**SEMANA 3 Week 1 ¿Qué es Ciencia de Datos?**

**Plan de Estudios**

**Módulo 1 - Definición de Ciencia de Datos**

* ¿Qué es la Ciencia de Datos?
* Los Muchos Caminos Hacia la Ciencia de Datos
* Consejos para nuevas científicas de datos
* Lectura: "El Trabajo más Sexy del Siglo XXI"
* Quiz: "El Trabajo más Sexy del Siglo XXI"

**Módulo 2 - Qué hacen los científicos de datos**

* Un Día en la Vida de un Científico de Datos
* Temas de Ciencia de Datos y Algoritmos
* Nube Para Ciencia de Datos
* Lectura: "¿Qué Hace que Alguien sea un Científico de Datos?"
* Quiz: "¿Qué Hace que Alguien sea un Científico de Datos?"

**Módulo 3 - Big Data y minería de datos**

* ¿Cómo deberían las empresas iniciarse en la ciencia de datos?
* Reclutamiento Para Ciencia de Datos
* Aplicaciones de Aprendizaje Automático
* Lectura: "Regresión"
* Quiz: "Regresión"

## Módulo 4 - Temas de ciencia de datos

* Científicos de Datos en la Universidad de Nueva York
* Habilidades de Ciencia de Datos y Big Data
* Qué es Hadoop
* Lectura: "Establecimiento de Objetivos de Minería de Datos"
* Quiz: "Establecimiento de Objetivos de Minería de Datos"

## Módulo 5 - Temas de ciencia de datos (Continuación)

* Redes Neuronales y Aprendizaje Profundo
* Cómo Puede Alguien Convertirse en un Científico de Datos
* Estudiantes de Secundaria y Carreras de Ciencias de Datos
* Aplicaciones del aprendizaje automático
* Reading: "The Final Deliverable"
* Quiz: "The Final Deliverable"

## Módulo 6 - Asignación final

* Lectura: "La estructura del informe"
* Quiz: "La estructura del informe"
* Trabajo final

# Las Claves del Éxito: Consejos Profesionales para Aprender en Coursera

* Bienvenido a Coursera.
* Para ayudarte a aprender m�s r�pido y mejor,
* este video te mostrar�
* siete consejos profesionales para usar la plataforma.
* Nuestros consejos son de personas que han completado el curso con �xito
* y de mentores de la comunidad.
* As� que rev�salos para convertirte en un s�per aprendiz hoy.
* N�mero uno, mot�vate con los plazos de entrega.
* Puede que ya te hayas dado cuenta
* de que hay un plazo de entrega para cada asignaci�n calificada.
* Son plazos sugeridos para ayudarte a
* planificar el aprendizaje y mantenerte motivado.
* Nuestros datos muestran que tener plazos de entrega pueden
* ayudarte a aumentar la tasa de completaci�n en un 17 por ciento.
* Sin embargo, no te asustes si
* tienes que perder un de los plazos de entrega,
* no hay grado de penalizaci�n.
* Si pierdes dos plazos de entrega seguidos o pierdes un plazo en
* dos semanas, tu ver�s la opci�n que
* dice, restablecer mis plazos de entrega,
* aunque no veas esa opci�n
* a�n, puedes enviar el trabajo.
* N�mero dos, �nete a la discusi�n.
* El foro de discusi�n es
* un gran lugar para buscar ayuda para entender
* el contenido del curso o hacer preguntas
* sobre algo que hayas le�do o aprendido en el curso.
* Los foros del curso te permiten buscar y ver si tu pregunta
* ha sido hecha antes o publicar tu propia pregunta.
* Play video starting at :1:11 and follow transcript1:11
* Tambi�n puedes leer y responder a
* las preguntas de otras personas y discusiones.
* Las publicaciones del foro ser�n le�das y respondidas por tus
* compa�eros y por mentores activos en el curso.
* Las investigaciones demuestran que los alumnos que publican en
* los foros de discusi�n, tienen
* un 37 por ciento m�s de probabilidades de completar un curso.
* As� que haz una publicaci�n hoy.
* N�mero tres, toma notas.
* Tal vez quieras tomar notas mientras aprendes
* en l�nea como lo hac�as en la escuela.
* Ahora puedes, tu puedes hacerlo completamente en
* Coursera sin abrir una segunda pesta�a o herramienta.
* Haz clic en Guardar nota debajo de cualquier v�deo del curso
* podr�s a�adir
* tus propias notas junto a el recorte de las transcripciones.
* Tambi�n puedes resaltar y
* asegurar los trozos de texto de la transcripci�n.
* N�mero cuatro, aprende con tus compa�eros.
* Puede que veas asignaciones
* calificadas por tus compa�eros en algunos cursos de Coursera.
* Esa es una gran oportunidad para que recibas
* comentarios de otras personas que est�n tomando el mismo curso.
* Tambi�n necesitar�s revisar
* las asignaciones de otras personas y darles tus propios puntos de vista.
* Tal vez te preguntes, si la calificaci�n de los compa�eros es acertada y �til.
* La investigaci�n dice que s�.
* Estudios de la Universidad de Penn State
* indican que los puntajes de calificaciones hechas por los compa�eros son
* bastante consistentes y muy
* similares a los puntajes de calificaciones de los instructores.
* Cuando veas una asignaci�n calificada por los compa�eros
* deber�as trabajar en ella lo antes posible.
* Puedes dejar suficiente tiempo para que otros revisen
* tu trabajo y escriban comentarios interesantes.
* Si env�as una asignaci�n calificada por tus compa�eros despu�s
* del plazo de entrega personalizado,
* es posible que no recibas suficientes comentarios de tus compa�eros.
* Si necesitas m�s revisiones de tus compa�eros, puedes publicar
* en los foros, pidi�ndole m�s comentarios a tus compa�eros.
* N�mero cinco, utiliza la aplicaci�n Coursera.
* Puedes descargar la aplicaci�n Coursera para
* acceder a tu curso desde cualquier lugar.
* Est� disponible para dispositivos iOS y Android.
* �Tienes algo de tiempo mientras esperas por tu caf�?
* Mira los videos de Coursera en tu tel�fono.
* N�mero seis, demuestra tus habilidades con un certificado.
* Un certificado de curso es una excelente prueba
* de que has completado el curso y dominas las habilidades.
* Puedes ver todos los certificados de los cursos
* que has obtenido en tu p�gina de logros.
* Tambi�n puedes compartirlos en tu perfil de LinkedIn o
* usando una URL segura de Coursera.
* Entonces, �c�mo puedes obtener tu certificado?
* Hay tres simples requisitos.
* Completar la verificaci�n del nombre,
* pasa todas asignaciones requeridas y paga el curso.
* Recuerda que si tomas un curso en modo de prueba gratuita
* no podr�s acceder a las asignaciones calificadas
* ni obtener el certificado.
* Pero puedes cambiar a un certificado pagado
* en cualquier momento durante o despu�s de realizar el curso.
* N�mero siete, el centro de ayuda al estudiante es tu amigo.
* �Tienes m�s preguntas sobre la plataforma Coursera?
* Puedes encontrar la respuesta en nuestro Centro de ayuda al estudiante.
* Hay grandiosos art�culos
* que cubren cada paso de tu aprendizaje.
* Play video starting at :3:53 and follow transcript3:53
* Bien. Espero que estos consejos profesionales te
* puedan ayudar a aprender mejor y m�s r�pido en Coursera.
* Por �ltimo, pero no menos importante, recuerda
* que la plataforma est� construida para ayudarte con el perfeccionamiento de tu aprendizaje.
* As� que contin�a intentando y aprendiendo de los fallos
* hasta que tengas �xito, feliz aprendizaje.

# Qué es la Ciencia de Datos

La ciencia de datos es un proceso, no un evento.

Es el proceso de usar datos para entender diferentes cosas,

para entender el mundo.

Para mi es cuando tienes un modelo o hipótesis de un problema,

y tratas de validar esa hipótesis o modelo con tus datos.

La ciencia de datos es el arte de

descubrir el conocimiento y las tendencias que se esconden tras los datos.

Es cuando traduces los datos a una historia.

Así que cuentas una historia para generar conocimiento.

Y con este conocimiento,

puedes hacer decisiones estratégicas para alguna compañía o institución.

La ciencia de datos es un campo a procesos y sistemas para extraer

datos de varios formas ya sea de manera estructurada o no.

La ciencia de datos es el estudio de los datos.

Así como las ciencia biológica es el estudio de la biología,

la ciencia física el estudio de las reacciones físicas.

Los datos son reales, tienen propiedades reales,

y necesitamos estudiarlos si vamos a trabajar con ellos.

La ciencia de datos involucra datos y algunos signos.

La definición o el nombre se forjó

en los 80s y 90s cuando algunos profesores estaban trabajando en los planes de estudio de estadística,

y decidieron que seria mejor llamarlo ciencia de datos.

¿Pero que es la Ciencia de Datos?

Yo vería la ciencia de datos como el intento de uno por trabajar con los datos,

para encontrar respuestas a las preguntas que se están explorando.

Más concretamente, es más sobre datos de lo que es sobre ciencia.

Si tienes datos y tienes curiosidad,

y trabajas con datos,

los manipulas, los exploras,

el ejercicio de realizar el análisis de datos,

tratar de obtener algunas respuestas de ello, eso es ciencia de datos.

La ciencia de datos es relevante hoy porque tenemos toneladas de datos disponibles.

Solíamos preocuparnos sobre la falta de datos.

y ahora tenemos un diluvio de datos.

En el pasado, no teníamos algoritmos, ahora los tenemos.

En el pasado el software era caro,

ahora es de libre uso y gratis.

En el pasado, no podíamos almacenar grandes cantidades de datos,

ahora por una fracción del costo,

podemos tener millones de conjuntos de datos a un costo muy bajo.

Así que, las herramientas para trabajar con los datos,

la variabilidad de los datos,

la habilidad para almacenar y analizar datos,

todo es barato, esta disponible,

es ubicuo y está aquí.

Nunca ha habido una época mejor para ser un científico de datos.

# Los Muchos Caminos hacia la Ciencia de Datos

[SONIDO].

[MÚSICA] Soy todo oídos.

La ciencia de los datos ni siquiera existía cuando yo estaba creciendo.

No es algo con lo que haya despertado y

dicho, quiero ser una científico de datos cuando sea grande.

No, eso no existía.

No sabía que estaría trabajando en la ciencia de los datos.

Cuando crecí, no había ese campo llamado ciencia de datos.

Y creo que realmente nuevo.

La ciencia de datos no existió hasta el 2009, 2011.

Alguien como DJ Patil o Andrew Gelman acuñaron el término.

Antes de eso, eran solo estadísticas.

Y yo no quería ser nada de eso.

Yo quería estar en el negocio.

Entonces encontré la ciencia de datos algo mucho mas interesante.

Yo estudié estadística, así es como empecé.

Play video starting at ::45 and follow transcript0:45

Pasé por muchas diferentes etapas de mi vida en donde quería ser cantante y

luego doctor.

Y entonces me di cuenta que era buena en matemáticas.

Así que escogí un área que se enfocaba en los análisis cuantitativos.

Y desde entonces creo que quería trabajar con datos.

No necesariamente ciencia de datos como la conocemos hoy.

La primera vez que tuve contacto con la ciencia de datos,

fue en mi primer año como ingeniero mecánico.

Y las firmas de consultoría estratégica usan la ciencia de datos para tomar decisiones.

Así que fue mi primer contacto con la ciencia de datos.

Tenía un problema complicado que necesitaba resolver, y

las técnicas usuales que teníamos no podían ayudarme a resolver el problema.

Me gradué con un título en matemáticas en la peor época posible,

justo después de la crisis económica, y tu realmente tienes que ser útil para obtener un trabajo.

Así que fui a obtener un título en estadísticas.

Y entonces trabajé los suficientes empleos que llamaban a científicos de datos que

de pronto me convertí en uno.

Mi licenciatura fue en negocios, y mi especialización en política,

filosofía y economía.

Y después hice una maestría en negocios analíticos en

La Universidad de Nueva York en la escuela de negocios Stern.

Cuando salí de mi licenciatura, la primera compañía a la que me uní, resultó que

que estaban analizando datos de puntos de venta electrónicos para fabricantes minoristas.

Y lo que estábamos haciendo era ciencia de datos.

Pero realmente comenzamos a utilizar ese término mucho después.

De hecho, diría que hace cuatro o cinco años es cuando empezamos a llamarlo analistas y

la ciencia de datos.

Tuve varias opciones para mi internado aquí en Canadá.

Y una de las opciones era trabajar con la ciencia de datos.

Yo solía trabajar en desarrollo de proyectos.

Pero creo que fue una buena opción.

Y entonces comencé mi internado como científico de datos.

Soy ingeniero civil por capacitación, así que todos los ingenieros trabajan con datos.

Diría que el uso convencional de ciencia

de datos en mi vida empezó con la investigación de transporte.

Empecé construyendo grandes modelos intentando pronosticar el tráfico en las calles, intentando

determinar la congestión y emisión de gases de efecto invernadero o las emisiones de escape.

