# Data requirements.

¡Bienvenido a Data Science Methodology 101, desde las necesidades hasta los requisitos de recopilación de datos! Si su objetivo es hacer un plato de espaguetis, pero no tiene los ingredientes adecuados para hacerlo, su éxito se verá comprometido. Piense en esta parte de la metodología de la ciencia de datos como cocinar con datos. Cada paso en la cocina es muy importante. Entonces, si el problema a resolver es, por así decirlo, y los datos están contenidos, entonces el científico de datos debe determinar: qué componentes se requieren, cómo obtenerlos o recopilarlos, cómo entenderlos o trabajar con ellos y cómo preparar los datos para lograr el resultado deseado. Sobre la base de una comprensión del problema en cuestión y luego utilizando el enfoque analítico elegido, el científico de datos está listo para comenzar. Veamos algunos ejemplos de requisitos de datos en la metodología de ciencia de datos. Antes de emprender las fases de recopilación y preparación de datos de la metodología, es vital definir los requisitos de datos para la clasificación del árbol de decisiones. Esto incluye determinar el contenido de los datos, los formatos y las fuentes requeridas para la recopilación inicial de datos. Ahora veamos el caso de estudio de la implementación de "Requisitos de datos". La primera tarea en el estudio de caso fue definir los requisitos de datos para el enfoque de clasificación del árbol de decisión elegido. Esto implicó seleccionar un grupo de pacientes apropiado de la base de miembros del proveedor de seguros de salud. Se determinaron tres criterios de inclusión en la cohorte para compilar historias clínicas completas. Primero, un paciente tenía que ser considerado un paciente hospitalizado dentro del área de servicio del proveedor para que pudiera acceder a la información necesaria. En segundo lugar, se centraron en pacientes diagnosticados con insuficiencia cardíaca congestiva primaria durante un año completo. En tercer lugar, un paciente debe tener un registro continuo durante al menos seis meses antes de la presentación inicial de insuficiencia cardíaca congestiva para que se pueda compilar un historial médico completo. Los pacientes con insuficiencia cardíaca congestiva que también fueron diagnosticados con otras condiciones médicas significativas fueron excluidos de la cohorte ya que estas condiciones darían como resultado tasas de reingreso más altas que el promedio y, por lo tanto, sesgarían los resultados. Luego, se definen el contenido, el formato y las representaciones de los datos necesarios para la clasificación del árbol de decisión. Esta técnica de modelado requiere un registro por paciente, con columnas que representan las variables del modelo. Para modelar el resultado de la readmisión, se requirieron datos que cubrieran todos los aspectos de la historia clínica del paciente. Este contenido incluirá admisiones, diagnósticos primarios, secundarios y terciarios, procedimientos, recetas y otros servicios prestados durante la hospitalización o visitas del paciente/médico. Así, un paciente dado puede tener miles de registros que representan todas sus características relevantes. Para llegar a un formato de registro por paciente, los científicos de datos agregaron los registros de transacciones al nivel del paciente y crearon un nuevo conjunto de variables para representar esta información. Este fue un trabajo para la fase de preparación de datos, por lo que es importante pensar en el futuro y anticipar las próximas fases. Finalizó la sección Requisitos de datos para este curso. Gracias por ver. (música)

