Preguntas y Respuestas sobre la Generación y Uso del Dataset

1. ¿Cómo se puede medir la dificultad de un dataset en el contexto de pruebas de software?

La dificultad de un dataset puede evaluarse a partir de diversos criterios. En primer lugar, el número de atributos o variables determina la complejidad dimensional: a mayor cantidad de variables, mayor dificultad para los modelos de aprendizaje. En segundo lugar, el balance de clases es un factor crítico; un dataset equilibrado entre pruebas aprobadas y fallidas facilita el entrenamiento, mientras que un fuerte desbalance lo hace más difícil de modelar. En tercer lugar, la claridad de los patrones influye directamente: si los atributos se relacionan de forma evidente con el resultado, el dataset es más sencillo; si las relaciones son ruidosas o poco lineales, la dificultad aumenta. Finalmente, pueden utilizarse medidas estadísticas como la entropía, que cuantifica la incertidumbre, o métricas de separabilidad de clases (ejemplo: F1-score, AUC), que reflejan qué tan distinguibles son los resultados para un modelo de predicción.

2. ¿Por qué generar un dataset sintético en lugar de utilizar datos reales?

El uso de datos reales en pruebas de software enfrenta limitaciones importantes: suelen ser confidenciales, heterogéneos y difíciles de compartir entre organizaciones. Por esta razón, se optó por la generación sintética del dataset, asegurando el control sobre los atributos y su distribución. Este enfoque permite modelar escenarios realistas sin comprometer información sensible, garantizar un balance adecuado entre clases y reproducir patrones de error frecuentes observados en entornos de pruebas reales.

3. ¿Qué limitaciones tiene un dataset sintético frente a uno real?

Un dataset sintético no captura en su totalidad la complejidad y las particularidades de un entorno real de pruebas. Factores como la interacción entre módulos del sistema, defectos inesperados o sesgos propios de los equipos de calidad pueden no estar reflejados en la simulación. Sin embargo, su ventaja radica en la validez metodológica que aporta: permite experimentar con distribuciones controladas, balancear los datos y verificar la factibilidad de un modelo predictivo de manera inicial, antes de aplicar la propuesta sobre información real.

4. ¿Cómo se garantiza que el dataset generado sea representativo?

El diseño del dataset se fundamentó en factores que influyen directamente en el resultado de una prueba de software, tales como la complejidad del caso, el origen de la ejecución (manual o automatizada), la metodología empleada y la prioridad asignada. A cada combinación se le asignó una probabilidad de fallo coherente con la práctica profesional (ejemplo: mayor probabilidad de fallar en pruebas manuales, críticas o de alta complejidad). De esta forma, aunque sintético, el dataset mantiene una representatividad funcional frente a escenarios de la industria.

5. ¿Cómo puede una empresa externa generar su propio dataset útil para este modelo?

Una empresa interesada en aplicar este enfoque puede construir su dataset a partir de sus registros históricos de pruebas. El proceso consiste en mapear sus datos internos a los atributos del modelo propuesto (tipo de prueba, metodología, fase, origen, complejidad, prioridad y resultado). Si ciertos atributos no están disponibles en su formato original, es posible realizar equivalencias, por ejemplo: la “severidad del defecto” puede mapearse a “prioridad”, y el “esfuerzo estimado” puede interpretarse como “complejidad”. Posteriormente, la información debe normalizarse para mantener consistencia (unificar escalas y nomenclaturas), y asegurarse de que los resultados (aprobada/fallida) estén correctamente etiquetados. En caso de datos insuficientes, la empresa puede generar instancias sintéticas siguiendo reglas similares a las del modelo propuesto, garantizando un dataset balanceado y representativo.

6. ¿Cómo evaluar la calidad del dataset generado?

La calidad del dataset se evalúa en tres dimensiones principales:

1. Completitud: cada registro debe contener todos los atributos definidos, sin valores vacíos ni nulos.
2. Consistencia: no deben existir contradicciones internas (por ejemplo, pruebas críticas que aparecen sistemáticamente como aprobadas).
3. Balance de clases: la proporción de pruebas aprobadas y fallidas debe ser razonablemente equilibrada para evitar sesgos en el aprendizaje del modelo.

De esta manera, se asegura que el dataset no solo cumpla con la estructura formal, sino que también sea confiable y útil en un contexto predictivo.

7. ¿Cómo medir la utilidad del dataset para el modelo de predicción propuesto?

La utilidad del dataset se mide mediante la capacidad de entrenar modelos de aprendizaje automático con resultados consistentes. Esto implica que los atributos seleccionados realmente aportan información diferenciadora y que los patrones en los datos permiten predecir con cierta fiabilidad el resultado de una prueba. En la práctica, esto puede verificarse evaluando métricas como precisión, recall y F1-score del modelo entrenado.