Fundamentos de datos, estadística descriptiva y seguridad de la información

Breve descripción:

Este componente está diseñado para desarrollar habilidades técnicas en el reconocimiento de fuentes y tipos de datos. Asimismo, fortalece las capacidades de los aprendices en un entorno digital en constante evolución, con fundamentos estadísticos. Además, permite adquirir competencias en el análisis, validación y preparación de datos para modelos de Inteligencia Artificial (IA).

Abril 2025

Tabla de contenido

Introducción

En una sociedad donde la tecnología evoluciona constantemente, las organizaciones dependen de los datos de sus sistemas de información para la toma de decisiones. La calidad y coherencia de estos datos son fundamentales para el rendimiento óptimo de los algoritmos de aprendizaje automático. La preparación de los datos es un paso esencial en el desarrollo de sistemas inteligentes y en la extracción de información relevante mediante técnicas de aprendizaje automático, las cuales tienen un impacto significativo en áreas como la visión por computadora, el procesamiento del habla, la comprensión del lenguaje natural, la neurociencia, la salud y el Internet de las cosas.

Las organizaciones requieren personal altamente capacitado en análisis y preparación de datos para desarrollar soluciones de inteligencia artificial (IA). Esto permite adaptarse a un ecosistema tecnológico en constante transformación, aportando valor estratégico a las empresas y respondiendo a los desafíos del mercado. La gestión eficiente de los datos facilita la conversión de la información en conocimiento estratégico, impulsando la innovación y la competitividad empresarial.

En este contexto, la IA se consolida como un motor de innovación, por lo que es fundamental que las organizaciones aprovechen de manera efectiva los datos recopilados en sus sistemas de información y gestionen grandes volúmenes de información en sectores tanto públicos como privados. La escasez de talento especializado representa un obstáculo para la digitalización y el desarrollo tecnológico.

Este programa se alinea con la visión del SENA de fortalecer el entorno productivo mediante el desarrollo de competencias técnicas y actitudinales que favorecen una economía basada en el conocimiento y la tecnología. La transformación digital en Colombia y la creciente implementación de IA destacan la necesidad de contar con profesionales especializados en datos. Los indicadores de desarrollo tecnológico y la inversión en innovación evidencian la integración de técnicas de aprendizaje automático en la gestión y toma de decisiones.

Para profundizar en la importancia de estos temas, se recomienda acceder al siguiente video:

Video 1. Fundamentos de datos, estadística descriptiva y seguridad de la información

Enlace de reproducción del video

Síntesis del video: Fundamentos de datos, estadística descriptiva y seguridad de la información
En este componente formativo, se exploran los principales tipos de datos utilizados en sistemas de información. Se presentan los datos estructurados, organizados en formatos predefinidos; los datos no estructurados, que carecen de una estructura fija; y los datos semiestructurados, que combinan características de ambos. Además, se analiza la calidad de los datos, considerando aspectos como la exactitud, completitud, coherencia y validez. La estadística descriptiva permite comprender y analizar los datos mediante la identificación de niveles de medición y la clasificación en variables categóricas y numéricas. Se abordan las medidas de tendencia central y dispersión, incluyendo media, mediana, moda, varianza y desviación estándar, elementos fundamentales para interpretar la distribución de los datos. Asimismo, se introducen técnicas de visualización y análisis exploratorio, esenciales para la detección de patrones y la toma de decisiones informadas. Por otro lado, la gobernanza y seguridad de los datos son clave para garantizar la integridad y confidencialidad de la información. Se revisan las principales políticas y normativas que rigen su gestión, la gobernanza de datos como estrategia para su administración eficiente y los mecanismos de seguridad que protegen su acceso y uso. Este conocimiento es fundamental en la era digital, donde los datos representan un recurso estratégico para la innovación y la toma de decisiones.

1. Tipo de datos

En el ámbito de la Inteligencia Artificial (IA), el análisis de datos es un proceso fundamental, ya que la calidad y relevancia de la información analizada inciden directamente en el rendimiento y la precisión de los modelos desarrollados. Un Análisis Exploratorio de Datos (AED) eficaz permite identificar patrones y tendencias clave para el entrenamiento de algoritmos de aprendizaje automático y la mejora en la toma de decisiones basada en datos. Para ello, se emplean técnicas estadísticas que facilitan la comprensión de las relaciones entre las variables analizadas.

Según García (2017), la esencia del AED radica en permitir que los datos “hablen” y, a partir de ellos, identificar los patrones y modelos correspondientes. Por esta razón, el AED se considera una herramienta de gran utilidad en la generación de modelos que representan fenómenos específicos.

El volumen de datos generado en los distintos sistemas de información supera la capacidad humana para analizarlos sin el uso de técnicas automatizadas. Los datos recopilados en los medios transaccionales de las organizaciones han experimentado un crecimiento constante, reflejando una tendencia al alza.

Existen diversas definiciones sobre el análisis de datos. Una de las más reconocidas es la propuesta por Hair y otros (2010), quienes lo definen como el proceso de convertir datos en bruto en información útil para la toma de decisiones. Este proceso no solo implica la aplicación de razonamiento estadístico y herramientas informáticas, sino también la capacidad de interpretar los resultados en el contexto del problema de investigación.

El análisis de datos es un componente esencial en la era del big data, ya que abarca la examinación, limpieza, transformación y modelado de datos con el objetivo de descubrir información valiosa, extraer conclusiones y respaldar la toma de decisiones. Este proceso incluye una variedad de técnicas y herramientas, que van desde métodos tradicionales de minería de datos y aprendizaje automático hasta enfoques avanzados como el aprendizaje profundo (Wang, 2017).

Un aspecto clave en el análisis de datos es la identificación de sus fuentes, las cuales pueden clasificarse en:

- ✓ **Internas**: generadas dentro de la organización, como registros de transacciones, bases de datos de clientes y datos operativos.
- ✓ **Externas**: provenientes de fuentes externas a la organización, como información de mercado, datos demográficos, estadísticas gubernamentales y registros meteorológicos.

1.1. Datos estructurados

Los sistemas de información actuales manejan diversos tipos de datos, los cuales pueden provenir de múltiples fuentes, incluyendo dispositivos IoT, sistemas operativos, aplicaciones y procesos de curación manual (Gehani y Tariq, 2012).

Los datos estructurados se encuentran organizados en formatos predefinidos, como tablas en bases de datos relacionales, donde cada campo tiene un tipo de dato específico (Pyle, 1999). Su organización sistemática permite realizar consultas eficientes, aplicar filtros y establecer relaciones entre diferentes conjuntos de datos. Se almacenan en bases de datos relacionales, hojas de cálculo y archivos con formato fijo,

lo que facilita su procesamiento y análisis en diversos sectores, como el comercio electrónico, la banca y la gestión empresarial.

Por ejemplo, en un sistema de administración de inventarios, los datos estructurados incluyen información como el código del producto, nombre, precio, cantidad disponible y ubicación en el almacén. Esta estructura facilita la gestión del inventario, el análisis de tendencias de venta y la toma de decisiones estratégicas. Además, los datos estructurados son fundamentales para la implementación de sistemas de inteligencia empresarial y análisis predictivo, ya que proporcionan una base sólida para extraer información clave.

En este contexto, los datos estructurados pueden contener tanto variables cualitativas como cuantitativas. Por ejemplo, una columna puede indicar la categoría de un producto (cualitativo) y otra su precio (cuantitativo). Esta organización facilita la aplicación de análisis estadísticos, como distribuciones de frecuencia y medidas de tendencia central.

