PRUEBA TÉCNICA CIELUM | ANALYTICS LEADER

CESAR RAMIREZ GOMEZ

INGENIERO ELECTRÓNICO, ESTUDIANTE MAESTRÍA EN INTELIGENCIA ARTIFICIAL

I. ENTENDIMIENTO DEL NEGOCIO

El virus de inmunodeficiencia humana "VIH" es una enfermedad que ataca el sistema inmunitario del cuerpo humano, permite el desarrollo de infecciones oportunistas y cánceres potencialmente mortales, cuando los niveles de linfocitos T CD4+ están por debajo de 200 por mililitro. La transmisión se da por sangre, semen, flujo vaginal, líquido preseminal y leche de lactancia. En países desarrollados transmuta en 10 años a síndrome de inmunodeficiencia adquirida (SIDA). Predecir que pacientes presentaran fracaso virológico permite a los sistemas de salud centrar esfuerzos en poblaciones con mayor riesgo de transmutación en una ventana de tiempo variable. Modelos de inteligencia artificial permiten realizar predicciones basados en datos de los usuarios por lo cual se plantea desarrollar modelos de IA para contribuir a su temprana detección.

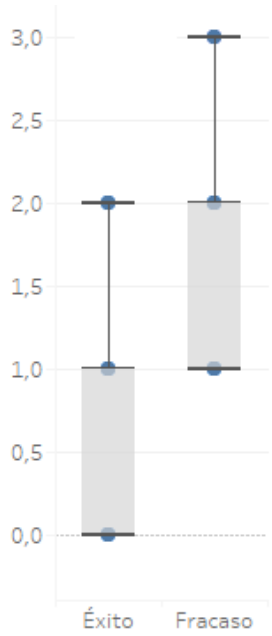
II. PREGUNTA A RESOLVER

¿Qué pacientes con VIH tendrán un desenlace fracaso virológico en los próximos seis meses?

III. VISUALIZACION DE VARIABLES

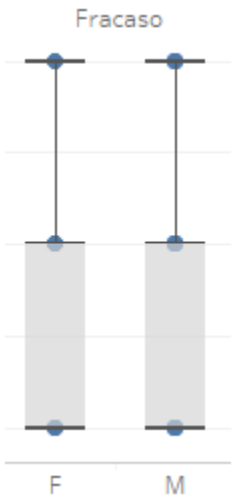
Se seleccionaron las variables con mayor correlación en la matriz de confusión desarrollada mas adelante para graficar e identificar hipótesis que ayuden a solucionar el propósito del modelo.

- Desenlace virológico vs número fracaso



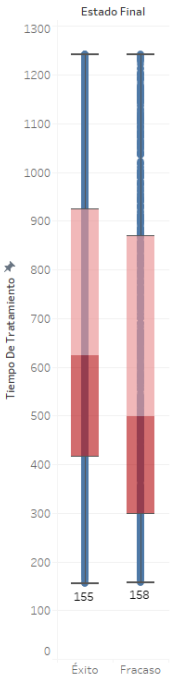
Grafica 1. Desenlace virológico vs número de fracasos en el tratamiento del paciente.

Los pacientes con éxito virológico están en un rango de numero de fracasos previos desde 0 hasta 2 fracasos con una media de 1 fracaso. Los pacientes con éxito virológico han tenido un promedio de 1 fracaso virológico en su tratamiento. Mientras que los pacientes fracaso virológico han tenido un promedio de 2 fracasos virológico en todo su tratamiento. Esta tendencia se presenta indiscriminadamente del género.



Grafica 2. Desenlace virológico vs número de fracasos en el tratamiento del paciente vs género.

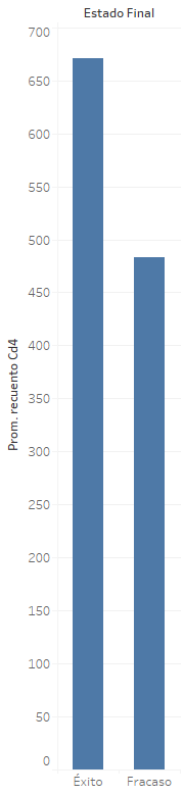
- Desenlace virológico vs tiempo en tratamiento



Grafica 3. Diagrama desenlace virológico vs tiempo en tratamiento.

