

# “Implementación de un clasificador para asistir el diagnóstico clínico de Lupus Eritematoso Sistémico en pacientes mexicanos”

Carlos Díaz Montejo<sup>1</sup>, Carlos Alberto Reyes García<sup>2</sup>, Donají Domínguez Zúñiga<sup>3</sup>.

<sup>1,2</sup> Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla Pue. 7200

<sup>3</sup> Universidad Autónoma del Estado de Morelos, Cuernavaca Mor. 62209

E-mail: ph\_carlos@live.com

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

## RESUMEN

El lupus es una enfermedad autoinmune de la cual se desconocen tanto su etiología como su cura. Se estima que aproximadamente 1.5 millones de personas en México la padecen, pero debido a que el lupus es una enfermedad heterogénea y que esta puede simular a otras enfermedades durante su etapa temprana, se obtienen dificultades para generar un diagnóstico médico acertado sin una gran cantidad de estudios clínicos; además, existen 4 tipos de lupus siendo el eritematoso sistémico (LES) el más común, complicando aún más la diagnosis. Lo anterior deriva en grandes costos de tiempo y capital para el diagnóstico diferencial de esta enfermedad. Durante los últimos años se ha incrementado el uso de métodos analíticos y automáticos que permiten hacer inferencias sobre la enfermedad de un paciente a partir de los registros médicos electrónicos. Entre los métodos más destacados se encuentra la implementación de clasificadores usando aprendizaje máquina. En este trabajo se implementaron 8 tipos de clasificadores sobre una base de datos de pacientes mexicanos que considera las variables clínicas empleadas por el criterio de la Liga Europea Contra el Reumatismo y el Colegio Americano de Reumatología en 2019 (EULAR/ACR-2019). De los clasificadores generados cinco otorgan resultados comparables, en cuestión de sensitivity (91.6%) y specificity (93.4%), con el criterio EULAR/ACR-2019.

## INTRODUCCIÓN

En 2017 un reporte publicado por CEVECE [1] determinó que aproximadamente 1.5 millones de mexicanos padecían lupus eritematoso sistémico (LES), pero debido al alto costo de las pruebas, su heterogeneidad, complejidad y a que suele confundirse con otras enfermedades se estima que pudieron existir casos no registrados. Debido a lo anterior en 2019 el Senado de la República declaró al lupus como una enfermedad causante de gastos catastróficos [2] por lo que se invitó al Consejo de Salubridad General a realizar nuevas investigaciones sobre esta enfermedad con el objetivo de reducir costos. En 2018 Murray [3] reportó obtener un clasificador para identificar individuos con LES en Estados Unidos, su clasificador alcanzó el 83% de sensitivity usando el criterio ACR-1997, aunque el trabajo de Murray alcanza un rendimiento aceptable, este no se compara con el nuevo criterio EULAR/ACR-2019 [4]; por otro lado, el estudio de Atisha, Ocampo y Llorente [5] indica que en América Latina existen diferencias entre las medidas de las variables clínicas usadas para el diagnóstico de Lupus, por lo tanto debe diagnosticarse lupus considerando medidas diferentes de los síntomas. En este trabajo se reporta la implementación de clasificadores usando una base de datos con 520 pacientes mexicanos, 245 diagnosticados con lupus y 255 descartados, generada con apoyo de la Fundación Proayuda Lupus Morelos A.C y considerando las variables del criterio EULAR/ACR-2019. Se entrenaron 8 tipos de clasificadores obteniendo que algunos clasificadores cuentan con una sensibilidad y especificidad comparables a las alcanzadas por el criterio EULAR/ACR-2019.

## MARCO TEÓRICO.

Actualmente en México se usa el criterio EULAR/ACR-2019 para la clasificación de personas con LES. Este considera como entrada una prueba ANA de 1:80 y la suma ponderada de los síntomas mostrados en la tabla 1; se requiere además: alcanzar 10 puntos, considerar al menos un criterio clínico, contar solo el criterio más alto en cada categoría y que cada síntoma considerado no tenga una explicación más razonable.

