METODOLOGÍA EN DATA SCIENCE

TABLA DE CONTENIDO

1. INTRODUCCIÓN

El objetivo de la ciencia de datos es responder las siguientes 10 preguntas en la secuencia prescrita:

Del problema a abordar:

1. Cuál es el problema que intenta resolver?

2. Cómo puede usar los datos para responder la pregunta?

Trabajando con los datos:

3. Qué datos se necesitan para responder la pregunta?

4. De dónde vienen los datos (identificar todas las fuentes) y cómo los obtendrá?

5. Son los datos coleccionados representativos del problema a ser resuelto?

6. Qué trabajo adicional se necesita para manipular y trabajar los datos?

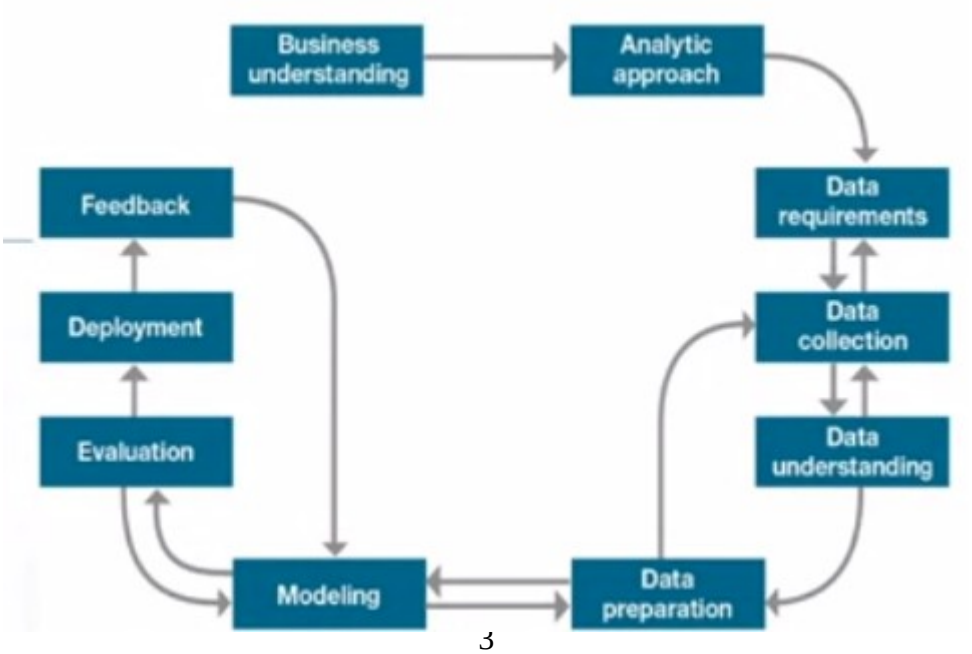
Deduciendo la respuesta:

7. De qué forma los datos pueden ser visualizados para obtener la respuesta requerida?

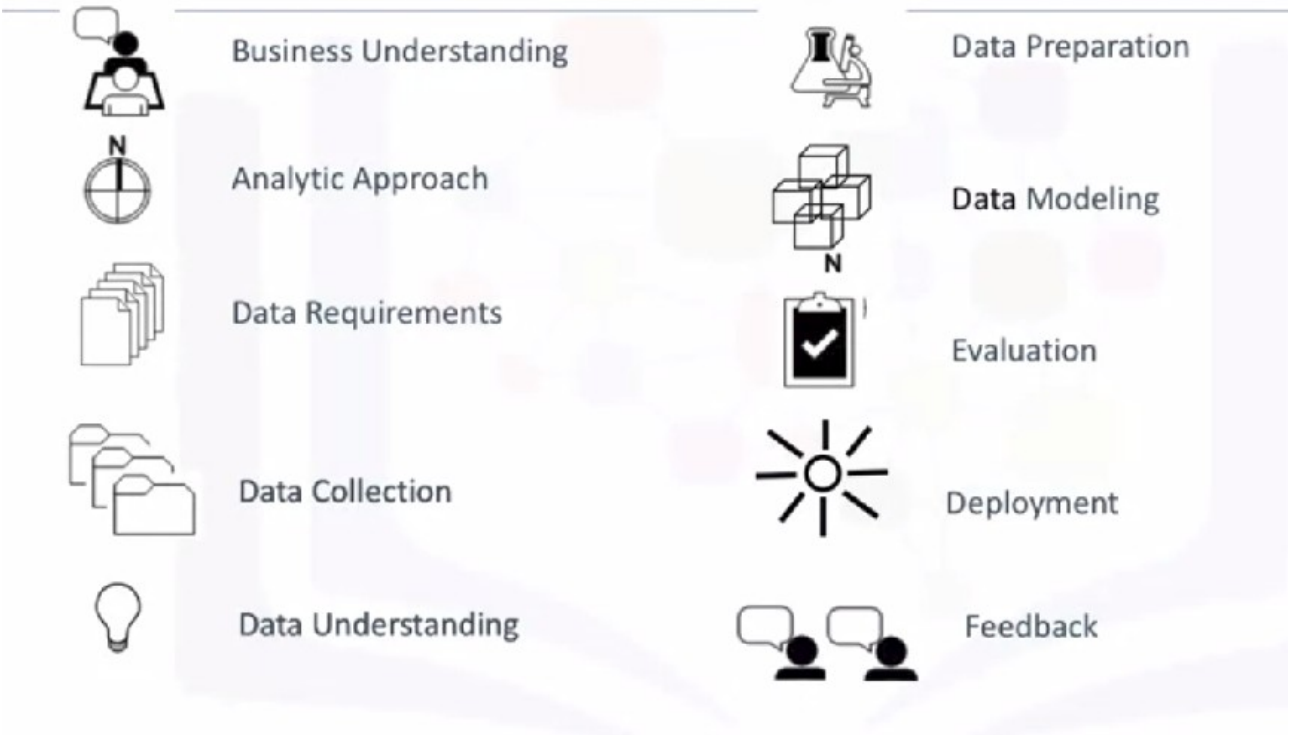
8. El modelo responde realmente la pregunta o necesita ser ajustado?

9. Puede el modelo llevarse a la práctica?

10. Puede obtener retroalimentación constructiva al responder la pregunta?



Nuestro **caso de estudio** es el siguiente:
Hay un presupuesto limitado para proveer servicio de salud público.
Las re-admisiones en el hospital por problemas que vuelven a ocurrir pueden verse como un signo de falla del sistema para tratar apropiadamente la condición del paciente antes del alta inicial del mismo.
La **pregunta clave** es: cuál es la mejor forma de asignar estos recursos para maximizar su uso al proveer una salud de calidad?
Si el programa es exitoso, se le brindarán a los médicos mejores herramientas, basadas en datos, para tomar mejores decisiones en lo que respecta a la salud del paciente.
Los íconos usados en el caso de estudio serán:



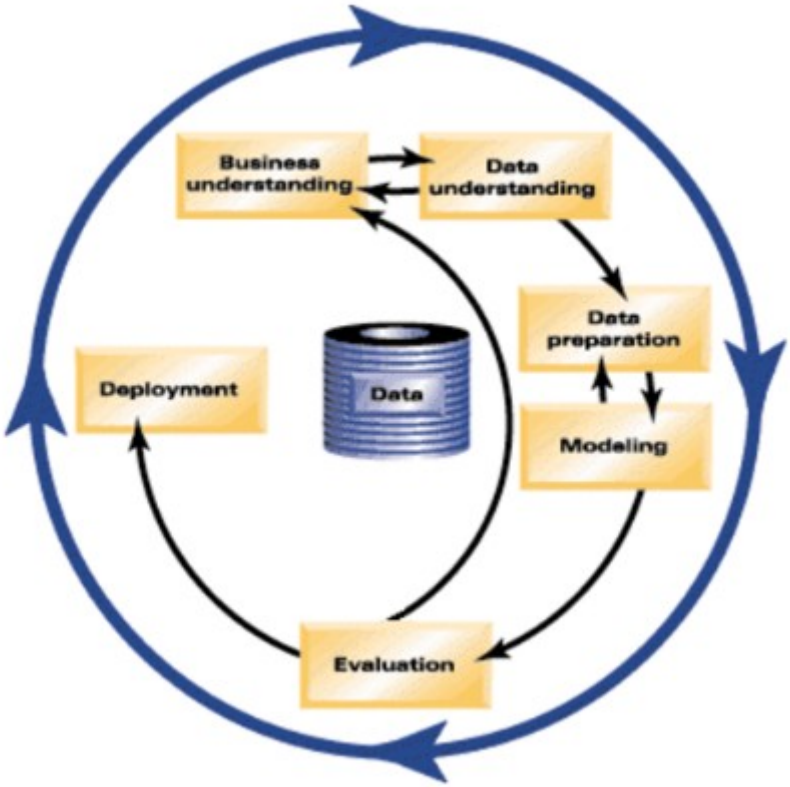
Este curso se basa en la **Foundational Methodology for Data Science de Jhon Rollins,** pero no es la única metodología que hay. Por ejemplo, en data mining, la metodología CRISP-DM (Cross Industry Process for Data Mining) es ampliamente utilizada.