El promedio de tiempo que los usuarios con desenlace virológico exitoso son de 626 días. En contraparte los usuarios que tiene desenlace virológico fracaso en promedio son de 500 días. Esta medida indica que es necesario en promedio un filtro de 500 días para detectar el desenlace virológico en vez de 180 días (6 meses aproximadamente).

- Desenlace virológico vs Recuento Cd4

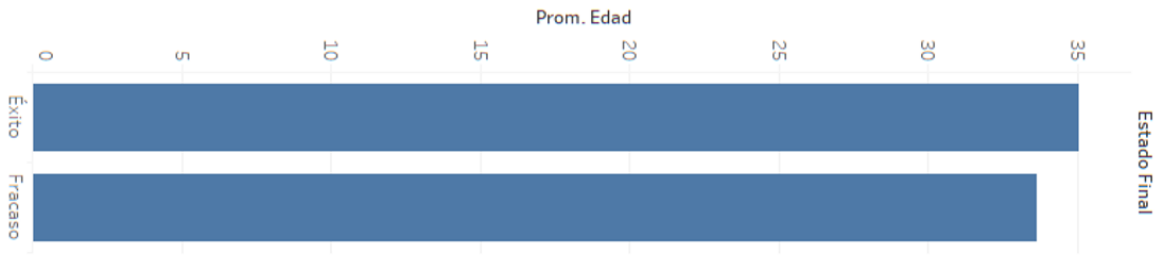


Grafica 4. Estado final vs promedio recuento Cd4

Los pacientes con fracaso virológico exitoso tienen un promedio de recuento Cd4 de 671; los pacientes con fracaso virológico Fracaso tiene un promedio de recuento Cd4 de 483,7.

- Fracaso virológico vs edad

El promedio de los usuarios que tiene fracaso virológico exitoso es de 35 años. Mientras que los usuarios con fracaso virológico tienen una edad promedio de 33 años. Se formula como hipótesis que la edad no es de relevancia en la predicción del fracaso virológico.



Grafica 5. Fracaso virológico vs promedio edad

IV. EXPLORACION DE DATOS

El siguiente cogido desarrollado se encuentra en el archivo “Exploracion_limpieza_datos.ipynb ”adjunto en la carpeta.

1. Análisis descriptivo - tipo de variable

   El conjunto datos está integrado por 55 variables, de las cuales 14 son de naturaleza numérica y 41 de tipo categórico. La variable objetivo se denomina “estado_final” de estructura categórica.

TIPO DE VARIABLE			
VARIABLE	**NUMERICO**	**CATEGORICO**	**NUMERO DE CLASES**
id	X		-
estado_final		X	2
ocupacion		X	7
regimen		X	2
acompanante		X	3
zona_cox		X	3
estrato_social	X		-
preferencia		X	3
estado_civil	X		-
escolaridad	X		-
edad	X		-
genero		X	2
esquema_arv		X	59
grupos_farmacologicos		X	5
numero_tomas_dia	X		-
numero_unidades_dia	X		-
persistencia	X		-
antecedentes_retrazo_autorizacion_cox		X	2
antecedentes_ram		X	2
antecedentes_omision		X	2

antecedentes_suspension		X	2
antecedentes_desafiliacion		X	2
antecedentes_no_adherencia		X	2
numero_de_no_adherencia	X		-
grado_maximo_de_no_adherencia		X	5
fecha_inicio_arv	Variable tiempo		-
tiempo_en_tratamiento_arv	X		-
recuento_Cd4	X		-
CT_categorico		X	2
Rango_CT		X	2
TG_categorico		X	2
Rango_TG		X	3
HDL_categorico		X	2
LDL_categorico		X	2
Rango_Riesgo_cardivascular		X	4
antecedente_fracaso		X	2
tiempo_a_1er_fracaso	X		-
numero_de_fracasos	X		-
tiempo_de_tratamiento	X		-
Aspergilosis		X	1
Micosis		X	2
Candidiasis		X	2
CMV		X	2
Enteritis		X	2
Herpes		X	2
Histoplasmosis		X	2
Sarcoma_Kaposi		X	2
Linfoma_carcinoma		X	2
Nocardiosis		X	1
Encefalitis		X	2
Neumonia		X	2
Toxoplasmosis		X	2
Tuberculosis		X	2
Hepatitis_viral		X	2
varicela		X	2

Tabla 1. Clasificación tipo de variable.