Así que creo que así mi comienzo.

Y comencé construyendo estos modelos cuando aun era un estudiante de

la Universidad de Toronto.

Empecé trabajando con una gran cantidad de sets de datos, buscando muestras de hogares,

digamos, 150,000 hogares de medio millón de viajes.

Y so, también, estoy hablando de mitad de los 90s cuando esto era

supuestamente una gran cantidad de set de datos, pero no en los términos de hoy en día.

Pero así es como comencé.

Y he continuado trabajando con esto.

Y entonces me cambié a la Universidad McGill en donde era profesor de ingeniería de

transporte.

Y construí aun modelos de datos mas grandes que involucraban datos y análisis.

Y por lo que diría, si, la investigación de transporte me trajo a la ciencia de datos.

[MÚSICA] Soy todo oídos.

# Consejos Para los Científicos de Datos

Mi consejo para un aspirante a científico de datos es que debe ser curioso,

bueno para argumentar y crítico.

La curiosidad es una cualidad obligatoria.

Si no eres curioso, no sabrías qué hacer con los datos.

El ser crítico es necesario porque si no tienes

ideas preconcebidas sobre las cosas no sabrías por dónde empezar.

Debes ser bueno argumentando pues si tienes esa habilidad, y debes abordar un tema,

por lo menos puedes empezar en un punto y luego aprender de los datos, para luego

modificar tus suposiciones e hipótesis y tus datos te ayudarán a aprender.

Y podrías empezar en el punto equivocado.

Podrías decir: "Yo pensé, yo creí ésto,

pero ahora gracias a los datos yo sé ésto".

Por lo tanto, esto permite un proceso de aprendizaje.

Así, la curiosidad te ayuda a tomar una posición,

una fuerte posición, para partir de ahí y seguir adelante.

La otra cosa que el científico de datos necesita es

algo de comodidad y flexibilidad con las plataformas de análisis: algunos programas,

alguna plataforma de computación, pero eso es secundario.

Lo más importante es la curiosidad y la capacidad de tomar posiciones.

Una vez que has hecho esto, una vez que has analizado,

entonces ya tienes algunas respuestas.

Y esto es lo último que necesita un científico de datos:

la capacidad para contar una historia.

Que una vez que tengas tu análisis,

una vez que tengas tus tablas,

deberías ser capaz de contar una gran historia de ello.

Porque si no logras contar una gran historia,

tus hallazgos permanecerán ocultos,

enterrados, y nadie los sabría.

Tu ascenso a la trascendencia depende bastante de tu capacidad de contar grandes historias.

Un punto de partida sería ver cuál es tu ventaja competitiva.

¿Quieres ser un científico de datos en cualquier campo o un campo específico?

Porque, supongamos que quieres ser un científico de datos y trabajas para

una empresa de IT o una empresa de Internet o basada en la web,

entonces necesitas un conjunto diferente de habilidades.

Y si quieres ser un científico de datos en la industria de la salud,

entonces necesitas un conjunto diferente de habilidades.

Así que primero averigua lo que te interesa,

y cuál es tu ventaja competitiva.

Tu ventaja competitiva no es necesariamente tu capacidad de análisis.

Tu ventaja competitiva es tu comprensión de algún aspecto de la

vida donde tus habilidades exceden las de los demás.

Tal vez sean las películas, tal vez sea el comercio,

tal vez es la salud, o quizás la computación.

Una vez has averiguado donde se encuentra tu experiencia,

entonces empiezas a adquirir habilidades analíticas.

Qué plataformas aprender y esas plataformas,

estas herramientas, deben ser específicas para el sector que te interesa.

Y una vez que has adquirido algún dominio de las herramientas,

lo siguiente debe ser aplicar tus habilidades a problemas reales,

y mostrarle al resto del mundo lo que puedes hacer.

# Ciencia de Datos: El Trabajo más Sexy del Siglo XXI

**Ciencia de Datos: El Trabajo más Sexy del Siglo XXI**

En el mundo basado en datos, los científicos de datos han surgido como un producto de moda. El objetivo es encontrar el mejor talento en ciencia de datos. Los expertos estiman que millones de empleos en ciencia de datos podrían quedar vacantes por falta de talento disponible. La búsqueda global de científicos expertos en datos no es simplemente una búsqueda de estadísticos o informáticos. De hecho, las empresas están buscando personas completas que posean la experiencia en el tema, cierta experiencia en programación y análisis de software y habilidades excepcionales de comunicación.

Nuestra huella digital se ha expandido rápidamente en los últimos 10 años. El tamaño del universo digital era de aproximadamente 130 billones de gigabytes en 1995. Para 2020, este número aumentará a 40 trilliones de gigabytes. Las empresas competirán por cientos de miles, sino millones, de nuevos trabajadores necesarios para navegar por el mundo digital. No es de extrañar que la prestigiosa Harvard Business Review califique a la ciencia de datos como "el trabajo más sexy del siglo XXI".

Un informe del Instituto Global McKinsey advierte sobre una gran escasez de talento para datos y análisis. "Para 2018, solo Estados Unidos podría enfrentar una escasez de 140,000 a 190,000 personas con profundas habilidades analíticas, así como 1.5 millones de gerentes y analistas con los conocimientos para usar el análisis de big data para tomar decisiones efectivas".

Debido a que la revolución digital ha tocado todos los aspectos de nuestras vidas, la oportunidad de beneficiarse al aprender sobre nuestros comportamientos es más que nunca. Con los datos correctos, los especialistas en marketing pueden echar un vistazo a nuestra formación de hábitos. La investigación en neurología y psicología revela cómo se forman los hábitos y las preferencias y los minoristas como Target se benefician de ello. Sin embargo, los minoristas solo pueden hacerlo si tienen científicos de datos trabajando para ellos. Por esta razón, es "como una carrera armamentista contratar estadísticos hoy en día", dijo Andreas Weigend, el ex jefe científico de Amazon.com.

Todavía existe la necesidad de convencer a los ejecutivos de C-suite de los beneficios de los datos y el análisis. Parece que la alta gerencia podría estar uno o dos pasos detrás de la gerencia media para estar informada sobre el potencial de la planificación basada en análisis. El profesor Peter Fader, que gestiona la Customer Analytics Initiative en Wharton, sabe que los ejecutivos llegan al C-suite sin tener que interactuar con los datos. Él cree que el cambio real sucederá cuando los ejecutivos estén bien versados en datos y análisis.

SAP, líder en datos y análisis, informó en una encuesta que el 92% de las empresas que respondieron en su muestra experimentaron un aumento significativo en sus tenencias de datos. Al mismo tiempo, tres cuartos identificaron la necesidad de nuevas habilidades de ciencia de datos en sus empresas. Accenture cree que la demanda de científicos de datos puede superar la oferta en 250,000 solo en 2015. Una encuesta similar de 150 ejecutivos realizada por KPMG en 2014 encontró que el 85% de los encuestados no sabían cómo analizar los datos. "La mayoría de las organizaciones no pueden conectar los puntos porque no entienden completamente cómo los datos y los análisis pueden transformar sus negocios", Alwin Magimay, jefe de digitales y analítica de KPMG UK, dijo en una entrevista en mayo de 2015.

Bernard Marr escribiendo para Forbes también plantea preocupaciones sobre el talento analítico insuficiente. "Simplemente no hay suficientes personas con las habilidades necesarias para analizar e interpretar esta información, transformándola de datos numéricos en bruto (u otros) en ideas procesables, el objetivo final de cualquier iniciativa impulsada por Big Data", escribió. Bernard cita una encuesta realizada por Gartner de líderes empresariales de los cuales más del 50% informó la falta de experiencia interna en ciencia de datos.

Bernard informó sobre Walmart, que recurrió al crowdsourcing para sus necesidades analíticas. Walmart se acercó a Kaggle para organizar una competencia para analizar sus datos propietarios. El minorista proporcionó datos de ventas de una lista restringida de tiendas y pidió a los competidores que desarrollaran mejores pronósticos de ventas basados en esquemas de promoción.

Dada la escasez de científicos de datos, los empleadores están dispuestos a pagar grandes cantidades por el talento. Michael Chui, director de McKinsey, lo sabe muy bien. La ciencia de datos "se ha vuelto relevante para todas las empresas... Hay una guerra por este tipo de talento", dijo en una entrevista. Tomemos a Paul Minton, por ejemplo. Estaba ganando $ 20,000 sirviendo mesas en un restaurante. Se había especializado en Matemáticas en la universidad. El Sr. Minton tomó un curso de programación de tres meses que lo cambió todo. Hizo más de $ 100,000 en 2014 como científico de datos para un startup web en San Francisco. "Seis cifras, desde el principio... Para mí, fue sorprendente", dijo el Sr. Minton.

¿Podría el Sr. Minton ser excepcionalmente afortunado, o son salarios tan altos la norma? La suerte tuvo poco que ver con eso; el New York Times reportó $ 100,000 como el salario base promedio de un ingeniero de software y $ 112,000 para científicos de datos.

# Un Día en la Vida de un Científico de Datos

He construido antes un motor de recomendaciones como parte de una organización muy grande y trabajado

con todo tipo de ingenieros y contables para diferentes partes del problema.

Es con uno de los que estoy más contento porque en última instancia,

encontré una solución muy simple que era fácil de entender a todos los niveles,

desde el estudio de ejecutivos, hasta los ingenieros y desarrolladores.

Al final, fue tan eficiente como algo realmente más complejo

y pudieron haber gastado más tiempo en ello

Cuando estaba en la universidad,

tuvimos un problema al querer predecir el florecimiento de algas.

Este florecimiento de algas puede causar un aumento

de toxicidad en el agua y puede causar problemas a la compañía de tratamiento de agua.

No podíamos predecirlo con nuestra preparación de ingeniería química.

Así que usamos redes neuronales artificiales para predecir cuando ocurriría este florecimiento.

De tal manera que las compañías de tratamiento de agua pudieran manejar mejor este problema.

En Toronto, el tránsito público es operado por la Comisión de Tránsito de Toronto.

Los llamamos TTC. Es una de

las autoridades más grandes en la región de Norte América.

Y un día me contactaron y dijeron:"Tenemos un problema".

Yo dije:"Bien, ¿cuál es el problema?",

y ellos dijeron:"Bueno, pues tenemos datos de las quejas

y nos gustaría analizarlos, así que necesitamos tu ayuda".

Yo dije:"Bien, me haría muy feliz ayudar".

Y pregunté:"¿Cuántas quejas tienen?",

"Algunas" dijeron. Y dije:

"¿Cuántas?" Tal vez medio millón.

Así que dije:"Bueno, empecemos a trabajar en ello".

Así que recibí los datos y empecé a analizarlos.

Básicamente, ellos habían hecho un gran trabajo guardando

algunos datos en formato tabular y otros datos sin estructura.

Y en ese caso, los datos tabulares eran sobre cuando se había recibido la queja,

quien la recibió, cual era el tipo de queja,

estaba resuelta, de quien era la falta.

Y la parte sin estructura era el intercambio de correos electrónicos y faxes.

Así que, imagínate mirar

como medio millón de intercambios de correo y tratar de obtener algunas respuestas de ello.

Así que empece a trabaja con eso.

La primera cosa que quería saber por que la gente se quejaba

y si había un patrón o si había algunos días cuando hay más quejas que en otros.

Y mire a los datos y los analicé de diferentes formas

y no pude encontrar las razones

para que las quejas fueran más en cierto día y menos en otros.

Y seguí así como un mes más o menos.

Y entonces, un día estaba bajándome del autobús en Toronto,

y aun estaba pensando al respecto.

Salí sin fijarme en el suelo,

Y piso un charco de agua.

Así que ahora, ahí estaba mi tobillo en el agua,

y estaba entonces con un pie seco y el otro mojado.

Así que estaba extremadamente molesto.

Y estaba caminando de regreso cuando caí en cuenta,

y dije:"Bien, espera un momento.

Hoy llovió inesperadamente,

y no estaba preparado.

Es por eso que estoy mojado, y no estaba mirando hacia adelante".

¿Qué tal si hay una relación entre

el clima extremo y el tipo de quejas que recibe el TTC?

Así que fui al sitio web del medio ambiente de Canada,

y obtuve datos sobre lluvias y precipitaciones,

el viento y la luz.

Y allí encontré algo muy interesante.

Los 10 días mas excesivos para las quejas.

Los 10 días cuando la gente se quejaba más eran los días cuando el clima era malo.

Había lluvia inesperada,

una caída extrema de temperatura,

mucha nieve, un día con mucho viento.

Así que regrese con los ejecutivos del TTC y dije:

"Tengo buenas noticias y malas noticias".

La buena noticia es,

Sé por qué la gente se queja excesivamente en ciertos días.

Y se la razón de ello. La mala noticia es,

es que no hay nada que puedan hacer al respecto.

# Temas de Ciencia de Datos y Algoritmos

Me encantan las regresiones, podría decir que fue uno de los

primeros conceptos que me ayudaron a entender los datos y por eso me gustan.

