
Universidad
Tecmilenio®

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Reducción de la dimensión

Introducción

La cantidad de datos que se crean cada segundo está creciendo a un ritmo exponencial, resultando en una explosión de datos. Asimismo, se espera que este crecimiento sea cada vez mayor, representando uno de los recursos más valiosos en la actualidad.

El objetivo principal del aprendizaje no supervisado es crear representaciones compactas de los datos, detectando su estructura intrínseca y oculta para aplicaciones en visualización de datos y análisis exploratorio de los mismos.

La necesidad de analizar grandes volúmenes de datos multivariados para tareas plantea el problema fundamental de descubrir representaciones compactas de datos de alta dimensión. Esta es la principal motivación detrás del renovado interés en formular el problema de la reducción de dimensionalidad y el análisis de componentes principales (ACP en español y PCA en inglés), ya que representa un enfoque clásico de reducción de dimensionalidad lineal no supervisada que se limita a datos lineales en términos de efectividad.

Explicación

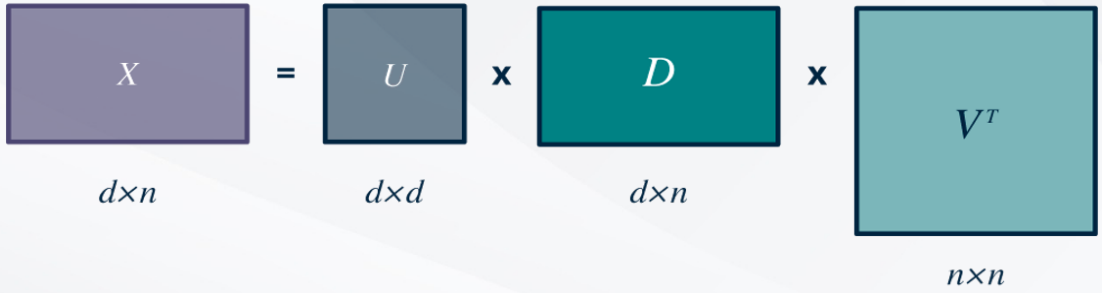
El **análisis de componentes principales** (ACP) es un método de extracción de características para la reducción de dimensionalidad, representando una de las técnicas de reducción de dimensionalidad más populares. El objetivo consiste en reducir la cantidad de características del conjunto de datos (dimensionalidad del conjunto de datos) y preservar la máxima información posible del conjunto de datos original al mismo tiempo.

El **ACP** resuelve este problema combinando las variables de entrada para representarlo con menos variables ortogonales (no correlacionadas) que capturan la mayor parte de su variabilidad.

Explicación

Cuando se trabaja con ACP o cualquiera de sus variantes, se utiliza una matriz de entrada estandarizada. Por lo tanto, *X* representa la matriz de datos de entrada estandarizada, a menos que se especifique lo contrario.

La **descomposición en valores singulares** (SVD por sus siglas en inglés) permite resolver rápidamente el problema de autodescomposición de una manera eficiente en términos computacionales.



La figura anterior representa el caso en donde el número de puntos de datos es mayor que la dimensionalidad de cualquier punto de datos (n > d). Este método permite encontrar fácilmente la solución para el ACP, utilizando la descomposición de valores singulares en la matriz de entrada X.

Explicación

A continuación, se enlistan las principales ventajas y limitaciones al aplicar el método de ACP.

