Borcelle Hospital

Equity Post-HCT Survival Predictions

"Equipo O+"

El Desafío del Trasplante de Células Hematopoyéticas (TCH)

¿Qué es TCH?

Procedimiento que reemplaza células madre defectuosas con células sanas de un donante o del propio paciente.
Existen dos tipos principales:

- Autotrasplante.
- Alotrasplante.

Pacientes que lo requieren

El TCH se utiliza para tratar a pacientes con enfermedades que afectan a la sangre, el sistema inmunitario o algunos tipos de cáncer. Por ejemplo, Leucemias agudas y crónicas, Linfomas, Mieloma múltiple, Enfermedades de la médula ósea, Inmunodeficiencias, Lupus, etc.

Impacto de la Inequidad en Datos de TCH

Problema

Sesgo en los datos médicos afecta la disponibilidad de donantes compatibles, la investigación y los resultados en trasplantes de células hematopoyéticas.

Consecuencias

Subrepresentación en registros de donantes, por lo que ciertos grupos étnicos tienen menos probabilidad de encontrar donantes.

Desigualdades en diagnóstico y acceso a centros especializados

Introducción al problema

Descripción del problema

Los modelos predictivos actuales frecuentemente no logran abordar las disparidades relacionadas con el estatus socioeconómico, la raza y la geografía.
Resolver estas brechas es fundamental para:

- Mejorar la atención al paciente
- Optimizar el uso de recursos

Solución propuesta

Desarrollar modelos predictivos basados en inteligencia artificial que mejoren la estratificación de riesgo, anticipe recaídas y optimice decisiones clínicas en pacientes de trasplante alogénico, eliminando sesgos socioeconómicos, raciales y geográficos

Descripción del Dataset

Fuente y Naturaleza del Dataset

El dataset utilizado proviene de datos sintéticos generados para reflejar situaciones reales de pacientes sometidos a trasplante alogénico de células hematopoyéticas (HCT). Se compone de variables clínicas, demográficas y de seguimiento, que permiten evaluar la probabilidad de supervivencia post-HCT. Aunque los datos son sintéticos, han sido diseñados para mantener las características y complejidades de los datos reales, garantizando la privacidad de los pacientes sin comprometer la validez del análisis.

Fuente de los datos: Kaggle - Equity in Post-HCT Survival

Tamañoo y Limitaciones

El dataset contiene 28,800 registros y 60 columnas, representando diferentes características de los pacientes. Entre las principales limitaciones del dataset se encuentran:

- **Datos censurados**: Algunas observaciones pueden no incluir el desenlace final del paciente, lo que complica la modelación de la supervivencia.
- **Posible sesgo en la generación de datos**: Aunque los datos han sido diseñados para reflejar la realidad, pueden no capturar completamente las relaciones complejas presentes en los datos clínicos reales.

Columnas clave

Entre las columnas clave del dataset, destacan:

- **efs_time**: Tiempo transcurrido hasta el evento de fallo o censura en la supervivencia del paciente.
- **efs**: Variable indicadora de evento (1 si ocurrió el evento, O si el paciente no sufrió un evento hasta ese momento).
- **y_nel**: Transformación del tiempo de fallo basada en la función de Nelson-Aalen para mejorar la modelación de la variable objetivo.

```
from lifelines import NelsonAalenFitter

def create_nelson(data):
    data = data.copy()
    naf = NelsonAalenFitter(nelson_aalen_smoothing=0)
    naf.fit(durations=data['efs_time'],
    event_observed=data['efs'])
    return naf.cumulative_hazard_at_times(data
    ['efs_time']).values * -1

train["y_nel"] = create_nelson(train)
train.loc[train.efs == 0, "y_nel"] = (-(-train.loc
[train.efs == 0, "y_nel"])**0.5)
```

Limpieza de datos

Se aplican conversiones de tipo y codificación de variables categóricas para que el dataset sea compatible con el modelo XGBoost.

