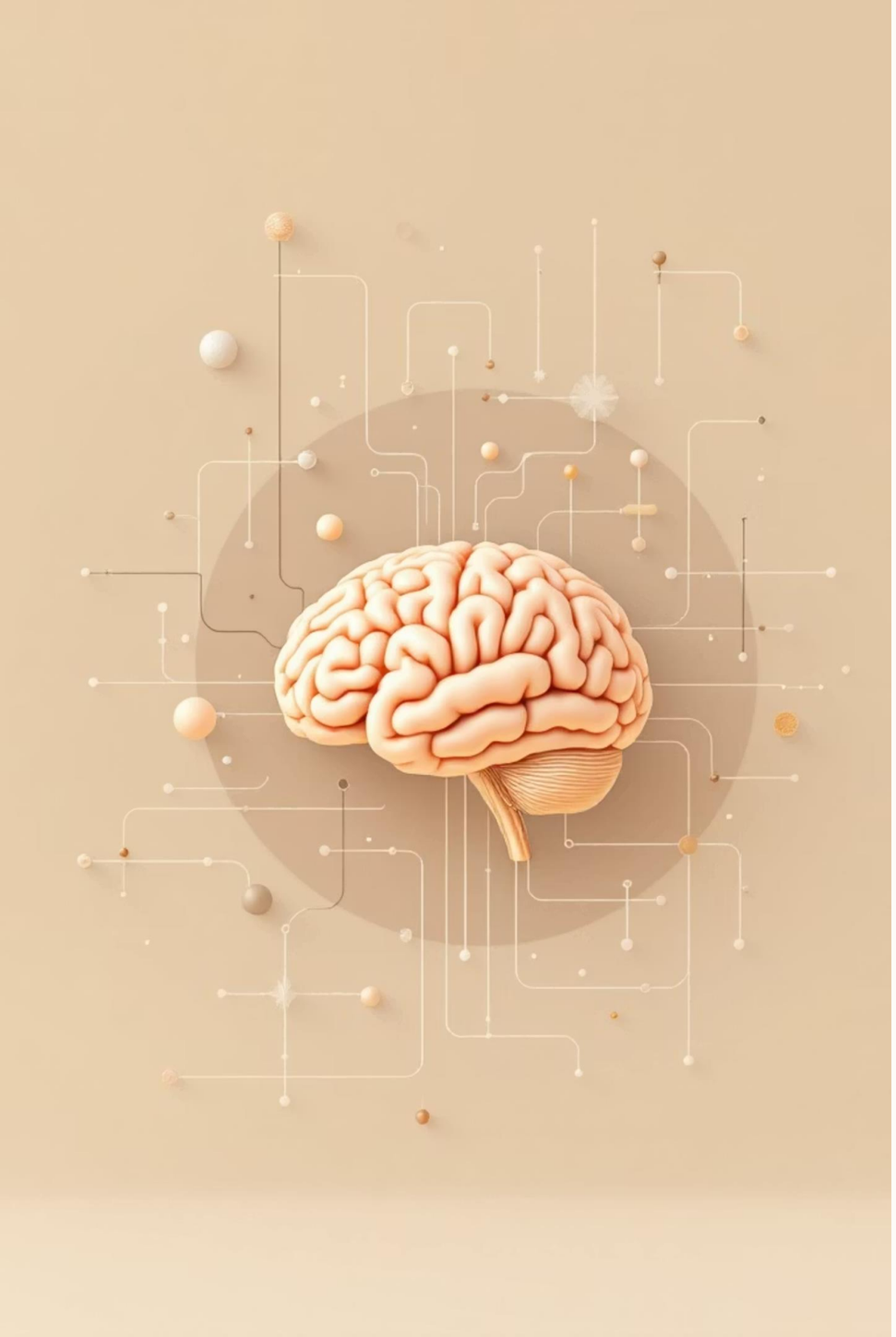


🧠 SkinMatch AI: Modelo Multilabel para Compatibilidad Compatibilidad Cosmética

Descubre cómo nuestra arquitectura de IA optimiza la selección de productos cosméticos, analizando la arquitectura, métricas, ingeniería de características y comparación de modelos.

