Metodología Microsoft

¿Qué es el Proceso de ciencia de datos en equipo (TDSP)?

El proceso de ciencia de datos en equipo (TDSP) es una metodología de ciencia de datos ágil e iterativa para proporcionar soluciones de análisis predictivo y aplicaciones inteligentes de manera eficiente. TDSP ayuda a mejorar la colaboración y el aprendizaje en equipo al sugerir cómo los roles de equipo funcionan mejor juntos. TDSP incluye procedimientos recomendados y estructuras de Microsoft y otros líderes del sector para ayudar a implementar correctamente iniciativas de ciencia de datos. El objetivo es ayudar a las empresas a que se den cuenta de las ventajas de su programa de análisis.

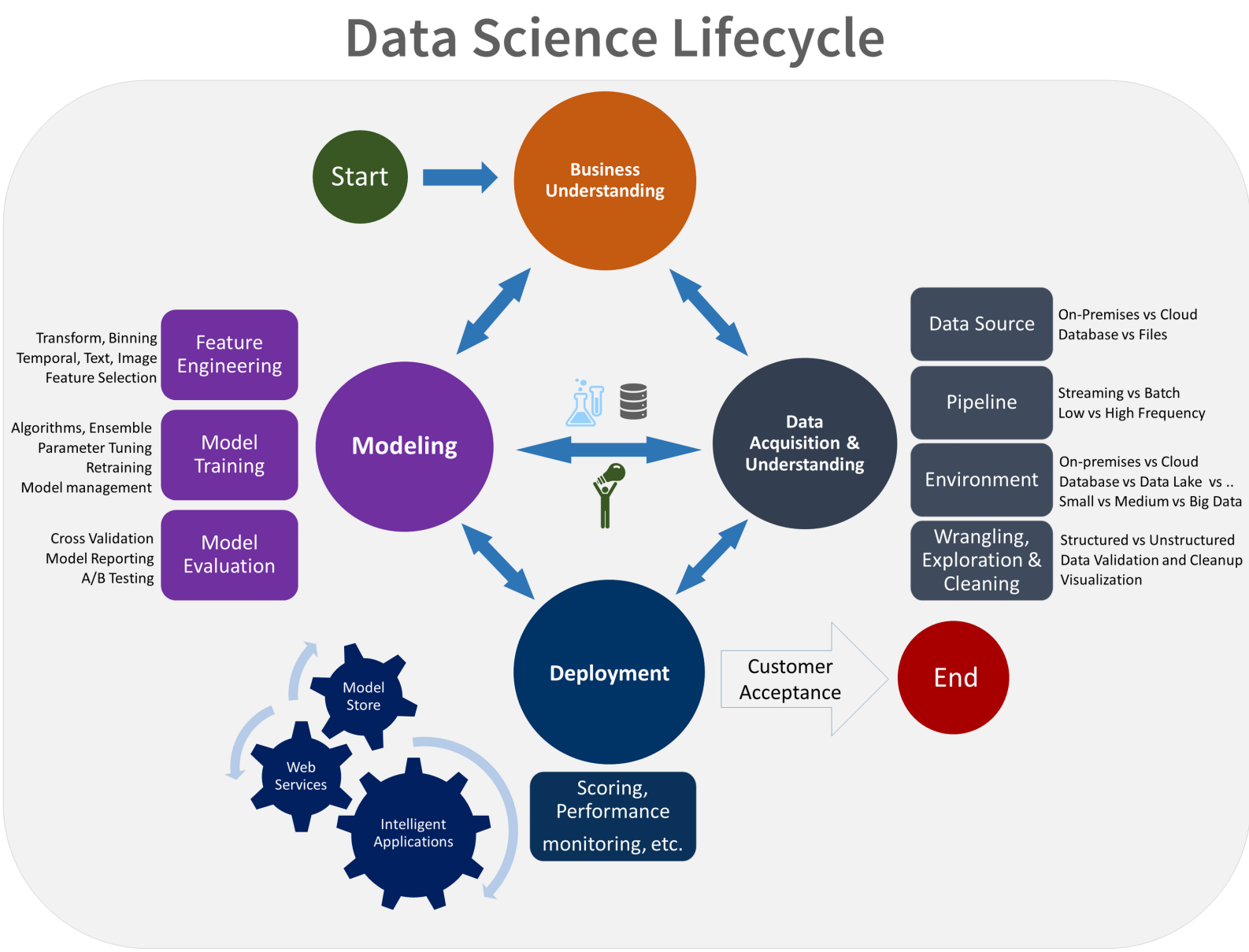
Ciclo de vida de ciencia de datos

El proceso de ciencia de datos en equipo (TDSP) proporciona un ciclo de vida para estructurar el desarrollo de los proyectos de ciencia de datos. En el ciclo de vida se describen todos los pasos que siguen los proyectos correctos.
Aunque esté usando otro ciclo de vida de ciencia de datos, como CRISP-DM, KDD o el proceso personalizado de su organización, puede usar también el TDSP basado en tareas en el contexto de esos ciclos de vida de desarrollo. En un nivel alto, estas distintas metodologías tienen mucho en común.

Este ciclo de vida se ha diseñado para proyectos de ciencia de datos que se enviarán como parte de aplicaciones inteligentes. Estas aplicaciones implementan modelos de aprendizaje o inteligencia artificial de máquina para realizar un análisis predictivo. Los proyectos de ciencia de datos exploratorios o proyectos de análisis improvisados también se pueden beneficiar del uso de este proceso. Pero, en estos casos, puede que algunos de los pasos descritos no sean necesarios.

El ciclo de vida describe las fases principales por las que pasan normalmente los proyectos, a menudo de forma iterativa:

- Conocimiento del negocio
- Adquisición y comprensión de los datos
- Modelado
- Implementación