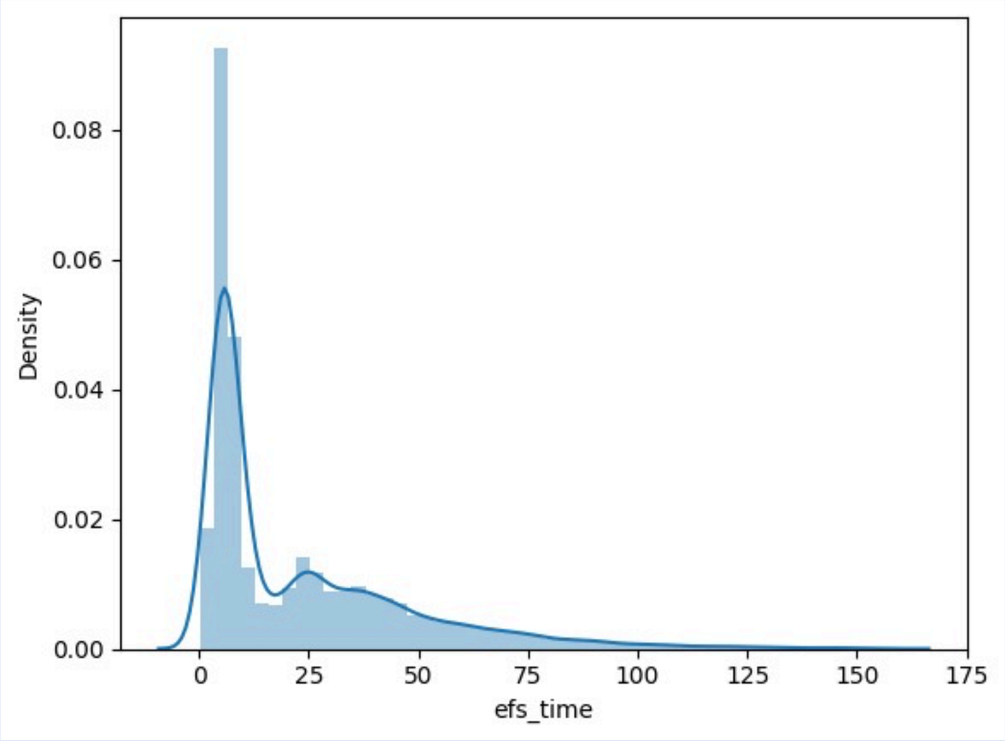
```
from sklearn.model_selection import KFold,StratifiedKFold
from xgboost import XGBRegressor, XGBClassifier
import xgboost as xgb
print("Using XGBoost version",xgb.__version__)
```

Tras cargar los datos, y observarlos notamos que existen datos nulos, se utilizan métodos como fillna() para imputar o reemplazar esos valores.

```
CATS = []
for c in FEATURES:
    if train[c].dtype=="object":
        CATS.append(c)
        train[c] = train[c].fillna("NAN")
        test[c] = test[c].fillna("NAN")
print(f"In these features, there are {len(CATS)} CATEGORICAL FEATURES: {CATS}")
```

