LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA

Identificación y pronóstico de sífilis congénita mediante técnicas de Aprendizaje Automático para las localidades de Usme, Tunjuelito, Ciudad Bolívar y Sumapaz (Bogotá D.C.)

Identification and prognosis of Congenital syphilis Using Machine Learning Techniques in Usme, Tunjuelito, Ciudad Bolívar and Sumapaz (Bogotá D.C.)

Carlos Fernando Macana González, cfmacanag@libertadores.edu.co
José John Fredy González Veloza, jjgonzalezv02@libertadores.edu.co

RESUMEN

La sífilis congénita es una infección bacteriana grave transmitida en un recién nacido de una madre que no fue tratada o fue tratada de manera inadecuada para la sífilis durante el embarazo; las consecuencias de esta infección en el bebé están relacionadas con una afectación en la calidad de vida y enfermedades como masas abdominales, bajo peso, anormalidades esqueléticas y dolores óseos, inflamación articular, ceguera, sordera, entre otros, e inclusive la muerte, por lo que constituye un problema de interés en salud pública a nivel mundial; esto ha llevado a los gobiernos y científicos a la búsqueda de estrategias para la reducción de nuevos casos de sífilis en bebés; por ello la importancia de contar con modelos predictivos como herramienta para la identificación temprana de factores de riesgo o variables en las embarazadas y así realizar una acción en salud para evitar el contagio de sífilis al recién nacido. Desde este punto de gravedad e impacto que la sífilis congénita genera, el presente trabajo utilizó técnicas de aprendizaje automático para la elaboración de modelos predictivos que apoyen la identificación de variables relacionadas con la aparición de nuevos casos de recién nacidos infectados y que sean útiles en las instituciones de salud para el manejo oportuno del tratamiento en la mujer gestante; esto a partir del conocimiento sobre variables sociodemográficas y de salud de la madre y su contexto. Se contó con un conjunto de datos que recopilan información sociodemográfica y de salud de una cohorte de 451 mujeres gestantes con diagnóstico positivo para sífilis de las localidades de Usme, Tunjuelito, Ciudad Bolívar y Sumapaz (Bogotá D.C.); se contó con información básica acerca del recién nacido en cuanto a su peso y estado de contagio de sífilis; se identificó en el conjunto de datos que el 21,5% (n=97) de los nacimientos de madres con sífilis también nacieron con sífilis (sífilis congénita); se entrenaron 12 modelos de predicción de sífilis congénita mediante técnicas de aprendizaje automático supervisado. El principal resultado ha sido generar cuatro modelos predictivos, *K Neighbors Classifier, Light Gradient Boosting*

Machine, Gradient Boosting Classifier y *Random Forest Classifier*. Sobre los modelos de predicción se evaluaron sus métricas de desempeño para seleccionar el mejor de ellos, logrando un F1-Score del 77,28% en el modelo basado en *K Neighbors Classifier* , del 73,69% en el modelo basado en *Light Gradient Boosting Machine,* del 73,76% en el modelo basado en *Gradient Boosting Classifier* y del 68,38% en el modelo basado en *Random Forest Classifier,* además con sensibilidad por encima del 70%, superando las métricas de desempeño de un modelo inicial basado en reglas; se consideran como variables relevantes en el potencial predictivo del modelo basado en algoritmos de aprendizaje automático: el número de semanas de gestación al momento del primer control prenatal; la edad de la madre y del; procedencia de la madre y el número de controles prenatales totales esperados.

Palabras clave: sífilis congénita, sífilis gestacional, aprendizaje automático, modelos predictivos, *K Neighbors Classifier*

ABSTRACT

Congenital syphilis is a serious bacterial infection transmitted in a newborn from a mother who was not treated or was inadequately treated for syphilis during pregnancy; the consequences of this infection in the baby are related to an affectation in the quality of life and diseases such as abdominal masses, low weight, skeletal abnormalities and bone pain, joint inflammation, blindness, deafness, among others, and even death, so it is a problem of interest in public health worldwide; This has led governments and scientists to search for strategies to reduce new cases of syphilis in infants; hence the importance of having predictive models as a tool for early identification of risk factors or variables in pregnant women and thus perform a health action to prevent the transmission of syphilis to the newborn. From this point of gravity and impact that congenital syphilis generates, the present work used machine learning techniques for the elaboration of predictive models that support the identification of variables related to the appearance of new cases of infected newborns and that are useful in health institutions for the timely management of treatment in pregnant women; this from the knowledge of sociodemographic and health variables of the mother and her context. A data set was available that compiles sociodemographic and health information of a cohort of 451 pregnant women with positive diagnosis for syphilis; basic information was available about the newborn in terms of weight and syphilis infection status; it was identified in the data set that 21.5% (n=97) of the births of mothers with syphilis were also born with syphilis (congenital syphilis); 12 prediction models of congenital syphilis were trained using supervised automatic learning techniques. The main result has been to generate four predictive models, K Neighbors Classifier, Light Gradient Boosting Machine, Gradient Boosting Classifier and Random Forest Classifier. The performance metrics of the predictive models were evaluated to select the best of them, achieving an F1-Score of 77.28% in the model based on K Neighbors Classifier, 73.69% in the model based on Light Gradient Boosting Machine, 73.76% in the model based on Gradient Boosting Classifier and 68.38%

in the model based on Random Forest Classifier, also with sensitivity above 70%, exceeding the performance metrics of an initial model based on rules; are considered as relevant variables in the predictive potential of the model based on machine learning algorithms: the number of weeks of gestation at the time of the first prenatal checkup; the age of the mother and the; origin of the mother and the number of expected total prenatal checkups.

Keywords: congenital syphilis, gestational syphilis, machine learning, predictive models, *K Neighbors Classifier*, *Light Gradient Boosting Machine*, *Gradient Boosting Classifier*, *Random Forest Classifier*.

