Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos

Rodolfo Mosquera^{(1)*}, Omar D. Castrillón⁽¹⁾ y Liliana Parra⁽²⁾

(1) Universidad Nacional de Colombia, Facultad de Ingeniería y Arquitectura, Bloque Q, Campus La Nubia, Manizales-Colombia. (E-mail: rmosqueran@unal.edu.co, odcastrillong@unal.edu.co)

(2) Universidad Libre, Seccional Bogotá, Grupo de Investigación Esculapio, Docente Investigador, Centro de Investigaciones Socio jurídicas, Colombia.

Recibido Abr. 11, 2018; Aceptado Jun. 4, 2018; Versión final Jul. 9, 2018, Publicado Dic. 2018

Resumen

En este artículo se presenta una nueva metodología basada en la aplicación de algoritmos Máquina de Soporte Vectorial, Naïve Bayes y Algoritmos Genéticos en diagnósticos de evaluaciones psicosociales para la identificación del grado de riesgo psicosocial en docentes de colegios públicos en Colombia. Se realizó un estudio comparativo del modelo de aprendizaje de máquina para la predicción: Máquinas de Soporte Vectorial (SVM) y Naïve Bayes, en dos etapas, primero con todas las variables y segundo, reduciendo la dimensionalidad de los datos aplicando: algoritmos genéticos. Se seleccionaron las cuarenta mejores variables con mejor eficiencia en la precisión de la predicción. La base de datos utilizada se componía de 3000 registros epidemiológicos, que correspondían a docentes de colegios públicos del área metropolitana de una ciudad colombiana. El uso de SVM reconoce fácilmente variables de tipo fisiológico y el mejor desempeño de predicción se obtuvo con una eficiencia en la precisión del 96,3%.

Palabras clave: aprendizaje de máquina; máquinas de soporte vectorial; Naïve Bayes; algoritmos genéticos

Support Vector Machines, Naïve Bayes Classifier and Genetic Algorithms for the Prediction of Psychosocial Risks in Teachers of Colombian Public Schools.

Abstract

This paper presents a new methodology based on the application of Support Vector Machine algorithms, Naïve Bayes and Genetic Algorithms in diagnostics of psychosocial evaluations for the identification and prediction of the psychosocial risk level of public-school teachers in Colombia. A comparative study of the model of machine learning for prediction was carried out: Support Vector Machines (SVM) and Naïve Bayes, in two stages, first with all the variables and second, reducing the dimensionality of the database applying genetic algorithms. The best forty variables with the best efficiency in prediction accuracy were selected. The database used consisted of 3000 epidemiological records, which corresponded to teachers from public schools in the metropolitan area of a Colombian city. The use of SVM easily detected variables of physiological type and the best prediction performance was obtained with accuracy of 96.3%.

Keywords: machine learning; support vector machine; Naïve Bayes; genetic algorithm

INTRODUCCIÓN

Según lo describe la Organización Internacional del Trabajo (OIT), las enfermedades derivadas del estrés relacionadas con el trabajo se desarrollan debido a la falta de capacidad que pueden tener los trabajadores para enfrentar las condiciones y demandas que se derivan de la exigencia de la organización del trabajo, de las relaciones en el trabajo, de las condiciones ambientales que se transforman en factores de riesgo por la mala adecuación que las organizaciones les dan (ILO, 2016). Los sistemas integrados de gestión pretenden mejorar cada día estas situaciones que son inherentes al trabajo y motivo de preocupación (Soler et al., 2014) y en busca de mejoramiento continuo se hace necesario el innovar con técnicas en áreas que aún no son exploradas y que contribuyan a la toma de decisiones estratégicas, como las técnicas de machine learning o aprendizaje de máquina (Mosquera et al., 2016). En los sistemas de gestión de la seguridad y salud en el trabajo es preponderante realizar los respectivos seguimientos y controles de proceso a que haya lugar en cualquier tipo de industria y organización cuya relación sea directa (Arias 2014; Saizarbitoria et al., 2014). Es decir, el monitoreo constante de absolutamente todos los procesos organizacionales y de organización del trabajo se hacen necesarios, fundamentalmente donde existen demasiadas variables que tienen relación directa en el desempeño de los trabajadores (Huerta et al., 2016), como es el caso de los docentes de colegios públicos.