La tabla 1 presenta un ejemplo de datos estructurados en una tienda en línea, donde se organizan en filas y columnas con tipos de datos definidos.

Tabla 1. Datos estructurados en una tienda en línea

ID Cliente	Nombre Completo	Edad	Género	Fecha de registros
C001	Laura Gómez	28	F	12/05/2023
C002	Jorge Ramírez	35	M	8/11/2022

ID Cliente	Nombre Completo	Edad	Género	Fecha de registros
C003	Ana Pérez	42	F	19/07/2021
C004	Manuel Ortega	31	M	25/01/2023
C005	Isabel Fernández	24	F	14/09/2022

De manera similar, la tabla 2 describe un conjunto de datos estructurados en un contexto de atención médica, organizado en filas y columnas predefinidas.

Tabla 2. Datos estructurados en una historia clínica

ID Paciente	Nombre Completo	Edad	Género	Fecha Ingreso	Diagnóstico Principal	Nivel de Riesgo
P001	Carmen Rodríguez	56	F	5/11/2024	Hipertensión arterial	Medio
P002	Luis González	43	M	1/12/2024	Diabetes tipo 2	Alto
P003	Daniela Martínez	30	F	15/01/2025	Asma leve	Bajo

ID Paciente	Nombre Completo	Edad	Género	Fecha Ingreso	Diagnóstico Principal	Nivel de Riesgo
P004	Jorge Herrera	65	M	7/02/2025	Enfermedad pulmonar crónica	Alto
P005	Lucía Torres	49	F	12/03/2025	Dolor lumbar	Medio

La estructuración de estos datos en un formato tabular permite una gestión eficiente en distintos ámbitos, como la atención médica. En la tabla 2, se describe un ejemplo de historia clínica con información clave sobre los pacientes, incluyendo su identificación, diagnóstico y nivel de riesgo. Esta organización facilita el acceso rápido a los datos, el seguimiento de la evolución clínica y la toma de decisiones informadas en el tratamiento de cada paciente. Además, la disposición estructurada de la información posibilita el análisis estadístico y la implementación de herramientas de inteligencia artificial para mejorar la precisión en los diagnósticos y optimizar la gestión hospitalaria.

1.2. Datos no estructurados

Los datos no estructurados son aquellos que no siguen un modelo predefinido ni una organización tabular fija, a diferencia de los datos estructurados. No pueden almacenarse ni procesarse fácilmente en bases de datos relacionales sin una transformación previa. Se estima que representan aproximadamente el 80 % de los datos generados a nivel global (Gartner, citado en Katal et al., 2013). Ejemplos comunes

incluyen correos electrónicos, documentos de texto, imágenes, videos y publicaciones en redes sociales.

En su estado original, estos datos no se ajustan directamente a las categorías de cualitativos o cuantitativos según los criterios del análisis estadístico tradicional. No obstante, para que los modelos de aprendizaje automático, como los Modelos de Lenguaje Grandes (LLMs), puedan utilizarlos, es necesario transformarlos en representaciones estructuradas o numéricas. Por ejemplo, los textos se tokenizan y convierten en embeddings numéricos, mientras que las imágenes se representan mediante matrices de valores de píxeles.

La transformación de datos no estructurados permite su análisis y uso en diversas aplicaciones. En el caso de las imágenes, por ejemplo, cada píxel se representa con un valor numérico que refleja su intensidad o color. Una vez procesados, estos datos pueden analizarse como variables cuantitativas continuas o, tras una categorización, como variables cualitativas discretas. Gracias a esta conversión, los modelos de inteligencia artificial pueden identificar patrones y extraer información valiosa para tareas como la clasificación de imágenes, el análisis de sentimientos en texto y la generación de lenguaje natural.

En la tabla 3 se presentan ejemplos de datos no estructurados y sus formatos más comunes:

Tabla 3. Formatos de datos no estructurados

Tipo de Dato	Ejemplo concreto
Texto libre	Comentarios en redes sociales, opiniones en Amazon.
Imágenes	Fotografías almacenadas en el móvil o en Google Photos.
Audio	Grabaciones de llamadas, podcasts, notas de voz.
Video	Clases grabadas, transmisiones en vivo, videos de YouTube.
Documentos diversos	PDFs, presentaciones, archivos Word con texto sin estructura.
Correos electrónicos	Emails con texto, archivos adjuntos y formato libre.

Los datos no estructurados representan un reto y una oportunidad en el análisis de información. Su flexibilidad permite capturar detalles ricos y contextuales, pero también requiere herramientas avanzadas para su procesamiento. Tecnologías como el procesamiento de lenguaje natural, la visión por computadora y la inteligencia artificial han facilitado la extracción de valor a partir de estos datos. En sectores como la salud, el comercio y la ciberseguridad, su análisis ha permitido mejorar la toma de decisiones,

automatizar procesos y descubrir patrones que serían difíciles de detectar con métodos tradicionales.

1.3. Datos semiestructurados

Los datos semiestructurados representan una forma híbrida de información que combina elementos de datos estructurados y no estructurados. A diferencia de los datos estructurados, que se organizan de manera rígida en bases de datos relacionales, los datos semiestructurados no siguen un esquema fijo, aunque conservan ciertos elementos organizativos. Esta flexibilidad permite representar jerarquías y relaciones complejas dentro de los datos, facilitando su adaptación a diferentes necesidades y contextos.

Ejemplos comunes de datos semiestructurados son los archivos XML y JSON, que utilizan etiquetas o pares clave-valor para organizar la información. También se incluyen correos electrónicos con metadatos, logs de servidores y archivos YAML, los cuales contienen estructura, pero sin ajustarse a una forma tabular estricta.

Los datos semiestructurados son especialmente útiles en escenarios donde las estructuras de datos pueden evolucionar con el tiempo o cuando se trabaja con fuentes de información heterogéneas. Se emplean ampliamente en tecnologías web, integración de datos en plataformas diversas y sistemas de gestión documental. Además, son esenciales en entornos donde la interoperabilidad y el intercambio de información entre distintos sistemas requieren flexibilidad sin perder organización.

Su uso ha crecido con el auge del big data y la computación en la nube, ya que facilitan la recopilación y análisis de grandes volúmenes de información procedente de

diversas fuentes. En la tabla 4 se presentan algunos tipos de archivos comúnmente clasificados como semiestructurados.

Tabla 4. Ejemplo de tipos de datos semiestructurados

Tipo de archivo	Ejemplo de contenido
JSON (JavaScript Object Notation)	{"nombre": "Ana", "edad": 30, "activo": true}
XML (eXtensible Markup Language)	Ana30
Archivos YAML	nombre: Ana\nedad: 30\nactivo: true
Correos electrónicos con metadatos	Asunto, fecha, remitente, cuerpo del mensaje.
Logs de servidor	Estructurados por línea pero no en formato tabular.

A continuación, en la tabla 5 se presenta una comparación entre los diferentes tipos de datos analizados hasta ahora.