2. Exploración de datos

Se realiza un análisis estadístico del dataset teniendo en cuenta:

- Numero de clases por variable categóricas.
- Porcentaje perdida de datos.

- Filtrado de datos.
- Aproximación variable objetivo.
- Análisis inferencial multivariado - matriz de confusión
- Imputación de datos.

Numero de clases por variable categóricas

Las variables que están conformadas únicamente por 1 clase se eliminan del dataset al no proporcionar información, las cuales son:

- Aspergilosis
- Nocardiosis

Porcentaje perdida de datos

Se calcula el porcentaje perdida de datos seleccionando las variables a imputar o eliminar del dataset.

VARIABLE	PERDIDA DE DATOS	VARIABLE	PERDIAD DE DATOS
Ocupacion	0%	**Rango CT**	**14.2%**
Regimen	0%	**TG_categorico**	**20.9%**
Acompanante	0%	**Rango_TG**	**20.9%**
Zona_cox	0%	HDL categorico	16.6%
Estrato_social	0.3%	LDL categorico	**43.3%**
Preferencia	**22.1%**	Rango Riesgo cardivascular	0%
Estado_civil	0.3%	Antecedente fracaso	0%
Escolaridad	0.3%	Tiempo a 1er fracaso	0%
Edad	0.3%	Numero de fracasos	0%
Genero	0.3%	Tiempo de tratamiento	0%
Esquema arv	0.2%	Aspergilosis	0%
Grupos farmacológicos	0.2%	Micosis	0%
Numero tomas dia	0.2%	Candidiasis	0%
Numero unidades dia	0.2%	CMV	0%
Persistencia	0%	Enteritis	0%

Antecedentes retrazo autorización cox	0%		Herpes	0%
Antecedentes ram	0%		Histoplasmosis	0%
Antecedentes omision	0%		Sarcoma_Kaposi	0%
Antecedentes suspensión	0%		Linfoma_carcinoma	0%
Antecedentes desafiliación	0%		Nocardiosis	0%
Antecedentes no adherencia	0%		Encefalitis	0%
Numero de no adherencia	0%		Neumonia	0%
Grado máximo de no adherecia	0%		Toxoplasmosis	0%
Fecha de inicio arv	0%		Tuberculosis	0%
Tiempo en tratamiento arv	0%		Hepatitis_viral	0%
Recuento Cd4	6.8%		Varicela	0%
CT_categorico	**14.2%**			

Tabla 2. Perdida de datos de las variables.

En la tabla 2 se evidencia un alto porcentaje de datos perdidos en las variables:

- Preferencia: 22.1%
- Rango_CT: 14.2%
- TG_categorico: 20.9%
- Rango_TG: 20.9%
- LDL_categorico: 43.3%
- CT_categorico: 14.2%

Estadísticamente las variables con una perdida de datos inferior al 25% es posible imputarle datos con técnicas avanzadas como generación sintética de datos sin afectar su distribución. La variable LDL_categorico obtiene un 43.3% de perdida acercándose al 50%, por lo que imputar datos generaría una desviación de los datos afectando el desempeño del modelo, se elimina del dataset.

Filtrado de datos

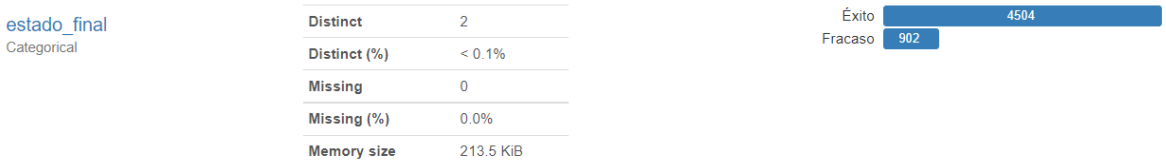
Las variables CT_categorico y Rango_CT presentan los mismos valores únicamente cambiando su estructura lingüística además de obtener el mismo porcentaje de perdida de datos 14.2% por lo que se elimina la variable Rango_CT al presentar un menor valor en la matriz de correlación.

Las variables TG_categorico y Rango_TG presentan los mismos valores únicamente cambiando su estructura lingüística además de obtener el mismo porcentaje de perdida de datos 20.9% por lo que se elimina la variable Rango_TG al presentar un menor valor en la matriz de correlación.