Análisis exploratorio de datos

Histogramas y Distribuciones

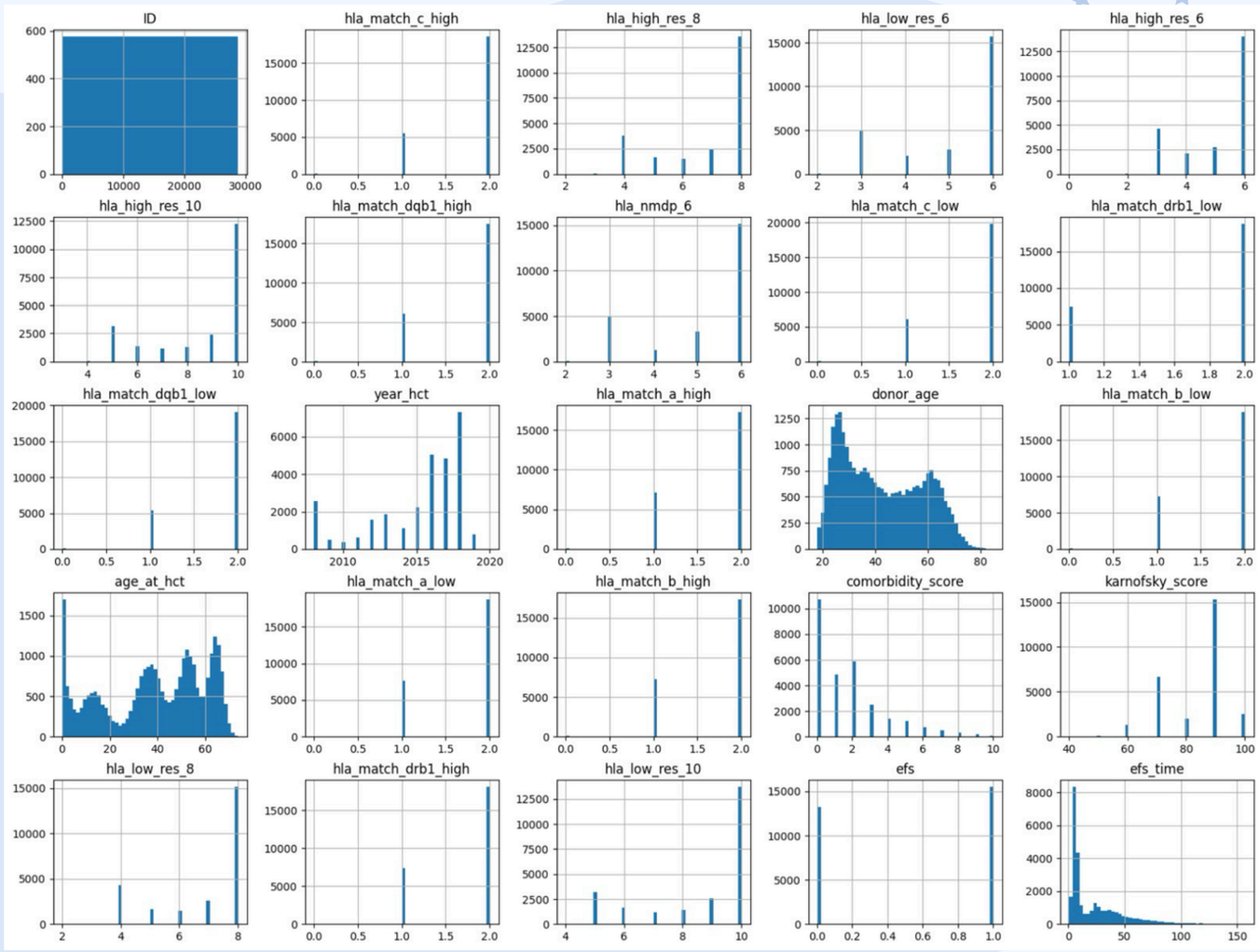


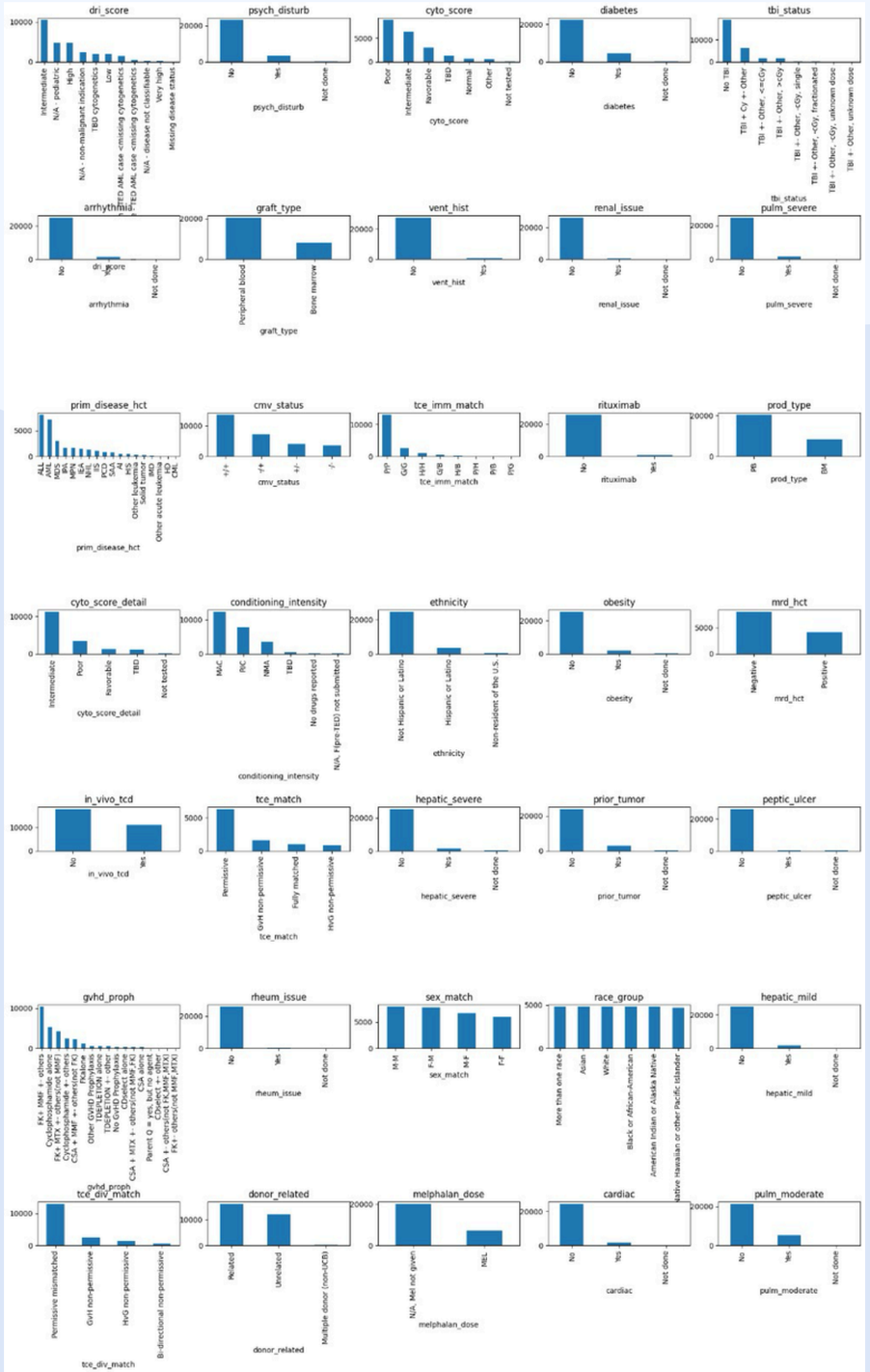