Tabla 5. Comparativa de datos

Característica	Estructurados	Semiestructurados	No estructurados
Formato	Tablas (SQL)	Etiquetas (XML, JSON)	Texto libre, imágenes
Esquema	Fijo	Flexible	Ausente

Característica	Estructurados	Semiestructurados	No estructurados
Ejemplo	Tabla de ventas	JSON de pedidos online	Opiniones en redes
Almacenamiento	Bases de datos SQL	Bases de datos NoSQL	Sistemas de archivos
Facilidad de análisis	Alta	Moderada	Baja (requiere IA, NLP)

Los datos no estructurados representan un reto y una oportunidad en el análisis de información. Su flexibilidad permite capturar detalles ricos y contextuales, pero también requiere herramientas avanzadas para su procesamiento. Tecnologías como el procesamiento de lenguaje natural, la visión por computadora y la inteligencia artificial han facilitado la extracción de valor a partir de estos datos. En sectores como la salud, el comercio y la ciberseguridad, su análisis ha permitido mejorar la toma de decisiones, automatizar procesos y descubrir patrones que serían difíciles de detectar con métodos tradicionales.

1.4. Calidad de los datos

La calidad de los datos se refiere a la idoneidad de los datos para servir a un propósito específico, como el análisis, la toma de decisiones o el entrenamiento de modelos de inteligencia artificial. En la metodología Data Assay, se describe como un proceso que "literalmente evalúa la calidad o el valor de los datos para la minería de datos" (Pyle, 1999). La calidad de los datos abarca múltiples dimensiones que determinan su fiabilidad y utilidad.

Las principales características de la calidad de los datos son:

- ✓ **Precisión y exactitud**: los datos deben reflejar la realidad sin errores. La exactitud implica que los valores registrados sean correctos, mientras que la precisión se refiere a la consistencia en la medición. Errores de medición, transcripción o interpretación pueden afectar estos aspectos (Del Pino, 2008).
- ✓ **Integridad y validez**: un dato de calidad es válido si pertenece a su espacio muestral correcto. Datos incompletos o con valores no válidos (como "No sabe/No contesta") pueden comprometer la fiabilidad del análisis (Viedma, 2018).
- ✓ **Consistencia y coherencia**: los datos deben mantenerse uniformes en diferentes sistemas y bases de datos. La falta de coherencia, como diferencias en nombres o formatos entre conjuntos de datos, puede generar errores en el análisis.
- ✓ **Completitud**: un conjunto de datos de calidad debe contener toda la información necesaria para su propósito. Datos faltantes o incompletos pueden afectar la validez de los resultados.
- ✓ **Actualidad y temporalidad**: la información debe estar actualizada y reflejar la realidad en el momento en que se usa. Datos desactualizados pueden generar decisiones erróneas.
- ✓ **Accesibilidad y usabilidad**: la calidad de los datos también depende de su disponibilidad para los usuarios adecuados en el momento oportuno. Los datos deben estar bien documentados y en formatos utilizables.

- ✓ **Adecuación al propósito**: la granularidad o nivel de detalle de los datos debe ser apropiado para el análisis que se pretende realizar. Datos demasiado generales o específicos pueden afectar la interpretación y los resultados.

Además, es importante considerar los diferentes tipos de calidad de datos, los cuales incluyen:

- ✓ **Calidad intrínseca**: relacionada con la precisión, consistencia y validez de los datos, independientemente de su uso.
- ✓ **Calidad contextual**: evalúa si los datos son adecuados para un propósito específico, considerando factores como la completitud y la actualidad.
- ✓ **Calidad de representación**: se refiere a la claridad con la que los datos están estructurados y documentados, incluyendo su formato y facilidad de comprensión.
- ✓ **Calidad operativa o de accesibilidad**: determina si los datos pueden ser obtenidos y utilizados sin restricciones, asegurando su disponibilidad y seguridad.

En síntesis, la calidad de los datos es un concepto multidimensional que garantiza que la información sea precisa, válida, completa, coherente y adecuada para el propósito al que está destinada. Evaluar la calidad de los datos es un paso esencial en la preparación de datos para cualquier análisis estadístico o aplicación de inteligencia artificial (Pyle, 1999).

2. Estadística descriptiva

Los métodos estadísticos se dividen en dos grandes categorías, estadística descriptiva y estadística inferencial. La estadística descriptiva se encarga de recopilar, organizar, analizar y presentar datos de manera resumida, sin realizar inferencias sobre la población de origen. En contraste, la estadística inferencial busca extraer conclusiones sobre una población a partir de una muestra, utilizando técnicas como la estimación de parámetros y el contraste de hipótesis (Aroca, 2009).

Este programa de formación complementario se centrará en la estadística descriptiva, la cual permite representar la información mediante tablas, gráficos y medidas estadísticas que sintetizan las características de los datos (Viedma, 2018). Dependiendo del origen de las fuentes, la estadística descriptiva organiza y clasifica los datos obtenidos a través de observaciones. Para ello, se utilizan herramientas como distribuciones de frecuencia, medidas de tendencia central (media, mediana, moda) y medidas de dispersión (rango, varianza, desviación estándar) (Del Pino, 2008).

Dentro de la estadística descriptiva, es posible diferenciar entre la estadística univariada y la estadística bivariada. La estadística univariada analiza una sola variable a la vez, mientras que la estadística bivariada estudia la relación entre dos variables, utilizando herramientas como tablas de contingencia, diagramas de dispersión y coeficientes de correlación.

Es importante señalar que la estadística descriptiva no emplea el cálculo de probabilidades ni permite realizar inferencias más allá de los datos analizados. Su objetivo principal es resumir y presentar la información de manera clara y comprensible, facilitando su interpretación y uso en la toma de decisiones (Del Pino, 2008).

2.1. Niveles de medición

Los niveles de medición en estadística describen las propiedades de los datos y determinan qué tipo de operaciones matemáticas y análisis estadísticos pueden aplicarse a ellos. Según la clasificación de Stevens (1946), existen cuatro niveles de medición: nominal, ordinal, de intervalo y de razón. Cada nivel tiene características particulares que determinan cómo se pueden interpretar y analizar los datos:

a) **Nivel nominal**: es el más básico y solo permite clasificar los datos en categorías sin establecer un orden entre ellas. No hay una jerarquía o secuencia lógica entre las categorías, y las únicas operaciones posibles son la clasificación y el conteo. A continuación, se presentan algunos ejemplos:
   - ✓ Género (masculino, femenino, otro).
   - ✓ Estado civil (soltero, casado, viudo, divorciado).
   - ✓ Nacionalidad (colombiana, argentina, mexicana).
   - ✓ Marcas de teléfonos (Samsung, Apple, Xiaomi).

b) **Nivel ordinal**: permite clasificar los datos en categorías con un orden o jerarquía, pero sin indicar diferencias cuantificables entre ellas. Se puede establecer una secuencia, pero no es posible medir con precisión la distancia entre los valores. A continuación, se presentan algunos ejemplos:

- ✓ Grados militares (soldado, cabo, sargento, teniente).
- ✓ Nivel de satisfacción en una encuesta (bajo, medio, alto)
- ✓ Clasificación en una competencia (primer lugar, segundo lugar, tercer lugar).
- ✓ Nivel educativo (primaria, secundaria, universitaria, posgrado).

c) **Nivel de intervalo**: no solo establece un orden, sino que también permite medir la distancia entre los valores. Sin embargo, no tiene un punto cero absoluto, lo que significa que no se pueden hacer afirmaciones proporcionales sobre los valores. A continuación, se presentan algunos ejemplos:

- ✓ Temperatura en grados Celsius o Fahrenheit (0 °C no significa ausencia de temperatura).
- ✓ Años en un calendario (el año 0 es arbitrario).
- ✓ Puntuaciones en un test de inteligencia (IQ).

d) **Nivel de razón:** es el más completo, ya que permite establecer un orden, medir la distancia entre los valores y tiene un punto cero absoluto, lo que permite realizar operaciones matemáticas como la multiplicación y la división. A continuación, se presentan algunos ejemplos:

- ✓ Ingresos mensuales de una persona ($0 indica ausencia de ingresos).
- ✓ Edad de una persona (20 años es el doble de 10 años).
- ✓ Peso en kilogramos.
- ✓ Distancia en kilómetros.