Tabla 1. Síntomas empleados por el criterio EULAR/ACR-2019 para clasificar a pacientes con LES.

Dominios clínicos	Peso	Dominios clínicos	Peso
<b>Constitucional</b>		<b>Mucocutáneo</b>	
Fiebre	2	Alopecia no cicatricial	2
<b>Hematológico</b>		Úlceras bucales	2
Leucopenia	3	Lupus discoide o cutáneo subagudo	4
Trombocitopenia	4	Lupus cutáneo agudo	6
Anemia hemolítica autoinmunitaria	4	<b>Serológico</b>	
<b>Musculoesquelético</b>		Derrame pleural o pericárdico	5
Compromiso articular	6	Pericarditis aguda	6
<b>Neuropsiquiátrico</b>		<b>Renal</b>	
Desvarío	2	Proteinuria > 0.5g/24h	4
Psicosis	3	Nefritis lúpica clase II o V	8
Convulsiones	5	Nefritis lúpica clase III o IV	10
<b>Dominios Inmunológicos</b>			
<b>Anticuerpos antifosfolípidos</b>			
Anticuerpos anti-cardiolipina o Anti-β2GPI o Anticogulante lúpico	2		
<b>Proteínas del complemento</b>			
Bajo C3 o Bajo C4	3		
Bajo C3 y Bajo C4	4		
<b>Anticuerpos específicos de lupus</b>			
Anti-ADNdc o Anti-Sm	6		

El proceso de clasificación se trata de un pronóstico basado en la información disponible [6], para esto es necesario diseñar modelos que aprendan de tal información y que aseguren además el poder trabajar con nuevos datos.

A partir del teorema de Bayes mostrado en la ecuación (1) es posible estimar la probabilidad de que una persona tenga una enfermedad a partir de sus síntomas.

$$p(enfermo|síntomas) = \frac{p(síntomas|enfermo)p(enfermo)}{p(síntomas)} \dots(1)$$

Si se aplica el supuesto de que los síntomas son independientes y que  $p(síntomas)$  es una constante para que la regla de clasificación sea la ecuación (2) entonces es posible usar una estimación de probabilidad máxima a posteriori para calcular  $p(enfermo)$  y  $p(síntoma_i|enfermo)$ . El resultado es que el primer término de la ecuación (2) es la frecuencia relativa de la clase “enfermo” y el segundo es algún tipo de distribución.

$$enfermo = \underset{enfermo}{\operatorname{argmax}} \left( P(enfermo) \prod_{i=1}^n P(síntoma_i|enfermo) \right) \dots(2)$$

Al algoritmo mostrado en la ecuación (2) se le conoce como Naïve Bayes. Si se propone que la distribución de  $p(síntoma_i|enfermo)$  sea Gausiano, como en la ecuación (3), entonces toma el nombre de Naïve Bayes Gausiano. Por otro lado, si se propone que  $p(síntoma_i|enfermo)$  sigue una distribución tipo Bernoulli multivariado, como en la ecuación (4), entonces toma el nombre de Naïve Bayes Bernoulli.

$$P(síntoma_i | enfermo) = \frac{1}{\sqrt{2\pi\sigma_{enfermo}^2}} \exp\left(-\frac{(síntoma_i - \mu_{enfermo})^2}{2\sigma_{enfermo}^2}\right) \dots(3)$$

$$P(síntoma_i | enfermo) = P(i | enfermo)síntoma_i + (1 - P(i | enfermo)) \dots(4)$$

El clasificador Random Forest es un ensemble de árboles de decisión donde cada árbol vota por una clase (enfermo o sano) basándose en los síntomas que toma como vector de características, así el resultado presentado por el bosque aleatorio será aquella clase que obtenga más votos. La figura 1 muestra un ejemplo de este clasificador.

