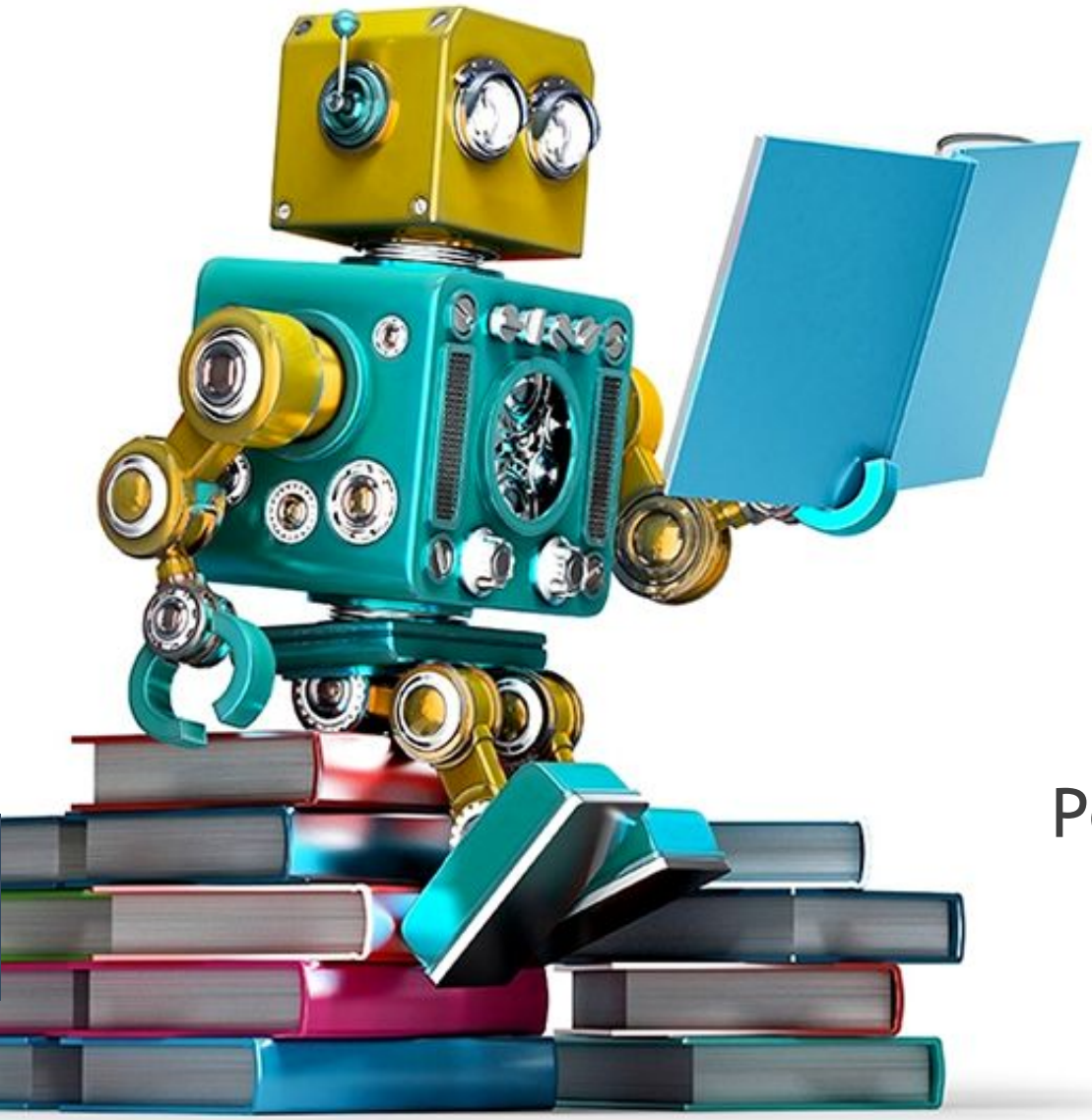


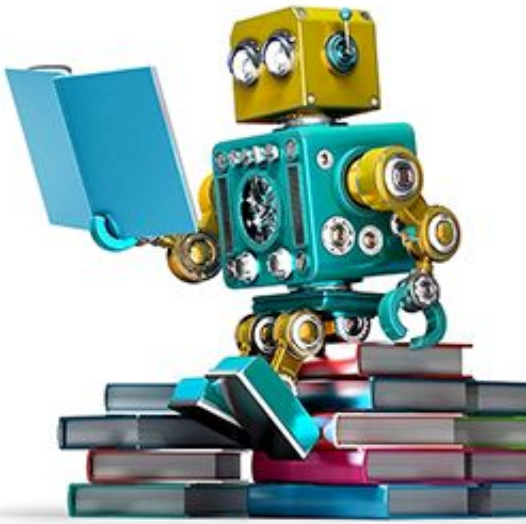
Proyecto ML

Clasificador de enfermedad del corazón

Por: Jaime León, Ivan Ávila & Rodrigo Ibarra

Agenda

1. Problema de Negocio
2. Dataset
3. Tipo de Modelo
4. Preprocesamiento de los datos
5. Optimización de Hiper Parámetros
6. Feature Engineering-Modelo Final
7. Metricas del modelo
8. Interfaz en Gradio
9. Conclusiones
10. Q&A



