Objetivo: Implementar un sistema comprensivo de validación que detecte errores, inconsistencias, y problemas de calidad en los datos.

Materiales: Archivo "ejercicio\_bloque6\_validacion.xlsx"

Instrucciones Detalladas:

Importe el archivo de datos con problemas de calidad intencionalmente introducidos. Este archivo contiene múltiples tipos de errores que debe detectar y documentar utilizando medidas DAX y consultas de validación sistemáticas.

Examine los datos para identificar visualmente algunos de los problemas presentes: fechas inválidas como "2023-02-30" y "2023-08-35", coordenadas fuera de rango como latitudes mayores a 90 grados, abundancias negativas, y campos de especies vacíos. Su tarea es crear un sistema automatizado que detecte estos y otros problemas de manera sistemática.

Desarrolle medidas DAX para validar diferentes aspectos de la calidad de datos. Comience con validaciones básicas de completitud creando medidas que calculen el porcentaje de valores no nulos en campos críticos. Por ejemplo, cree una medida llamada "Completitud\_Especies" que calcule qué porcentaje de registros tienen información de especies no vacía.

Implemente validaciones de rango para variables numéricas. Cree medidas que identifiquen coordenadas geográficas fuera de rangos válidos (latitud entre -90 y 90, longitud entre -180 y 180), abundancias negativas, y profundidades imposibles. Configure estas medidas para proporcionar tanto conteos de registros problemáticos como porcentajes del total.

Desarrolle validaciones de formato y consistencia para fechas. Cree medidas que identifiquen fechas inválidas, fechas futuras que no deberían existir en datos históricos, y secuencias temporales ilógicas. Utilice funciones DAX de fecha para verificar que los valores sean fechas válidas del calendario.

Implemente validaciones cruzadas entre variables relacionadas. Por ejemplo, cree medidas que identifiquen registros donde las coordenadas no correspondan a ubicaciones marinas cuando el tipo de ecosistema indica ambiente marino, o donde las abundancias sean extremadamente altas para las especies y métodos reportados.

Cree una tabla calculada llamada "Registros\_Problematicos" que compile todos los registros que fallan una o más validaciones. Esta tabla debe incluir el ID del registro, una descripción de cada problema identificado, y un nivel de severidad (crítico, moderado, menor) para cada tipo de error.

Desarrolle un sistema de puntuación de calidad que asigne puntuaciones numéricas a cada registro basándose en el número y severidad de los problemas identificados. Registros con puntuaciones bajas deben ser marcados para revisión prioritaria, mientras que registros con puntuaciones altas pueden ser considerados confiables para análisis.

Implemente detección avanzada de duplicados que considere no solo correspondencias exactas sino también registros que probablemente representen la misma observación con pequeñas variaciones. Utilice combinaciones de fecha, ubicación aproximada, y especie para identificar posibles duplicados que requieren revisión manual.

Cree medidas que calculen estadísticas de distribución para variables numéricas e identifiquen automáticamente valores que caen fuera de rangos esperados basándose en conocimiento ecológico. Por ejemplo, abundancias extremadamente altas que podrían indicar errores de transcripción o medición.

Desarrolle un dashboard de calidad de datos que presente visualmente todas las validaciones implementadas. Utilice gráficos de barras para mostrar la distribución de diferentes tipos de errores, mapas para mostrar la distribución espacial de problemas, y tablas para listar los registros más problemáticos.

Implemente un checklist automatizado que evalúe el sistema completo contra criterios de calidad predefinidos. Este checklist debe incluir verificaciones técnicas (tipos de datos correctos, relaciones apropiadas, ausencia de duplicados) y verificaciones científicas (coherencia taxonómica, plausibilidad ecológica, completitud de metadatos).

Configure alertas que puedan ser utilizadas para monitoreo continuo cuando se agreguen nuevos datos al sistema. Estas alertas deben notificar cuando nuevos registros violen validaciones existentes o cuando se detecten patrones anómalos que podrían indicar problemas sistemáticos.

Criterios de Evaluación:

•Implementación comprensiva de validaciones básicas de completitud y rango (25%)

•Desarrollo de validaciones cruzadas y detección de inconsistencias lógicas (25%)

•Creación de sistema de puntuación de calidad y tabla de registros problemáticos (25%)

•Desarrollo de dashboard visual de calidad de datos (15%)

•Implementación de checklist automatizado y sistema de alertas (10%)

Resultado Esperado: Un sistema robusto de control de calidad que identifique automáticamente todos los tipos de problemas presentes en los datos y proporcione herramientas para su monitoreo y corrección continua.