Análisis exploratorio de datos

Histogramas y Distribuciones

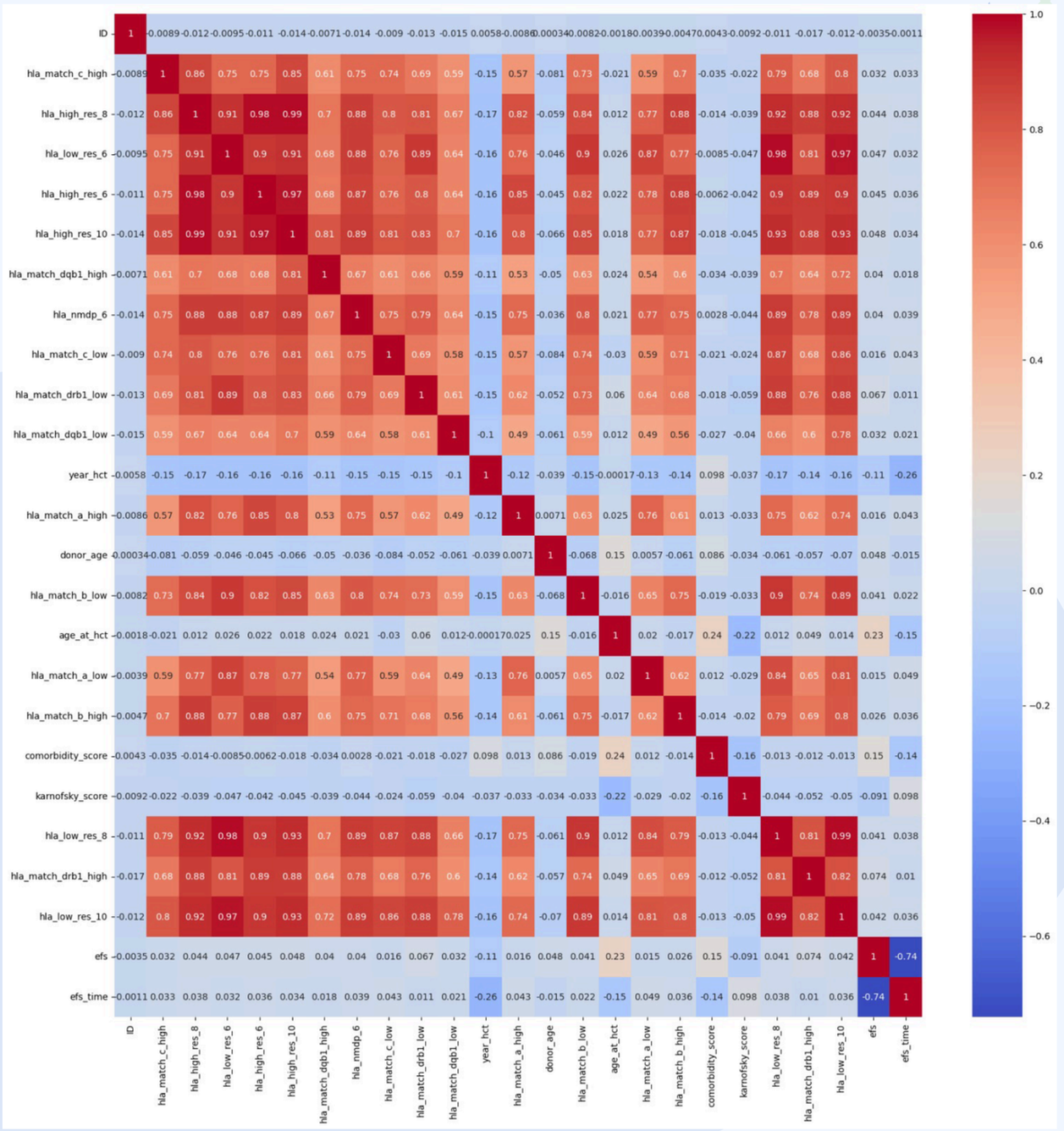
```
train.isna().sum().sort_values(ascending=False)[:10]
```

