ANÁLISIS AUTOMATIZADO DE CALIDAD DE VINOS BLANCOS

Herramienta de Análisis de Datos con Inteligencia Artificial

UNIVERSIDAD
Panamericana

Materia: Minería de Datos

Equipo:

TEAM SOPO

Integrantes:

Sebastián Avilez Hernández
Gabriel Zaid Gutiérrez González
Yahwthani Morales Gómez
Gabriel Torres Zacarias

Profesor:

Ronaldo Jesús Ruíz Álvarez
Oswaldo Rojas Bravo

Fecha:

5 de diciembre de 2025

Índice

1. Resumen Ejecutivo

El presente proyecto desarrolla una herramienta automatizada para el análisis de datos que genera reportes profesionales utilizando técnicas de minería de datos, machine learning y generación de contenido mediante la API de Claude (Anthropic).

1.1. Dataset Analizado

- **Nombre:** Wine Quality - White Wine
- **Fuente:** UCI Machine Learning Repository
- **Dimensiones:** 4,898 registros × 12 columnas
- **Variable objetivo:** Quality (escala 3-9)

1.2. Principales Hallazgos

1. El **contenido alcohólico** es el predictor más importante de calidad (correlación: 0.436)
2. Existe una fuerte correlación negativa entre **densidad y alcohol** (-0.780)
3. El 74.62 % de los vinos se concentra en calidad media (5-6)
4. El modelo **Random Forest** alcanza 67.55 % de accuracy
5. Se identificaron 1,063 outliers distribuidos en múltiples variables

2. Objetivo del Proyecto

Desarrollar una herramienta capaz de analizar automáticamente cualquier archivo CSV y generar un reporte profesional utilizando:

- Análisis estadístico descriptivo e inferencial
- Visualizaciones profesionales
- Modelos de Machine Learning
- Generación de insights mediante la API de Claude (Anthropic)

3. Metodología

3.1. Diagrama de Flujo del Proceso

El proceso de análisis sigue el siguiente flujo:

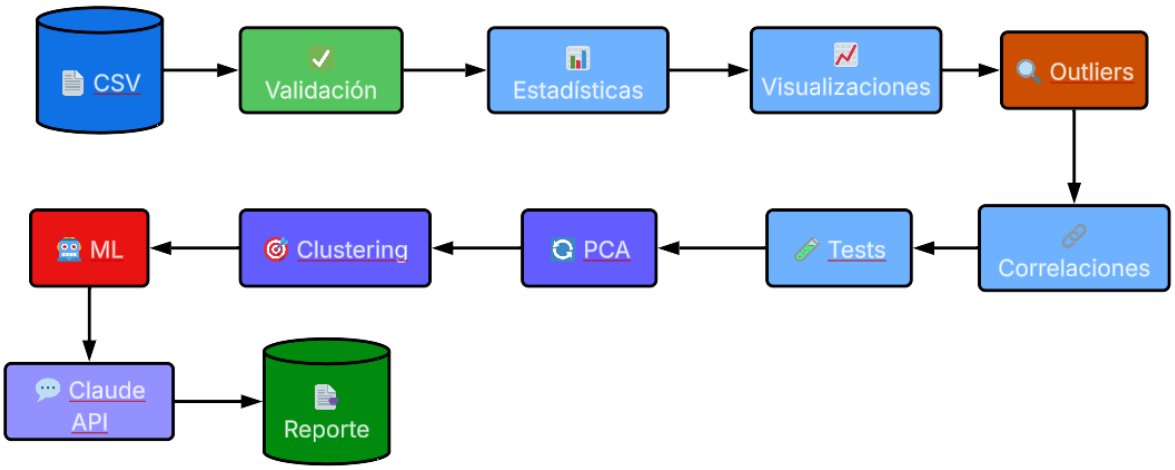

Figura 1: Diagrama de flujo del proceso de análisis

3.2. Fases del Análisis

Cuadro 1: Fases del análisis automatizado

Fase	Descripción	Herramientas
1	Carga y validación de datos	pandas
2	Análisis de valores faltantes	pandas, seaborn
3	Estadísticas descriptivas	pandas, numpy
4	Tests de normalidad	scipy.stats
5	Visualizaciones	matplotlib, seaborn
6	Detección de outliers	IQR, Z-score
7	Análisis de correlaciones	Pearson, Spearman
8	Tests estadísticos	ANOVA, Kruskal-Wallis
9	PCA	sklearn.decomposition
10	Clustering	K-Means
11	Machine Learning	sklearn (5 modelos)
12	Feature Importance	Random Forest
13	Regresión	Linear, RF Regressor
14	Generación de insights	API Claude (Anthropic)

