LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Variables de incidencia en la mortalidad de pacientes con tuberculosis en el departamento del Tolima

Variables of incidence in the mortality of patients with tuberculosis in the department of Tolima

Wilson Andrés Gómez Gutiérrez, wagomezg02@libertadores.edu.co, Fundación Universitaria Los Libertadores.

José John Fredy González, jjgonzalezv02@libertadores.edu.co, Fundación Universitaria Los Libertadores

RESUMEN

La tuberculosis (TB) es la décimo tercera causa de muerte en el mundo, en Colombia el país ha adoptado la “Estrategia Mundial denominada Fin a la TB 2016-2035” a fin de mitigar el contagio y muertes por la enfermedad. (abecé-tuberculosis. Minsalud).

En el análisis de la TB se han definido variables tales como la edad, sexo, las condiciones de salubridad y residencia del paciente. Variables que están asociadas al desarrollo más temprano de esta enfermedad, no obstante, no se ha confirmado que necesariamente sean estas quienes determinen una condición mortal en el paciente.

Es por ello que a través de un modelo machine learning se determinan las características más importantes que están relacionadas con la evolución de la enfermedad TB y caracterizar los perfiles de pacientes con TB, de acuerdo a la información de las bases de datos de la plataforma de notificación de eventos en salud pública Sivigila y así poder estimar el porcentaje de mortalidad que puede llegar a tener un paciente de TB.

Realizando la implementación se pudo mejorar el modelo base del modelo basado en reglas siendo el Quadratic Discriminant Analysis el mejor por sus métricas las cuales no son muy buenas pero tienen una tendencia de superar el modelo base.

Palabras clave: Tuberculosis, variables, salud pública, modelo de aprendizaje.

ABSTRACT
Tuberculosis (TB) is the thirteenth cause of death in the world, in Colombia the country has adopted the "Global Strategy called End TB 2016-2035" in order to mitigate the contagion and deaths from the disease. (ABCs-tuberculosis. Minsalud). In the analysis of TB, variables such as age, sex, health conditions and residence of the patient have been defined. Variables that are associated with the earliest development of this disease, however, it has not been confirmed that these are necessarily the ones who determine a fatal condition in the patient.
That is why through a machine learning model , the most important characteristics that are related to the evolution of TB disease will be determined and characterize the profiles of patients with TB, according to the information in the databases of the public health event notification platform Sivigila and thus be able to estimate the percentage of mortality that a TB patient can have.

Keywords: Tuberculosis, variables, public health, learning model, deaths.

INTRODUCCIÓN

La tuberculosis, actualmente es considerada una de las principales enfermedades causante de la mortalidad de millones de personas. Según el Informe Mundial de Tuberculosis 2019 de la Organización Mundial de la Salud (OMS), alrededor de 1700 millones de personas infectadas con Mycobacterium tuberculosis desarrollarán tuberculosis durante su vida.

Se han estudiado variables relacionadas con las condiciones del ambiente en el que viven los pacientes con TB que inciden en el desarrollo de la enfermedad y posiblemente incrementan las posibilidades de muerte en los pacientes, según la investigación realizada por el Dr Walter H. Curioso y Maria J, Brunette, publicada en la revista Scielo Perú, además se estudia la variable correspondiente a los ingresos económicos del paciente que le facilitan el acceso al servicio de salud, compra de medicamentos y el tratamiento pronto a la enfermedad.

Los pacientes de TB deben de adoptar medidas preventivas frente a su autocuidado e higiénicas para prevenir la propagación y aceleración de la enfermedad. En el desarrollo de tesis enfocada en “Conocimientos de las Medidas Preventivas y Actitudes en el Autocuidado de pacientes con Tuberculosis Pulmonar en el Centro de Salud Los Libertadores en Noviembre – Diciembre 2008”. Se concluye que la variable de ingreso en los pacientes con TB 83.75% (34) muestra a la que se aplicó el instrumento, tienen conocimientos sobre los cuidados en el hogar, pero si no cuentan con una vivienda adecuada y medios económicos necesarios no podrán cuidar su salud.

En el desarrollo del presente trabajo de investigación titulado “Variables de incidencia en la mortalidad de pacientes con tuberculosis en el Tolima” se busca identificar los factores individuales y del entorno que pueden pronosticar la mortalidad de un paciente con TB con la ayuda de modelos de aprendizaje

automático supervisado. Para tal fin, se realizará el análisis de estudio con los datos privados de las bases cerradas del año 2019, 2020 y 2021, proporcionados por SIVIGILA el Sistema de Vigilancia en Salud Pública.

