Proyectos de Machine Learning y Series Temporales

Análisis Completo de Datasets UCI

Tomás Travis Alonso Cremnitz

Septiembre 2025

Contenido

Planteamiento del Caso

Objetivo General

Desarrollar e implementar dos sistemas completos de Machine Learning utilizando datasets reales de UCI ML Repository

TEST 1: Clasificación

- **Dataset**: Wine Quality UCI (6,497 muestras)
- **Resultado**: 69.31 % accuracy con Random Forest
- **Features**: 11 propiedades fisicoquímicas
- **Modelos**: RF, SVM, XGBoost

TEST 2: Series Temporales

- **Dataset**: Gas Sensor Array Drift UCI (144 obs.)
- **Resultado**: 2.58 % MAPE con Random Forest
- **Target**: sensor_drift (degradación sensores)
- **Modelos**: RF, XGBoost, ARIMA

Metodología de Trabajo

1. **Análisis Exploratorio de Datos (EDA)**
   - Carga y validación de datasets UCI reales
   - Análisis estadístico descriptivo completo
   - Visualización de distribuciones y correlaciones
2. **Preprocesamiento de Datos**
   - Limpieza y tratamiento de valores faltantes
   - Escalado de características
   - División estratificada train/test
3. **Modelado y Evaluación**
   - Implementación de múltiples algoritmos
   - Validación cruzada y métricas robustas
   - Comparación de rendimiento

Dataset UCI Wine Quality

Características del Dataset

- **Fuente**: UCI ML Repository
- **Total**: 6,497 muestras reales
- **Vinos tintos**: 1,599 (24.6 %)
- **Vinos blancos**: 4,898 (75.4 %)
- **Features**: 11 propiedades fisicoquímicas
- **Target**: Calidad (escala 3-9)

Calidad	Muestras
3	30
4	216
5	2,138
6	2,836
7	1,079
8	193
9	5

Cuadro: Distribución de calidades

Principales Hallazgos del EDA

- Distribución centrada en calidades medias (5-7)
- Desbalance significativo en clases extremas (3, 9)

Modelos de Clasificación Implementados

Modelo	Accuracy	F1-Score	Tiempo
Random Forest	**69.31 %**	**68.30 %**	∼2s
XGBoost	66.62 %	65.72 %	∼5s
SVM	57.15 %	53.80 %	∼8s

Cuadro: Resultados de clasificación en Wine Quality UCI

Random Forest (Mejor Modelo)

- **Hiperparámetros**: 100 árboles, max_depth=10
- **Ventajas**: Robusto, interpretable
- **Feature Importance**: Alcohol, volatil acidity principales

Métricas de Validación

- **Validación estratificada**: 5-fold CV
- **F1-Score weighted**: Apropiado para desbalance
- **Matriz de confusión**: Análisis por clase

Dataset UCI Gas Sensor Array Drift

Características del Dataset

- **Fuente**: UCI ML Repository
- **Total**: 144 observaciones semanales
- **Período**: Enero 2008 - Octubre 2010
- **Variables**: 16 sensores químicos
- **Target**: sensor_drift (degradación)
- **Frecuencia**: Mediciones semanales

Compliance TEST 2

- ✓ Dataset NO financiero
- ✓ Variable NO estacional
- ✓ Forecasting implementado

Métrica	Valor
Media drift	3.40
Std drift	1.23
Min drift	1.00
Max drift	6.00
Sensores	16

Cuadro: Estadísticas sensor_drift

Modelos de Forecasting Implementados

Modelo	MAE	RMSE	MAPE	Tiempo
Random Forest	**0.128**	**0.159**	**2.58 %**	∼0.06s
XGBoost	0.144	0.171	2.88 %	∼0.11s
ARIMA(1,1,1)	0.479	0.551	10.12 %	∼0.02s

Cuadro: Resultados de forecasting en Gas Sensor Array Drift

Random Forest (Modelo Principal)

- **Configuración**: 100 estimadores
- **Performance**: 2.58 % MAPE excelente
- **Ventajas**: Robusto, maneja no-linealidad

Performance Destacada

- **MAPE 2.58 %**: Excelente para series temporales
- **144 observaciones**: Dataset compacto
- **Reproducible**: Scripts CLI automatizados

Arquitectura de los Proyectos

Estructura Modular

- **src/**: Código fuente organizado en paquetes
- **notebooks/**: Análisis interactivos (EDA + Modeling)
- **tests/**: Framework de testing con pytest
- **data/**: Pipeline raw → processed → final

Herramientas CLI

TEST 1 - Clasificación:

- `train_model.py`: Entrenamiento
- `inference.py`: Predicciones

TEST 2 - Series Temporales:

- `train_model.py`: Múltiples modelos
- `forecast.py`: Predicción de drift

Stack Tecnológico

- **Core**: pandas, scikit-learn, statsmodels
- **Deep Learning**: TensorFlow/Keras
- **Viz**: matplotlib, seaborn, plotly

Comparación de Resultados

Proyecto	Mejor Modelo	Métrica Principal	Dataset	Compliance
TEST 1	Random Forest	69.31 % Accuracy	Wine Quality (6,497)	✓ UCI Real
TEST 2	Random Forest	2.58 % MAPE	Gas Sensor Drift (144)	✓ No financiero

TEST 1: Clasificación

Fortalezas:

- Accuracy superior al 65 % baseline
- F1-Score balanceado para clases desbalanceadas
- Feature importance interpretable

Desafíos:

- Desbalance en clases extremas (3, 9)
- Variabilidad en evaluación humana

TEST 2: Forecasting

Fortalezas:

- MAPE ¡ 3 % para datos industriales
- Dataset compacto pero representativo
- Modelos ML superan métodos tradicionales

Desafíos:

- Dataset pequeño (144 observaciones)
- Necesidad de más datos temporales

Análisis de Features y Variables

Wine Quality - Feature Importance

Top 5 Variables (Random Forest):

1. **alcohol**: 18.3 %
2. **volatile acidity**: 14.7 %
3. **sulphates**: 12.1 %
4. **total sulfur dioxide**: 10.8 %
5. **density**: 9.4 %

Insight: El alcohol es el predictor más fuerte de calidad, seguido de propiedades relacionadas con acidez y conservantes.

Gas Sensor Drift - Análisis Temporal

Características sensor_drift:

- **Dataset**: 144 observaciones semanales (2008-2010)
- **Rango valores**: 1.0 - 6.0 (degradación sensores)
- **Media**: 3.40, **Std**: 1.23
- **Compliance**: Variable no estacional confirmada

Insight: Random Forest captura patrones no lineales mejor que ARIMA tradicional (2.58 % vs 10.12 % MAPE).

Trabajo Futuro y Mejoras

Mejoras Inmediatas

TEST 1:

- Ensemble methods (RF + XGBoost)
- SMOTE para balancear clases extremas
- Feature engineering: ratios químicos

TEST 2:

- Ensemble methods (RF + XGBoost)
- Análisis con más datos temporales
- LSTM para patrones complejos

Extensiones Avanzadas

- **MLOps**: CI/CD con GitHub Actions
- **API REST**: FastAPI para serving
- **Dashboard**: Streamlit interactivo
- **AutoML**: Hyperparameter optimization
- **Monitoring**: Model drift detection

Gracias