Técnicas de limpieza en modelos de aprendizaje automático

Breve descripción:

La inteligencia artificial (IA) está permeando en muchas área y sectores, esto ha impulsado nuevas ideas de negocio en las organizaciones y empresas. De ahí que la calidad de los datos sea crucial para el entrenamiento de modelos y así obtener predicciones precisas. Si los datos son desordenados o incorrectos, el modelo también lo será.

Julio de 2025

Tabla de Contenido

Introducción

La sociedad actual está caracterizada por el predominio de los datos. Desde la optimización de las cadenas de suministro hasta el diagnóstico médico personalizado y la creación de arte generativo, la inteligencia artificial (IA) se ha establecido como el motor de la innovación más disruptiva de nuestro tiempo. No obstante, detrás de cada algoritmo "inteligente", de cada predicción sorprendentemente precisa y de cada sistema autónomo, subyace un pilar fundamental que a menudo se subestima: la calidad de los datos.

El adagio "basura entra, basura sale" nunca ha sido más pertinente. Un modelo de inteligencia artificial, por más sofisticado que sea, es inherentemente un reflejo del universo de datos con el que ha sido entrenado. Si dicho universo es caótico, incompleto o impreciso, el modelo resultante será, en el mejor de los casos, mediocre y, en el peor, peligrosamente erróneo.

En este componente, se busca darle al aprendiz el conocimiento necesario para que el éxito de la aplicación de la inteligencia artificial que desarrolle, no dependa únicamente de la elección del algoritmo más avanzado, sino de la maestría en el arte y la ciencia de preparar los datos. El preprocesamiento de datos no es un mero trámite preliminar; es el conjunto de etapas estratégicas que transforman datos crudos y ruidosos en un activo de alto valor, listo para ser explotado por el aprendizaje automático.

Partiendo de lo anterior, se invita a que acceda al siguiente video, el cual relaciona la temática a tratar durante este componente formativo:

Video 1. Técnicas de limpieza en modelos de aprendizaje automático

Enlace de reproducción del video

Video 1. Síntesis del video: Técnicas de limpieza en modelos de aprendizaje automático
En el mundo actual, los datos son el nuevo oro... pero ¿sabe cómo transformar datos crudos en inteligencia artificial que realmente funcione? El proceso comienza con la limpieza de datos, que consiste en identificar y corregir errores como valores faltantes, registros duplicados o inconsistencias. Aplicar técnicas adecuadas en esta etapa mejora la precisión y robustez de los modelos. Una vez depurados los datos, es posible reducir su complejidad mediante la reducción de dimensionalidad. Este procedimiento permite conservar la información

esencial del conjunto de datos original, eliminando variables redundantes o irrelevantes. Técnicas como el Análisis de Componentes Principales (PCA) ayudan a simplificar el modelo sin comprometer su efectividad.

El enfoque continúa con el aprendizaje automático, cuyo propósito es construir sistemas capaces de identificar patrones y tomar decisiones a partir de los datos. Existen diversas técnicas, entre ellas los algoritmos supervisados, no supervisados y de aprendizaje por refuerzo. La elección del algoritmo adecuado depende del tipo de problema, la naturaleza de los datos y los objetivos de análisis.

Para cerrar el ciclo, es fundamental evaluar el rendimiento del modelo. En tareas de clasificación, se utilizan métricas como la precisión, la sensibilidad y la matriz de confusión. Para modelos de regresión, se aplican indicadores como el error cuadrático medio o el coeficiente de determinación.

Dominar estas técnicas garantiza modelos más eficientes, interpretables y aplicables a problemas reales.

1. Limpieza de datos

El campo del análisis de datos y el aprendizaje automático está experimentado un crecimiento exponencial, lo que les ha permitido consolidarse como elementos esenciales para la toma de decisiones, basadas en datos. La transformación de datos en bruto en información relevante, implica un proceso que incluye la limpieza de datos, la reducción de dimensionalidad, la implementación de algoritmos y la evaluación de los modelos generados.

En este tema se abordará la tarea esencial de la limpieza de datos, aprendiendo a identificar y corregir los errores que plagan casi todos los conjuntos de datos del mundo real.

1.1 Concepto

A continuación, se detalla lo expresado por algunos autores:

- **Dasu & Johnson (2003)**

  La limpieza de datos, conocida en inglés como `data cleaning`, constituye un proceso fundamental en la ciencia de datos. Este proceso implica la identificación, corrección o eliminación de datos "sucios", es decir, registros incorrectos, corruptos, incompletos, irrelevantes o duplicados dentro de un conjunto de datos. Lejos de ser una tarea trivial, a menudo representa la fase más demandante en términos de tiempo en un proyecto de aprendizaje automático, llegando a ocupar entre el 60 % y el 80 % del tiempo total de un especialista.

- **Rahm & Do (2000)**

  El concepto se define formalmente como un conjunto de operaciones destinadas a resolver anomalías de datos, que pueden originarse por diversas causas, como errores humanos durante la entrada de datos, fallos en los sensores de recolección, problemas durante la transmisión o el almacenamiento, o la integración de múltiples fuentes de datos con esquemas y formatos diferentes.

  En esencia, la limpieza de datos es un proceso de aseguramiento de la calidad. Su objetivo es maximizar la consistencia, precisión, completitud y uniformidad de los datos para que representen fielmente la realidad que se pretende modelar. Se podría utilizar la siguiente analogía para explicar la limpieza de datos.

Se tienen invitados a cenar en el hogar y antes de cocinar el plato favorito de ellos, se debe alistar los ingredientes. Hay que pensar en la limpieza de datos como la preparación de ingredientes antes de cocinar. Por ellos, un chef no usaría ingredientes vencidos, sucios o en mal estado para preparar un plato gourmet. De la misma manera en la construcción de un modelo predictivo fiable, no se pueden utilizar datos contaminados.

Asimismo, la limpieza de datos requiere una combinación de habilidades técnicas y conocimiento del dominio para identificar y corregir anomalías. Este proceso puede implicar la detección y eliminación de valores atípicos, el manejo de valores faltantes y la estandarización de formatos. Así como un chef experto, sabe cómo seleccionar y preparar los mejores ingredientes, un científico de datos competente debe ser capaz de

refinar y optimizar los datos para obtener los mejores resultados en sus análisis y modelos.

El no darle toda la importancia y atención necesaria en esta etapa, puede verse reflejada en un rendimiento poco óptimo del modelo, independientemente de lo avanzado que sea el algoritmo utilizado.