INTRODUCCIÓN

La sífilis es una infección bacteriana causada por *Treponema pallidum* que resulta en morbilidad y mortalidad de alto impacto en salud pública. Se trata de una infección de transmisión sexual (ITS) que generalmente se transmite por el contacto con úlceras infecciosas de los órganos genitales, el ano, el recto, los labios o la boca; por medio de las transfusiones de sangre, o mediante la transmisión vertical materno infantil durante el embarazo (OMS, 2022). Es más preocupante aún tratándose de una mujer en estado de gestación con diagnóstico positivo dado el impacto en salud pública que genera este evento, tanto a nivel humano, social y económico y generando con ello al tiempo otras condiciones de salud, tanto en la mujer como en el recién nacido; las consecuencias de esta infección en el bebé están relacionadas con una afectación en la calidad de vida y enfermedades como masas abdominales, bajo peso, anormalidades esqueléticas y dolores óseos, inflamación articular, ceguera, sordera, entre otros, e inclusive la muerte (Instituto Nacional de Salud, 2018; OMS, 2019). Según el Observatorio de datos en salud de Bogotá la incidencia de sífilis congénita para el 2015 era de 1,1 casos por cada 1.000 nacimientos (nacidos vivos + Muertes fetales), para el año 2021 fue de 1,5 casos por cada 1.000 nacimientos (nacidos vivos + Muertes fetales), se identifica una tendencia al aumento; en el primer semestre de 2022 ya el resultado de esta situación era de 1,4 casos de sífilis congénita por cada 1.000 nacimientos (nacidos vivos + Muertes fetales) (Secretaria Distrital de Salud, 2022).

Diferentes son las estrategias o mecanismos para la identificación de la población de mujeres gestantes infectadas con sífilis, los cuales pasan por la búsqueda en las ciudades y territorios de las gestantes para valorar sus condiciones de salud y realizar pruebas rápidas para la detección de la enfermedad, atención integral en salud desde los gobiernos y empresas administradoras de planes de beneficio, además de acciones de promoción de la salud y prevención de la enfermedad con base en la educación y la comunicación; así mismo son diferentes los mecanismos para avanzar en el cumplimiento al tratamiento de la enfermedad, principalmente aplicación de penicilina benzatínica en oportunidad a la gestante (tres dosis, la primera al menos 30 días antes del parto), la cual es de fácil acceso. Todas estas actividades generan datos acerca de la población (sexo, edad, ciudad de residencia, lugar de origen), sus

condiciones de salud (resultados de laboratorio, antecedentes clínicos, aseguramiento en salud) y los factores de riesgo para que aparezca la enfermedad (Adhikari, 2020). Sin embargo, los sistemas de información y gestión del riesgo carecen de técnicas sofisticadas para el análisis y clasificación/predicción de población según riesgo en salud y condiciones de vida, y en general, de toda la información disponible.

Por su parte, el aprendizaje automático, como paradigma en el marco las ciencias de la información y los datos, proporciona técnicas que permiten la identificación de patrones en el comportamiento de múltiples situaciones económicas, sociales, políticas, geográficas, de salud, entre otros campos del conocimiento, con el fin de poder abordar posibles soluciones a los problemas que ha generado la falta de comprensión de la realidad de donde se recolectaron los datos (Lakshmanan et al., 2021). Los algoritmos de aprendizaje automático están diseñados para lograr predecir con éxito las enfermedades; varios investigadores ya han trabajado en la predicción de algunas enfermedades con base en algoritmos de aprendizaje automático como (Gomathy, 2021) que usa los algoritmos de aprendizaje automático como referencia para el modelo de predicción de enfermedades cardiacas, por ejemplo, diferentes algoritmos de aprendizaje automático como Máquinas de Soporte Vectorial, árboles de decisión y *Random Forest*, han tenido el mejor desempeño en comparación con los demás, con precisión superior al 90%. En otros estudios los modelos aplicados a la predicción de enfermedades se han usado los algoritmos *K Nearest Neighbor* y regresión logística, que obtuvieron una precisión del 80% lo que es mayor a las obtenidas en investigaciones anteriores; la precisión puede mejorarse con el aumento de las variables médicas que se usan en el conjunto de datos (Jeyaganesan et al., 2020).

Este trabajo busca, a partir de técnicas y flujos de trabajo del aprendizaje automático, el desarrollo de modelos predictivos, e identificar los principales factores sociodemográficos y de salud de la madre, para identificar la sífilis en bebés (sífilis congénita), respondiendo a la pregunta de investigación ¿Los modelos basados en técnicas de aprendizaje automático pueden predecir la sífilis congénita como instrumento para el tratamiento e intervención en salud en las gestantes, a partir de datos sociodemográficos y de salud en las localidades de Usme, Tunjuelito, Ciudad Bolívar y Sumapaz?; estos resultados son relevantes en la toma de decisiones en cuanto a la prestación de servicios de salud, que pasa por capacidades humanas, insumos y recursos financieros.

El presente trabajo se organiza de la siguiente manera: un apartado metodológico que indica los recursos y técnicas utilizadas para el desarrollo de modelos predictivos con base en algoritmos de aprendizaje automático, incluida la evaluación de desempeño de los modelos; posterior se presentan los principales resultados con énfasis en las métricas de desempeño de los modelos seleccionados para predecir la sífilis congénita, luego se presenta la discusión y conclusiones, y al final se encontrarán las referencias bibliográficas.