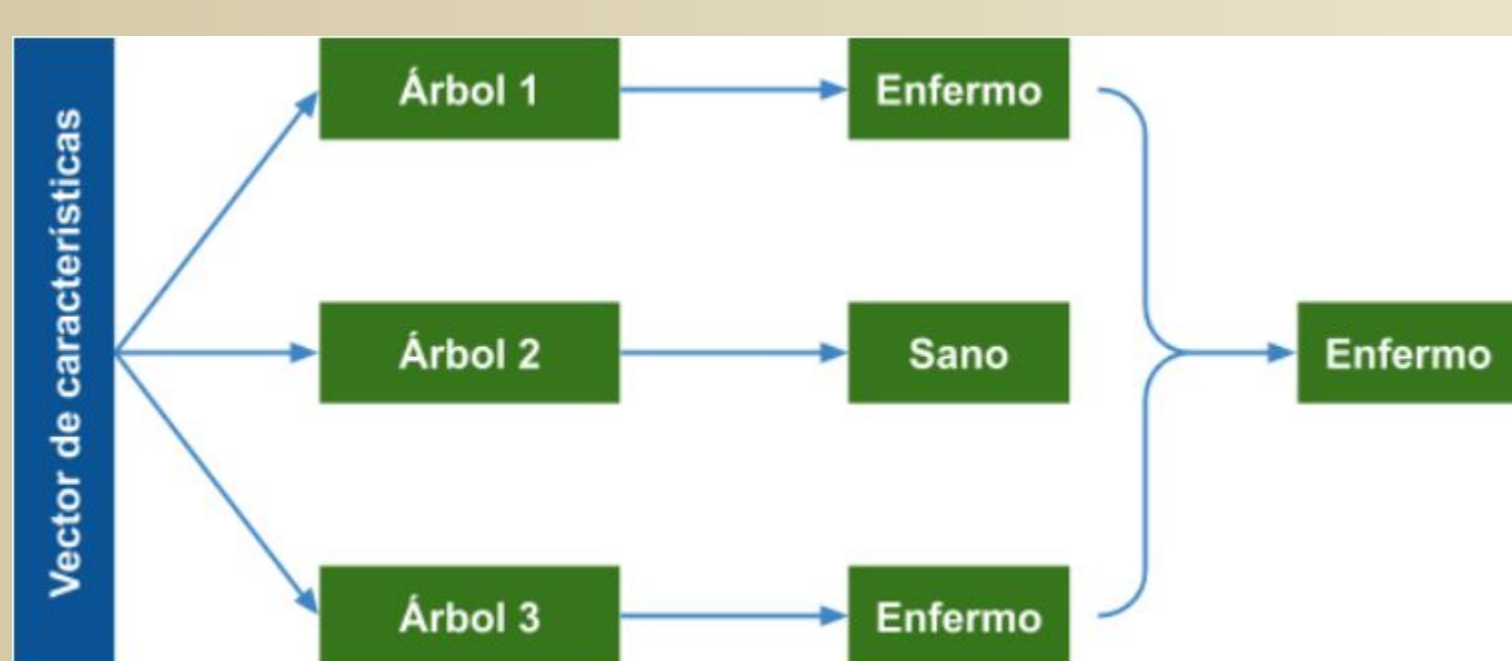


Figura 1. Ejemplo de Random Forest con tres árboles tomando como resultado final que una persona está enferma.

Support vector machine es un clasificador que toma ejemplos del conjunto con el que entrena para crear hiperplanos que ayuden a distinguir entre las clases que clasifica. Para realizar lo anterior requiere de una función llamada función kernel. La figura 2 muestra un ejemplo de este tipo de clasificadores.

Quando se entrenan modelos clasificadores es necesario verificar su rendimiento usando métricas evaluadoras para evitar que el clasificador entrenado solo sea funcional con los datos que le fueron presentados. Entre las herramientas que se pueden emplear para realizar lo anterior están las matrices de confusión a partir de las cuales se obtienen las medidas de Accuracy, Precision, Sensitivity y Specificity. En aprendizaje supervisado esta herramienta compara la etiqueta que predice el clasificador contra la etiqueta real de cada vector de características. Cuando se tiene un problema binario, la matriz es de 2x2 cuyos elementos en la diagonal es el número de vectores clasificados correctamente como Sanos y como Enfermos, la figura 3 muestra la forma de una matriz de confusión. Note que las entradas (1,2) y (2,1) son Falso Enfermo y Falso Sano, respectivamente.