4. Construcción del Prompt para Claude

4.1. Estrategia de Diseño del Prompt

El prompt fue diseñado siguiendo las mejores prácticas de prompt engineering para obtener respuestas estructuradas y profesionales:

4.1.1. Elementos Clave del Prompt

1. **Rol definido:** Se establece que Claude actúa como experto en análisis de datos
2. **Estructura clara:** Se especifican las 8 secciones requeridas
3. **Datos concretos:** Se incluye el resumen completo del análisis
4. **Formato de salida:** Se solicita respuesta en español y formato profesional

4.2. Código del Prompt

```
def generar_insights_con_claude(resumen, api_key):
    client = Anthropic(api_key=api_key)

    prompt = f"""
    Eres un experto en analisis de datos y ciencia de datos.
    Analiza el siguiente resumen completo de un analisis
    de datos sobre calidad de vinos blancos.

    Genera un REPORTE EJECUTIVO PROFESIONAL que incluya:

    1. RESUMEN EJECUTIVO (3-4 parrafos)
    2. ANALISIS DE LA CALIDAD DE DATOS
    3. HALLAZGOS ESTADISTICOS CLAVE (5 hallazgos)
    4. ANALISIS DE MACHINE LEARNING
    5. ANALISIS DE CLUSTERING Y PCA
    6. RECOMENDACIONES DE PREPROCESAMIENTO (5)
    7. LIMITACIONES DEL ANALISIS
    8. CONCLUSIONES Y PROXIMOS PASOS

    DATOS DEL ANALISIS:
    {resumen}

    Responde en espanol, de manera profesional.
    """

    response = client.messages.create(
        model="claude-sonnet-4-20250514",
        max_tokens=6000,
        messages=[{"role": "user", "content": prompt}]
    )

    return response.content[0].text
```

Listing 1: Función para generar insights con Claude

4.3. Resumen Enviado a Claude

El resumen incluye las siguientes secciones con datos específicos:

- Información del dataset (dimensiones, tipos de variables)
- Calidad de datos (valores faltantes, outliers)
- Distribución de la variable objetivo
- Estadísticas descriptivas completas
- Resultados de tests de normalidad
- Matriz de correlaciones y top correlaciones
- Resultados de tests estadísticos (ANOVA, Kruskal-Wallis)
- Análisis de PCA (varianza explicada, loadings)
- Resultados de clustering (perfiles de clusters)
- Métricas de modelos de Machine Learning
- Feature importance
- Resultados de regresión
- Comparación entre grupos de calidad

5. Análisis Técnico Detallado

5.1. Descripción del Dataset

Cuadro 2: Variables del dataset Wine Quality

Variable	Descripción	Media	Std
fixed acidity	Acidez fija (g/L)	6.85	0.84
volatile acidity	Acidez volátil (g/L)	0.28	0.10
citric acid	Ácido cítrico (g/L)	0.33	0.12
residual sugar	Azúcar residual (g/L)	6.39	5.07
chlorides	Cloruros (g/L)	0.046	0.022
free sulfur dioxide	SO2 libre (mg/L)	35.31	17.01
total sulfur dioxide	SO2 total (mg/L)	138.36	42.50
density	Densidad (g/cm^3)	0.994	0.003
pH	pH	3.19	0.15
sulphates	Sulfatos (g/L)	0.49	0.11
alcohol	Alcohol (% vol)	10.51	1.23
quality	Calidad (0-10)	5.88	0.89

5.2. Distribución de la Variable Objetivo

Cuadro 3: Distribución de calidad de vinos

Calidad	Frecuencia	Porcentaje	Acumulado
3	20	0.41 %	0.41 %
4	163	3.33 %	3.74 %
5	1,457	29.75 %	33.49 %
6	2,198	44.88 %	78.37 %
7	880	17.97 %	96.34 %
8	175	3.57 %	99.91 %
9	5	0.10 %	100.00 %

