Modelo Predictivo de Puntuación de Vinos

¡Bienvenidos! En esta presentación, exploraremos un modelo predictivo que utiliza características de los vinos para predecir su puntuación en catas a ciegas.

Definición del Problema y Objetivos

El Desafío

Predecir la puntuación que un vino obtendría en una cata a ciegas de expertos, en base a características al alcance de todos.

Objetivo Principal

Desarrollar un modelo que pueda ayudar a los consumidores a evaluar la calidad potencial de los vinos que tienen a su alcance.

Contextualización técnica

Dataset

Dataset de Kaggle (webscrapping a Wine Enthusiast):

+80.000 registros

Selección

Registros de vinos españoles:

3. 455 registros

Variables

```
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   title                 3455 non-null   object
 1   vintage               3455 non-null   object
 2   winery                3455 non-null   object
 3   variety               3455 non-null   object
 4   country               3455 non-null   object
 5   description           3455 non-null   object
 6   designation           2868 non-null   object
 7   points                3455 non-null   int64
 8   price                 3356 non-null   float64
 9   province              3455 non-null   object
 10  region_1              3453 non-null   object
 11  region_2              0 non-null      object
 12  taster_name           3455 non-null   object
 13  taster_photo          3455 non-null   object
 14  taster_twitter_handle 3455 non-null   object
```

Preprocesamiento y Exploración de los Datos

1

1. Limpieza de Datos

Eliminar valores faltantes, corregir errores, transformar datos categóricos.

2

2. Análisis Exploratorio

Identificar patrones, correlaciones, distribución de variables.

3

3. Selección de Variables

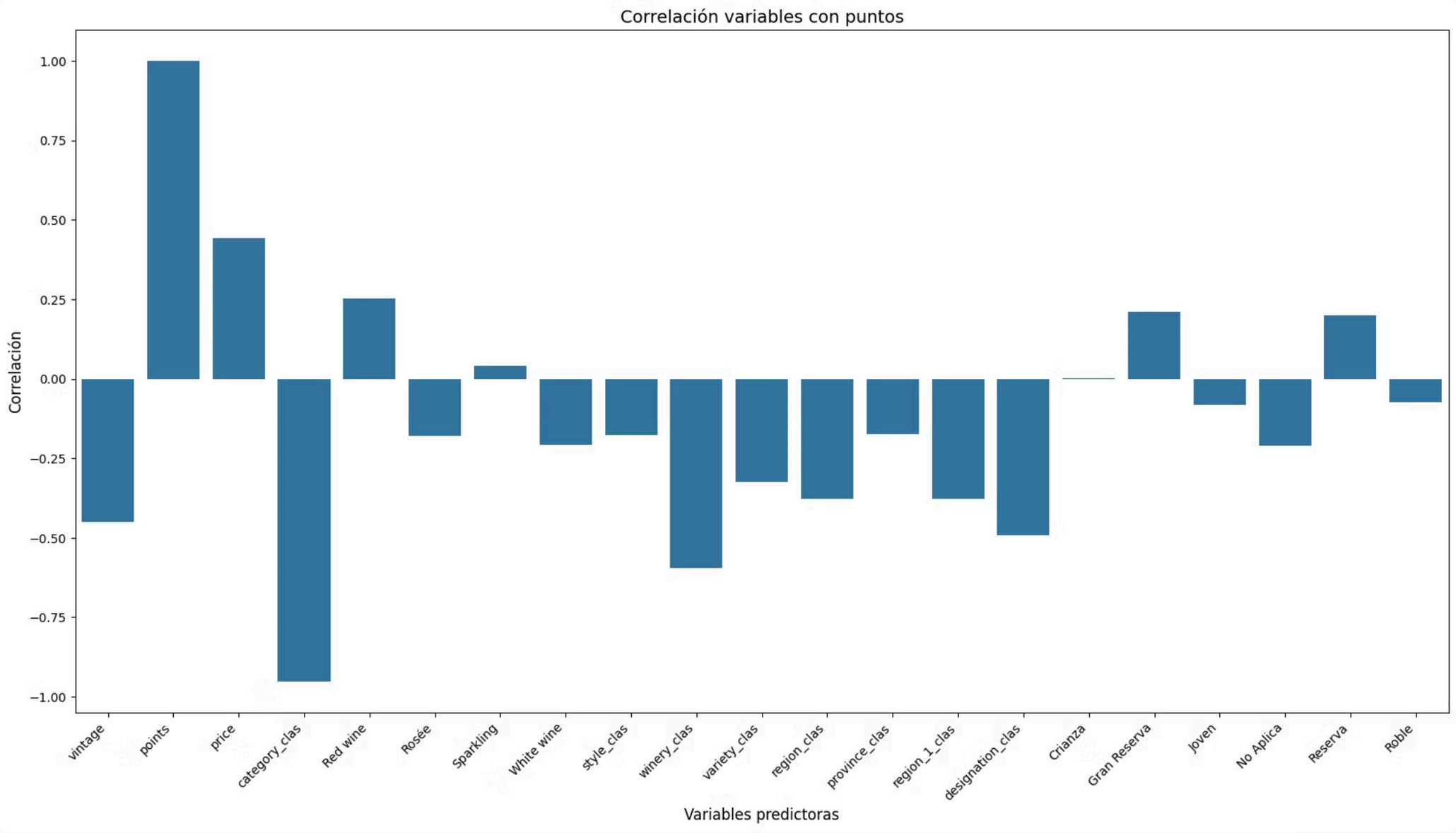
Identificar las variables más relevantes para la predicción de la puntuación del vino.