1.2 Tipos de errores

Los conjuntos de datos a menudo contienen diversos tipos de errores que deben ser identificados y corregidos durante el proceso de limpieza. Estos errores pueden clasificarse en varias categorías:

- **Datos faltantes**

  Valores ausentes en registros debido a problemas de captura o transmisión, que pueden introducir sesgos si no se manejan adecuadamente (Little & Rubin, 2019). Estos valores suelen representarse como NULL, N/A, NaN (Not a Number) o simplemente mediante una celda vacía. Las causas de estas representaciones son diversas: un usuario que se niega a proporcionar su edad, un sensor que experimenta un fallo temporal o un error en el proceso de extracción de datos.

- **Datos duplicados**

  Registros repetidos que distorsionan las estadísticas y pueden sesgar los resultados de modelos predictivos (Rahm & Do, 2000). Ocurren cuando un mismo registro o entidad aparece más de una vez en el conjunto de datos. Esto puede suceder al combinar datos de diferentes fuentes o por errores en la recolección. Pueden sesgar gravemente los análisis y el entrenamiento, dando un peso indebido a las observaciones repetidas.

- **Datos incorrectos o inválidos**

  Este término se refiere a valores que, aunque presentes, son evidentemente erróneos. Esto incluye errores tipográficos (por ejemplo, "Colobia" en lugar de "Colombia"), valores que se encuentran fuera de un rango plausible (como una edad de 350 años) o datos que no cumplen con un formato estándar (como un correo electrónico sin el símbolo "@").

- **Inconsistencias estructurales y de formato**

  Este tipo de error se relaciona con la falta de uniformidad. (Aggarwal, 2017). Por ejemplo, en una columna de "país", se podría encontrar "Colombia", "COL" y "Co.", refiriéndose todos a la misma nación. Del mismo modo, las fechas pueden estar en formatos dispares (DD/MM/AAAA, YYYY-MM-DD, Mon, Day, Year) o las unidades de medida pueden estar mezcladas (kilogramos y libras en la misma columna).

- **Datos sesgados**

  Se caracterizan por una alta varianza o una distribución no representativa, lo que puede llevar al desarrollo de modelos que favorecen ciertos resultados sobre otros. El sesgo presente en los datos puede perpetuarse en los modelos y afectar la equidad de las decisiones.

- **Datos irrelevantes**

  Se refiere a la información que no contribuye de manera significativa al análisis o que puede desviar la atención del objetivo principal. La identificación y eliminación de estos datos, puede mejorar la eficiencia y precisión de los modelos.

1.3 Técnicas

Una vez se han identificado los tipos de errores, se aplican técnicas específicas para subsanarlos. La elección de la técnica adecuada es un acto de equilibrio que depende del contexto del problema, el tipo y la cantidad de datos sucios, así como el impacto potencial en el modelo final. Para abordar los diversos tipos de errores mencionados, existen varias técnicas de limpieza de datos que pueden implementarse de manera sistemática.

A continuación, se detallan cada uno de ellos:

A. Tratamiento de valores faltantes

En esta técnica se suelen aplicar tres formas:

- **Eliminación**

  La estrategia más básica, consiste en suprimir los registros (filas) que presentan valores faltantes, conocida como eliminación por lista completa. Sin embargo, esta técnica debe aplicarse con precaución, ya que puede conllevar una pérdida considerable de datos si los valores faltantes son numerosos, introduciendo un posible sesgo si los datos no faltan de manera completamente aleatoria (Schafer & Graham, 2002). Una alternativa es la eliminación de la variable (columna), si más de un cierto umbral (por ejemplo, 60 %) de sus valores están ausentes.

- **Imputación mediante medidas de tendencia central**

  Una técnica común, consiste en sustituir el valor faltante con una medida estadística derivada del resto de la columna. Para variables numéricas, se emplea frecuentemente la media o la mediana, siendo esta última generalmente más robusta frente a valores atípicos. En el

caso de variables categóricas, se utiliza la moda; es decir, el valor más frecuente.

- **Imputación avanzada**

  Existen métodos más avanzados para la imputación de datos. La imputación por regresión estima los valores faltantes basándose en otras variables del conjunto de datos. Por otro lado, la imputación mediante K-NN (K-Nearest Neighbors) utiliza la media o la moda de los 'k' registros más similares (vecinos) al registro con el valor faltante para realizar la imputación. Estos métodos suelen ofrecer una mayor precisión, aunque son computacionalmente más costosos (Van Buuren, 2018).

B. Tratamiento de datos duplicados

En esta técnica lo que se busca es identificar y eliminar los registros duplicados. El primer paso consiste en definir qué se considera un duplicado. ¿Debe ser idéntica toda la fila o solo un subconjunto de columnas clave (por ejemplo, ID_cliente y fecha_compra)? Una vez establecidos estos criterios, se pueden emplear funciones de programación para identificar dichos registros y eliminar todas las ocurrencias, excepto una.

C. Tratamiento de datos incorrectos y de formato

Para esta técnica se utilizan dos maneras:

- **Primera**

  Estandarización y normalización.

- **Segunda**

  Aplicación de reglas de validación y expresiones regulares.

  En el caso de la estandarización y normalización de datos, implica la transformación de estos a un formato común. Este proceso incluye la conversión de texto a minúsculas o mayúsculas para evitar duplicados, debido a la capitalización, la estandarización de categorías (por ejemplo, mapear "COL" y "Co." a "Colombia") y la unificación de formatos de fecha y unidades de medida.

D. Reglas de validación y expresiones regulares

Es posible establecer reglas para la validación de datos. Por ejemplo, una columna que contenga edades debe ser un número entero comprendido entre 0 y 120. Las expresiones regulares (`regex`) son herramientas sumamente eficaces para validar y corregir formatos de texto, tales como códigos postales, números de teléfono o direcciones de correo electrónico.

E. Tratamiento de valores atípicos

En esta técnica lo primero es detectar los datos que no se comportan como la gran mayoría de los datos, posteriormente se les da un manejo. Para la detección los métodos visuales como los diagramas de caja (`box plots`), son excelentes para una primera aproximación. Estadísticamente, se pueden usar el rango intercuartílico (IQR) o el Z-score. Un punto de dato se considera atípico si cae por debajo de $Q1-1.5\times IQR$ o por encima de $Q3+1.5\times IQR$, o si su Z-score (número de desviaciones estándar desde la media) excede un umbral, típicamente 3. (Han, Pei & Tong, 2022).