Realmente me gusta la visualización de datos porque pienso que es un elemento clave

para transmitir el mensaje hacia aquellas personas que no estan familiarizados

con la ciencia de datos. Redes neuronales artificiales. Soy un verdadero apasionado

de la redes neuronales porque tenemos mucho que aprender de la naturaleza y por eso

tratamos de imitar nuestro cerebro. Pienso que podemos hacer muchas aplicaciones

partiendo del comportamiento biológico y crear algoritmos. Me encanta

visualizar datos usando R. El algoritmo de vecinos cercanos es el mas sencillo pero

muchas veces genera mejores resultados que otros

algoritmos mas complicados que tienden a sobre ajustar los resultados

Los datos estructurados son parecidos a los datos tabulados con los cuales estamos familiarizados

al usar formato Excel donde tienes filas y columnas, y son llamados

datos estructurados. Datos no estructurados son básicamente aquellos que vienen

principalmente de la web donde encontramos datos no tabulados y los textos no estan en filas y columnas,

y algunas veces son video y audio sobre los que querrías implementar

algoritmos sofisticados para extraer datos. Realmente muchas veces tenemos

datos no estructurados y dedicamos mucho tiempo y esfuerzo

para poder analizarlos. Por lo tanto, si tienes algo que encaja bien en

tablas y columnas, adelante, son datos estructurados, pero si

tienes un log de la web o estas intentando extraer información

de paginas web, tienes una cantidad extrema de paginas web

con datos no estructurados que requieren un mas esfuerzo para conseguir

la información contenida en ellas. Dejame explicarte la regresión en términos sencillos.

Si alguna vez has tomado un taxi entonces entiendes una regresión.

Asi es como funciona. En el momento que te sientas en un taxi, puedes

ver un importe fijo de 2,5 dolares . Tanto si el taxi se mueve o si sales del taxi

es la cantidad que debes. En el momento que entras dentro del taxi

es un importe fijo que tienes que pagar una vez estas dentro del taxi

Cuando comienza a moverse, cada cien metros, la tarifa se incrementa

en un cierto importe, una fracción, por lo que hay una relación entre

la distancia y el importe que vas a pagar según aumenta. Y, si

no estas en movimiento y estas parado, entonces cada minuto adicional

vas a pagar mas, por lo que los minutos

y la distancia incrementan la tarifa, y mientras todo esta ocurriendo

además has de pagar la tarifa base, la cual es constante. Esto es la regresión,

la regresión te dice cual es la tarifa base y cual es la relación entre

tiempo y tarifa que has pagado, así como la distancia que has recorrido

y el temor generado por no conocer dichas relaciones,

saber cuanta gente viajó y cuanto pagaron. Las regresiones te permiten

computar esa constante que no sabias era 2,50 dolares y debías computar la

relación entre la tarifa y la distancia, y entre la tarifa y el tiempo

Esto es una regresión

puedes

# Nube Para Ciencia de Datos

La Nube es un regalo del cielo para los científicos de datos principalmente porque toma sus datos, toma su información

y la pone en la Nube, la pone en el sistema de almacenamiento central.

Le permite evitar las limitaciones físicas de las computadoras y los sistemas que está utilizando,

y le permite implementar las capacidades de análisis y almacenamiento de máquinas avanzadas

que no necesariamente tienen que ser su máquina o la máquina de su empresa.

La Nube le permite no solo almacenar grandes cantidades de datos en servidores en algún lugar de California

o Nevada, sino que también le permite implementar algoritmos informáticos muy avanzados y la

capacidad de realizar cómputos de alto rendimiento utilizando máquinas que no son suyas.

Entonces, piense que tiene información, no puede almacenarla, por lo que la envía al espacio de

almacenamiento, llamémosla Nube.

Y los algoritmos que necesita usar, no los tiene con usted.

Pero luego, en la Nube, tiene esos algoritmos disponibles.

Entonces, lo que hace es implementar esos algoritmos en conjuntos de datos muy grandes y puede

hacerlo a pesar de que sus propios sistemas, sus propias máquinas y su propio entorno informático

no le permitirían hacerlo.

La Nube es hermosa.

Y la otra cosa por la que la Nube es hermosa es que permite que varias entidades trabajen

con los mismos datos al mismo tiempo.

Entonces, puede estar trabajando con los mismos datos que sus colegas en, digamos, Alemania, y

otro equipo en India, y otro equipo en Ghana, están trabajando colectivamente y pueden

hacerlo debido a la información y los algoritmos, las herramientas, las respuestas

y los resultados, todo lo que necesitaban está disponible en un lugar central que llamamos Nube.

El uso de la Nube le permite obtener acceso instantáneo a tecnologías de código aAerto como Apache

Spark, sin la necesidad de instalarlas y configurarlas localmente.

El uso de la nube también le da acceso a las herramientas y bibliotecas más actualizadas sin

la preocupación de mantenerlas y asegurarse de que estén actualizadas.

Se puede acceder a la Nube desde cualquier lugar y en cada zona horaria.

Puede usar tecnologías basadas en la Nube desde su computadora portátil, desde su tableta e incluso desde

su teléfono, lo que permite la colaboración más fácil que nunca.

Múltiples colaboradores o equipos pueden acceder a los datos simultáneamente, trabajando juntos

para producir una solución.

Algunas grandes compañías tecnológicas ofrecen plataformas en la nube, lo que le permite familiarizarse con las

tecnologías basadas en la nube en un entorno preconstruido; IBM ofrece IBM Cloud, Amazon ofrece Amazon

Web Services (o AWS), y Google ofrece Google Cloud Platform.

IBM también proporciona Skills Network Labs (o SN Labs) a estudiantes registrados en cualquiera de

los portales de aprendizaje en IBM Developer Skills Network, donde tiene acceso a herramientas

como Jupyter Notebooks y clusters Spark, para que pueda crear su propio proyecto de ciencia de datos

y desarrollar soluciones.

Con práctica y familiaridad, descubrirá cómo la nube mejora drásticamente la productividad

de los científicos de datos.

# ¿Qué Hace que Alguien sea un Científico de Datos?

**¿Qué Hace que Alguien sea un Científico de Datos?**

Ahora que sabe lo que hay en el libro, es hora de poner algunas definiciones. A pesar de su uso ubicuo, el consenso evade las nociones de big data y data science. La pregunta, “¿quién es un científico de datos?” Está muy viva y es cuestionada por individuos, algunos de los cuales simplemente están interesados en proteger su disciplina o territorio académico. En esta sección, intento abordar estas controversias y explicar por qué una definición estrechamente interpretada de big data o data science dará como resultado la exclusión de cientos de miles de personas que recientemente se han volcado al campo emergente.

"Todo el mundo ama a un científico de datos", escribió Simon Rogers (2012) en The Guardian. El sr. Rogers también rastreó el nuevo amor por el cálculo de números a una cita de Hal Varian de Google, quien declaró que "el trabajo sexy en los próximos diez años serán los estadísticos".

Mientras que Hal Varian nombró a los estadísticos “sexy”, se cree ampliamente que lo que realmente quiso decir eran científicos de datos. Esto plantea varias preguntas importantes:

•       ¿Qué es la ciencia de datos?

•       ¿Cómo difiere de las estadísticas?

•       ¿Qué hace que alguien sea un científico de datos?

En tiempos de big data, una pregunta tan simple como, “¿qué es la ciencia de datos?” puede dar lugar a muchas respuestas. En algunos casos, la diversidad de opiniones sobre estas respuestas raya en la hostilidad.

Defino al científico de datos como alguien que encuentra soluciones a los problemas mediante el análisis de datos grandes o pequeños utilizando las herramientas apropiadas y luego cuenta historias para comunicar sus hallazgos a las partes interesadas relevantes. No uso el tamaño de los datos como una cláusula restrictiva. Un dato por debajo de cierto umbral arbitrario no lo convierte en un científico de datos menos. Mi definición de científico de datos tampoco se limita a herramientas analíticas particulares, como el aprendizaje automático. Mientras uno tenga una mente curiosa, fluidez en el análisis y la capacidad de comunicar los hallazgos, considero a la persona un científico de datos.

Defino ciencia de datos como algo que hacen los científicos de datos. Hace años, como estudiante de ingeniería en la Universidad de Toronto, me quedé atrapado con la pregunta: ¿Qué es la ingeniería? Escribí mi tesis de maestría sobre el pronóstico de los precios de la vivienda y mi tesis doctoral sobre el pronóstico de las elecciones de los constructores de viviendas relacionadas con lo que construyen, cuándo construyen y dónde construyen nuevas viviendas. En el departamento de ingeniería civil, otros trabajaban en el diseño de edificios, puentes, túneles y se preocupaban por la estabilidad de las pendientes. Mi trabajo, y el de mi supervisor, no era su ingeniería tradicional de variedades de jardín. Obviamente, otros me preguntaron repetidamente si mi investigación era realmente ingeniería.

Cuando compartí estas preocupaciones con mi supervisor doctoral, el profesor Eric Miller, se rió. El Dr. Miller pasó toda una vida investigando el uso del suelo urbano y el transporte, y anteriormente obtuvo un doctorado del MIT. "La ingeniería es lo que hacen los ingenieros", respondió. Durante los siguientes 17 años, me di cuenta de la sabiduría en su declaración. Primero se convierte en ingeniero al obtener un título y luego registrarse en el organismo profesional local que regula la profesión de ingeniería. Ahora eres ingeniero. Puedes cavar túneles; escribir códigos de software; componentes de diseño de un iPhone o un jet supersónico. Eres un ingeniero. Y cuando lidera la respuesta global a la crisis financiera en su papel de economista jefe del Fondo Monetario Internacional (FMI), como lo hizo el Dr. Raghuram Rajan, usted es ingeniero.

El profesor Raghuram Rajan realizó su primer título en ingeniería eléctrica del Instituto Indio de Tecnología. Se dedicó a la economía en estudios de posgrado, más tarde se convirtió en profesor en una prestigiosa universidad y finalmente aterrizó en el FMI. Actualmente se desempeña como el 23er Gobernador del Banco de la Reserva de la India. ¿Podría alguien argumentar que su destreza intelectual se basa solo en su formación como economista y que los fundamentos que aprendió como estudiante de ingeniería no desempeñaron ningún papel en el desarrollo de sus habilidades para resolver problemas?

El profesor Rajan es ingeniero. También lo son Xi Jinping, presidente de la República Popular de China, y primer ministro griego que está obligando al mundo a repensar los fundamentos de la economía global. Es posible que no estén diseñando nuevos circuitos, equipos de destilación o puentes, pero están ayudando a construir mejores sociedades y economías y no puede haber una mejor definición de ingeniería e ingenieros, es decir, individuos dedicados a construir mejores economías y sociedades.

Brevemente, diría que la ciencia de datos es lo que hacen los científicos de datos.

Otros tienen definiciones muy diferentes. En septiembre de 2015, un co-panelista en una reunión organizada por BigDataUniversity.com en Toronto limitó la ciencia de datos al aprendizaje automático. Ahí tiene. Si no está utilizando las cajas negras que componen el aprendizaje automático, según algunos expertos en el campo, no es un científico de datos. Incluso si descubriera la cura de una enfermedad que amenaza la vida de millones, los colegas que protegen su territorio lo excluirán del club de ciencia de datos.

El Dr. Vincent Granville (2014), autor de ciencia de datos, ofrece ciertos umbrales que cumplir para ser un científico de datos. En las páginas 8 y 9 en Developing Analytic Talent, el Dr. Granville describe al nuevo profesor de ciencias de datos como un instructor no titular en una universidad no tradicional, que publica resultados de investigación en blogs en línea, no pierde el tiempo escribiendo subvenciones, trabaja desde casa, y gana más dinero que los profesores tradicionales. Baste decir que la próspera comunidad académica de científicos de datos podría estar en desacuerdo con el Dr. Granville.

El Dr. Granville utiliza restricciones en el tamaño y los métodos de datos para definir qué es la ciencia de datos. Define a un científico de datos como alguien que puede "procesar fácilmente un conjunto de datos de 50 millones de filas en un par de horas: y que desconfía de los modelos (estadísticos). Distingue la ciencia de datos de las estadísticas. Sin embargo, enumera álgebra, cálculo y capacitación. en probabilidad y estadística como antecedentes necesarios para comprender la ciencia de datos" (página 4).

Algunos creen que big data se trata simplemente de cruzar un cierto umbral en el tamaño de los datos o el número de observaciones, o se trata del uso de una herramienta en particular, como Hadoop. Dichos umbrales arbitrarios en el tamaño de los datos son problemáticos porque con la innovación, incluso las computadoras normales y el software comercial han comenzado a manipular conjuntos de datos muy grandes. Stata, un software comúnmente utilizado por científicos de datos y estadísticos, anunció que ahora se pueden procesar entre 2 y 24,4 mil millones de filas utilizando sus soluciones de escritorio. Si Hadoop es la contraseña para el club de big data, la capacidad de Stata para procesar 24.4 mil millones de filas, bajo ciertas limitaciones, acaba de colarse en esa fiesta de big data.