METODOLOGÍA

El presente trabajó desarrolló una metodología que consideró los recursos técnicos y operativos para el logro de elaborar un modelo predictivo con base en técnicas de aprendizaje automático, por lo que se realizó una adaptación de las metodologías propuestas por (Alanazi, 2022; Bao et al., 2021; Forradellas et al., 2021), logrando establecer cinco etapas como se presenta a continuación; se implementó la metodología del proyecto con lenguaje Python usando librerías para la lectura de datos (*pandas*, *numpy*), elaboración de gráficos y análisis descriptivo (*matplotlib*, *sweetviz*), preprocesamiento de datos (*sklearn*) y desarrollo y análisis de desempeño de modelos predictivos con base en algoritmos de aprendizaje automático (*pycaret*, *sklearn*).

Etapa 1. Selección de datos de estudio y definición de variable de estudio

Se usó un conjunto de datos de registros de mujeres gestantes residentes de cuatro localidades del sur de la ciudad de Bogotá D.C. (Usme, Tunjuelito, Ciudad Bolívar y Sumapaz) que fueron notificadas al Sistema de Información de Vigilancia en Salud Pública del Instituto Nacional de Salud (SIVIGILA) al ser un caso con diagnóstico confirmado de sífilis (Evento SIVIGILA: 750), definido como "*toda mujer gestante, puérpera o con aborto en los últimos 40 días con o sin signos clínicos sugestivos de sífilis (ejemplo: úlcera genital, erupción cutánea, placas en palmas y plantas), con prueba treponémica rápida positiva acompañada de una prueba no treponémica reactiva (VDRL, RPR) a cualquier dilución, que no ha recibido tratamiento adecuado para sífilis durante la presente gestación o que tiene una reinfección no tratada*" (Instituto Nacional de Salud, 2022). El conjunto de datos recopila información sociodemográfica y de salud (20 variables iniciales, incluida la variable de estudio) de una cohorte de 451 mujeres gestantes con diagnóstico positivo para sífilis; se identificó en el conjunto de datos que el 21,5% (n=97) de los nacimientos de madres con sífilis también nacieron con sífilis (sífilis congénita, evento SIVIGILA: 740), según la estructura de los registros para la notificación de los datos de eventos de interés en salud pública (Instituto Nacional de Salud, 2014). Se presenta en Tabla 1 la descripción de las variables de la base de datos y se guarda reserva de los datos de identidad de las mujeres gestantes según la normatividad vigente en el país.

Tabla 1. Descripción de variables

Variable	Descripción	Rango/medida
ANIO	Año de identificación del evento de interés en salud pública	2019-2020-2021

Variable	Descripción	Rango/medida
SUBRED_NOT	ESE que notifica el evento	CENTRO ORIENTE, NORTE, SUR, SUR OCCIDENTE
UPGD _NOT	IPS que notifica el evento	133 IPS NOTIFICARON EVENTO 750
EVENTO	evento 750: Sífilis gestacional	750 TODO ES SÍFILIS GESTACIONAL
VENZ	Si se identifica como población venezolana	SI/NO
EDAD_MADRE	Edad de la madre en años	14-42 AÑOS
NIVEL_EDU_MADRE	Nivel educativo de la madre	10 CATEGORÍAS INCLUIDO NINGUNO
EDAD_PADRE	Edad del padre en años	17-57 AÑOS
NIVEL_EDU_PADRE	Nivel educativo del padre	10 CATEGORÍAS INCLUIDO NINGUNO
EST_CONY_MADRE	Estado conyugal de la madre	5 CATEGORÍAS
IEC	Si se hizo Investigación Epidemiológica de Campo a la gestante con diagnóstico de sífilis	SI/NO VISITA
DOSIS_TRAT_IEC	Número de dosis de tratamiento con penicilina benzatínica en la gestante al momento de la IEC	0-6 DOSIS
PAREJA_TRATAMIENTO	La pareja de la gestante con diagnóstico de sífilis cuenta con tratamiento de penicilina benzatínica	SI/NO
EFECTIVO_HOGAR	Visita efectiva en la vivienda por equipo de salud	SI/NO
PESO_NV	Peso del recién nacido en gramos	1110-4385 GRAMOS
EVENTO_740	Evento 740: Sífilis congénita **(variable objetivo de estudio)**	SI/NO
SEM_GES	Número de semanas de gestación al momento del primer control prenatal	29-41 SEMANAS
CONTROL_PRENATAL	Número de controles prenatales	0-21 CONTROLES
REGIMEN	Régimen de afiliación al Sistema General de Seguridad Social y de Salud (SGSSS)	CONTRIBUTIVO, NO ASEGURADO, SUBSIDIADO, VINCULADO
LOC_RES	Localidad de residencia de la gestante	CIUDAD BOLÍVAR, USME, TUNJUELITO, SUMAPAZ

Etapa 2. Selección de variables explicativas para la predicción de riesgo de sífilis congénita y su relación con la variable de estudio.

La selección de características relevantes se basó en la revisión de la literatura, con especial atención en el trabajo de (Ahsan et al., 2022), (Fitzpatrick et al., 2020) y (Uddin et al., 2019), así como en la opinión de expertos. Las variables con potencial predictivo para el modelado incluyeron la procedencia de la madre, edad y nivel educativo de madre y padre, estado conyugal de la madre, dosis de tratamiento para sífilis, número de controles prenatales, régimen de seguridad social, localidad de residencia y número de semanas de gestación al iniciar controles prenatales.

Etapa 3. Transformación de variables

Para el desarrollo de los modelos predictivos a partir de técnicas de aprendizaje automático se realizan las siguientes transformaciones a los datos dada la necesidad de mejorar la estructura de los datos:

- Variable EVENTO_740 que permite respuestas SI/NO se transformó a 0 = SI y 1 = NO que crea una nueva variable EVENTO_740_SI.
- Variable VENZ que permite respuestas SI/NO se transformó a 0 = SI y 1 = NO que crea una nueva variable VENZ_SI.
- Variable DOSIS_TRAT_IEC se transformó a entero (*int*).
- Variables EDAD_MADRE y EDAD_PADRE dado el sesgo que presenta cada variable en el extremo superior (sesgo a la derecha), creando nuevas variables, LOG10_EDAD_MADRE y LOG10_EDAD_PADRE, respectivamente.