Dataset Real: Más de 1.400 Cosméticos Analizados

Columnas Iniciales

Comenzamos con más de 11 columnas de datos ricos por producto.

Variable Objetivo Multilabel

5 targets binarios para una categorización precisa.

Variabilidad INCI

Gran diversidad en la longitud de las listas de ingredientes (INCI).

Etiquetas Desbalanceadas

Desafío particular en la detección de pieles sensibles.

Limpieza y Exploración de Datos: Preparando el Terreno

Una fase crucial para asegurar la calidad de los datos, donde realizamos una normalización exhaustiva y un análisis profundo de las correlaciones entre variables.

→ Normalización y Estandarización

Aplicamos el uso de minúsculas (lowercasing) y expresiones regulares (regex) para eliminar caracteres no útiles, garantizando uniformidad en el texto.

→ Eliminación de Duplicados

Identificamos y eliminamos entradas duplicadas para evitar sesgos y asegurar que cada producto se represente una única vez.

→ Análisis de Longitud INCI

Estudiamos la longitud de las listas de ingredientes para entender la complejidad formulativa de los productos.

→ Conteo de Ingredientes

Contabilizamos el número de ingredientes por producto, una métrica clave para entender la densidad de la formulación.

→ Correlación Numérica

Analizamos la correlación entre las características numéricas existentes para identificar posibles redundancias o interacciones.