	0
tce_match	18996
mrd_hct	16597
cyto_score_detail	11923
tce_div_match	11396
tce_imm_match	11133
cyto_score	8068
hla_high_res_10	7163
hla_high_res_8	5829
hla_high_res_6	5284
hla_match_dqb1_high	5199




dri_score
Intermediate
N/A - pediatric
High
N/A - non-malignant indication
TBD cytogenetics
Low
Missing disease status
psych_disturb
No
Yes
Not done
cyto_score
Poor
Intermediate
Favorable
TBD
Normal
Other
Not tested
diabetes
No
Yes
Not done
tbi_status
No TBI
TBI + Cy +- Other
TBI + Other, <=cGy
TBI + Other, >cGy
TBI + Other, <Gy, single
TBI +- Other, <Gy, fractionated
TBI + Other, <Gy, unknown dose
arrhythmia
No
Yes
Not done
graft_type
Peripheral blood
Bone marrow
vent_hist
No
Yes
renal_issue
No
Yes
Not done
pulm_severe
No
Yes
Not done
prim_disease_hct
ALL
AML
MDS
IEA
MPN
NHL
HS
PCD
SAA
IA
HIS
Other leukemia
Solid tumor
IMD
Other acute leukemia
HD
CML
cmv_status
+/+
-/+
+/-
-/-
tce_imm_match
P/P
G/G
H/H
G/B
H/B
P/H
P/B
P/G
rituximab
No
Yes
prod_type
PB
BM
cyto_score_detail
Intermediate
Poor
Favorable
TBD
Not tested
conditioning_intensity
MAC
RIC
NMA
TBD
No drugs reported
N/A, F(pre-TED) not submitted
ethnicity
Not Hispanic or Latino
Hispanic or Latino
Non-resident of the U.S.
obesity
No
Yes
Not done
mrd_hct
Negative
Positive
in_vivo_tcd
No
Yes
tce_match
Permissive
GvH non-permissive
Fully matched
HvG non-permissive
hepatic_severe
No
Yes
Not done
prior_tumor
No
Yes
Not done
peptic_ulcer
No
Yes
Not done
gvhd_proph
FK + MMF + others
Cyclophosphamide alone
FK + MTX +- others(not MMF)
Cyclosporine alone + others
CSA + MMF +- others(not FK)
Other
Other GVHD Prophylaxis
TDEPLETION alone
TDEPLETION + other
No GVHD Prophylaxis
CSA + MTX + -others(not MMF, FK)
CSA alone
Parent Q + yes, but no agent
CDselect +- other
CSA + -others(not FK,MMF,MTX)
FK + others(not MMF,MTX)
rheum_issue
No
Yes
Not done
sex_match
M-M
F-M
M-F
F-F
race_group
More than one race
Asian
White
Black or African American
American Indian or Alaska Native
Native Hawaiian or other Pacific Islander
hepatic_mild
No
Yes
Not done
tce_div_match
Permissive mismatched
GvH non-permissive
HvG non-permissive
Bi-directional non-permissive
donor_related
Related
Unrelated
Multiple donor (non-UCB)
melphalan_dose
N/A, Mel not given
MEL
cardiac
No
Yes
Not done
pulm_moderate
No
Yes
Not done

Análisis de correlaciones



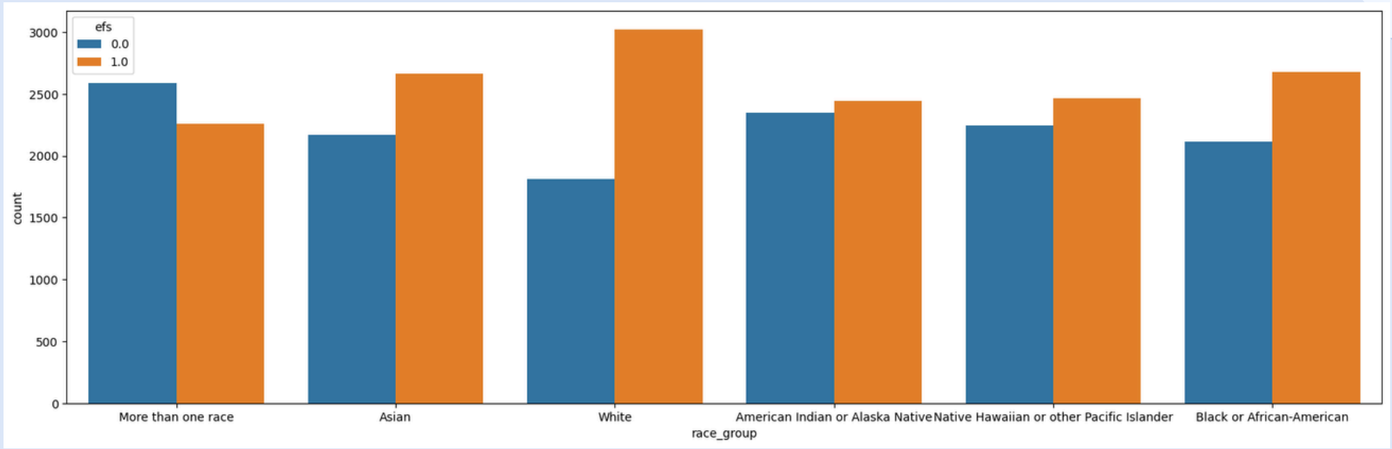
Visualizaciones Adicionales

```
train["efs"].value_counts()
```

efs	count
1.0	15532
0.0	13268

Visualizaciones Adicionales

race_group	count
More than one race	4845
Asian	4832
White	4831
Black or African-American	4795
American Indian or Alaska Native	4790
Native Hawaiian or other Pacific Islander	4707



Análisis Profundos Propuestos

Análisis de regresión

Es una técnica estadística utilizada para modelar y analizar la relación entre una variable dependiente y una o más variables independientes. Su objetivo es predecir valores y comprender la influencia de las variables independientes sobre la dependiente.