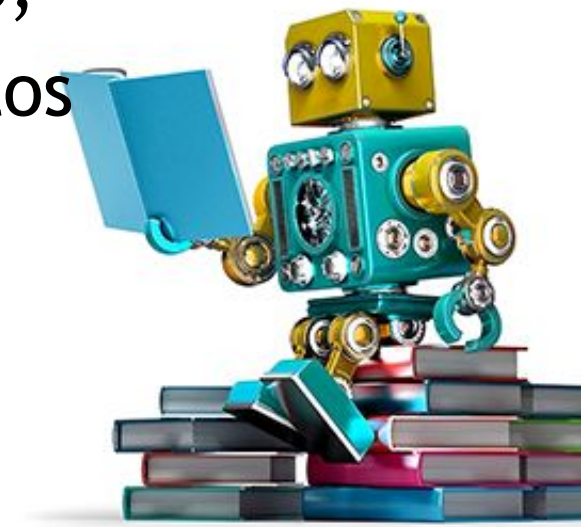
Problema de Negocio

Pharma Inc, una compañía farmacéutica desea saber qué medicamentos deben promocionar en una campaña de márketing basada en salud cardíaca.

Buscamos encontrar un modelo que nos diga qué variables son las que mejor predicen el padecimiento. Por ejemplo,

si la falta de sueño está relacionado con padecimientos cardiacos, entonces vamos a dirigir la campaña a

fármacos que ayuden a combatir el insomnio.



DiffWalking	Binario
Smoking	Binario
AlcoholDrinking	Binario
Stroke	Binario
Sex	Binario
Diabetic	Binario
Asthma	Binario
Physical Activity	Binario
KidneyDisease	Binario
SkinCancer	Binario
BMI	Numérico
PhysicalHealth	Numérico
MentalHealth	Numérico
SleepTime	Numérico
Age Category	Categoría
Gen Health	Categoría
Race	Categoría

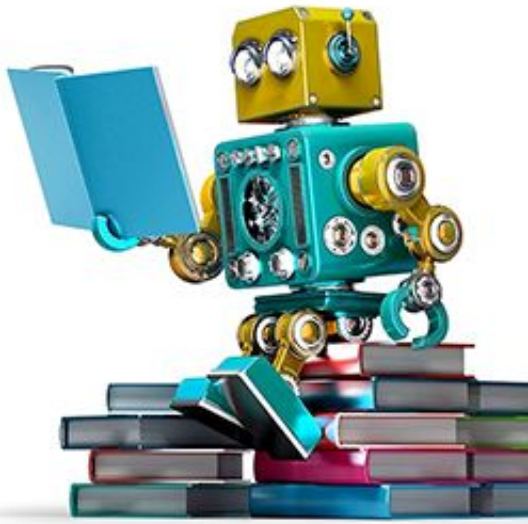
Dataset

El conjunto de datos proviene de los CDC(Centers for Disease Control and Prevention), que realizan encuestas telefónicas anuales para recopilar datos sobre el estado de salud de los residentes de EE. UU. 2020

Target
True: 27.4k
False: 292k

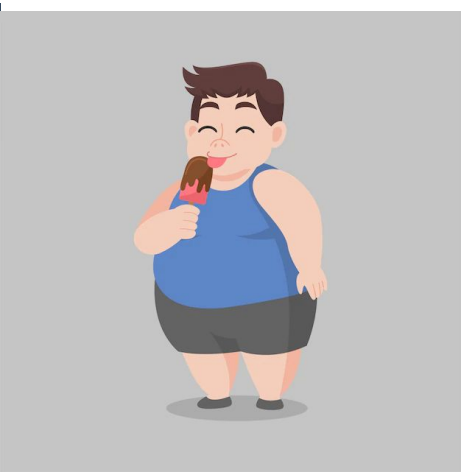
320,00 Records

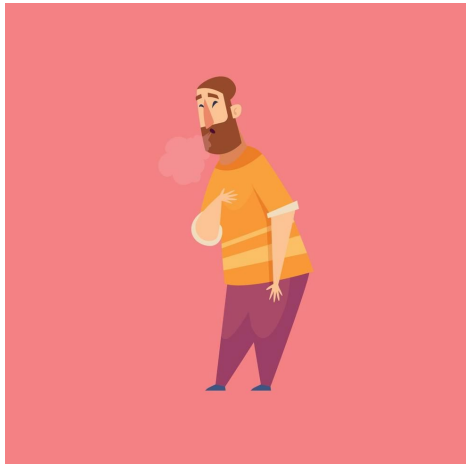
La base de datos cuenta con 17 columnas, las cuales 10 son de tipo booleano, 3 son tipo texto y 4 son tipo numérico.