En la tabla 6 se presenta una comparación de los niveles de medición con sus características principales:

Tabla 6. Comparación de los niveles de medición

Nivel de medición	Clasificación	Orden	Diferencia cuantificable	Punto cero absoluto	Ejemplo
Nominal	Sí	No	No	No	Color de ojos.
Ordinal	Sí	Sí	No	No	Nivel socioeconómico.
Intervalo	Sí	Sí	Sí	No	Temperatura en °C.
Razón	Sí	Sí	Sí	Sí	Peso en kg.

La correcta identificación del nivel de medición de una variable es crucial para seleccionar las técnicas estadísticas adecuadas y evitar interpretaciones erróneas de los datos.

2.2. Variables categóricas

En el campo de la estadística descriptiva, una variable categórica clasifica los datos en grupos o categorías según cualidades o atributos, sin asignarles un valor numérico inherente. Sin embargo, en algunos casos, pueden recibir códigos numéricos solo para facilitar su análisis.

Las variables categóricas se dividen en dos tipos principales:

1. **Nominales: n**o tienen un orden específico.

Ejemplos: género, país de origen y carrera universitaria

2. **Ordinales**: tienen un orden o jerarquía, pero la diferencia entre categorías no es necesariamente uniforme.

Ejemplos: nivel de satisfacción y nivel educativo.

En la tabla 7, se presentan ejemplos de registros de personas con variables categóricas nominales y ordinales.

Tabla 7. Ejemplo de tabla de datos

ID de la persona	Nombre	Género (Nominal)	Carrera (Nominal)	Nivel de satisfacción (Ordinal)	Nivel educativo (Ordinal)
E001	Laura Gómez	Femenino	Ingeniería	Alto	Universitario
E002	Pedro Martínez	Masculino	Psicología	Medio	Secundario
E003	Ana Torres	Femenino	Medicina	Bajo	Universitario
E004	Juan Pérez	Masculino	Derecho	Alto	Técnico
E005	Silvia Ramírez	Femenino	Ingeniería	Medio	Universitario

Las variables categóricas suelen organizarse en tablas de frecuencia, que permiten visualizar la cantidad de veces que aparece cada categoría dentro de un conjunto de datos. Por ejemplo, en una encuesta realizada a 20 personas sobre su nivel

de satisfacción con un curso, se obtuvieron los siguientes resultados: 4 personas indicaron un nivel de satisfacción bajo, 9 personas lo calificaron como medio y 7 personas como alto. La tabla 8 muestra la distribución de frecuencias correspondiente, facilitando el análisis de estos datos:

Tabla 8. Representación de variables categóricas

Nivel de satisfacción	Frecuencia absoluta (f)	Frecuencia relativa (%)	Frecuencia acumulada
Bajo	4	20 %	4
Medio	9	45 %	13
Alto	7	35 %	20
Total	20	100 %	—

Las variables categóricas se pueden visualizar mediante:

- ✓ **Gráficos de barras**: representan la frecuencia de cada categoría mediante barras de diferente altura.
- ✓ **Gráficos circulares o de pastel**: presentan la proporción de cada categoría como un sector de un círculo.

A continuación, se presenta la tabla 9, que servirá como base para graficar los datos:

Tabla 9. Tabla de frecuencias variable categórica nominal

Lenguaje de programación	Frecuencia absoluta (f)	Frecuencia relativa (%)
Python	12	40 %
Java	8	26.7 %
C++	6	20 %
JavaScript	4	13.3 %
Total	30	100 %

Estos datos pueden representarse con un gráfico de barras o un gráfico circular para una mayor claridad en su interpretación para la frecuencia absoluta:

Figura 1. Gráfico de barras

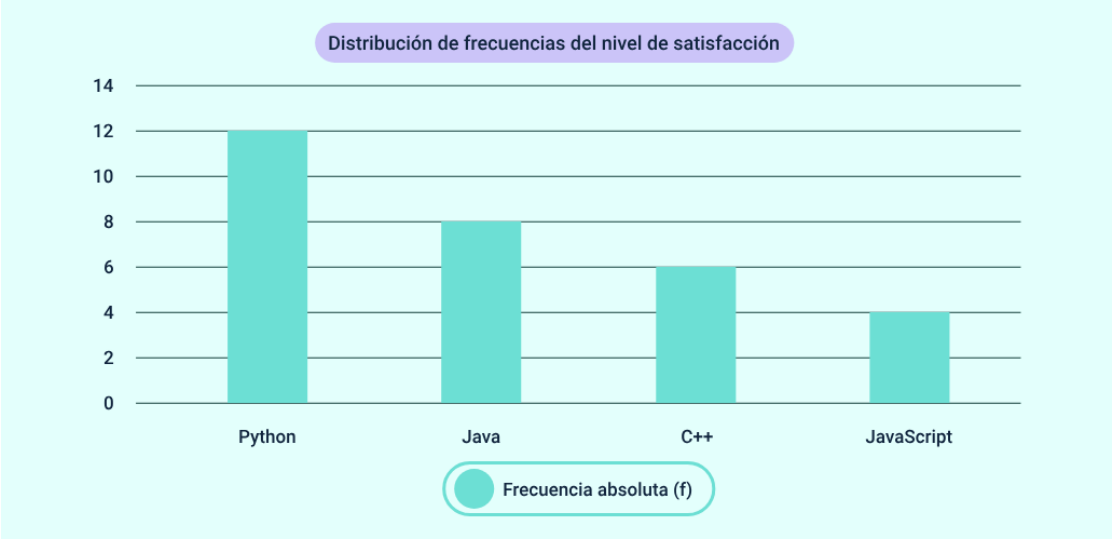
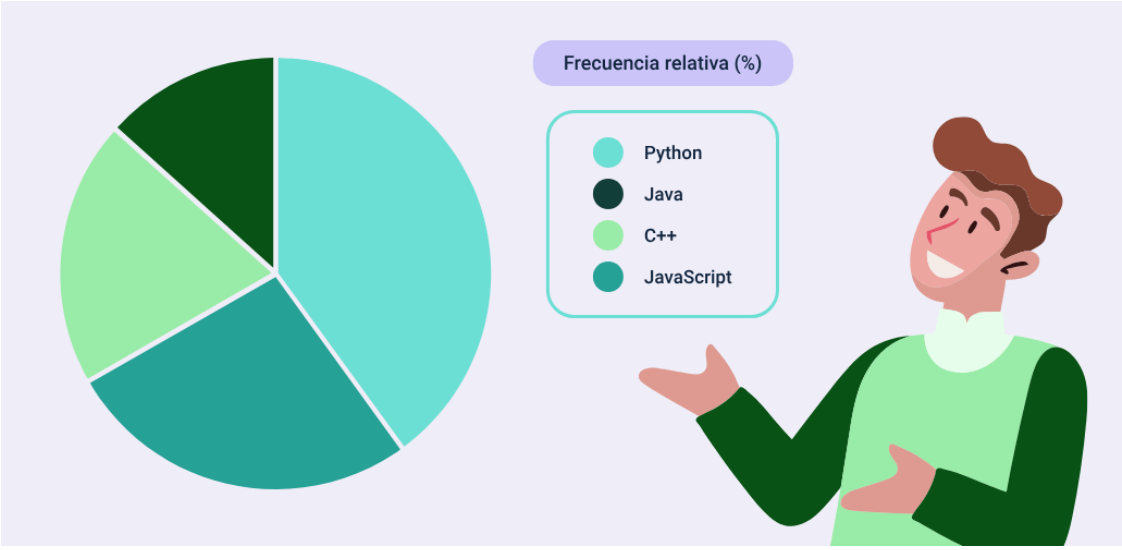


Figura 2. Gráfico circular



Las variables categóricas permiten clasificar datos cualitativos en grupos diferenciados, ya sean nominales (sin jerarquía) u ordinales (con jerarquía). Para analizarlas, se pueden usar tablas de frecuencia y gráficos visuales que faciliten su interpretación

2.3. Variables numéricas

En estadística descriptiva, una variable numérica o cuantitativa es aquella cuyos valores son números que representan cantidades o medidas. Estos valores tienen un significado intrínseco y permiten realizar operaciones aritméticas, como sumas, restas o cálculos de promedios. Las variables numéricas son fundamentales para el análisis estadístico, ya que permiten medir y comparar características dentro de un conjunto de datos.