# Data collection.

Metodología de ciencia de datos 101 ¡Bienvenido a Recopilación de datos a partir de requisitos! Una vez que se realiza la recopilación de datos inicial, el científico de datos realiza una evaluación para determinar si tienen lo que necesitan. Al igual que con la compra de ingredientes para cocinar, algunos ingredientes pueden estar fuera de temporada y ser más difíciles de obtener o más caros de lo que se pensó originalmente. En esta etapa, se revisan los requisitos de datos y se decide si la recopilación requiere más o menos datos. Después de recopilar los componentes de datos, durante la fase de recopilación de datos, el científico de datos comprenderá bien con qué trabajará. Se pueden aplicar técnicas como estadísticas descriptivas y visualización al conjunto de datos para evaluar el contenido, la calidad y los conocimientos iniciales sobre los datos. Se identificarán las lagunas en los datos y será necesario hacer un plan para llenar o sustituir. En esencia, los ingredientes ahora están sentados en la tabla de cortar. Veamos ahora algunos ejemplos de la fase de recopilación de datos en la metodología de ciencia de datos. Esta fase se realiza como seguimiento de la fase de requisitos de datos. Ahora, veamos el caso de estudio de la práctica de "Recopilación de datos". La recopilación de datos requiere conocer la fuente o saber dónde encontrar los elementos de datos necesarios. En el contexto de nuestro estudio de caso, estos pueden incluir: información demográfica, clínica y de cobertura de los pacientes, registros de solicitud de información del proveedor, así como información farmacéutica y de otro tipo relacionada con todos los diagnósticos de pacientes con insuficiencia cardíaca congestiva. También se necesitaba cierta información sobre medicamentos para este estudio de caso, pero esta fuente de datos aún no estaba integrada con otras fuentes de datos. Esto nos lleva a un punto importante: está bien retrasar las decisiones sobre datos desfavorables e intentar obtenerlos en una etapa posterior. Por ejemplo, esto se puede hacer incluso después de que se hayan obtenido algunos resultados intermedios del modelado predictivo. Si estos resultados muestran que la información sobre medicamentos puede ser importante para obtener un buen modelo, entonces se habrá invertido tiempo tratando de derivarlo. Sin embargo, resultó que pudieron construir un modelo bastante bueno sin este conocimiento de las drogas. Los DBA y los programadores a menudo trabajan juntos para extraer datos de varias fuentes y luego combinarlos. Esto permite la eliminación de datos redundantes, haciéndolos utilizables para la siguiente fase de la metodología, la comprensión de datos. En esta etapa, si es necesario, los científicos de datos y los miembros del equipo de análisis pueden discutir varias formas de administrar mejor sus datos, incluida la automatización de ciertos procesos en la base de datos para que la recopilación de datos sea más fácil y rápida. Gracias por ver. (música).

# Data understanding.

¡Bienvenido a la Metodología de ciencia de datos 101, desde la comprensión hasta la preparación, comprensión de datos! La comprensión de datos abarca todas las actividades relacionadas con la construcción del conjunto de datos. Esencialmente, la sección de comprensión de datos de la metodología de ciencia de datos responde a la pregunta: ¿Son los datos que recopiló representativos del problema a resolver? Apliquemos la etapa de comprensión de datos de nuestra metodología al estudio de caso que hemos estado examinando. Para comprender los datos relacionados con los ingresos por insuficiencia cardíaca congestiva, era necesario ejecutar estadísticas descriptivas en las columnas de datos que se convertirían en variables en el modelo. En primer lugar, estas estadísticas incluían Hearst, univariantes y estadísticas sobre cada variable, como la media, la mediana, el mínimo, el máximo y la desviación estándar. En segundo lugar, se usaron correlaciones por pares para ver qué tan estrechamente estaban relacionadas ciertas variables y cuáles, si las había, estaban altamente correlacionadas, lo que significa que serían esencialmente redundantes, por lo que solo una sería relevante para el modelo. En tercer lugar, se examinaron los histogramas de las variables para comprender sus distribuciones. Los histogramas son una buena manera de comprender cómo se distribuyen los valores o una variable, y qué tipo de preparación de datos puede ser necesaria para que la variable sea más útil en un modelo. Por ejemplo, para una variable categórica que tiene demasiados valores distintos para ser informativa en un modelo, el histograma les ayudaría a decidir cómo consolidar esos valores. Las variables univariadas, las estadísticas y los histogramas también se utilizan para evaluar la calidad de los datos. A partir de la información proporcionada, ciertos valores se pueden volver a codificar o incluso descartar si es necesario, como cuando a una determinada variable le faltan valores. La pregunta entonces es, ¿"perdido" significa algo? A veces, un valor faltante puede significar "no" o "0" (cero), o en otras ocasiones simplemente significa "no sabemos". O bien, si una variable contiene valores no válidos o engañosos, como una variable numérica llamada "edad" que contiene del 0 al 100 y también 999, donde ese "triple-9" en realidad significa "ausente", pero se trataría como un valor válido a menos que lo corrijamos. Inicialmente, el significado de ingreso por insuficiencia cardíaca congestiva se decidió sobre la base de un diagnóstico primario de insuficiencia cardíaca congestiva. Pero trabajar en la etapa de comprensión de los datos reveló que la definición inicial no capturaba todas las admisiones por insuficiencia cardíaca congestiva que se esperaban, según la experiencia clínica. Esto significó volver a la etapa de recopilación de datos y agregar diagnósticos secundarios y terciarios, y construir una definición más completa de ingreso por insuficiencia cardíaca congestiva. Este es solo un ejemplo de los procesos interactivos en la metodología. Cuanto más se trabaja con el problema y los datos, más se aprende y, por lo tanto, más refinamiento se puede hacer dentro del modelo, lo que en última instancia conduce a una mejor solución al problema. Esto finaliza la sección Comprensión de datos de este curso. ¡Gracias por ver! (música)