Se eliminan la variable "tiempo_de_tratamiento" la cual presenta los mismos datos que la variable "tiempo_en_tratamiento_arv".

Aproximación variable objetivo

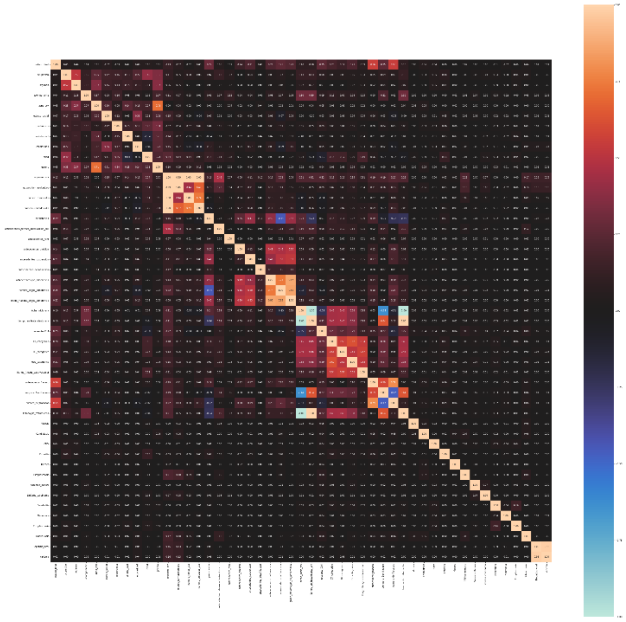
La variable objetivo inicial sin filtros temporales (> 6 meses) cuenta con 4504 registros de Éxito (83.3%) y 902 registros de fracaso (16.7%). Se evidencia un desbalanceo de clases por lo que se implementara técnicas de balanceo sintético de datos. No se encuentra errores gramaticales como combinación de mayúsculas o minúsculas.



Grafica 6. Distribución estadística para la variable Estado final.

Análisis inferencial multivariado - Matriz de correlación

Con el conjunto de datos filtrado se aplica una matriz de correlación para determinar las variables mas relevantes con la variable objetivo “estado final”. El conjunto de datos se encuentran variables categorías y numéricas; se empleó la librería Nominal-dython que permite encontrar la correlación entre variables tipo numéricas y/o categóricas. Los valores inferiores a ±0.00 o ±0.09 representan una correlación nula por lo que se descartan del conjunto de datos a implementar en el modelo de inteligencia artificial.



Grafica 7. Matriz de correlación.

Centrándose en la variable objetivo se simplifican los resultados en la siguiente tabla:

VARIABLE	VALOR CORRELACION
estado_final	1
ocupacion	0,05
regimen	0,04
acompanante	0,04
zona_cox	0,05
estrato_social	0,07
preferencia	0,03
estado_civil	0,02
escolaridad	0,06
edad	0,06
genero	0,05
esquema_arv	0,13
grupos_farmacologicos	0,07
numero_tomas_dia	0,02
numero_unidades_dia	0,05
persistencia	0,29
antecedentes_retrazo_autorizacion_cox	0,06
antecedentes_ram	0,01
antecedentes_omision	0,14
antecedentes_suspension	0,21
antecedentes_desafiliacion	0,05
antecedentes_no_adherencia	0,21
numero_de_no_adherencia	0,25
grado_maximo_de_no_adherencia	0,22
fecha_inicio_arv	0,1
tiempo_en_tratamiento_arv	0,1
recuento_Cd4	0,25
CT_categorico	0,2
TG_categorico	0,14
HDL_categorico	0,21
Rango_Riesgo_cardivascular	0,09
antecedente_fracaso	0,56
tiempo_a_1er_fracaso	0,25
numero_de_fracasos	0,55
Micosis	0,03
Candidiasis	0,06
CMV	0,04
Enteritis	0

Herpes	0
Histoplasmosis	0,02
Sarcoma_Kaposi	0
Linfoma_carcinoma	0
Encefalitis	0,01
Neumonia	0,02
Toxoplasmosis	0
Tuberculosis	0,04
Hepatitis_viral	0,01
varicela	0,01

Tabla 3. Resumen matriz correlación con variable objetivo.