Es importante darse cuenta de que alguien que intenta establecer umbrales arbitrarios para excluir a otros es probable que encuentre inconsistencias. El objetivo debería ser definir la ciencia de los datos en un contexto más exclusivo, independiente de la disciplina y la plataforma, sin tamaño, donde la resolución de problemas centrada en los datos y la capacidad de tejer narrativas fuertes sean el centro de atención.

Dada la controversia, preferiría consultar a otros para ver cómo describen a un científico de datos. ¿Por qué no volvemos a consultar al Científico en Jefe de Datos de los Estados Unidos? Recordemos que el Dr. Patil le dijo al periódico Guardian en 2012 que un "científico de datos es esa combinación única de habilidades que pueden desbloquear los conocimientos de los datos y contar una historia fantástica a través de los datos". Lo que es admirable de la definición del Dr. Patil es que incluye a personas de diversos antecedentes académicos y capacitación, y no restringe la definición de un científico de datos a una herramienta en particular ni la somete a un cierto umbral mínimo arbitrario de tamaño de datos.

El otro ingrediente clave para un científico de datos exitoso es un rasgo de comportamiento: la curiosidad. Un científico de datos tiene que ser uno con una mente muy curiosa, dispuesta a dedicar mucho tiempo y esfuerzo para explorar sus corazonadas. En periodismo, los editores lo llaman tener nariz para las noticias. No todos los periodistas saben dónde están las noticias. Solo aquellos que tienen la nariz para las noticias entienden la historia. La curiosidad es igualmente importante para los científicos de datos como lo es para los periodistas.

Rachel Schutt es la Científica de Datos en Jefe de News Corp., imparte un curso de ciencias de datos en la Universidad de Columbia. También es autora de un excelente libro, Doing Data Science. En una entrevista con el New York Times, la Dra. Schutt definió a un científico de datos como alguien que es en parte científico de la computación, en parte ingeniero de software y en parte estadístico (Miller, 2013). Pero esa es la definición de un científico de datos promedio. "Los mejores", sostuvo, "tienden a ser personas realmente curiosas, pensadores que hacen buenas preguntas y están bien lidiando con situaciones no estructuradas y tratando de encontrar estructura en ellas".

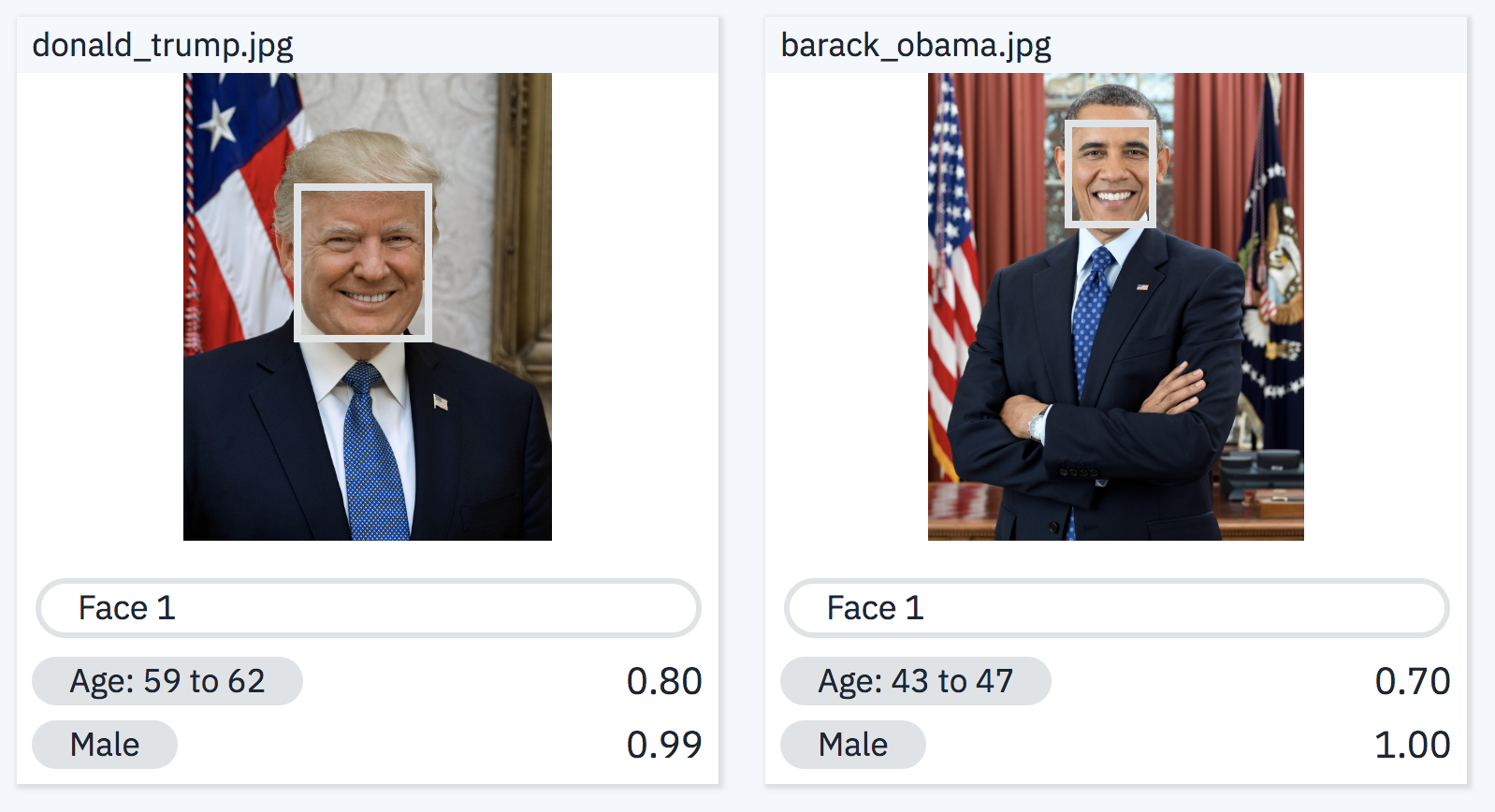
**Ejercicio: Identifica los objetos en imágenes con IBM Watson**

**ACERCA DE ESTE EJERCICIO:**

En esta lección, nos gustaría llevarlo a un viaje secundario. Es un ejercicio divertido y esperamos que lo disfrute tanto como nosotros disfrutamos armando este breve ejercicio para usted. Este ejercicio no está calificado, ¡así que siéntete libre de divertirte!

**CLASIFICACIÓN DE IMAGEN CON IBM WATSON VISUAL RECOGNITION:**

¿Qué mejor manera de comprender las aplicaciones de la ciencia de datos que probarlo usted mismo? Subirá imágenes y verá cómo IBM Watson identifica los diversos objetos y caras (¡incluso el sexo y la edad!) En sus imágenes.



Para aprender a clasificar sus propias imágenes, ¡continúe con la siguiente lectura!

# Ejercicio: ¡Subir y clasificar tus imágenes!

En este ejercicio práctico utilizarás el IBM Watson Visual Recognition (VR) para subir y clasificar tus imágenes. Watson VR es un servicio que utiliza algoritmos de aprendizaje profundo para identificar objetos, rostros y otros contenidos en una imagen. Sigue los pasos que se indican a continuación para completar este ejercicio.

NOTA: Para completar este ejercicio crearás una cuenta en IBM Cloud y te aprovisionará de una instancia para el servicio Watson Visual Recognition. NO se requiere una tarjeta de crédito para registrarse en la cuenta IBM Cloud Lite y no hay ningún cargo asociado a la creación de una instancia de plan Lite del servicio Watson VR.

**Regístrate a IBM Watson Visual Recognition en IBM Cloud**

En este curso, utilizaremos Watson Assistant y Watson Discovery, junto con otros servicios como IBM Cloud Functions, Speech to Text y Text to Speech. Creará instancias de todos estos servicios a lo largo del curso, pero para comenzar, al menos debe tener una instancia de Watson Assistant y Watson Discovery ejecutándose.

### **Crear una instancia de Watson Assistant**

Si ya tiene un servicio Watson Assistant ejecutándose en su cuenta de IBM Cloud, ya está todo listo. Si no lo tiene, siga estos pasos para crear uno. (Puede verificar qué servicios ya creó desde su IBM Cloud Dashboard).

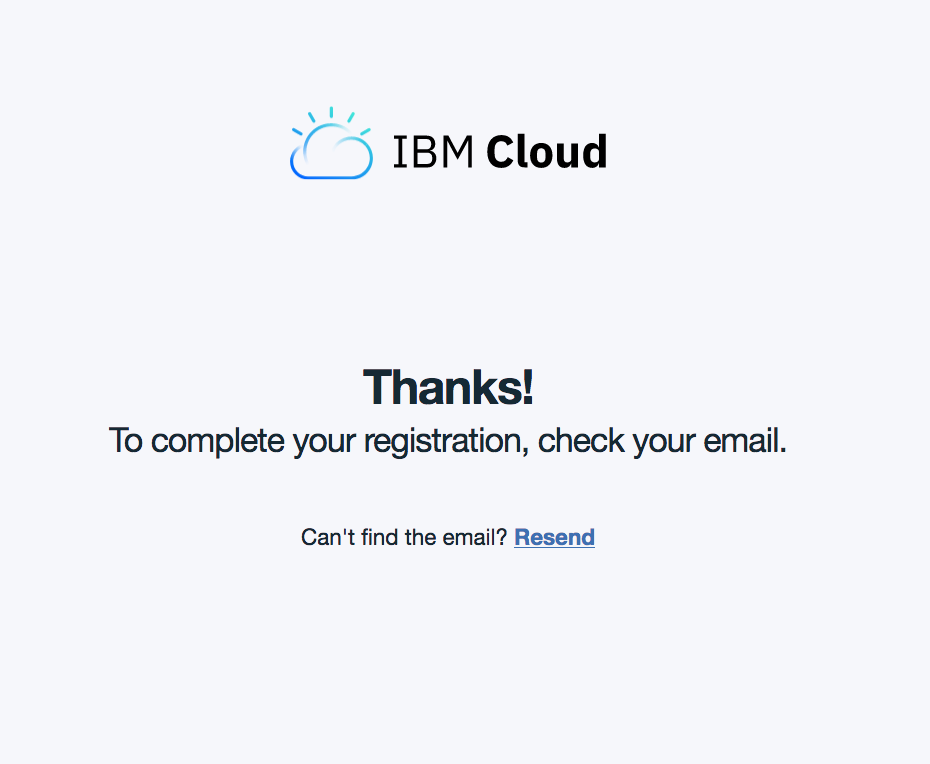
En este curso, utilizaremos el servicio Watson Assistant alojado en la plataforma IBM Cloud. Como resultado, antes de avanzar, necesitaremos asegurarnos de que esté configurado con una cuenta. Siga las instrucciones a continuación para obtener su propia instancia de Watson Assistant y comenzar a construir chatbots.

**Requisito previo:**

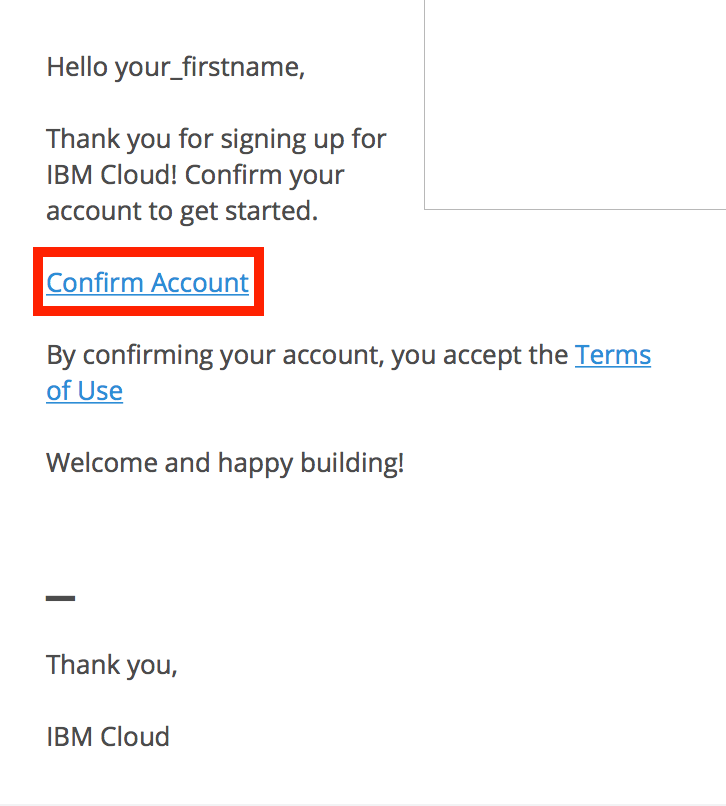
Debe tener una cuenta de IBM Cloud para realizar este laboratorio. Si no tiene una cuenta de IBM Cloud, haga clic y abra este enlace y siga las instrucciones para crear una cuenta de IBM Cloud.