2. CRISP-DM

La metodología **CRISP-DM** es un proceso destinado a aumentar el uso de la minería de datos sobre una amplia gama de aplicaciones de negocios e industrias.

El objetivo es tomar escenarios específicos y comportamientos generales para hacerlos de dominio neutral.

CRISP-DM consiste de 6 pasos con una entidad que debe implementarse para tener una posibilidad razonable de éxito. Los 6 pasos se muestran en el diagrama siguiente:



- **Business Understanding (entendimiento del negocio)**
  - Es donde se perfila la intención del proyecto.
  - Foundational Methodology y CRISP-DM están alineadas aquí.
  - Requiere comunicación y claridad.
  - La dificultad aquí es que los interesados tienen diferentes objetivos, sesgos y formas de relacionar la información. No ven todo de la misma manera.
  - Sin una perspectiva clara, concisa y completa de los objetivos del proyecto, los recursos serán gastados innecesariamente.

- **Data understanding (entendimiento de los datos)**
  - Se basa en el entendimiento del negocio.
  - Los datos se coleccionan en esta etapa.
  - El entendimiento de lo que quiere el negocio y sus necesidades determinará qué datos son recolectados, de qué fuentes y mediante qué métodos.
  - CRISP-DM combina las etapas de Data Requirements, Data Collection y Data Understanding de la Foundational Methodology.
- **Data Preparation (preparación de datos)**
  - Una vez que los datos han sido coleccionados, deben ser transformados en un subconjunto utilizable, a menos que se determine que se necesitan más datos.
  - Una vez que se selecciona un dataset, debe ser chequeado para casos cuestionables, ambiguos o faltantes.
  - Es común con el Foundational Methodology.
- **Modelling (Modelado)**
  - Una vez preparados para el uso, los datos deben ser expresados a través de modelos apropiados, brindar conocimientos significativos y ojalá nuevos conocimientos.
  - Este es el propósito del data mining: crear información de conocimiento que tenga significado y utilidad.
  - El uso de modelos revela patrones y estructuras dentro de los datos para proveer conocimiento dentro de las características de interés.
  - Los modelos son seleccionados sobre una porción de los datos y se ajustan de ser necesario.
- **Evaluación**
  - El modelo seleccionado debe ser testeado.
  - Esto usualmente se realiza teniendo un test pre-seleccionado, establecido para ejecutarse sobre el modelo entrenado.
  - Esto permite ver la efectividad del modelo sobre un conjunto que ve como si fuera nuevo.
  - Los resultados de aquí se usan para determinar la eficacia del modelo y presagiar su rol en la etapa siguiente.
- **Deployment (despliegue)**
  - El modelo es utilizado sobre nuevos datos fuera del alcance del dataset y por nuevos interesados.

- Las nuevas interacciones en esta fase pueden revelar nuevas variables y necesidades para el dataset y el modelo.
- Pueden revisarse necesidades o acciones del negocio, el modelo y los datos, o ambos.

CRISP-DM es altamente flexible y cíclico.
En cada etapa puede ser necesario revisar una etapa anterior y realizar cambios.
El punto clave es que es cíclico, por eso, incluso al final, se puede tener otro entendimiento del negocio para discutir la viabilidad luego del despliegue.

3. ENTENDIMIENTO DEL NEGOCIO

Alguna vez le ha pasado esto? Su jefe lo llama para tener una reunión, y éste lo hace consciente de una tarea importante con un deadline muy ajustado que debe cumplirse sí o sí. Hablan para asegurarse de que todos los aspectos de la tarea han sido considerados y la reunión concluye con ambos confiados en que las cosas están encaminadas. Más tarde, en la tardecita, luego de examinar varios problemas en juego, se da cuenta de que necesita realizar varias preguntas extra para cumplir con la tarea. Desafortunadamente, su jefe no estará disponible hasta mañana de mañana. Ahora, con un deadline tan ajustado, usted comienza a sentir cierta inseguridad. Qué hace? Se arriesga a seguir adelante o se detiene para buscar claridad.

La metodología de data science comienza dedicándole tiempo a buscar claridad, para obtener lo que puede ser llamado como "entendimiento del negocio".

Tener este entendimiento se ubica al comienzo de la metodología porque obtener claridad acerca del problema a resolver, le permite determinar qué datos serán utilizados para responder la pregunta clave.

Rollins sugiere que tener una pregunta claramente definida es vital porque finalmente dirige el enfoque analítico que será necesitado para abordar la pregunta.

A menudo, se pone mucho esfuerzo en responder lo que la gente CREE es la pegunta, y los métodos utilizados para tratarla no resuelven el problema verdadero.

Establecer una definición clara de la pregunta comienza con entender el OBJETIVO de la persona que la realiza. Por ejemplo, si un dueño de un negocio pregunta: "Cómo podemos reducir el costo de realizar una actividad" tenemos que entender si el objetivo es mejorar la eficiencia o la actividad; o incrementar la rentabilidad del negocio.

Una vez que el objetivo está claro, la siguiente pieza del puzzle es descubrir los objetivos que apoyan la meta.

Al desglosar los objetivos, discusiones estructuradas pueden tomar lugar donde las prioridades pueden ser identificadas de una forma que lleve a organizar y planificar cómo atacar el problema.

Dependiendo del problema, diferentes interesados necesitarán participar de las discusiones para ayudar a determinar los requerimientos y aclarar las preguntas.

Volvamos al **caso de estudio**: la pregunta es: "cuál es la mejor forma de asignar recursos limitados del presupuesto de salud para maximizar su uso para proveer un cuidado de calidad".

Conforme los fondos públicos para re-admisiones bajaron, las compañías aseguradoras corrían el riesgo de tener que compensar por la diferencia de costos, lo que potencialmente

incrementaría las tasas para sus clientes. Sabiendo que incrementar las tasas no sería una medida popular, la aseguradora hizo una reunión con autoridades del sistema de salud y científicos de datos de IBM para ver cómo resolver el problema en cuestión.

Antes de comenzar a recolectar los datos, las metas y objetivos debían ser definidos. Luego de cierto tiempo, el equipo priorizó “re-admisiones de pacientes” como un área efectiva para revisar. Con las metas y objetivos claros, se encontró que:

- Aproximadamente el 30% de los individuos que terminaron su tratamiento de rehabilitación serían re-admitidos en en el centro en el correr del año, y el 50% dentro de los 5 años.
- Los pacientes con fallas cardíacas congestivas estaban al comienzo de la lista de re-admisiones.

Se vio luego que un modelo de árbol de decisión podría ser aplicado al escenario, para determinar qué estaba ocurriendo.

Para ganar entendimiento del negocio, IBM propuso y entregó un workshop on-site para empezar.

La participación de los patrocinadores comerciales clave a lo largo del proyecto fue crítica. El patrocinador:

- Estableció la dirección general.
- Permaneció comprometido y brindó orientación.
- Aseguró el soporte necesario, donde se necesitaba.

4 requerimientos del negocio fueron identificados para cualquiera sea el modelo se fuese a construir:

- Predecir los resultados de re-admisión para aquellos pacientes con falla cardíaca congestiva.
- Predecir el riesgo de re-admisión.
- Entender la combinación de eventos que llevaron al resultado predicho.
- Aplicar un proceso simple de entender para nuevos pacientes, con respecto a su riesgo de re-admisión.

4. ENFOQUE ANALÍTICO

Seleccionar el enfoque analítico adecuado dependerá de la pregunta a responder. El enfoque incluye buscar aclaraciones de la persona que hace la pregunta, para poder elegir la ruta o enfoque más apropiado.

Una vez definido el problema a resolver, el enfoque analítico apropiado para este problema se selecciona en el contexto de las necesidades del negocio. Esta es la segunda etapa de la metodología de la ciencia de datos.

Habiendo logrado una sólida comprensión de la pregunta, se puede seleccionar el enfoque analítico. Esto significa identificar el tipo de patrones necesarios para responder la pregunta del modo más efectivo.

- Si la pregunta es determinar las probabilidades de una acción, podría usarse un **modelo predictivo.**
- Si la pregunta es mostrar relaciones, podría necesitarse un **enfoque descriptivo.**
  - Éste observaría grupos de actividades similares con base en eventos y preferencias.
- El **análisis estadístico** aplica para los problemas que requieren conteos. Por ejemplo, si la pregunta requiere una respuesta "sí" o "no", sería adecuado un **enfoque de clasificación** para predecir una respuesta.
- El **Aprendizaje Automático** es un campo de estudio que dota a las computadoras de capacidad para aprender sin ser explícitamente programadas. El Aprendizaje Automático puede usarse para identificar relaciones y tendencias en los datos, que de otro modo no serían accesibles o identificadas.
- En casos donde la pregunta sea aprender sobre el comportamiento humano, una respuesta apropiada sería usar enfoques de **Análisis de Grupos**.

Ahora, veamos el **caso de estudio** en relación con la aplicación del Enfoque Analítico.

Para este caso, se usó un **modelo de clasificación de árbol de decisión** para identificar la combinación de condiciones que conducen al resultado de cada paciente.