Las variables seleccionadas para conformar el dataset para el entrenamiento y prueba del modelo de IA se resaltaron en amarrillo por tener un nivel de correlación con la variable objetivo, las demás se eliminan. Se elimina la variable "Fecha de inicio arv" al no representa un valor alto de correlación y no presenta una ventana de tiempo variable.

Imputación de datos

A las variables de la matriz de confusión seleccionadas y que presentan perdida de datos se realizara procesos de imputación de datos que no desequilibren las medidas de tendencia central en cada variable. A variables categóricas se aplicará la moda en los datos faltantes y en variables numéricas se reemplazarán por la media aritmética.

- Esquema ARV: La variable tiene una perdida 0.2% de los datos, se calcula la moda contando los elementos en cada clase. La mayor clase es de "Abacavir/Lamivudina+Efavirenz" con 12988 filas. Se reemplaza los valores NaN con la clase de mayor frecuencia.
- Recuento_Cd4: La variable tiene una perdida 6.8% de los datos, se calcula la media para reemplazar los valores vacíos.
- CT_categorico: La variable tiene una perdida 14.2% de los datos, se calcula la moda contando los elementos en cada clase. Se reemplaza los valores NaN con la clase de mayor frecuencia.
- TG_categorico: La variable tiene una perdida 20.9% de los datos, se calcula la moda contando los elementos en cada clase. Se reemplaza los valores NaN con la clase de mayor frecuencia.
- HDL_categorico: La variable tiene una perdida 16.6% de los datos, se calcula la moda contando los elementos en cada clase. Se reemplaza los valores NaN con la clase de mayor frecuencia.

Generación variable objetivo

La IPS requiere predecir con 6 meses de anticipación el desenlace "fracaso virológico" por lo cual se realiza un filtro a los datos permitiendo definir la clase objetivo como:

- Clase objetivo = “1”

Pacientes menor o igual a seis meses en tratamiento

Paciente con estado final Fracaso

- Clase no objetivo = 0

  Pacientes con mayor a seis meses en tratamiento

  Pacientes con estado final Exitoso

La variable objetivo genera 22 registros mientras que la variable no objetivo obtiene 4484 registros. Se presenta un claro desbalanceo de datos por lo cual se aplicarán técnicas de Underfitting y/o Overfitting.

ONE HOT ENCODING

Se requiere convertir las variables categóricas a formato binario para que los modelos de inteligencia artificial puedan emplean sus registros. Se crea por cada variable y clase una nueva columna indicando su pertenencia con un "1" y su ausencia con un "0".

V. GENERACION DE MODELOS IA PARA LA PREDICCION DE VARIABLE FRACASO VIROLOGICO

El siguiente desarrollo se encuentra en el archivo "Modelos_IA.ipynb" adjunto en la carpeta.

Se divide el dataset en conjunto de prueba (80%) y test (20%) con una semilla aleatoria fija para poder ser reproducible los resultados. Se estandarizan las variables numéricas para no generar un sesgo en el entrenamiento. Para el correcto desarrollo de modelos de IA se requiere un conjunto de datos balanceado para evitar sesgos en la predicción del resultado. Se evidencia un marcado desbalanceo en la clase objetivo "Éxito = 0,488237905%" y "Fracaso = 99,51176209%". Se combinan técnicas de Overfitting y Underfitting. Como primer paso se selecciona de manera aleatoria 500 registros de la clase mayoritaria "fracaso". Empleando la librería SMOTE se generan datos sintéticos que igualen la clase mayoritaria y minoritaria.

- Antes de balanceo de datos:
  Variable objetivo: 13
  Variable no objetivo: 404
- Después de balanceo de datos:
  Variable objetivo: 404
  Variable no objetivo: 387

<u>Random Forest</u>

Se emplea el modelo Random Forest basada en árboles de decisión. Su principal ventaja se centra en un óptimo rendimiento de generalización para un rendimiento en entrenamiento. Se busca los mejores hiperparametros realizando un barrido en los parámetros empleando el criterio de Gini y Entropia.