Categorías agregadas:

- Baja (3-4): 183 vinos (3.74 %)
- Media (5-6): 3,655 vinos (74.62 %)
- Alta (7-9): 1,060 vinos (21.64 %)

5.3. Análisis de Correlaciones

5.3.1. Top 5 Correlaciones más Fuertes

Cuadro 4: Correlaciones más fuertes en el dataset

#	Variable 1	Variable 2	Correlación
1	residual sugar	density	**0.839**
2	density	alcohol	**-0.780**
3	free sulfur dioxide	total sulfur dioxide	0.616
4	total sulfur dioxide	density	0.530
5	residual sugar	alcohol	-0.451

5.3.2. Correlaciones con Calidad

Cuadro 5: Correlación de variables con calidad del vino

Variable	Pearson	Spearman
alcohol	**0.436**	**0.440**
density	-0.307	-0.348
chlorides	-0.210	-0.315
volatile acidity	-0.195	-0.197
total sulfur dioxide	-0.175	-0.197

5.4. Resultados de Machine Learning

Cuadro 6: Comparación de modelos de clasificación

Modelo	Accuracy	Precision	Recall	F1	CV
Random Forest	**67.55 %**	68.40 %	67.55 %	66.44 %	66.28 %
Decision Tree	59.39 %	59.54 %	59.39 %	59.42 %	57.53 %
Gradient Boosting	57.55 %	57.98 %	57.55 %	56.20 %	59.11 %
Logistic Regression	54.90 %	53.14 %	54.90 %	51.55 %	53.52 %
KNN	52.65 %	51.99 %	52.65 %	51.91 %	54.24 %

5.5. Feature Importance

Cuadro 7: Importancia de variables (Random Forest)

Ranking	Variable	Importancia
1	**alcohol**	**11.65 %**
2	density	10.30 %
3	volatile acidity	10.05 %
4	free sulfur dioxide	9.22 %
5	total sulfur dioxide	9.08 %
6	pH	8.66 %
7	chlorides	8.64 %
8	residual sugar	8.63 %
9	citric acid	8.18 %
10	sulphates	7.83 %
11	fixed acidity	7.76 %

6. Visualizaciones Generadas

El análisis genera 16 visualizaciones profesionales guardadas en la carpeta `outputs/`:

Cuadro 8: Lista de visualizaciones generadas

#	Archivo	Descripción
1	01_heatmap_valores_faltantes_v2.png	Mapa de valores nulos
2	02_qq_plots_v2.png	Q-Q plots de normalidad
3	03_histogramas_kde_v2.png	Distribuciones con KDE
4	04_boxplots_v2.png	Boxplots con outliers
5	05_violin_plots_v2.png	Violin por categoría
6	06_barras_quality_v2.png	Distribución de calidad
7	07_heatmaps_correlacion_v2.png	Pearson y Spearman
8	08_correlacion_con_quality_v2.png	Barras horizontales
9	09_pca_varianza_v2.png	Scree plot y acumulada
10	10_pca_scatter_v2.png	PC1 vs PC2
11	11_clustering_elbow_v2.png	Método del codo
12	12_clustering_pca_v2.png	Clusters en espacio PCA
13	13_confusion_matrix_v2.png	Matriz de confusión
14	14_comparacion_modelos_v2.png	Métricas de modelos
15	15_feature_importance_v2.png	Importancia de variables
16	16_regression_predictions_v2.png	Predicciones vs reales