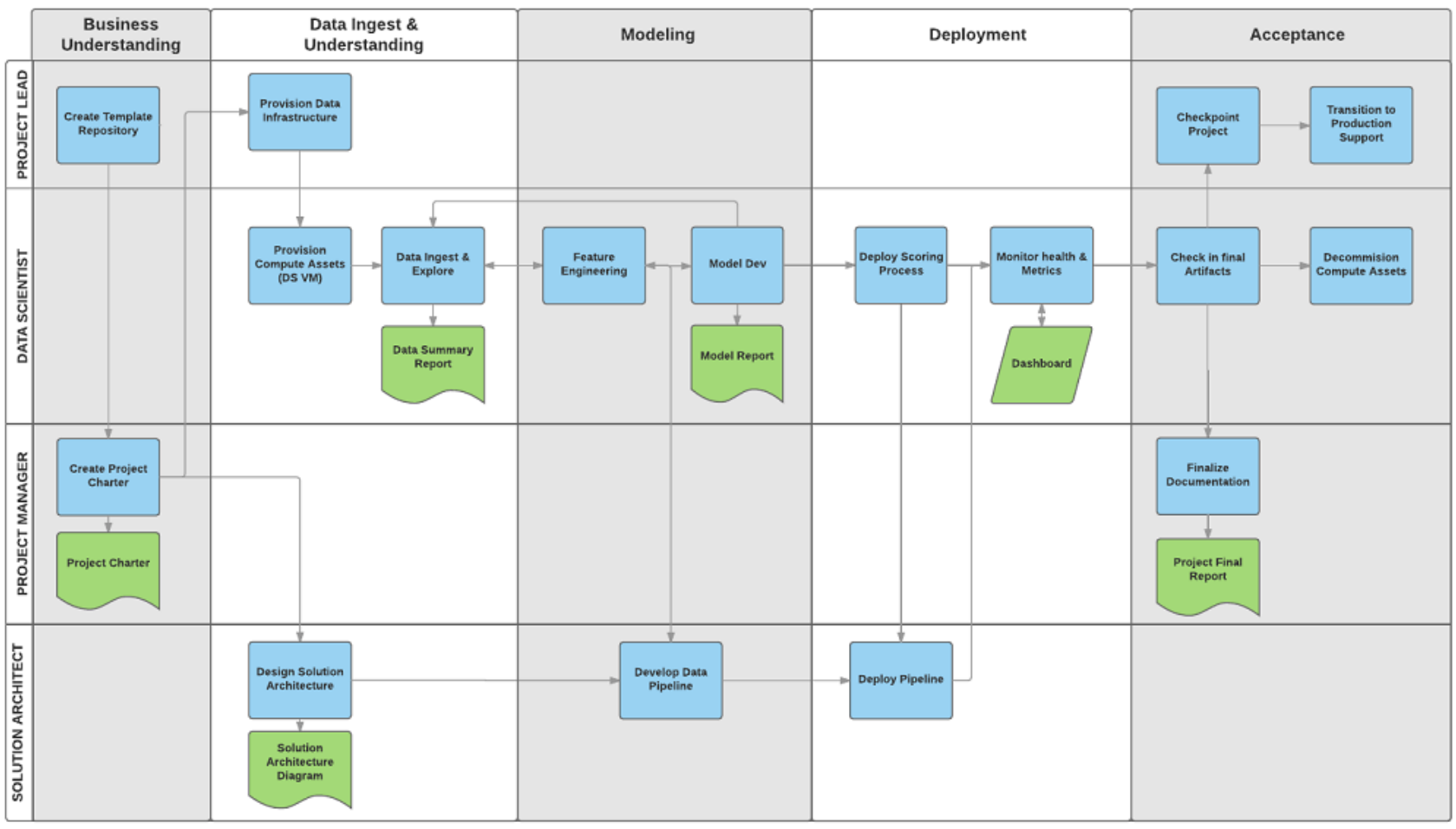


Esta es una representación visual del ciclo de vida del proceso de ciencia de datos en equipo.

En el tema Team Data Science Process lifecycle (Ciclo de vida del proceso de ciencia de datos en equipo) se describen los objetivos, las tareas y los artefactos de documentación de cada fase del ciclo de vida de TDSP. Estas tareas y artefactos están asociados con roles de proyecto:

- Arquitecto de soluciones
- Jefe de proyecto
- Ingeniero de datos
- Científico de datos
- Desarrollador de aplicaciones
- Responsable de proyecto

En el siguiente diagrama se proporciona una vista de cuadrícula de las tareas (en azul) y los artefactos (en verde) asociados con cada fase del ciclo de vida (eje horizontal) de estos roles (eje vertical).



TDSP-roles-and-tasks
Basado en la documentación de Microsoft
https://learn.microsoft.com/es-es/azure/architecture/data-science-process/overview