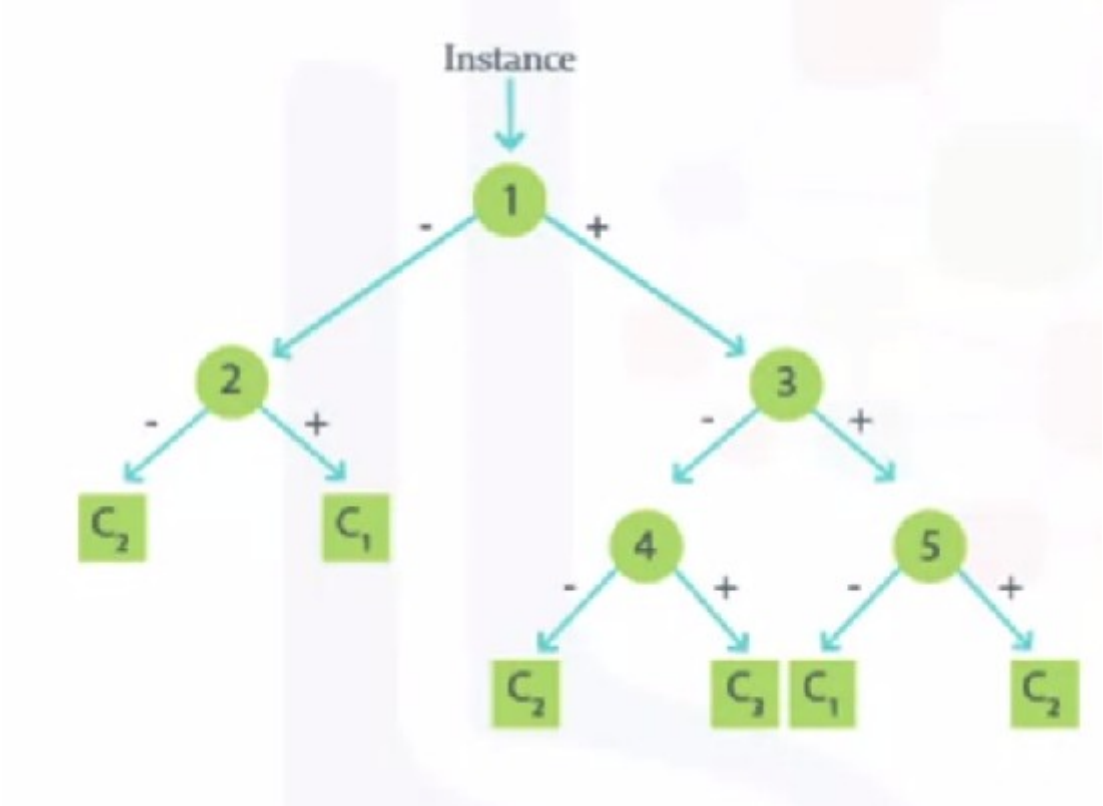
En este enfoque, examinar las variables de cada nodo a lo largo de la ruta de cada hoja, condujo a un respectivo valor de umbral. Esto significa que la clasificación de árbol de decisión indica:

- El resultado predicho.
- La probabilidad de cada desenlace, con base en la proporción de un resultado dominante, "sí" o "no", en cada grupo.

- Con esta información, los analistas pueden obtener el riesgo de re-admisión, o la probabilidad de un "sí" para cada paciente.
  - Si el resultado dominante es "sí", entonces el riesgo es simplemente la proporción de pacientes "sí" de la hoja.
  - Si es "no", entonces el riesgo es 1 menos la proporción de pacientes "no" de la hoja.

Un modelo de clasificación de árbol de decisión es fácil de aplicar y entender para quiénes no son científicos de datos, y así puntuar a nuevos pacientes según su riesgo de re-admisión.

Los médicos ven fácilmente las condiciones que hacen que un paciente sea puntuado como de alto riesgo, y pueden construirse y aplicarse múltiples modelos en varios puntos de la estadía en el hospital. Esto brinda una imagen móvil del riesgo del paciente y cómo éste evoluciona con los diversos tratamientos aplicados. Por ello, se eligió el enfoque de clasificación de árbol de decisión para construir el modelo de re-admisión de insuficiencia cardíaca.



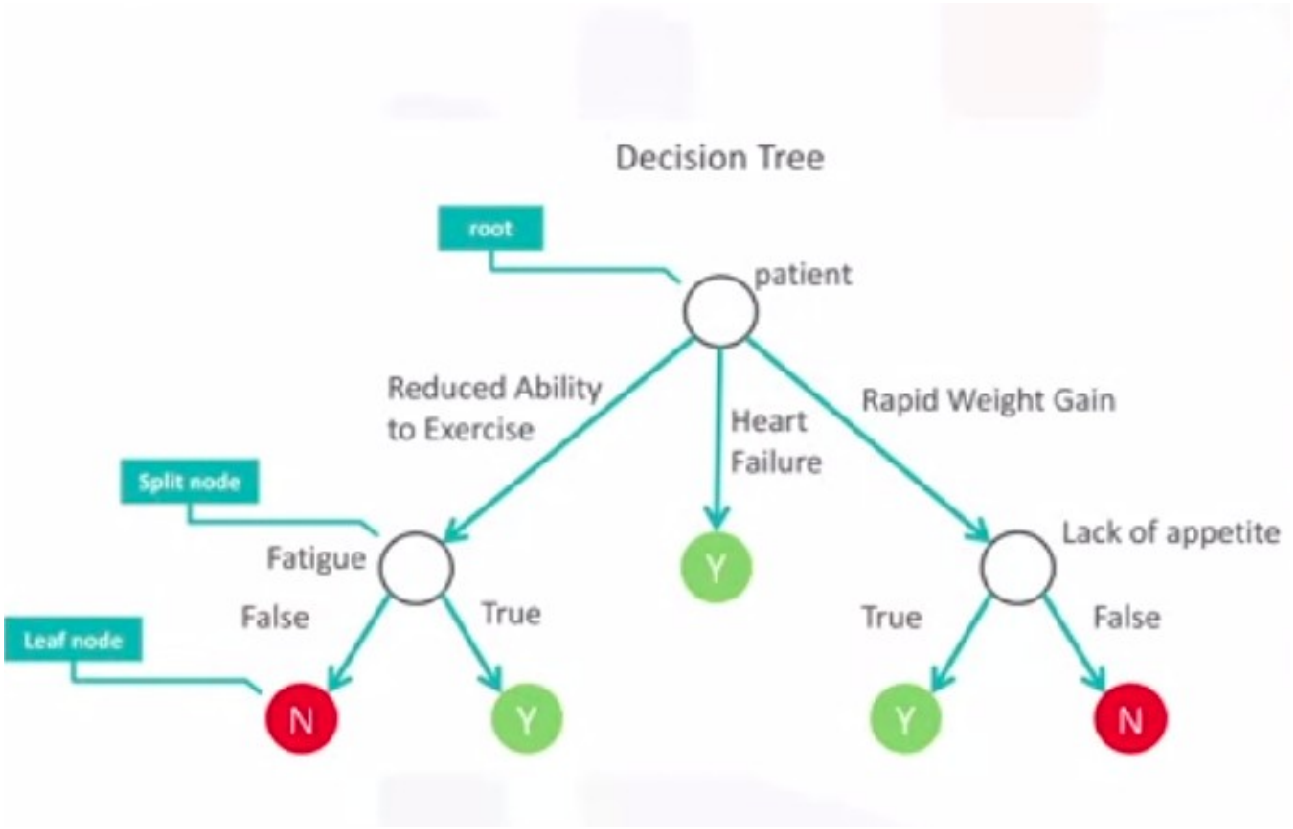


Figura. En el caso de estudio trabajaremos con un modelo de árbol de decisión.

5. REQUERIMIENTOS DE LOS DATOS

Si su meta es hacer espagueti para la cena pero no tiene los ingredientes correctos para realizarlo entonces su éxito se verá comprometido. Piense en esta sección como si fuera cocinar con los datos. Cada paso es crítico al preparar la comida. Entonces, si el problema a resolver es la receta y los datos son los ingredientes, entonces el científico de datos necesita identificar:

- Qué ingredientes se necesitan.
- Cómo recolectar los ingredientes.
- Cómo entender o cómo trabajar con los ingredientes.
- Cómo preparar los datos para cumplir con el resultado deseado.

Al construir el entendimiento del problema y luego usar el enfoque analítico seleccionado, el científico de datos está pronto para comenzar.

Veamos ejemplos de requerimientos de datos dentro de la metodología de ciencia de datos.

Antes de emprender la colección y preparación de datos es vital definir los requerimientos para el árbol de clasificación. Esto incluye identificar los contenidos de datos, formato y fuentes necesarios para la colección inicial de datos.

Veamos nuestro **caso de estudio**.

La primer tarea fue definir los requerimientos de datos para el árbol de decisión seleccionado. Esto incluye seleccionar un grupo de pacientes adecuado de los miembros base del proveedor de salud.