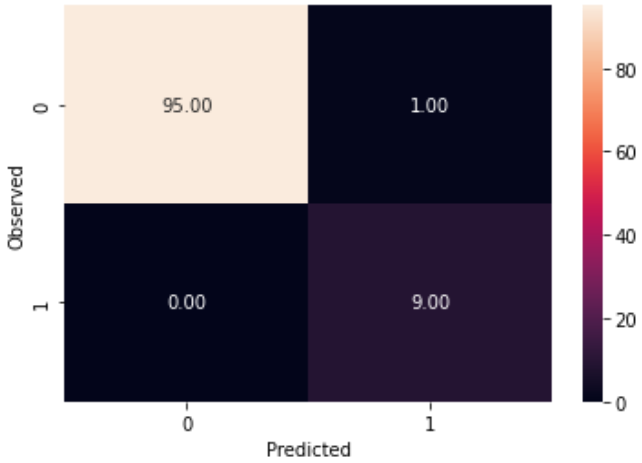
- n_estimators: 150

- max_features: 5, 7, 9
- max_depth: 0, 3, 10, 20

	oob_accuracy	criterion	max_depth	max_features	n_estimators
0	1.0	gini	NaN	5	150
11	1.0	gini	20.0	9	150
22	1.0	entropy	20.0	7	150
21	1.0	entropy	20.0	5	150

Grafica 8. Resultados sintonización Arboles de decisión.

El mejor resultado se da con 5 características y 0 poda de hojas.



Grafica 9. Matriz de decisión arboles de decisión.

- Accuracy entrenamiento: 1.0
- Accuracy testeo: 0.9904761904761905
- Recall: 1.0
- Precisión: 0.9
- F1 Score: 0.9473684210526316
- Roc Auc Score: 0.9947916666666667

La métrica Acurracy en entrenamiento y prueba presenta un buen desempeño acompañado de la métrica F1 score. La matriz de decisión presenta un sesgo en la predicción de la variable no objetivo con 95 casos mientras que la variable objetivo presenta 9 casos. Se da como hipótesis un desbalanceo en las clases por la no correcta generación sintética de clases en la variable minoritaria en este caso objetivo

REGRESION LINEA

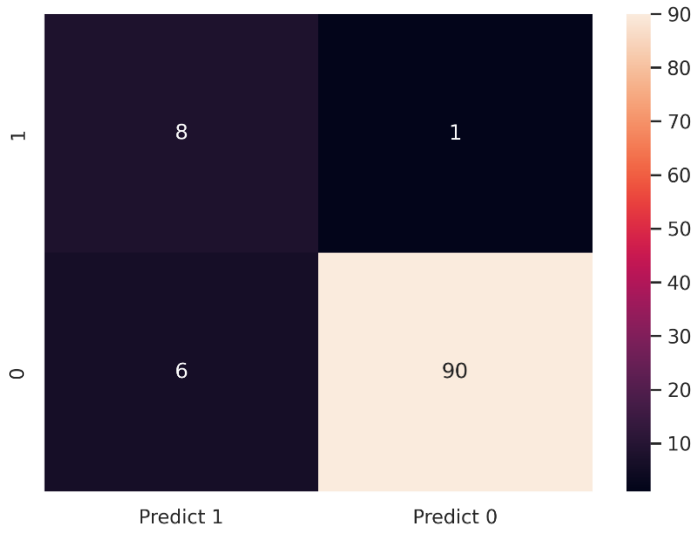
Se emplea una función que busca los mejores parámetros para una regresión lineal con los siguientes parámetros:

- solvers = ['newton-cg', 'lbfgs', 'liblinear']
- penalty = ['l2']
- c_values = [100, 10, 1.0, 0.1, 0.01]

Los resultados óptimos son: 'C': 100, 'penalty': 'l2', 'solver': 'newton-cg.

	precision	**recall**	**f1-score**
0	0,99	0,94	0,96
1	0,57	0,89	0,7

Tabla 4. Métricas regresión lineal.



Grafica 10. Matriz de confusión regresión lineal.

El modelo presenta un 57% de precisión, un recall del 89% y f1-score del 70% en la variable objetivo, son valores aceptables. La matriz de correlación presenta un sesgo a la variable no objetivo con 90 casos vs 8 casos de verdaderos positivos.

