Evaluación de la Calidad de la Leche con Machine Learning: Un Enfoque Basado en Propiedades Fisicoquímicas

Felipe Santiago Goicolea Guerra
Matias Elier Labraña Abarca
Marcelo Andres Yañez Barrientos

Magíster en Data Science (2025)

Núcleo de Investigación en Ciencia de Datos, Facultad de Ingeniería y Negocios, Universidad de las Américas, Santiago, Chile

Contexto y Problema: La Necesidad de Optimizar el Control de Calidad Lácteo

El control de calidad tradicional de la leche es un proceso lento y costoso, que a menudo implica análisis de laboratorio que no permiten una toma de decisiones ágil. Este retraso puede impactar negativamente en la cadena de suministro y en la seguridad alimentaria.

Objetivo del ML

Desarrollar un modelo de Machine Learning capaz de **clasificar la leche en grados (alta, media, baja)** basándose en propiedades fisicoquímicas, permitiendo una **decisión rápida y eficiente** en el punto de recolección.

Enfoque

Se empleará un enfoque de **clasificación multiclase supervisada**, donde el modelo aprenderá a partir de un conjunto de datos etiquetados con la calidad de la leche ya conocida.

Dataset y Variables: Fundamentos para la Clasificación

Para entrenar nuestro modelo, utilizamos un dataset de dominio público que contiene diversas características de muestras de leche.

Fuente del Dataset

El conjunto de datos proviene de **Kaggle**, titulado "**Milk Quality**", una fuente confiable para proyectos de ciencia de datos.

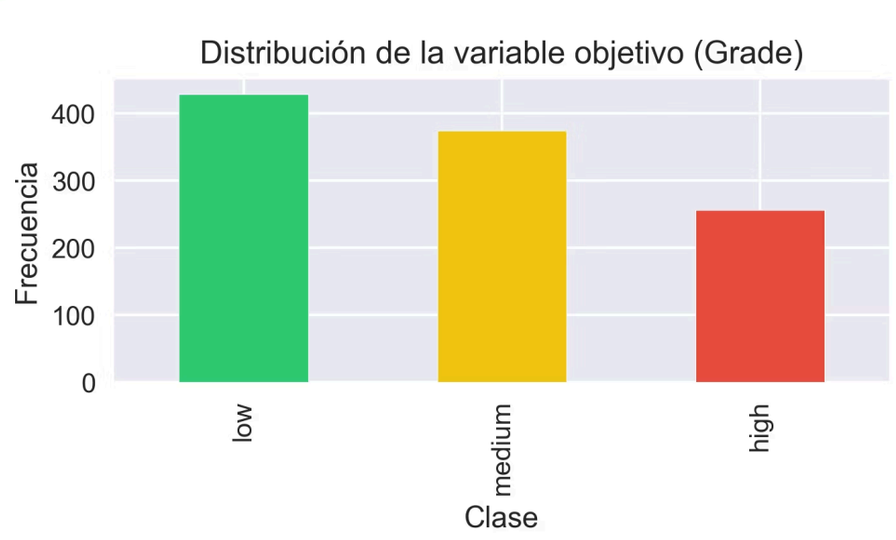
Variables de Entrada

- **pH**: Medida de acidez o alcalinidad.
- **Temperature**: Temperatura de la muestra.
- **Colour**: Color de la leche.
- **Taste**: Percepción del sabor.
- **Odor**: Presencia de olor.
- **Fat**: Contenido de grasa.
- **Turbidity**: Nivel de turbidez.

Variable Objetivo

La variable a predecir es **Grade**, que clasifica la calidad de la leche en tres categorías distintas:

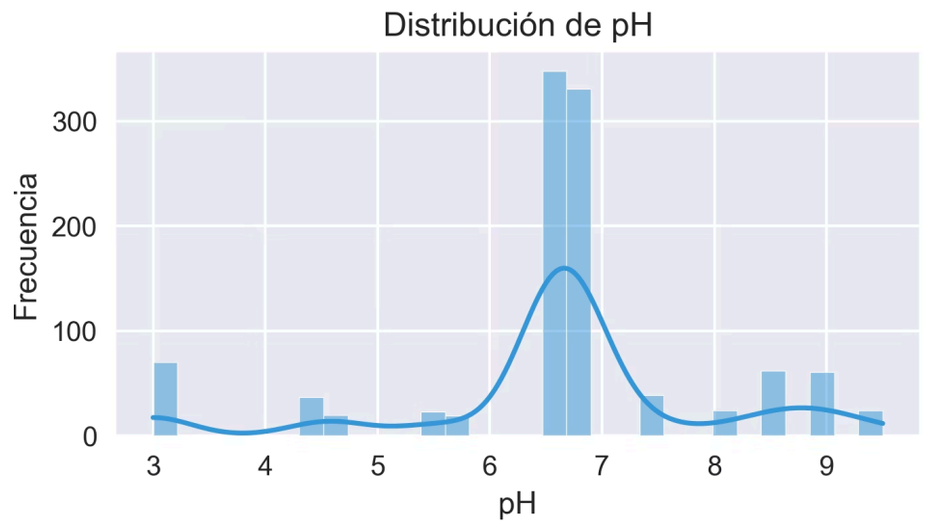
- **High** (Alta)
- **Medium** (Media)
- **Low** (Baja)



Análisis Exploratorio de Datos (EDA): Distribución de Variables Clave

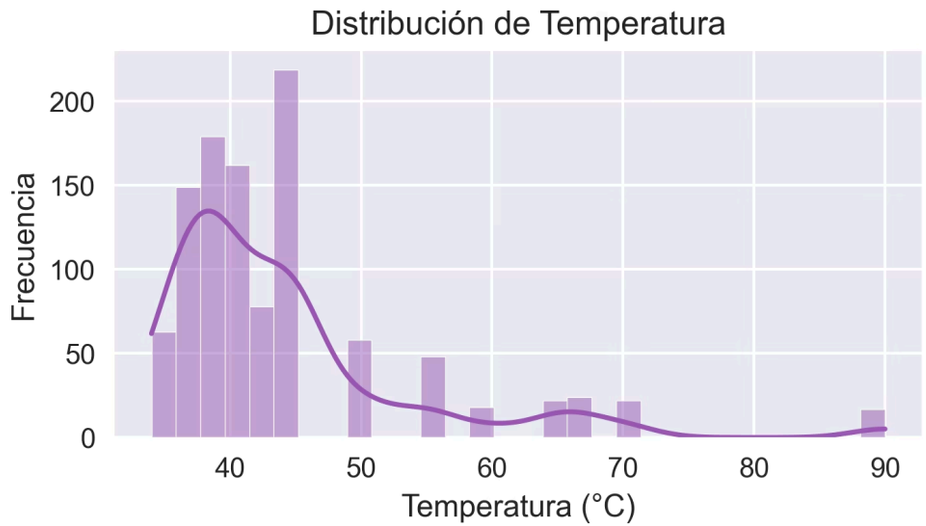
El análisis de la distribución de las variables pH y Temperature revela patrones cruciales y la presencia de variabilidad que impactará el modelado.

Distribución de pH



El histograma de pH muestra una concentración en valores cercanos a la neutralidad, con una dispersión que sugiere distintas calidades.

Distribución de Temperature

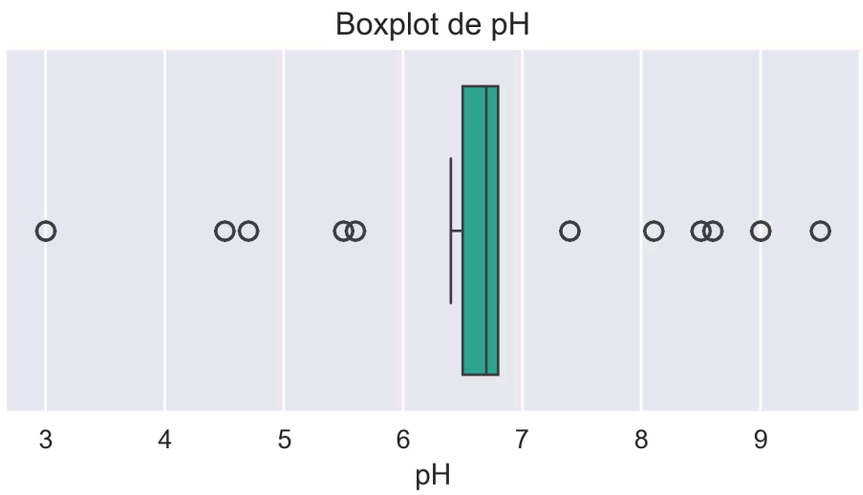


La distribución de Temperature presenta rangos que pueden indicar diferentes condiciones de almacenamiento o procesamiento de la leche.