Para recopilar las historias clínicas completas, 3 criterios fueron identificados para la inclusión en el grupo:

- El paciente debía ser admitido como paciente dentro del área del proveedor de servicio, para así tener acceso a la información necesaria.
- Se puso el foco en pacientes con un diagnóstico primario de falla cardíaca congestiva durante un año completo.
- El paciente debe tener una inscripción continua de al menos 6 meses, antes de la admisión primaria por falla cardíaca congestiva, para que la historia médica completa pudiese ser recopilada.

Los pacientes con falla cardíaca congestiva que también tuvieron diagnosticados otras condiciones médicas fueron excluidos del grupo porque esas condiciones causaron tasas de re-admisiones mayores al promedio, y por tanto, podrían sesgar los resultados.

Luego, el contenido, formato y representaciones de los datos necesarios para el árbol fueron definidos.

Esta técnica de modelado requiere un registro por paciente, con columnas representando las variables en el modelo.

Para modelar el resultado de re-admisiones, era necesario que hubiera datos que cubrieran todos los aspectos de la historia clínica del paciente, este contenido incluye:

- Admisiones.
- Diagnósticos primarios, secundarios y terciarios.
- Procedimientos.
- Prescripciones.
- Otros servicios que hayan sido provistos durante la hospitalización o visitas paciente/ doctor.

Así, un paciente particular podría tener miles de registros, representando todos sus atributos relacionados.

Para obtener un registro por paciente, los científicos de datos acumularon los registros transaccionales a nivel del paciente, creando un número de variables nuevas para representar la información. Este fue un trabajo para la etapa de preparación, por lo que pensar hacia adelante y anticiparse a subsecuentes etapas es importante.

6. RECOLECCIÓN DE DATOS

Luego de hacer la recolección de datos inicial, el científico de datos hace una evaluación para determinar si tiene o no tiene lo que necesita. Como ocurre cuando se compran ingredientes para la cena, algunos ingredientes pueden estar fuera de temporada y ser más difíciles de obtener, o más costosos de lo pensado.

En esta fase se revisan los requerimientos de datos y se decide si la recolección requiere más, o menos datos.

Una vez recolectados los ingredientes de datos en la etapa de recolección, el científico de datos tendrá una buena comprensión de los insumos con los cuales trabajará.

Técnicas como la estadística descriptiva y la visualización pueden aplicarse al conjunto de datos para evaluar su contenido, calidad y conclusiones iniciales.

Se identificarán vacíos en los datos y se harán planes, ya sea para llenarlos o hacer sustituciones.

En esencia, los ingredientes están ahora sobre la tabla de corte.

Ahora veamos algunos ejemplos de la etapa de recolección de la metodología de la ciencia de datos. Esta etapa se emprende al finalizar la etapa de requerimientos de datos. Veamos el caso de estudio en relación con la aplicación de la "Recolección de datos". Para recolectar datos, necesitas saber la fuente o el modo de hallar los elementos de datos que se requieren. En el contexto de nuestro caso de estudio, estos pueden incluir:

- Información demográfica, clínica o de cobertura de los pacientes.
- información del proveedor.
- registros de reclamos.
- Información farmacéutica y adicional relacionada con diagnósticos de pacientes de insuficiencia cardíaca.

En este caso de estudio, se necesitaba cierta información de medicamentos, pero esa fuente de datos aún no estaba integrada al resto de las fuentes de datos. Esto nos lleva a un punto importante: **Está bien aplazar decisiones acerca de datos no disponibles, e intentar adquirirlos en una etapa posterior**. Por ejemplo, esto puede hacerse aún después de obtener resultados intermedios del modelado predictivo. Si esos resultados indican que la información de medicamentos será importante para lograr un buen modelo, se invertiría el tiempo para intentar conseguirla. Sin embargo, resultó que pudieron construir un modelo razonablemente bueno sin esta información de medicamentos.

Los DBA y programadores a menudo trabajan juntos para extraer datos de diversas fuentes, y luego los combinan. Esto permite eliminar datos redundantes, dejándolos a disposición

para la próxima etapa de la metodología, que es la comprensión de datos. En esta etapa, si es necesario, científicos de datos y miembros del equipo de análisis pueden discutir maneras de manejar mejor sus datos, como automatizar ciertos procesos en la base de datos, para que la recolección de datos sea más fácil y rápida.

7. ENTENDIMIENTO DE LOS DATOS

La comprensión de los datos involucra todas las actividades de construcción de un conjunto de datos.

En esencia, la sección de comprensión de los datos de la metodología de la ciencia de datos responde a esta pregunta: **¿Los datos recolectados son representativos del problema a resolver?**

Apliquemos la etapa de comprensión de los datos de nuestra metodología, al **caso de estudio** que hemos venido examinando.

Para comprender los datos relacionados con admisiones por insuficiencia cardíaca, se debía correr estadísticas descriptivas frente a las columnas de datos que se volverían variables en el modelo.

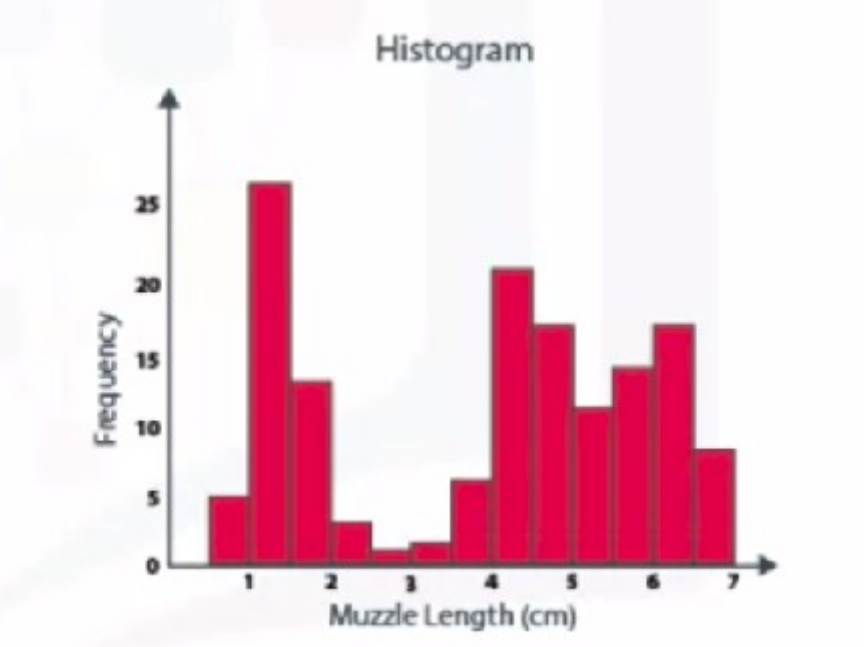
Primero, estas estadísticas incluyeron univariantes y estadísticas en cada variable tales como:

- Media
- Mediana
- Mínimo
- Máximo
- Desviación estándar.

Segundo, se usaron correlaciones por pares, para ver qué tan cerca se relacionaban ciertas variables, y cuáles, si las había, estaban muy altamente correlacionadas (de modo que serían esencialmente redundantes), haciendo que sólo una fuera relevante para el modelado.

Tercero, se examinaron **histogramas** de las variables para entender sus distribuciones.

Los histogramas son un buen modo de entender cómo se distribuyen los valores de una variable, y cuáles tipos de preparación se necesitarían para volver la variable más útil en un modelo. Por ejemplo, para que una variable categórica con demasiados valores distintos sea informativa en un modelo, el histograma les ayudaría a decidir cómo consolidar esos valores.



Univariantes, estadísticas e histogramas también se usan para evaluar la calidad de los datos. Con la información conseguida, ciertos valores pueden re-codificarse o incluso descartarse de ser necesario, como cuando cierta variable tiene demasiados valores faltantes. La pregunta es, entonces, **¿"faltante" significa algo?**

- A veces un valor faltante puede significar "no", o "0" (cero), o en ocasiones sólo significa "no sabemos".
- O, si una variable contiene valores inválidos o confusos, tales como una variable numérica llamada "edad" que contiene 0 a 100 y también 999, donde ese "triple-9" realmente significa "faltante", pero se trataría como un valor válido a menos que lo corrigiéramos.

Inicialmente, el significado de admisión por insuficiencia cardíaca se decidió sobre la base de un diagnóstico primario de insuficiencia cardíaca. Pero al avanzar en la etapa de comprensión de los datos se reveló que la definición inicial no capturaba todas las admisiones por insuficiencia cardíaca que se esperaban, según la experiencia clínica. Esto implicó volver a la etapa de recolección de datos y agregar diagnósticos secundarios y terciarios, construyendo una definición más completa de la admisión por insuficiencia cardíaca.

Entre más se trabaja con el problema y los datos, más se aprende y por lo tanto más refinamientos pueden hacerse dentro del modelo, llevando en últimas a una mejor solución del problema.

8. PREPARACIÓN DE LOS DATOS

En cierta forma, la preparación de datos se parece a lavar los vegetales recién elegidos, pues elimina los elementos indeseados, como la tierra e imperfecciones.