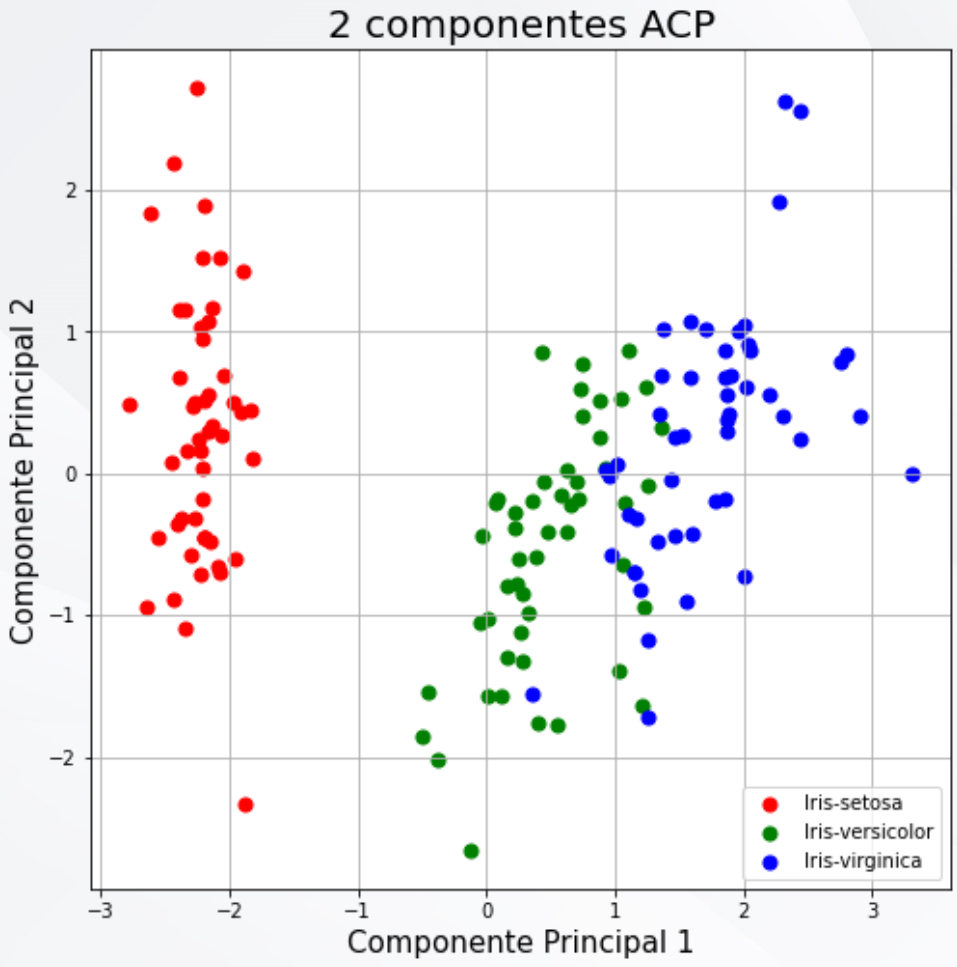
Ventajas	Limitaciones
• Se eliminan las funciones correlacionadas después de implementar ACP en el conjunto de datos. Asimismo, todos los componentes principales son independientes entre sí y no existe correlación entre ellos. • El rendimiento de los algoritmos mejora, ya que el ACP es una técnica que puede ayudar a acelerar los algoritmos de aprendizaje automático al eliminar las variables correlacionadas que no contribuyen a la toma de decisiones. • Se reduce el sobreajuste. • Se mejora la visualización.	• Las variables independientes se vuelven menos interpretables. • Los componentes principales pueden sesgarse si los datos no se estandarizan antes, conduciendo a resultados falsos. • Si no se selecciona el número de componentes principales con cuidado, existe la posibilidad de perder información. • El rendimiento del modelo se reduce si los conjuntos de datos tienen una correlación de características muy baja o nula.

Explicación

La visualización de datos en dos o tres dimensiones no representa un desafío. Sin embargo, incluso un conjunto de datos de flores del tipo iris, que se usa en el siguiente ejemplo, es de cuatro dimensiones.

En la figura se muestra la visualización de datos del conjunto de flores tipo iris, basada en lenguaje Python

El conjunto de datos de iris está dentro de la librería Scikit-learn. El objetivo consistió en importar el conjunto de datos de iris, por lo que, mediante ACP, se redujeron los datos de cuatro dimensiones a dos o tres, permitiendo trazar y comprenderlos de una mejor manera.



Explicación

En la siguiente tabla se mencionan varias aplicaciones del ACP en distintos campos.