La investigación resultará ser un instrumento de bastante importancia en el sector de la salud, porque permitirá caracterizar al paciente de acuerdo con las variables y esto servirá para establecer estrategias de detección e intervención temprana según el perfil del paciente.

METODOLOGÍA

El presente estudio se basó en los datos privados de las bases cerradas del año 2019, 2020 y 2021, proporcionados por SIVIGILA el Sistema de Vigilancia en Salud Pública. Solicitados al coordinador de SIVIGILA y estadísticas vitales del Tolima. (ver datos)

Sivigila es la plataforma estatal de notificaciones de eventos en salud pública que se encarga de recopilar toda la información de eventos en salud pública de las entidades de salud prestadoras del servicio; siendo estas las encargadas de alimentar todas bases y así tener consolidado los casos que se presenten para realizar posteriormente el control epidemiológico llegase al caso y poder establecer planes de contingencia y demás protocolos que se requieran para el manejo de los eventos.

Estas bases son descargadas de manera anual el primer trimestre del año siguiente a la construcción de la misma y correspondiente a los 3 años mencionados se contó con 1.378 registros de personas con TB.

Las fases para llevar a cabo el presente estudio fueron:

- limpieza de la base de datos y selección de los registros relevantes
- realización de análisis descriptivos
- identificación de posibles modelos y la respectiva evaluación de su desempeño.

En el análisis descriptivo de la base de tuberculosis (TB) encontramos, n= 1372 con 24 variables. Las variables están distribuidas en variables dicotómicas y categóricas.

Al realizar un análisis descriptivo de los datos encontramos que la variable edad no cuenta con una normalidad; por ello se realiza una transformación aplicando logaritmo base 10 creando una nueva variable edad_log10 en la tabla df2 creada a través de python.

La variable estrato presenta una ausencia del 8% de sus datos respecto al total de registros con pacientes de TB, pero según análisis no es representativo para continuar con esa variable que explica un comportamiento del estrato respecto a la mortalidad por TB que es la variable objetivo.

Para este estudio las variables categóricas cuentan con un valor cuando 1 es afirmativo al significado de la variable. A continuación, se exponen las variables, sus significados y transformaciones realizadas.

Tabla 1. Transformación de algunas de las variables previa al análisis

Variable	***Descripción y unidades***
año	año de notificación del paciente
edad	edad de la persona (Se aplica Log 10)
*uni_med_*	Unidad de medida de la edad
*sexo_*	sexo de la persona
*area_*	Área de residencia. 1: Cabecera Municipal 2: Centro poblado 3: Rural disperso
*estrato_*	Estrato de la persona
*pac_hos_*	paciente hospitalizado. 1: SÍ 0: No
*con_fin_*	condición final del paciente 0: vivo 1: Muerto
con_tuber	condición de la tuberculosis 1: Sensible 0: Resistente
tip_tub	Tipo de tuberculosis 1: Pulmonar 0: Extrapulmonar
clas_ant	Clasificación anterior 1: Nuevo 0: Previamente tratado
vih_confirmado	Confirmación VIH 1: SI 0:NO

peso _act	Peso Actual persona en Kg
talla_act	Talla actual persona en Metros
imc	Índice de masa corporal persona
diabetes	diabetes 1: SI 0:NO
silicosis	Silicosis 1: SI 0: NO
enfe_renal	Enfermedad renal 1: SI 0: NO
epoc	Epoc 1: SI 0: NO
enfe_hepat	Enfermedad hepática 1: SI 0: NO
cancer	Cáncer 1: SI 0: NO
astritis	Artritis 1: SI 0: NO
desnutrici	desnutrición 1: SI 0: NO
nmun_proce	Municipio de Procedencia

Se puede evidenciar la relación o correlación que tienen algunas variables respecto a la variable objetivo *con_fin_* para determinar seguidamente los posibles modelos a implementar. el Ranking de variables que resultó del análisis fueron:

1. Silicosis = silicosis_
2. Enfermedad Hepática = enfe_hepat
3. Cancer = cancer_
4. Edad = edad_
5. Enfermedad renal = enfe_renal
6. Epoc = epoc_
7. Artritis = artritis_