Tipo de Modelo

- Supervisado
- Clasificación
- Binario
- Regresion Logistica vs Random Forest Classifier
- Velocidad / Preciso / Escalable / Interpretable
- Linealmente Separable / Tipos de Variables /Numercias Extrapolables / Multiclase

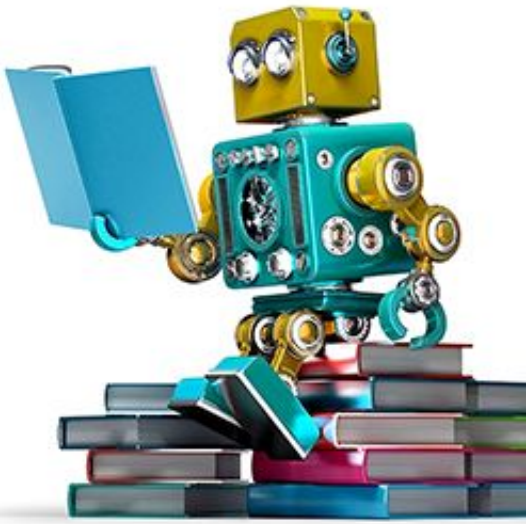




Preprocesamiento de los datos

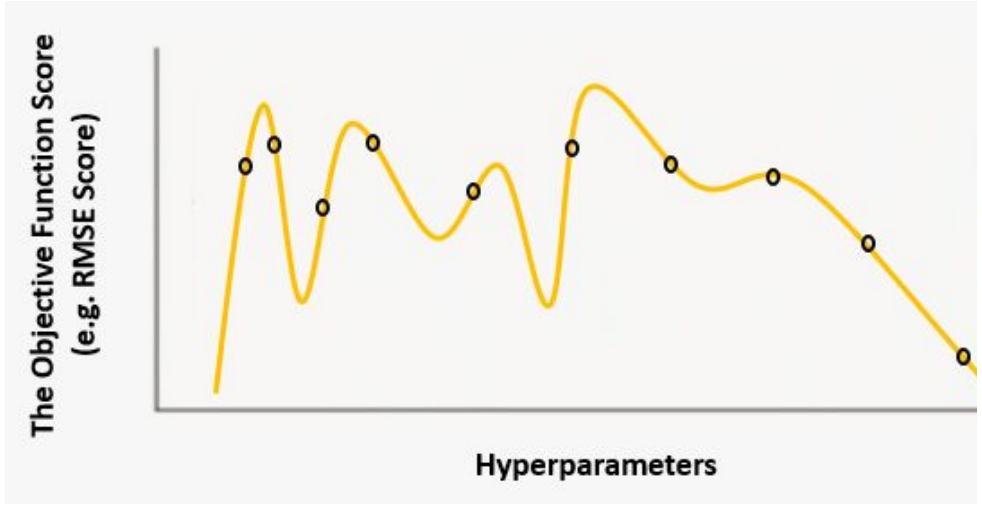
Como los datos están altamente desbalanceados:

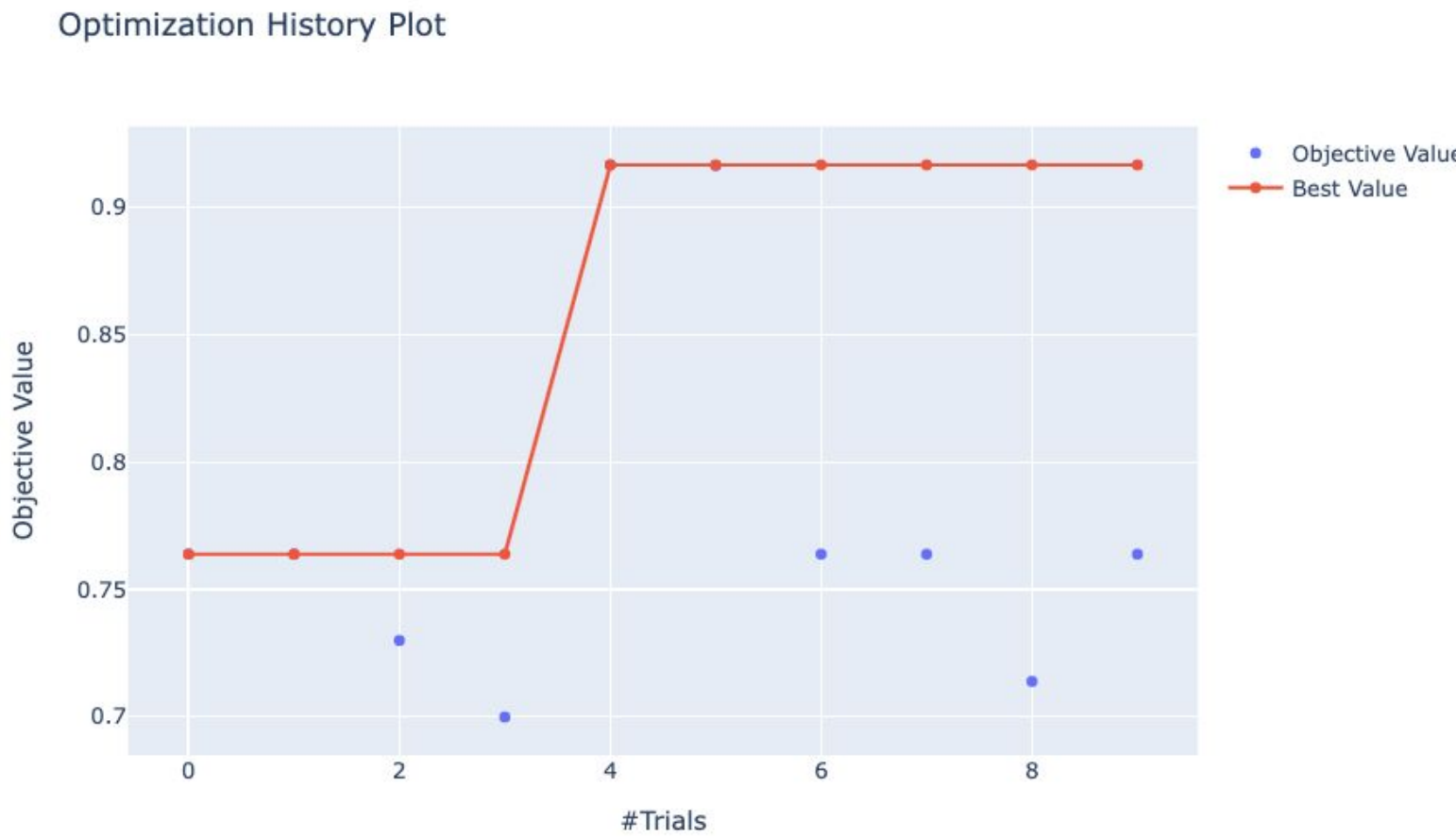
- Usamos over sampling: 0.7 - 0.3
- Luego under sampling: 0.5 - 0.5
- One Hot Encoding para variables categóricas
- Label Encoding para variables binarias
- Estandarizamos variables numéricas



Optimizacion de Hiperparametros

La optimización bayesiana construye un modelo de probabilidad de la función objetivo y lo usa para seleccionar hiper parámetros para evaluar en la verdadera función objetivo.