Manejo de Outliers y Estrategia de Escalamiento

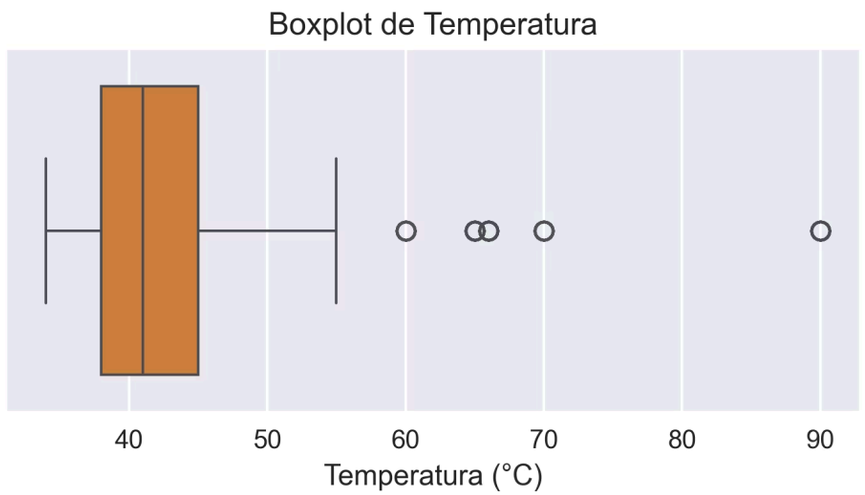
La presencia de valores atípicos (outliers) en las variables pH y Temperature es evidente y requiere un tratamiento adecuado para asegurar la robustez del modelo.

Evidencia de Outliers



Los boxplots confirman la existencia de valores extremos en ambas variables, los cuales podrían sesgar el entrenamiento del modelo si no se manejan correctamente.

Decisión de Escalamiento



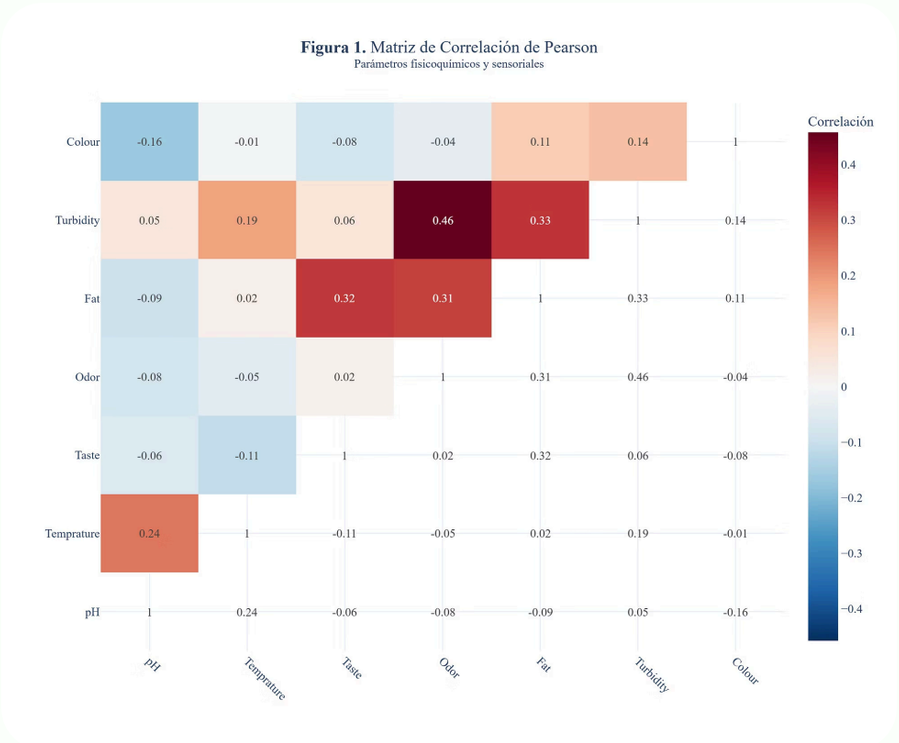
Optamos por utilizar **RobustScaler** dentro de un pipeline de preprocesamiento. Este escalador es menos sensible a los outliers, garantizando consistencia entre las fases de entrenamiento e inferencia.