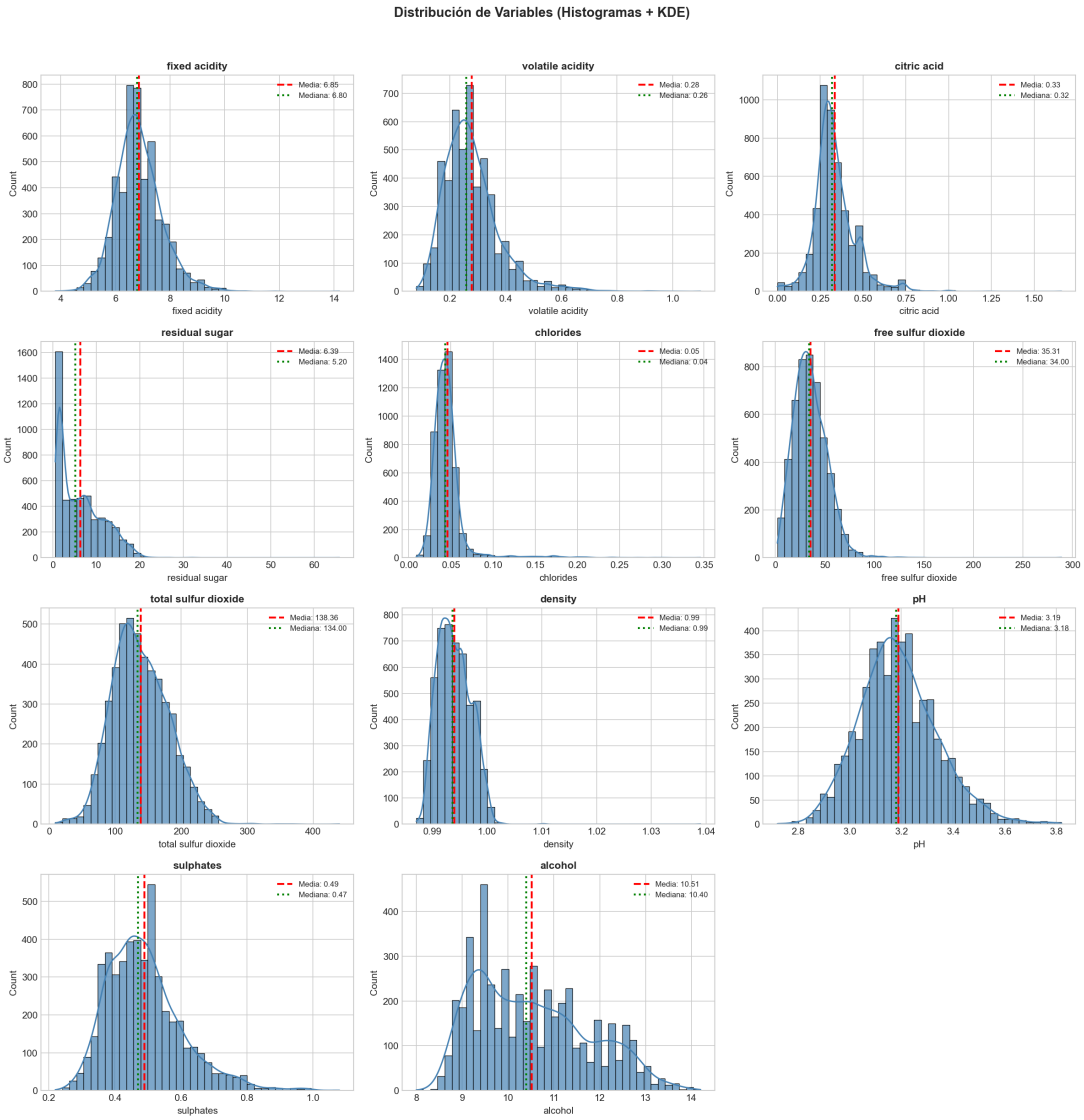

Figura 2: Histogramas de distribución de variables

7. Insights Generados por Claude

7.1. Resumen del Reporte Ejecutivo

El análisis con Claude API generó un reporte ejecutivo de 8 secciones con los siguientes puntos destacados:

7.1.1. Hallazgos Principales

1. **Supremacía del alcohol como predictor:** El contenido alcohólico muestra la correlación más fuerte con calidad (r=0.436) y la mayor importancia en modelos predictivos (11.65 %).

2. **Relación inversa densidad-alcohol:** Existe una correlación negativa muy fuerte (-0.780) que explica el 29.29 % de la varianza en el primer componente principal.

3. **Significancia estadística universal:** Las 11 variables muestran diferencias significativas entre grupos de calidad ($p<0.05$).

4. **Ausencia de normalidad:** Ninguna variable sigue distribución normal, siendo chlorides la más asimétrica (skewness=5.02).

5. **Estructura de correlación compleja:** Se identifican tres ejes principales: densidad-alcohol-azúcar, sulfuros, y acidez-pH.