OPTUNA

Optimizacion de Hiperparametros

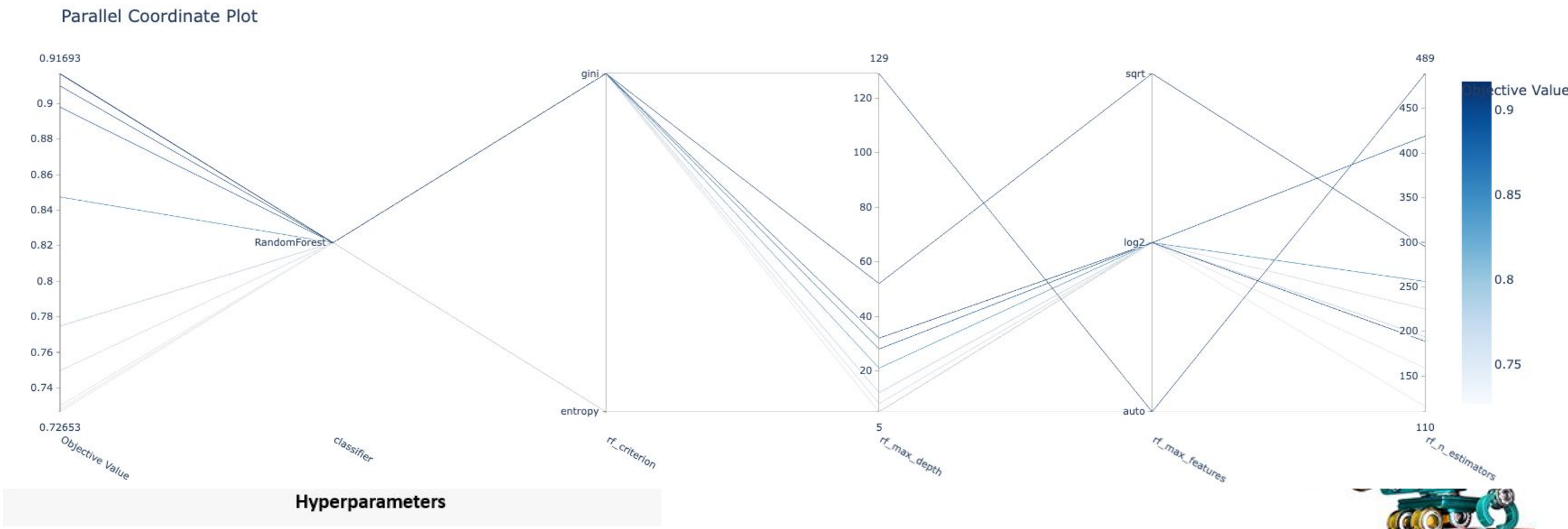
La optimización bayesiana construye un modelo de probabilidad de la función objetivo y lo usa para seleccionar hiper parámetros para evaluar en la verdadera función objetivo.

```
classifier_name = trial.suggest_categorical('classifier', ['RandomForest', 'LogisticRegression'])
if classifier_name == 'LogisticRegression':
  LogisticRegression_penalty = trial.suggest_categorical('LogisticRegression_penalty', ['l2', 'l1',])
  LogisticRegression_c = trial.suggest_float('LogisticRegression_c', 1e-4, 1e2, log=False)
  LogisticRegression_solver = trial.suggest_categorical('LogisticRegression_solver', [ 'liblinear','saga'])
  LogisticRegression_fit_intercept = trial.suggest_categorical('LogisticRegression_fit_intercept', [False, True])
  classifier_obj = sklearn.linear_model.LogisticRegression(
      penalty=LogisticRegression_penalty, C=LogisticRegression_c, solver=LogisticRegression_solver, fit_intercept =

elif classifier_name == 'RandomForest':
    rf_max_depth = trial.suggest_int('rf_max_depth', 2, 200, log=True)
    rf_n_estimators = trial.suggest_int('rf_n_estimators', 100, 500, log=True)
    rf_criterion = trial.suggest_categorical('rf_criterion', ['gini', 'entropy'])
    rf_max_features = trial.suggest_categorical('rf_max_features', ['auto', 'sqrt', 'log2'])
    classifier_obj = sklearn.ensemble.RandomForestClassifier(
        max_depth=rf_max_depth, n_estimators=rf_n_estimators, criterion=rf_criterion,
    max_features=rf_max_features)


score = sklearn.model_selection.cross_val_score(classifier_obj, x,
                                y, n_jobs=-1, cv=3
```

The Objective Function Score (e.g. RMSE Score)

