guayerd

Fundamentos IA

Análisis con Python

Criterios

En colaboración con

IBM SkillsBuild

Objetivo

- Criterios para la evaluación de Datos en la realización del ETL

guayerd

En colaboración con

IBM **SkillsBuild**

Criterios de aceptación

•Cumplen con las reglas de negocio:

•Los datos deben alinearse con los requisitos y la lógica definidos para el proyecto.

•Tienen un formato y tipo de dato correctos:

•Se debe validar que los datos se ajustan al tipo de dato esperado, por ejemplo, números en columnas numéricas.

•Son consistentes y únicos:

•Los datos deben ser consistentes entre sí y no deben incluir registros duplicados, a menos que sea apropiado para el análisis.

•Son completos:

•Los datos esenciales no deben estar nulos o incompletos

Criterios de descarte

•**Datos irrelevantes:**

•Se eliminan los datos que no contribuyen a los objetivos del análisis, así como las columnas innecesarias.

•**Datos inconsistentes o erróneos:**

•Se descartan los datos que contienen valores incorrectos o que no cumplen con los estándares de calidad.

•**Datos incompletos:**

•Se pueden desechar registros que carecen de información esencial, especialmente si la ausencia de esta información puede invalidar el análisis posterior.

•**Datos duplicados:**

•Si se detectan registros duplicados que no son válidos para el análisis, se descartan para mantener la unicidad de los datos.

Implementación práctica

•**Establecer reglas de validación claras:**

•Definir las reglas de calidad y validación de los datos antes de comenzar el proceso ETL.

•**Implementar validaciones en cada etapa:**

•Realizar controles de calidad en cada paso del proceso (extracción, transformación y carga) para detectar y corregir errores tempranamente.

•**Usar una estrategia de limpieza de datos:**

•Realizar pasos de limpieza y transformación definidos para gestionar errores, valores nulos y datos innecesarios.

Implementación práctica

•Documentar los procesos:

•Mantener una documentación clara de los criterios y las transformaciones aplicadas para garantizar la transparencia y el mantenimiento del proceso.

•Realizar pruebas:

•Utilizar pruebas para validar que el proceso ETL está funcionando correctamente y que los datos finales cumplen con los estándares de calidad.

Tipo de columna	% de valores faltantes	Acción recomendada	Justificación
ID o clave primaria	> 0 %	❌ Descartar o revisar fuente	No se pueden imputar identificadores. Un solo nulo rompe la unicidad.
Campos críticos (fecha, monto, categoría obligatoria, etc.)	≤ 5 %	✅ Imputar (si posible) o dejar nulo	Pequeños porcentajes pueden corregirse sin sesgo.
	5 – 20 %	⚠️ Evaluar impacto; si el campo es necesario → imputar; si no → eliminar columna o registros.	Riesgo moderado de distorsión.
	> 20 %	❌ Descartar columna o registros según relevancia.	El campo deja de ser confiable.
Campos no críticos (comentarios, opcionales, etc.)	≤ 20 %	👍 Dejar nulos o imputar.	No afecta métricas clave.
	20 – 50 %	⚠️ Imputar si el campo es útil para modelos; sino eliminar.	Balance entre cobertura y ruido.
	> 50 %	❌ Eliminar columna.	No tiene suficiente información útil.

Columnas numéricas	≤ 10 %	Imputar con media, mediana o KNN.	Bajo sesgo.
	> 10 – 30 %	Imputar con mediana o modelo predictivo si el campo es importante; si no, eliminar.	Evaluar impacto en correlaciones.
	> 30 %	Eliminar o marcar como "poco confiable".	Riesgo alto de distorsión.
Columnas categóricas	≤ 10 %	Imputar con moda o categoría "Desconocido".	Conserva estructura de clases.
	10 – 40 %	Crear categoría “Missing” explícita.	No sesga distribución.
	> 40 %	Eliminar columna o fusionarla con otras.	Exceso de información faltante.
Duplicados (filas idénticas)	≤ 1 %	Eliminar.	Normal en fuentes múltiples.
	> 1 – 5 %	Revisar origen; eliminar si no son eventos válidos.	Posible error de integración.
	> 5 %	Auditar fuente o redefinir claves de unicidad.	Alta probabilidad de error.