# Data preparation.

¡Bienvenido a la Metodología de ciencia de datos 101, "Preparación de datos: conceptos" desde la comprensión hasta la preparación! En cierto sentido, la preparación de datos es similar a lavar vegetales recién cosechados, siempre que se eliminen los elementos indeseables como la suciedad o las imperfecciones. Junto con la recopilación y la comprensión de datos, la preparación de datos es la fase que consume más tiempo de un proyecto de ciencia de datos y, por lo general, ocupa hasta el setenta o incluso el noventa por ciento del tiempo total del proyecto. La automatización de algunos de los procesos de recopilación y preparación de datos en la base de datos puede reducir ese tiempo hasta en un 50 %, lo que significa más tiempo para que los científicos de datos se concentren en la creación de modelos. Continuando con nuestra metáfora de la cocina, sabemos que adelgazar la cebolla permitirá que sus sabores se extiendan a través de una salsa más fácilmente que si simplemente echamos la cebolla entera en la olla. De manera similar, la preparación de datos es el proceso de transformar los datos en un estado en el que puede ser más fácil trabajar. Específicamente, la fase de preparación de datos de la metodología responde a la pregunta: ¿Cuáles son las formas en que se prepararon los datos? Para trabajar de manera efectiva con datos, debe estar preparado para manejar valores faltantes o no válidos y eliminar duplicados para garantizar que todo tenga el formato correcto. La ingeniería de características también forma parte de la preparación de datos. Es el proceso de utilizar el conocimiento del dominio de los datos para crear funciones que permitan el funcionamiento de los algoritmos de aprendizaje automático. Una característica es una característica que puede ser útil para resolver un problema. Las características de los datos son importantes para los modelos predictivos y afectarán los resultados que desea lograr. La ingeniería de funciones es fundamental cuando se aplican herramientas de aprendizaje automático para analizar datos. Cuando se trabaja con texto, se requieren pasos de análisis de texto para codificar los datos a fin de poder manipularlos. El científico de datos debe saber qué está buscando en los conjuntos de datos para abordar la pregunta. El análisis de texto es fundamental para garantizar que se establezcan las agrupaciones adecuadas y que la programación no pase por alto lo que se esconde en su interior. La fase de preparación de datos sienta las bases para los próximos pasos para abordar el problema. Esta fase puede llevar algún tiempo, pero si se hace correctamente, los resultados respaldarán el proyecto. Si esto se omite, el resultado será desigual y puede volver a la mesa de dibujo. Es esencial que invierta su tiempo en esta área y utilice las herramientas disponibles para automatizar los pasos comunes para acelerar la preparación de datos. Asegúrese de prestar atención a los detalles en esta área. Después de todo, un mal ingrediente es todo lo que se necesita para arruinar una buena comida. Esto concluye la sección de preparación de datos de este curso donde revisamos los conceptos clave. Gracias por ver. (Música)