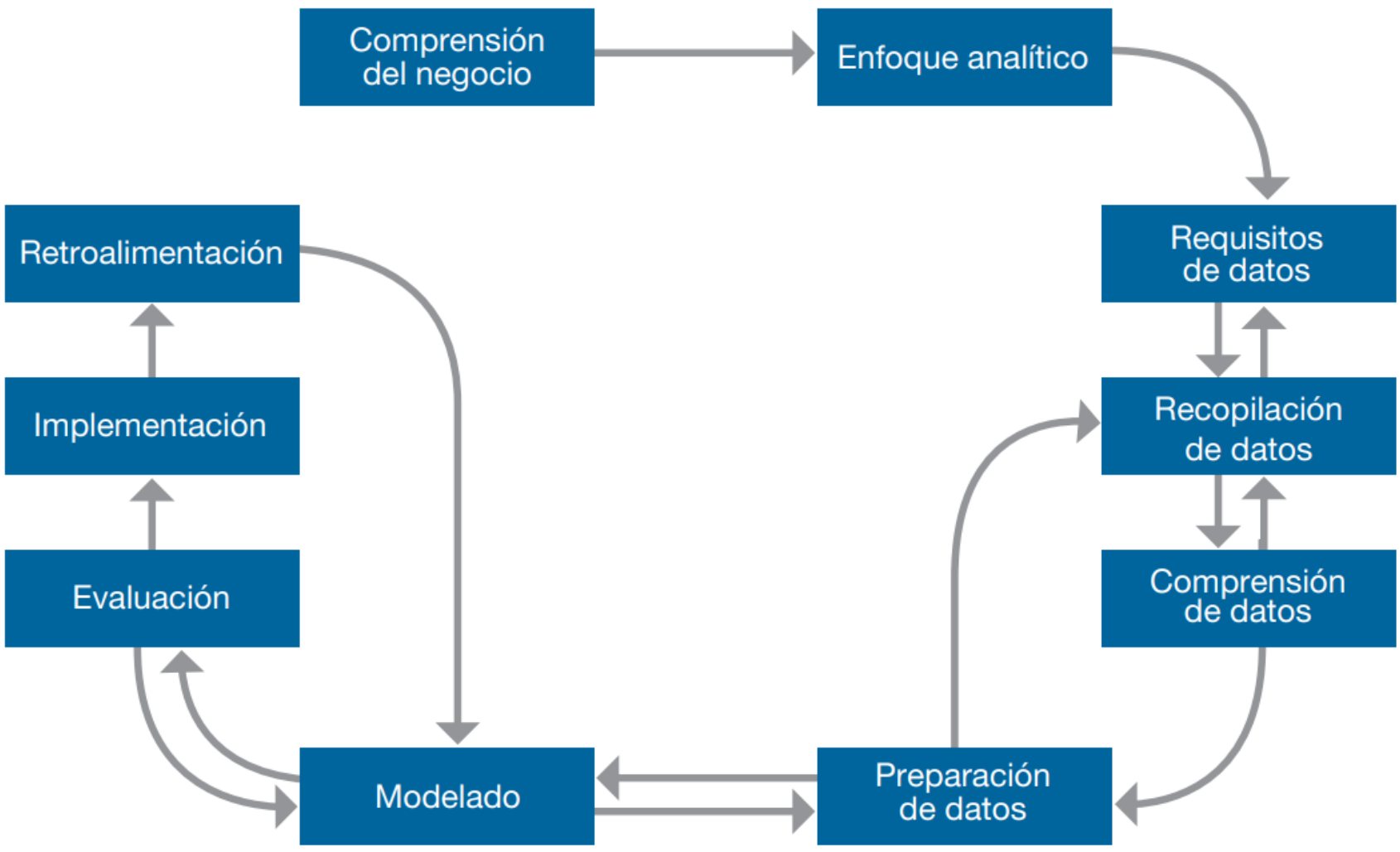
Metodología IBM

Metodología Fundamental para la Ciencia de Datos

En el dominio de la ciencia de datos, resolver problemas y responder preguntas a través del análisis de datos es una práctica estándar. A menudo, los científicos de datos construyen modelos para predecir resultados o para descubrir patrones subyacentes, con la meta de obtener insights. Después, las organizaciones pueden usar estos insights para tomar medidas que mejoren los siguientes resultados. Para analizar los datos y construir modelos existen numerosas tecnologías que evolucionan rápidamente. En un tiempo extraordinariamente corto, han pasado de utilizar escritorios a almacenes que están masivamente en paralelo con enormes volúmenes de datos y funcionalidad analítica en las bases de datos relacionales y de Apache Hadoop.

