Rain Prediction Model

Stephanie Grotewold
20210567

Agenda

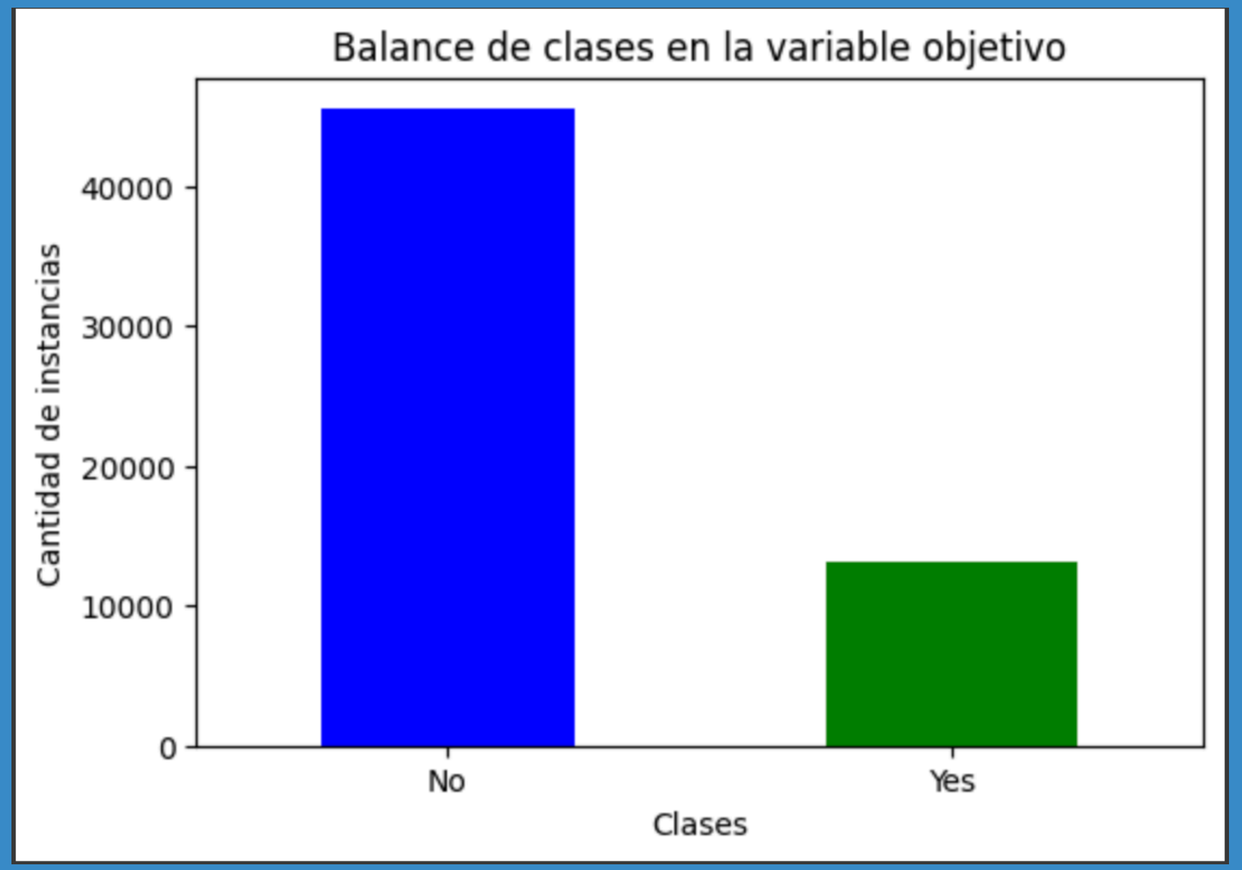
01 Modelos

02 Resultados

03 Conclusiones

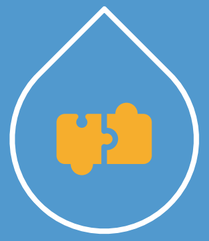
EDA

1. **19 de las 23 features tenian nulos**
2. **48.03% en sunshine fue la variable con más datos nulos**
3. **Un gran desbalance en el dataset con el 75% siendo “NO”**



Modelos

Random Forest



SVM

Logistic Regression

XGBoost

	Yes	No	Promedio
XGBoost	59%	93%	67.16%
SVC	58%	93%	66.55%
Random Forest	71%	89%	64.27%
Logistic Regression	52%	92%	62.09%

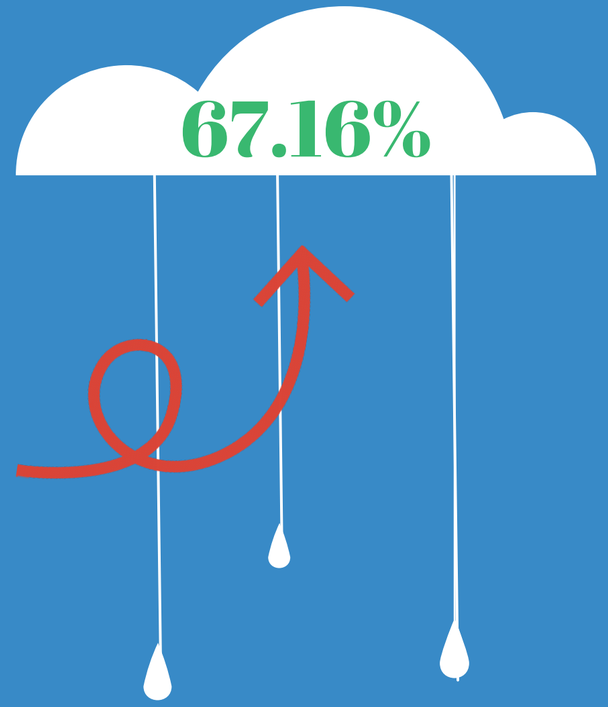
Pasos del Algoritmo

1. **Drop Rows with Missing Target Values**
2. **Encode the Target Variable**
3. **Separate Features and Target**
4. **Identify Numerical and Categorical Features**
5. **Numeric Preprocessor:**
    a. SimpleImputer
    b. StandardScaler
6. **Categorical Preprocessor**
    a. SimpleImputer
    b. OneHotEnconder
7. **ColumnTransformer:** combina el procesamiento
8. **Pipeline de XGBoost:**
    a. Preprocesamiento ('preprocessor'): Prepara los datos antes de entrenar el modelo.
    b. Sobremuestreo ('oversampler'): Aborda el desequilibrio de clases usando RandomOverSampler.
    c. Clasificador ('classifier'): Utiliza XGBoost para entrenar el modelo predictivo.
9. **Entrenar Modelo, Predecir, Evaluar**

Grid Search

Learning rate: 0.2
max_depth: 7
n_estimators: 300
subsample: 0.7
Con un cross validation de 8

66%



Conclusiones y recomendaciones

- ¿Por qué se eligió XGBoost?
  - Eficiencia computacional
  - Manejo de datos faltantes: Capacidad integrada para manejar valores faltantes.
  - Regularización integrada: Evita sobreajuste y mejora generalización.

- Optimización de hiperparámetros
- Probar otros modelos

Thanks :P