Una vez detectado el dato, se procede a manejarlo; si se determina que el valor atípico es un error, puede ser tratado como un valor faltante. En caso de que se trate de un dato genuino, las opciones incluyen:

- **Transformación**

  Aplicar una transformación matemática, como el logaritmo, a la variable para mitigar el efecto del valor atípico.

- **Winsorización (Capping)**

  Limitar los valores extremos, por ejemplo, reemplazando cualquier valor por encima del percentil 99 con el valor del percentil 99.

- **Eliminación**

  Si el outlier es tan extremo que podría desestabilizar el modelo y no se puede corregir, se puede optar por eliminarlo, documentando siempre la justificación.

En conclusión, se puede afirmar que el proceso de limpieza de datos debe seguir un enfoque estructurado que incluya los siguientes pasos:

- Eliminar observaciones duplicadas o irrelevantes.
- Corregir errores estructurales en los datos.
- Filtrar valores atípicos no deseados.
- Manejar adecuadamente los datos faltantes.
- Realizar validación y control de calidad.

2. Reducción de dimensionalidad

Una vez que se han depurado los datos, asegurando su calidad y consistencia, se enfrenta a un nuevo desafío, sutil, pero de gran impacto: la complejidad. En la era del Big Data, es común trabajar con conjuntos de datos que no solo contienen millones de registros (filas), sino también cientos o miles de características (columnas o dimensiones). Aunque podría parecer que "más datos es siempre mejor", un exceso de dimensiones puede ser perjudicial. Este fenómeno, conocido como la "maldición de la dimensionalidad", puede degradar el rendimiento del modelo, aumentar la complejidad computacional y oscurecer los patrones verdaderamente significativos.

La reducción de dimensionalidad constituye un conjunto de técnicas de ingeniería de características empleadas para disminuir el número de variables de entrada en un conjunto de datos, transformándolo en un espacio de menor dimensionalidad, sin perder una cantidad significativa de información relevante. El objetivo no es simplemente eliminar columnas de manera aleatoria, sino hacerlo de una forma inteligente y estructurada que simplifique el problema para los algoritmos de aprendizaje automático.

Este proceso no solo mejora la eficiencia computacional, sino que también puede revelar estructuras latentes en los datos que no eran evidentes en su forma original de alta dimensionalidad.

Algunas técnicas populares de reducción de dimensionalidad incluyen:

- Análisis de Componentes Principales (PCA).
- Análisis de Componentes Independientes (ICA).
- Métodos de selección de características.

Le elección de la técnica adecuada, dependerá de la naturaleza específica de los datos y los objetivos del análisis, requiriendo una comprensión profunda, tanto del dominio del problema como de las matemáticas subyacentes.

2.1 Concepto

La reducción de dimensionalidad se refiere al proceso de transformar un conjunto de datos con un elevado número de variables originales en otro de menor dimensión, preservando al máximo la información relevante. Esta práctica mejora la eficiencia computacional, mitiga la "maldición de la dimensionalidad" y facilita la visualización de datos. (Jolliffe, 2002)

El concepto central de la reducción de dimensionalidad se fundamenta en la mitigación de la "maldición de la dimensionalidad", un término acuñado por el matemático Richard Bellman (1961). Este fenómeno describe cómo, al incrementarse el número de dimensiones, el volumen del espacio de características crece de manera exponencial, generando varios efectos adversos:

- **Dispersión de datos**

  En un espacio de alta dimensión, los puntos de datos se tornan extremadamente dispersos. La distancia entre cualquier par de puntos tiende a ser similar, lo que dificulta que los algoritmos basados en la distancia, como K-NN, puedan agrupar o diferenciar observaciones de manera eficaz.

- **Aumento del costo computacional**

  Un mayor número de dimensiones implica más parámetros que un modelo debe aprender, lo cual se traduce directamente en tiempos de

entrenamiento más prolongados, mayor consumo de memoria y una infraestructura más costosa.

- **Riesgo de sobreajuste (overfitting)**

  Con una gran cantidad de características, es más probable que un modelo aprenda del "ruido" y de las peculiaridades específicas del conjunto de entrenamiento, en lugar de generalizar los patrones subyacentes. El modelo se vuelve excesivamente complejo y funciona muy bien con los datos que ya ha visto, pero falla al predecir sobre datos nuevos.

- **Multicolinealidad**

  En muchos conjuntos de datos, algunas características están altamente correlacionadas entre sí (por ejemplo, las variables "altura en metros" y "altura en centímetros"). Esta redundancia no aporta nueva información y puede desestabilizar algunos modelos de aprendizaje automático, como la regresión lineal.

La reducción de dimensionalidad aborda estos desafíos al identificar y retener solo las características más relevantes y significativas del conjunto de datos original. Este proceso no solo mitiga los problemas mencionados, sino que también puede revelar estructuras latentes en los datos que no eran evidentes en su forma de alta dimensión. Además, la visualización de datos se vuelve más factible y comprensible cuando se reduce a dos o tres dimensiones, permitiendo a los analistas y científicos de datos obtener insights valiosos que de otra manera podrían pasar desapercibidos.

2.2. Técnicas

Las técnicas para reducir la dimensionalidad se dividen en dos grandes familias, cada una con su propia filosofía y aplicabilidad:

A. Selección de características (Feature selection)

Este enfoque se centra en la identificación y selección de un subconjunto de las características originales, descartando las restantes. La principal ventaja de este método radica en que preserva la interpretabilidad de las variables originales, ya que se retienen únicamente las columnas más informativas. Estas técnicas pueden clasificarse en tres categorías, (Guyon & Elisseeff, 2003):

I. Métodos de filtro

Evalúan la relevancia de las características, mediante el uso de métricas estadísticas, independientemente del modelo de aprendizaje automático que se empleará posteriormente. Estos métodos son rápidos y eficientes desde el punto de vista computacional y hacen parte de ello:

- **Prueba de Chi-cuadrado (X2)**

  Se emplea para determinar si existe una dependencia significativa entre dos variables categóricas. Se utiliza comúnmente para seleccionar las características categóricas más relevantes para una variable objetivo también categórica.

- **ANOVA (Análisis de varianza)**

  La prueba F de ANOVA permite comparar las medias de una variable continua entre dos o más grupos categóricos. Es útil para

seleccionar características numéricas que tienen una relación fuerte con una variable objetivo categórica.