Junto con la recolección de datos y la comprensión de datos, la preparación de datos es la fase más dispendiosa de un proyecto de ciencia de datos, ocupando por lo general el setenta por ciento y aún el noventa por ciento del tiempo total del proyecto. Automatizar algunos procesos de recolección y preparación de datos en la base datos puede reducir este tiempo a tan sólo el 50 por ciento. Este ahorro se traduce en más tiempo para que los científicos de datos se centren en crear modelos.

Continuando con la metáfora culinaria, sabemos que el proceso de cortar cebollas en trozos más pequeños permitirá que su sabor se esparza por la salsa más fácil que si dejáramos caer la cebolla entera en la olla de la salsa. Así mismo, transformar los datos en la etapa de preparación es el proceso de llevar los datos a un estado en que sea más fácil trabajar con ellos.

En particular, la etapa de preparación de datos de la metodología responde a esto: **¿Cuáles son los modos de preparación de los datos?**

Para trabajar efectivamente con datos, deben prepararse de una forma que aborde los datos faltantes o inválidos y remueva duplicados, para asegurar que todo está debidamente formateado.

La **ingeniería de características t**ambién es parte de la preparación de datos. Es el proceso de usar conocimiento del dominio de los datos para crear características que hagan funcionar algoritmos de aprendizaje automático.

Una **característica** es una propiedad que puede ayudar a resolver un problema. Las características de los datos son importantes en los modelos predictivos e influyen en los resultados que quieras conseguir.

La ingeniería de características es crítica al aplicar herramientas de aprendizaje automático para analizar los datos. Al trabajar con texto, se necesitan pasos de análisis textual para codificar los datos y así poder manipular los datos. El científico de datos debe saber qué está buscando en su conjunto de datos para abordar la pregunta. El análisis textual es crucial para fijar las agrupaciones apropiadas, y asegurar que la programación no pase por alto lo que se oculta en el interior.

La fase de preparación de datos prepara el terreno para los próximos pasos en torno a la pregunta.

Si bien esta fase puede ser demorada, si se hace bien los resultados respaldarán el proyecto.

Si se deja de lado, el resultado no estará a la altura y podría obligarte a comenzar desde cero.

Es vital invertir tiempo en esta etapa, y usar las herramientas disponibles para automatizar los pasos comunes y acelerar la preparación de datos. Asegúrate de prestarle atención a los detalles en esta fase. Después de todo, un sólo ingrediente malo puede arruinar una buena receta.

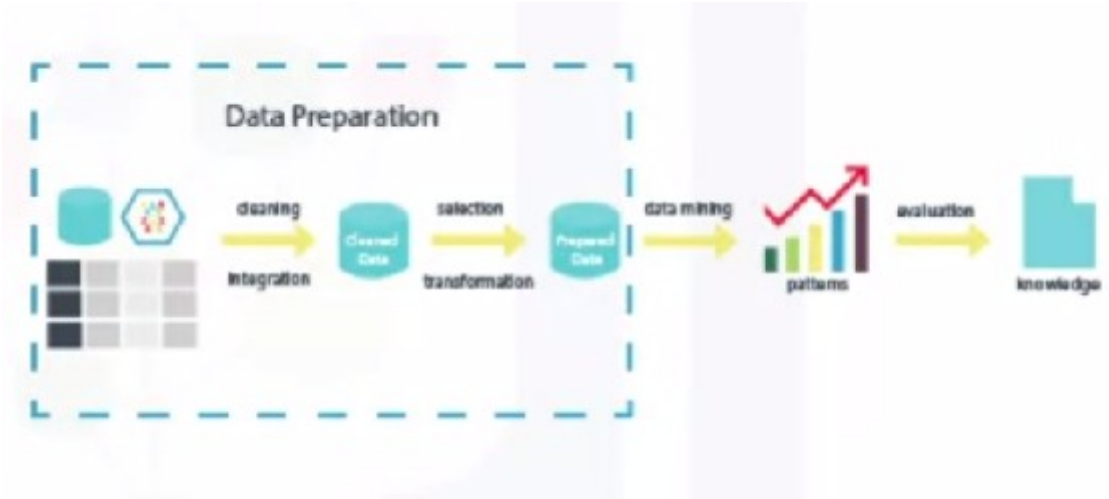


Figura. Preparación de datos. Cuáles son las formas en las que se preparan los datos.

8.1. Caso de estudio

En cierta forma, la preparación de datos se parece a lavar los vegetales recién elegidos pues elimina los elementos indeseados, como la tierra e imperfecciones.

Ahora, veamos el **caso de estudio** en relación con la aplicación de conceptos de Preparación de Datos.

En este estudio de caso, un primer paso importante en la etapa de preparación de datos era definir la insuficiencia cardíaca. Esto parecía fácil al inicio, pero definirlo con precisión no fue tan simple. Primero, el conjunto de códigos grupales de diagnóstico debía ser identificado, pues la insuficiencia cardíaca trae cierto tipo de retención de fluidos. También debíamos tener en cuenta que la insuficiencia cardíaca es tan sólo un tipo de afección cardíaca. Necesitamos ayuda médica para obtener los códigos correctos de insuficiencia cardíaca.

El próximo paso incluyó definir los criterios de re-admisión para esta dolencia. La secuencia de eventos debía evaluarse para definir si una admisión por insuficiencia cardíaca en particular era un evento inicial, denominado admisión índice, o una re-admisión relacionada

con insuficiencia cardíaca. Con base en la experiencia médica, se fijó un periodo de 30 días como ventana de re-admisión relevante para pacientes de insuficiencia cardíaca, luego de recibir el alta por la admisión inicial.

Luego se agregaron los registros que estaban en formato transaccional, es decir, que los datos incluían múltiples registros por cada paciente. Los registros transaccionales incluían reclamos en instalaciones profesionales del proveedor por servicios médicos, de laboratorio, hospitalarios y clínicos. También se incluían registros de todos los diagnósticos, procedimientos, prescripciones, y más información de pacientes hospitalizados y ambulatorios.

Un paciente dado podía tener cientos o incluso miles de estos registros, dependiendo de su historia clínica.

Luego se agregaron todos los registros transaccionales a nivel del paciente, obteniendo un sólo registro para cada paciente, como exigía el método de clasificación de árbol de decisión que se usaría para el modelado.

Como parte del proceso de agregación, se crearon muchas columnas nuevas para la información de las transacciones. Por ejemplo, la frecuencia y últimas visitas a doctores, clínicas y hospitales con diagnósticos, procedimientos, prescripciones, y demás.

También se tuvieron en cuenta las comorbilidades de la insuficiencia cardíaca, como diabetes, hipertensión, y muchas otras afecciones y dolencias crónicas que podrían afectar el riesgo de re-admisión por insuficiencia cardíaca.

Como parte del debate sobre la preparación de datos, también se hizo una revisión bibliográfica sobre la insuficiencia cardíaca para evitar omitir elementos de datos importantes, como comorbilidades que no se hubieran tenido en cuenta aún. La revisión bibliográfica implicó volver a la etapa de recolección de datos para agregar algunos indicadores adicionales para dolencias y procedimientos. Agregar los datos transaccionales a nivel del paciente, significó combinarlos con otros datos del paciente, incluyendo su información demográfica, como edad, género, tipo de seguro, entre otros. El resultado fue la creación de una tabla con un sólo registro por paciente, y muchas columnas que representaban los atributos del paciente en su historia clínica. Estas columnas se usarían como variables en el modelado predictivo. Esta es una lista de las variables que se utilizaron para construir el modelo.

La variable dependiente, o resultado, era la re-admisión por insuficiencia cardíaca en los 30 días siguientes a recibir el alta de una hospitalización por insuficiencia cardíaca, con un resultado de "sí" o "no". La etapa de preparación de datos arrojo una grupo de 2.343

pacientes que cumplían los criterios para este estudio de caso. Esta grupo luego se dividió en grupos de entrenamiento y prueba para construir y validar el modelo, respectivamente.

9. CONCEPTOS DE MODELADO

El modelado es la etapa de la metodología de la ciencia de datos donde el científico de datos tiene la oportunidad de probar la salsa y decidir si está en el punto o necesita más condimentos!

Esta porción del curso está dirigida a responder dos preguntas claves:

- ¿ Cuál es el propósito del modelado de datos ?
- ¿ Cuáles son algunas características de este proceso?

El Modelado de Datos se enfoca en desarrollar modelos que sean o descriptivos, o predictivos.

Un **modelo descriptivo** podría examinar, por ejemplo, cosas como: si una persona hace esto, entonces es probable que prefiera aquello.