1. Limpieza de Datos

- **Limpieza**:
- Manejo de valores faltantes mediante eliminación si faltaba precio.
- **Transformaciones**:
- Creación de nuevas categorías: tipo de vino, denominación de origen y crianza.
- One-hot encoding para las variables categóricas tipo de vino y crianza.
- OrdinalEncoder para las variables categóricas bodega, variedad de uva y denominación.
- StandarScaler a las variables numéricas precio y año.
- **Análisis Exploratorio**:
-
- Distribución por año de cosecha y clasificación.

2. Análisis Exploratorio

Correlación entre variables predictoras y puntuación.



3. Selección de Variables

1 Variable objetivo:

"Points"

2 Características con muchas categorías únicas

OrdinalEncoder:

'winery', 'variety', 'denominacion'

3 Características con pocas categorías únicas

OneHotEncoder:

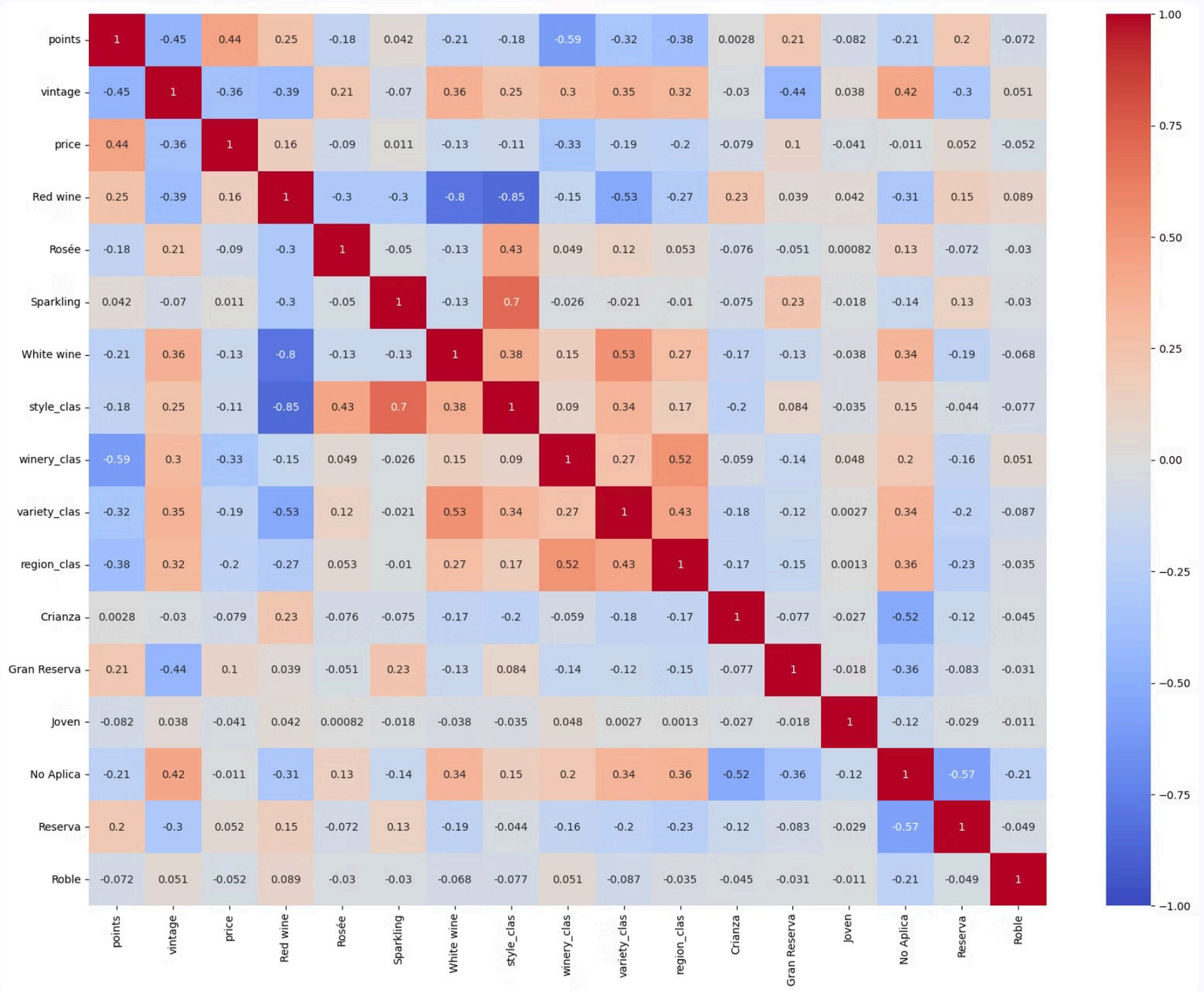
'style', 'aging_1'

4 Variables numéricas

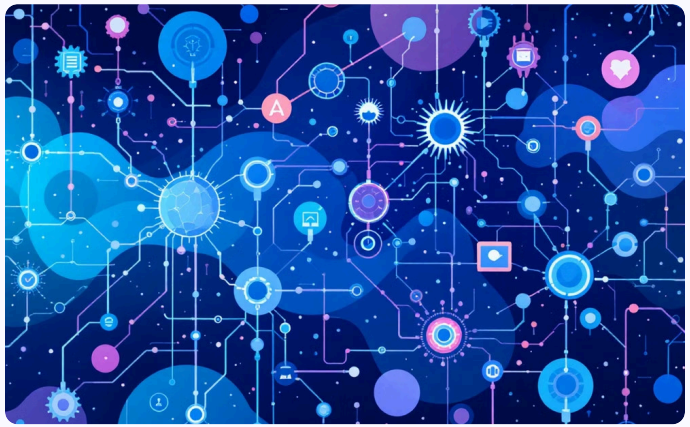
StandarScaler:

'price', 'vintage'