Tzeng et al., (2004), en un proyecto sobre la intencionalidad de las enfermeras de abandono del puesto de trabajo en Taiwan, encontró en la base de datos de MEDLINE que previamente nadie había trabajado la aplicación de máquinas de soporte vectorial (SVM) en el área de gestión de recursos humanos. Se formuló la hipótesis de que este enfoque podría ser útil para establecer un mecanismo de alerta temprana para la administración de recursos de enfermería, utilizando un nuevo método llamado “support vector learning machine”. Se sabe con claridad que las máquinas de soporte vectorial pueden establecer una relación no lineal entre la (s) variable (s) independiente (s) y la (s) variable (s) dependiente (s) y pronosticar el (los) valor (es) de la (s) variable (s) dependiente (s). Las máquinas de vectores de soporte se pueden equiparar a los análisis estadísticos de regresión lineal o no lineal comúnmente utilizados, como la regresión múltiple, pero son mucho más robustos, la calidad de las máquinas de soporte vectorial se determina con el entrenamiento y la capacidad de generalización (Cristianini et al., 2000). Yasodha et al., (2012), tomaron factores tales como la experiencia humana, el conocimiento y la capacidad de juicio y aplicaron técnicas de minería de datos utilizando un método híbrido entre coeficiente de contingencia de atributo de clase (CACC) y máquinas de soporte vectorial (SVM) para la clasificación de datos para la gestión y selección de recursos humanos para determinados puestos laborales.

Es por eso que la aplicación de técnicas de aprendizaje de máquina (machine learning) como los algoritmos máquina de soporte vectorial (SVM) (Cortes et al., 1995; Vapnik 1998), el algoritmo Naïve Bayes (Chandra et al., 2007), se utilizan como herramienta para la predicción de afecciones de índole psicosocial, con el fin de evaluar su eficiencia en el grado de precisión con respecto a la identificación del tipo de grado de riesgo y mejorar los procesos de administración y gestión de recursos humanos para la prevención de riesgos. Para la reducción de la dimensionalidad se utilizan los algoritmos genéticos, los cuales ya han sido probados como técnicas de inteligencia artificial (Coca et al., 2013) en otros procesos para optimizar diferentes objetivos. El desarrollo y validación de nuevos instrumentos desde la rama de la ingeniería que contribuyan a la detección de riesgos psicosociales en las organizaciones de toda índole es posible gracias a los patrones que facilitan la clasificación de las variables (atributos) que denotan un resultado (clase), actualmente se trabaja en el desarrollo de nuevos algoritmos que faciliten el trabajo para la administración y la gerencia de seguridad y salud en el trabajo en todo tipo de actividades (Mosquera et al., 2016; Mosquera et al., 2018).

En este artículo se presentan los pasos para la aplicación de los algoritmos Naïve Bayes y Maquinas de Soporte Vectorial (SVM) a la Batería para la evaluación de factores de riesgo psicosocial intralaboral (Villalobos et al., 2010, Anexo 1) más las variables fisiológicas evaluadas (electroencefalografía, electrocardiograma, frecuencia cardíaca, actividad eléctrica de la piel y electromiografía) en primera instancia, posteriormente, a estas mismas variables para reducir el error de clasificación a los atributos, se le redujo la dimensionalidad de los datos a través de la aplicación de algoritmos genéticos para observar el grado de precisión de los algoritmos para la predicción del grado de riesgo psicosocial en docentes de colegios públicos de una ciudad de Colombia. En este artículo se presentan los pasos para obtener un modelo cuantitativo con el cual se pueda evaluar y predecir el grado de riesgo psicosocial aplicando técnicas de aprendizaje de máquina adicionando variables fisiológicas al modelo existente.

MATERIALES Y MÉTODOS

Para el desarrollo del artículo se presenta la información concerniente a la base de datos a intervenir, las variables que la componen, la descripción de los modelos de clasificación Naïve Bayes, máquinas de soporte vectorial (SVM) y el método de reducción de la dimensionalidad de los datos concernientes a este estudio:

algoritmo genético (Whitley 1994), dejando como variables a intervenir los cuarenta atributos de mayor impacto para la correcta clasificación y con menor tasa de error. La metodología descrita que se utilizó para el desarrollo se muestra en la figura 1. El conjunto de datos que se analizó proviene de datos de docentes, información de variables sociodemográficas, físicas, psicosociales, ambientales y fisiológicas, como se muestra en la tabla 1, se verifican los campos con la información correcta en la base de datos de 3000 registros. Los antecedentes de los algoritmos a desarrollar en la investigación, se describen en los que sigue.

Naïve Bayes

El algoritmo clasificador Naïve-Bayes (NBC), es un clasificador probabilístico simple con fuerte suposición de independencia. Aunque la suposición de la independencia de los atributos es generalmente una suposición pobre y se viola a menudo para los conjuntos de datos verdaderos. A menudo proporciona una mejor precisión de clasificación en conjuntos de datos en tiempo real que cualquier otro clasificador. También requiere una pequeña cantidad de datos de entrenamiento. El clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior. También es útil para datos dimensionales altos ya que la probabilidad de cada atributo se estima independientemente (Chandra et al., 2007).

Fig. 1. Visión general de la metodología propuesta.