La analítica de texto en datos no estructurados o semiestructurados se está volviendo cada vez más importante como forma de incorporar a modelos predictivos la percepción y otra información útil de los textos, lo que a menudo conlleva mejoras significativas en la calidad y precisión del modelo. Los enfoques analíticos emergentes buscan automatizar muchos de los pasos de la creación y aplicación de modelos, lo que hace que la tecnología de aprendizaje automático sea más accesible para quienes carecen de profundas habilidades cuantitativas. Además, en contraposición al enfoque "de arriba a abajo" por el que primero se define el problema empresarial y luego se analizan los datos para obtener una solución, algunos científicos de datos pueden usar un enfoque "de abajo a arriba". Con este último enfoque, el científico de datos analiza grandes volúmenes de datos para saber cuál es el objetivo empresarial que pueden sugerir los datos y, luego, aborda ese problema. Dado que la mayoría de los problemas se abordan de manera descendente, la metodología de este documento refleja esa visión.

Una metodología de ciencia de datos de 10 etapas que abarca tecnologías y enfoques A medida que las capacidades de analítica de datos se vuelven más accesibles y prevalentes, los científicos de datos necesitan una metodología fundamental capaz de proporcionar una estrategia de orientación, que sea independiente de las tecnologías, los volúmenes de datos o los enfoques involucrados (vea la Imagen 1). Esta metodología tiene algunas similitudes con las metodologías reconocidas 1-5 para la minería de datos, pero pone el énfasis en varias de las nuevas prácticas en la ciencia de datos, como el uso de grandes volúmenes de datos, la incorporación de la analítica de texto en el modelado predictivo y la automatización de algunos procesos. La metodología consta de 10 etapas que forman un proceso iterativo para el uso de datos para descubrir insights. Cada etapa juega un papel vital en el contexto de la metodología general.



Etapa 1: Comprensión del negocio Todos los proyectos comienzan con la comprensión del negocio. Los promotores de negocios que necesitan la solución analítica desempeñan el papel más importante en esta etapa, al definir el problema, los objetivos del proyecto y los requisitos de la solución desde una perspectiva empresarial. Esta primera etapa sienta las bases para que el problema empresarial sea resuelto con éxito. Para ayudar a garantizar el éxito del proyecto, los promotores deben participar mientras dure el proyecto para proporcionar experiencia en el dominio, revisar los hallazgos intermedios y garantizar que el trabajo siga su curso para generar la solución deseada.

Etapa 2: Enfoque analítico Cuando el problema empresarial se haya establecido claramente, el científico de datos podrá definir el enfoque analítico para resolver el problema. Esta etapa implica expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático, para que la organización pueda identificar las más adecuadas para el resultado deseado. Por ejemplo, si el objetivo es predecir una respuesta como "sí" o "no", el enfoque analítico podría definirse como la construcción, las pruebas y la implementación de un modelo de clasificación.

Etapa 3: Requisitos de datos El enfoque analítico elegido determina los requisitos de datos. Más concretamente, los métodos analíticos a utilizar requieren de determinados contenidos de datos, formatos y representaciones, orientados por el conocimiento en el dominio.

Etapa 4: Recopilación de datos En la etapa inicial de recopilación de datos, los científicos de datos identifican y reúnen los recursos de datos disponibles (estructurados, no estructurados y semiestructurados) y relevantes para el dominio del problema. Por lo general, deben elegir si realizan inversiones adicionales para obtener elementos informativos menos accesibles. Lo mejor puede ser aplazar la decisión de inversión hasta que se sepa más sobre los datos y el modelo. Si hay algunas lagunas en la recopilación de datos, es posible que el científico tenga que revisar los requisitos de datos y recopilar más datos o nuevos datos. Aunque el muestreo y la subdivisión de datos siguen siendo importantes, las plataformas actuales de alto rendimiento y la funcionalidad analítica en la base de datos permiten que los científicos de datos utilicen conjuntos de datos mucho más grandes que contienen gran parte de los datos disponibles, o incluso todos. Al incorporar más datos, los modelos predictivos pueden representar mejor los eventos raros, como la incidencia de una enfermedad o un fallo del sistema.

Etapa 5: Comprensión de datos Después de la recopilación de datos inicial, los científicos de datos suelen utilizar estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar su calidad y descubrir insights iniciales sobre ellos. Para llenar los huecos es posible que sea necesario volver a recopilar datos.