3. Selección de Variables



Selección y Entrenamiento del Modelo



Modelos de Machine Learning

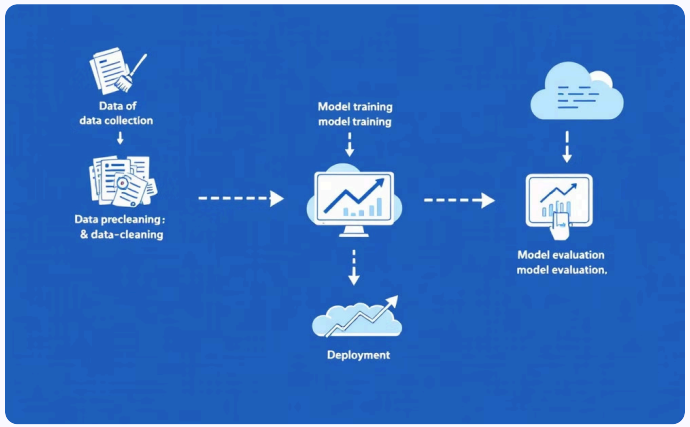
Implementación de diversos algoritmos:

- KNN y SVR para regresión
- LightGBM y XGBRegressor
- Random Forest y Random Tree
- Linear Regressor y Keras

Selección de Variables

Técnicas de selección implementadas:

- SelectKBest para variables relevantes
- PCA para reducción dimensional
- KMeans para agrupamiento



Proceso de Entrenamiento

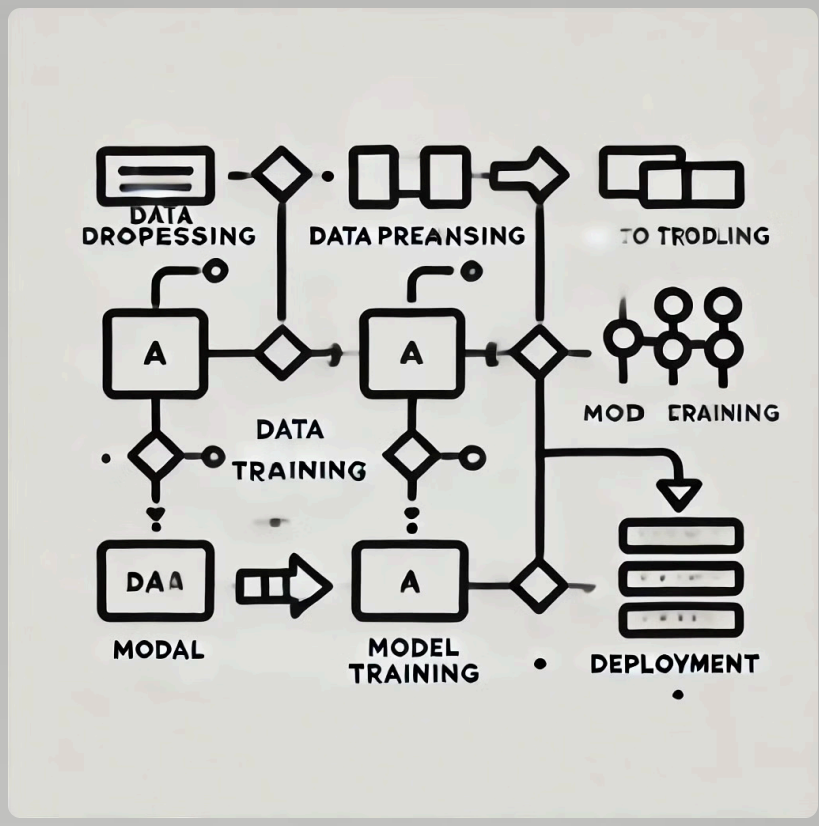
Optimización mediante:

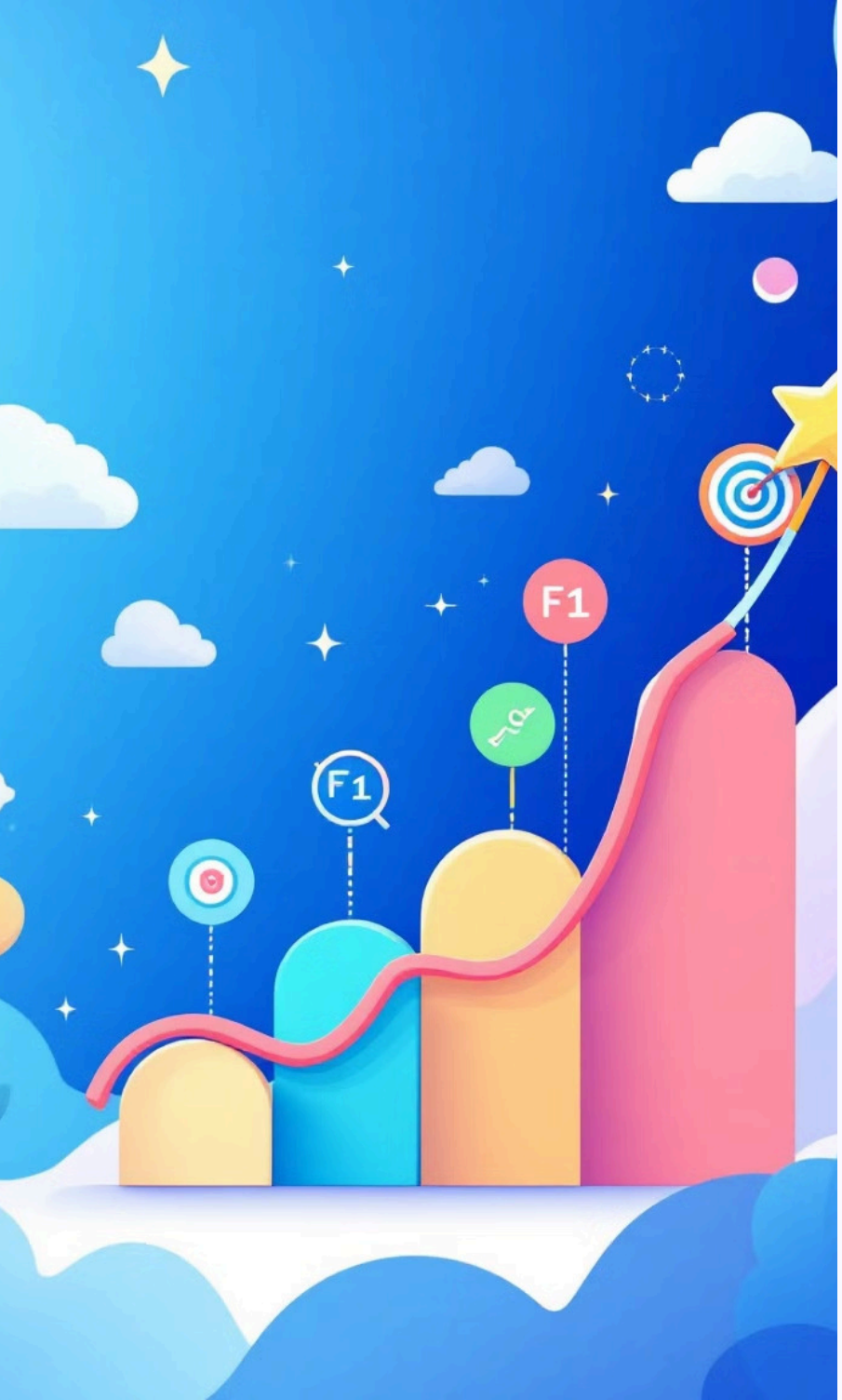
- RandomizedSearchCV para búsqueda
- Pipeline para procesamiento
- Ajuste específico de parámetros

Pipeline del modelo final

Pipeline del Modelo Final:

1. **Preprocesamiento:**
   - **Transformación Categórica:**
     - Variables con muchas categorías (`winery`, `variety`, `denominacion`) codificadas mediante `OrdinalEncoder`.
     - Variables con pocas categorías (`style`, `aging_1`) codificadas mediante `OneHotEncoder`.
   - **Transformación Numérica:**
     - Variables continuas (`price`, `vintage`) escaladas mediante `StandardScaler`.
2. **Escalado Global:**
   - Uso de `RobustScaler` para mitigar el impacto de valores atípicos.
3. **Selección de Características:**
   - `SelectKBest` con `f_regression` como función de evaluación, seleccionando las 14 características más relevantes.
4. **Modelo de Regresión:**
   - **Algoritmo:** `RandomForestRegressor`
   - **Parámetros del Modelo:**
     - Profundidad máxima: 11 (`max_depth=11`).
     - Número de estimadores: 121 (`n_estimators=121`).
     - Número máximo de características a considerar en cada división: `sqrt` (`max_features='sqrt'`).
     - Semilla para reproducibilidad: 63 (`random_state=63`).