Tabla 1: Variables de los dominios de factores de riesgo psicosocial intralaboral. (Adaptado de: Batería para la evaluación de factores de riesgo psicosocial. Villalobos et al., 2010, Anexo 1)

Ítems	Dominios	Variables
S ₁	Sociodemográficas	S ₁ ... S ₁₀
D ₁	Demandas del trabajo	D ₁ ... D ₅₀
W ₁	Control sobre el trabajo	W ₁ ... W ₂₁
L ₁	Liderazgo y relaciones sociales en el lugar de trabajo (Workplace)	L ₁ ... L ₃₂
R ₁	Recompensas	R ₁ ... R ₁₁
P ₁	V. Fisiológicas	P ₁ ... P ₅

El clasificador se denota de la siguiente manera: Sea C la variable aleatoria que denota la clase de una instancia que $X = \langle X_1, X_2, \dots, X_m \rangle$ un vector de variables aleatorias que denote los valores de los atributos observados. Sea c representando una etiqueta de clase particular $x = \langle X_1, X_2, \dots, X_m \rangle$ representan un vector

de valor de atributo observado particular. Para predecir la clase de una instancia de prueba x , el teorema de Bayes se utiliza para calcular la probabilidad (Chandra et al., 2007, p.12):

$$p(C = c | X = x) = \frac{p(C = c)p(X=x | C = c)}{p(X=x)} \quad (1)$$

Luego, predecir la clase de instancia de prueba con mayor probabilidad. Aquí $X = x$ representa el evento que $X_1 = x_1 \wedge X_2 = x_2 \wedge \dots \wedge X_m = x_m$. * $P(X = x)$ puede ser ignorada ya que es invariante a través de las clases, entonces la ecuación (1) se convierte en:

$$p(C = c | X = x) \propto p(C = c)p(X = x | C = c) \quad (2)$$

" $p(C = c)$ y $p(X = x | C = c)$ " se calculan a partir de los datos de entrenamiento. Como los atributos $X_1, X_2, \dots, X_m$ son condicionalmente independientes entre sí para una clase determinada, entonces la ecuación (2) se convierte en:

$$p(C = c | X = x) \propto p(C = c) \prod_{i=1}^m p(X_i = x_i | C = c) \quad (3)$$

Que es simple de calcular para instancias de prueba y estimar a partir de datos de entrenamiento. Los clasificadores que utilizan la ecuación (3) se llaman clasificador Naïve Bayes. El algoritmo Naïve Bayes puede manejar atributos categóricos y numéricos. Así, para un atributo numérico de una instancia de prueba $x <X_1, X_2, X_m>$, la probabilidad se calcula mediante la ecuación (Chandra et al., 2007, p.13):

$$p(C = c | X = x) \propto p(C = c) \prod_{i=1}^m p(a_i < x_i \leq b_i | C = c) \quad (4)$$

Para reducir el error de clasificación de datos ruidosos y superar los problemas de la estimación de probabilidades en el clasificador Naïve Bayes, un enfoque simple novedoso basado en la máxima ocurrencia del número para el cual las probabilidades condicionales de cualquiera de los atributos son cero para una instancia dada, ha sido propuesto por Chandra et al., (2007) como sigue a continuación.

Para la instancia de prueba dada $X <X_1, X_2, \dots, X_m>$, si $p(C = c | X = x)$ para cada clase es cero, entonces para cada clase cuenta las ocurrencias de valores de atributos (digamos, n_i) para los cuales $p(a_i < x_i \leq b_i | C = c) = 0$. Aquí, n_i significa el número de atributos para los cuales la instancia de entrenamiento con $X_i = x_i \wedge C = c$ no está presente en los datos de entrenamiento. Cuanto mayor es el número n_i de una clase c , menor es la probabilidad de que la instancia de prueba x pertenezca a esa clase c . N_i depende también de la probabilidad de esa clase en los datos de entrenamiento. Por lo tanto, en lugar de tomar n_i como el número significativo en la decisión de la clase de tal instancia de prueba, se calcula para cada atributo (Chandra et al., 2007, p.14):

$$N_i = n_i / p(C = c) \quad (5)$$

Máquina de Soporte Vectorial (SVM)

La clasificación multi-clase está presente en muchos problemas del mundo real, inicialmente las máquinas de soporte vectorial fueron diseñadas para tratar con problemas binarios ± 1 . Ahora veremos cómo tratar este problema. Podría decirse que el algoritmo multi-clase más elegante, y sin duda el método más alineado con el principio de Vapnik de tratar siempre de resolver problemas directamente, implica la modificación de la función objetivo SVM de tal manera que, simultáneamente permite el cálculo de un clasificador multi-clase (smola et al., 1998, p.233) podemos usar la siguiente configuración cuadrática:

$$\text{minimizar}_{w_r \in H, \epsilon^r \in R^m, b_r \in R} \frac{1}{2} \sum_{r=1}^M \|w_r\|^2 + \frac{C}{m} \sum_{i=1}^m \sum_{r \neq y_1} \epsilon_i^r, \quad (6)$$

Sujeto a:

$$\langle W_{y_i}, X_i \rangle + b_{y_i} \geq \langle W_r, X_i \rangle + b_r + 2 - \epsilon_i^r, \epsilon_i^r \geq 0, \quad (7)$$

Donde:

$m \in \{1, \dots, M\} \setminus Y_i$ y $Y_i \in \{1, \dots, M\}$ es la etiqueta multi-clase del patrón X_i .

