Identificación de zonas de riesgo en Barcelona

Anna Moreno
@alohaanna

Resumen

1. Análisis y limpieza de los datos
2. EDA (exploratory data analyst)
3. Modelos
4. Predicción
5. Conclusiones

Datos

Datos privados

Categóricas

- Districte
- Barri
- Espai
- Data
- Hora
- Franja Horaria
- Semàfor

Numéricas

- Homes
- Dones
- No Binari

Binarias

- Consum
- Pernocta
- Proximitat recurs
- Punt Trobada

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 21344 entries, 0 to 21343
Data columns (total 30 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   districte_id            21344 non-null  int64
 1   districte               21344 non-null  object
 2   barri                   21344 non-null  object
 3   espai_id                21344 non-null  int64
 4   data                    21344 non-null  object
 5   hora                    21344 non-null  object
 6   franja_horaria          21344 non-null  object
 7   semafor                 21344 non-null  object
 8   num_menors_homes        21344 non-null  int64
 9   num_adults_homes        21344 non-null  int64
 10  num_per_determ_homes_18 21344 non-null  int64
 11  num_per_determ_homes_21 21344 non-null  int64
 12  num_menores_dones       21344 non-null  int64
 13  num_adultes_dones       21344 non-null  int64
 14  num_per_determ_dones_18 21344 non-null  int64
 15  num_per_determ_dones_21 21344 non-null  int64
 16  num_no_binari           21344 non-null  int64
 17  punt_trobada            3405 non-null   object
 18  proximitat_recurs       1362 non-null   object
 19  lloc_pernocta           1511 non-null   object
 20  altres                  163 non-null    object
 21  consum_cannabis         685 non-null    object
 22  consum_alcohol          182 non-null    object
 23  consum_altres_toxics    153 non-null    object