8. Desnutrición = desnutricion_

Figura 1. Gráfica de correlación de variables por TB

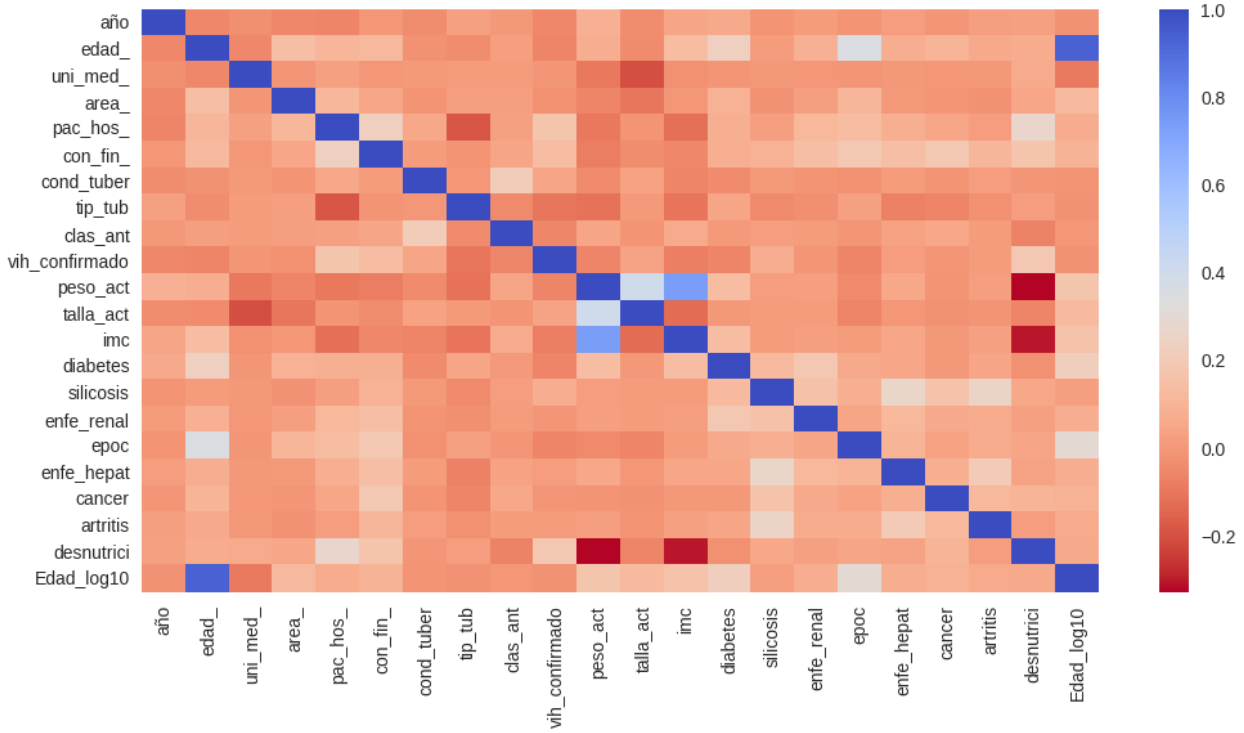
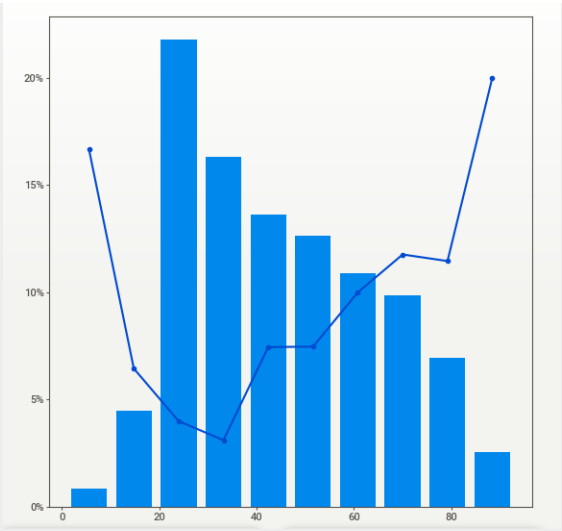


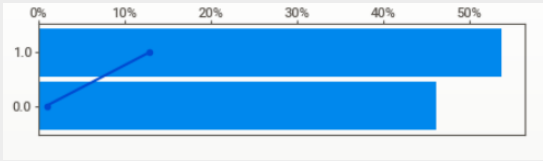
Figura 2. Variables relevantes de patologías por TB incidentes en la variable objetivo

Simbología:

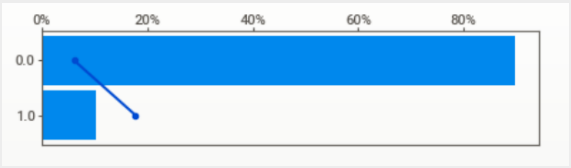
Datos:
% cond_fin_:



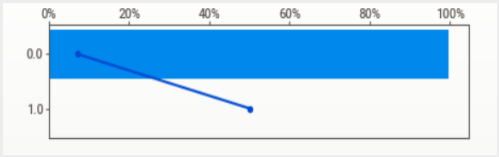
Variable: Edad



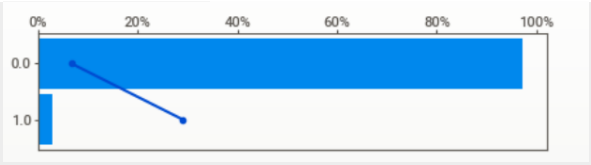
Variable: Pac_hosp



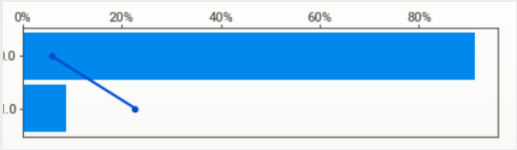
Variable: VIH_confirmad



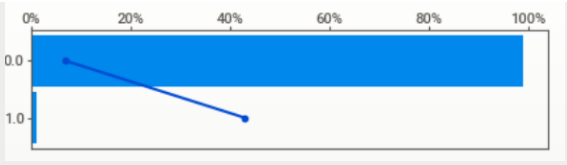
Variable: Silicosis



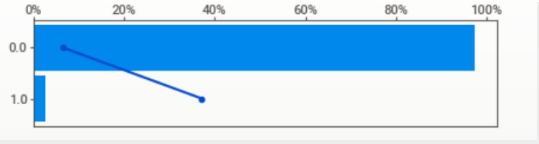
Variable: enfe_renal



Variable: epoc



Variable: enfe_hepat



Variable: cancer

Al observar las variables estas no presentan una gran variabilidad ya que existen muy pocos registros donde la persona fallece a causa de la variable. Se estima conveniente aplicar el balanceo de clases.

A través de Python y la librería PyCaret se realizan dos modelos el cual muestra que a mayor número de variables se mejora poco a poco el modelo ya que se tienen pocos registros de fallecimientos.

RESULTADOS

Creación inicial de modelo basado en reglas:

Se realizo la separación de la tabla en entrenamiento y testeo donde obtenemos los X_train, y_train, X_test, y_test. Seguidamente se crea un modelo de ajuste y predicción con una sola variable para obtener modelo de referencia y partir a mejorar y comprar con el modelo machine learning que se piensa realizar; donde obtenemos un acuracy = 94.92 pero con ella una precisión = 0.00

Tabla 2. Resultados iniciales modelo basado en reglas

	precision	recall	f1-score	support
0	0.95	1.00	0.97	262
1	0.00	0.00	0.00	14

A continuación, se presentan los resultados del modelo machine learning el cual arrojó mejores resultados en comparación con el modelo basado en reglas.

Se comparan los modelos y obtenemos la siguiente tabla de modelos:

Tabla 3. Comparación de modelo final

Model	Accuracy	AUC	Recall	Prec.	F1
Quadratic Discriminant Analysis	0.8897	0.6535	0.2095	0.3071	0.2385
Linear Discriminant Analysis	0.9001	0.6525	0.1048	0.2583	0.1433
Ada Boost Classifier	0.9157	0.6523	0.0429	0.25	0.0722
Logistic Regression	0.9157	0.6488	0.0286	0.2	0.05
Naive Bayes	0.8884	0.6478	0.2095	0.3018	0.2357