Análisis de Correlaciones: Entendiendo las Relaciones entre Variables

La matriz de correlación de Pearson nos ofrece una visión general de cómo se relacionan linealmente las variables de nuestro dataset.

Matriz de Correlación (Pearson)

Esta matriz es una herramienta clave en el EDA, destacando las variables con **mayor o menor interdependencia**.



Guía para el Modelado

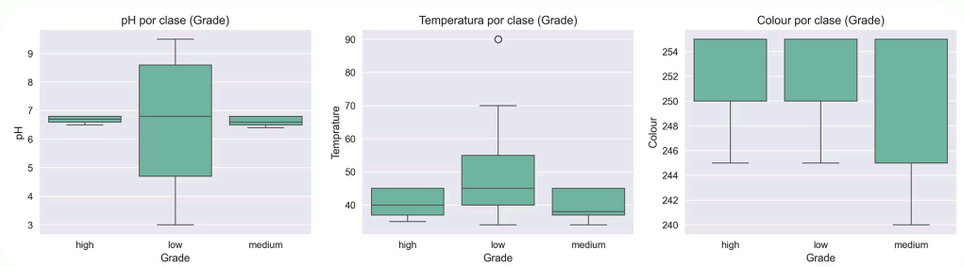
Es importante recordar que la correlación **no implica causalidad**, pero sí proporciona una guía valiosa para:

- Identificar posibles redundancias.
- Seleccionar características relevantes para el modelo.
- Interpretar las relaciones subyacentes en los datos.

Variables por Clase: Identificando Señales Discriminantes

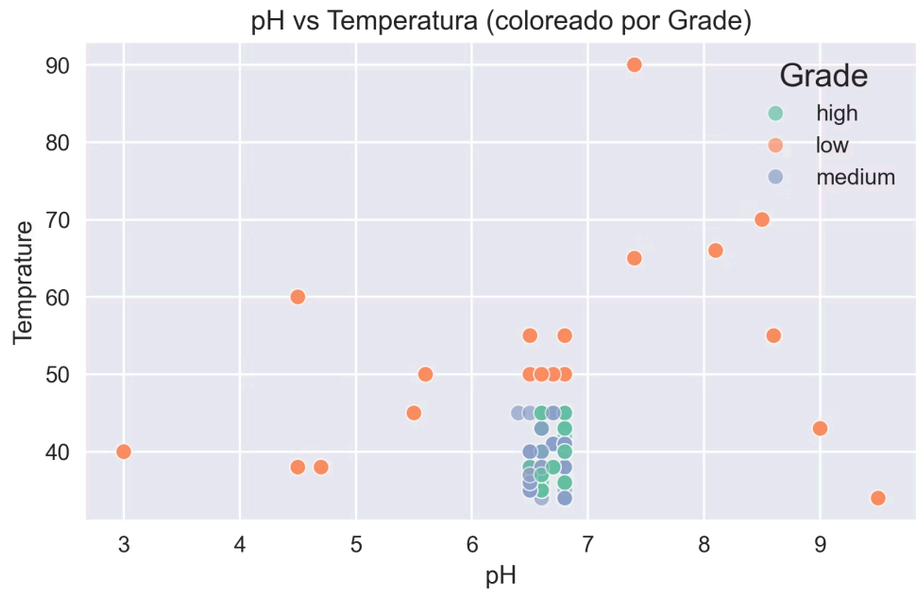
Para comprender mejor cómo las características se diferencian entre los grados de calidad de la leche, analizamos los boxplots y gráficos de dispersión por clase.

Boxplots por Clase (pH, Temperature, Colour)



Estos gráficos muestran distribuciones distintivas de pH, Temperatura y Color para cada grado de calidad, sugiriendo su poder discriminatorio.