Aplicaciones de ACP			
Neurociencia	**Finanzas cuantitativas**	**Reconocimiento facial**	**Otras aplicaciones**
• Una técnica comúnmente conocida como análisis de covarianza espiga-disparado, utiliza una variante de ACP en la neurociencia para identificar y comprender el comportamiento de las neuronas. • El ACP se utiliza para reducir la dimensionalidad y detectar actividades coordinadas de grandes conjuntos neuronales, determinando variables colectivas y parámetros de orden durante las transiciones de fase en el cerebro.	• El ACP se utiliza para reducir la dimensionalidad en las finanzas cuantitativas. Por ejemplo, si un administrador de fondos tiene 200 acciones en su cartera y desea analizar estas poblaciones cuantitativamente, requerirá una matriz de correlacional de 200 × 200, lo que hace que el problema se vuelva realmente complicado. Sin embargo, si se extraen 10 principales componentes que mejor representan la varianza de los datos, se reduciría significativamente la complejidad del problema.	• EigenFaces (un método clásico en visión artificial) se utiliza para el reconocimiento facial y el ACP forma la base su enfoque, ya que el conjunto de datos para esta tecnología se genera utilizando ACP, reduciendo la complejidad estadística en la representación de imágenes faciales. • Otros investigadores han aumentado la precisión del reconocimiento facial mediante el uso de una combinación de Wavelet, ACP y redes neuronales.	• El ACP se utiliza en datos médicos para mostrar la correlación del colesterol con las lipoproteínas de baja densidad. • También se utiliza en datos HVSR (relación espectral horizontal a vertical) destinados a la caracterización sísmica de áreas propensas a terremotos. • Otra aplicación es la detección y visualización de ataques a redes informáticas. • Detección de anomalías. • Compresión de imágenes.

Actividad

Después de haber estudiado el tema puedes abordar las siguientes cuestiones:

- ¿Qué otro tipo de problemas puedes resolver empleando el algoritmo de análisis de componentes principales?
- ¿Por qué considerarías que ACP corresponde a un método no supervisado?

Cierre

En este tema se abordó una técnica importante del aprendizaje no supervisado: el **análisis de componentes principales**.

Asimismo, se mencionó que el ACP tiene muchas aplicaciones, por ejemplo, como herramienta para el estudio de algunas ciencias médicas, así como en el campo de la visión artificial. Del mismo modo, se enumeraron varias ventajas y limitaciones que pueden presentarse al aplicar este método.

Además, se presentó un ejemplo de implementación en donde pudiste comprobar la efectividad del algoritmo al reducir las dimensiones de un conjunto de datos.

Finalmente, resultaría conveniente preguntarte lo siguiente: ¿qué otro tipo de aplicaciones podrías darle al algoritmo de ACP?

Universidad Tecmilenio

Aprendizaje Automático No Supervisado

Construcción de sistemas completos

Introducción

El estudio de las técnicas de aprendizaje automático no puede estar completo sin comprender adecuadamente las métricas que se utilizan para evaluar el desempeño de un modelo.

Este proceso de validación garantiza que las implementaciones y el uso de dichos modelos en los entornos de producción incorporen la menor cantidad de errores, manteniendo su validez durante un periodo relativamente largo.

Cada método de aprendizaje (supervisado o no supervisado) utiliza un conjunto de métricas específicas, por lo que es de suma importancia reconocer y diferenciar cuáles son las que corresponden a cada tipo para ser capaces de aplicarlas e interpretarlas correctamente.

Explicación

Existen diversas métricas que se utilizan para evaluar los resultados producidos por un modelo de aprendizaje automático, pero su selección puede resultar un poco confusa si no se tiene claridad sobre cuál es la más adecuada para cada tipo de aplicación. En la tabla se muestra un resumen de algunas métricas asociadas a un tipo de uso específico.

Métricas de desempeño		
Aprendizaje supervisado		Aprendizaje no supervisado
Regresión	**Clasificación**	**Agrupamiento**
Métricas de error MSE RMSE MAE MAPE **Métricas R^2** R^2 R^2 ajustado	**Métricas de desempeño** Exactitud Precisión Exhaustividad Especificidad AUC Puntuación F1 **Gráficas de desempeño** Curva ROC Curva precision/recall	Puntuación de silueta Índice Rand Índice Rand ajustado Información mutua Índice Calinski-Harabasz Índice Davies-Bouldin