En términos de precisión, los resultados obtenidos con esta aproximación son comparables a los obtenidos directamente usando el método uno contra el resto. Para problemas prácticos, la elección del enfoque dependerá de las limitaciones disponibles, los factores relevantes incluyen la precisión requerida, el tiempo disponible para el desarrollo, el tiempo de procesamiento y la naturaleza del problema de clasificación (Smola et al., 1998, p.234).

Algoritmos genéticos

Los algoritmos evolutivos son técnicas de optimización metaheurística inspiradas en la evolución biológica. Se mantiene una población de soluciones candidatas en cada generación, y cada solución candidata se codifica en un espacio apropiado para aplicar operadores como selección, reproducción y mutación. Se define una función de fitness para medir la calidad de los individuos (Whitley, 1994). En nuestro caso, esta función lo que hace es entrenar clasificadores usando las 40 características contenidas en los padres y la información de las muestras que se extrae de la base de datos. El método de entrenamiento es el de k vecinos más cercanos y para entrenar se usa solo el 60% de las muestras (Mosquera et al., 2018):

$$\text{Fitness}_{G.\text{Riesgo}} = \min \left\{ \sum_{i=1}^N P_{ij} \right\} \quad (8)$$

Dónde " i " representa cada variable, " N " representa el número total de variables con sus respectivos valores y " P_{ij} " hace referencia a la eficiencia de cada variable i con respecto al grado de riesgo j . Las subpoblaciones corresponden a un conjunto de 10 individuos, donde cada individuo representa un conjunto de 40 características. Entonces:

$$\text{Individuo}_i = \{\text{conjunto de 40 variables de las 123}\}, (i=1,2,3,\dots,10) \quad (9)$$

Paso 3. Evaluación con las técnicas de aprendizaje de máquina

Se realizó nuevamente la evaluación a los datos con las 40 variables seleccionadas con los dos algoritmos de aprendizaje de máquina, Naïve Bayes (NBC) y máquinas de soporte vectorial (SVM), para obtener los resultados de precisión, sensibilidad, especificidad y tasa de error de cada uno.

$$\text{Sensibilidad} = \frac{TP}{TP + FN} 100\% \quad (10)$$

$$\text{Especificidad} = \frac{TN}{TN + FP} 100\% \quad (11)$$

$$\text{Precisión} = \frac{TP + TN}{TP + TN + FP + FN} 100\% \quad (12)$$

Donde TP (Verdaderos Positivos) = Casos positivos correctamente clasificados; TN (Verdaderos Negativos) = Casos negativos correctamente clasificados; FP (Falsos Positivos) = Casos negativos incorrectamente clasificados; FN (Falsos Negativos) = Casos positivos incorrectamente clasificados; y Sensibilidad (Fracción positiva verdadera) es la probabilidad de que una prueba de diagnóstico sea positiva, dado que la persona tiene la afección. La especificidad (fracción negativa verdadera) es la probabilidad de que una prueba de diagnóstico sea negativa, dado que la persona no tiene la enfermedad. La precisión es la probabilidad de que una prueba de diagnóstico se realice correctamente.

Paso 4. Análisis de estabilidad

Con la finalidad de establecer la estabilidad del sistema en los resultados que se obtuvieron, se usó un análisis de varianza (ecuación 13) para establecer si existen diferencias significativas en el valor de los resultados. En este análisis, Y_i representa la respuesta a las variables; T_i es el efecto causado por el tratamiento i -ésimo y ϵ_i el i -ésimo error experimental. La información obtenida cumple los requerimientos de independencia y normalidad requeridos. El análisis de varianza se realizó en un intervalo de confianza del 99.5% (Rodríguez 2014).

$$Y_i = \mu + T_i + \epsilon_i \quad (13)$$

RESULTADOS

Primero se organizaron los Datasets, luego se realizó la aplicación de los algoritmos, luego se procede a realizar la reducción de la dimensionalidad, aplicación de los algoritmos y por último se realiza el análisis de estabilidad.

Paso 1. Datasets y comprobación de la información

Se consideraron 129 variables, con valores cuantitativos y cualitativos y 3000 registros, la aplicación de la evaluación se hizo de manera directa a cada uno de las personas, la que contenía 129 variables predictoras y 1 variable de clase.