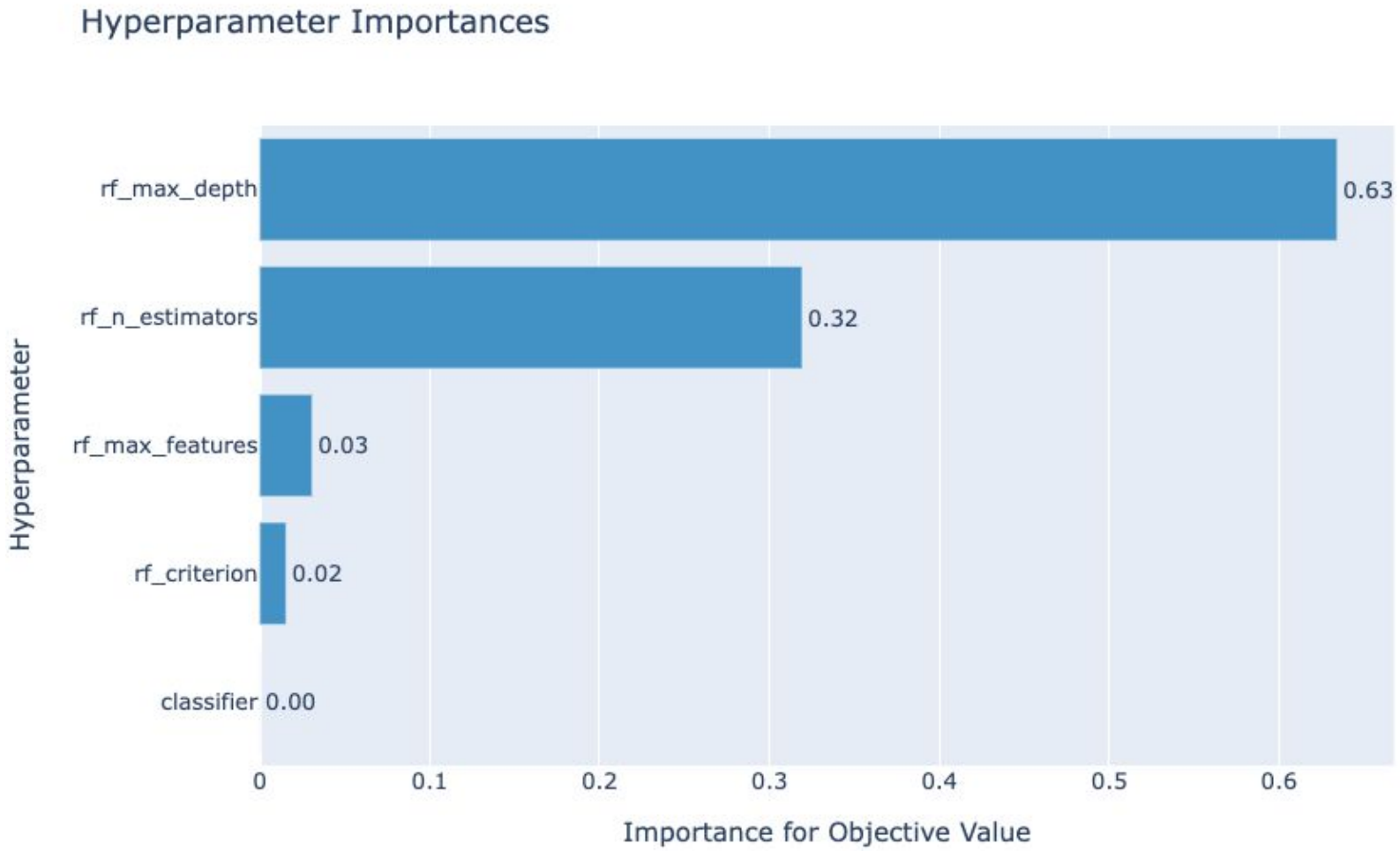
Objective Value
Best Value

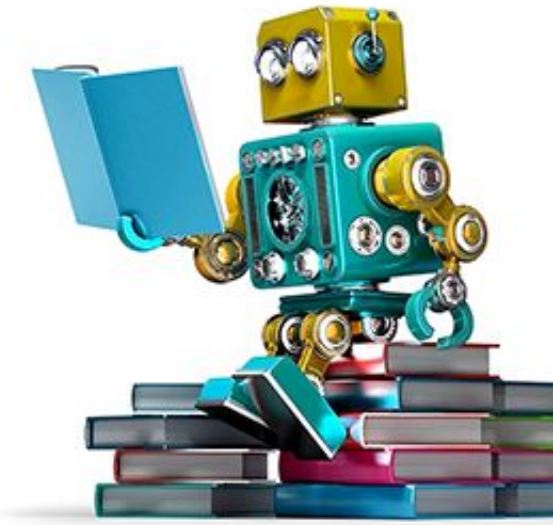
Optimizacion de Hiperparametros



OPTUNA

Optimizacion de Hiperparametros

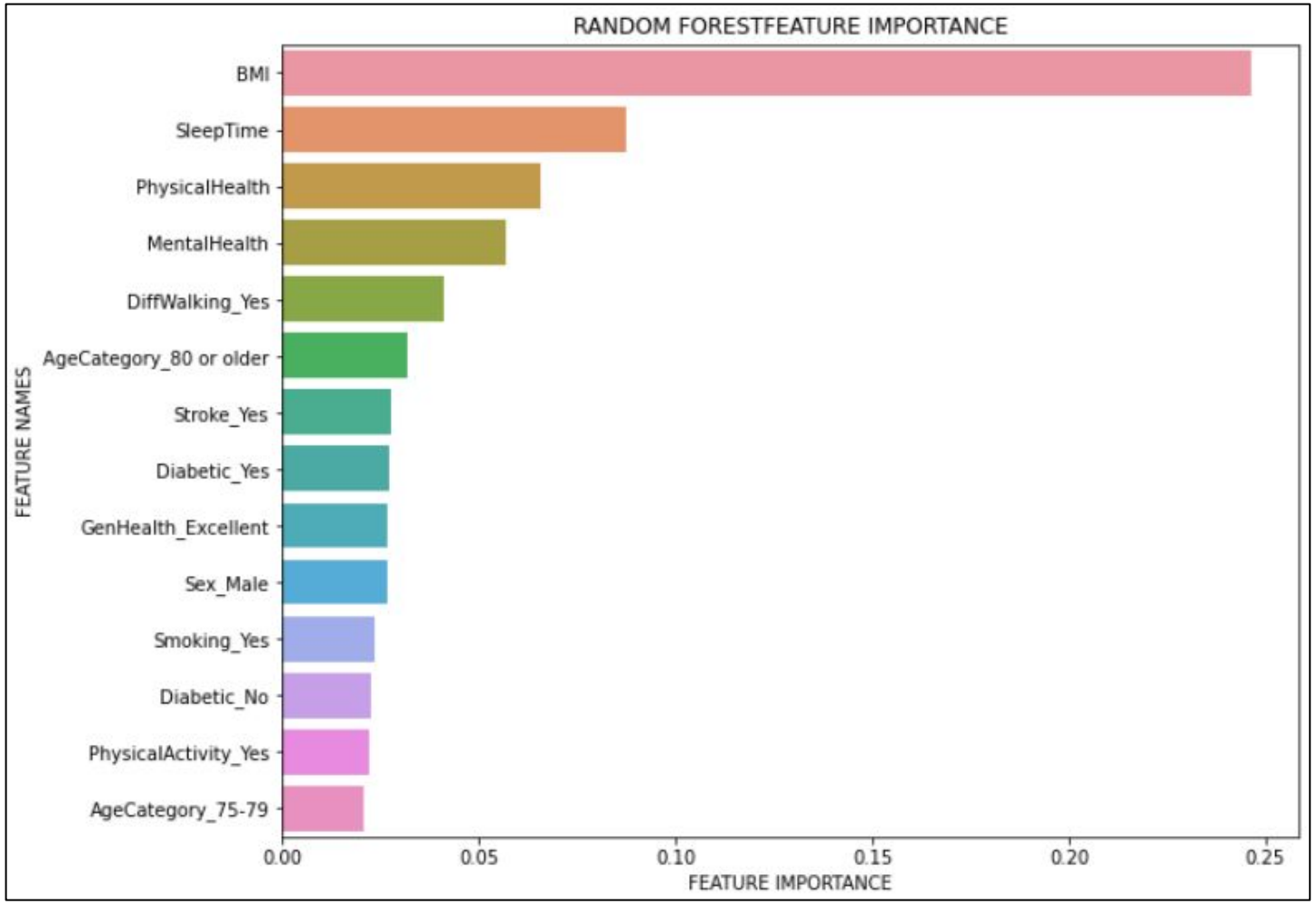




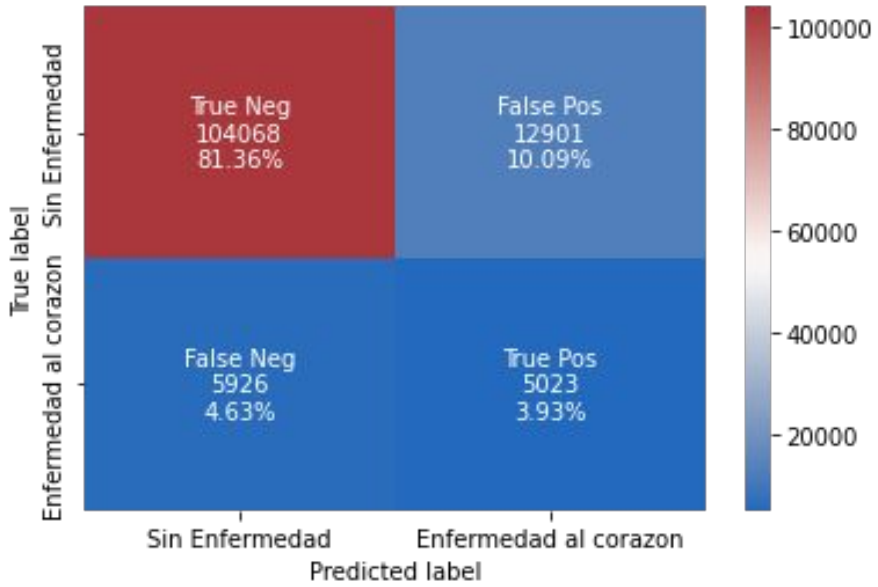
