

Optimización de la Calidad en la Producción de Chips

Un Enfoque Basado en Datos

Pasos a seguir en un proceso de análisis de datos

1. Preguntar

1. **¿Qué tipo de respuestas basadas en datos quiere su cliente y qué le pide que consiga?**
 - Cliente: Una fábrica de chips que busca identificar patrones en los defectos de producción para mejorar la calidad y reducir costos.
 - Objetivo: Presentar un análisis para predecir productos defectuosos y priorizar áreas de mejora.
 2. **¿Cuáles son los factores clave que intervienen en la tarea empresarial que está investigando?**
 - Factores: Sensores de producción, condiciones ambientales, y métricas del proceso (temperatura, presión, tiempo).
 3. **¿Qué tipo de datos serán apropiados para su análisis?**
 - Datos históricos de producción que incluyan parámetros de los sensores y etiquetas de calidad (defectuoso/no defectuoso).
 4. **¿Dónde obtendrá esos datos?**
 - Usaremos los datos previamente estudiados de <https://www.kaggle.com/datasets/paresh2047/uci-semcom> y simularemos un contexto realista de adquisición (p. ej., recopilados de sensores IoT en la planta).
 5. **¿Quién es su público y qué materiales le ayudarán a presentárselo efectivamente?**
 - Público: Directores de calidad y operaciones de la fábrica.
 - Materiales: Visualizaciones de datos claras, resultados de los modelos predictivos y recomendaciones prácticas.
-

2. Preparar

Revisaremos y documentaremos los datos existentes, asegurándonos de que sean adecuados para este caso práctico.

1. Descripción de las fuentes de datos:

- Conjunto de datos “Kaggle” <https://www.kaggle.com/datasets/paresh2047/uci-semcom>, Propietario: Paresh Mathur, uci-secom.csv. Contiene registros de sensores y etiquetas de calidad.
- Variables: Sensores (X1, X2, ..., X596) y etiqueta (1 = defectuoso, 0 = no defectuoso).

2. Integridad de los datos:

- Verificar valores faltantes, duplicados y rangos inconsistentes.
 - Documentar cualquier problema y su solución (p. ej., imputación, eliminación de outliers).
-

3. Procesar

1. Script básico:

- Entrenaremos un modelo simple (p. ej., Logistic Regression) para establecer una línea base de predicción de productos defectuosos.

2. Script con hiperparámetros:

- Optimización de parámetros clave en modelos más complejos, como Random Forest o Gradient Boosting.

3. Script XGBoost:

- Aplicación de XGBoost para maximizar la precisión y analizar la importancia de las características.

4. Preparación de los datos para los modelos:

- Se imputaron los valores faltantes con la mediana de cada columna para evitar sesgos en las características numéricas.
 - Random Forest puede manejar datos imputados con facilidad porque utiliza árboles de decisión.
 - XGBoost maneja valores faltantes automáticamente durante el entrenamiento. Busca el valor óptimo en los nodos de decisión cuando encuentra un valor NaN
-

4. Analizar

Realizaremos un análisis profundo de los resultados obtenidos:

1. **Métricas clave:**

- Precisión, recall, F1-score y matriz de confusión para evaluar cada modelo.

2. **Análisis de importancia de características:**

- Identificar qué sensores son críticos para predecir defectos.
-

5. Compartir

Prepararemos una presentación clara con los resultados y visualizaciones relevantes:

1. **Visualizaciones:**

- Gráficas de importancia de características.
- Comparación de desempeño entre los tres modelos.
- Ejemplo de predicciones en datos reales.

2. **Informe final:**

- Resumen de hallazgos clave y recomendaciones para mejorar la calidad del proceso.
-

6. Actuar

Basándonos en el análisis, propondríamos acciones específicas, como:

1. **Optimización de sensores:**

- Monitorear más de cerca los sensores críticos identificados.

2. **Automatización del control de calidad:**

- Implementar el modelo XGBoost en la línea de producción para alertar sobre posibles defectos en tiempo real.

3. **Próximos pasos:**

- Recopilar más datos y explorar nuevos factores (p. ej., cambios en el equipo o en el entorno).