 24  consum_inhalants        115 non-null    object
 25  consum_psicofarmacs     65 non-null     object
 26  consum_sense_determ     87 non-null     object
 27  activitat_delictiva     87 non-null     object
 28  salut_mental_impactant  28 non-null     object
 29  zonas                   21344 non-null  object
dtypes: int64(11), object(19)
memory usage: 4.9+ MB
```

Objetivo

Crear un modelo de clasificación supervisada.

- Riesgo: rojo, amarillo, verde.
- Dinámicas de impacto.

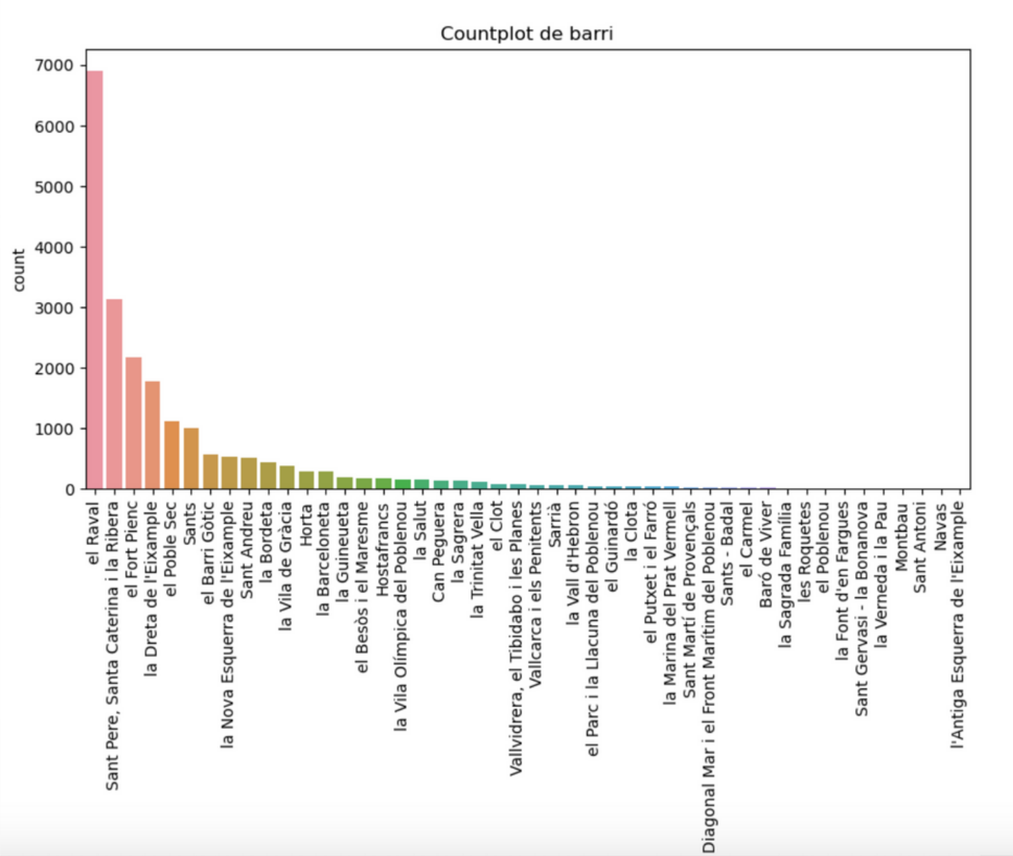
Limpieza de datos

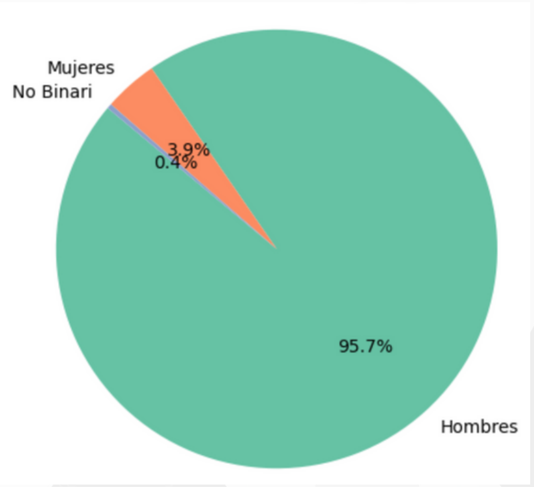
1. Valores duplicados
2. Valores absolutos
3. Agrupación de variables
    a. hombres, mujeres
    b. drogas

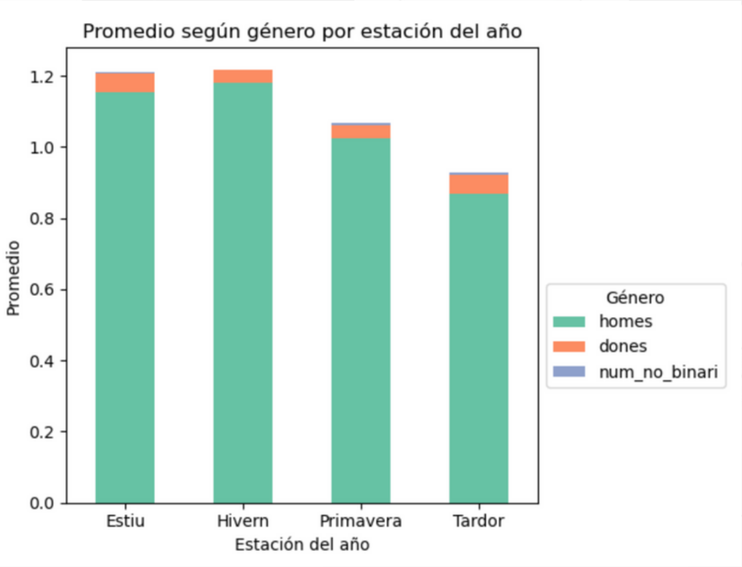
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 21054 entries, 0 to 21343
Data columns (total 17 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   districte          21054 non-null  object
 1   barri              21054 non-null  object
 2   espai_id           21054 non-null  int64
 3   data               21054 non-null  datetime64[ns]
 4   hora               21054 non-null  object
 5   franja_horaria     21054 non-null  object
 6   semafor            21054 non-null  object
 7   num_no_binari      21054 non-null  int64
 8   punt_trobada       21054 non-null  float64
 9   proximitat_recurs  21054 non-null  float64
 10  lloc_pernocta      21054 non-null  float64
 11  zonas              21054 non-null  object
 12  homes              21054 non-null  int64
 13  dones              21054 non-null  int64
 14  consum             21054 non-null  float64
 15  año                21054 non-null  int64
 16  estacio            21054 non-null  object
dtypes: datetime64[ns](1), float64(4), int64(5), object(7)
memory usage: 2.9+ MB
```