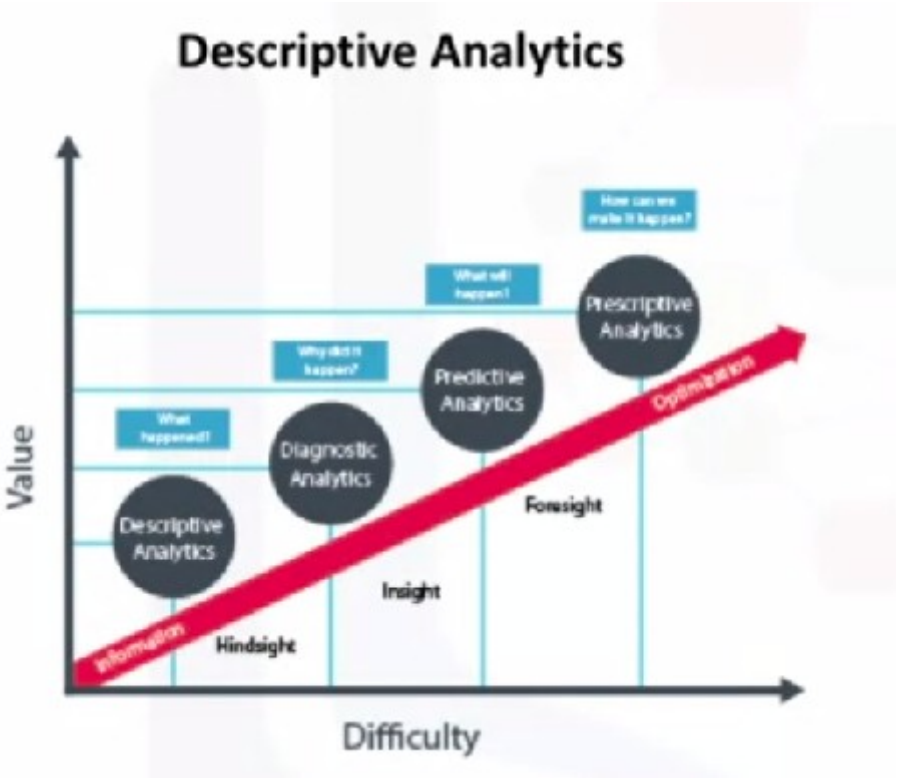


Figura. Análisis descriptivo.

Un **modelo predictivo** intenta dar resultados de tipo sí/no, o stop/go.
Estos modelos se basan en el enfoque analítico que se haya elegido, ya sea impulsado por la estadística, o por el aprendizaje automático.

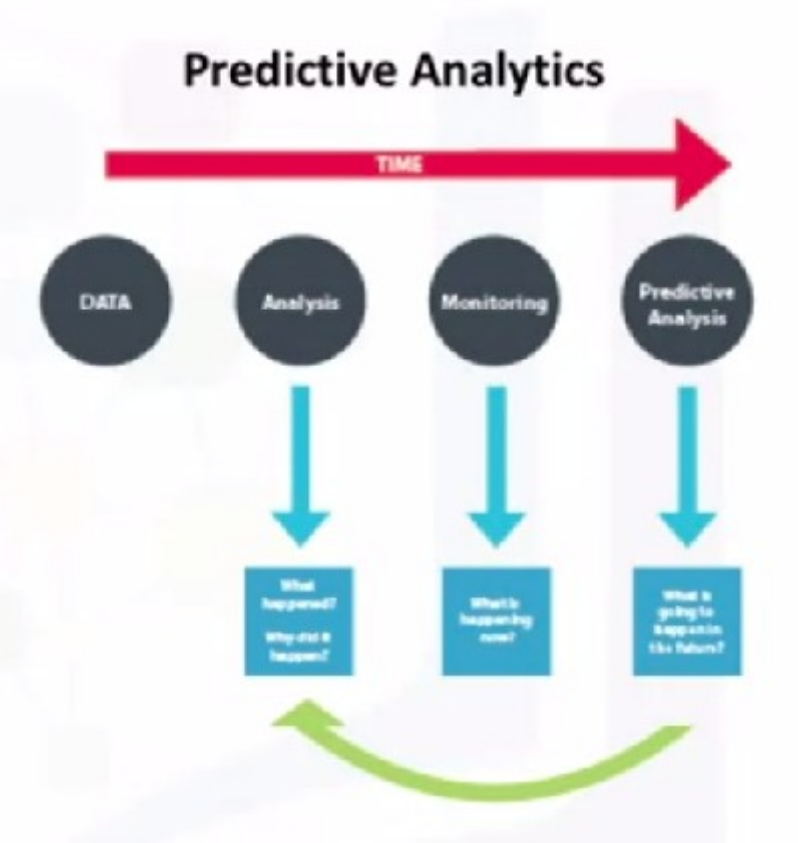


Figura. Análisis predictivo.

El científico de datos usará un conjunto de entrenamiento para el modelado predictivo. Un **conjunto de entrenamiento** es un conjunto de datos históricos para los cuales ya se conoce el resultado. El conjunto de entrenamiento funciona como un medidor para definir si el modelo debe ser calibrado.

En esta etapa, el científico de datos jugará con distintos algoritmos para asegurar que las variables en acción realmente se requieran.

El éxito de la compilación, preparación y modelado de datos, dependerá de:

- La comprensión del problema a mano y de
- La elección del enfoque analítico apropiado.

Los datos ayudan a la resolución de la pregunta, y como la calidad de los ingredientes al cocinar, preparan el terreno para el resultado.

Se debe refinar, ajustar y afinar constantemente en cada paso, para asegurar que el resultado sea uno sólido.

En la descriptiva Metodología de la Ciencia de Datos de John Rollins, el marco de referencia busca conseguir 3 cosas:

- Primero, comprender la pregunta a mano.
- Segundo, elegir un enfoque analítico o método para resolver el problema, y,
- tercero, obtener, comprender, preparar y modelar los datos.

La meta final es llevar al científico de datos a un punto donde pueda construir un modelo de datos para responder la pregunta.
Cuando la cena está lista para servir y el huésped hambriento está a la mesa, la pregunta clave es: ¿He preparado bastante para comer? Bien, esperemos que así sea.
En esta etapa de la metodología, los bucles de evaluación, despliegue y retroalimentación del modelo aseguran que la respuesta sea cercana y relevante. Esta relevancia es crucial para el campo de la ciencia de datos en general, por ser un campo de estudio más bien reciente, y nos interesan las posibilidades que tiene para ofrecer. Entre más gente se beneficie con los resultados de esta práctica, más lejos avanzará su desarrollo.

9.1. Caso de estudio

El modelado es la etapa de la metodología de la ciencia de datos donde el científico de datos tiene la oportunidad de probar la salsa y decidir si está en el punto o necesita más condimentos!
Ahora, apliquemos en el caso de estudio la etapa de modelado de la metodología de la ciencia de datos.
Aquí, discutiremos los muchos aspectos de la construcción de modelos, en este caso, el ajuste de parámetros para mejorar el modelo.
Con un conjunto de entrenamiento listo, se puede construir el primer modelo de clasificación de árbol de decisión para la re-admisión por insuficiencia cardíaca.
Buscamos pacientes con alto riesgo de re-admisión, así que el resultado que nos interesa será re-admisión por insuficiencia cardíaca igual a "sí".
En este primer modelo, la precisión general en la clasificación de resultados "sí" y "no" fue del 85%. Suena bien, pero sólo representa el 45% del "sí". Las re-admisiones reales están clasificadas correctamente, de manera que el modelo no es muy preciso.
La pregunta es, entonces: ¿Cómo se podría mejorar la precisión del modelo para predecir el resultado "sí"?
En la clasificación de árbol de decisión, el mejor parámetro para ajustar es el costo relativo de los resultados "sí" y "no" mal clasificados. Míralo así: Cuando una no-re-admisión

verdadera se clasifica mal, y se adoptan acciones para reducir el riesgo del paciente, el costo de aquel error es la intervención desperdiciada. Un estadístico llamaría a esto un **error tipo I, o un falso positivo**. Pero cuando una re-admisión verdadera se clasifica mal, y no se adoptan acciones para reducir ese riesgo, el costo de aquel error es la re-admisión y todos sus costos relacionados, además del traumatismo para el paciente. Este es un **error tipo II, o un falso negativo.**

Como podemos ver, los costos de los dos tipos de errores de mala clasificación pueden ser bastante diferentes.

Por esta razón, es razonable ajustar los pesos relativos de la mala clasificación de resultados "sí" y "no". Por defecto es de 1 a 1, pero el algoritmo de árbol de decisión permite fijar un valor más alto para "sí".

En el segundo modelo, el costo relativo se fijó en 9 a 1. Esta es una relación muy alta, pero permite explorar el comportamiento del modelo. Esta vez, el modelo clasificó correctamente el 97% de los "sí", pero a costa de una muy baja precisión para el "no", con una precisión general de sólo 49%. Claramente, este no era un buen modelo. El problema con este resultado es el alto número de falsos positivos, que recomendarían intervenciones costosas e innecesarias, para pacientes que no hubieran sido readmitidos de ninguna manera. Por lo tanto, el científico de datos debe volver a tratar de hallar un mejor equilibrio entre las precisiones del "sí" y el "no".

En el tercer modelo, el costo relativo se fijó en un más razonable 4 a 1. Esta vez se obtuvo una precisión del 68% sólo en el "sí", denominada sensibilidad por los estadísticos, y una precisión del 85% para el "no", llamada especificidad, con una precisión general del 81%. Este es el mejor balance que se puede conseguir con un conjunto de entrenamiento más bien pequeño a través del ajuste del parámetro de costo relativo de mala clasificación de resultados "sí" y "no".