Figura 2: Curva ROC

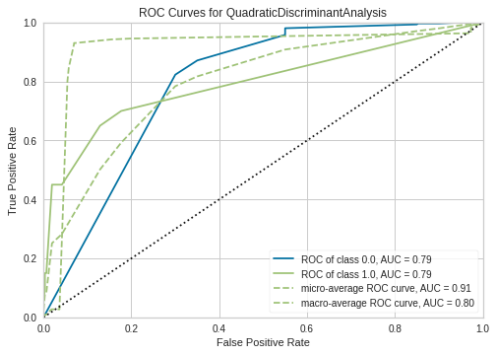


Figura 3. Matriz de confusión

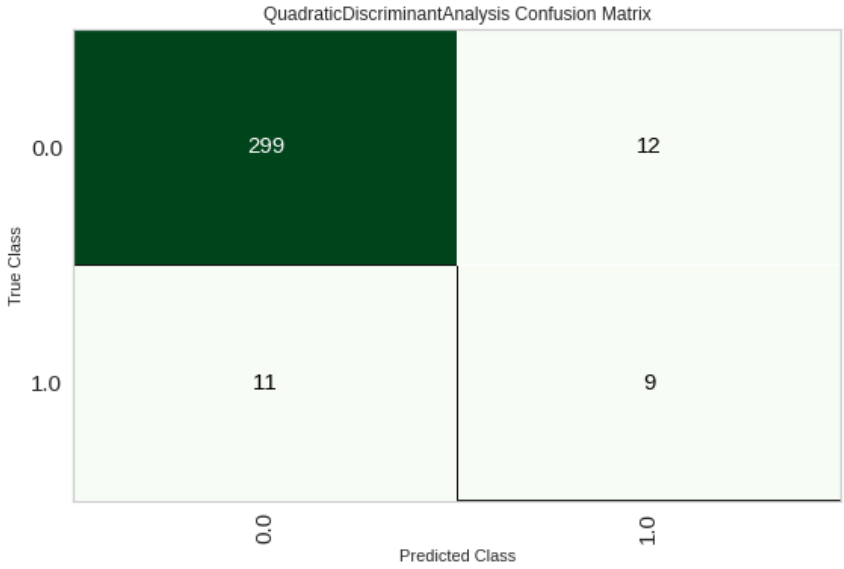


Tabla 5 : Predicción del modelo

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Quadratic Discriminant Analysis	0.9305	0.7947	0.45	0.4286	0.439	0.402	0.4021

DISCUSIÓN DE RESULTADOS

De acuerdo con la metodología planteada se mejora el modelo basado en reglas; ya que este no mostraba ninguna precisión al respecto para evaluar las condiciones de los pacientes por TB, este mostraba que de 27 personas fallecidas solo podía predecir a 1.

Se crea un modelo inicial el cual tiene una pequeña mejora ilustrada en una matriz de confusión donde de 20 casos el modelo alcanza a clasificar correctamente 9 de ellos.

Estos resultados tienen un gran problema y es el sesgo que tienen los datos ya que no se contaba con un número significativo de muertes por TB lo que sesgaba como tal la base. Se aplicó en un momento dado el balanceo de clases, pero este empeoró como tal el modelo haciendo que su predicción respecto a las muertes y a las personas que vivieron fuera de 0.

CONCLUSIONES

Para este ejercicio este instrumento se puede tomar en cuenta para un pequeño análisis inicial de prioridad a un paciente diagnosticado con TB, ya que arroja resultados de certeza de más del 50%; pero aclarando que no debe ser decisivo en la priorización de pacientes con TB. Este instrumento debe ser evaluado nuevamente para poder establecer a través de los reportes de las instituciones de salud, que variables pueden ser más incisivas en la predicción de mortalidad por TB; esto con el fin de poder actuar de manera pronta y poder clasificar la prioridad de paciente creando una oportunidad de que podamos disminuir el riesgo de mortalidad y salvar más vidas con una atención temprana y priorizada.

REFERENCIAS BIBLIOGRÁFICAS

[1] World Health Organization, “Global report Tuberculosis 2019,” WHO., Geneva., GE, Switzerland, WHO/CDS/TB/2019. 15, 2019.

[2] D. Moher et al. “Preferred reporting items for systematic review and meta-analysis protocols

(PRISMA-P) 2015 statement,” Syst Rev, vol. 4, no. 1, pp. 2046-4053, 2015.

[3] J. Melendez, et al. “An automated tuberculosis screening strategy combining X-ray-based

computer-aided detection and clinical information,” Sci Rep, vol. 6, no. 1, pp. 25265, 2016.

[4] A. S. Becker, et al. “Detection of tuberculosis patterns in digital photographs of chest X-ray images using Deep Learning: feasibility study,” Int J Tuberc Lung Dis, vol. 22, no. 3, pp. 328-335, 2018.