Estas variables se clasifican en dos tipos principales:

a) Variables discretas

Toman valores aislados, generalmente números enteros, sin posibilidad de valores intermedios entre dos consecutivos. Suelen surgir de procesos de conteo, como el número de materias cursadas o la cantidad de hijos en una familia.

b) Variables continuas

Pueden asumir cualquier valor dentro de un rango o intervalo determinado, incluidos valores decimales. Son comunes en mediciones como la estatura, la temperatura o el tiempo empleado en una tarea.

Para analizar las variables numéricas, se utilizan herramientas como medidas de tendencia central (media, mediana y moda) y medidas de dispersión (rango, desviación estándar y varianza). Además, pueden representarse mediante histogramas, diagramas de caja o polígonos de frecuencia para facilitar su interpretación.

En la siguiente tabla, se presentan 10 registros con datos que explican el concepto de variables numéricas en estadística descriptiva, diferenciando entre discretas y continuas:

Tabla 10. Ejemplo de valores de variables numéricas

ID de la persona	Edad (años)	Estatura (cm)	Cantidad de materias cursadas	Promedio académico
E001	19	165.2	5	3.8

ID de la persona	Edad (años)	Estatura (cm)	Cantidad de materias cursadas	Promedio académico
E002	21	172.4	6	4.2
E003	20	160.8	4	3.5
E004	22	168.0	6	4.0
E005	23	175.5	7	3.9
E006	19	162.1	5	3.2
E007	24	178.3	7	4.5
E008	20	169.6	5	4.1
E009	21	171.0	6	3.7
E010	22	166.7	6	4.3

En la tabla se presentan ejemplos de variables numéricas discretas y continuas:

- ✓ **Variables discretas**: la edad (en años) y la cantidad de materias cursadas son variables discretas, ya que toman valores enteros sin fracciones intermedias.

- ✓ **Variables continuas**: la estatura (en centímetros) y el promedio académico son variables continuas, pues pueden tomar cualquier valor dentro de un rango, incluyendo valores decimales.

2.4. Medidas de tendencia central y dispersión

En estadística descriptiva, las medidas de tendencia central (MTC) son valores representativos de un conjunto de datos. También se les conoce como promedios, ya que proporcionan un valor típico que resume la distribución de los datos (Mesa Guerrero, 2020). Las principales MTC son:

A. **Media aritmética**: es el resultado de sumar todos los valores y dividirlos por el número total de observaciones. Su fórmula es:

$$\bar{x} = \frac{\sum_{i=1}^{n} Xi}{n}$$

Donde:

$\bar{x}$: es la media aritmética.

x_i: representa cada valor del conjunto.

n: es el número total de observaciones.

Por ejemplo, si se tienen las edades de cinco aprendices: 20, 22, 24, 21 y 23, la media se calcula como:

$\bar{x}$ = (20+22+24+21+23) /5

$\bar{x}$ =22

B. **Mediana**: es el valor central cuando los datos están ordenados. Si el número de datos es impar, la mediana es el valor del centro; si es par, se calcula como el promedio de los dos valores centrales.

Ejemplo con un número impar de datos:

Datos originales: 20, 22, 21, 23, 25.

Ordenados: 20, 21, 22, 23, 25.

Mediana: 22 (el valor central).

Ejemplo con un número par de datos:

Datos ordenados: 15, 18, 20, 25, 30, 35.

Se toman los dos valores centrales: 20 y 25.

$$Mediana = \frac{20+5}{2} = 22.5$$

C. **Moda**: es el valor que ocurre con mayor frecuencia en un conjunto de datos. Puede ser:

- ✓ **Unimodal**: un solo valor más frecuente. Ejemplo: en el conjunto {18, 19, 20, 20, 20, 21, 22, 23}, la moda es 20.
- ✓ **Bimodal**: dos valores con la misma mayor frecuencia. Ejemplo: en {14, 15, 15, 16, 16, 17, 18}, las modas son 15 y 16.
- ✓ **Multimodal**: más de dos valores con la misma mayor frecuencia. Ejemplo: en {12, 13, 14, 14, 15, 15, 16, 17, 17, 18, 18, 19}, las modas son 14, 15, 17 y 18.

- ✓ **Sin moda**: todos los valores aparecen con la misma frecuencia. Ejemplo: en {10, 11, 12, 13, 14, 15}, no hay moda.

Estas medidas ayudan a describir la distribución de los datos y se eligen según la estructura de la muestra:

- ✓ La media aritmética es útil en conjuntos grandes y con distribución normal.
- ✓ La mediana es preferida cuando hay valores extremos o en muestras pequeñas.
- ✓ La moda es especialmente relevante en datos cualitativos o cuando se desea conocer el valor más común.

Además, existen otras medidas relacionadas, como cuartiles, deciles y percentiles, que ayudan a analizar la dispersión y posición de los datos dentro de la distribución.

En conclusión, la estadística descriptiva proporciona herramientas para resumir y comprender un conjunto de datos a través de métodos como la tabulación, la representación gráfica y el cálculo de estadísticos descriptivos (Del Pino, 2008).

2.5. Visualización de datos y análisis exploratorio

La visualización de datos y el análisis exploratorio (AED) son pasos fundamentales en la estadística descriptiva, ya que permiten examinar la distribución, el comportamiento y la estructura de los datos antes de aplicar técnicas más complejas. Su objetivo es ofrecer una representación clara y comprensible de la información recolectada, facilitando la identificación de patrones, tendencias, valores atípicos y posibles errores.

Dependiendo del tipo de variable, se seleccionan diferentes tipos de gráficos. Para variables categóricas, los gráficos de barras y los diagramas de sectores permiten comparar frecuencias de manera visual. En cambio, para variables numéricas, se utilizan histogramas, diagramas de caja (boxplot) y gráficos de dispersión, que facilitan el análisis de distribución, variabilidad y relaciones entre variables.

El análisis exploratorio también contribuye a verificar la calidad de los datos, ya que permite detectar valores inconsistentes, datos faltantes o atípicos que podrían afectar los resultados estadísticos. Esta revisión previa es esencial para decidir si se requiere depuración, transformación o imputación de valores antes de proceder con análisis inferenciales.

Actualmente, existen múltiples herramientas tecnológicas que facilitan estos procesos. Programas como Excel, SPSS, R o Python (utilizando bibliotecas como Matplotlib, Seaborn o Pandas), así como plataformas como Power BI y Tableau, permiten generar visualizaciones dinámicas y realizar análisis exploratorio de manera efectiva.