Feature Engineering Dermatológico: Aumentando la Señal

Contadores Específicos

- `cnt_irritantes`
- `cnt_calmantes`
- `cnt_emolientes`
- `cnt_aceites`

Ratios Clave

- `ratio_irritantes`
- `ratio_calmantes`

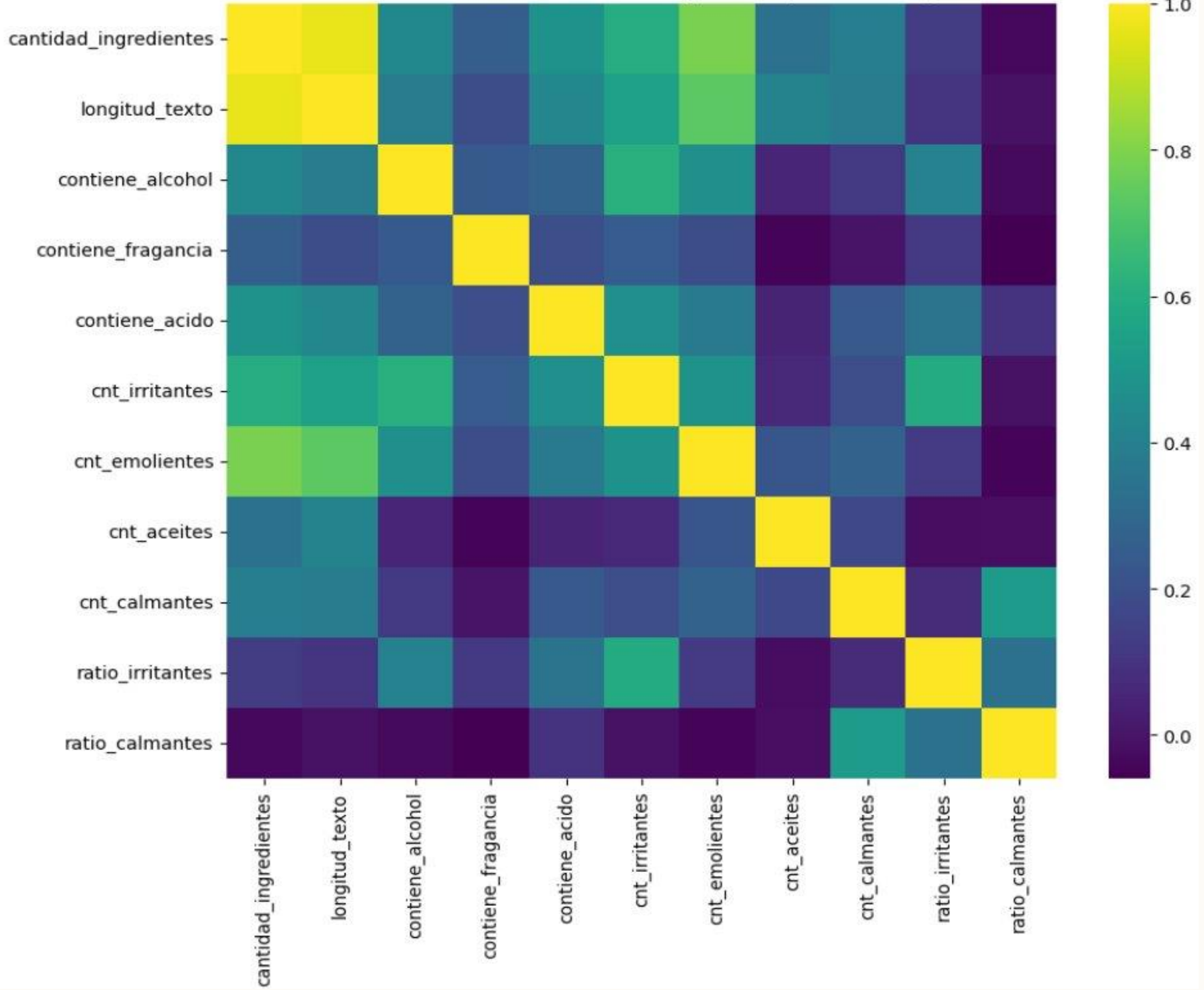
Indicadores Binarios

- `contiene_alcohol`
- `contiene_fragancia`
- `contiene_acido`

Características Textuales

- `longitud_texto`
- `cantidad_ingredientes`

La incorporación de estas variables dermatológicas permite que el modelo comprenda no solo *qué ingredientes aparecen*, sino *cómo afectan realmente a la piel*. Esto enriquece el TF-IDF tradicional con información clínica y mejora significativamente la calidad de las predicciones.


cantidad_ingredientes
longitud_texto
contiene_alcohol
contiene_fragancia
contiene_acido
cnt_irritantes
cnt_emolientes
cnt_aceites
cnt_calmantes
ratio_irritantes
ratio_calmantes
1.0
0.8
0.6
0.4
0.2
0.0

🎯 Selección del Target y Variables Predictoras

El objetivo del proyecto es predecir para qué tipo(s) de piel es apto un producto cosmético, por lo que seleccioné como target las cinco etiquetas dermatológicas del dataset: Dry, Oily, Normal, Sensitive y Combination. Estas etiquetas representan resultados reales anotados por expertos y son ideales para un enfoque multilabel, donde un mismo producto puede ser compatible con más de un tipo de piel.

Variables Predictoras Utilizadas:
Para construir el modelo, utilicé tanto el texto completo de ingredientes (INCI) como un conjunto de features dermatológicas creadas manualmente que aportan señal clínica:

- Cantidad de ingredientes
- Longitud del INCI
- Presencia de alcoholes, ácidos y fragancias
- Conteo de ingredientes irritantes, calmantes, aceites y emolientes
- Proporciones relativas de cada categoría

Estas variables describen el impacto potencial del producto sobre la piel y permiten que el modelo aprenda patrones químicos que influyen en la compatibilidad dermatológica.

Conclusión: Elegí este target porque representa la necesidad real del usuario y seleccioné estas variables predictoras porque capturan tanto la estructura del INCI como su relevancia clínica.

Modelos Evaluados

Exploracios de variedad de algoritmos para encontrar la mejor solución para nuestro problema multilabel, utilizando MultiOutputClassifier para una gestión eficiente de las etiquetas.

Regresión Logística

Un algoritmo lineal robusto, ideal como línea base para clasificación.

Random Forest

Conjunto de árboles de decisión, conocido por su capacidad para manejar datos complejos y reducir el sobreajuste.

Gradient Boosting

Otro potente algoritmo de ensamble que construye modelos de forma secuencial para corregir errores previos.