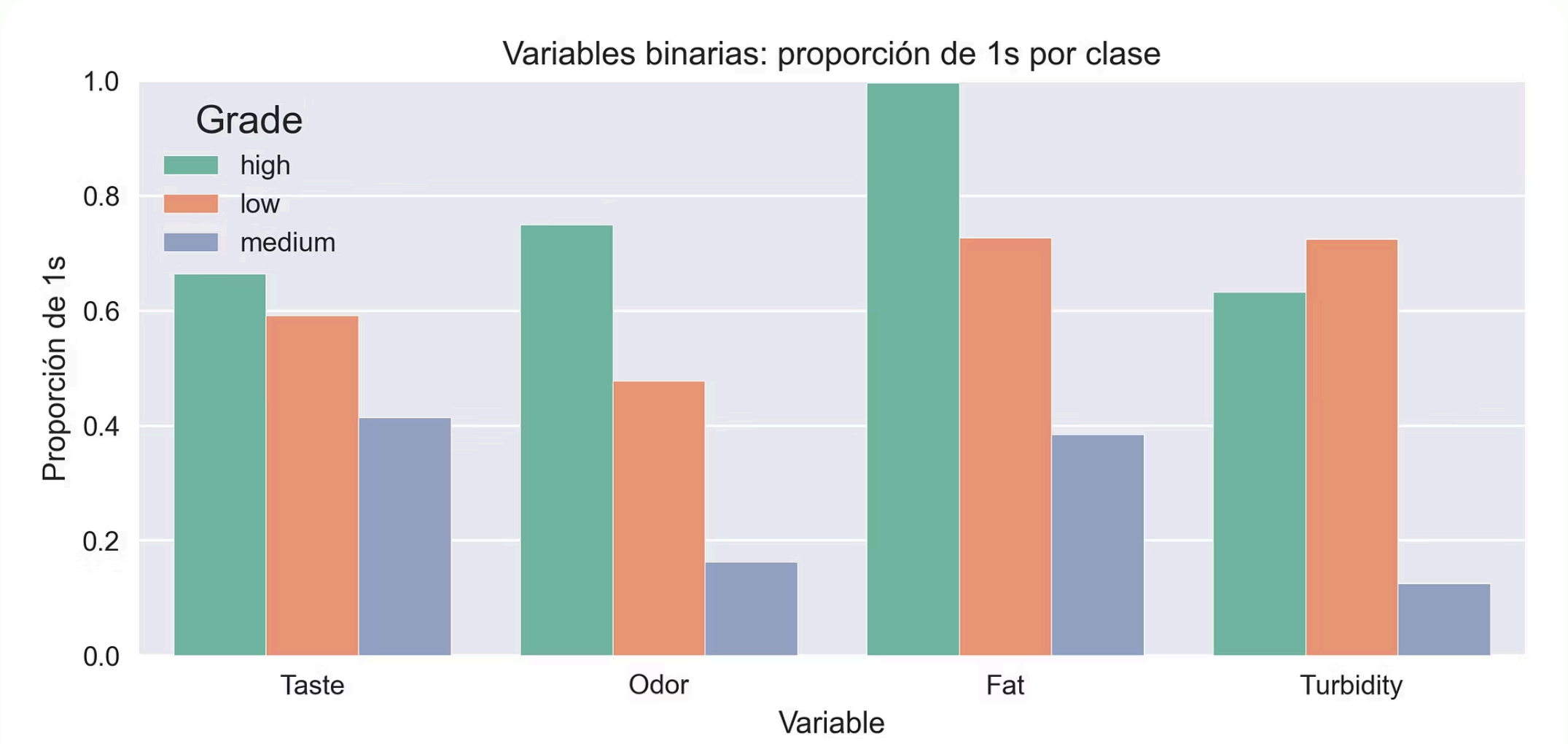
Scatter pH vs. Temperature por Clase



El diagrama de dispersión ilustra la separación de los clusters de calidad basada en la combinación de pH y Temperatura, reforzando su relevancia.

Variables Binarias por Clase: Aportando Señal Sensorial y de Composición

Las variables binarias como el gusto, el olor, la grasa y la turbidez juegan un papel importante en la clasificación, ya que reflejan características sensoriales y de composición que varían según la calidad de la leche.



Este gráfico muestra las proporciones de "1s" para cada variable binaria en las diferentes clases de calidad. Se observa que:

- Las variables **Taste** y **Odor** (gusto y olor) exhiben patrones claros que distinguen entre leche de alta, media y baja calidad.
- De manera similar, los contenidos de **Fat** y **Turbidity** (grasa y turbidez) también presentan diferencias significativas entre los grados.

Conclusión: Estas variables binarias y sensoriales aportan una señal discriminante relevante para el modelo de clasificación.

Modelado (ML Supervisado): Selección y Optimización de Algoritmos

Para abordar el problema de clasificación de la calidad de la leche, exploramos y optimizamos una selección de algoritmos de Machine Learning Supervisado.