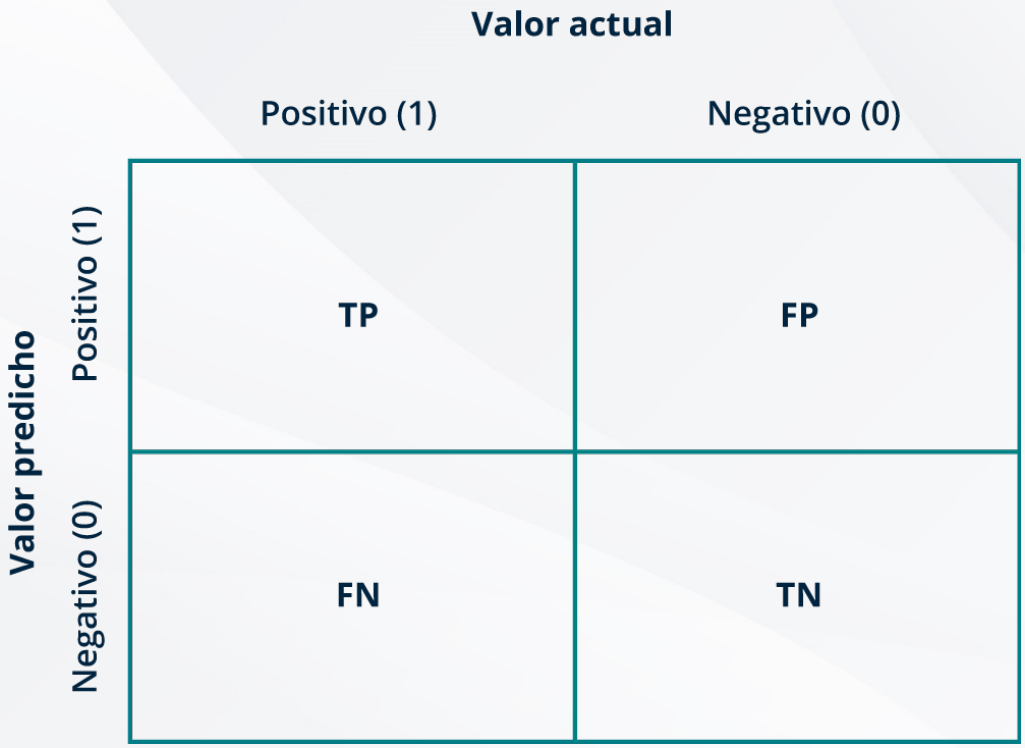
Explicación

Es importante comprender algunos conceptos básicos que se usan de forma regular en la mayoría de las métricas de desempeño. Cada predicción realizada por el modelo puede tomar cuatro formas respecto al rendimiento:

- **Verdadero positivo (TP):** la predicción indica que una muestra será positiva y su etiqueta refleja que es realmente positiva.
- **Verdadero negativo (TN):** en este caso, la predicción asigna como negativa a la muestra y su etiqueta indica que es realmente negativa.
- **Falso positivo (FP):** el resultado de la predicción indica que la muestra es positiva, pero su etiqueta es realmente negativa. En este caso, la muestra se predice falsamente como positiva.
- **Falso negativo (FN):** la predicción sugiere que la muestra es negativa, mientras que su etiqueta es realmente positiva. En este caso, la muestra se predice falsamente como negativa.

Explicación

Matriz de confusión: los valores verdadero positivo, verdadero negativo, falso positivo y falso negativo, generalmente se presentan de manera organizada en un formato tabular llamado matriz de confusión (ver la figura).



Explicación

Exactitud (*accuracy*): según Zucarrelli (2020), la exactitud es la fracción de predicciones correctas que el modelo obtuvo de todas las predicciones. Esto significa que sumamos el número de predicciones correctamente predichas como positivas (TP) o negativas (TN) y lo dividimos entre todas las predicciones realizadas:

$$Exactitud = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

En la vida real, la exactitud varía entre 0.5 y 1, ya que un valor debajo de 0.5 indica que los resultados no son buenos, por lo que es necesario modificar las etiquetas para obtener una mejor predicción.

Explicación

Precisión (precision): es una métrica que se utiliza para superar las limitaciones de la exactitud, ya que indica qué proporción de las predicciones positivas fue realmente correcta. Para lograr su función, considera la relación existente entre las muestras acertadamente predichas como positivas (TP) y el total de predicciones positivas, correctas o incorrectas (TP, FP):

$$Precisión = \frac{TP}{(TP + FP)}$$

Exhaustividad (recall): esta métrica también se conoce como sensibilidad y, de manera similar a la precisión, tiene como objetivo medir qué relación de correspondencia de positivos reales se identificó correctamente.

$$Exhaustividad = \frac{TP}{(TP + FN)}$$

Explicación

Especificidad (specificity): según Zucarrelli (2020), la especificidad se considera como la métrica simétrica a la exhaustividad. Por ende, tiene como objetivo medir qué proporción de negativos reales se identificó correctamente.