**Paso 1: Confirma Tu Dirección de Correo Electrónico**

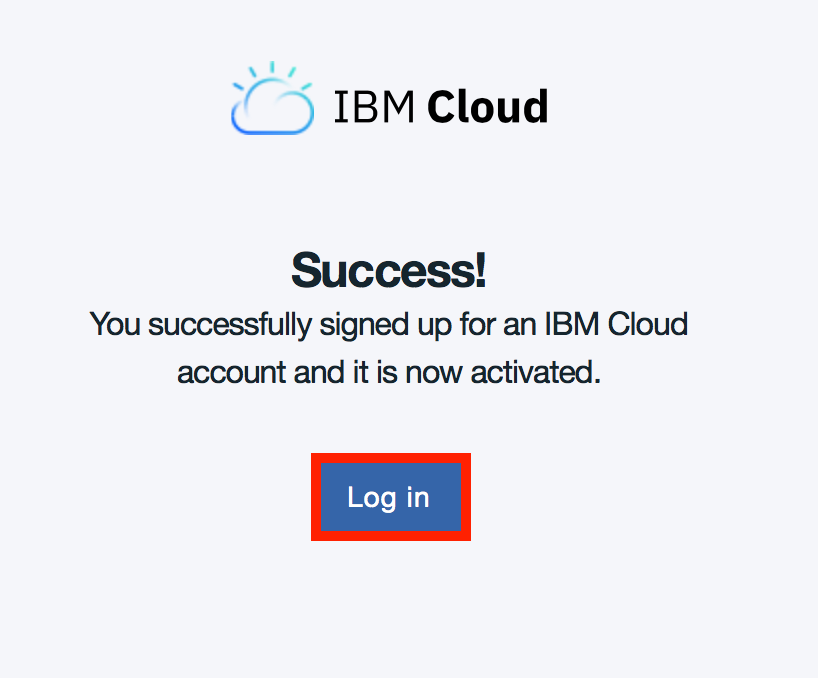
Se envía un correo electrónico a tu dirección de correo electrónico para confirmar tu cuenta.



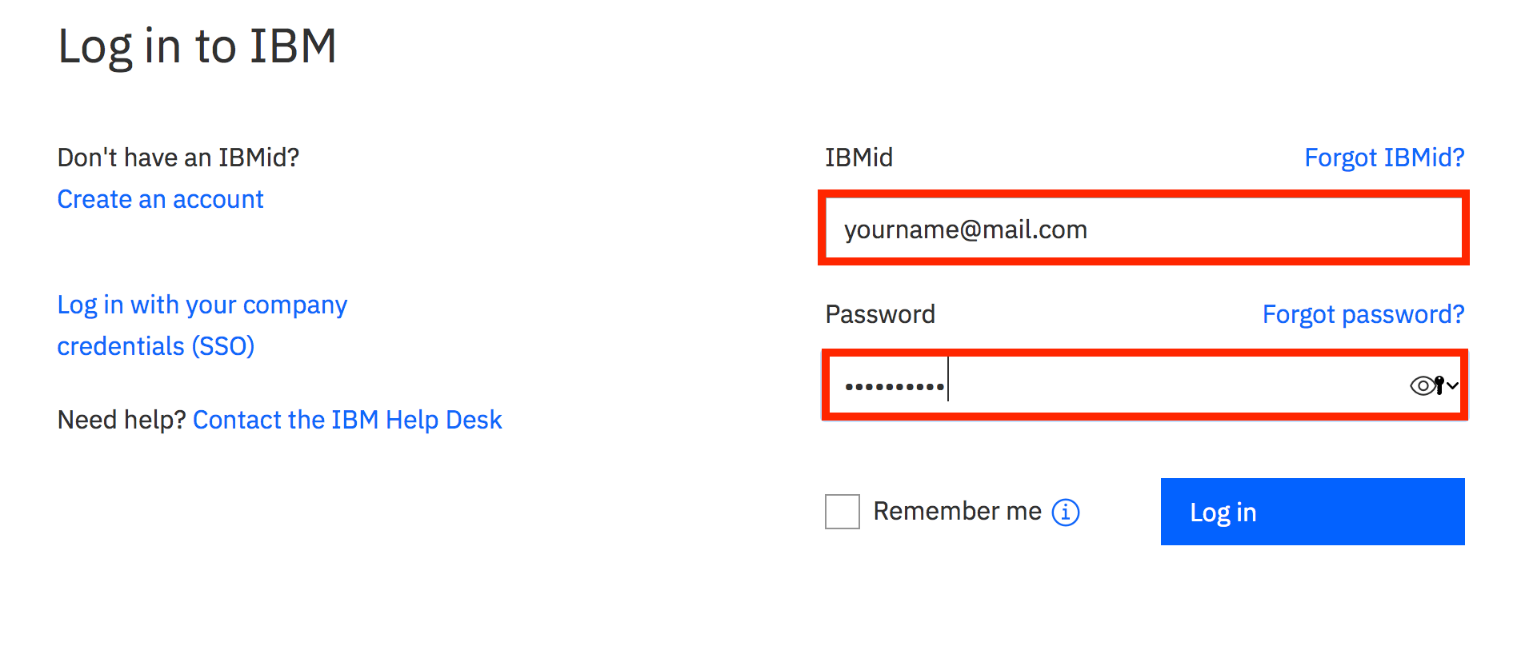
Ve a tu cuenta de correo electrónico y haz clic en el enlace "**Confirm Account**" del correo electrónico que se te envió.



**Paso 2: Ingresa a Tu Cuenta**

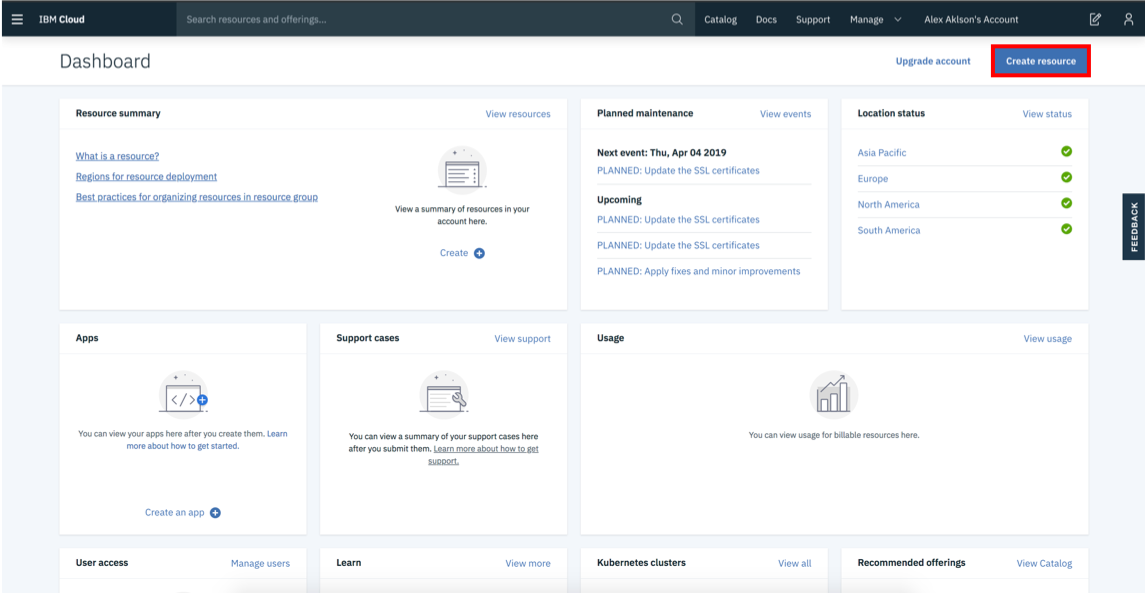


Al hacer clic en **Log in**, se te redirigirá a una página para iniciar sesión en tu cuenta IBM Cloud.



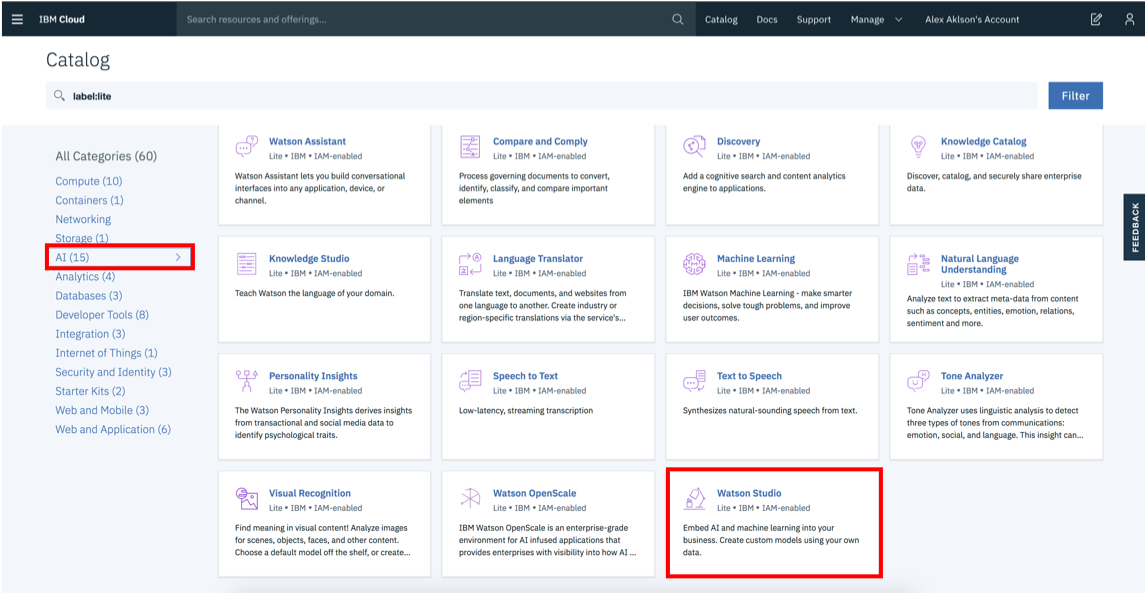
**Paso 3: Crear un Nuevo Recurso**

En la página Dashboard (tablero), haz clic en el botón **Create a resource**en la parte superiorderecha para crear una nueva fuente.