En conjunto, la visualización de datos y el AED no solo proporcionan una comprensión inicial de los datos, sino que también orientan la selección de métodos estadísticos adecuados en etapas posteriores del análisis.

3. Gobernanza y seguridad de datos

La gobernanza de los datos comprende el conjunto de políticas, procesos, normas, responsabilidades y estructuras organizacionales destinadas a garantizar que los datos sean confiables, seguros, accesibles, coherentes y utilizados de manera ética. Esta práctica permite gestionar los datos como un activo estratégico, asegurando su calidad, integridad y disponibilidad a lo largo de su ciclo de vida.

En los últimos años, el concepto de gobernanza se ha incorporado de manera sistemática en el lenguaje de las administraciones públicas, las organizaciones privadas y el ámbito académico (Cerrillo-Martínez, 2018). Esta incorporación refleja un reconocimiento creciente del valor de los datos para la formulación de políticas, la innovación institucional y la toma de decisiones basada en evidencia.

El crecimiento exponencial de los datos digitales, así como la transformación digital de las organizaciones, ha llevado a que los datos se posicionen como uno de los recursos más valiosos. Ante esta realidad, se vuelve imprescindible implementar una estrategia de gobernanza que permita su gestión eficaz, integrando procesos de control, supervisión, protección y aprovechamiento. Según la CEPAL (2023), la gobernanza de datos es la clave para organizar, custodiar y maximizar el valor de la información dentro de las instituciones.

Adicionalmente, la seguridad de los datos es un componente inseparable de la gobernanza. Esta implica establecer mecanismos técnicos y normativos para prevenir accesos no autorizados, proteger la confidencialidad, mantener la integridad y garantizar la disponibilidad de la información. El cumplimiento de marcos legales como la Ley de Protección de Datos Personales y las normativas internacionales refuerza esta

dimensión, promoviendo un entorno seguro para el intercambio y procesamiento de datos.

En resumen, una estrategia efectiva de gobernanza y seguridad de los datos permite no solo cumplir con las exigencias normativas, sino también potenciar el valor de los datos como base para la innovación, la eficiencia operativa y la transparencia institucional.

3.1. Políticas y normativas

En materia de seguridad de datos y protección de la información, diversas normativas nacionales e internacionales imponen obligaciones a las organizaciones. Estas regulaciones establecen directrices sobre el tratamiento de datos personales, garantizando derechos fundamentales y promoviendo medidas de seguridad para proteger la confidencialidad, integridad y disponibilidad de la información. A continuación, se presentan algunas normativas destacadas:

c) **Reglamento General de Protección de Datos (RGPD – Unión Europea)**

Reglamento (UE) 2016/679, vigente desde 2018 y su objetivo es otorgar a las personas un mayor control sobre sus datos personales y armonizar las leyes de protección de datos en Europa. A continuación, se describen sus aspectos clave:

Se aplica a organizaciones dentro y fuera de la UE que traten datos de residentes europeos.

- ✓ Establece principios como licitud, transparencia, integridad y confidencialidad.
- ✓ Reconoce derechos como acceso, rectificación, supresión y oposición.
- ✓ Exige notificación de brechas de seguridad.

- ✓ Requiere la implementación de controles técnicos y organizativos estrictos.
- ✓ Impone sanciones elevadas por incumplimiento.

d) **Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD – España)**

Ley Orgánica 3/2018, en vigor desde mayo de 2018, complementa al RGPD adaptando su aplicación al contexto español y ampliando derechos en el entorno digital. Entre sus aspectos clave se destacan:

- ✓ Desarrolla derechos como el olvido, la portabilidad y la desconexión digital.
- ✓ Regula el tratamiento de datos en entornos laborales y en el ámbito digital.
- ✓ Refuerza las obligaciones del RGPD en el contexto español.
- ✓ Establece medidas adicionales para garantizar la privacidad digital de los ciudadanos.
- ✓ Requiere políticas de seguridad, formación del personal y control de acceso a los datos.
- ✓ Establece el rol del Delegado de Protección de Datos (DPO) en ciertas organizaciones.
- ✓ Las sanciones son aplicadas por la Agencia Española de Protección de Datos (AEPD).

e) **Ley 1581 de 2012 – Protección de Datos Personales (Colombia)**

Establece disposiciones generales para la protección de datos personales y desarrolla el derecho constitucional al habeas data. Esta ley regula el tratamiento de datos por parte de entidades públicas y privadas. Sus puntos clave son:

- ✓ Clasifica los datos en públicos, privados y sensibles.
- ✓ Exige consentimiento previo, expreso e informado para el tratamiento de datos.
- ✓ Reconoce derechos como acceso, actualización, rectificación y cancelación.
- ✓ Obliga a las organizaciones a implementar medidas técnicas y administrativas de seguridad.
- ✓ Requiere la inscripción en el Registro Nacional de Bases de Datos.
- ✓ Impone la adopción de políticas internas para el manejo de datos personales.
- ✓ Supervisa el cumplimiento la Superintendencia de Industria y Comercio (SIC), que puede imponer sanciones por incumplimiento.

3.2. Gobernanza de datos

La gobernanza de datos consiste en el conjunto de normativas, procedimientos, estándares y roles diseñados para asegurar que los datos dentro de una organización sean confiables, seguros, accesibles, consistentes y se manejen de forma ética. Ha ganado relevancia a medida que las organizaciones reconocen el valor crucial de los datos como un activo fundamental para la toma de decisiones. Los componentes clave de la gobernanza de datos incluyen:

- ✓ Crear políticas y estándares para el tratamiento de los datos.
- ✓ Definir roles y responsabilidades en la gestión de la información.
- ✓ Implementar medidas que garanticen la calidad y seguridad de los datos.
- ✓ Asegurar el cumplimiento de las normativas legales vigentes.
- ✓ Facilitar el acceso adecuado y el uso responsable de la información.
- ✓ Gestionar el ciclo de vida de los datos, desde su creación hasta su eliminación.

Una gestión eficaz de los datos permite a las organizaciones maximizar el valor de su información, al tiempo que minimizan los riesgos asociados. Además, proporciona un marco coherente para administrar los datos en toda la empresa, mejorar su calidad y respaldar decisiones basadas en evidencia. En un entorno donde el volumen de datos crece constantemente, contar con una estrategia sólida de gobernanza es esencial para utilizar la información de manera efectiva y responsable.

Actualmente, existen diversos marcos de referencia que apoyan la implementación de programas de gobernanza de datos. Entre los más influyentes y utilizados por empresas y organizaciones se encuentran:

- ✓ **DAMA-DMBOK (Data Management Body of Knowledge)**

Desarrollado por DAMA International (Asociación de Gestión de Datos), este marco es reconocido como el estándar más completo para la gestión de datos. La Gobernanza de Datos es una de sus 11 áreas de conocimiento fundamentales, y se encarga de coordinar otras áreas como Calidad de Datos, Arquitectura de Datos, Seguridad de Datos y Metadatos. Incluye principios éticos y estratégicos, define estructuras organizativas con roles y responsabilidades claros, y establece flujos de

trabajo para la toma de decisiones, la resolución de problemas y la gestión del cambio. También incorpora métricas que permiten evaluar la efectividad del programa de gobernanza y la calidad de los datos.