Etapa 4. Desarrollo y entrenamiento de modelos para construir una herramienta de predicción de riesgo de sífilis congénita

Se establecieron una serie de modelos de aprendizaje automático lineales y no lineales que involucran algoritmos de regresión, incluida la regresión logística; así como modelos basados en algoritmos *K Neighbors Classifier, Light Gradient Boosting Machine, Random Forest Classifier*, entre otros. Se separó el conjunto de datos tanto en variable objetivo (y=EVENTO_740_SI) y variables explicativas (x= VENZ_SI, LOG10_EDAD_MADRE, NIVEL_EDU_MADRE, LOG10_EDAD_PADRE, NIVEL_EDU_PADRE, EST_CONY_MADRE, DOSIS_TRAT_IEC, EFECTIVO_HOGAR, SEM_GES, CONTROL_PRENATAL, REGIMEN, LOC_RES) para entrenamiento y para prueba. Los modelos de sífilis congénita se entrenaron utilizando el conjunto de datos de entrenamiento con un método de validación cruzada.

Etapa 5. Evaluación de desempeño de los modelos de predicción para sífilis congénita

Se evaluaron los modelos con base en las métricas de la matriz de confusión como la sensibilidad, precisión, exactitud, y al tener datos desbalanceados se analizan las métricas de AUC y F1-Score, con base en ello se revisaron otros aspectos sobre los modelos con el mejor desempeño para revisar otros elementos para el análisis de los resultados.

RESULTADOS

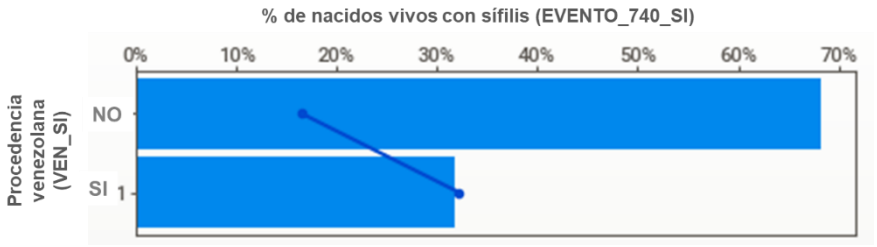
Análisis descriptivo del conjunto de datos

De las 451 mujeres gestantes (entre 14 y 42 años) con diagnóstico de sífilis del conjunto de datos se identifica que el 21,5% (n=97) de los nacimientos también nacieron con sífilis (sífilis congénita). Además, el resultado del coeficiente de incertidumbre indica relación con variables categóricas como la nacionalidad es o no venezolana, el régimen de seguridad social y la dosis de tratamiento; en cuanto a las variables numéricas se identifica relacionadas

con la variable de estudio el número de semanas de gestación al inicio de controles prenatales y el número de controles prenatales esperados al momento del parto.

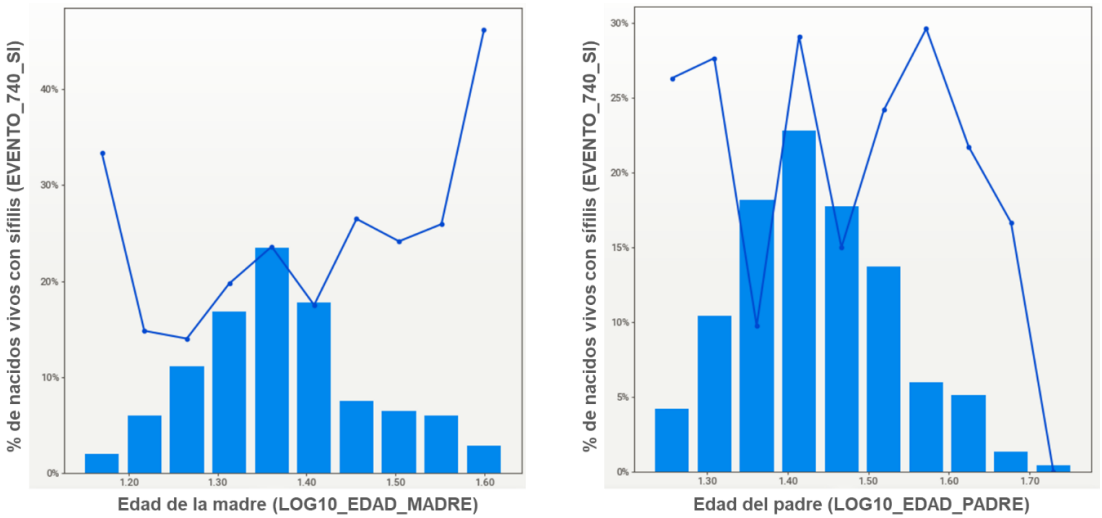
Se identifica en el conjunto de datos de mujeres embarazadas con sífilis que el 68,07% (n=307) de ellas son colombianas, mientras que el 31,93% (n=144) son venezolanas; a partir de ello, resulta que el grupo de mujeres colombianas con sífilis tuvo un 16,61% (n=51) de bebés infectados, mientras que en el conjunto de mujeres venezolanas la proporción de bebés que nacieron con sífilis se localizó en el 31,94% (n=46) (Figura 1), esto indica que el riesgo para una mujer gestante que no tiene una nacionalidad colombiana de desencadenar su parto en una sífilis congénita es mayor que la mujer nacional colombiana.

Figura 1. Relación entre gestantes con sífilis según su procedencia versus porcentaje de sífilis congénita.