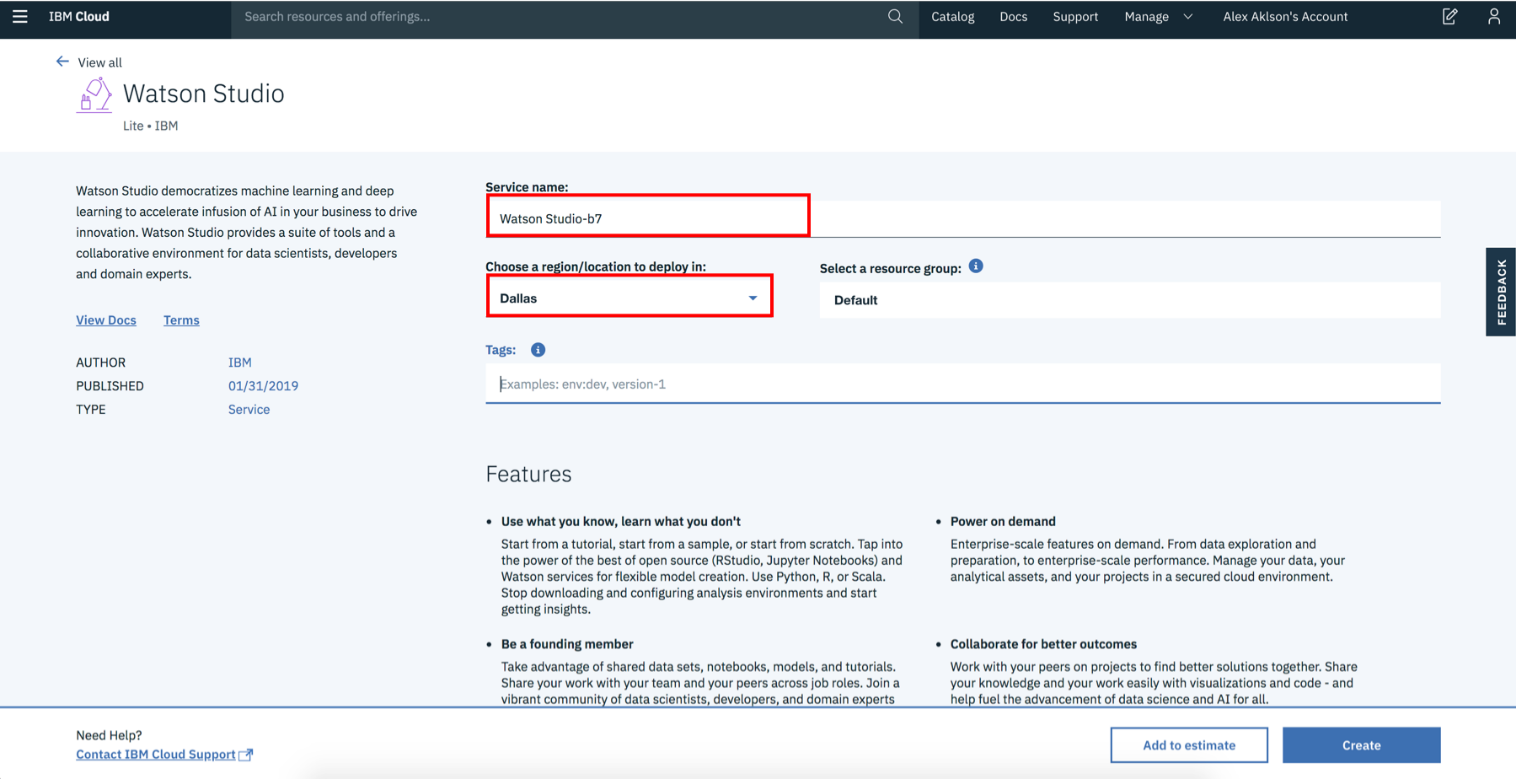


**Paso 4: Crear un Recurso en Watson Studio**

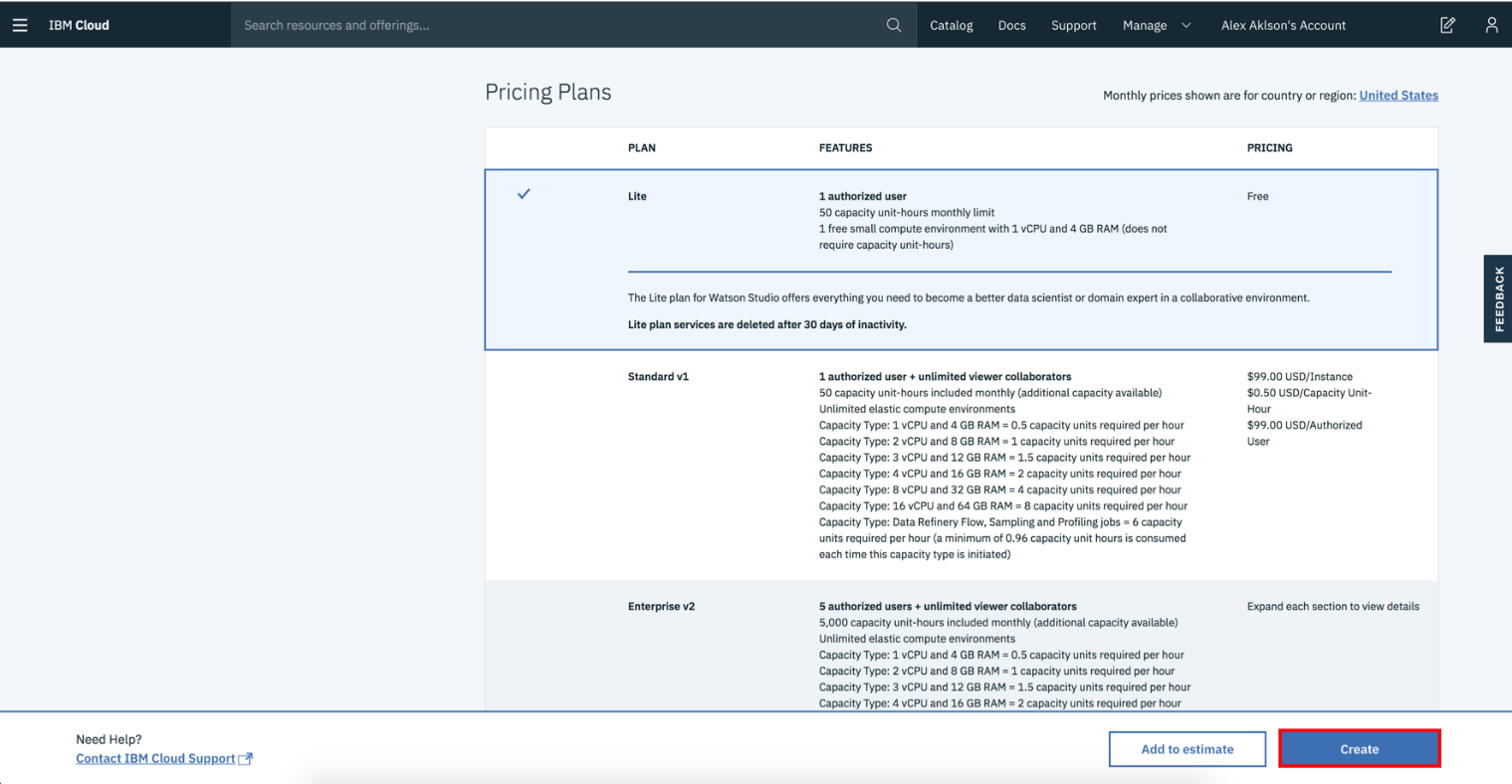
En la página **Catalog** (catálogo), selecciona la categoría de la **AI**en el panel izquierdo y luego selecciona el recurso de **Watson Studio**.



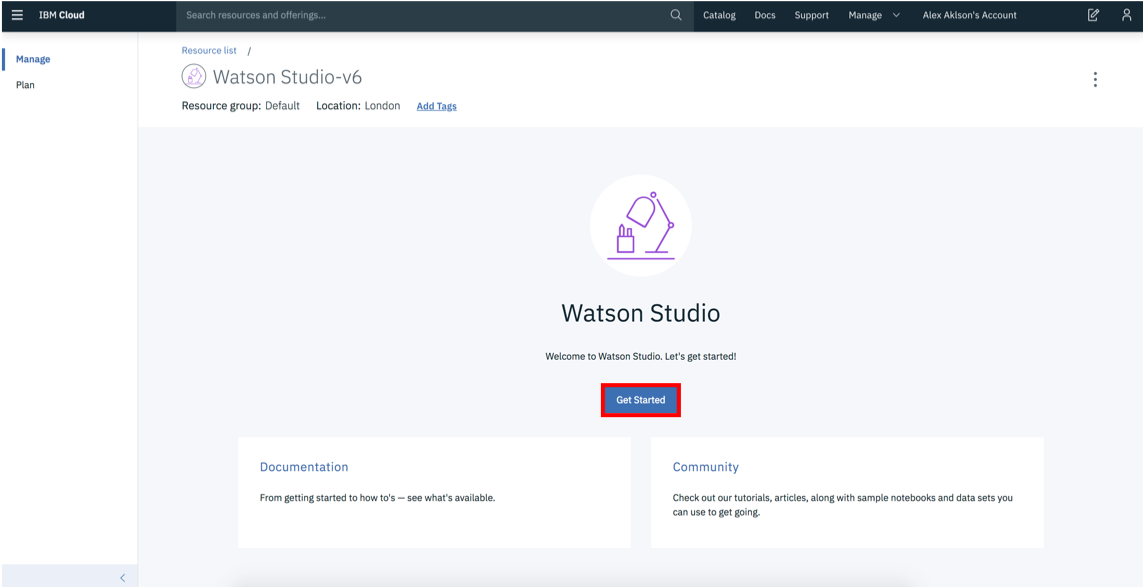
En la siguiente página, podrás nombrar tu instancia de servicio y elegir tu región. Haz clic en la flecha para ver el menú desplegable de regiones. Asegúrate de seleccionar la región más cercana a ti. Como me encuentro en Canadá, elijo **Dallas** como mi región, ya que es la más cercana a mí.



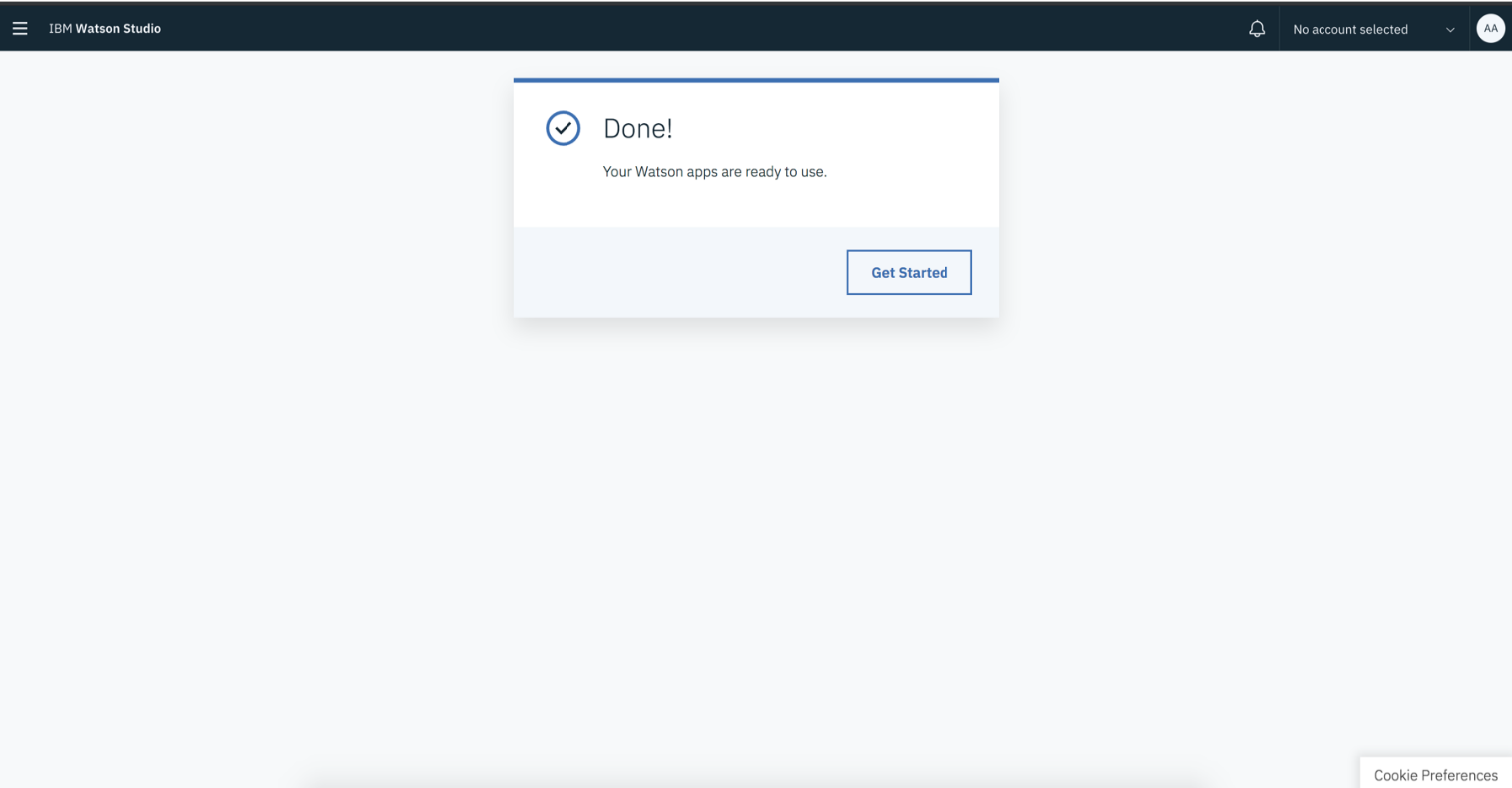
Luego desplázate hacia abajo y asegúrate de que el plan **lite** este seleccionado y haz clic en el botón **Create**.



En la siguiente página, haz clic en el botón **Get Started**para empezar a usar Watson Studio.



Esto comenzará a aprovisionar tu instancia de Watson Studio.



Una vez que el proceso de aprovisionamiento se haya completado, haz clic en el botón **Get Started**para comenzar a utilizarWatson Studio.