Objetivo del análisis

Se busca modelar la relación entre diversas variables del dataset y la variable objetivo **y_nel**, que representa la estimación del riesgo de los pacientes tras el trasplante. Este análisis permitirá evaluar qué factores influyen más en la predicción de la supervivencia.

Potencial impacto del análisis

Se espera identificar variables clave que afectan la supervivencia post-HCT y así mejorar la precisión del modelo predictivo y a desarrollar estrategias de intervención más efectivas para los pacientes en alto riesgo.

Análisis Profundos Propuestos

Análisis de clasificación

A diferencia de la regresión, que predice valores continuos, la clasificación se utiliza para predecir valores discretos.

Objetivo del análisis

La clasificación se centrará en predecir si un paciente tiene un alto o bajo riesgo de no sobrevivir tras el trasplante, basándose en la variable binaria **efs** (evento de falla del trasplante).

Potencial impacto del análisis

Diferenciar grupos de pacientes con diferentes niveles de riesgo, además de evaluar si ciertos subgrupos poblacionales presentan sesgos en la predicción, asegurando un modelo más equitativo.

Análisis Profundos Propuestos

Análisis de clusterización

Busca agrupar un conjunto de objetos en grupos (clusters) de tal manera que los objetos dentro del mismo grupo sean más similares entre sí.

Expectativas de Analisis

Se pretende segmentar a los pacientes en grupos según su nivel de riesgo post-HCT, identificando aquellos más propensos a eventos adversos y aquellos con mejor pronóstico de supervivencia.

Modelado

Modelos entrenados

Clasificación

XGBClassifier

Se usa para clasificación, es decir, predecir categorías.

Regresion

XGBRegressor

Se usa para regresión, es decir, predecir valores numéricos

Clusterización

k-means

Algoritmo de clusterización que agrupa datos en "k" grupos según su similitud.

Aprendizaje no supervisado: Clustering

```
# Train k_means with 4 clusters
kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10, random_state=0)
y_kmeans = kmeans.fit_predict(train_clusters_sclaed)
train_clusters_sclaed["cluster"] = y_kmeans
```

Los clusters se formaron utilizando las siguientes características:

1. **karnofsky_score** (Qué tan bien puede una persona realizar sus actividades diarias normales sin ayuda médica)
2. **dri_score** (puntuación de riesgo de donante)
3. **prim_disease_hct** (enfermedad primaria del paciente)
4. obesity (obesidad)
5. **donor_related** (donante relacionado o no)
6. **sex_match** (compatibilidad de sexo)
7. **cyto_score** (puntuación citogenética)
8. **donor_age** (edad del donante)
9. **age_at_hct** (edad al momento del trasplante)

Aprendizaje no supervisado: Clustering

Cluster 0

Peor estado funcional, mayor edad y riesgo general.

Cluster 1

Mejor estado funcional, menor edad y riesgo bajo.

Cluster 2

Características intermedias en edad, riesgo y enfermedad.