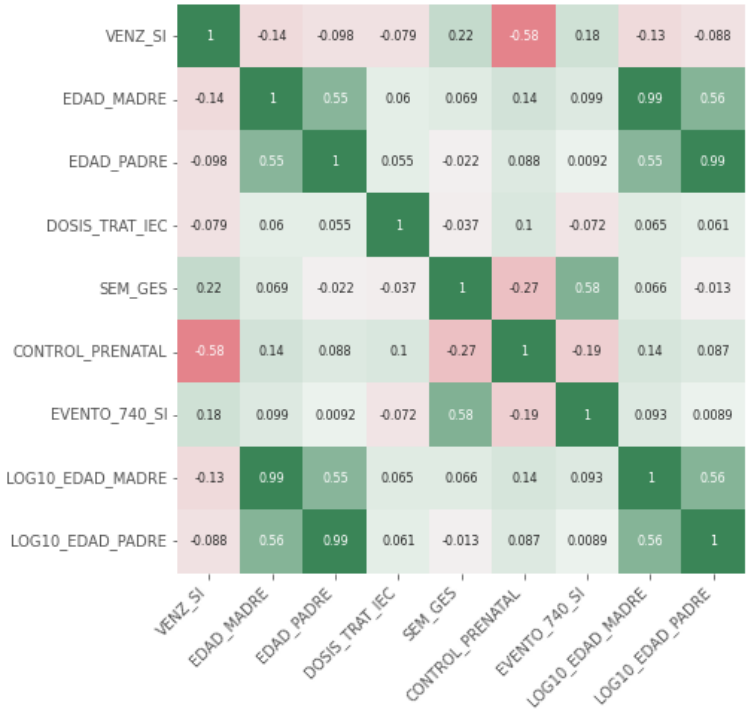
En relación con la edad de la madre, se identifica que en la mujeres embarazadas adolescentes y jóvenes (12 y 28 años) el promedio de casos de sífilis congénita es de 17,80% mientras que en mujeres adultas (entre 29 y 59 años) la proporción media de casos de recién nacidos que nacen con la infección es del 37,90% (Figura 2); respecto al porcentaje de padres de bebés con sífilis en el grupo de padres adolescentes y jóvenes en relación al grupo de padres adultos no hay diferencia representativa, siendo esto 21,13% y 22,16 respectivamente.

Figura 2. Relación entre gestantes con sífilis según edad y edad del padre versus porcentaje de sífilis congénita.



En cuanto a la correlación lineal de Pearson de la variable de estudio con demás variables numéricas se identifica que esta relación es débil; se tiene una correlación con la variable CONTROL_PRENATAL (número de controles prenatales totales esperados en la gestación) de -0,19, con la variable VEN_SI (si tiene procedencia venezolana) de 0,18 y con la variable SEM_GES (semanas de gestación al inicio de controles prenatales) de 0,58 (Figura 3).

Figura 3. Correlación de variables (Pearson).

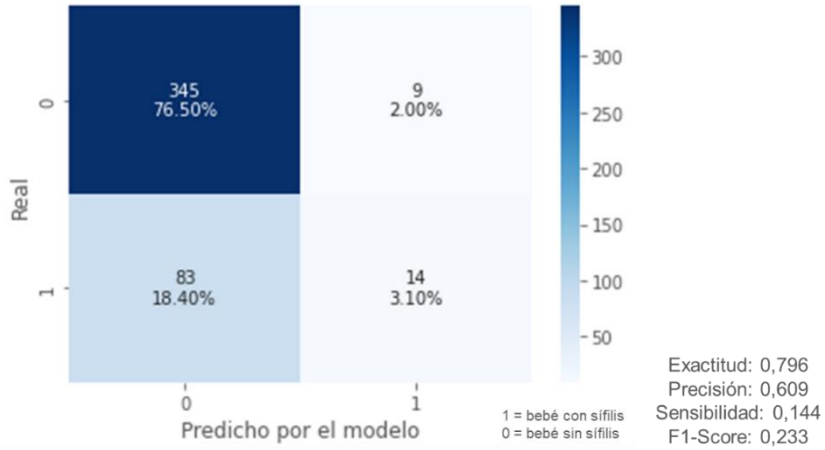


Modelo base inicial

De acuerdo con el análisis descriptivo anterior se identifican puntos de inflexión que permiten localizar umbrales para elaborar un modelo basado en reglas bajo las siguientes condiciones para predecir la sífilis congénita con el conjunto de datos disponible:

Mujer embarazada mayor de 26 años.
Mujer embarazada con procedencia venezolana.
Mujer embarazada con primer control prenatal mayor a la semana 12 de gestación.

Figura 3. Matriz de confusión modelo basado en reglas



De acuerdo con las métricas de desempeño del modelo inicial basado en reglas se puede indicar que el rendimiento general del modelo es del 23,3% (según F1-Score). Respecto a la precisión que se logra con el modelo basado en reglas se identifica que el 60,9% de las predicciones de casos positivos están correctamente predichos. En cuanto a la sensibilidad, la proporción de casos positivos que fueron correctamente identificados por el modelo es de 14,4%

Modelo basado en técnicas de aprendizaje automático para la predicción de casos de sífilis congénita

Se construyó un conjunto de 12 modelos de predicción de sífilis congénita (variable y) utilizando los predictores más importantes (variable x) mediante técnicas de aprendizaje automático y se analizan aquellos con las mejores métricas de desempeño (Tabla 2).