Esto lo hace dividiendo las muestras negativas predichas correctamente entre el número total de negativas, señaladas correctamente como negativas o reconocidas incorrectamente como positivas (TN, FP).

$$Especificidad = \frac{\boldsymbol{TN}}{(\boldsymbol{FP} + \boldsymbol{TN})}$$

Explicación

De la misma forma en que se abordaron las aplicaciones de clasificación, en este caso es necesario considerar un concepto de gran relevancia: el **error**. El error es una medida bastante intuitiva que necesita poca definición formal, ya que es un concepto ampliamente conocido. En términos de aprendizaje automático, cuando se menciona al error se refiere a la diferencia que existe entre el valor verdadero y el valor predicho. A continuación, se presentan las principales métricas que involucran este parámetro.

Error cuadrático medio (MSE): mide el promedio de los errores al cuadrado. Básicamente, calcula la diferencia entre el valor estimado y el actual, elevando ese resultado al cuadrado y calculando su promedio:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

El procedimiento de elevar al cuadrado garantiza que esta métrica solo pueda asumir valores positivos. Además, debido a la aleatoriedad y al ruido asociado en la mayoría de los procesos, suele tener un valor no nulo.

Explicación

Raíz del error cuadrático medio (RMSE): de manera similar al error cuadrático medio, esta métrica calcula el promedio de los errores cuadráticos en todas las muestras, pero, en este caso, tomando la raíz cuadrada del resultado:

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}{N}}$$

La RMSE proporciona una medida de error indicada en la misma unidad que la variable objetivo, por lo que esta propiedad facilita la interpretación de resultados que no tienen una representación válida en el dominio de la unidad exponencial (por ejemplo, el valor de una casa se indica en pesos, sin embargo, no tienen mucho sentido representarlo en *pesos*^{2}).

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n}$$

Explicación

Error absoluto medio (MAE): el error absoluto medio no toma el cuadrado de los errores. En este caso, simplemente calcula el valor absoluto de los errores y luego promedia estos valores:

A diferencia del MSE, el MAE no penaliza los errores más grandes sobre los más pequeños porque no aplica el cuadrado a los errores. Otra ventaja es que el MAE no modifica las unidades (de manera similar a la RMSE), lo que facilita la interpretación de los resultados.

Error de porcentaje absoluto medio (MAPE): el error de porcentaje absoluto medio mide en porcentaje el error entre los valores reales y predichos. Esta métrica se calcula de manera similar al MAE, pero incluyendo la división por el valor real:

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

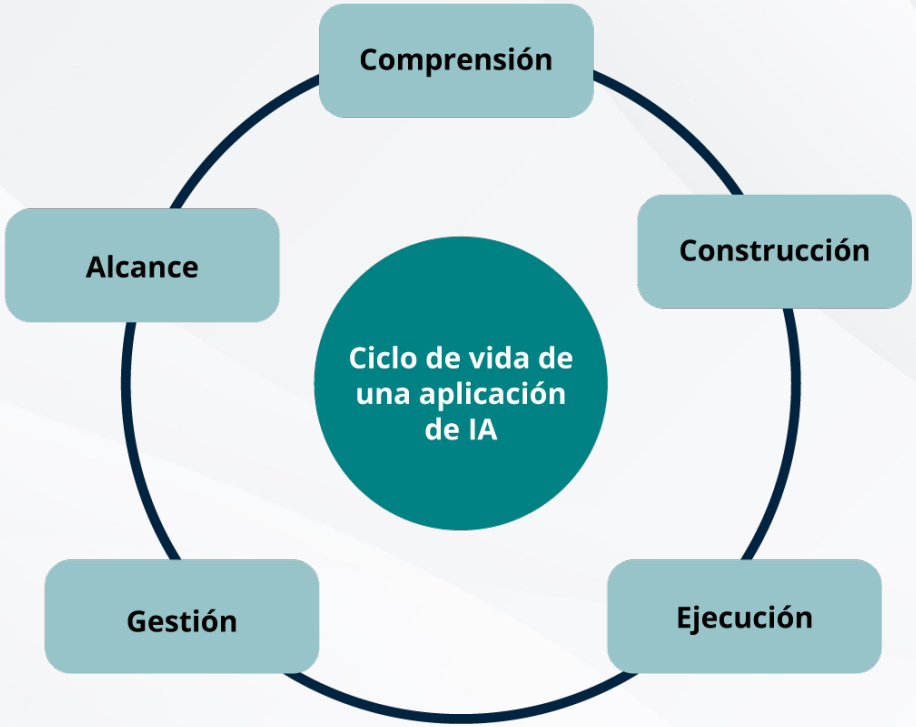
Al expresar el error como un porcentaje, se puede comprender mejor qué tan equivocadas están las predicciones en términos relativos.