**Paso 5: Crear un Proyecto**

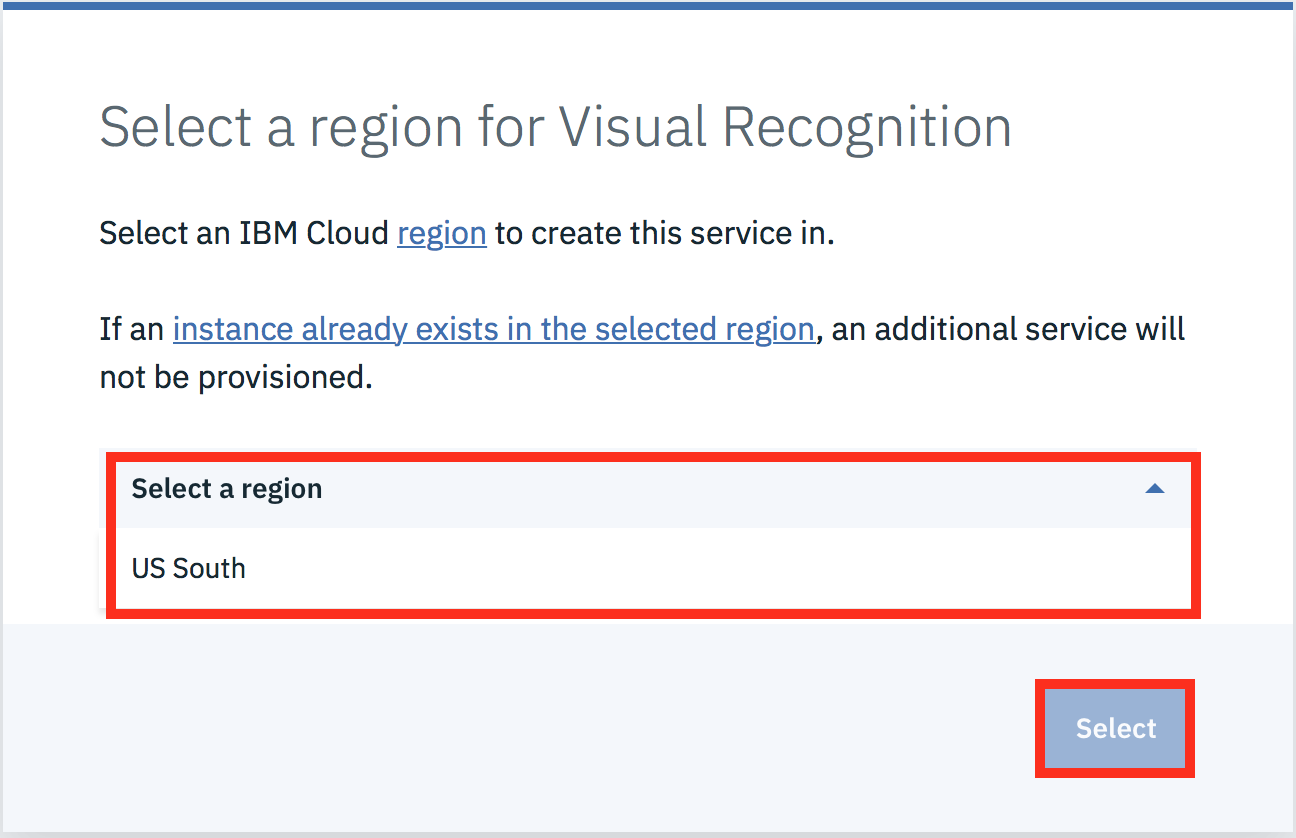
Una vez que llegues a la página principal de IBM Watson Studio, empieza por crear un proyecto.



Crea un **Visual Recognition project**.



Mientras más cerca esté la región de tu ubicación real, más rápido se podrán clasificar las imágenes. Si no estás seguro de cuál elegir, adelántate y selecciona **US South**.



**Paso 6: Establecer tu Proyecto**

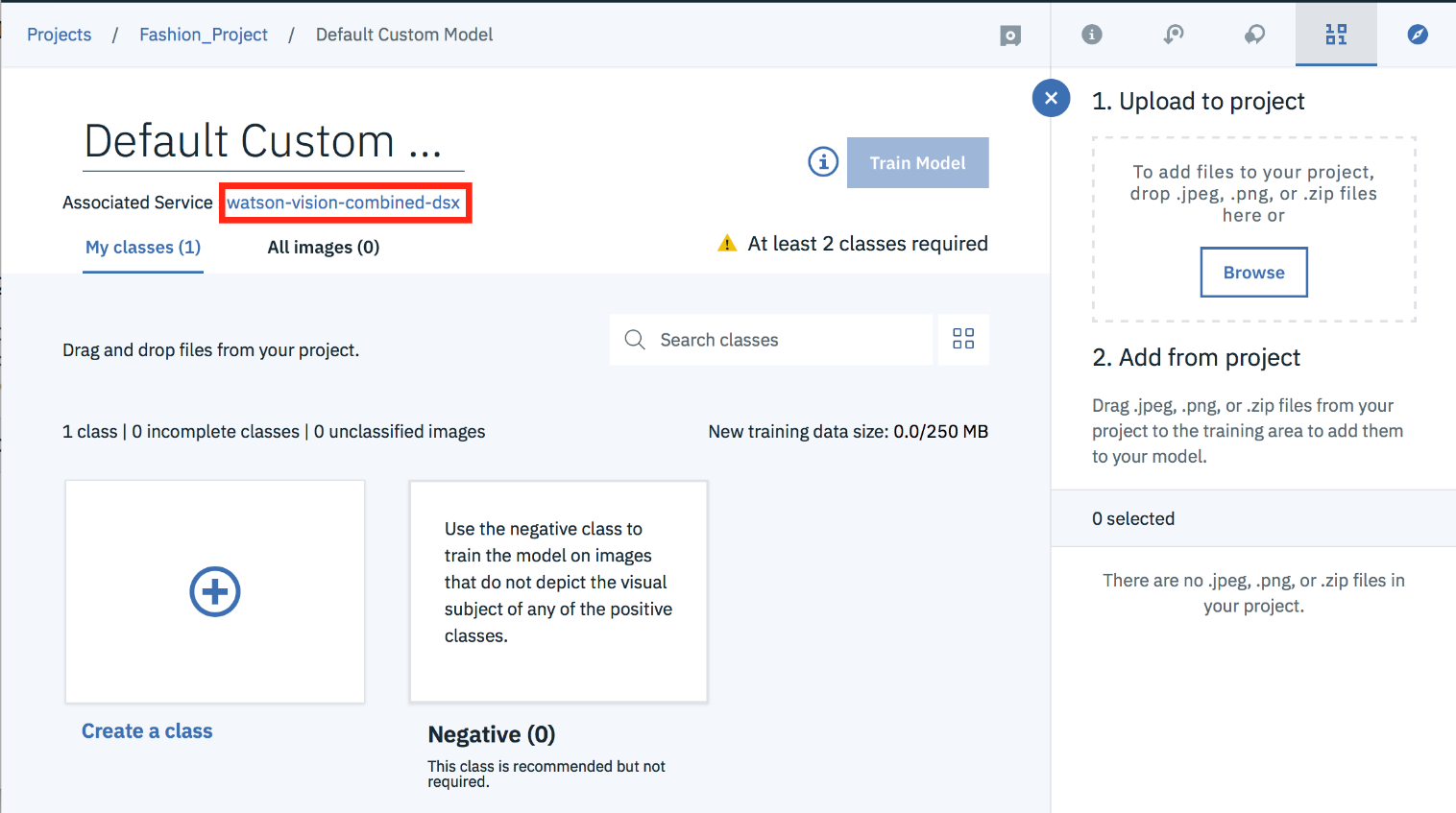
Ahora vamos a rellenar algunos detalles del proyecto y luego puedes hacer clic en **Create**. El IBM Cloud Object Storage, el cual proporciona almacenamiento para tus imágenes y debería ser creado automáticamente para ti.



**Paso 7: Selección de los Modelos Incorporados para Watson Visual Recognition**

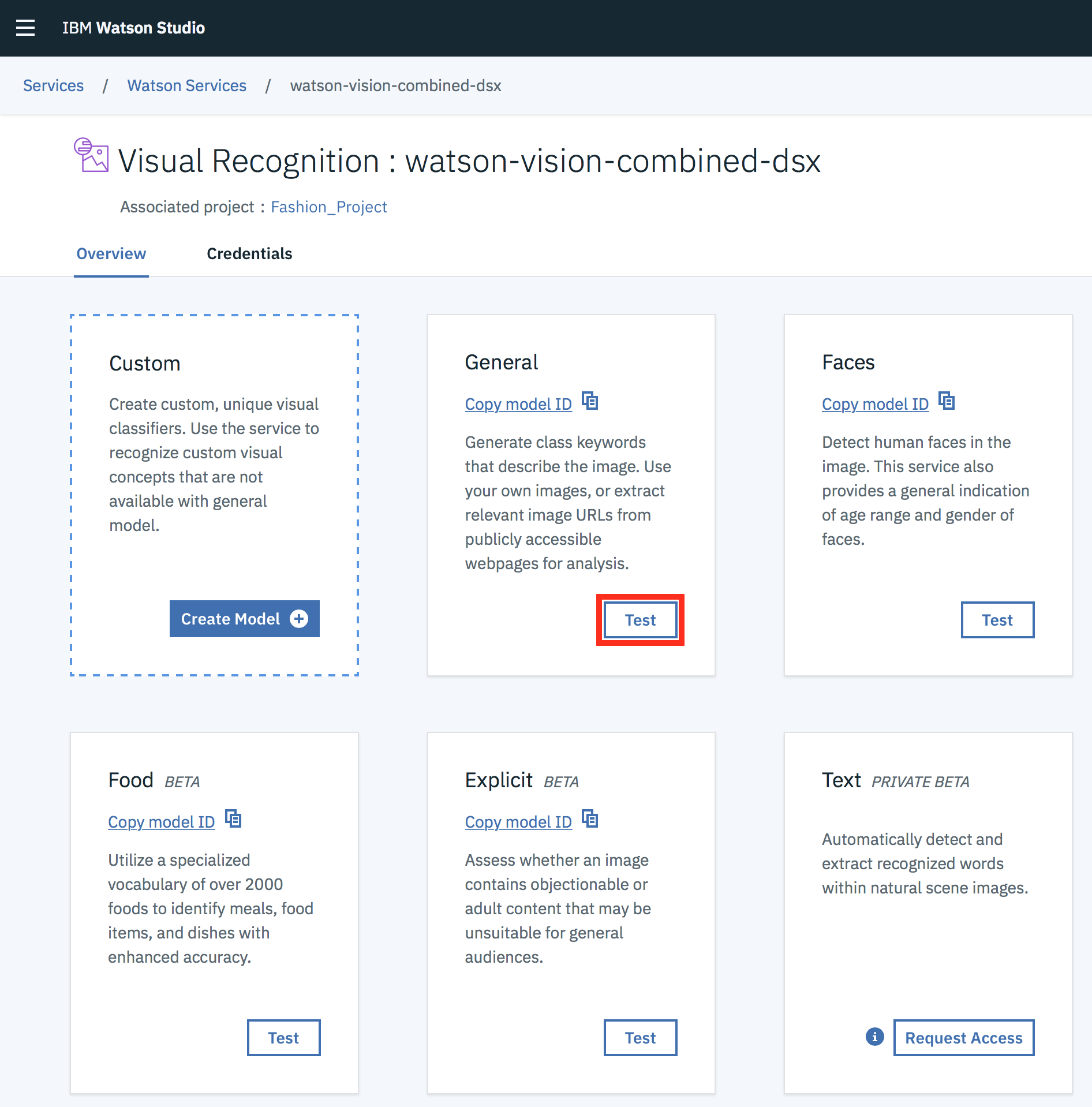
Después de crear tu proyecto, por defecto, pasarás a la página en la que puedes crear tus propios modelos personalizados para clasificar los objetos -- pero saltémonos este paso. ¡En vez de eso, usemos algunos modelos de clasificación de imágenes ya existentes que Watson Visual Recognition tiene por defecto!

Para acceder a los modelos incorporados, haz clic en el nombre del servicio, como se observa en el cuadro rojo de abajo:



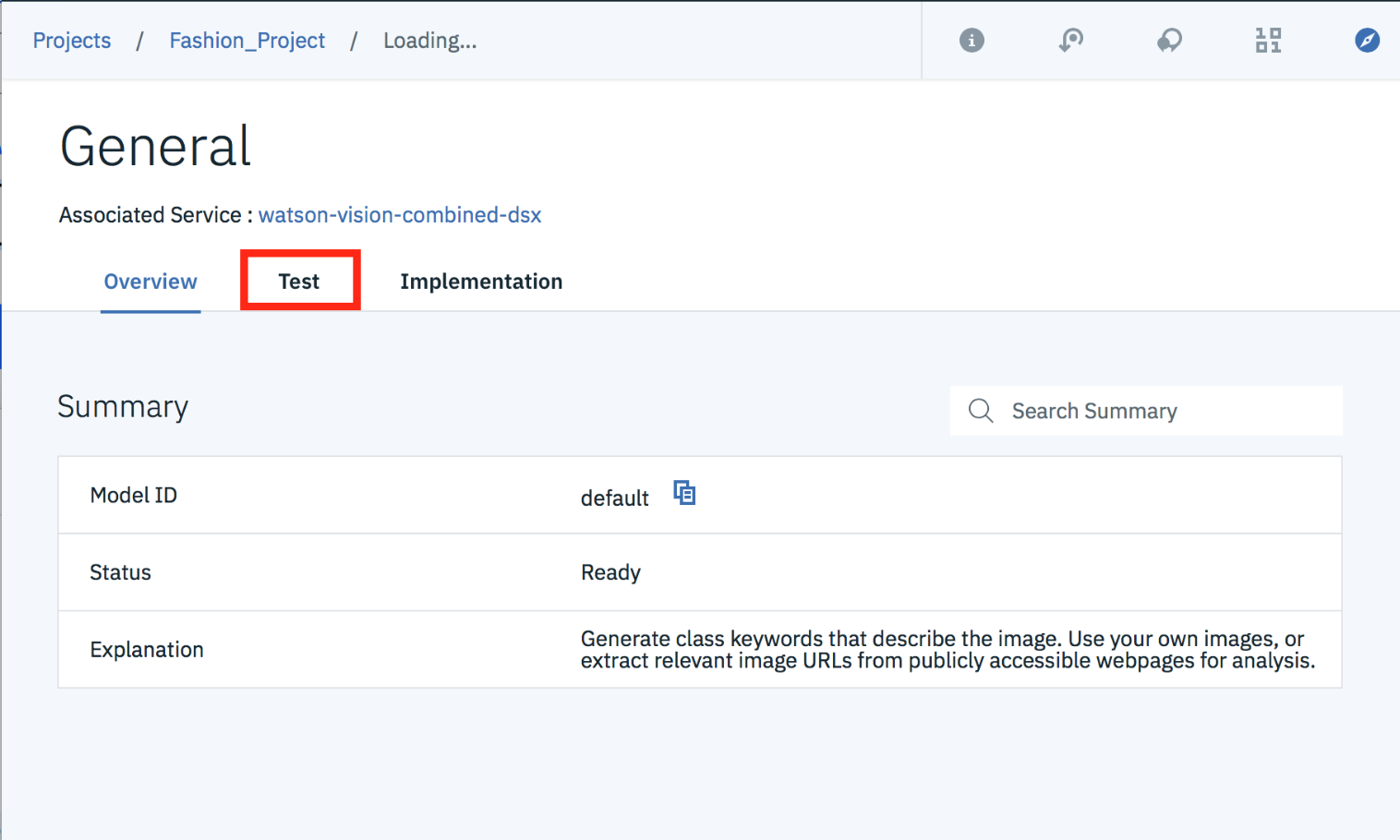
**Paso 8: Elegir el Modelo General.**

¡Ahora puedes ver todos los modelos de clasificación de imágenes incorporados que IBM Watson proporciona! Probemos el modelo **General**.