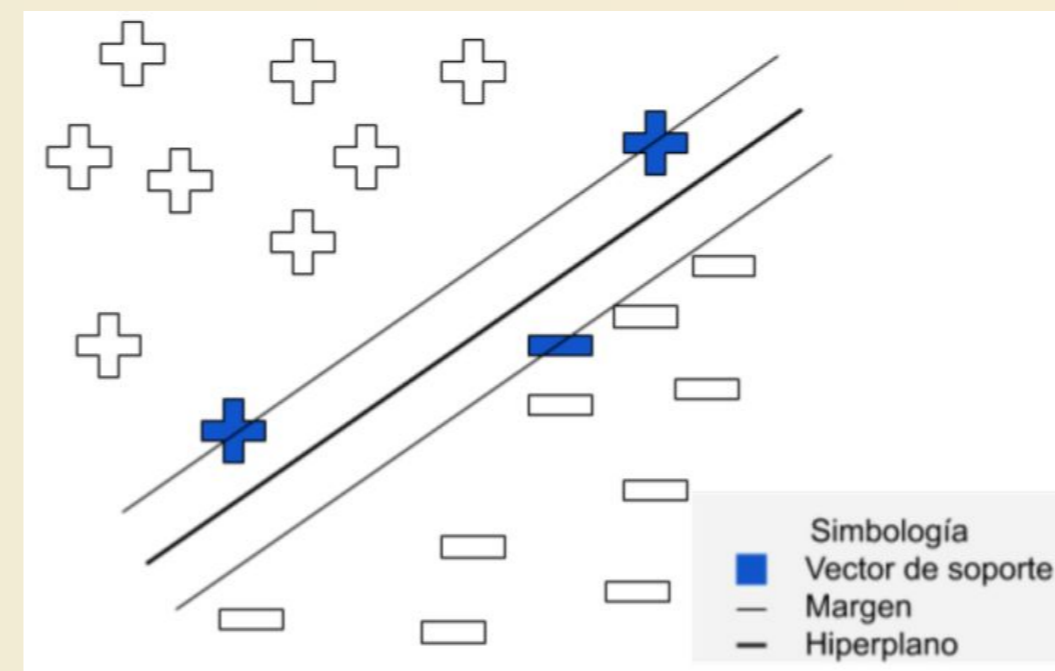


Figura 2. Hiperplano que separa dos clases (positivo y negativo) formado por SVM con kernel lineal

Etiqueta ↓	Clasificado en →	
	Sano	Enfermo
Sano	VS	FE
Enfermo	FS	VE

Figura 3. Matriz de confusión para clasificación binaria Enfermo/Sano.

## ARREGLO EXPERIMENTAL

Para el preprocesamiento, clasificación y evaluación de los modelos se ha hecho uso de Jupyter Notebook 6.0.3, Python™3.7.6 así como de las bibliotecas Scikit-learn 0.22.1, Pandas 1.0.1, Matplotlib 3.1.3 y Numpy 1.18.1.

Los experimentos fueron realizados usando una base de datos generada a partir de la entrevista vía formulario a pacientes con LES confirmado y pacientes que fueron descartados. En total se ha obtenido una base de datos con 245 personas con lupus y 255 sujetos descartados, se ha usado el 70% de la población total para entrenar los siguientes 8 modelos clasificadores: 1 Naïve Bayes Gausiano; 1 Naïve Bayes Bernoulli; 3 Random Forest con diferente cantidad de árboles así como diferente cantidad de ramas; 3 tipos de SVM con kernel RBF, variando un parámetro de corrección C y datos modificados para presentar distribución normal estándar.