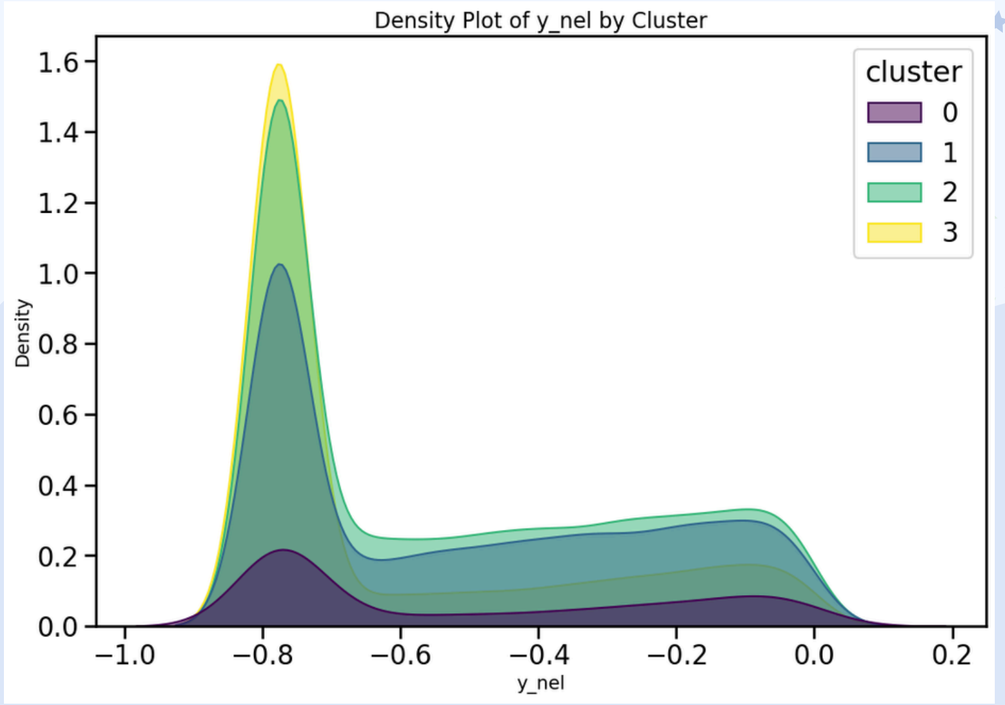
Cluster 3

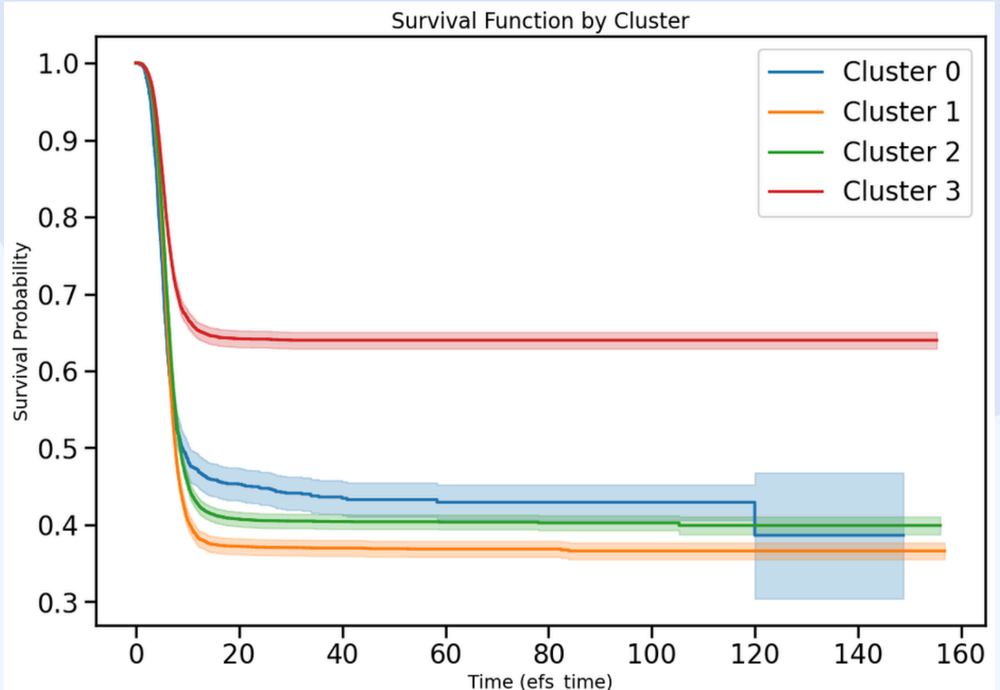
Alto riesgo genético, aunque buen estado funcional.

Supervivencia (y_nel)

El valor negativo de y_nel (que representa el hazard acumulado de Nelson-Aalen) muestra diferencias entre clusters:

Clusters 0 y 3: peor pronóstico.
Clusters 1 y 2: mejor pronóstico.





Aprendizaje supervisado: Regresión

Transformación del target

Se hizo uso de una herramienta estadística de riesgo acumulado (NelsonAalenFitter) para unificar las variables “efs” y “efs_time”, con el fin de utilizar el modelo XGBoost, ya que este no entiende datos de supervivencia.

Estratificación por grupo racial

Con el fin de obtener los resultados con el menor sesgo posible se utilizó Stratified K-Fold, una función que permite que cada fold tenga proporciones similares, en este caso aplicado al factor del grupo racial.

Aprendizaje supervisado: Regresión

XGBRegressor

El objetivo es predecir con exactitud el riesgo de un paciente que se somete a este tratamiento.