Etapa 6: Preparación de datos Esta etapa abarca todas las actividades para construir el conjunto de datos que se utilizará en la subsiguiente etapa de modelado. Entre las actividades de preparación de datos están la limpieza de datos (tratar con valores no válidos o que faltan, eliminar duplicados y dar un formato adecuado), combinar datos de múltiples fuentes (archivos, tablas y plataformas) y transformar los datos en variables más útiles. Los científicos de datos utilizan un proceso llamado ingeniería de características para crear variables explicativas adicionales, también conocidas como indicadores o características, a través de una combinación de conocimiento en el dominio y de variables estructuradas existentes. Cuando hay disponibles datos en texto, como los registros del centro de atención al cliente o las observaciones de los médicos en forma no estructurada o semiestructurada, la analítica de texto se puede utilizar para derivar nuevas variables estructuradas y, así, enriquecer el conjunto de indicadores y mejorar la precisión del modelo. La preparación de datos suele ser el paso más largo de los proyectos de ciencia de datos. En muchos dominios, algunos pasos de la preparación de datos son comunes para problemas diferentes. La automatización anticipada de determinados pasos de la preparación de datos puede acelerar el proceso al minimizar el tiempo de preparación a medida. Gracias al alto rendimiento, los sistemas masivamente paralelos y la funcionalidad analítica que reside donde se almacenan los datos de hoy en día, los científicos de datos pueden preparar los datos de forma más fácil y rápida utilizando conjuntos de datos muy grandes.

Etapa 7: Modelado La etapa de modelado utiliza la primera versión del conjunto de datos preparado y se enfoca en desarrollar modelos predictivos o descriptivos según el enfoque analítico previamente definido. En los modelos predictivos, los científicos de datos utilizan un conjunto de capacitación (datos históricos en los que se conoce el resultado de interés) para construir el modelo. El proceso de modelado normalmente es muy iterativo, ya que las organizaciones están adquiriendo insights intermedios, lo que deriva en ajustes en la preparación de datos y en la especificación del modelo. Para una técnica determinada, los científicos de datos pueden probar múltiples algoritmos con sus respectivos parámetros para encontrar el mejor modelo para las variables disponibles.

Etapa 8: Evaluación Durante el desarrollo del modelo y antes de su implementación, el científico de datos evalúa el modelo para comprender su calidad y garantizar que aborda el problema empresarial de manera adecuada y completa. La evaluación del modelo implica el cálculo de varias medidas de diagnóstico y de otros resultados, como tablas y gráficos, lo que permite al científico de datos interpretar la calidad y la eficacia del modelo en la resolución del problema. Para los modelos predictivos, los científicos de datos usan un conjunto de pruebas, que es independiente del conjunto de capacitación, pero sigue la misma distribución de probabilidad y tiene un resultado conocido. El conjunto de pruebas se utiliza para evaluar el modelo para ajustarlo según las necesidades. A veces, el modelo final también se aplica a un conjunto de validación para realizar una evaluación final. Además, los científicos de datos pueden asignar al modelo pruebas de significancia estadística como prueba adicional de su calidad. Esta prueba adicional puede ser fundamental para justificar la implementación del modelo o para tomar medidas cuando hay mucho en juego, como un costoso protocolo médico suplementario o un sistema crítico para vuelos en avión.

Etapa 9: Implementación Cuando el modelo satisfactorio ha sido desarrollado y aprobado por los promotores del negocio, se implementa en el entorno de producción o en un entorno de pruebas comparable. Por lo general, se implementa de forma limitada hasta que su rendimiento se haya evaluado completamente. Su implementación puede ser tan fácil como generar un informe con recomendaciones, o tan enrevesado como incrustar el modelo en un complejo proceso de puntuación y de flujo de trabajo administrado por una aplicación personalizada. La implementación de un modelo en un proceso operativo empresarial generalmente involucra a grupos, habilidades y tecnologías adicionales dentro de la empresa. Por ejemplo, un grupo de ventas puede implementar un modelo de propensión a la respuesta a través de un proceso de administración de campañas creado por un equipo de desarrollo y administrado por un grupo de marketing.

