en ***Kaggle***[12].

Un ejemplo de los retos más complejos es el proyecto para desarrollar un modelo de aprendizaje automático para detectar transacciones fraudulentas con tarjetas de crédito, lo cual es crucial para que las instituciones financieras mejoren la seguridad, protejan a los usuarios de actividades fraudulentas y faciliten el entorno para diferentes transacciones. Éste es un reto de aprendizaje supervisado en el que los participantes deben recopilar el conjunto de datos, que contiene datos de transacciones de tarjetas de crédito con casos etiquetados de transacciones fraudulentas y no fraudulentas.

En estos proyectos más complejos, ***Kaggle*** recomienda usar algoritmos de detección de anomalías, modelos de clasificación como ***Random Forest*** o ***Support Vector Machines*** y marcos de aprendizaje automático para su implementación. Para implementarlo se debe preprocesar los datos de la transacción, entrenar un modelo de detección de fraude, ajustar los parámetros para obtener un rendimiento óptimo y evaluar el modelo utilizando métricas de evaluación de clasificación como precisión, recuperación y *ROC-AUC*[13].

Atracción de talento

Observar y seguir las competencias y desafíos que se plantean en ***Kaggle*** se ha convertido en un mecanismo de preselección para la atracción de talento para las compañías de desarrollo de software, especialmente en ciencia de datos y aprendizaje automatizado, como es el caso de ***Machina.***

Kaggle permite a los usuarios colaborar con otros usuarios, encontrar y publicar conjuntos de datos, utilizar notebooks integrados GPU y competir con otros científicos de datos para resolver los desafíos de interés común.

Conocer y entender lo que ocurre dentro de la plataforma ***Kaggle*** es indispensable para entender los cambios y tendencias en el campo donde trabaja ***Machina***. Aquellas empresas que llegan a entender bien a la comunidad que participa en estos desafíos logra encontrar a científicos de datos capaces y adecuados para ayudarles a resolver tareas específicas para sus propios proyectos.

Casi tres lustros después de su fundación, es posible decir que se ha hecho realidad la visión de ***Anthony Goldbloom*** y ***Jeremy Howard*** de utilizar la competencia directa para estimular el desarrollo del ***machine learning*** y ***data science*** en todo el mundo.

Referencias

1 Aryan Garg. KD Nuggets. Top 10 Kaggle machine learning projects to become Data scientists in 2024. https://www.kdnuggets.com/top-10-kaggle-machine-learning-projects-to-become-data-scientist-in-2024

2 Kaggle. Número de usuarios únicos de Kaggle, hasta el 31 de diciembre de 2023. https://www.kaggle.com/discussions/general/458300

3 Kaggle. Zeeshan Ul Hassan Usmani. Getting started. What is Kaggle, Why I Participate, what is the impact? 2017. https://www.kaggle.com/discussions/getting-started/44916

4 Kaggle. Competencias vigentes. Marzo 2024. https://www.kaggle.com/competitions

5 Massel Data. ¿Qué es kaggle? Noviembre 2022. https://www.maseldata.com/post/que-es-kaggle#:~:text=Kaggle%20permite%20a%20los%20usuarios,la%20ciencia%20de%20los%20datos

6 Kaggle. Datasets and New Datasetts. Marzo 2024. https://www.kaggle.com/datasets

7 Kaggle. What are Kernels? 2017. https://www.kaggle.com/discussions/general/27328

8 CodersLink. Científicos de datos: los rockstar del mundo TI. 14 de septiembre de 2021. https://coderslink.com/talento/blog/cuanto-gana-data-scientist-en-mexico/

9 Kaggle. State of Machine learning and data science 2020. https://www.kaggle.com/kaggle-survey-2020

10 Kaggle. Desafío de identificación de números escritos a mano. https://www.kaggle.com/code/imdevskp/digits-mnist classification-using-cnn#

11 Kaggle. Desafío de segmentación de consumidores para marketing digital. https://www.kaggle.com/code/fabiendaniel/customer-segmentation

12 Kaggle. Desafío de detección de noticias falsas. https://www.kaggle.com/code/maxcohen31/nlp-fake-news-detection-for-beginners

13 Kaggle. Desafío de detección de fraudes con tarjetas de crédito. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

machina .ai

a reboot company.

Mayo 2024