Explicación

A medida que las organizaciones progresan en la adopción de la inteligencia artificial dentro de sus operaciones, enfocan sus esfuerzos en contratar personal capacitado en ciencias de datos para gestionar sus fuentes de información y desarrollar los algoritmos, marcos de trabajo y técnicas de aprendizaje automático que puedan sacarles un máximo de beneficio.

La necesidad de poner en funcionamiento las soluciones de IA ha devenido en un cambio de enfoque, el cual incluye la gestión completa del ciclo de vida de un proyecto llamado **operacionalización de la IA**.

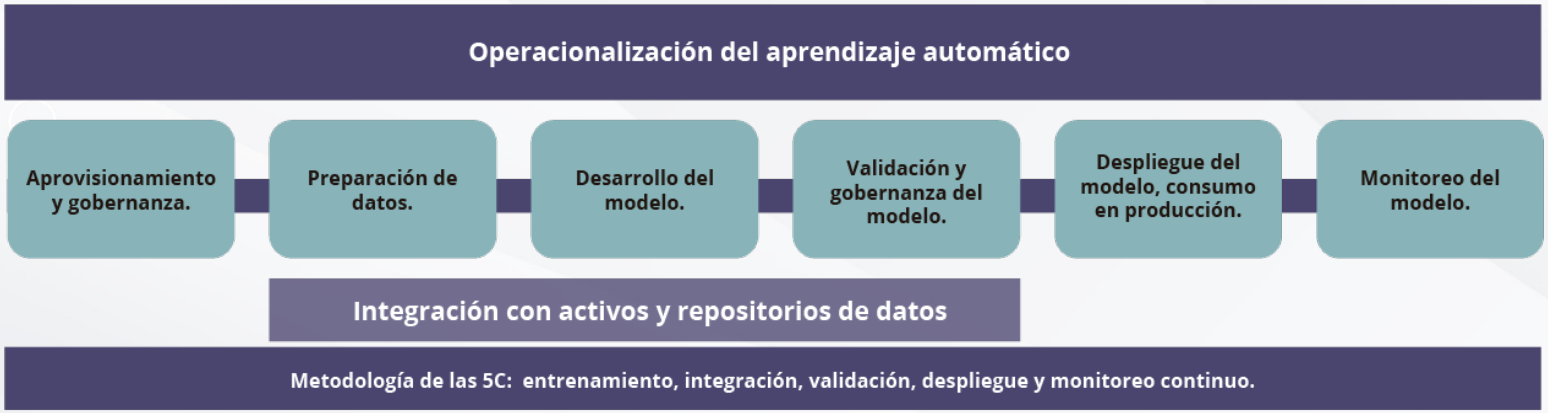
Este diagrama ilustra las consideraciones para implementar la operacionalización de la IA dentro de una organización



Explicación

Por otra parte, la **operacionalización del aprendizaje automático** (MLOps) es una rama de la operacionalización de la inteligencia artificial que se encarga específicamente de trabajar con los modelos de aprendizaje automático para el uso en producción, y garantizar la obtención del máximo valor comercial de estos.

El siguiente diagrama muestra los pasos que componen un canal de operacionalización (ML pipeline) de un modelo en producción:



A diferencia de DevOps, MLOps no solo se encarga de la integración y la implementación continuas, sino que también cubre el entrenamiento, la validación y el monitoreo continuo.