7.1.2. Recomendaciones de Preprocesamiento

1. Implementar detección de outliers multivariada (Isolation Forest)

2. Aplicar transformación logarítmica a chlorides

3. Combinar SMOTE con undersampling para balanceo de clases

4. Crear variables derivadas (ratios alcohol/azúcar)

5. Usar validación cruzada estratificada de 10 folds

8. Estructura del Código

8.1. Archivos del Proyecto

```
Ciencia_Datos_P3/
|-- .env                        # API key (no subir a Git)
|-- .gitignore                  # Archivos excluidos
|-- README.md                   # Documentacion
|-- requirements.txt            # Dependencias
|-- data/
|   |-- winequality-white.csv
|-- automated_analysis_v1.py    # Notebook sin ML
|-- automated_analysis_v2.ipynb # Notebook con ML
|-- outputs/                    # 16 visualizaciones
|-- reports/
|   |-- resumen_analisis.txt
|   |-- insights_claude_v2.txt
```

8.2. Dependencias

```
pandas>=2.0.0
numpy>=1.24.0
matplotlib>=3.7.0
seaborn>=0.12.0
scikit-learn>=1.3.0
scipy>=1.11.0
anthropic>=0.39.0
```

```
python-dotenv>=1.0.0
jupyter>=1.0.0
```

Listing 2: requirements.txt

9. Instrucciones de Replicación

9.1. Configuración del Entorno

```
# Clonar repositorio
git clone https://github.com/YahwthaniMG/Ciencia_Datos_P3.git
cd Ciencia_Datos_P3

# Crear entorno virtual
python -m venv venv
source venv/bin/activate  # Linux/Mac
# venv\Scripts\activate   # Windows

# Instalar dependencias
pip install -r requirements.txt

# Crea un archivo .env
example.env
# Editar .env con tu API key de Anthropic
```

Listing 3: Instalación del proyecto

9.2. Ejecución del Análisis

```
jupyter notebook automated_analysis_v2.ipynb
```

Listing 4: Ejecutar el análisis

10. Limitaciones del Proyecto

1. **Restricciones del dataset:**
   - Ausencia de información contextual (región, variedad de uva, año)
   - Solo variables fisicoquímicas, sin factores organolépticos
2. **Desbalance de clases:**
   - 74.62 % de datos en calidades medias (5-6)
   - Ratio de desbalance aproximado de 440:1 entre clase mayoritaria y minoritaria
3. **Subjetividad en evaluación:**

- La calidad se basa en puntuaciones sensoriales subjetivas
- Posible variabilidad entre evaluadores

4. **Limitaciones del modelo:**
   - Accuracy máximo de 67.55 %
   - Dificultad para predecir clases extremas

11. Conclusiones

11.1. Objetivos Cumplidos

✓ Desarrollo de herramienta automatizada de análisis de CSV

✓ Implementación de análisis estadístico completo

✓ Generación de 16 visualizaciones profesionales

✓ Entrenamiento y comparación de 5 modelos de ML

✓ Integración exitosa con API de Claude

✓ Generación automática de reportes ejecutivos

11.2. Hallazgos Principales

El análisis confirma que el contenido alcohólico es el predictor dominante de calidad en vinos blancos. La estructura de correlación revela relaciones fisicoquímicas coherentes con los principios enológicos. Los modelos de Machine Learning alcanzan rendimiento moderado (67.55 % accuracy), limitados principalmente por el desbalance severo de clases.

11.3. Trabajo Futuro

- Incorporar técnicas de deep learning para interacciones no lineales
- Implementar SMOTE para balanceo de clases
- Desarrollar modelos específicos por cluster
- Validar con expertos enólogos
- Extender a otros tipos de vino (tinto)

12. Referencias

1. Kaggle: White Wine Quality. https://www.kaggle.com/datasets/piyushagni5/white-wine-quality
2. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.
3. Documentación de Anthropic - Claude API
   https://docs.anthropic.com
4. Scikit-learn: Machine Learning in Python
   https://scikit-learn.org
5. Pandas Documentation
   https://pandas.pydata.org/docs/