- **Coeficiente de correlación de Pearson**

  Mide la relación lineal entre dos variables numéricas. Se utiliza para identificar características que están altamente correlacionadas con la variable objetivo (en problemas de regresión) y también para identificar y eliminar características redundantes (multicolinealidad).

II. Métodos de envoltura (Wrapper methods)

Estos métodos emplean un algoritmo de aprendizaje automático específico, para evaluar la utilidad de diferentes subconjuntos de características. Consideran la selección de características como un problema de búsqueda, donde cada estado representa un conjunto de variables. Aunque son más precisos que los métodos de filtro, su costo computacional es significativamente mayor e incluyen:

- **Eliminación recursiva de características (RFE)**

  Este es un método iterativo que comienza entrenando un modelo con todas las características disponibles. Posteriormente, se evalúa la importancia de cada característica, por ejemplo, mediante los coeficientes en una regresión o la impureza en un árbol de decisión. La característica menos importante se elimina y el proceso se repite hasta alcanzar el número deseado de características.

- **Selección hacia adelante (Forward selection)**

  Este método inicia sin características y procede a añadirlas una a una, seleccionando aquella que más mejora el rendimiento del modelo, hasta que no se observe una mejora significativa.

III. Métodos integrados (Embedded methods)

Realizan la selección de características, como parte integral del proceso de entrenamiento del modelo. Estos métodos ofrecen un equilibrio entre la precisión de los métodos de envoltura y la eficiencia de los métodos de filtro; además, incluyen:

- **Regularización L1 (Lasso)**

  Los modelos como la regresión Lasso, incorporan una penalización basada en la suma del valor absoluto de los coeficientes del modelo. Esta penalización induce a que los coeficientes de las características menos informativas se reduzcan a exactamente cero, eliminándolas efectivamente del modelo (Tibshirani, 1996).

- **Modelos basados en árboles**

  Algoritmos como Random Forest o Gradient Boosting, calculan de manera inherente la "importancia de las características" durante su construcción. Esta métrica puede emplearse para clasificar las características y seleccionar las más relevantes. Estos métodos pueden utilizarse para seleccionar las características más relevantes y mejorar el rendimiento del modelo. La importancia de las características en los modelos basados en árboles, se calcula generalmente, mediante la reducción en la impureza (como el índice Gini o la entropía) que proporciona cada característica.

Además, estos métodos son robustos frente a características no lineales y pueden capturar interacciones complejas entre variables.

B. Extracción de características

La segunda técnica que se utiliza es la extracción de características, conocida también como proyección de características; la cual convierte los datos originales en un espacio de menor dimensión al crear nuevas variables que son combinaciones de las originales. Este método es particularmente beneficioso, cuando las características originales presentan alta correlación o son redundantes. Entre las técnicas más destacadas se encuentran:

- **El análisis de componentes principales (PCA)**

  Es una de las más empleadas que proyecta los datos en las direcciones de máxima varianza. PCA genera nuevas variables denominadas componentes principales, las cuales son combinaciones lineales de las variables originales. Estos componentes se ordenan, de acuerdo con la cantidad de varianza que explican, lo que permite reducir la dimensionalidad al seleccionar únicamente los componentes más significativos. Con PCA se busca la proyección que mejor representa los datos en términos de mínimos cuadrados. Es particularmente eficaz para la visualización y exploración de conjuntos de datos de alta dimensión, ya que permite identificar fácilmente tendencias, patrones o valores atípicos. Además, reduce la complejidad del modelo y minimiza problemas como la multicolinealidad y el sobreajuste.

- **El análisis discriminante lineal (LDA)**

  A diferencia de PCA se centra en maximizar la separabilidad entre clases. Este es un método de clasificación supervisado en el que se conocen de antemano dos o más grupos, y las nuevas observaciones se clasifican en uno de ellos en función de sus características. LDA estima la probabilidad de que una observación, dado un valor específico de predictores, pertenezca a cada una de las clases de la variable cualitativa.

- **Métodos no lineales**

  Para datos que residen en una variedad no lineal, se emplean técnicas como t-SNE (t-Distributed Stochastic Neighbor Embedding), UMAP (Uniform Manifold Approximation and Projection) y autocodificadores. Estas técnicas son capaces de capturar relaciones más complejas entre las variables en comparación con los métodos lineales como PCA y LDA.

La decisión entre la selección y la extracción de características, así como la técnica específica a emplear, dependerá del objetivo final: si la interpretabilidad es fundamental, se prefiere la selección. En cambio, si el objetivo es la máxima compactación de la información, la extracción puede resultar más eficaz.

3. Aprendizaje automático

El aprendizaje automático o Machine Learning (ML), constituye una subdisciplina de la inteligencia artificial y la informática, centrada en la utilización de datos y algoritmos para emular el proceso de aprendizaje humano, mejorando progresivamente su precisión. La definición canónica, propuesta por Tom M. Mitchell (1997), establece que un programa de computadora aprende de la experiencia E en relación con una clase de tareas T y una medida de rendimiento P, si su desempeño en las tareas T, evaluado por P, mejora con la experiencia E.

En términos más simples, en lugar de programar explícitamente un conjunto de reglas para resolver una tarea, se proporciona a un algoritmo una gran cantidad de datos, permitiéndole "aprender" las reglas y patrones de manera autónoma. Esto permite a los sistemas de ML adaptarse y mejorar continuamente a medida que se exponen a nuevos datos. Los algoritmos de ML pueden identificar patrones complejos y tomar decisiones con mínima intervención humana. Esta capacidad ha revolucionado diversos campos, desde la visión por computadora, hasta el procesamiento del lenguaje natural, abriendo nuevas posibilidades en la resolución de problemas y la automatización de tareas.

3.1. Objetivos

Antes de elegir cualquier técnica, es crucial definir el objetivo del modelo. ¿Qué pregunta de negocio o problema científico se está tratando de resolver? Los objetivos del aprendizaje automático generalmente se dividen en dos categorías principales:

A. **Predicción o inferencia predictiva**

Es el objetivo más común en el ámbito de la modelización. Busca desarrollar un modelo que pueda realizar predicciones precisas sobre datos nuevos y no observados previamente. Este objetivo se subdivide según la naturaleza de lo que se desea predecir en:

- **Clasificación**

  El objetivo es predecir una etiqueta o categoría discreta, abordando preguntas como "¿a qué clase pertenece?" o "¿es esto A o B?". Los problemas de clasificación pueden ser binarios (dos clases, como "es spam" o "no es spam") o multiclase (más de dos clases, como clasificar un animal en "perro", "gato" o "pájaro").