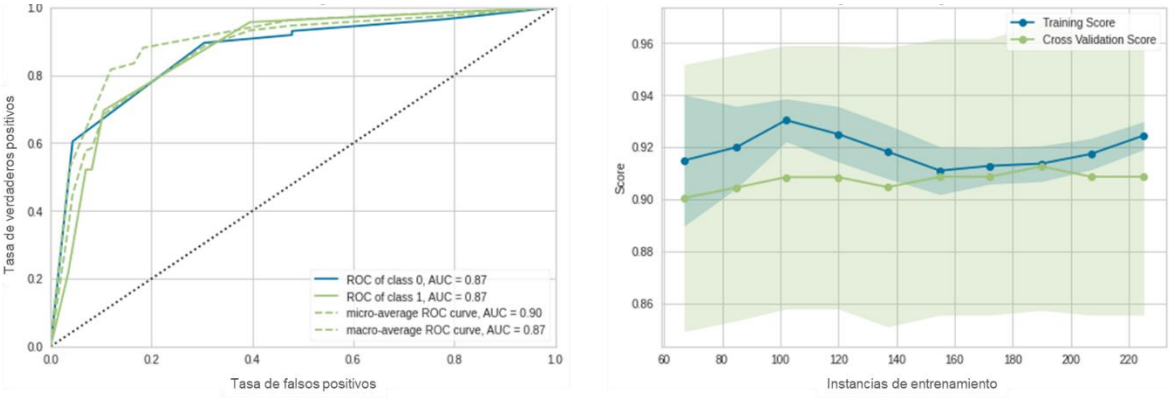
Tabla 2. Modelos basados en técnicas de aprendizaje automático para la predicción de casos de sífilis congénita.

Modelo	Exactitud	AUC	Sensibilidad	Precisión	F1
K Neighbors Classifier	0.9086	0.8476	0.7467	0.8188	0.7728

Modelo	Exactitud	AUC	Sensibilidad	Precisión	F1
Light Gradient Boosting Machine	0.9003	0.8515	0.7067	0.8105	0.7369
Gradient Boosting Classifier	0.8926	0.8598	0.7267	0.7738	0.7376
Random Forest Classifier	0.8883	0.8627	0.6067	0.8183	0.6838
Logistic Regression	0.8763	0.8638	0.6467	0.7471	0.6656
Ada Boost Classifier	0.8645	0.7790	0.6067	0.6871	0.6306
Ridge Classifier	0.8606	0.0000	0.5900	0.7250	0.6125
Decision Tree Classifier	0.8406	0.7625	0.6300	0.6529	0.6300
Extra Trees Classifier	0.8406	0.8368	0.3733	0.6267	0.4461
Linear Discriminant Analysis	0.8405	0.8442	0.6467	0.6186	0.6125
Quadratic Discriminant Analysis	0.5582	0.4442	0.2533	0.1268	0.1663
Naive Bayes	0.2391	0.6152	0.9433	0.2033	0.3341

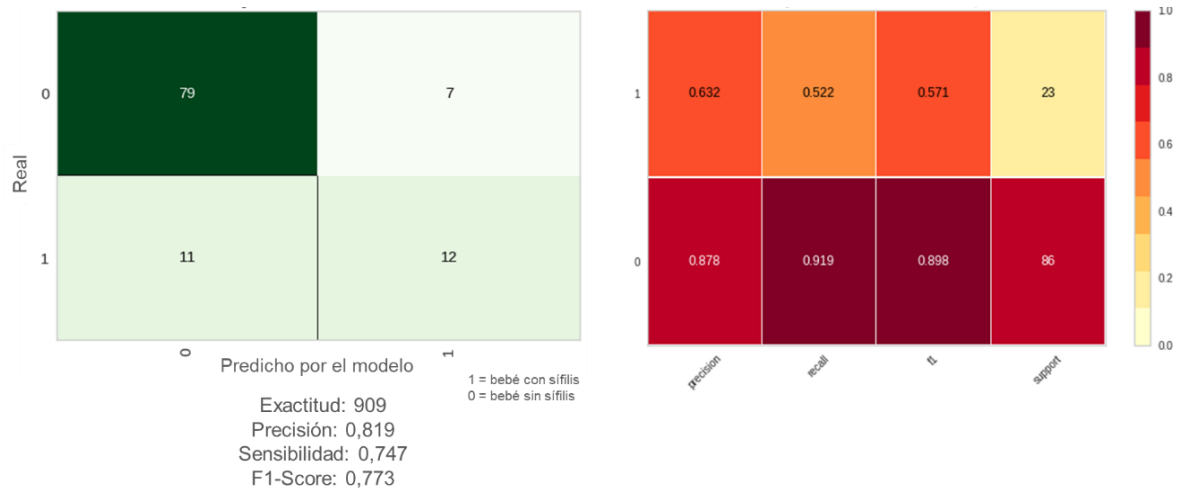
El modelo de predicción que resultó con base en el algoritmo *K Neighbors Classifier* obtuvo un rendimiento óptimo para la predicción de sífilis gestacional; según F1-Score el rendimiento general del modelo en datos de desarrollo es del 77,28% superior en un 100% el desempeño general del modelo original basado en reglas, y del 78,0% en datos de testeo; el modelo presenta un AUC de 84,76% (Figura 4).

Figura 4. Curva ROC y curva de aprendizaje para el modelo de predicción de sífilis congénita basado en *K Neighbors Classifier*.



Respecto a la precisión que se logra en este modelo se identifica que el 81,88% de las predicciones de casos positivos están correctamente predichos. En cuanto a la sensibilidad, la proporción de casos positivos que fueron correctamente identificados por el modelo es de 74,67%; en datos de testeo estas métricas cambian y se localizan en 90,1% y 69,6% respectivamente, teniendo un buen modelo (Figura 5).