A. Código Completo del Prompt

```
print("[16/16] Generando insights con Claude...")

api_key = os.getenv("ANTHROPIC_API_KEY")

if api_key:
    try:
        from anthropic import Anthropic

        client = Anthropic(api_key=api_key)

        prompt = f"""
Eres un experto en analisis de datos, ciencia de datos y machine learning.
Analiza el siguiente resumen completo de un analisis de datos sobre calidad de vinos blancos.

Genera un REPORTE EJECUTIVO PROFESIONAL Y DETALLADO que incluya:

1. RESUMEN EJECUTIVO (3-4 parrafos)
   - Descripcion del dataset y contexto del problema
   - Calidad general de los datos
   - Principales descubrimientos

2. ANALISIS DE LA CALIDAD DE DATOS
   - Evaluacion de valores faltantes
   - Analisis de outliers y su impacto
   - Distribucion de la variable objetivo y sus implicaciones

```

```
3. HALLAZGOS ESTADISTICOS CLAVE (5 hallazgos)
   - Patrones y tendencias descubiertos
   - Relaciones significativas entre variables
   - Interpretacion de los tests estadisticos

4. ANALISIS DE MACHINE LEARNING
   - Evaluacion del rendimiento de los modelos
   - Interpretacion del feature importance
   - Capacidad predictiva y limitaciones

5. ANALISIS DE CLUSTERING Y PCA
   - Interpretacion de los clusters encontrados
   - Significado de los componentes principales
   - Patrones de agrupamiento de vinos

6. RECOMENDACIONES DE PREPROCESAMIENTO (5 recomendaciones)
   - Tecnicas para mejorar la calidad de datos
   - Estrategias para manejar outliers
   - Transformaciones sugeridas
   - Estrategias para el desbalance de clases

7. LIMITACIONES DEL ANALISIS
   - Restricciones del dataset
   - Consideraciones metodologicas
   - Posibles sesgos

8. CONCLUSIONES Y PROXIMOS PASOS
   - Sintesis de hallazgos principales
   - Recomendaciones para investigacion futura
   - Aplicaciones practicas

DATOS DEL ANALISIS:
{resumen_completo}

Responde en espanol, de manera profesional y detallada. Usa datos
   especificos del analisis para respaldar cada punto.
"""

        response = client.messages.create(
            model="claude-sonnet-4-20250514",
            max_tokens=8000,
            messages=[{"role": "user", "content": prompt}]
        )

        insights = response.content[0].text

        with open('reports/insights_claude_v2.txt', 'w', encoding='
           utf-8') as f:
```

```
            f.write("REPORTE EJECUTIVO - ANALISIS DE CALIDAD DE
                VINOS BLANCOS\n")
            f.write("="*70 + "\n\n")
            f.write(insights)

        print("   Insights guardados en: reports/insights_claude_v2
            .txt")
        print("\n" + "="*70)
        print("REPORTE EJECUTIVO GENERADO")
        print("="*70)
        print(insights[:2000] + "...\n[Ver archivo completo en
            reports/insights_claude_v2.txt]")

    except Exception as e:
        print(f"   Error al generar insights: {e}")
else:
    print("   API key no encontrada. Configura ANTHROPIC_API_KEY en
        .env")
```

Listing 5: Llamada a Claude

El prompt completo enviado a Claude incluye todas las estadísticas generadas por el análisis. La estructura del prompt sigue el formato detallado en la Sección 4.

B. Ejemplo de Salida de Claude

REPORTE EJECUTIVO: ANÁLISIS DE CALIDAD DE VINOS BLANCOS

1. RESUMEN EJECUTIVO
El dataset analizado contiene información de 4,898 vinos blancos con 12 variables fisicoquímicas que determinan su calidad. Los datos presentan una excelente integridad con 0 % de valores faltantes, lo que indica una recolección de datos sistemática y completa. Las variables incluyen medidas de acidez, azúcares, sulfitos, densidad, pH, sulfatos y contenido alcohólico, todas ellas fundamentales en la caracterización enológica.

La distribución de calidad muestra un patrón típicamente normal, concentrándose en valores medios (calidades 5-6 representan el 74.63 % de la muestra), con muy pocos ejemplos de vinos de calidad excepcional (calidad 9: 0.10 %) o deficiente (calidad 3: 0.41 %). Esta distribución sugiere que el dataset representa principalmente vinos de calidad comercial estándar, lo que podría limitar la capacidad de modelar vinos premium o de baja calidad. .....
Ver el reporte ejecutivo completo generado por la API de Claude en la carpeta `reports` como `insights_claude_v2.txt`.