- **Regresión**

  El objetivo es predecir un valor continuo, respondiendo a preguntas del tipo "¿cuánto?" o "¿cuál será el valor?". Ejemplos incluyen la predicción del precio de una vivienda, la temperatura del día siguiente o la demanda de un producto.

B. **Descubrimiento de patrones (o inferencia descriptiva)**

En este contexto, el objetivo no es prever un resultado o valor numérico, sino identificar patrones, estructuras y agrupaciones de interés que sean previamente desconocidas en los datos. Se suelen clasificar en:

- **Agrupamiento (Clustering)**

  El propósito es organizar las observaciones en clústeres o segmentos, de manera que los integrantes de un mismo clúster presenten una

alta similitud entre sí y una marcada diferencia respecto a los integrantes de otros clústeres. Un ejemplo de aplicación sería la segmentación de clientes basada en su comportamiento de compra para desarrollar campañas de marketing personalizadas.

- **Reglas de asociación**

  Su objetivo identificar relaciones significativas entre variables en grandes bases de datos. Un ejemplo clásico de su aplicación es el "análisis de la cesta de mercado", que puede revelar patrones como que los clientes que compran pañales también tienden a adquirir cerveza (Agrawal, Imieliński, Swami, 1993).

3.2. Técnicas

Las técnicas de aprendizaje automático se organizan tradicionalmente en tres paradigmas principales:

A. Aprendizaje supervisado

Utiliza datos etiquetados para entrenar modelos que pueden realizar predicciones sobre nuevos datos. Incluye algoritmos de clasificación como Support Vector Machines, Random Forest, Gradient Boosting y redes neuronales, así como algoritmos de regresión como regresión lineal, regresión polinomial y regresión logística.

Este tipo de técnica es quizás el paradigma más ampliamente difundido y estudiado. En este enfoque, el algoritmo aprende a partir de un conjunto de datos de entrenamiento que ha sido "etiquetado" por un experto humano. Cada punto de datos de entrada, conocido como vector de características (X), tiene una etiqueta o resultado

de salida correspondiente (y). El objetivo del algoritmo es aprender una función de mapeo “f” tal que y=f(X). Durante el proceso de entrenamiento, el modelo realiza predicciones y las compara con la etiqueta correcta, ajustando sus parámetros internos para minimizar el error.

Los algoritmos que suelen utilizarse en esta técnica son de dos tipos:

I. Algoritmos de regresión

Entre los algoritmos de regresión se tiene:

- **Regresión lineal**

  El modelo más simple, que asume una relación lineal entre las características de entrada y la variable de salida.

- **Support Vector Regression (SVR)**

  Adapta los Support Vector Machines para problemas de predicción de valores continuos.

II. Algoritmos de clasificación

Por otra parte, dentro de los algoritmos de clasificación se encuentran:

- **Regresión Logística**

  Es un método de clasificación que modela la probabilidad de que una entrada pertenezca a una clase particular.

- **k-Nearest Neighbors (k-NN)**

  Clasifica un nuevo punto de datos basándose en la clase mayoritaria de sus 'k' vecinos más cercanos en el espacio de características.

- **Support Vector Machines (SVM)**

  Encuentra el hiperplano que mejor separa las clases en el espacio de características.

- **Árboles de Decisión y Random Forest**

  Los árboles de decisión aprenden una serie de reglas de "si... entonces..." para dividir los datos. Random Forest es un método de ensemble que construye múltiples árboles de decisión y combina sus predicciones para obtener un resultado más robusto y preciso.

- **Naive Bayes**

  Clasificador probabilístico basado en el teorema de Bayes con una suposición "ingenua" (naive) de independencia entre las características.

B. Aprendizaje no supervisado

Trabaja con datos sin etiquetas, buscando descubrir patrones ocultos en la estructura de los datos. En este paradigma, el algoritmo trabaja con datos que no han sido etiquetados. No hay una "respuesta correcta". El objetivo es explorar los datos y encontrar alguna estructura o patrón inherente por sí mismo. Dentro de esta técnica se pueden encontrar los siguientes algoritmos:

I. Algoritmos de Agrupamiento *(Clustering)*

Estos algoritmos tienen presente los siguientes tipos:

- **K-Means**

  Este es un algoritmo iterativo que organiza los datos en un número predefinido 'k' de clústeres, asignando cada punto de datos al clúster cuyo centroide (media) se encuentra más próximo.

- **DBSCAN**

  Este método, basado en la densidad, es capaz de identificar clústeres de formas arbitrarias y detectar el ruido.

II. Algoritmos de reglas de asociación

El tipo de algoritmo representativo dentro de este tipo es:

- **Apriori**

  Es el más reconocido para extraer conjuntos de ítems frecuentes y derivar reglas de asociación.

  Estos algoritmos de aprendizaje no supervisado son fundamentales en el análisis exploratorio de datos y la identificación de patrones ocultos. El K-Means, por ejemplo, es ampliamente utilizado en segmentación de clientes y análisis de mercado, mientras que el *DBSCAN* es particularmente útil en la detección de anomalías y agrupamiento espacial. Por otro lado, el algoritmo Apriori encuentra aplicaciones en el análisis de canasta de mercado y sistemas de recomendación, ayudando a descubrir relaciones interesantes entre productos o elementos.

C. Aprendizaje por refuerzo

Este es un paradigma diferente, inspirado en la psicología conductista que involucra a un agente que aprende a comportarse en un entorno realizando acciones y observando los resultados. Se enfoca en entrenar agentes que aprenden a tomar decisiones óptimas, a través de la interacción con un entorno, recibiendo recompensas o penalizaciones por sus acciones. Las técnicas avanzadas incluyen:

- **Ensemble methods**

  Combinan múltiples modelos para mejorar el rendimiento.
- **Técnicas de deep learning**

  Utilizadas para problemas complejos con grandes volúmenes de datos.
- **Métodos de transfer learning**

  Aprovechan conocimiento de dominios relacionados.

3.3. Selección del algoritmo

Con una amplia variedad de algoritmos disponibles, la elección del más adecuado para un problema específico constituye una de las decisiones más críticas. No existe un algoritmo "mejor" de manera universal, un principio formalizado en el teorema "No Free Lunch" (Wolpert, 1996), que establece que ningún algoritmo puede superar a todos los demás en todos los problemas posibles. La elección implica un compromiso entre varios factores. Por ejemplo, la naturaleza del problema es el primer filtro a considerar. ¿Se trata de un problema de regresión, clasificación o agrupamiento? Esto limita significativamente las opciones disponibles.