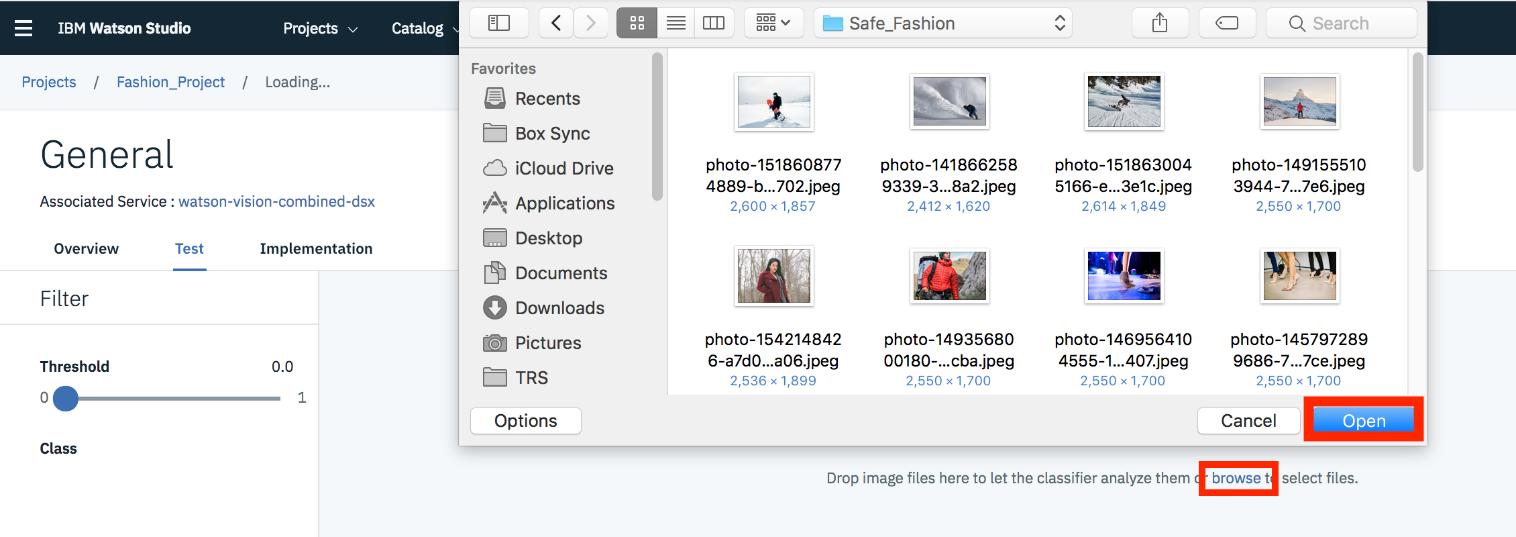
**Paso 9: Prueba el modelo General**

Para probar el modelo General, haz clic en **Test**.



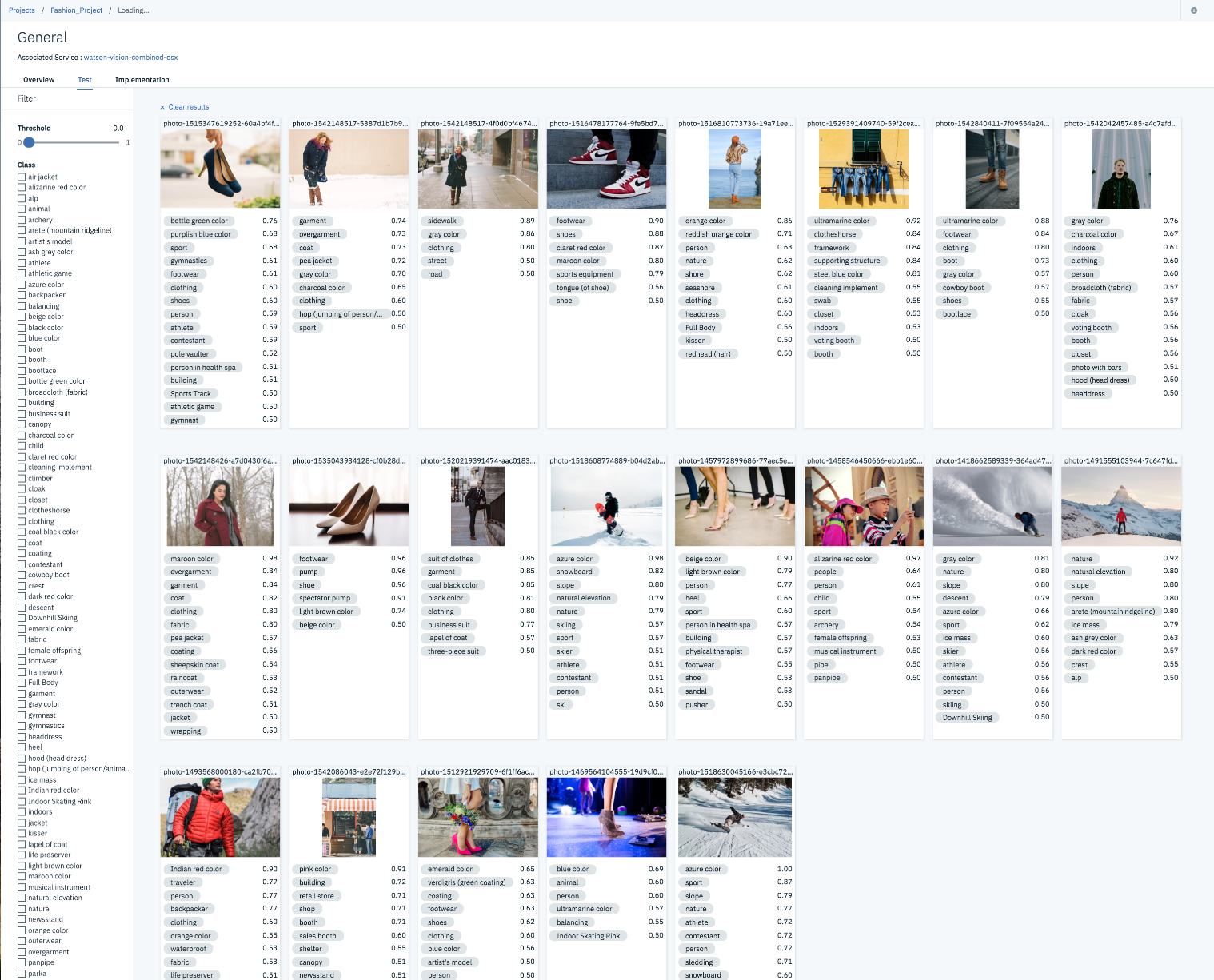
**Paso 10: ¡Subir Tus Imágenes!**

Ahora puedes subir las imágenes que desees haciendo clic en **Browse.**

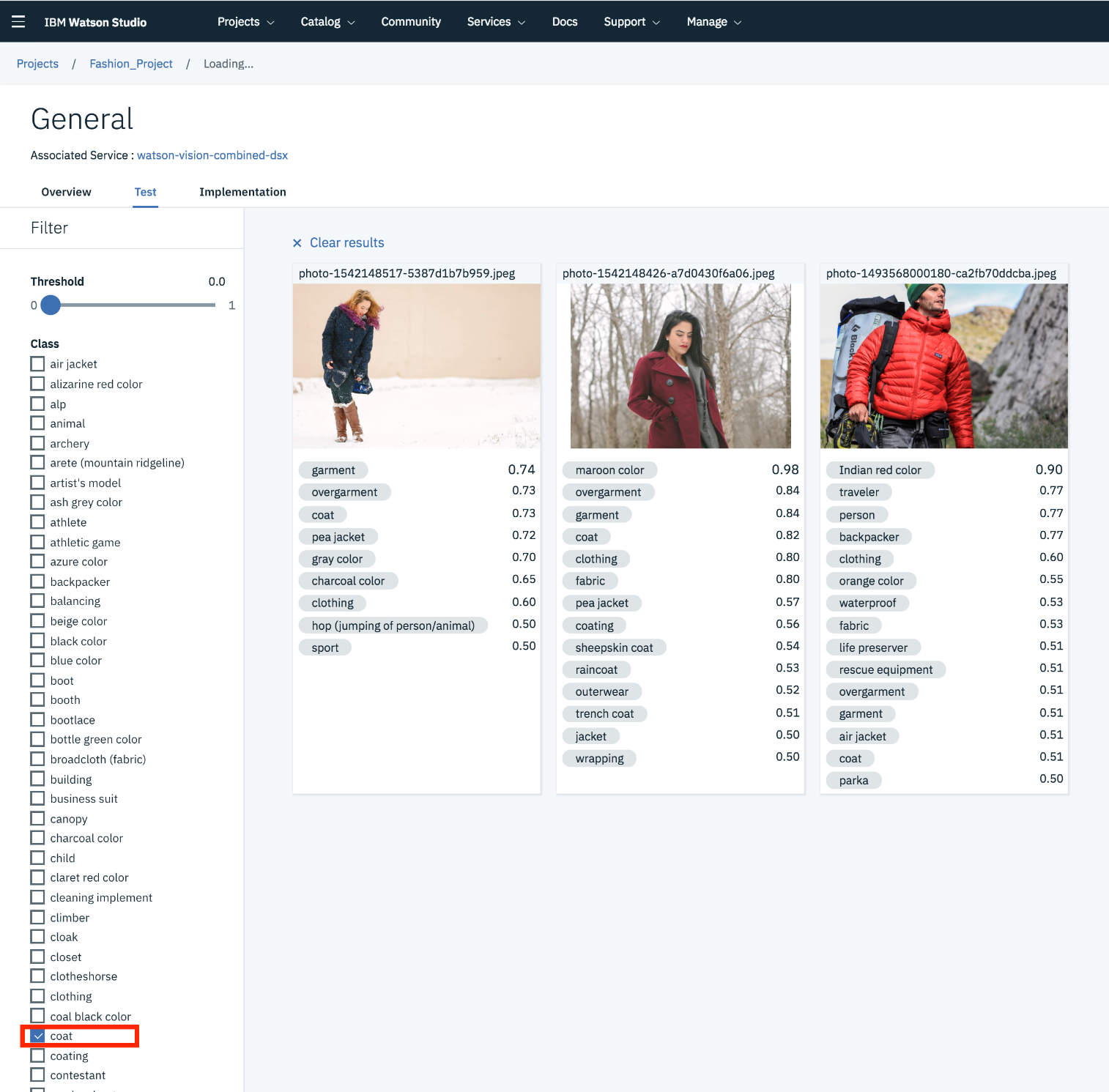


**Paso 11: ¡Comprueba los Resultados!**

Una vez que hayas subido tus imágenes, ¡el Watson Studio Visual Recognition te dirá lo que cree que encontró en tus imágenes! Junto a cada clase de objeto (o color, edad, entre otras características), también te dará una puntuación de seguridad (entre 0 y 1) sobre la seguridad que cree haber encontrado ese objeto en particular en tu imagen (0 para la seguridad más baja y 1 para la más alta).



**Paso 12: Filtrar Tus Resultados**



**Paso 13: ¡Comparte tus Resultados con Nosotros en Twitter!**

¡Síguenos en Twitter y envíanos algunos de los resultados más divertidos e interesantes que hayas encontrado con IBM Watson Visual Recognition!

Sigue a Polong Lin

Sigue a Alex Aklson