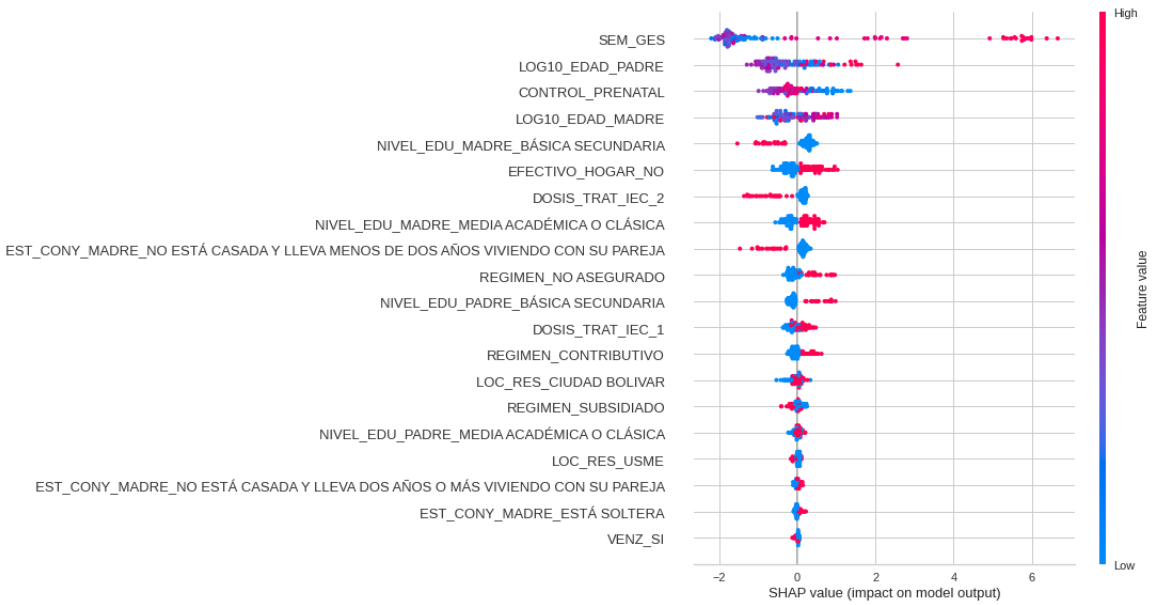
Figura 5. Matriz de confusión y reporte de clases para el modelo de predicción de sífilis congénita.



Respecto a los demás modelos relevantes en los resultados de las métricas de desempeño se encuentra el modelo de predicción de casos de sífilis congénita que resultó con base en el algoritmo *Light Gradient Boosting Machine* obtuvo un rendimiento óptimo para la predicción de sífilis gestacional; según F1-Score el rendimiento general del modelo en datos de desarrollo es del 73,69% superior en un 100% el desempeño general del modelo original basado en reglas, y del 54,5% en datos de testeo; el modelo de predicción que resultó con base en el algoritmo *Gradient Boosting Classifier* obtuvo también un rendimiento óptimo para la predicción de sífilis gestacional; según F1-Score el rendimiento general del modelo en datos de desarrollo es del 73,76% y del 61,1% en datos de testeo; el modelo presenta un AUC de 85,98%. Para finalizar, el modelo de predicción de casos de sífilis congénita que resultó con base en el algoritmo *Random Forest Classifier* obtuvo también un rendimiento óptimo para la predicción de sífilis gestacional; según F1-Score el rendimiento general del modelo en datos de desarrollo es del 68,38% superior en un 100% el desempeño general del modelo original basado en reglas, y del 61,1% en datos de testeo; el modelo presenta un AUC de 86,27%; sin embargo la curva de aprendizaje de estos modelos permitieron identificar sobre ajuste en el proceso de modelamiento y predicción.

Respecto a las características o variables que contribuyen a la predicción de sífilis congénita se identifica el número de semanas de gestación al momento del primer control prenatal, siendo aquellos valores más altos, o lo mismo, las gestantes que inician sus controles prenatales con una gestación avanzada tiene mayor riesgo de tener un bebé con la infección por sífilis (Figura 6); de otro lado, la edad del padre también se identifica como una variable que impulsa un pronóstico de sífilis congénita, en especial aquellos padres jóvenes. El número de controles prenatales esperados totales, la edad y educación de la madre, también son variables que contribuyen al pronóstico del diagnóstico de sífilis en los recién nacidos, así como la procedencia venezolana aumenta el riesgo de la situación en salud del recién nacido.

Figura 6. Variables representativas del modelo para predicción de sífilis congénita



DISCUSIÓN

La elaboración de modelos para predecir la sífilis congénita mediante técnicas de aprendizaje automático y con datos relacionados con la situación sociodemográfica y de salud de la madre, en este caso mujer gestante con diagnóstico de sífilis, resulta en un instrumento clínico y de salud pública para la detección temprana de factores de riesgo que pueden ser abordados o intervenidos a nivel individual y colectivo con el propósito de minimizar la incidencia de este situación de interés en salud pública. Al comparar los resultados de los modelos construidos y evaluar las métricas de desempeño se resalta que aquellos con base en algoritmos de aprendizaje automático presentan mejores resultados respecto al modelo base inicial (modelo basado en reglas) hasta del 100% en varios indicadores de desempeño, lo que indica que se cuenta con mejores instrumentos para identificar cada vez a más mujeres gestantes con la infección con mayor riesgo de transmitirla a su bebé al momento de nacer para intensificar las atenciones en salud y el seguimiento social.