ANEXOS

Gráfico Anexo 1: Matriz de confusión modelo inicial

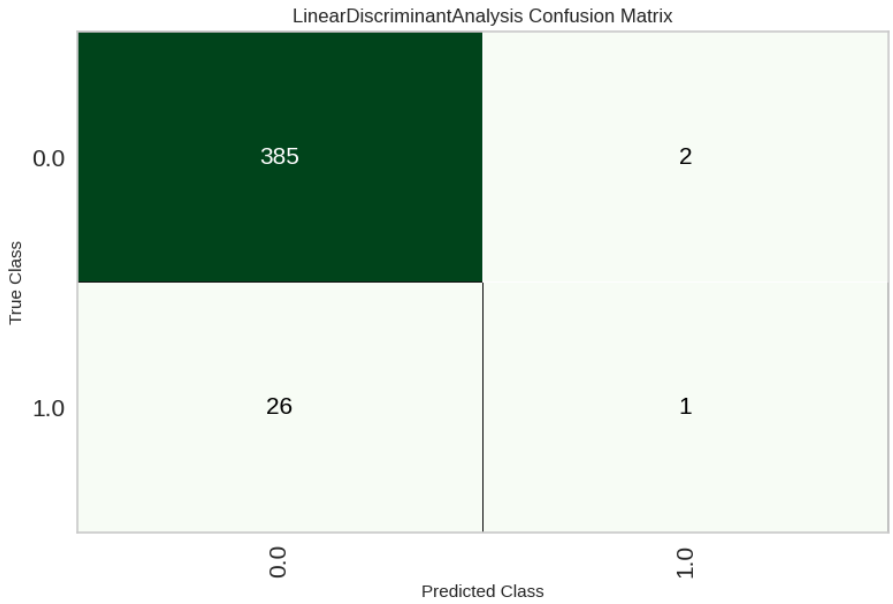


Gráfico Anexo 2: Informe de clasificación de análisis discriminante

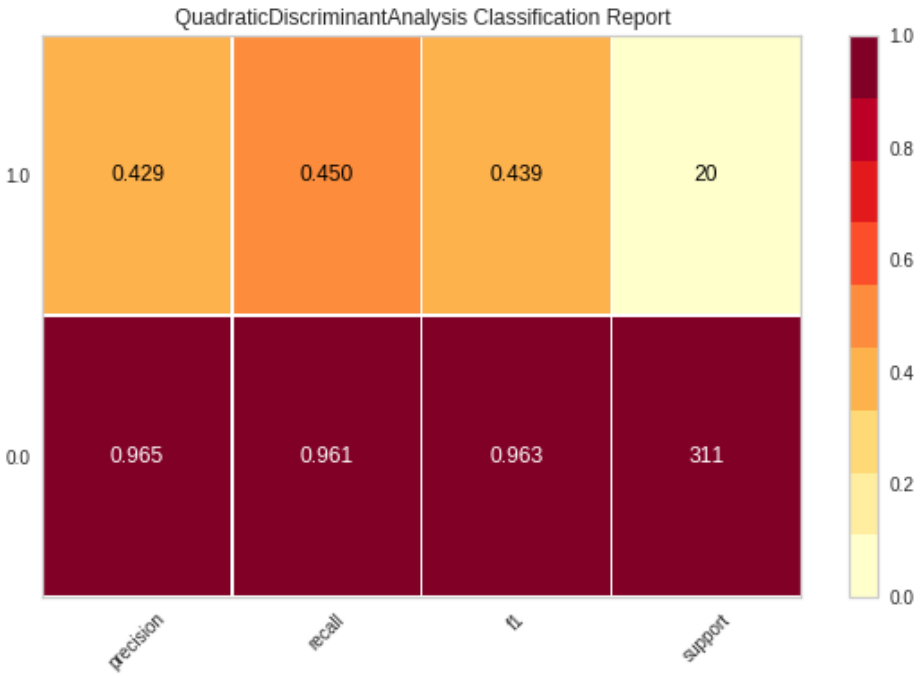


Tabla anexo 1. Desempeño promedio de los mejores modelos

	Accuracy	AUC	Recall	Prec.	F1
Fold					
0	0.9103	0.7153	0.2857	0.5000	0.3636
1	0.9091	0.5927	0.1667	0.3333	0.2222
2	0.8961	0.6655	0.3333	0.3333	0.3333
3	0.8571	0.5951	0.1667	0.1429	0.1538
4	0.8961	0.6745	0.4286	0.4286	0.4286
5	0.8961	0.6643	0.1429	0.3333	0.2000
6	0.9091	0.7163	0.4286	0.5000	0.4615
7	0.8571	0.6857	0.0000	0.0000	0.0000
8	0.9091	0.5541	0.1429	0.5000	0.2222
9	0.8571	0.6714	0.0000	0.0000	0.0000

Mean	0.8897	0.6535	0.2095	0.3071	0.2385
Std	0.0220	0.0518	0.1475	0.1855	0.1528

Anexo 3: Códigos de la investigación, Link: Google Colab