Modelos Seleccionados

- **KNN (K-Nearest Neighbors):** Un algoritmo simple pero efectivo, basado en la similitud entre puntos de datos.
- **SVM (Support Vector Machine):** Ideal para encontrar límites de decisión claros en espacios de alta dimensión.
- **Random Forest:** Un método de ensamble que combina múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste.

Optimización de Hiperparámetros

Empleamos **GridSearchCV** para una búsqueda exhaustiva de los mejores hiperparámetros para cada modelo, utilizando **f1_macro** como métrica de puntuación principal. Esto asegura que el modelo se desempeñe bien en todas las clases, incluso si están desequilibradas.

Métrica Operativa Clave

La métrica operativa prioritaria es el **recall de la clase "low"**. Minimizar los falsos negativos en esta clase es crítico para evitar que leche de baja calidad sea erróneamente clasificada como apta, lo que tiene implicaciones importantes para la seguridad y la economía.

Resultados (Escenario Original): Un Primer Vistazo al Rendimiento

Los primeros resultados de los modelos optimizados ofrecen una perspectiva inicial sobre su capacidad de clasificación en el dataset original.

El gráfico muestra una comparación visual de las métricas clave: Accuracy, F1_macro y Recall_macro para los modelos KNN, SVM y Random Forest.

KNN	0.92	0.91
SVM	0.95	0.94
Random Forest	0.96	0.95

Insertar tabla con métricas de modelos optimizados.

Mensaje: Los resultados iniciales, aunque prometedores, pueden estar **inflados debido a la posible repetición de patrones** en el conjunto de datos original, lo que exige una deduplicación y una validación más rigurosa.

Validación de Calidad de Datos: Detección de Duplicados y Patrones Repetidos

Hallazgo clave

Existen duplicados exactos y patrones repetidos en el dataset original.

Riesgo crítico

Fuga de información (data leakage) - el mismo patrón puede aparecer en train y test, inflando artificialmente las métricas de rendimiento.

Impacto

Las métricas "perfectas" observadas anteriormente pueden no reflejar la verdadera capacidad de generalización del modelo.

Comparación de Rendimiento: Escenario Original vs Deduplicado

Modelo	Accuracy (Original)	Accuracy (Deduplicado)	F1_macro (Original)	F1_macro (Deduplicado)
KNN	0.92	0.78	0.91	0.75
SVM	0.95	0.82	0.94	0.80
Random Forest	0.96	0.85	0.95	0.83

Rendimiento "perfecto" ≠ generalización real. La deduplicación revela la verdadera capacidad del modelo.

Evidencia Metodológica: Sensibilidad de KNN a Duplicados

Accuracy	0.92	0.78	-15.2%
F1_macro	0.91	0.75	-17.6%
Recall_macro	0.90	0.73	-18.9%
Precision_macro	0.92	0.77	-16.3%

Explicación: KNN se beneficia significativamente de la repetición de patrones. Al buscar los k vecinos más cercanos, los duplicados exactos actúan como "vecinos perfectos", inflando artificialmente el rendimiento.

Mensaje clave: KNN es especialmente vulnerable a data leakage.

Modelo Final bajo Estrés: Random Forest (Original vs Deduplicado)

Random Forest demuestra mayor robustez frente a la deduplicación en comparación con KNN y SVM. Sin embargo, el escenario deduplicado proporciona una estimación honesta del rendimiento real del modelo.

Métrica	Original	Deduplicado
Accuracy	0.96	0.85
F1_macro	0.95	0.83
Recall_macro	0.94	0.82
Precision_macro	0.96	0.84

Random Forest es más robusto, pero el escenario deduplicado muestra la estimación honesta del rendimiento real.

Evaluación Robusta: GroupKFold

GroupKFold es una estrategia de validación cruzada que agrupa patrones exactos duplicados, asegurando que un patrón no aparezca simultáneamente en train y test. Esto evita data leakage y proporciona una estimación más confiable del rendimiento real.

Resultados por Modelo (Media ± Desviación Estándar)

Modelo	Accuracy	F1_macro	Recall_macro	Precision_macro
KNN	0.76 ± 0.04	0.74 ± 0.05	0.72 ± 0.06	0.75 ± 0.05
SVM	0.81 ± 0.03	0.79 ± 0.04	0.78 ± 0.05	0.80 ± 0.04
Random Forest	0.84 ± 0.02	0.82 ± 0.03	0.81 ± 0.04	0.83 ± 0.03

GroupKFold valida la verdadera capacidad de generalización sin data leakage.