OPTUNA

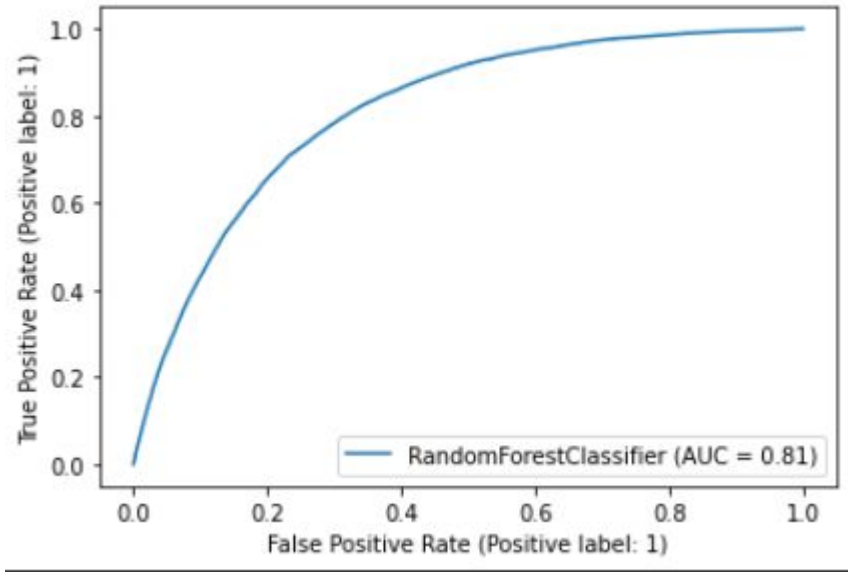
Feature Engineering/Modelo Final

```
Best trial:
  Value: 0.9165009808599756
  Params:
    classifier: RandomForest
    rf_max_depth: 106
    rf_n_estimators: 146
    rf_criterion: entropy
    rf_max_features: auto
```