La mejor precisión obtenida fue: 0.690316

```python
for i in range(FOLDS):

    print("#"*25)
    print(f"### Fold {i+1}")
    print("#"*25)

    x_train = train.loc[train.fold!=i,FEATURES].copy()
    y_train = train.loc[train.fold!=i,"y_nel"]
    x_valid = train.loc[train.fold==i,FEATURES].copy()
    y_valid = train.loc[train.fold==i,"y_nel"]
    x_test = test[FEATURES].copy()

    model_xgb = XGBRegressor(
        # device="cuda",
        max_depth=4,
        colsample_bytree=0.55,
        subsample=0.8,
        n_estimators=5000,
        learning_rate=0.02,
        enable_categorical=True,
        min_child_weight=80,
        early_stopping_rounds=200,
        n_jobs=4
    )
    model_xgb.fit(
        x_train, y_train,
        eval_set=[(x_valid, y_valid)],
        verbose=500
    )
```

Aprendizaje supervisado: Clasificación

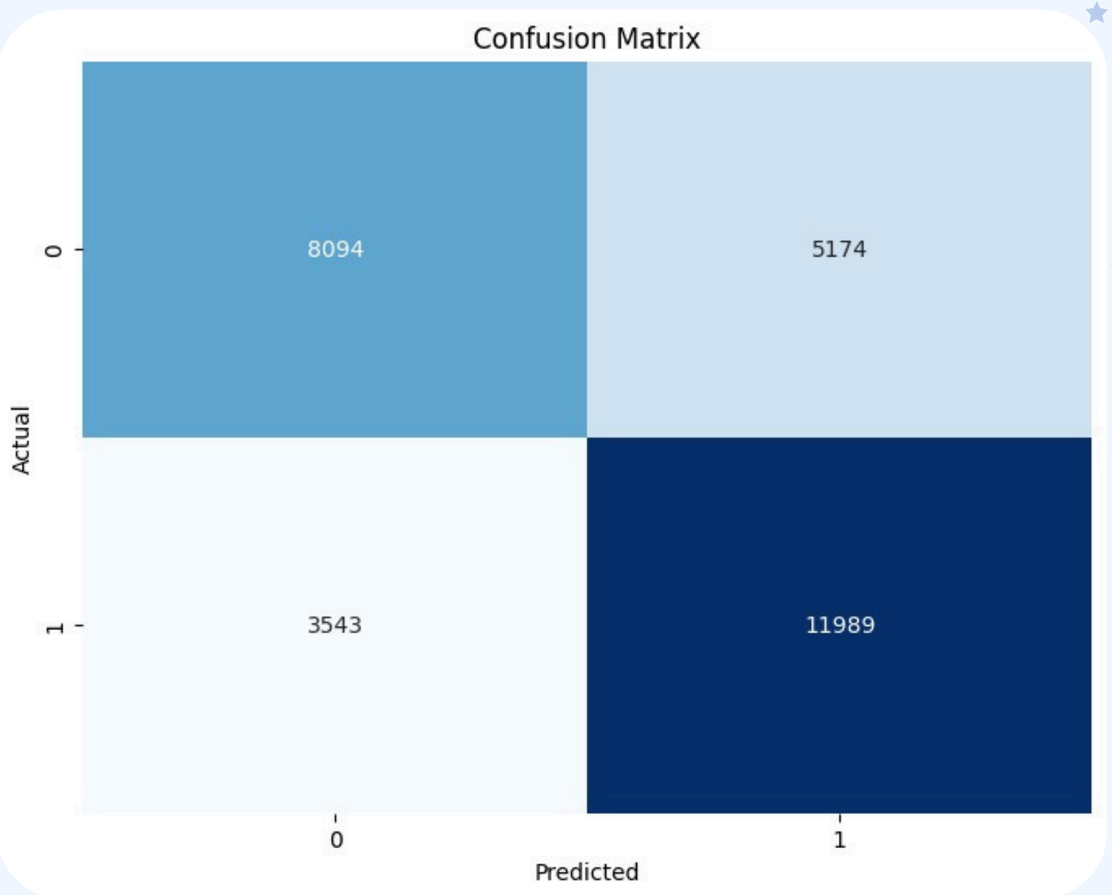
XGBClassifier

El objetivo es anticipar el resultado clínico y así poder actuar a tiempo en pacientes con alto riesgo.

```python
# Print classification report
from sklearn.metrics import classification_report

print(classification_report(y_true, y_pred))
```

[ ] Python

```
...                precision    recall  f1-score   support

         0.0       0.70      0.61      0.65     13268
         1.0       0.70      0.77      0.73     15532

    accuracy                           0.70     28800
   macro avg       0.70      0.69      0.69     28800
weighted avg       0.70      0.70      0.69     28800
```



Conclusión

Resumen de aprendizaje

La mayoría de los eventos adversos ocurren en los primeros días post-trasplante, destacando la importancia de este periodo. Identificamos subgrupos de riesgo y variables HLA altamente correlacionadas, lo que sugiere optimizar la selección de características. efs_time es clave en la predicción de supervivencia. Aplicamos XGBoost y clustering para segmentar pacientes y mejorar la toma de decisiones clínicas.

Pasos Futuros

Mejorar el desempeño de los modelos, mediante distintas técnicas, como *feature engineering*, *PCA*, etc. Ayudar al desarrollo de aplicaciones/páginas que permitan acercar este tipo de tecnología a las personas que más las necesitan.