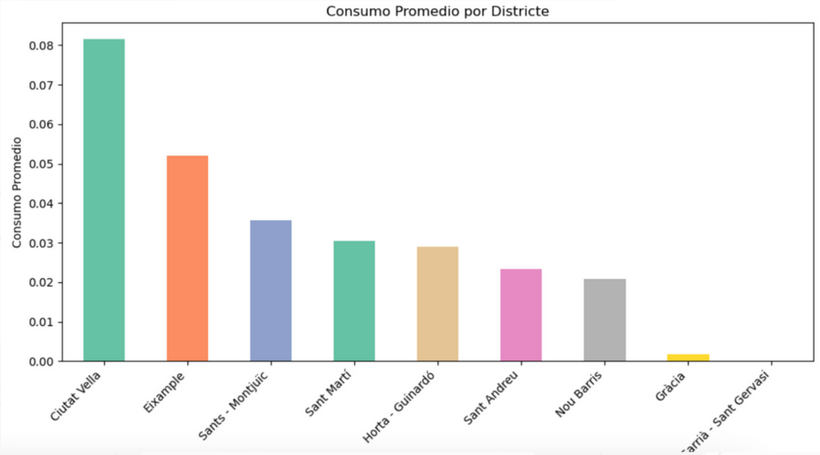
Análisis de datos

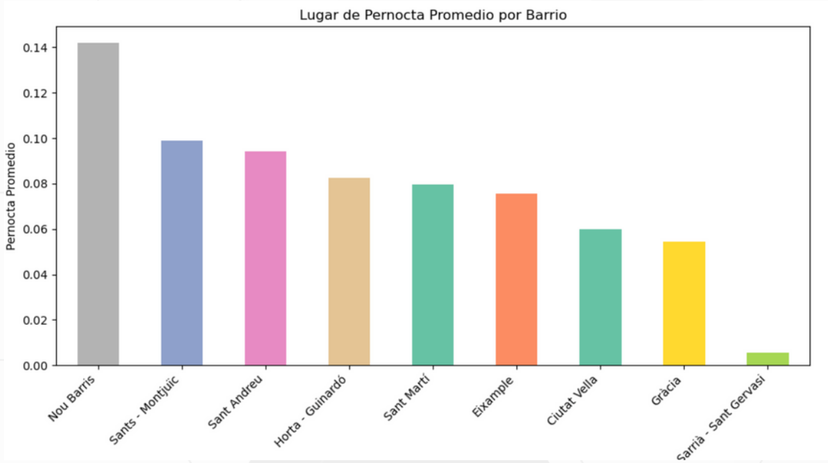


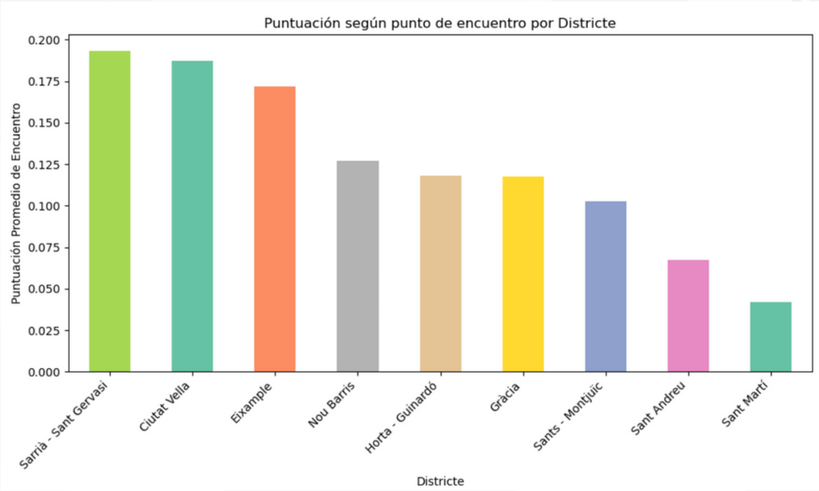


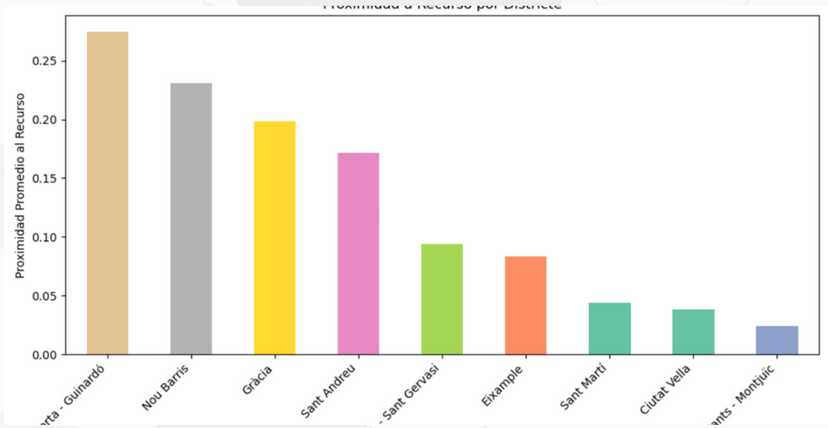