Se ha usado como etiqueta a la variable “Lupus” la cual tiene las clases “Sí” y “Fui descartado” mientras que los nombres de cada entrada del vector se presentan en el orden siguiente: Leucopenia, Trombocitopenia, Nefritis\_II\_V, Nefritis\_III\_IV, AntiCar, AntiB2GPI, AntiCoa\_Lup, C3, C4, AntiADNDC, AntiSm, Fiebre, Anemia\_hemol, Desvario, Psicosis, Convulsiones, Alopecia\_no\_cica, Ulceras\_bucales, Lupus\_discoide, Lupus\_cut\_agu, Derrame\_pleu card, Pericarditis, Comp\_artic, Proteinuria.

## RESULTADOS

Las figuras 4 a) h) muestran las matrices de confusión obtenidas al evaluar cada clasificador y a su vez la tabla 2 indica los porcentajes alcanzados por cada modelo en las métricas de Accuracy, Precision, Sensitivity y Specificity.

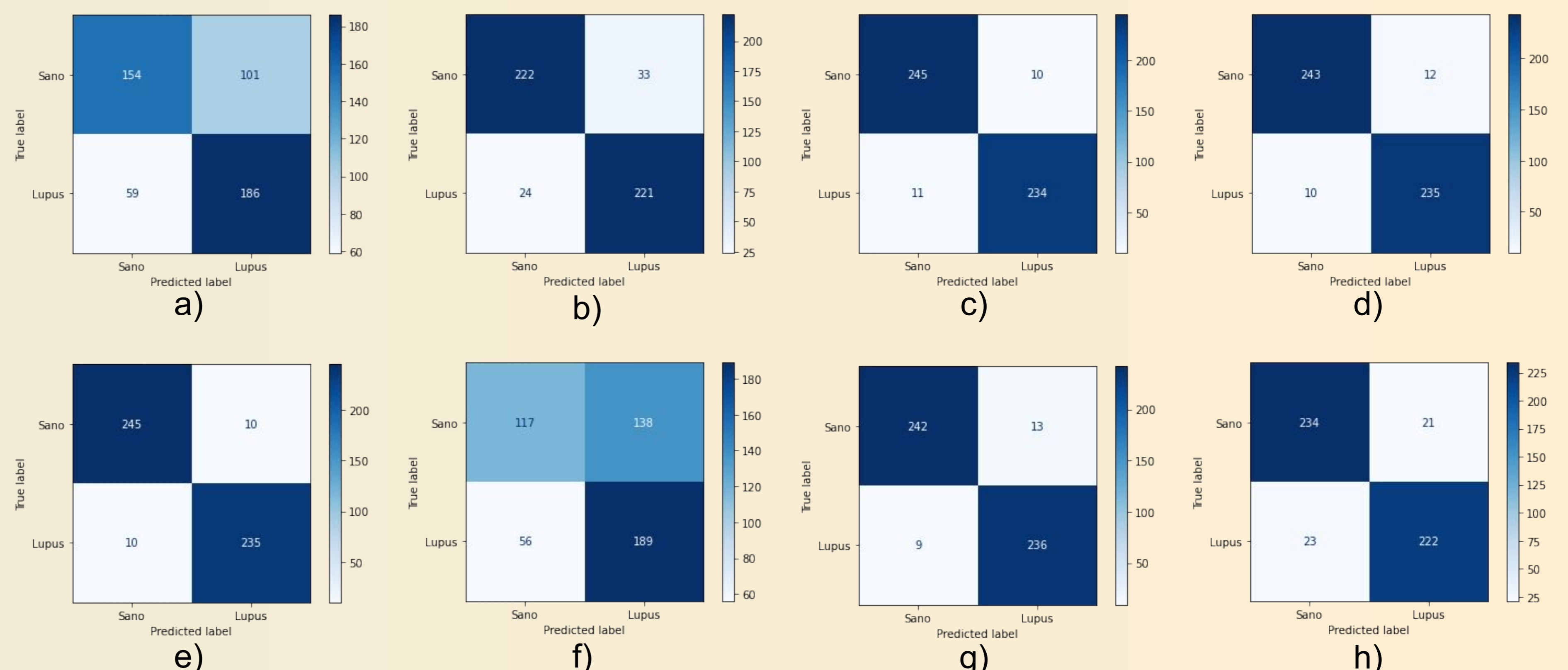


Figura 4. Matriz de confusión obtenida del entrenamiento del clasificador a) Naïve Bayes Gausiano, b) Naïve Bayes Bernoulli, c) Random forest con 100 árboles y 24 ramas por árbol, d) Random forest con 200 árboles y 50 ramas por árbol, e) Random forest con 150 árboles y sin límite de ramas, f) SVM con C=1.0, g) SVM con C=1.0 y distribución normal estándar, h) SVM con C=0.1 y distribución normal estándar.