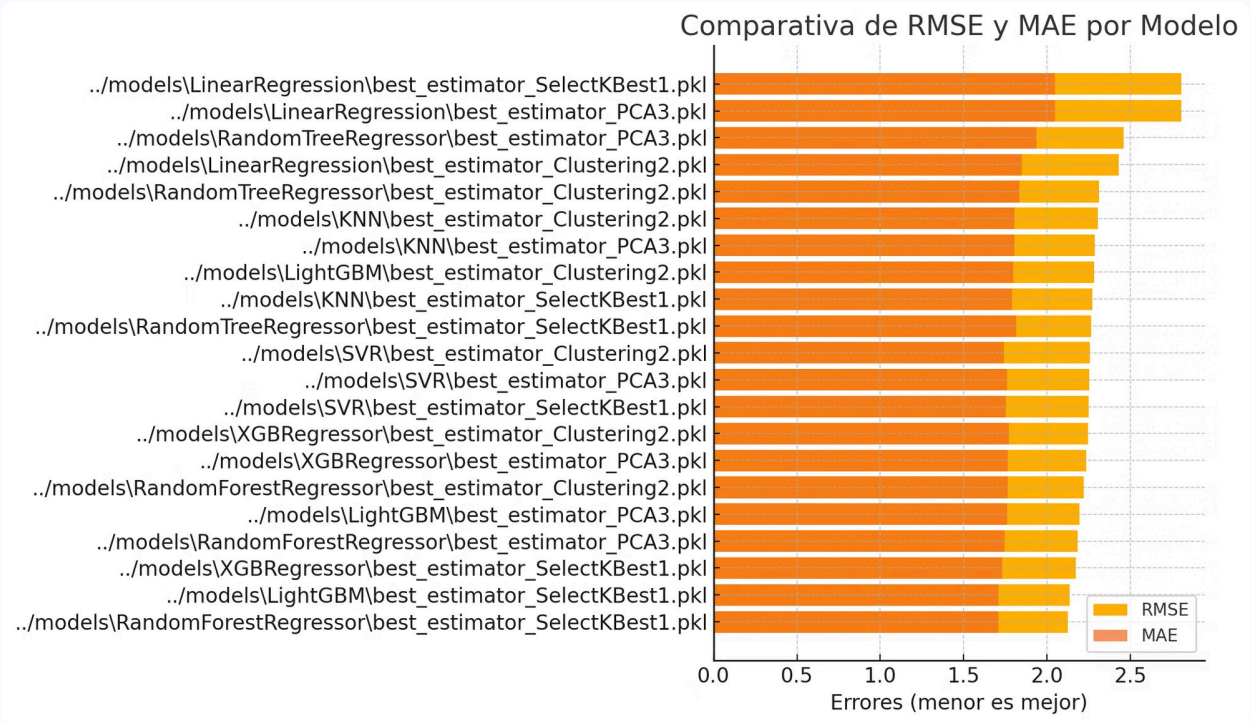




Evaluación del Rendimiento del Modelo

Error

Diferencia entre las predicciones y los valores reales.



Aplicación del Modelo a Nuevos Vinos

1

Paso 1

Recopilación de datos sobre las características del nuevo vino.