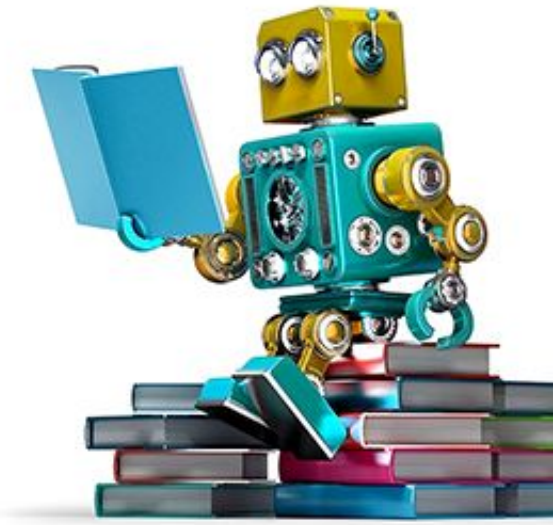


Métricas del Modelo





	precision	recall	f1-score	support
0	0.95	0.89	0.92	116969
1	0.28	0.46	0.35	10949
accuracy			0.85	127918
macro avg	0.61	0.67	0.63	127918
weighted avg	0.89	0.85	0.87	127918



Heart Disease Prediction Using a Random Forest Classifier

What´s your BMI? 1

Do you smoke?

Yes

Do you drink?

Yes

Have you had an stroke?

Yes

How many time have you thought about your physical health in the last month 0

How many time have you thought about your mental health in the last month 0

Do you have any difficulty for walking?

Yes

How dou you identify yourself?

Male

output

Avisar

output

Heart Disease Prediction: Yes

output

Heart Disease Prediction: No

Conclusiones

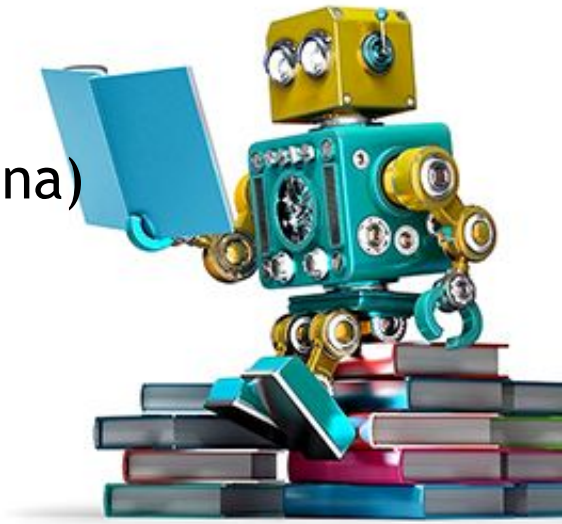
Medicamentos contra insomnio, a favor del control de peso o la salud mental como ansiolíticos son buenos candidatos para las campañas de márketing.

A pesar de lo mucho que estaba desbalanceado el dataset, se logró alcanzar un buen resultado gracias a las técnicas del preprocesamiento de datos y la optimización de los hiperparametros.

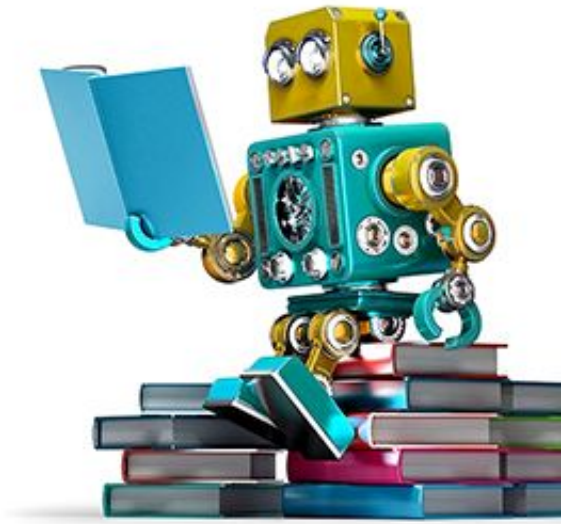
Aún así, con estos resultados, al tratarse del sector de salud se espera mejorar el modelo y llegar a un mejor resultado bajando el caso de Falsos Negativos.

Bajando el threshold y aumentando el número de features como:

- Alimentación de la persona (Cuántas veces come comida rapida/semana)
- Pasos a la semana/dia
- Litros de soda/semana
- Gramos de azúcar consumidos/semana
- Tiempo de uso del celular/semana
-



Q&A



Gracias :)


THANK YOU!
WALL·E
www.generalprints.com
Thank you for coming
to my birthday party
and for the great gift.
I had a great time
and I hope you did too!
Love,
Wilson