Análisis de datos



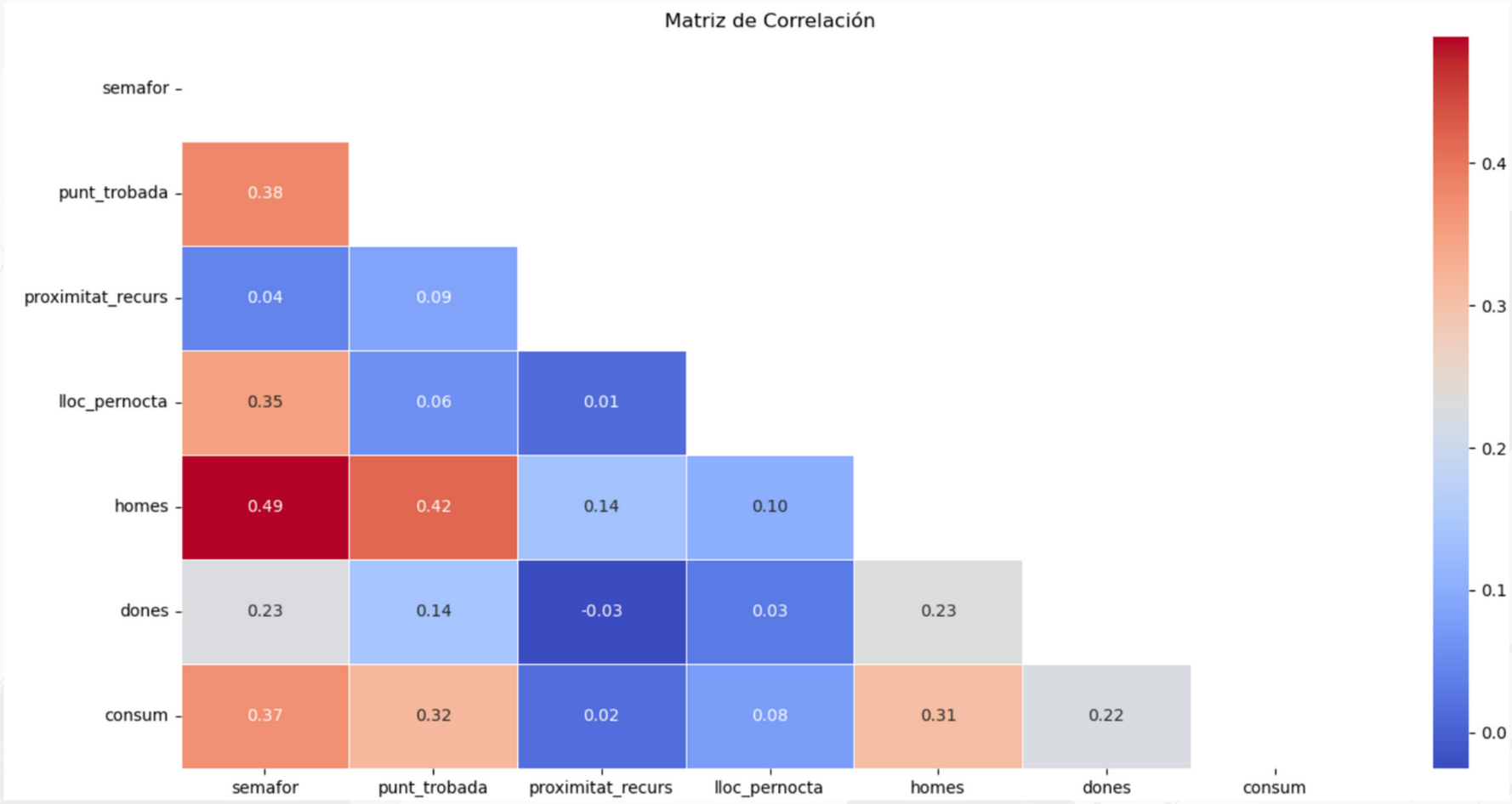




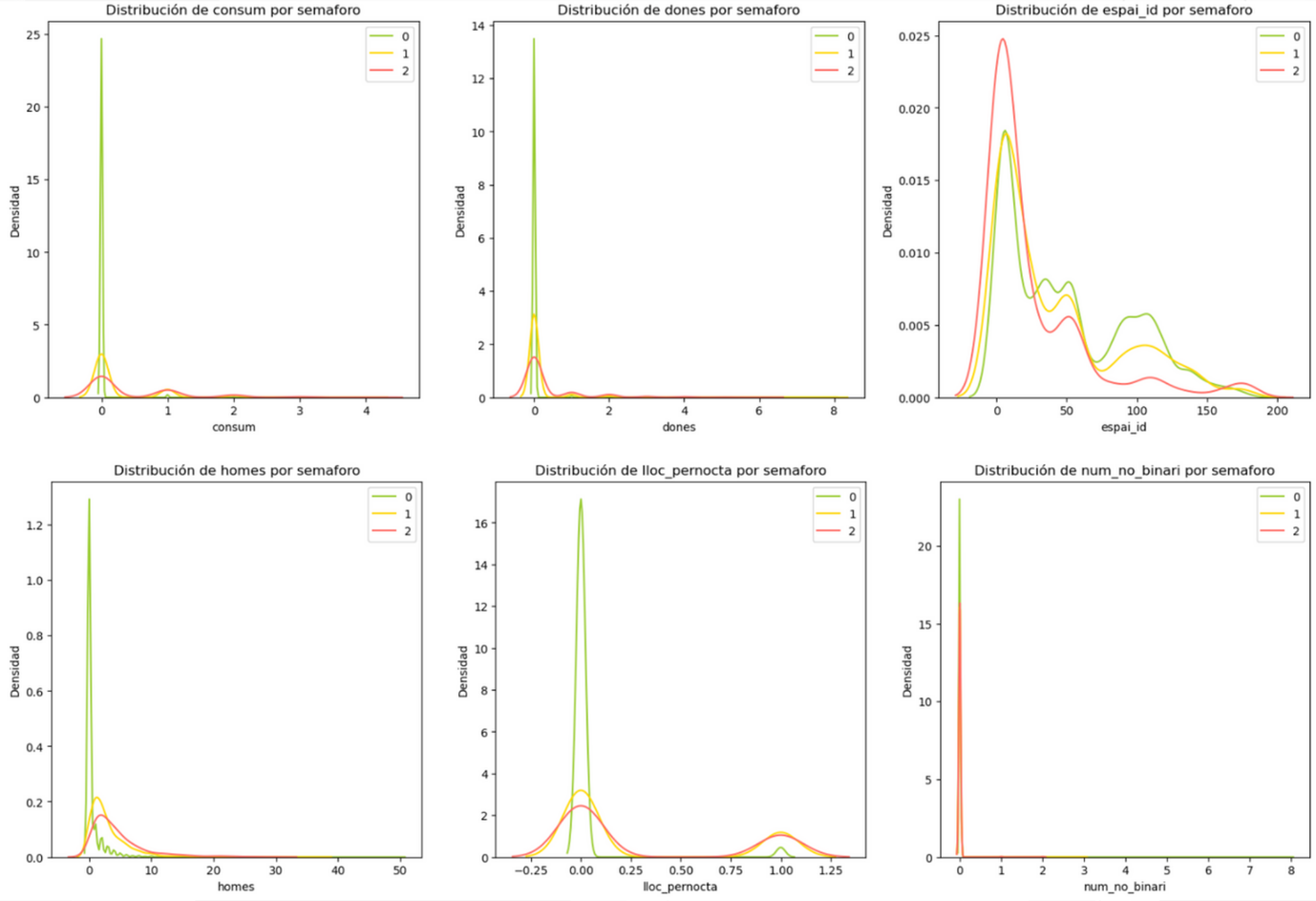


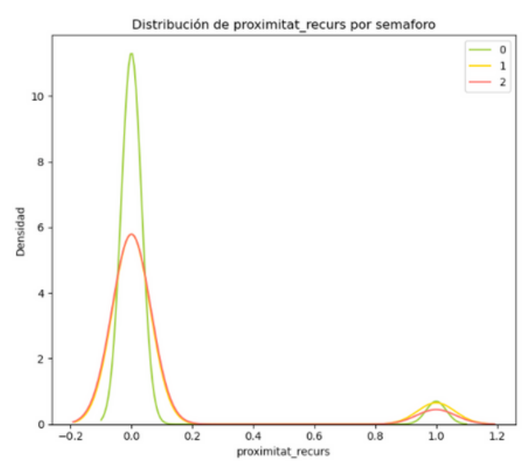