Paso 2. Aplicación de técnicas de aprendizaje de máquina

Luego de aplicar en primera instancia los algoritmos Naïve Bayes y máquinas de soporte vectorial al conjunto de datos de entrenamiento, se obtuvieron los resultados que se muestran en la tabla 2, la cual presenta de manera comparativa las instancias correctas generadas por cada algoritmo.

Tabla 2: Comparación entre algoritmos de clasificación con los datos iniciales.

Técnica	TP	TN	FP	FN	Sensibilidad $\pm$ SD	Especificidad	Precisión $\pm$ SD
<i>Naïve Bayes</i>	30 $\pm$ 1	15	0	2 $\pm$ 1	90,5 $\pm$ 2,6	100	90,55 $\pm$ 1,85
<i>SVM</i>	31 $\pm$ 1	15	0	2 $\pm$ 1	91,6 $\pm$ 2,5	100	93,3 $\pm$ 1,85

Esta comparación muestra que el sistema SVM tiene una alta precisión de clasificación y menos computación debido a la extracción de características. De hecho, los resultados experimentales muestran una menor tasa de error y la tasa de sensibilidad varía entre 89.1 a 94.1% con una media $\pm$ Desviación Estándar de 91.6 $\pm$ 2.5%.

Paso 3. Reducción de la dimensionalidad

Se redujo la dimensionalidad de los datos utilizando: algoritmo genético, del método se extrajo las variables en común para conformar una nueva organización de variables (40), como se muestra en la tabla 3. En el Algoritmo genético se utilizaron subconjuntos de poblaciones de 10 individuos usando 40 características, las más preponderantes contenidas en los padres, acorde a la información que se extrajo de la base de datos, como se muestra en la tabla 3.

Paso 4. Evaluación del nuevo conjunto de datos con técnicas de aprendizaje de máquina

Se realizaron 30 réplicas con cada algoritmo y se seleccionaron los 10 mejores resultados para ser expuestos. Los resultados obtenidos se exponen en la Tabla 4. Los valores obtenidos se utilizaron para realizar el análisis de estabilidad utilizando el análisis de varianza con un intervalo de confianza del 99%. Se realizó nuevamente la evaluación del conjunto de datos con los dos algoritmos, Naïve Bayes y SVM para obtener un promedio en la precisión respectivamente de 94,44% con una desviación estándar de 1,85 y 96,29% con una desviación estándar de 1,85 en la predicción del grado de riesgo psicosocial, como se ve en la tabla 5.

Tabla 3: Algoritmo genético. Generación aleatoria de individuos.

X ₉₉	X ₂	X ₂₁	X ₁₀	X ₁₅	X ₈	X ₂₀	X ₁₁	X ₃₀	X ₁₉
X ₁₈	X ₄₃	X ₄₄	X ₅₀	X ₆₄	X ₇	X ₉₁	X ₃₂	X ₂₅	X ₁₂
X ₆₅	X ₁₁₇	X ₁₀₀	X ₂₁	X ₁₆	X ₄₇	X ₆₆	X ₇₇	X ₉₀	X ₁₁₆
X ₂₉	X ₅₄	X ₈₆	X ₁₁₅	X ₁₀₆	X ₃₃	X ₁₁₉	X ₁₁₈	X ₃₁	X ₄₅

Tabla 4: Réplicas con cada algoritmo y su promedio general.

Algoritmos	Replicas										Promedio General
SVM	95,5	95,54	96,6	96,66	94,63	96,76	96,2	96,9	96,5	96	96,29
Naïve Bayes	89,9	89,25	93,98	92,45	93,57	94,93	96,3	96,58	96,9	96,9	94,44

Tabla 5: Comparación entre algoritmos de clasificación

Técnica	TP	TN	FP	FN	Sensibilidad $\pm$ SD	Especificidad	Precisión $\pm$ SD
SVM	35 $\pm$ 1	17	0	3 $\pm$ 1	94,6 $\pm$ 2,7	100	96,29 $\pm$ 1,85
Naïve Bayes	34 $\pm$ 1	17	0	2 $\pm$ 1	91,87 $\pm$ 2,69	100	94,44 $\pm$ 1,85

Paso 5. Análisis de estabilidad

Tomando como referencia las soluciones encontradas en el paso 4 (Tabla 4), se realizó el análisis de varianza, con un intervalo de confianza del 99.5%. Dado que en la Tabla 6, el F tabulado es mayor que el F calculado, se puede expresar que no existen diferencias significativas en los resultados entregados por el sistema; es decir, bajo un intervalo de confianza del 99.5% los datos son estadísticamente iguales.

Tabla 6: Análisis de varianza (Anova).