SVC (Support Vector Classifier)

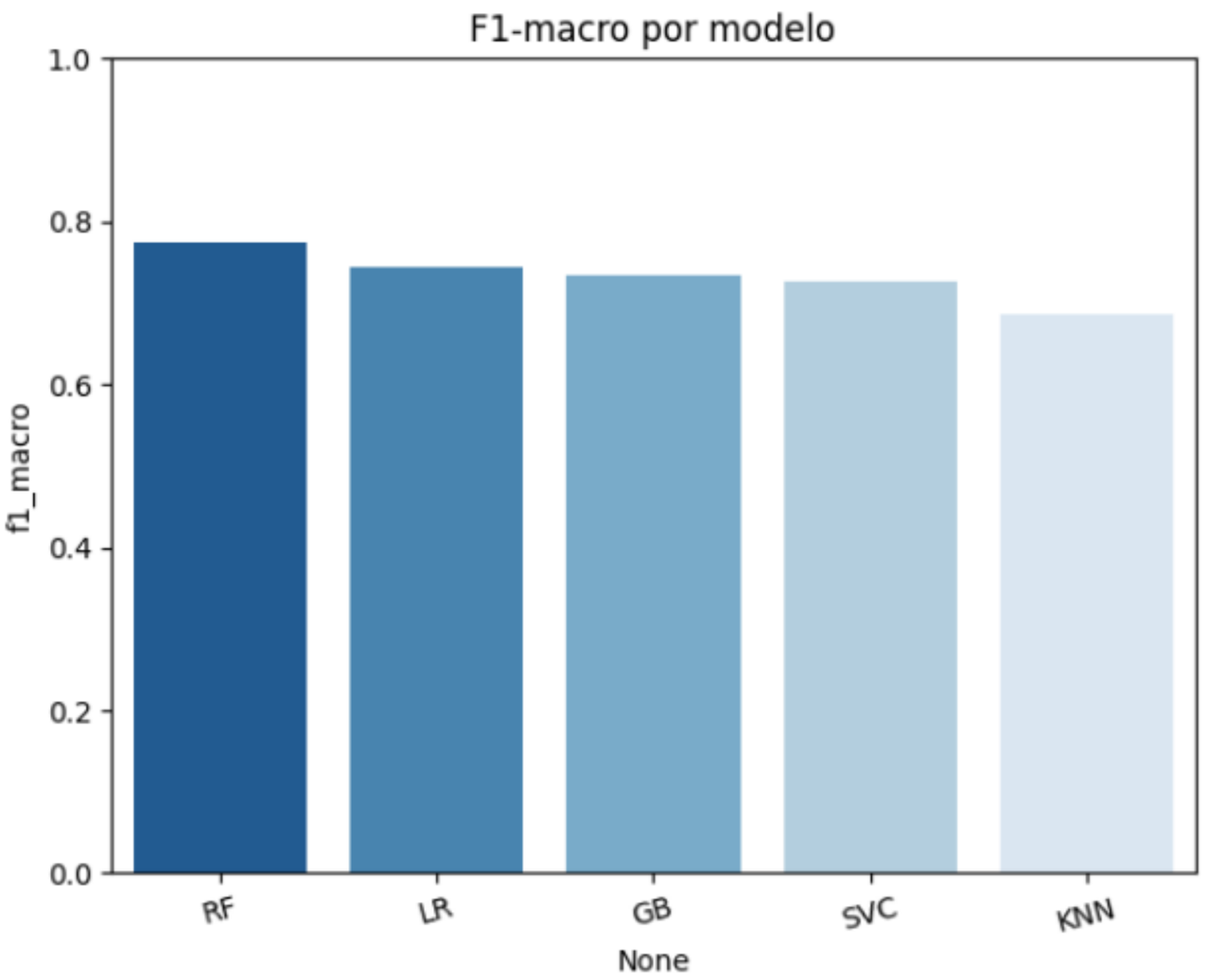
Efectivo en espacios de alta dimensión y útil cuando el número de dimensiones es mayor que el número de muestras.