Etapa 10: Retroalimentación Al recopilar los resultados del modelo implementado, la organización obtiene retroalimentación sobre el rendimiento del modelo y su impacto en el entorno en el que se implementó. Por ejemplo, la retroalimentación puede ser en forma de porcentajes de respuesta a una campaña promocional dirigida a un grupo de clientes que ha sido identificado por el modelo como respondedores de alto potencial. Los científicos de datos pueden analizar esta retroalimentación para ajustar el modelo para mejorar su precisión y utilidad. Pueden automatizar algunos o todos los pasos de la evaluación del modelo y de la recopilación de retroalimentación, el ajuste y la reimplementación del modelo para acelerar el proceso de actualización del modelo para obtener mejores resultados.

Brindar un valor continuo a la organización

El flujo de la metodología ilustra la naturaleza iterativa del proceso de resolución de problemas. Los científicos de datos vuelven frecuentemente a etapas previas para realizar ajustes a medida que van aprendiendo más sobre los datos y el modelado. Los modelos no se crean una vez, se implementan y se dejan en su lugar tal como están; en vez de eso, se mejoran y se adaptan constantemente a las condiciones cambiantes a través de retroalimentación, ajustes y reimplementaciones. De esta manera, tanto el modelo como su trabajo pueden proporcionar un valor continuo a la organización mientras la solución sea necesaria

.

Basado en: https://www.ibm.com/downloads/cas/6RZMKDN8

Metodología AWS

¿Qué es el proceso de la ciencia de datos?

Un problema empresarial suele iniciar el proceso de la ciencia de datos. Un científico de datos trabajará con las partes interesadas del negocio para entender las necesidades del mismo. Una vez definido el problema, el científico de datos puede resolverlo con el proceso que consiste en obtener, depurar, explorar y modelar datos e interpretar los resultados (OSEMN):

Obtener datos

Los datos pueden ser preexistentes, recién adquiridos o un repositorio descargable de Internet. Los científicos de datos pueden extraerlos de las bases de datos internas o externas, del software CRM de la empresa, de los registros del servidor web, de las redes sociales o adquirirlos de terceros de confianza.

Depurar datos

La depuración o limpieza de datos consiste en el proceso de normalizarlos según un formato predeterminado. Incluye la gestión de los datos que faltan, la corrección de errores en estos y la eliminación de datos atípicos. Algunos ejemplos de la depuración de datos son:

- Cambiar todos los valores de fecha a un formato estándar común.
- Corregir las faltas de ortografía o los espacios adicionales.
- Corregir inexactitudes matemáticas o eliminar comas de números grandes.

Explorar datos

La exploración de datos es un análisis preliminar de estos que se utiliza para planificar otras estrategias para su modelado. Los científicos de datos obtienen una comprensión inicial de los datos mediante estadísticas descriptivas y herramientas de visualización de los mismos. A continuación, exploran los datos para identificar patrones interesantes que se puedan estudiar o utilizar.

Modelar datos

El software y los algoritmos de machine learning se utilizan para obtener información más profunda, predecir resultados y prescribir el mejor curso de acción. Las técnicas de machine learning, como la asociación, clasificación y agrupación, se aplican al conjunto de datos de entrenamiento. El modelo podría probarse con datos de prueba predeterminados para evaluar la precisión de los resultados. El modelo de datos se puede ajustar muchas veces para mejorar los resultados.

Interpretar los resultados

Los científicos de datos trabajan junto a los analistas y las empresas para convertir la información de datos en acción. Hacen diagramas, gráficos y tablas para representar tendencias y predicciones. La síntesis de datos ayuda a las partes interesadas a comprender y aplicar con eficacia los resultados.