AdaBoosting

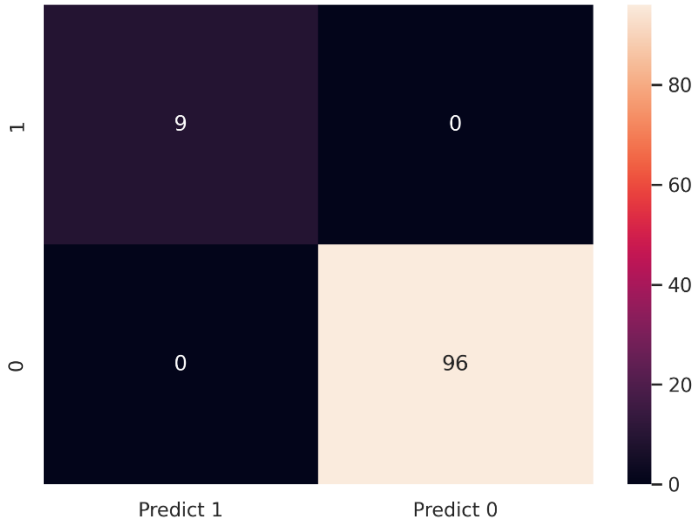
Se estructura un modelo para la sintonización de los hiperparametros:

- n_estimators: 10, 50, 100, 500
- learning_rate: 0.0001, 0.001, 0.01, 0.1, 1.0

Los hiperparametros con mejor resultado fueron learning_rate: 0.0001 y n_estimators: 10

	precisión	**recall**	**f1-score**
0	1	1	1
1	1	1	1

Tabla 5. Métricas AdaBoosting



Grafica 11. Matriz de confusión AdaBoosting.

Las métricas tienen una correlación perfecta evidencia de un desbalanceo de clases, se descarta el modelo.

XGBoost

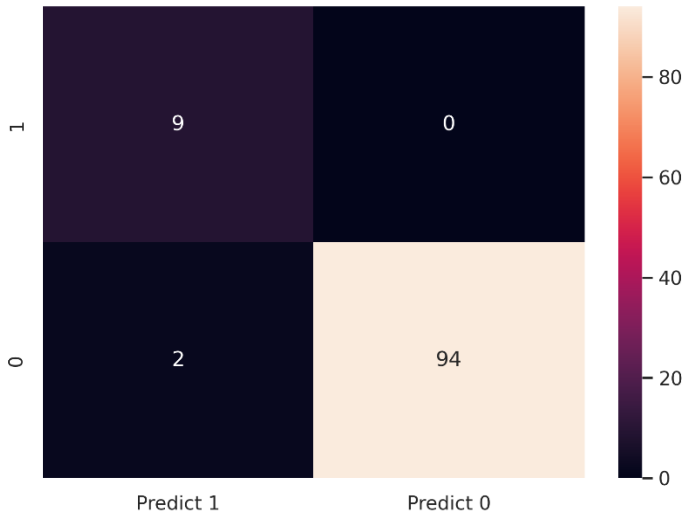
Para este modelo se sintoniza los hiperparametros en los siguientes intervalos

- learning_rate: 0.05, 0.10, 0.15, 0.20, 0.25, 0.30
- max_depth: 3, 4, 5, 6, 8, 10, 12, 15
- min_child_weight: 1, 3, 5, 7
- gamma: 0.0, 0.1, 0.2, 0.3, 0.4
- colsample_bytree: 0.3, 0.4, 0.5, 0.7

Los hiperparametros con mejor puntaje son: {'min_child_weight': 5, 'max_depth': 4, 'learning_rate': 0.25, 'gamma': 0.2, 'colsample_bytree': 0.7}

	precision	**recall**	**f1-score**
0	1	0,98	0,99
1	0,82	1	0,9

Tabla 6. Métricas XGBoost



Grafica 12. Matriz de confusión XGBoost

El modelo presenta un optimo desempeño en los parámetros de precisión, recall y F1-scrore. La matriz de confusión evidencia un desbalanceo de clases

VI. CONCLUSION DESARROLLO DE MODELOS

Los modelos empleados un buen desempeño en las métricas de precisión, recall y F1-score resaltando el modelo XGBoost. Por otra parte, las matrices de confusión evidencian un sesgo en la variable no objetivo “0”. Esto se debe a la incorrecta creación sintética de la variable minoritaria. Para solucionar este problema se propone adquirir mas datos de pacientes que su estado virológico sea fracaso y su tiempo de tratamiento sea menor o igual a 6 meses. Se propone aumentar el tiempo de predicción de 6 meses a 12 meses, el cual es cerca al promedio que tienen los pacientes con fracaso virológico en desarrollar su tratamiento.