# Data preparation: Case study

Bienvenido a Metodología de ciencia de datos 101 Desde la comprensión hasta la preparación de datos - ¡Estudio de caso! En cierto sentido, la preparación de datos es similar a lavar vegetales recién cosechados en la medida en que se eliminan los elementos no deseados, como la suciedad o las imperfecciones. Ahora, veamos el caso de estudio relacionado con la aplicación de los conceptos de preparación de datos. En el estudio de caso, un primer paso importante en la etapa de preparación de datos fue definir realmente la insuficiencia cardíaca congestiva. Esto sonaba fácil al principio, pero definirlo con precisión no fue sencillo. En primer lugar, era necesario identificar el conjunto de códigos de grupo relacionados con el diagnóstico, ya que la insuficiencia cardíaca congestiva implica ciertos tipos de acumulación de líquido. También necesitábamos considerar que la insuficiencia cardíaca congestiva es solo un tipo de insuficiencia cardíaca. Se necesitaba orientación clínica para obtener los códigos correctos para la insuficiencia cardíaca congestiva. El siguiente paso consistió en definir los criterios de readmisión por la misma condición. Era necesario evaluar el momento de los eventos para definir si un ingreso por insuficiencia cardíaca congestiva en particular era un evento inicial, que se denomina ingreso índice, o un reingreso relacionado con insuficiencia cardíaca congestiva. Con base en la experiencia clínica, se estableció un período de tiempo de 30 días como la ventana para la readmisión relevante para los pacientes con insuficiencia cardíaca congestiva, luego del alta de la admisión inicial. A continuación, se agregaron los registros que estaban en formato transaccional, lo que significa que los datos incluían múltiples registros para cada paciente. Los registros de transacciones incluyeron reclamos de instalaciones de proveedores profesionales presentados para servicios médicos, de laboratorio, hospitalarios y clínicos. También se incluyeron registros que describían todos los diagnósticos, procedimientos, recetas y otra información sobre pacientes hospitalizados y ambulatorios. Un paciente determinado fácilmente podría tener cientos o incluso miles de estos registros, dependiendo de su historial clínico. Luego, todos los registros transaccionales se agregaron al nivel del paciente, lo que generó un solo registro para cada paciente, según lo requerido por el método de clasificación del árbol de decisión que se usaría para el modelado. Como parte del proceso de agregación, se crearon muchas columnas nuevas que representan la información de las transacciones. Por ejemplo, frecuencia y visitas más recientes a médicos, clínicas y hospitales con diagnósticos, procedimientos, recetas, etc. También se consideraron las comorbilidades con la insuficiencia cardíaca congestiva, como la diabetes, la hipertensión y muchas otras enfermedades y afecciones crónicas que podrían afectar el riesgo de reingreso por insuficiencia cardíaca congestiva. Durante las discusiones sobre la preparación de los datos, también se llevó a cabo una revisión literaria sobre la insuficiencia cardíaca congestiva para ver si se pasó por alto algún elemento importante de los datos, como las comorbilidades que aún no se habían tenido en cuenta. La revisión literaria implicó volver a la etapa de recopilación de datos para agregar algunos indicadores más para condiciones y procedimientos. Agregar los datos transaccionales a nivel de paciente significaba fusionarlos con los demás datos del paciente, incluida su información demográfica, como edad, sexo, tipo de seguro, etc. El resultado fue la creación de una tabla que contenía un solo registro por paciente, con muchas columnas que representaban los atributos del paciente en su historial clínico. Estas columnas se utilizarían como variables en el modelo predictivo. Aquí hay una lista de las variables que finalmente se usaron en la construcción del modelo. La variable dependiente, u objetivo, era el reingreso por insuficiencia cardíaca congestiva dentro de los 30 días posteriores al alta de una hospitalización por insuficiencia cardíaca congestiva, con un resultado de sí o no. La etapa de preparación de datos resultó en una cohorte de 2343 pacientes que cumplieron con todos los criterios para este estudio de caso. Luego, la cohorte se dividió en conjuntos de entrenamiento y prueba para construir y validar el modelo, respectivamente. Esto finaliza la sección de preparación de datos de este curso, en la que aplicamos los conceptos clave al estudio de caso. ¡Gracias por ver! (Música)