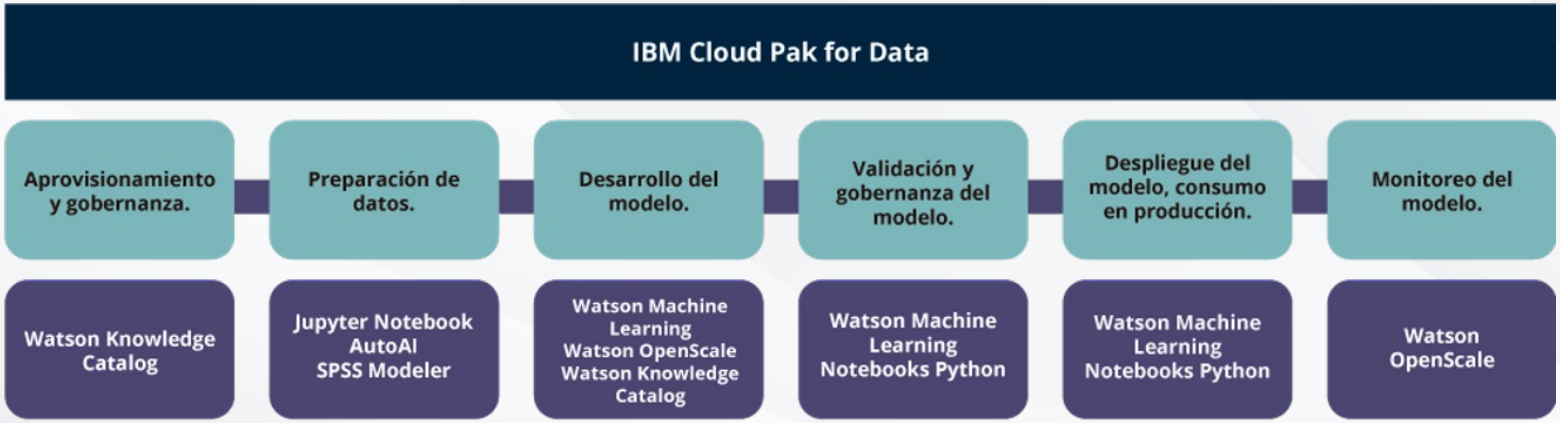
Explicación

En el mercado existen varias plataformas y marcos de trabajo (*frameworks*) que permiten implementar la operacionalización de los modelos de aprendizaje automático a diferentes niveles. Entre las propuestas comerciales más destacadas podemos mencionar las siguientes: **IBM Cloud Pak For Data, Microsoft Azure, Google Cloud** y **Amazon SageMaker**.

Por otro lado, de la mano del software libre también existen varias herramientas relevantes, entre las cuales destacan las siguientes: **mlFlow, Kubeflow, Apache Airflow** y **Open Data Hub.**

Explicación

Cloud Pak For Data es la plataforma de extremo a extremo de IBM que incluye las capacidades funcionales y no funcionales requeridas para implementar la operacionalización del aprendizaje automático de manera integral. En la figura se muestra el proceso de integración de los diversos componentes dentro de Cloud Pak para diseñar un pipeline de MLOps:



Actividad

Piensa en los conceptos vistos durante el tema y responde las siguientes preguntas:

- Considera algún proyecto de aprendizaje automático que hayas implementado antes y aplica las métricas vistas durante el tema.
- Evalúa los resultados e interprétalos.
- Elige algún problema nuevo y plantéate cómo implementarías Cloud Pak For Data en el diseño de la solución.

Cierre

En este tema se abordaron dos aspectos clave para todo experto en **IA** y **aprendizaje automático**. El primero está directamente relacionado con las diversas métricas que se pueden utilizar para evaluar el desempeño de un modelo y tomar las decisiones adecuadas para sacarles un máximo provecho o simplemente descartar como válido.

Por su parte, las **plataformas de software** usan estas mismas métricas para realizar sus comprobaciones internas, por lo que son la piedra angular de la mayoría de los sistemas de automatización de IA. El segundo tema es la **operacionalización**, la cual nos aporta los conocimientos necesarios para llevar los modelos que se construyen (como experimentos de laboratorio) a un entorno real, donde se entrenan continuamente y mejoran sus resultados al tiempo que son utilizados por miles de usuarios.

La **explicabilidad** es uno de los factores que la mayoría de las empresas en todos los niveles necesitan implementar en sus soluciones inteligentes, por tanto, como experto en aprendizaje automático, tienes la responsabilidad de dominarla y usarla de manera adecuada y ética.

Tecmilenio no guarda relación alguna con las marcas mencionadas como ejemplo. Las marcas son propiedad de sus titulares conforme a la legislación aplicable, estas se utilizan con fines académicos y didácticos, por lo que no existen fines de lucro, relación publicitaria o de patrocinio.