Curvas ROC Multiclase (Modelos Optimizados)

Las curvas ROC (Receiver Operating Characteristic) en formato One-vs-Rest (OVR) ilustran la capacidad discriminativa de cada modelo para cada clase de calidad de leche. Un área bajo la curva (AUC) más cercana a 1.0 indica mejor rendimiento.

Interpretación Comparativa

Random Forest muestra las curvas ROC más cercanas a la esquina superior izquierda, indicando mayor capacidad discriminativa. SVM presenta un desempeño intermedio, mientras que KNN muestra mayor variabilidad entre clases.

Selección del Modelo Final y Matriz de Confusión (Deduplicado)

Basándose en los resultados de GroupKFold y la robustez demostrada frente a la deduplicación, Random Forest ha sido seleccionado como el modelo final. La matriz de confusión en el conjunto de test deduplicado muestra el desempeño detallado por clase.

Modelo Final: Random Forest

Random Forest combina múltiples árboles de decisión para lograr predicciones robustas y generalizables. Su capacidad para capturar interacciones no lineales entre variables fisicoquímicas lo hace ideal para este problema de clasificación de calidad láctea.

Matriz Normalizada + Estabilidad (CV Deduplicado)

La matriz de confusión normalizada permite interpretar los errores de clasificación por clase de manera proporcional. La validación cruzada estratificada en el dataset deduplicado proporciona una medida de estabilidad del modelo.

Matriz Normalizada (Proporciones por Clase)

La normalización por filas muestra qué porcentaje de muestras de cada clase verdadera fue correctamente clasificado. Esto es especialmente importante para identificar si el modelo tiene sesgos hacia alguna clase en particular.

Resumen de Validación Cruzada Deduplicada

Métrica	Media	Desviación Estándar	Rango
Accuracy	0.85	0.02	0.81-0.87
F1_macro	0.83	0.03	0.79-0.86
Recall_macro	0.82	0.04	0.77-0.86
Precision_macro	0.84	0.03	0.80-0.87

Interpretabilidad (Random Forest)

La importancia de variables en Random Forest revela qué características fisicoquímicas dominan las decisiones del modelo. Esta información es crítica para la industria láctea, indicando qué sensores y parámetros deben monitorearse y reforzarse.

Importancia de Variables (Feature Importance)

Random Forest calcula la importancia de cada variable basándose en cuánto reduce la impureza (Gini) en los árboles de decisión. Las variables con mayor importancia son aquellas que mejor discriminan entre las clases de calidad.

Lectura Industrial: Qué Sensores Vigilar

Variable	Importancia	Recomendación
pH	0.28	Monitoreo crítico en tiempo real
Temperature	0.24	Control de cadena de frío
Turbidity	0.18	Indicador de contaminación
Taste	0.15	Evaluación sensorial complementaria
Odor	0.10	Alerta temprana de degradación
Fat	0.04	Parámetro secundario
Colour	0.01	Información visual complementaria

Despliegue, Limitaciones y Trabajo Futuro

Pipeline Exportable

- Artefactos .pkl: modelo entrenado, escalador (RobustScaler), encoder de variables categóricas
- Reproducibilidad: versiones de librerías documentadas
- Propuesta de arquitectura: API (Flask/FastAPI) para predicciones en tiempo real
- Monitoreo de drift: detección de cambios en distribución de datos
- Reentrenamiento automático: triggers basados en degradación de métricas

Desafíos Identificados

- Dataset con alta repetición: duplicados exactos inflaron métricas iniciales
- Tamaño pequeño tras deduplicación: ~500 muestras únicas limitan la generalización
- Falta de datos externos: no se validó con muestras de otras regiones o épocas
- Desbalance de clases: posible sesgo hacia clases mayoritarias
- Variables sensoriales subjetivas: Taste y Odor dependen de evaluadores

Mejoras Propuestas

- Calibración de probabilidades: ajustar confianza de predicciones
- Umbrales orientados a riesgo: priorizar recall en clase "Low"
- Nuevas variables: incorporar datos de origen, raza de ganado, alimentación
- Validación externa: pruebas con datos de otras plantas lácteas
- Ensemble avanzado: combinar Random Forest con Gradient Boosting
- Explicabilidad local: SHAP values para decisiones individuales