Asimismo, el tamaño y la dimensionalidad del conjunto de datos también son cruciales: un conjunto de datos con millones de muestras y características (Big Data) puede hacer que algunos algoritmos no sean viables debido a su falta de escalabilidad.

Por otro lado, algoritmos complejos como las redes neuronales profundas *(*Deep Learning) a menudo necesitan grandes volúmenes de datos para evitar el sobreajuste. Existe un equilibrio entre interpretabilidad y precisión.

En el otro extremo, se encuentran modelos más sencillos como la regresión lineal o los árboles de decisión, los cuales son fácilmente interpretables, lo que significa que es sencillo entender por qué hacen una predicción específica. Sin embargo, modelos más complejos como las redes neuronales o los ensambles de Gradient Boosting suelen ser "cajas negras": ofrecen alta precisión, pero su lógica interna es difícil de entender.

La elección depende de si el "por qué" es tan importante como el "qué". Los supuestos del modelo también son necesarios: muchos algoritmos se basan en ciertas suposiciones sobre los datos (por ejemplo, la regresión lineal asume linealidad y baja multicolinealidad; Naive Bayes asume independencia de las características). Es esencial tener un conocimiento básico de estos supuestos para evitar aplicar un modelo de manera incorrecta. Las restricciones computacionales también juegan un papel: ¿cuánto tiempo se tiene para entrenar el modelo? ¿Qué tan rápidas deben ser las predicciones en producción?

Como algunas ejemplificaciones, se pueden tener las siguientes:

- **k-NN**

  Algunos modelos son rápidos de entrenar, pero lentos para predecir.

- **SVM**

  Son lentos de entrenar, pero instantáneos para predecir.

En la práctica, el proceso de selección a menudo implica experimentar con varios algoritmos candidatos, ajustarlos y compararlos de manera rigurosa, un tema que exploraremos en el siguiente capítulo sobre la evaluación de modelos.

4. Evaluación del modelo

En los numerales previos se han abordados conceptos como la limpieza de datos, se ha simplificado su complejidad, mediante la reducción de dimensionalidad y se han empleado poderosos algoritmos para que aprendan de ellos. El resultado es un "modelo entrenado". Pero, ¿cómo se sabe si este modelo es bueno? ¿Qué significa "bueno" en este contexto? ¿Es preciso? ¿Es fiable? ¿Comete errores costosos? La evaluación del modelo es el proceso sistemático, mediante el cual se pueden responder esas preguntas, cuantificando el rendimiento de un modelo de manera objetiva.

Antes de dar inicio al concepto de las métricas, se debe introducir el principio más importante para una evaluación honesta: la separación de los datos. Nunca se debe evaluar un modelo con los mismos datos que se usaron para entrenarlo. Hacerlo sería como darle a un estudiante las respuestas de un examen, antes de que lo tome; obtendría una puntuación perfecta, pero no habríamos medido su capacidad real de generalización.

Para evitar este sesgo, los datos se dividen típicamente en tres conjuntos:

- **Entrenamiento**

  Se utiliza para enseñar al modelo.

- **Validación**

  Útil para ajustar hiperparámetros y seleccionar el mejor modelo.

- **Prueba**

  Se utiliza para la evaluación final.

Esta separación asegura que esta evaluación sea justa y representativa del rendimiento real del modelo en datos nuevos y no vistos.

Las métricas que se abordan a continuación, deben calcularse sobre el conjunto de prueba, para obtener una estimación real de cómo se comportará el modelo en el mundo real con datos nuevos.

4.1. Métricas de clasificación

En los problemas de clasificación, el objetivo es predecir una etiqueta categórica. Las métricas de evaluación ayudan a entender la naturaleza de la predicciones correctas e incorrectas. La matriz de confusión es una herramienta fundamental para visualizar estas predicciones, mostrando la cantidad de aciertos y errores para cada clase. A partir de esta matriz, se pueden calcular métricas más específicas como la precisión, la sensibilidad y el `F1-score`. Estas métricas permiten evaluar el rendimiento del modelo desde diferentes perspectivas, considerando el equilibrio entre los falsos positivos y los falsos negativos.

Para comenzar se tiene la matriz de confusión. La cual es fundamental para la evaluación de la clasificación. Esta tabla desglosa el rendimiento de un modelo al comparar las clases predichas con las clases reales. En un problema de clasificación binaria, que incluye una clase "Positiva" y una "Negativa", la matriz consta de cuatro celdas:

- **Verdaderos Positivos**

  El modelo predijo "Positivo" y la clase real era "Positivo". (Ej: El modelo detectó correctamente una transacción fraudulenta).

- **Verdaderos Negativos**

  El modelo predijo "Negativo" y la clase real era "Negativo". (Ej: El modelo identificó correctamente una transacción legítima).

- **Falsos Positivos**

  El modelo predijo "Positivo", pero la clase real era "Negativo". (Ej: El modelo marcó una transacción legítima como fraude, causando una molestia al cliente).

- **Falsos Negativos**

  El modelo predijo "Negativo", pero la clase real era "Positivo". (Ej: El modelo no detectó una transacción fraudulenta, causando una pérdida financiera).

A partir de esta matriz se derivan todas las demás métricas de clasificación:

- **Exactitud (Accuracy)**

  La métrica más intuitiva. Representa la proporción de predicciones correctas. Aunque es fácil de entender, puede ser muy engañosa en conjuntos de datos desbalanceados. Por ejemplo, si el 99 % de las transacciones no son fraudulentas, un modelo que siempre predice "no fraude" tendrá un 99 % de exactitud, pero será completamente inútil para detectar el fraude.

- **Precisión (Precision)**

  De todas las veces que el modelo predijo "Positivo", ¿cuántas acertó? Es crucial cuando el costo de un Falso Positivo es alto.

- **Sensibilidad (Recall o Exhaustividad)**

  De todos los casos que eran realmente "Positivos", ¿cuántos fue capaz de identificar el modelo? Es la métrica más importante cuando el costo de un Falso Negativo es alto (como en el diagnóstico médico).

- **Puntuación F1 (F1-Score)**

  Es la media armónica de la Precisión y la Sensibilidad. Proporciona una única métrica que equilibra ambas, y es especialmente útil cuando se tienen clases desbalanceadas.