Dinàmicas vinculadas a distritos

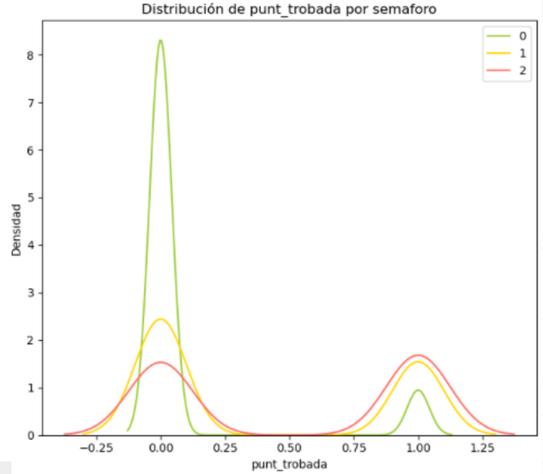
Correlación entre variables



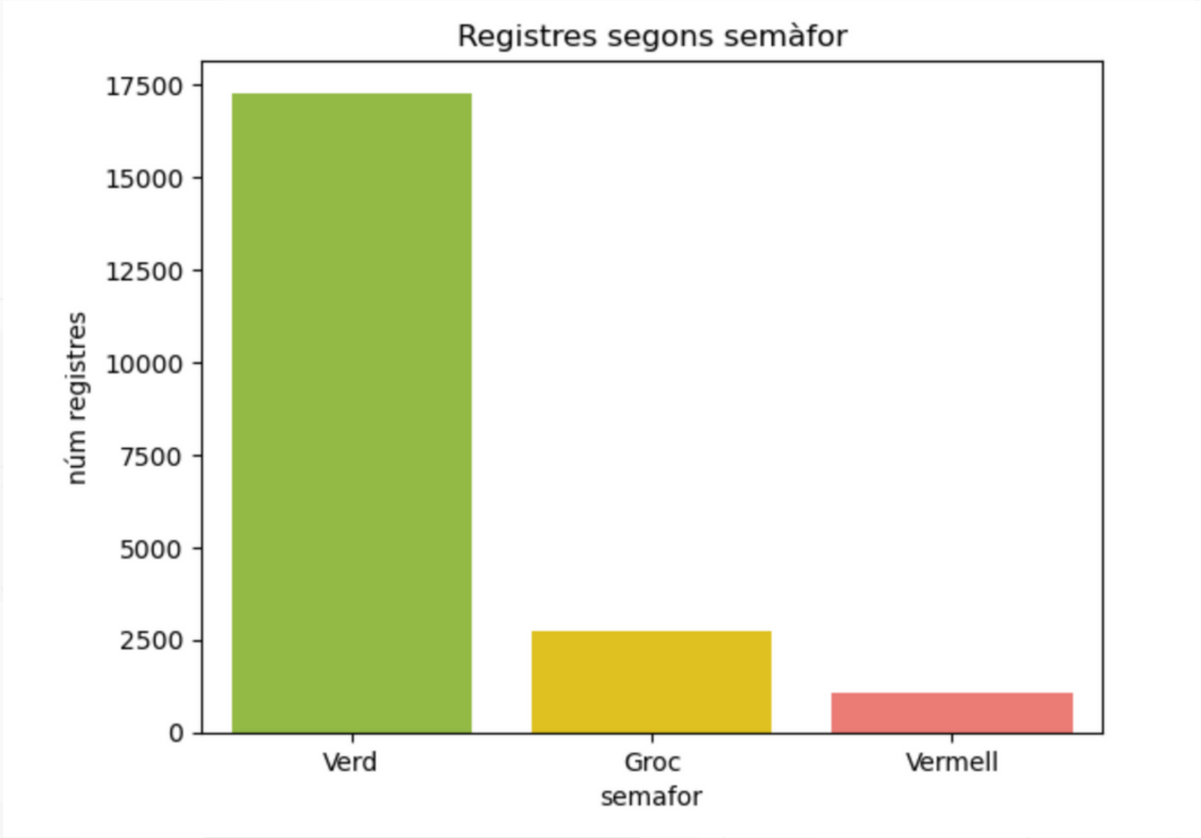
Distribución variables