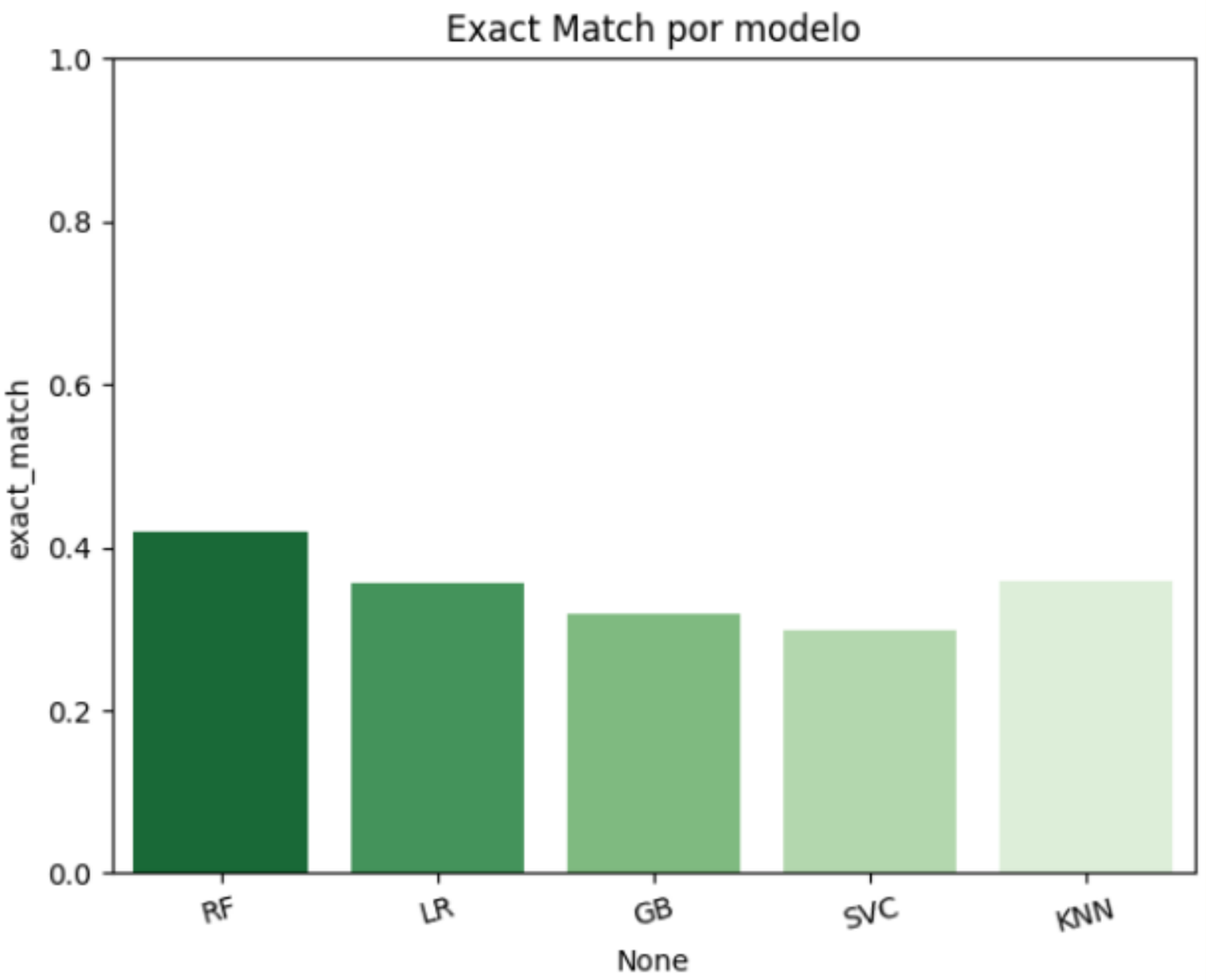
KNN (K-Nearest Neighbors)