- ✓ **COBIT (Control Objectives for Information and Related Technologies)**

Desarrollado por ISACA, este marco está orientado a la gobernanza y gestión de la información y tecnología (I&T) en las organizaciones. Aunque no se centra exclusivamente en los datos, los considera un componente clave. COBIT se enfoca en el control, la gestión de riesgos y la alineación con los objetivos estratégicos del negocio. Es ampliamente adoptado por áreas de auditoría, gestión de riesgos y tecnología, especialmente en contextos donde es necesario asegurar el cumplimiento normativo y la creación de valor a partir del uso responsable de la información.

- ✓ **Serie ISO/IEC 38500 – Gobernanza de TI**

Producida por la Organización Internacional de Normalización (ISO) y la Comisión Electrotécnica Internacional (IEC), esta serie ofrece directrices para que los líderes utilicen de manera eficiente y efectiva la tecnología de la información. Si bien su eje central es la TI, su impacto se extiende a la gobernanza de los datos, considerados un activo esencial. Entre sus principios destacan la responsabilidad, la estrategia, la adquisición, el rendimiento, el cumplimiento y el comportamiento humano, todos aplicables a la toma de decisiones relacionadas con la información.

- ✓ **ISO 8000 – Calidad de Datos**

Este estándar, aunque no constituye un marco de gobernanza completo, es fundamental por su enfoque específico en la calidad de los datos. Define principios, conceptos y procesos para medir y mejorar dicha calidad, siendo una herramienta clave

para las organizaciones que desean fortalecer su estrategia de gobernanza a partir de datos confiables y precisos.

DAMA-DMBOK y COBIT son los marcos más citados y empleados como referencia, complementados por normas ISO y adaptados a los requisitos regulatorios específicos de cada sector y región. La tendencia actual es adoptar y personalizar estos marcos según las necesidades, capacidades y objetivos de cada organización.

3.3. Seguridad de los datos

La seguridad de los datos se refiere al conjunto de políticas, prácticas, herramientas y controles diseñados para proteger la información digital contra accesos no autorizados, alteraciones, pérdida o destrucción. Su objetivo principal es garantizar la confidencialidad, integridad y disponibilidad de los datos, independientemente del medio en el que se almacenen o transmitan.

En un entorno empresarial cada vez más digitalizado e interconectado, la seguridad de los datos se ha convertido en una prioridad estratégica. Las amenazas cibernéticas, los errores humanos y los fallos técnicos representan riesgos constantes que pueden comprometer la información crítica de las organizaciones. Por ello, se requiere una gestión proactiva y sistemática que abarque todos los niveles de la organización.

Los pilares fundamentales de la seguridad de los datos incluyen:

- ✓ **Confidencialidad**: asegura que solo las personas autorizadas puedan acceder a la información.
- ✓ **Integridad**: garantiza que los datos no sean alterados de manera no autorizada o accidental.

- ✓ **Disponibilidad**: permite que los datos estén accesibles cuando se necesiten, por los usuarios pertinentes.

Las prácticas comunes de seguridad de los datos comprenden:

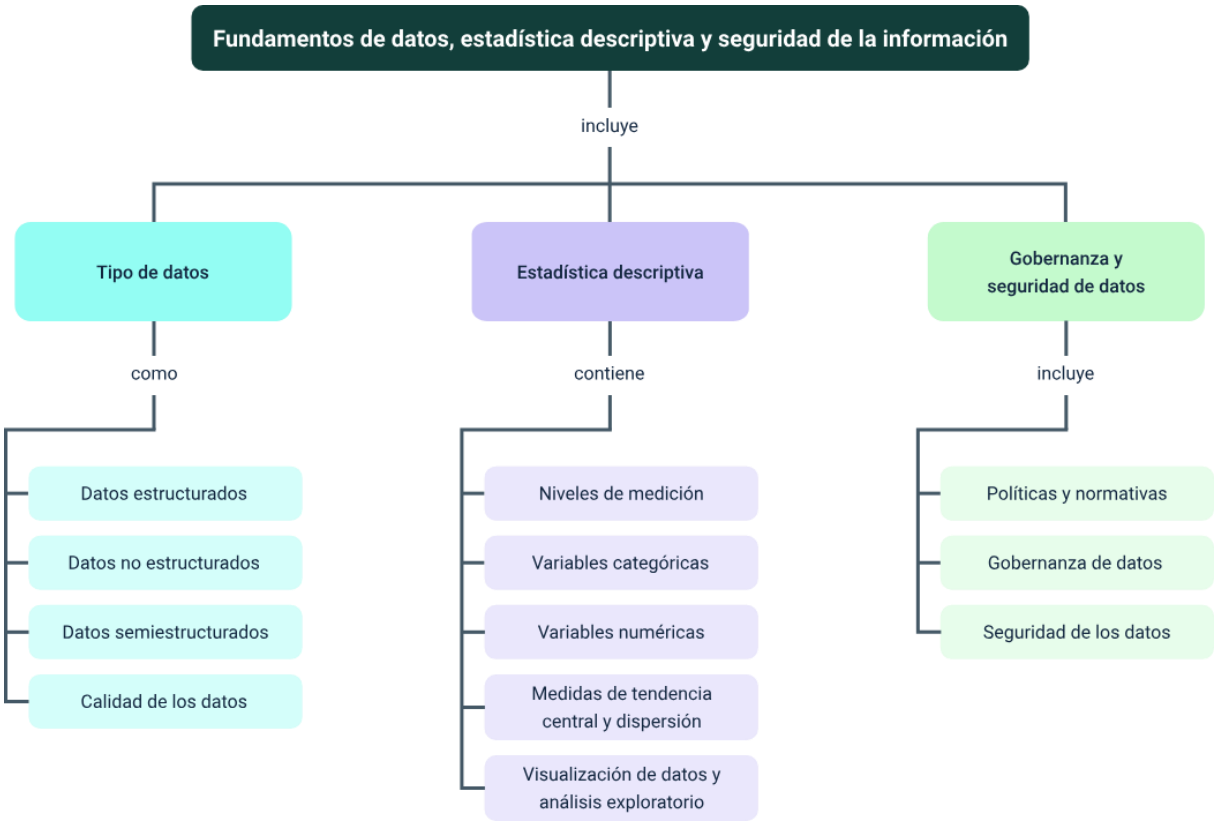
1) Implementación de controles de acceso y autenticación.
2) Uso de cifrado en tránsito y en reposo.
3) Copias de seguridad periódicas y planes de recuperación ante desastres.
4) Monitorización continua y detección de incidentes de seguridad.
5) Capacitación del personal en buenas prácticas de ciberseguridad.
6) Actualización constante de software y sistemas para corregir vulnerabilidades.

La seguridad de los datos también está estrechamente vinculada con el cumplimiento normativo. Legislaciones como el Reglamento General de Protección de Datos (RGPD) en Europa o la Ley de Protección de Datos Personales en Colombia exigen a las organizaciones implementar medidas técnicas y organizativas que aseguren la protección de la información.

Una estrategia robusta de seguridad de los datos no solo protege los activos digitales de la organización, sino que también fortalece la confianza de los usuarios, clientes y socios comerciales, lo cual es fundamental en la economía digital actual.

Síntesis

En el entorno digital actual, los datos se han convertido en un recurso estratégico fundamental para la toma de decisiones en las organizaciones. Su calidad, disponibilidad y preparación son esenciales para el funcionamiento de los algoritmos de inteligencia artificial (IA), especialmente en áreas como la visión por computadora, la salud, el procesamiento de lenguaje natural y el Internet de las cosas. Las organizaciones, tanto públicas como privadas, demandan personal capacitado en análisis, transformación y gobernanza de datos, con el fin de convertir la información en un activo de alto valor para la innovación, la eficiencia operativa y la adaptación al cambio. La IA se consolida como un motor clave de la transformación digital, cuyo desempeño depende del manejo adecuado de grandes volúmenes de datos, aunque la escasez de talento especializado representa una barrera para su adopción efectiva. Este programa formativo se articula con la misión del SENA al fortalecer competencias técnicas en ciencia de datos, impulsando una economía basada en el conocimiento, la tecnología y la innovación.