Variable target

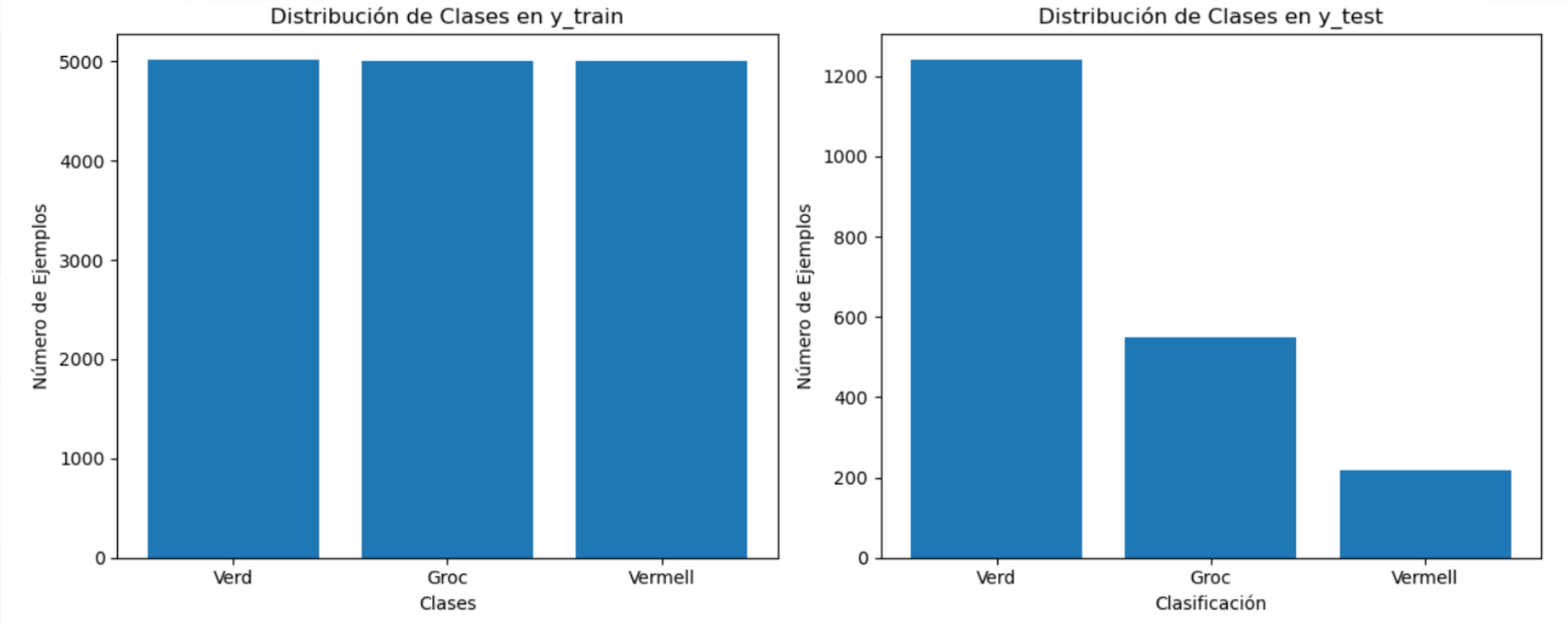


- Descompensación de la variable target
- Sobre representación de riesgo bajo
- Rendimiento eficiente