Fuente Variación	SS	df	MS	F	Prob>F
Columnas	239.55	2	119.22	106.16	-1.8E-17
Error	30.61	27	1.134		
Total	269.067	29			

DISCUSIÓN

El uso de clasificadores lineales como Naïve Bayes en la implementación de sistemas de recomendación (Zhang et al., 2002) permite obtener buenos resultados frente a otros sistemas como SVM, otros autores como (Varando et al., 2015) utilizan los algoritmos Naïve Bayes para potenciar los sistemas de redes bayesianas con el fin de optimizar los límites de decisión de las mismas, otros autores utilizan las redes bayesianas con el fin de poder hacer predictivo el padecer estrés debido a las condiciones laborales (Cárdenas, 2016), algunos autores utilizan las técnicas de Naïve Bayes para poder predecir la aparición de diabetes y poder estimar un tratamiento adecuado para los pacientes (Castrillón et al., 2017).

Allwein et al., (2000) propusieron un marco unificador para estudiar la solución de problemas de categorización multiclase al reducirlos a múltiples problemas binarios que luego se resuelven utilizando un algoritmo de aprendizaje binario basado en márgenes. Los resultados experimentales con SVM y AdaBoost muestran que el esquema proporciona una alternativa viable a los algoritmos multiclase más comúnmente utilizados. Huang et al., (2013) proponen extender la discriminación binaria ponderada por distancia (DWD) a los casos multi-clase. Además de las extensiones habituales que combinan varios clasificadores DWD binarios, proponen un DWD multiclase global (MDWD) que encuentra un único clasificador que considera todas las clases a la vez. Los resultados teóricos muestran que MDWD es estable con la consistencia de Fisher, incluso en el caso particularmente desafiante cuando no hay una clase dominante, es decir, una clase con probabilidad mayor a 0.5. El rendimiento de diferentes métodos DWD multiclase se evaluaron mediante estudios de simulación y aplicación a conjuntos de datos de microarrays reales. También proporcionaron una comparación con las máquinas de vectores de soporte.

Qiao et al., (2015a) proponen un novedoso método de clasificación lineal que posee los méritos tanto de la máquina de vectores de soporte (SVM) como de la discriminación ponderada por distancia (DWD). El método de máquina de vectores de soporte ponderado por distancia propuesto se puede ver como un híbrido de SVM y DWD que encuentra la dirección de clasificación minimizando principalmente la pérdida de DWD, y determina el término de intercepción en la forma de SVM. Muestran que el mérito del método es la aplicación de DWD y, por lo tanto, supera el problema de superposición de datos y superposición de SVM. Por otro lado, el nuevo método no está sujeto a problemas de datos desequilibrados, que era una ventaja principal de SVM sobre DWD. Utiliza una pérdida inusual que combina la pérdida de bisagra (de SVM) y la pérdida de DWD a través de un truco de hiperplano axilar. Se desarrollan varias propiedades teóricas, que incluyen la consistencia de Fisher y la normalidad asintótica de la solución DWSVM. Utilizaron algunos ejemplos simulados para mostrar que el nuevo método puede competir con DWD y SVM en el rendimiento de la clasificación y la interpretabilidad.

Sun et al., (2017) desarrollaron un nuevo método de clasificación basado en ángulos multicategoría de máquinas de vectores de soporte ponderado por distancia (MD-WSVM) que está motivado por el método de clasificación binaria de vector de soporte ponderado por distancia (DWSVM). El nuevo método tiene los méritos tanto de la máquina de vectores de soporte (SVM) como de la discriminación ponderada por distancia (DWD), pero también alivia tanto el problema de acumulación de datos de SVM como el problema de datos desequilibrados de DWD. Los estudios teóricos y numéricos desarrollados por ellos demuestran las ventajas

del método MDWSVM sobre los métodos basados en ángulos existentes, lo que permite una mejor clasificación. Hong et al., (2005), decidieron comparar las máquinas de soporte vectorial con los modelos de regresión logístico y probabilístico para evaluar la predicción en la rotación de empleados basado en el desempeño laboral, donde, los resultados empíricos demostraron que el uso de las máquinas de soporte vectorial (SVM) es una alternativa prometedora para predecir la rotación de empleados en la gestión de recursos humanos. Con base en esto, nuestro modelo demuestra que la utilización de máquinas de vectores de soporte en la clasificación del riesgo psicosocial tiene un gran futuro para la implementación de modelos ya probados y para el desarrollo de un nuevo método basado en máquinas de vector de soporte que se adapte al modelo de datos que se generan en la identificación de riesgo psicosocial en población docente de colegios públicos colombianos. El uso de variables fisiológicas en el estudio y su uso con algoritmos como las máquinas de vectores de soporte permiten generalizar e inferir que se adaptan fácilmente al tratamiento del modelo matemático que trata SVM. Se recomienda seguir utilizando este tipo de algoritmos con este tipo de variables fisiológicas para su implementación en SVM.