El modelado, por supuesto, requiere mucho más trabajo, incluyendo iterar de vuelta a la etapa de preparación de datos para redefinir algunas otras variables, para así representar mejor la información subyacente, y en consecuencia mejorar el modelo.

Model	Relative Cost Y:N	Overall Accuracy (% correct Y & N)	Sensitivity (Y accuracy)	Specificity (N accuracy)
1	1:1	85%	45%	97%
2	9:1	49%	97%	35%
3	4:1	81%	68%	85%

Figura. Analizando la precisión de los distintos modelos.

10. EVALUACIÓN

La evaluación de un modelo va de la mano de su misma construcción. Las etapas de modelado y evaluación se hacen iterativamente. La evaluación de un modelo se hace durante su desarrollo y antes de ser desplegado. La evaluación permite evaluar la calidad del modelo pero también es una oportunidad para ver si cumple con la solicitud inicial. La evaluación responde a la pregunta: **¿El modelo usado realmente responde la pregunta inicial, o necesita ser ajustado?**

La evaluación del modelo puede tener dos fases principales.

- La **primera es la fase**
  - Es la de medidas de diagnóstico, que se usa para asegurar que el modelo funcione como se pretendía. Si es modelo es predictivo, se puede usar un árbol de decisión para evaluar si la respuesta que arroja el modelo está alineada con el diseño inicial. Se puede usar para ver dónde hay áreas que requieran ajustes. Si el modelo es descriptivo, uno donde se evalúen relaciones, entonces puede aplicarse un conjunto de prueba con resultados conocidos, y el modelo puede refinarse como sea requerido.
- La **segunda fase**
  - Es la prueba de importancia estadística. Este tipo de evaluación puede aplicarse al modelo para asegurar que los datos estén siendo manejados e interpretados adecuadamente dentro del modelo. Está diseñado para evitar dudas innecesarias cuando la respuesta sea revelada.

Ahora, volvamos a nuestro estudio de caso para aplicar el componente de Evaluación de la metodología de la ciencia de datos.

Veamos una forma de hallar el modelo óptimo con una medida de diagnóstico con base en el ajuste de uno de los parámetros de construcción del modelo. En particular, veremos cómo ajustar el costo relativo de mala clasificación de resultados "sí" y "no". Como indica esta tabla, se hicieron cuatro modelos con cuatro costos relativos diferentes de mala clasificación. Como vemos, cada valor de este parámetro de construcción de modelos aumenta la tasa de verdaderos positivos, o la sensibilidad, de la precisión para predecir "sí", a expensas de una menor precisión para predecir "no". Esto es, una creciente tasa de falsos positivos.

La pregunta es, ¿cuál modelo es mejor con base en el ajuste de este parámetro? Por razones de presupuesto, la intervención de reducción de riesgo no podía aplicarse a la mayoría o la totalidad de los pacientes de insuficiencia cardíaca, muchos de los cuales no habrían sido readmitidos de cualquier manera. Por otra parte, la intervención no sería tan

efectiva para mejorar la atención al paciente como debería, si no se enfocaran suficientes pacientes de insuficiencia cardíaca de alto riesgo. Entonces, ¿cómo definir cuál era el modelo óptimo? Como puedes ver en esta dispositiva, el modelo óptimo es el que brinda la máxima separación entre la **curva ROC** azul con respecto a la línea base roja.

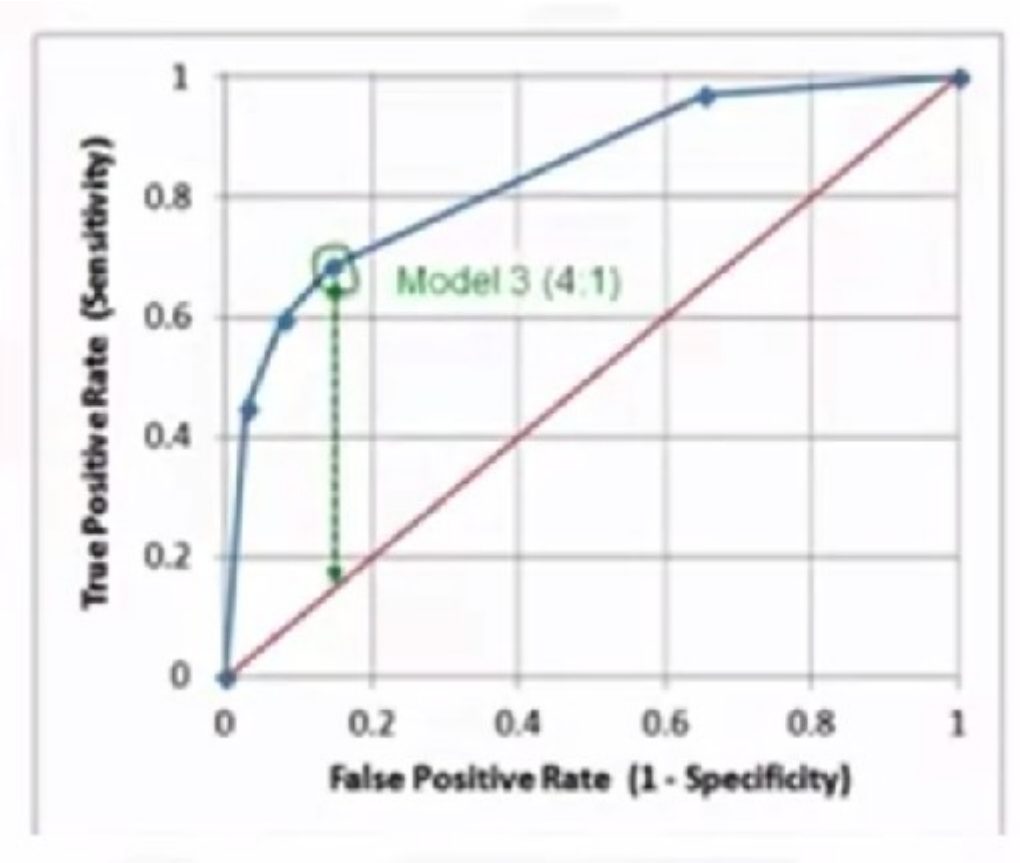


Figura. Curva ROC.

Vemos que el modelo 3, con un costo relativo de mala clasificación de 4 a 1, es el mejor de los 4 modelos.

Y por si acaso tenías curiosidad, el acrónimo ROC se traduce como Característica Operativa del Receptor, creada en la Segunda Guerra Mundial para detectar aviones enemigos en el radar. Desde entonces ha sido usada en muchos otros campos. Hoy es de uso común en el aprendizaje automático y la minería de datos. La curva ROC es una útil herramienta de diagnóstico para definir el modelo óptimo de clasificación. Esta curva cuantifica el desempeño de un modelo binario de clasificación, desclasificando los resultados "sí" y "no" cuando cambia algún criterio de discriminación. En este caso, el criterio es el costo relativo

de mala clasificación. Al graficar la tasa de verdaderos positivos contra la tasa de falsos positivos para distintos valores del costo relativo de mala clasificación, la curva ROC ayudó a seleccionar el modelo óptimo.

11. DESPLIEGUE

Si bien un modelo de ciencia de datos dará una respuesta, la clave para hacerla relevante y útil para abordar la pregunta inicial, implica familiarizar a los actores con la herramienta producida.

En un escenario de negocios, los actores tienen distintos especialistas que ayudarán a este objetivo, como el dueño de la solución, mercadeo, desarrolladores y administración de IT.

Cuando el modelo ha sido evaluado y el científico de datos confía en que funcionará, se despliega y se somete a la prueba definitiva.

Dependiendo del propósito del modelo, puede lanzarse entre un grupo limitado de usuarios o en un entorno de prueba, creando confianza en la aplicación del resultado con miras a un uso generalizado.

Ahora, veamos el estudio de caso en relación con la aplicación del "Despliegue".

Como preparación para el despliegue de la solución, el próximo paso era asimilar el conocimiento al grupo de negocios que diseñaría y administraría el programa de intervención para reducir el riesgo de re-admisión. En este escenario, la gente de negocios tradujo los resultados del modelo para que el equipo médico entendiera cómo identificar pacientes de alto riesgo y diseñara acciones de intervención adecuadas.

La meta, claro, era reducir la probabilidad de que aquellos pacientes fueran readmitidos dentro de 30 días después del alta.

En la etapa de requisitos del negocio, la Directora del Programa de Intervención y su equipo querían una aplicación que brindara evaluaciones de riesgo de insuficiencia cardíaca automáticas, casi en tiempo real. También debía ser de fácil uso para el equipo médico, ojalá mediante una aplicación de tablet basada en navegador, que cada miembro del equipo pudiera portar.

Los datos del paciente se generarían a lo largo de la hospitalización. Automáticamente se prepararían en el formato que el modelo requería y cada paciente sería puntuado cerca de la hora del alta. Así los médicos tendrían la evaluación de riesgo más actualizada de cada paciente, ayudándoles a escoger cuáles pacientes recibirían intervenciones después del alta.