- **Curva ROC y AUC**

  Es un gráfico que muestra el rendimiento de un clasificador, a través de todos los umbrales de clasificación. Grafica la Tasa de Verdaderos Positivos (Recall) frente a la Tasa de Falsos Positivos (FP/(FP+TN)). Un modelo perfecto se ubicaría en la esquina superior izquierda (100 % de sensibilidad, 0 % de falsos positivos). Se puede decir que es una métrica agregada que representa la capacidad general del modelo para discriminar entre las clases. Un AUC de 1.0 es un clasificador perfecto, mientras que un AUC de 0.5 representa un modelo que no es mejor que una elección al azar (Fawcett, 2006).

Ejemplo de interpretación: en un modelo de detección de fraude, una precisión alta pero en un recall bajo indicaría que, aunque la mayoría de las transacciones marcadas como fraude son realmente fraudalentas, el modelo está pasando por alto muchos casos de fraude reales (Chawla, Boyer, Kegelmeyer, 2002).

4.2. Métricas de regresión

Para problemas de regresión, donde la variable objetivo es continua, se utilizan:

- **Error Absoluto Medio (MAE - Mean Absolute Error)**

  Se define como el promedio de las diferencias absolutas entre las predicciones y los valores reales. Es fácil de interpretar se expresa en las mismas unidades que la variable objetivo.

- **Error Cuadrático Medio (MSE - Mean Squared Error)**

  Se define como el promedio de los errores elevados al cuadrado. Al elevar el error al cuadrado, se penalizan de manera más significativa los errores grandes en comparación con los pequeños. Es de las más utilizadas; sin embargo, su resultado se expresa en unidades al cuadrado (por ejemplo, "dólares al cuadrado"), lo que complica su interpretación directa.

- **Raíz del Error Cuadrático Medio (RMSE - Root Mean Squared Error)**

  Se define como la raíz cuadrada del MSE. Aborda el problema de interpretabilidad del MSE, dado que sus unidades coinciden con las de la variable objetivo, al tiempo que conserva la característica de penalizar de manera más severa los errores de mayor magnitud.

- **Coeficiente de Determinación (R^2)**

  A diferencia de las métricas de error, el coeficiente de determinación R^2 evalúa la adecuación del modelo a los datos. Indica la proporción de la varianza en la variable dependiente que puede ser explicada por las variables independientes. Un valor de R^2=0.75 implica que el modelo

explica el 75 % de la variabilidad en los datos de respuesta. Aunque es una medida útil, puede resultar engañosa, ya que su valor tiende a incrementarse con la adición de cada variable al modelo, independientemente de su utilidad. Por esta razón, a menudo se prefiere el uso del R^2 ajustado, el cual penaliza la inclusión de predictores que no aportan valor.

La elección de la métrica de evaluación adecuada es tan importante como la elección del algoritmo. Debe estar directamente alineada con los objetivos del proyecto. Un modelo no es "bueno" o "malo" en el vacío; es bueno o malo para la tarea específica para la que fue diseñado y estas métricas son el lenguaje que se usan para describir y defender esa conclusión.

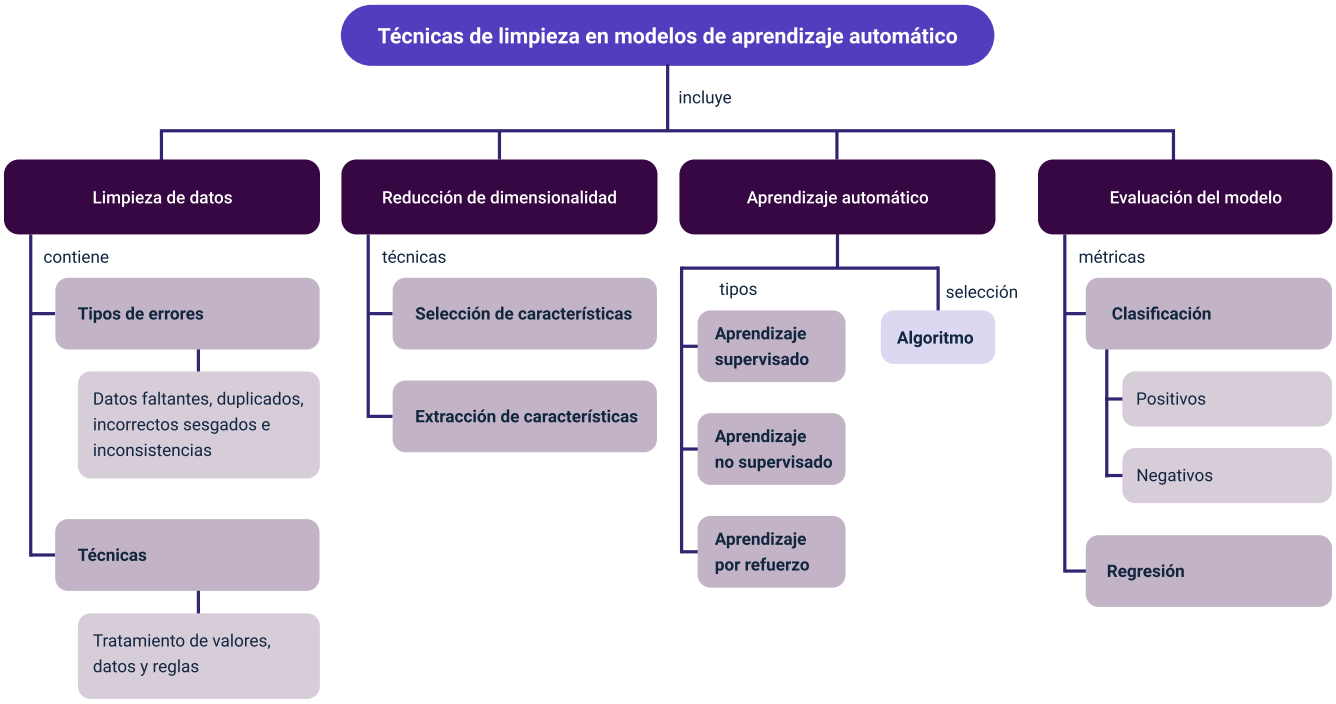
La selección de la métrica también debe considerar las características específicas del conjunto de datos y el contexto del problema. Es crucial entender las implicaciones y limitaciones de cada métrica para interpretar correctamente los resultados del modelo. Además, en muchos casos, es beneficioso utilizar múltiples métricas complementarias para obtener una evaluación más completa y robusta del rendimiento del modelo.