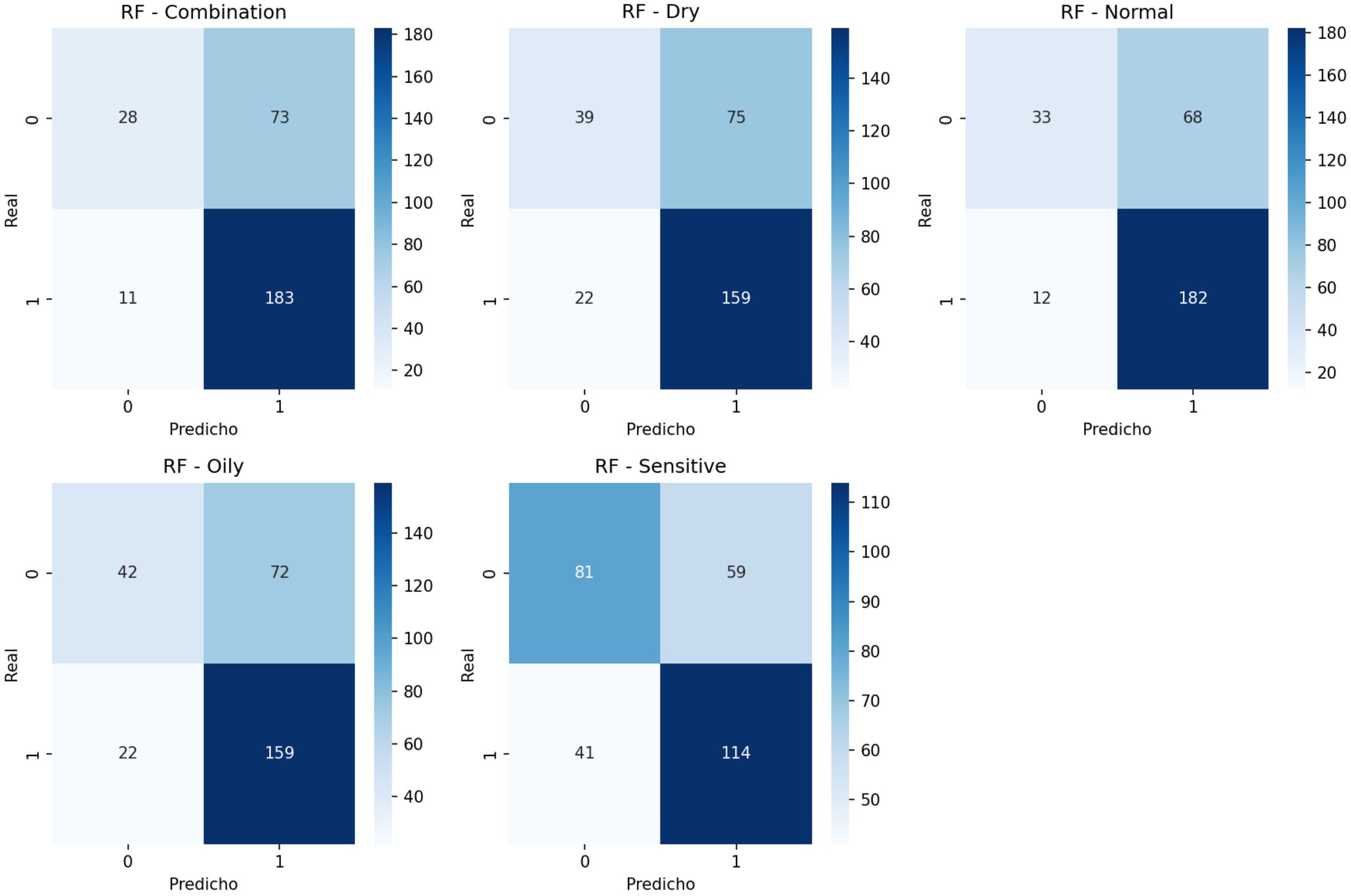
Clasificador no paramétrico que clasifica un punto de datos en función de la mayoría de sus vecinos.

KMeans

Para explorar la estructura no supervisada del dataset.