Fundamentos de datos, estadística descriptiva y seguridad de la información
incluye
Tipo de datos
como
Datos estructurados
Datos no estructurados
Datos semiestructurados
Calidad de los datos
Estadística descriptiva
contiene
Niveles de medición
Variables categóricas
Variables numéricas
Medidas de tendencia central y dispersión
Visualización de datos y análisis exploratorio
Gobernanza y seguridad de datos
incluye
Políticas y normativas
Gobernanza de datos
Seguridad de los datos

Material Complementario

Tema	Referencia	Tipo de material	Enlace del recurso
Tipo de datos	Ecosistema de Recursos Educativos SENA. (2022). Recursos y herramientas para el análisis efectivo de datos: introducción [Video]. YouTube.	Video	https://www.youtube.com/watch?v=BP8OeszBSCc
Tipo de datos	1. Tipo de datos Ecosistema de Recursos Educativos SENA. (2022). Modelo de análisis de datos [Video]. YouTube.	Video	https://www.youtube.com/watch?v=KMRGyi1ZB9k
Visualización de datos y análisis exploratorio	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

Glosario

Algoritmo: conjunto finito y ordenado de instrucciones, reglas o pasos bien definidos que se siguen para realizar una tarea específica o resolver un problema.

Aprendizaje automático (Machine Learning - ML): subcampo de la inteligencia artificial que se centra en el desarrollo de algoritmos que permiten a los sistemas informáticos aprender de los datos y mejorar su rendimiento en una tarea específica sin ser programados explícitamente para ello.

Aprendizaje profundo (Deep Learning - DL): subcampo especializado del aprendizaje automático que utiliza redes neuronales artificiales con múltiples capas (profundas) para analizar y aprender representaciones complejas directamente de grandes volúmenes de datos (como imágenes, sonido o texto).

Datos (Data): información cruda, hechos, cifras, observaciones o señales que se recopilan y registran.

Entrenamiento (Training): proceso mediante el cual un algoritmo de aprendizaje automático ajusta los parámetros internos de un modelo utilizando un conjunto de datos específico (datos de entrenamiento). El objetivo es que el modelo aprenda a identificar patrones o realizar la tarea deseada con precisión.

Modelo (de IA/ML): representación matemática o computacional que simula un proceso o tarea, creada a partir de datos mediante algoritmos de inteligencia artificial o aprendizaje automático.

Red neuronal artificial (Artificial Neural Network - ANN): modelo computacional inspirado en la estructura y funcionamiento de las redes neuronales

biológicas del cerebro. Consiste en nodos interconectados (“neuronas”) organizados en capas, que procesan información y aprenden a reconocer patrones complejos.

Referencias bibliográficas

Almeida, F., & Calistru, C. (2013). The main challenges and issues of big data management. International Journal of Research Studies in Computing, 2(1), 11–20.

Aroca, P. R., García, C. L., & López, J. J. G. (2009). Estadística descriptiva e inferencial. Revista El Auge de la Estadística en el Siglo XX, 22, 165–176.

Capa, L., García, M., Crespo, E., Palmero, D., López, R., Crespo, T., ... & Fadul, J. (2017). Análisis exploratorio de datos con SPSS. Editorial Universo Sur, Res, 3(15), 315.

CEPAL. (2023). Análisis de los modelos de gobernanza de datos en el sector público: una mirada desde Bogotá, Buenos Aires, Ciudad de México y São Paulo. Documentos de Proyectos (LC/TS.2023/71). Santiago: Comisión Económica para América Latina y el Caribe (CEPAL).

Chen, M., Mao, S., & Liu, Y. (2014). Big data: una encuesta. Redes y Aplicaciones Móviles, 19(2), 171–209.

Congreso de Colombia. (2012). Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. Diario Oficial No. 48.587.

Del Pino, S. B. (2008). Estadística descriptiva e inferencial. Innovación y Experiencias Educativas, 2–10.

Elgendy, N., & Elragal, A. (2014). Big data analytics: A literature review paper. En Advances in Data Mining. Applications and Theoretical Aspects: 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16–20, 2014. Proceedings (pp. 214–227). Springer International Publishing.

España. (2018). Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. Boletín Oficial del Estado, núm. 294, de 6 de diciembre de 2018. https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673

Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C., & Iyengar, S. S. (2016). Informática computacional de la salud en la era de los grandes datos: una encuesta. ACM Surveys on Computing (CSUR), 49(1), 12.

FreeCodeCamp.org. (2024). Data Analyst Bootcamp for Beginners (SQL, Tableau, Power BI, Python, Excel, Pandas, Projects, more). [Curso en línea].

Gehani, A., & Tariq, D. (2012). SPADE: Support for Provenance Auditing in Distributed Environments (pp. 101–120). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-35170-9_6

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis. (8.ª ed.).

Maltby, D. (2011, October). Big data analytics. En 74th Annual Meeting of the Association for Information Science and Technology (ASIST) (pp. 1–6).

Mesa Guerrero, J. A., & Caicedo Zambrano, S. J. (2020). Introducción a la estadística descriptiva.

Pyle, D. (1999). Data preparation for data mining. Morgan Kaufmann.

Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3–13.

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering, 8(12), 4727–4735.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. Science, 103(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Tsai, C. W., Lai, C. F., Chao, H. C., & Vasilakos, A. V. (2015). Big data analytics: A survey. Journal of Big Data, 2, 1–32.

Unión Europea. (2016). Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. Diario Oficial de la Unión Europea, L 119/1. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

Viedma, C. D. L. P. (2018). Estadística descriptiva e inferencial. Madrid: Ediciones IDT.

Wang, L. (2017). Heterogeneous Data and Big Data Analytics. Automatic Control and Information Sciences, 3(1), 8–15. https://doi.org/10.12691/acis-3-1-3

Zakir, J., Seymour, T., & Berg, K. (2015). Big data analytics. Issues in Information Systems, 16(2).

Créditos

Nombre	Cargo	Centro de Formación y Regional
Milady Tatiana Villamil Castellanos	Responsable Ecosistema de Recursos Educativos Digitales (RED)	Dirección General
Diana Rocío 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
Viviana Esperanza Herrera Quiñonez	Evaluadora instruccional	Centro de Comercio y Servicios - Regional Tolima
Oscar Ivan Uribe Ortiz	Diseñador web	Centro de Comercio y Servicios - Regional Tolima
José Jaime Luis Tang Pinzón	Diseñador web	Centro de Comercio y Servicios - Regional Tolima
Veimar Celis Meléndez	Desarrollador fullstack	Centro de Comercio y Servicios - Regional Tolima
Gilberto Junior Rodríguez Rodríguez	Animador y productor audiovisual	Centro de Comercio y Servicios - Regional Tolima
Jorge Eduardo Rueda Peña	Evaluadora de contenidos inclusivos y accesibles	Centro de Comercio y Servicios - Regional Tolima

Nombre	Cargo	Centro de Formación y Regional
Jorge Bustos Gómez	Validador y vinculador de recursos educativos digitales	Centro de Comercio y Servicios - Regional Tolima