2

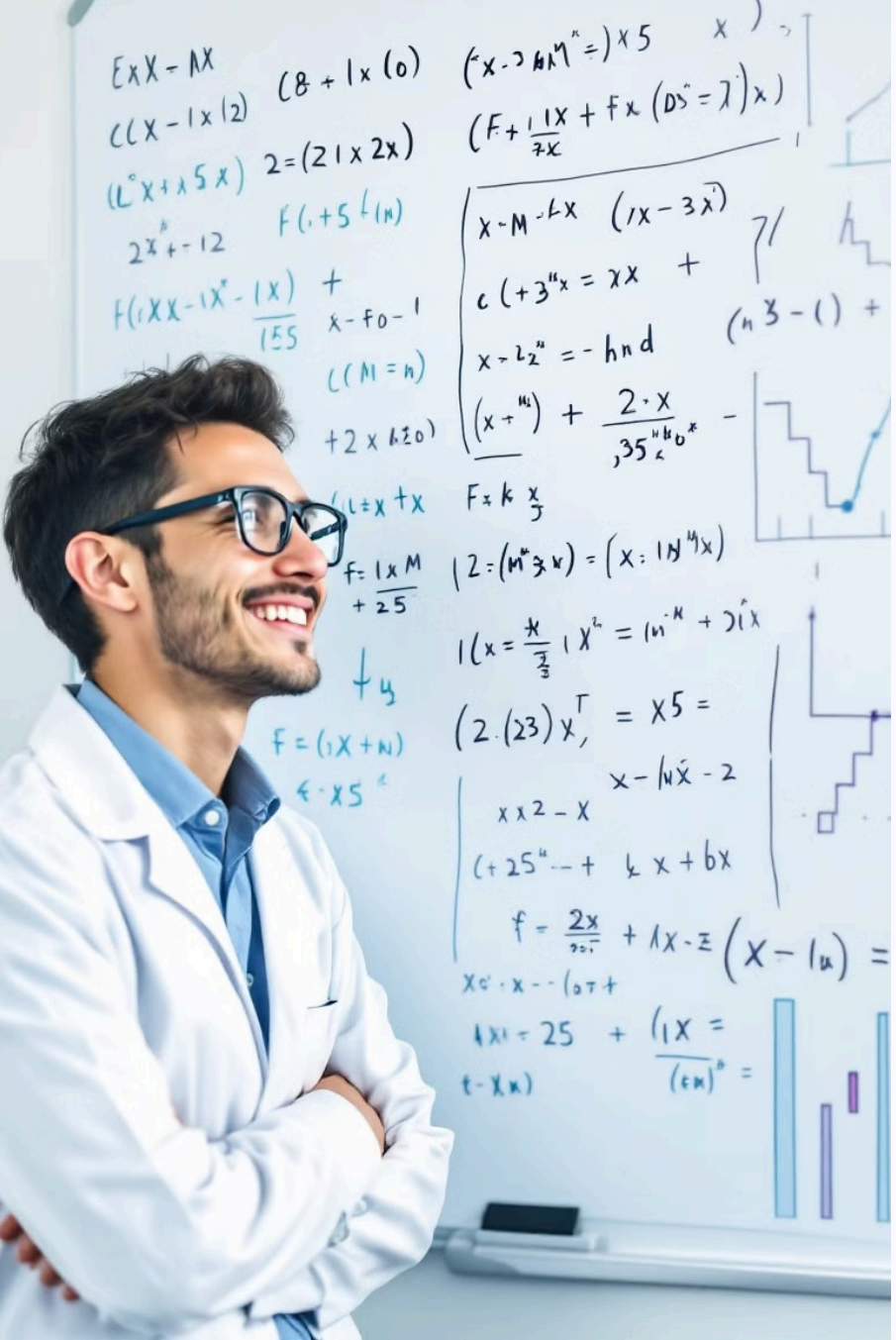
Paso 2

Introducción de los datos al modelo predictivo a través de la app de streamlit.

3

Paso 3

Obtención de la puntuación predicha para el nuevo vino.



Limitaciones y Mejoras

1

Limitaciones

Desequilibrio en las Clases: Pocas observaciones para categorías como "Clásico".

Imputación de Valores Faltantes: Podría haber introducido sesgos en ciertas variables clave.

2

Mejoras

Ampliar el conjunto de datos de entrenamiento.

3

Posibles Direcciones

Integración de análisis técnico para mejorar la precisión.