F1-macro por modelo
f1_macro
1.0
0.8
0.6
0.4
0.2
0.0
RF
LR
GB
SVC
KNN
None


Exact Match por modelo
exact_match
1.0
0.8
0.6
0.4
0.2
0.0
RF
LR
GB
SVC
KNN
None


RF - Combination
28
73
11
183
Real
Predicho
RF - Dry
39
75
22
159
Real
Predicho
RF - Normal
33
68
12
182
Real
Predicho
RF - Oily
42
72
22
159
Real
Predicho
RF - Sensitive
81
59
41
114
Real
Predicho

Matrices de Confusión: Comportamiento por Clase

El análisis de las matrices de confusión nos permite comprender en detalle el comportamiento de nuestro modelo Random Forest Multilabel para cada una de las etiquetas. Este desglose es crucial para identificar puntos fuertes y áreas de mejora.

Normal, Oily, Combination

Estas clases muestran una alta tasa de detección y una baja incidencia de falsos negativos y positivos, indicando un excelente rendimiento.

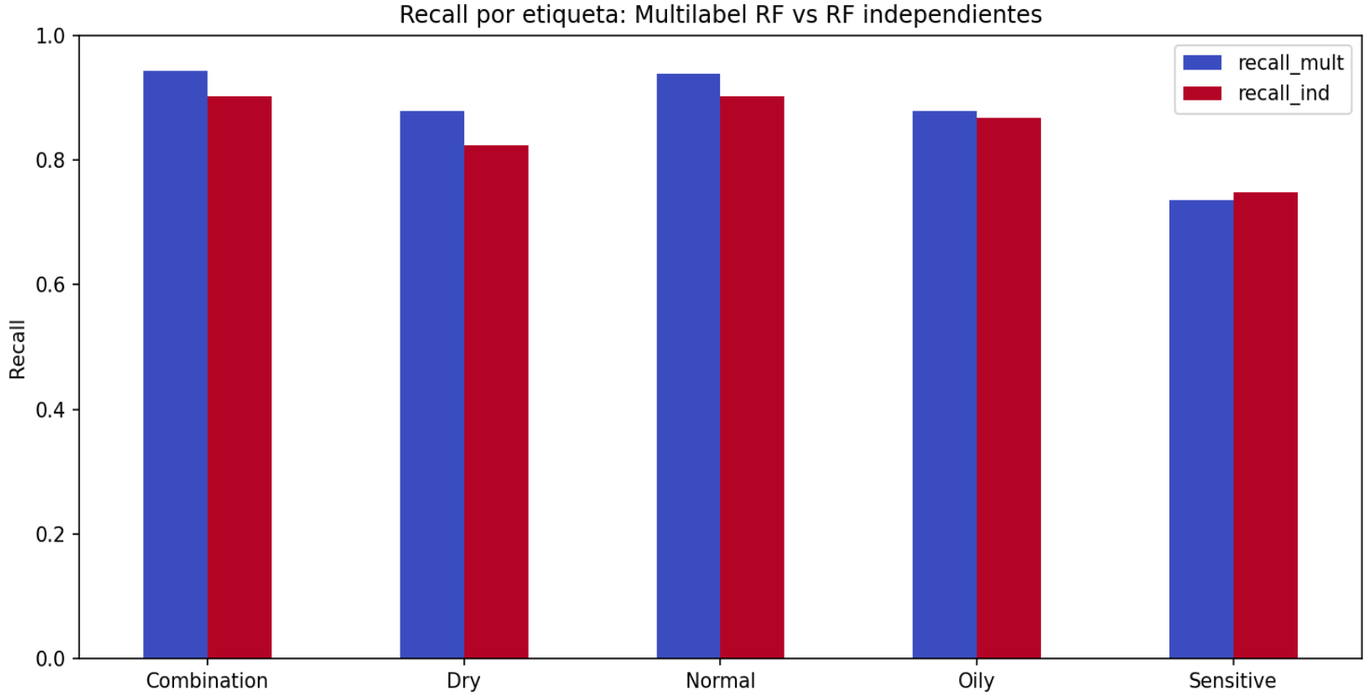
Sensitive: Un Desafío

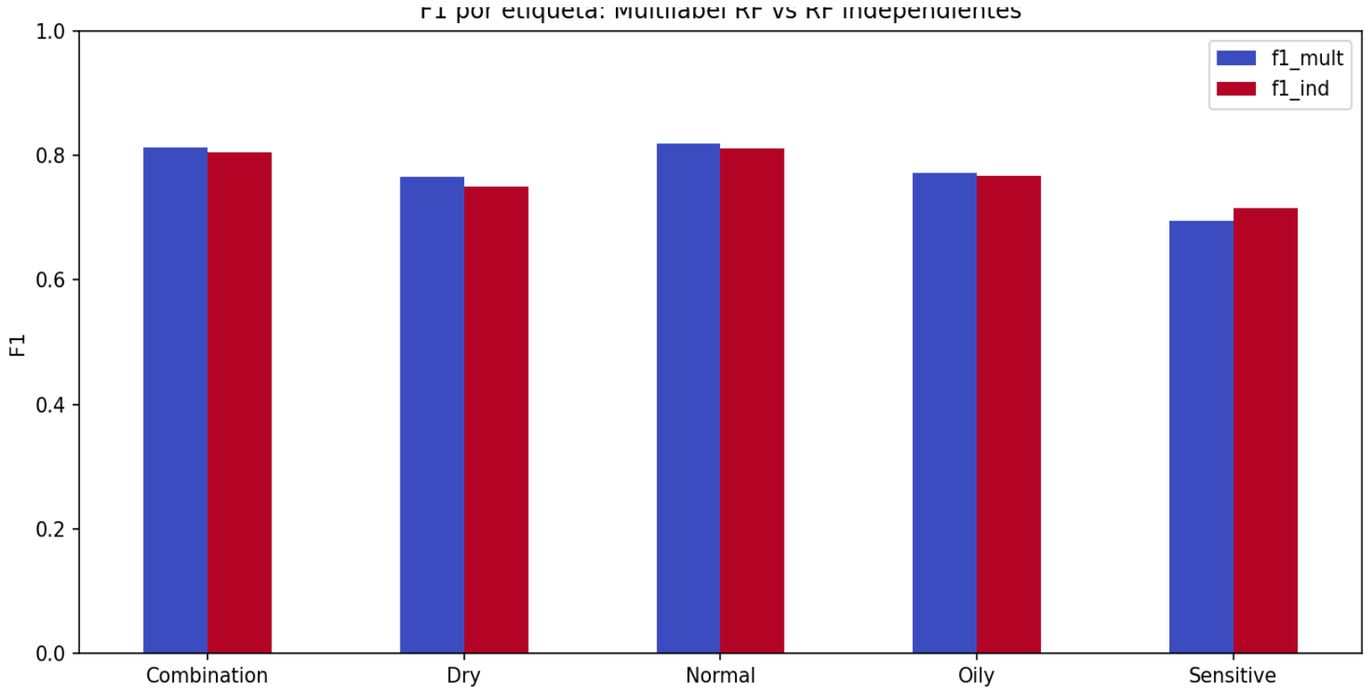
La clase 'Sensitive' es intrínsecamente más difícil debido a su menor representación en el dataset, lo que se refleja en una detección ligeramente inferior, aunque sigue siendo robusta.

Bajo Falsos Negativos

Es vital destacar el bajo nivel de falsos negativos en todas las clases, especialmente en aquellas críticas donde no identificar una incompatibilidad podría tener implicaciones negativas para el usuario final.

Las matrices confirman la capacidad del modelo para discernir entre los distintos tipos de piel, proporcionando una base sólida para recomendaciones precisas.adecuados pasen desapercibidos.





Random Forest Multilabel: La Elección Óptima

El análisis comparativo revela que el enfoque multilabel con Random Forest ofrece un rendimiento superior en varios aspectos clave, superando a los modelos independientes.

Mejor Balance de Clases

Distribución más equitativa del rendimiento entre las etiquetas.

Mayor Recall Promedio

Mejor detección de verdaderos positivos en todas las categorías.

Rendimiento Estable

Menor varianza y mayor consistencia en las predicciones.

Implementación Sencilla

Un solo pipeline vs. múltiples modelos individuales.

SkinMatch AI: ¡Gracias!