Al comparar los modelos basados en técnicas de aprendizaje automático resulta favorable para el pronóstico de nuevos casos de sífilis congénita el modelo que resulta del algoritmo *K Neighbors Classifier* no sólo por que posee las mejores métricas de desempeño de predicción en conjunto y además de ser las mejores son buenas, en relación a los demás modelos, sino por que al predecir nuevos casos de bebés infectados con datos de testeo las métricas mejoran o se mantienen cercanas, la exactitud cambia de 90,8% a 90,1%; la sensibilidad cambia de 74,6% a 69,6%; la precisión cambia de 81,88% a 88,9% y el F1-Score pasa de 77,28% a 78,0% a diferencia de los demás modelos elaborados, en donde resulta ser la sensibilidad la métrica que se ve afectada entre los datos de entrenamiento y los datos de testeo en la predicción que generan los modelos. Se considera relevante mejorar la métrica de sensibilidad, dado que se de gran interés captar a las gestantes con diagnóstico de sífilis que

tienen mayor riesgo de transmitir la infección al bebé, sin embargo, la exactitud presenta un resultado favorable para la investigación y para la práctica en salud, lo que minimiza el costo de la inversión en intensificar acciones en salud con base en los resultados de la predicción del modelo con base en técnicas de aprendizaje automático. Con estos resultados es posible vincular al análisis de riesgo en salud instrumentos para la predicción de casos nuevos de sífilis a partir de los datos disponibles de la madre y su contexto para focalizar una acción que prevenga este contagio al bebé, que pasa tanto por aumentar la adherencia al tratamiento con penicilina benzatínica, hasta por la educación diferencial según edad, procedencia, nivel educativo y cercanía con los mismos servicios de salud, elementos que en la actualidad son débiles al momento de prestar un servicio de salud a una gestante con sífilis para gestionar el riesgo en salud del bebé.

CONCLUSIONES

Con el creciente número de casos nuevos de bebés con sífilis y las complicaciones que ello trae, se ha vuelto obligatorio desarrollar un sistema para su predicción de manera efectiva y precisa. La motivación de este trabajo fue buscar el algoritmo de aprendizaje automático más eficiente para la detección de la sífilis congénita. El resultado de este estudio indica que el algoritmo *K Neighbors Classifier* es el algoritmo más eficiente con un puntaje de desempeño general en el modelo de predicción de la sífilis congénita con un F1-Score del 77,28% y sensibilidad del 74,68%, métricas mejores que las de los trabajos de referencia, lo que permite implementar un sistema de predicción en producción en las instituciones de salud materno infantil para impactar cada vez más el riesgo de contagio de sífilis al bebé, instrumentos o técnicas que son poco utilizadas salud para tomar decisiones, dado que se demuestra que los mejores modelos presentan métricas de desempeño óptimas para avanzar en la gestión del riesgo en salud, especialmente la sensibilidad y la exactitud. En el futuro, el trabajo se puede mejorar vinculando otras variables en el conjunto de datos que, según la literatura, tiene un mayor potencial predictivo como lo es la reinfección e información acerca de sus modos de vida, e inclusive mayor cantidad de datos en comparación con el utilizado en este análisis, que puede proporcionar mejores resultados y ayudar a los profesionales de la salud a predecir la enfermedad de manera efectiva y eficiente.

REFERENCIAS BIBLIOGRÁFICAS

Adhikari, E. H. (2020). Syphilis in Pregnancy. *Obstetrics and Gynecology*, *135*(5), 1121–1135. https://doi.org/10.1097/AOG.0000000000003788

Ahsan, M., Akter Luna, S., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis : A Comprehensive Review. *Healthcare*, *10*(541), 1–30. https://doi.org/https://doi.org/10.3390/healthcare10030541

Alanazi, R. (2022). Identification and Prediction of Chronic Diseases Using Machine Learning Approach. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/2826127

Bao, Y., Medland, N. A., Fairley, C. K., Wu, J., Shang, X., Chow, E. P. F., Xu, X., Ge, Z., Zhuang, X., & Zhang, L. (2021). Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *Journal of Infection*, *82*(1), 48–59. https://doi.org/10.1016/j.jinf.2020.11.007

Fitzpatrick, F., Doherty, A., & Lacey, G. (2020). Using Artificial Intelligence in Infection Prevention. *Current Treatment Options in Infectious Diseases*, *12*(2), 135–144. https://doi.org/10.1007/s40506-020-00216-7

Forradellas, R. F. R., Alonso, S. L. N., Rodriguez, M. L., & Jorge-Vazquez, J. (2021). Applied machine learning in social sciences: Neural networks and crime prediction. *Social Sciences*, *10*(1), 1–20. https://doi.org/10.3390/socsci10010004

Gomathy, C. K. (2021). The prediction of disease using machine learning. *International Journal of Scientific Research in Engineering and Management (IJSREM)*, *05*(10).

Instituto Nacional de Salud. (2014). *Estructura de los registros para notificación de datos de eventos de interés en salud pública*.

Instituto Nacional de Salud. (2018). *Protocolo de vigilancia en salud publica: Sifilis Gestacional y congenita.* 1–17.

Instituto Nacional de Salud. (2022). *Protocolo de Vigilancia de Sífilis Gestacional y Congénita*. https://www.ins.gov.co/buscador-eventos/SitePages/Evento.aspx?Event=55

Jeyaganesan, J., Sathiya, A., Keerthana, S., & Aiyer, A. (2020). Diagnosis And Prediction Of Heart Disease Using Machine Learning Techniques. *Ilkogretim Online - Elementary Education Online*, *19*(2), 1817–1827. https://doi.org/10.17051/ilkonline.2020.02.696765

Lakshmanan, V., Robinson, S., & Michael, M. (2021). *Machine Learning Design Patterns*.

OMS. (2019). *Detección y tratamiento de la sífilis en embarazadas*.

OMS. (2022). *Sífilis - OPS/OMS | Organización Panamericana de la Salud*. https://www.paho.org/es/temas/sifilis

Secretaria Distrital de Salud. (2022). *Sífilis congénita 2008-2022ps | SALUDATA Bogotá*. https://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/salud-sexual-y-reproductiva/sifilis-congenita/

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. https://doi.org/10.1186/s12911-019-1004-8