Se puede para futuros trabajos de investigación seguir implementando las máquinas de vectores de soporte (SVM) incluyendo nuevas variables a parte de las fisiológicas ya demostradas aquí, como reconocimiento de patrones visuales o patrones de imágenes, desarrollar un nuevo método basado en las variables fisiológicas junto a variables organizacionales y variables asociadas a técnicas lean con técnicas de aprendizaje de máquina (machine learning) para automatizar el proceso de reconocimiento de presencia de riesgo psicosocial en la cultura organizacional de colegios y compañías. La optimización del espacio de búsqueda a través de métodos de optimización matemática como algoritmos de optimización perteneciente a la clase de técnicas de búsqueda local (hill climbing) y de búsqueda global (simulated annealing) adicional a las técnicas de SVM y NBC, puede mejorar el desarrollo y la aplicación de técnicas de clasificación respecto a identificación de riesgo psicosocial en diferentes modelos organizacionales.

Como vemos las disciplinas relacionadas con las técnicas de machine learning, incluyen el estudio psicológico del aprendizaje humano, el estudio de la evolución, la teoría del control adaptativo, el estudio de las prácticas educativas, la neurociencia, el comportamiento organizacional y la economía. Aunque la última década ha visto un mayor crecimiento con estos otros campos, recién estamos comenzando a aprovechar las posibles sinergias y la diversidad de formalismos y métodos experimentales utilizados en estos campos múltiples para estudiar sistemas que mejoran con la experiencia (Jordan et al., 2015), como son los psicosociales y organizacionales.

CONCLUSIONES

Con base en los resultados encontrados y su discusión se concluye que:

1) Se recomienda la utilización de algoritmos de máquinas de vectores de soporte para desarrollar mejores modelos que permitan mayor precisión para la predicción del riesgo psicosocial en este sector. 2) La utilización de modelos Naïve Bayes también podría ser fructífero en la predicción de este tipo de riesgo con menos variables incluidas en los estudios. 3) La experimentación con datos epidemiológicos de riesgo psicosocial y variables fisiológicas permitirá el desarrollo de herramientas tecnológicas que faciliten el mejoramiento de las condiciones laborales a través de la predicción. 4) La implementación de heurísticas poco utilizadas (Hill Climbing, Grasp) optimiza el espacio de búsqueda logrando la mejor solución posible local, que se puede mejorar realizando búsquedas locales iteradas para favorecer modelos de clasificación, como, las máquinas de soporte vectorial. 5) La creación de una línea de investigación dedicada a la implementación de algoritmos tipo SVM permitirá mejorar el desarrollo de la ingeniería y la aplicación de técnicas de soft computing en los niveles organizacionales, psicosociales y ergonómicos de las empresas en el país. 6) La implementación de machine learning en la toma de decisiones organizacionales favorecerá el mejoramiento de las condiciones de seguridad y salud en el trabajo en diferentes sectores económicos del país. 7) El mejor desempeño de predicción se obtuvo con el método SVM con una eficiencia en la precisión del 96,3%.

AGRADECIMIENTOS

Los más sinceros agradecimientos a la “Convocatoria Nacional para el Apoyo al Desarrollo de Tesis de Posgrado o de Trabajos Finales de Especialidades en el área de la Salud de la Universidad Nacional de Colombia 2017-2018” mediante Resolución 21 de 2017 de la Vicerrectoría de Investigación (21 de diciembre de 2017), como propuesta seleccionada para financiación del proyecto de investigación de la U. Nacional de Colombia bajo ID 40976, que financia la estancia doctoral del estudiante en la UPV.

REFERENCIAS

Allwein, E.L., R.E. Schapire e Y. Singer, *Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers*, *Journal of machine learning research*, 1(12), 113-141 (2000)