Preprocesamiento de datos

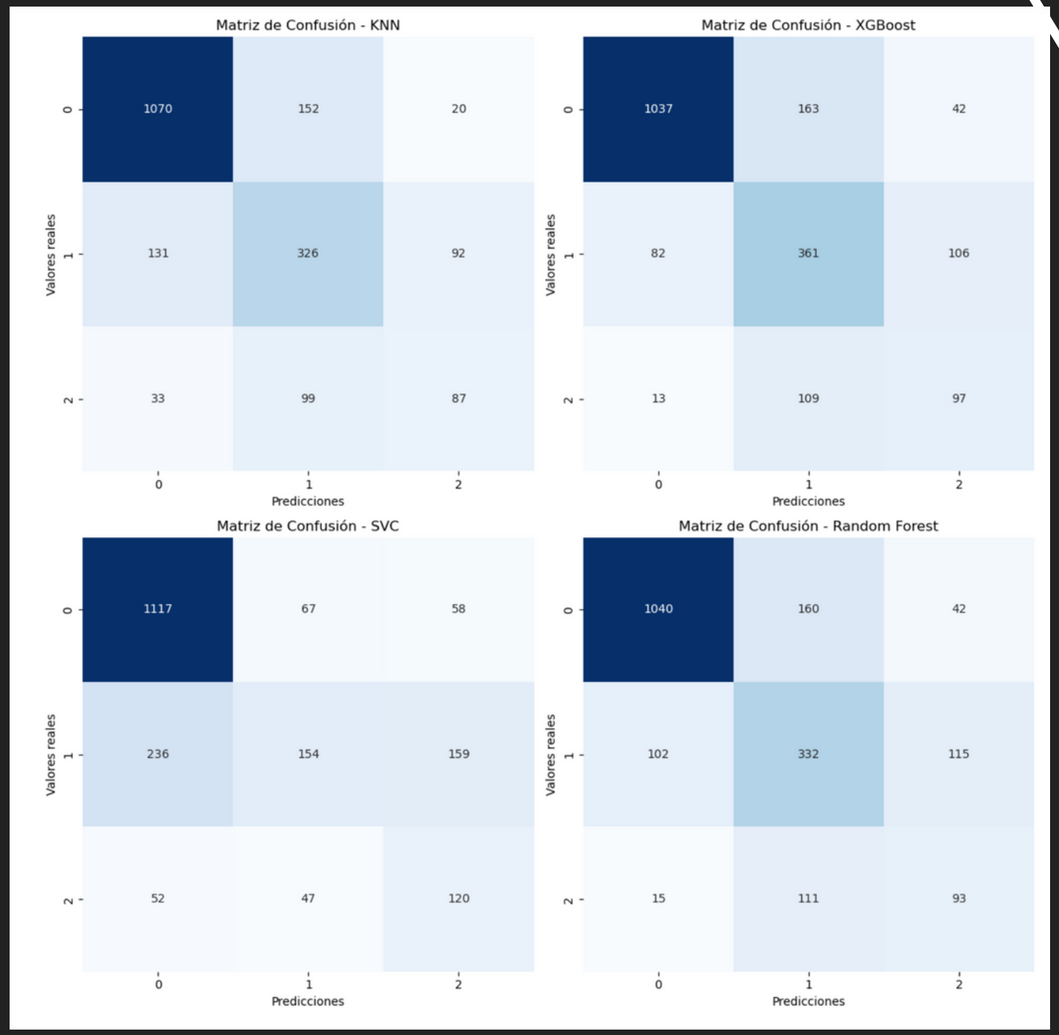
Técnicas undersampling y oversampling

- **SMOTE:** muestras sintéticas de la clases minoritarias.
- Combinadas - el reequilibrio y el rendimiento eficiente del modelo