Como parte del despliegue de la solución, el equipo de Intervención crearía y brindaría entrenamiento al equipo médico. También, se desarrollarían procesos para seguir y monitorear a los pacientes que recibieran la intervención, en colaboración con desarrolladores de IT y administradores de bases de datos, para que los resultados pasaran por la etapa de retroalimentación y el modelo fuera refinado con el tiempo.

Este mapa ejemplifica el despliegue de una solución por una aplicación Cognos.

Aquí, el estudio de caso era el riesgo de hospitalización de pacientes con diabetes juvenil. Como el caso de uso de insuficiencia cardíaca, este usó una clasificación de árbol de decisión para crear un modelo de riesgo que serviría como base para esta aplicación. El mapa muestra un resumen del riesgo nacional de hospitalización, con un análisis interactivo del riesgo previsto según varias afecciones del paciente y otras características. Esta dispositiva muestra un reporte interactivo del riesgo según la población del paciente dentro de cada nodo del modelo, para que los médicos pudieran entender la combinación de condiciones para este subgrupo de pacientes.

Y este reporte entrega un resumen detallado de un paciente individual, incluyendo el riesgo previsto del paciente y detalles de su historia clínica, brindando una síntesis concisa para el doctor.

12. RETROALIMENTACIÓN

Sobre la marcha, la retroalimentación de los usuarios ayudará a refinar el modelo y evaluar su desempeño e impacto.

El valor del modelo dependerá de

- incorporar exitosamente la retroalimentación y
- hacer ajustes durante todo el tiempo que se requiera la solución.

En la Metodología de la Ciencia de Datos, cada fase le da paso a la siguiente. Volver cíclica la metodología, asegura refinamiento en cada etapa del juego.

El proceso de retroalimentación nace de la noción de que, entre más conozcas, más querrás conocer.

Cuando el modelo ha sido evaluado y el científico de datos confía en que funcionará, se despliega y se somete a la prueba definitiva: el verdadero uso de campo en tiempo real. Entonces, miremos de nuevo nuestro estudio de caso, para ver cómo se aplica la porción de retroalimentación de la metodología.

El plan para la etapa de retroalimentación incluyó estos pasos:

- Primero, se definiría y prepararía el proceso de revisión, con la responsabilidad general de medir los resultados de un modelo de "vuelo al riesgo" de la población de riesgo por insuficiencia cardíaca. Los ejecutivos de manejo clínico serían los responsables del proceso de revisión.
- Segundo, se haría seguimiento a los pacientes de insuficiencia cardíaca que recibieran intervención y se registrarían sus resultados de re-admisión.
- Tercero, la intervención se mediría para determinar qué tan efectiva era para reducir las re-admisiones.

Por razones éticas, los pacientes de insuficiencia cardíaca no se dividirían en grupos de control y tratamiento. En cambio, las tasas de re-admisión se compararían antes y después de la implementación del modelo para medir su impacto.

Luego de las etapas de despliegue y retroalimentación, se revisaría el impacto del programa de intervención en las tasas de re-admisión después del primer año de su implementación. Entonces se refinaría el modelo, con base en los datos compilados después de la implementación del modelo y el conocimiento obtenido a lo largo de estas etapas.

Otros refinamientos incluían:

- Incorporar información de participación en el programa de intervención,
- Refinar el modelo para incorporar datos farmacéuticos detallados.

- Si recuerdas, la recolección de datos se aplazó inicialmente porque los datos farmacéuticos no estaban disponibles en ese momento. Pero después de la retroalimentación y de experimentar el modelo en la práctica, podría decidirse que agregar esos datos ameritaría la inversión de tiempo y esfuerzo.

También cabe la posibilidad de que otros refinamientos se presentaran por sí mismos durante la etapa de retroalimentación.

Además, las acciones y procesos de intervención serían revisados y muy seguramente refinados también, con base en la experiencia y el conocimiento adquiridos en el despliegue y la retroalimentación iniciales.

Finalmente, el modelo y las acciones de intervención refinadas se desplegarían de nuevo, manteniendo el proceso de retroalimentación a lo largo de todo el programa de intervención.

ANEXO. Enfoques TOP-DOWN y BOTTOM-UP

Los proyectos de datos generalmente se organizan de 2 formas:

- Top-down
  - Comenzando con la pregunta del negocio.
- Bottom-up
  - Comenzando con los datos y trabajando hasta obtener información.

Es un enfoque más efectivo que el otro?
Tomemos como ejemplo que queremos saber qué transacciones son fraudulentas.

1. Top-down

En este enfoque, se comienza con una creencia inicial de que las transacciones fraudulentas son un problema serio para la empresa.
Para testear la hipótesis necesita datos, por lo que elige perfiles representativos de clientes y sigue cuidadosamente sus transacciones para estar seguro de cuáles son fraudulentas y cuáles no.
Luego de coleccionar suficientes datos y enriquecerlos con datos de terceros (como información acerca de las ubicaciones donde se realizaron las transacciones, la fecha, la hora, si fue en vacaciones, etc), puede:

- Estimar el costo del fraude entre estos grupos y luego extrapolarlo a su negocio completo.
- Construir un modelo para predecir si una transacción es fraudulenta, y desplegar el modelo para alertar sobre transacciones sospechosas para investigación intermedia.

2. Bottom-up

Aquí comienza con todos los datos y se pregunta si hay relaciones interesantes en ellos.
Luego de realizar varias etapas de estadística descriptiva, llega a un mapa geo-temporal que sugiere hay un pico en las compras de clientes luego de las 5 p.m. dentro de las 10 millas de la dirección de facturación del cliente.

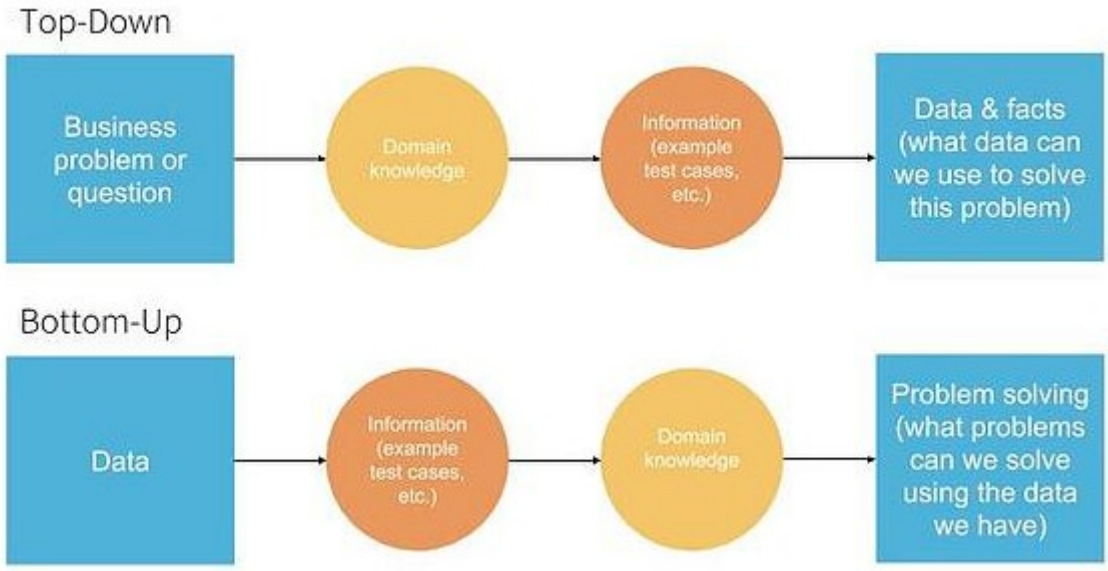
3. Cómo los enfoques son diferentes (y aún así complementarios)

El método bottom-up tiende a ser no estructurado y exploratario. Permite que los datos lleven al resultado.
El método top-down define el problema a resolver y construye un experimento para resolverlo.

Así, el enfoque top-down está más alineado con el método científico; sin embargo puede ser costoso diseñar y llevar a cabo un experimento apropiado. Además se necesita suficiente información para comenzar. El enfoque bottom-up hace un buen uso de los datos disponibles, pero puede llevarlo al proverbial viaje de pesca y debe tener cuidado de no perseguir resultados falsos.

Al final, ningún enfoque es el mejor en todos los casos. A menudo son complementarios, con un enfoque llevando al otro de forma cíclica. Por ejemplo:

- Top-down a bottom-up
  - Luego de formular la creencia inicial de que el fraude es un problema para la compañía, puede usar los datos disponibles pata realizar un chequeo de sanidad para obtener una estimación aproximada de la escala del problema y si el experimento completo es necesario.
- Bottom-up a top-down
  - Luego de que el análisis exploratorio sugiere cuándo y dónde ocurre el incremento en las compras de un cliente, puede confirmar esto con un análisis top-down más riguroso.


Top-Down
Business problem or question
Domain knowledge
Information (example test cases, etc.)
Data & facts (what data can we use to solve this problem)
Bottom-Up
Data
Information (example test cases, etc.)
Domain knowledge
Problem solving (what problems can we solve using the data we have)