- Arias-González, M., *Integración de los Sistemas de Gestión de Calidad, el Medio Ambiente y la Seguridad y Salud del Trabajo*, Ciencias Holguín, 20(2), 39-48 (2014)
- Cárdenas-Gonzalo, D., *Probability Calculation of Suffering Stress Associated with Physical Health Problems through Bayesian Networks*, doi: 10.6036/7904, DYNA, 91(3), 248 (2016)
- Castrillón, O.D., W. Sarache y E. Castaño, *Sistema Bayesiano para la Predicción de la Diabetes*, doi: 10.4067/S0718-07642017000600017, Inf. Tecnol, 28(6), 161-168 (2017)
- Chandra, B., M. Gupta y M.P. Gupta, *Robust Approach for Estimating Probabilities in Naive-Bayes Classifier*. In International Conference on Pattern Recognition and Machine Intelligence (pp. 11-16), Springer, Berlin, Heidelberg, December (2007)
- Coca, G.A., O.D. Castrillón y S. Ruiz, *Programación Multiobjetivo en un Sistema de Fabricación Tipo "Contra Existencia" (JOB SHOP)*, doi: 10.4067/S0718-07642013000600011, Inf. Tecnol, 24(6), 79-92 (2013)
- Cortes, C. y V. Vapnik, *Support-Vector Network*, Mach. Learn., 20, 273-297 (1995)
- Cristiniani, N. y J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge, Cambridge University Press (2000)
- Hong W.C., P.F. Pai., Y.Y. Huang y S.L. Yang, *Application of Support Vector Machines in Predicting Employee Turnover Based on Job Performance*. In: Wang L., Chen K., Ong Y.S. (Eds.) *Advances in Natural Computation*. ICNC 2005. Lecture Notes in Computer Science, Vol. 3610. Springer, Berlin, Heidelberg (2005)
- Huang, H., Y. Liu y otros cinco autores, *Multiclass Distance-Weighted Discrimination*, Journal of Computational and Graphical Statistics, 22(4), 953-969 (2013)
- Huerta, L.R.J., N. Cabrera, L. Montiel, F. Benítez y V. Ramírez, *Data Mining: Impact of Daily Activities on Student Performance*, International Journal of Innovation and Applied Studies, 14(4), 927-935 (2016)
- International Labour Organization, *"Workplace Stress: A Collective Challenge"*. Occupational Safety and Health Service (LABADMIN/OSH), Ginebra, ILO, 2016. 62 p. (2016)
- Jordan, M.I. y T.M. Mitchell, *Machine learning: Trends, Perspectives, and Prospects*, Science, 349(6245), 255-260 (2015)
- Mosquera, R., L. Parra-Osorio y O.D. Castrillón, *Metodología para la Predicción del Grado de Riesgo Psicosocial en Docentes de Colegios Colombianos utilizando Técnicas de Minería de Datos*, doi: 10.4067/S0718-07642016000600026, Inf. Tecnol. 27(6), 259-272 (2016)
- Mosquera, R., L. Parra-Osorio y O.D. Castrillón, *Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos Utilizando Técnicas de Inteligencia Artificial*, <http://dx.doi.org/10.4067/S0718-07642018000400267>, Inf. Tecnol., 29(4), 267-280 (2018)
- Qiao, X. y L. Zhang, *Distance-Weighted Support Vector Machine*, arXiv preprint arXiv:1310.3003 (2013)
- Rodríguez Ojeda, L. *"Probabilidad y Estadística Básica para Ingenieros"*, Guayaquil, Ecuador: Escuela Superior Politécnica del Litoral, Instituto de Ciencias Matemáticas (2014)
- Saizarbitoria, I.H., M.B. Vilamitjana y M.C. Fa, *La Integración de Sistemas de Gestión Basados en Estándares Internacionales: Resultados de un Estudio Empírico Realizado en la Capv.*, Revista de dirección y administración de empresas, 1(14), 155-174 (2014)
- Smola, A.J. y B. Schölkopf, *Learning With Kernels* (Vol. 4). GMD-Forschungszentrum Informationstechnik. (1998)
- Soler, V.G y R. Esengeldiev, *Sistemas Integrados de Gestión y los Beneficios*, 3C Empresa, 3(4), 246-257 (2014)
- Sun, H., B.A. Craig y L. Zhang, *Angle-Based Multicategory Distance-Weighted SVM*, The Journal of Machine Learning Research, 18(1), 2981-3001 (2017)
- Tzeng, H. M., J.G. Hsieh e Y.L. Lin, *Predicting Nurses' Intention to Quit with a Support Vector Machine: A New Approach to Set up an Early Warning Mechanism in Human Resource Management*, CIN: Computers, Informatics, Nursing, 22(4), 232-242 (2004)
- Vapnik, V., *The Nature of Statistical Learning*, 2nd Ed., Springer, New York (1998)
- Varando, G., M.C. Bielza Lozoya y P. Larrañaga-Múgica, *Decision Boundary for Discrete Bayesian Network Classifiers*, Journal of Machine Learning Research, (16), 2725-2749 (2015)
- Villalobos, G., A.M. Vargas, J. Escobar, M.L. Jiménez y M.A. Rondón, *Batería de Instrumentos para la Evaluación de Factores de Riesgo Psicosocial*, Bogotá, Ministerio de la Protección Social (2010)
- Whitley, D.A., *Genetic Algorithm Tutorial*, Statistics and Computing, 4(2), 65-85 (1994)
- Yasodha, S. y P.S. Prakash, *Data Mining Classification Technique for Talent Management Using SVM*. In Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on (pp. 959-963). IEEE. (2012)
- Zhang, T. y V.S. Iyengar, *Recommender Systems Using Linear Classifiers*, Journal of Machine Learning Research, 2 (Feb), 313-334 (2002)