Clasificador	Resultados			
	accuracy (%)	precision (%)	sensitivity (%)	specificity (%)
Naïve Bayes Gausiano	68.0	64.8	75.9	60.4
Naïve Bayes Bernoulli	88.6	87.0	90.2	87.0
Random Forest 1	95.8	95.9	95.5	96.0
Random Forest 2	95.6	95.1	95.9	95.2
Random Forest F	96.0	95.9	95.9	96.0
SVM-C	61.2	57.8	77.1	45.9
SVM-C Escalado	95.6	94.8	96.3	94.9
SVM-C2 Escalado	91.2	91.4	90.6	91.8

Tabla 2 (Izquierda). Resultados obtenidos por cada clasificador. Random Forest 1=Random forest con 100 árboles y 24 ramas por árbol; Random Forest 2=Random forest con 200 árboles y 50 ramas por árbol; Random Forest F=Random forest con 150 árboles y sin límite de ramas; SVM-C=SVM con C=1.0; SVM-C Escalado=SVM con C=1.0 y distribución normal estándar; SVM-C2 Escalado=SVM con C=0.1 y distribución normal estándar.

## CONCLUSIONES

De los resultados obtenidos, se debe notar que aunque el criterio EULAR/ACR-2019 es un criterio de clasificación y no de diagnóstico, el considerar medir los síntomas de solo pacientes mexicanos, indica que implementar esta clase de modelos permite desarrollar un estándar de plata. De los modelos clasificadores presentados se observa que el mejor tipo de clasificador es el random forest. Más aún el Random Forest con 150 árboles y sin límite de ramas por árbol alcanza el mejor rendimiento en la mayoría de las métricas además sus medidas de sensitivity y specificity son comparables al criterio EULAR/ACR-2019, superando al trabajo publicado por Murray. Se planea aumentar la base de datos para seguir entrenando los clasificadores buscando observar si es posible mejorar a los que obtuvieron rendimientos bajos. De igual manera se busca poder aplicar estos modelos en prácticas de reumatología, donde puedan compararse las predicciones con la opinión médica con el fin de que sirvan de apoyo en el diagnóstico médico.

## REFERENCIAS

- [1] S. de Salud, “Lo que debes saber del lupus,” tech. rep., Centro Estatal de Vigilancia Epidemiológica y Control de Enfermedades, Estado de Mexico, Mexico, 2017.
- [2] “DEFINIR Y PRIORIZAR AL LUPUS COMO UNA ENFERMEDAD CAUSANTE DE GASTOS CATASTRÓFICOS,” Gaceta Parlamentaria del Senado de la República de México, 2019.
- [3] S. G. Murray, A. Avati, G. Schmajak, J. Yazdany, “Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling,” Journal of the American Medical Informatics Association, vol. 26, pp. 61–65, Jan. 2019.
- [4] M. . Aringer, et al, “2019 european league against rheumatism/american college of rheumatology classification criteria for systemic lupus erythematosus,” Arthritis & Rheumatology, vol. 71, pp. 1400–1412, Sept. 2019.
- [5] Atisha-Fregoso, Y., Jakez-Ocampo, J., & Llorente, L. “Systemic lupus erythematosus in Hispanics”. Autoimmunity, Vol 44, pp. 555-561, 2011.
- [6] A. A. Torres García, Clasificación de palabras no pronunciadas presentes en Electroencefalogramas (EEG). PhD thesis, INAOE, Puebla, Mexico, 2011.