Síntesis

Este componente analiza los principios universales que sustentan las técnicas más usadas en la limpieza y preparación de datos para modelos de inteligencia artificial. La gestión eficaz de datos en modelos de inteligencia artificial, exige un enfoque sistemático y exhaustivo que abarque desde la limpieza inicial de datos hasta la evaluación rigurosa de los modelos. Cada etapa del proceso contribuye de manera significativa al éxito final del proyecto y la negligencia en cualquier fase puede comprometer la efectividad y confiabilidad del sistema en su totalidad.

La limpieza de datos constituye la base sobre la cual se desarrollan modelos robustos, mientras que la reducción de dimensionalidad optimiza la eficiencia computacional sin comprometer información esencial. La selección adecuada de técnicas de aprendizaje automático y algoritmos específicos determina la capacidad del modelo para identificar patrones significativos y generar predicciones precisas.

La evaluación sistemática del modelo mediante métricas adecuadas garantiza que los sistemas desarrollados, no solo funcionen correctamente con los datos de entrenamiento, sino que también mantengan su rendimiento en aplicaciones del mundo real. La integración efectiva de todos estos componentes resulta en sistemas de inteligencia artificial que pueden proporcionar un valor tangible y confiable en una amplia gama de aplicaciones y dominios.


Técnicas de limpieza en modelos de aprendizaje automático
incluye
Limpieza de datos
Reducción de dimensionalidad
Aprendizaje automático
Evaluación del modelo
contiene
Tipos de errores
Datos faltantes, duplicados, incorrectos sesgados e inconsistencias
Técnicas
Tratamiento de valores, datos y reglas
técnicas
Selección de características
Extracción de características
tipos
selección
Aprendizaje supervisado
Aprendizaje no supervisado
Aprendizaje por refuerzo
Algoritmo
métricas
Clasificación
Positivos
Negativos
Regresión

Material complementario

Tema	Referencia APA del Material	Tipo de material (Video, capítulo de libro, artículo, otro)	Enlace del Recurso o Archivo del documento o material
1. Limpieza de datos.	Ecosistema de Recursos Educativos SENA. (2023). Proceso de normalización de datos [Video]. YouTube.	Video.	https://www.youtube.com/watch?v=hKwuc-JJisI
2. Reducción de dimensionalidad.	Ecosistema de Recursos Educativos SENA. (2023). Etapas del procesamiento de datos y métodos estadísticos Introducción [Video]. YouTube.	Video.	https://www.youtube.com/watch?v=ndzj15PQEVw

Tema	Referencia APA del Material	Tipo de material (Video, capítulo de libro, artículo, otro)	Enlace del Recurso o Archivo del documento o material
3. Aprendizaje automático.	Ecosistema de Recursos Educativos SENA. (2025). Preparación y modelado de datos para algoritmos de machine learnig [Video]. YouTube.	Video.	https://www.youtube.com/watch?v=cDIla4TZWoU

Glosario

Bias-Variance Tradeoff (Compensación Sesgo-Varianza): describe la relación entre sesgo (errores por suposiciones simplificadoras) y varianza (sensibilidad a fluctuaciones en datos de entrenamiento).

Cross-Validation (Validación Cruzada): técnica de evaluación que divide los datos en múltiples subconjuntos para entrenar y evaluar el modelo repetidamente, proporcionando una estimación más robusta del rendimiento y ayudando a detectar sobreajuste.

Ensemble Methods (Métodos de Ensamble): técnicas que combinan múltiples modelos de aprendizaje automático para crear un predictor más fuerte que cualquiera de los modelos individuales. Incluyen métodos como bagging, boosting y stacking.

Ingeniería de Características: conjunto de técnicas para transformar y seleccionar variables derivadas de datos brutos con el fin de mejorar el rendimiento de los modelos de IA. Incluye creación de nuevas características, codificación de categorías y escalado de valores.

Overfitting (Sobreajuste): fenómeno donde un modelo aprende demasiado específicamente de los datos de entrenamiento, incluyendo ruido, resultando en excelente rendimiento en entrenamiento, pero pobre generalización a datos nuevos.

Regularización: conjunto de técnicas para prevenir sobreajuste añadiendo un término de penalización a la función de pérdida del modelo. Incluye regularización L1 (Lasso), L2 (Ridge) y elástica, controlando la complejidad del modelo.

Referencias bibliográficas

Aggarwal, C. C. (2017). An introduction to outlier analysis (pp. 1-34). Springer International Publishing.

Agrawal, R., Imieliński, T. & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).

Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Han, J., Pei, J. & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

Dasu, T. & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3, 1157-1182.

Jolliffe, I. T. (2002). Principal component analysis for special types of data (pp. 338-372). Springer New York.

Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.

Mitchell, T. M. (1997). Machine learning (Vol. 1, No. 9). New York: McGraw-hill.

Rahm, E. & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

Schafer, J. L. & Graham, J. W. (2002). Missing data: our view of the state of the art. Psychological methods, 7(2), 147.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

Van Buuren, S. (2018). Flexible Imputation of Missing Data. CRC Press.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. Neural computation, 8(7), 1341-1390.

Créditos

Nombre	Cargo	Centro de Formación y Regional
Milady Tatiana Villamil Castellanos	Responsable Ecosistema de Recursos Educativos Digitales (RED)	Dirección General
Diana Rocio Possos Beltrán	Responsable de línea de producción	Centro de Comercio y Servicios - Regional Tolima
Deivis Eduard Ramírez Martínez	Experto temático	Centro de Comercio y Servicios - Regional Tolima
Andrés Felipe Velandia Espitia	Evaluador instruccional	Centro de Comercio y Servicios - Regional Tolima
Oscar Iván Uribe Ortiz	Diseñador web	Centro de Comercio y Servicios - Regional Tolima
Juan Daniel Polanco Muñoz	Diseñador web	Centro de Comercio y Servicios - Regional Tolima
Diego Fernando Velasco Güiza	Desarrollador Full stack	Centro de Comercio y Servicios - Regional Tolima
Francisco José Vásquez Suárez	Desarrollador Full stack	Centro de Comercio y Servicios - Regional Tolima
Ernesto Navarro Jaimes	Animador y productor audiovisual	Centro de Comercio y Servicios - Regional Tolima
Norma Constanza Morales Cruz	Evaluadora de contenidos inclusivos y accesibles	Centro de Comercio y Servicios - Regional Tolima

Nombre	Cargo	Centro de Formación y Regional
Javier Mauricio Oviedo	Validador y vinculador de recursos educativos digitales	Centro de Comercio y Servicios - Regional Tolima