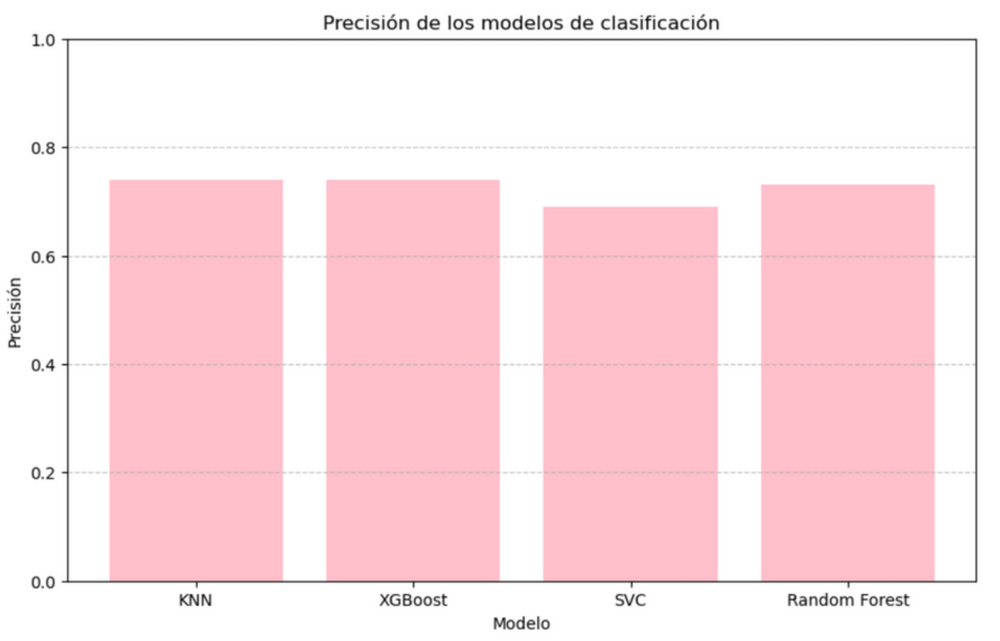


Modelos

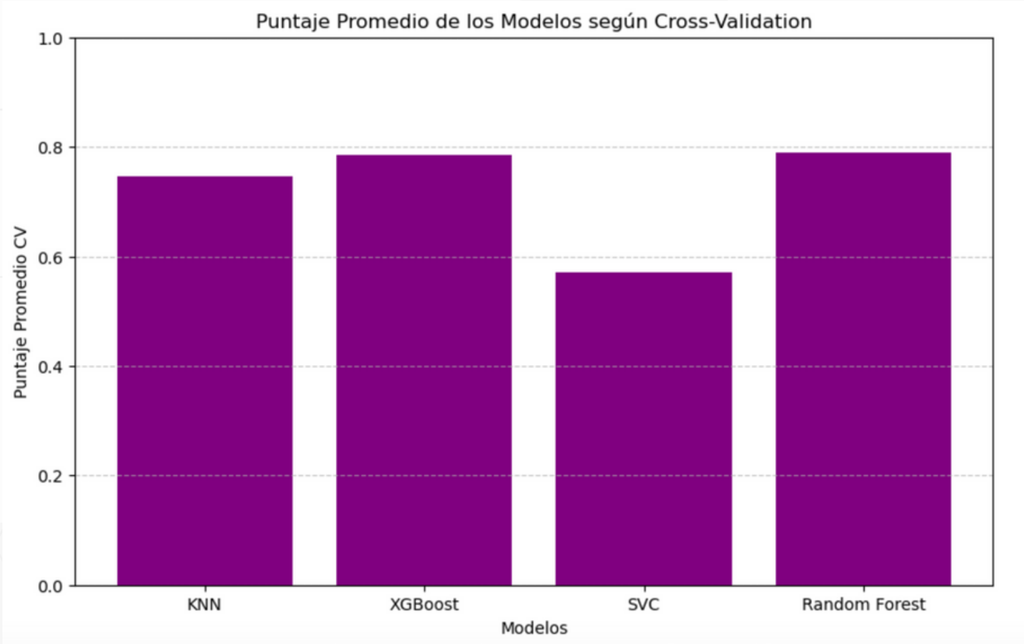
- KNN
- XGBoost
- SVC
- Random Forest



Evaluación de los modelos



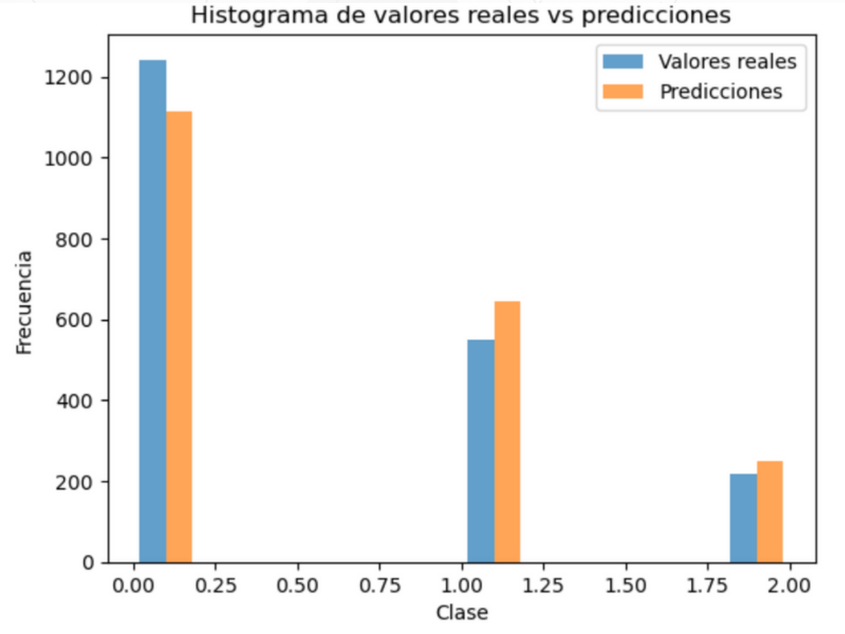
Accuracy de los modelos:
KNN: 0.74
XGBoost: 0.74
SVC: 0.69
Random Forest: 0.73



KNN: Media CV: 0.746 Desviación estándar CV: 0.021
XGBoost: Media CV: 0.785 Desviación estándar CV: 0.036
SVC: Media CV: 0.572 Desviación estándar CV: 0.009
Random Forest: Media CV: 0.790 Desviación estándar CV: 0.032

Modelo

XGBoost



Predicción con streamlit

Predicciones

Streamlit

Conclusiones

- Falta de datos socio económicos
- Variables no discriminantes
- Subjetividad del registrador
- Registro sesgado - modelo sesgado

Gracias

alohanna

Anna Moreno