

## **Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos utilizando Técnicas de Inteligencia Artificial**

**Rodolfo Mosquera<sup>(1)</sup>, Omar D. Castrillón<sup>(1)</sup> y Liliana Parra<sup>(2)</sup>**

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

(2) Universidad Libre, Seccional Bogotá, Facultad de Derecho, Bogotá, Colombia.

*Recibido Sep. 28, 2017; Aceptado Nov. 30, 2017; Versión final Ene. 18, 2018, Publicado Ago. 2018*

---

### **Resumen**

Este artículo presenta una nueva metodología basada en técnicas de aprendizaje de máquina en diagnósticos de evaluaciones psicosociales para la identificación del grado de riesgo en docentes de colegios públicos en Colombia. Se realizó un estudio comparativo de tres importantes modelos de aprendizaje de máquina para la predicción: redes neuronales artificiales, arboles de decisión y Naïve Bayes, reduciendo la dimensionalidad de los datos. Esto se hace aplicando algoritmos genéticos, algoritmo de la cantidad esperada de la información, algoritmo GainRatioAttributeEval, coeficiente de relación de Pearson y análisis principal de componentes. Se utilizó una base de datos con 5340 registros epidemiológicos, correspondientes a evaluaciones psicosociales de docentes de colegios públicos de una ciudad colombiana. El mejor desempeño de predicción se obtuvo con el modelo de redes neuronales artificiales con una efectividad del 93%.

*Palabras clave: aprendizaje de máquina; redes neuronales artificiales; algoritmos genéticos; análisis principal de componentes*

## **Prediction of Psychosocial Risks in Colombian Teachers of Public Schools using Machine Learning Techniques**

### **Abstract**

This paper presents a new methodology based on machine learning techniques in diagnostics of psychosocial assessments to identify the risk level in teachers of public schools in Colombia. A comparative study of three important models of machine learning for prediction was done: artificial neural networks, decision trees and naive bayes, reducing the dimensionality of the data. This was done by applying genetic algorithms, algorithm of the expected amount of information, the algorithm GainRatioAttributeEval, Pearson's relation coefficient and principal components analysis. A database was used with 5340 epidemiological records, corresponding to psychosocial evaluations of teachers from public schools in the metropolitan area of a Colombian city. The best predictive performance was obtained with the model of artificial neural networks with an accuracy 93%.

*Keywords: machine learning; artificial neural network; genetic algorithm; principal components analysis*

## INTRODUCCIÓN

Los factores psicosociales se refieren a aspectos del diseño, la gestión y la organización del trabajo, que tienen el potencial para causar daño en la salud y la seguridad individuales, así como otros resultados adversos de la organización como la ausencia por enfermedad, la productividad reducida o el error humano. Incluyen varios temas como las demandas de trabajo, la disponibilidad de apoyo organizacional, recompensas y relaciones interpersonales, incluyendo temas como el acoso y la intimidación en el lugar de trabajo. El riesgo psicosocial se refiere al potencial de los riesgos psicosociales para causar daño, estos factores no tienen inmediatamente una connotación negativa o positiva (Leka 2016), el hecho de que una organización tenga un ambiente de trabajo psicosocial positivo o negativo dependerá de la eficacia con la que maneje el riesgo asociado con sus diversas dimensiones (Leka et al., 2015). En tiempos de globalización, las organizaciones se caracterizan cada vez más por reducción de personal, intensificación del trabajo y racionalización de los recursos. Esto ha dado lugar a la diversificación y al surgimiento de nuevos riesgos en el campo de la salud y la seguridad ocupacional, con un impacto importante. Los riesgos psicosociales se incorporan en las organizaciones, tanto a nivel de negocio como de políticas, por lo que la barrera clave se relaciona con las dificultades prácticas de no saber manejar adecuadamente los riesgos psicosociales (Langenhan et al., 2013).

El aprendizaje automático (machine learning) involucra el supuesto de cómo construir computadoras que mejoren automáticamente a través de la experiencia. Es uno de los campos técnicos de más rápido crecimiento en la actualidad, situado en la intersección de la informática y la estadística, y en el centro de la inteligencia artificial y la ciencia de los datos. El progreso reciente en el aprendizaje de máquina ha sido impulsado tanto por el desarrollo de nuevos algoritmos de aprendizaje y la teoría y por la explosión en curso en la disponibilidad de datos en línea y de computación de bajo costo. La adopción de métodos de aprendizaje de máquina intensivos en datos se puede encontrar en toda la ciencia, la tecnología y el comercio, lo que lleva a una mayor toma de decisiones basadas en evidencia en muchos aspectos de la vida, incluyendo el cuidado de la salud, la manufactura, la educación, los modelos financieros, actividades de seguridad y el marketing (Jordan et al., 2015). No se puede negar que muchos sistemas de inteligencia artificial mejoran indiscutiblemente nuestras vidas, por ejemplo, facilitando la comunicación a través de la traducción automática, ayudando a diagnosticar enfermedades y proporcionando comodidades modernas, como los teléfonos inteligentes que actúan como asistentes personales (Stajic et al., 2015).

El desarrollo de nuevas estrategias para la predicción temprana de la identificación de problemas psicosociales para elaborar actividades de prevención está actualmente en curso (Mosquera et al., 2016). Con el desarrollo de las nuevas tecnologías en el campo de la ingeniería, nuevos datos se pueden recoger y ponerse a disposición de la comunidad de investigación científica. Sin embargo, la predicción exacta del resultado de un factor psicosocial es una de las tareas más interesantes y desafiantes para los ingenieros. Como resultado, los métodos de Machine Learning se han convertido en una herramienta popular para los investigadores. Estas técnicas pueden descubrir e identificar patrones y relaciones entre ellos, a partir de conjuntos de datos complejos, mientras que son capaces de predecir con eficacia resultados futuros de un tipo de riesgo.

En este artículo se presentan los pasos para obtener un modelo cuantitativo con el cual se pueda evaluar e identificar los riesgos psicosociales aplicando técnicas de aprendizaje de máquina, logrando disminuir la dimensionalidad de los datos, mejorando el grado de precisión en la predicción, teniendo en cuenta las áreas físicas, psicosociales, ambientales y demográficas clasificadas dentro de factores como son, liderazgo y relaciones sociales en el trabajo, control sobre el trabajo, demandas del trabajo y recompensas.

## MATERIALES Y MÉTODOS

Esta sección presenta la información de la base de datos a intervenir y las descripciones cortas de tres modelos de clasificación representativos: Redes Neuronales Artificiales, Naïve Bayes y Árboles de Decisión junto con la aplicación de cinco métodos para la reducción de la dimensionalidad de los datos como son: algoritmo genético (Whitley, 1994), coeficiente de correlación de Pearson (Wang, 2013), la cantidad de información esperada que proporcionan los atributos (Tuya, 2007), el algoritmo Gain Ratio Attribute Eval (Hall et al., 2003) y el análisis de componentes principales (Jolliffe et al., 2016), teniendo como punto de corte los cuarenta atributos más significativos presentes en los cinco métodos de reducción de dimensionalidad. Se presenta una visión general de la metodología propuesta en la Figura 1.

### *Paso 1 y 2. Datos a Intervenir*

El conjunto de datos que se analizó proviene de datos sociodemográficos de docentes, información de variables físicas, psicosociales y ambientales y del grado de riesgo, como se muestra en la tabla 1, se verifican

los campos con la información correcta en la base de datos. Cabe anotar que para la evaluación en la denotación de las ecuaciones se muestra como se realizó la evaluación y como se programaron los algoritmos con las 40 variables que resultaron de la reducción de la dimensionalidad.

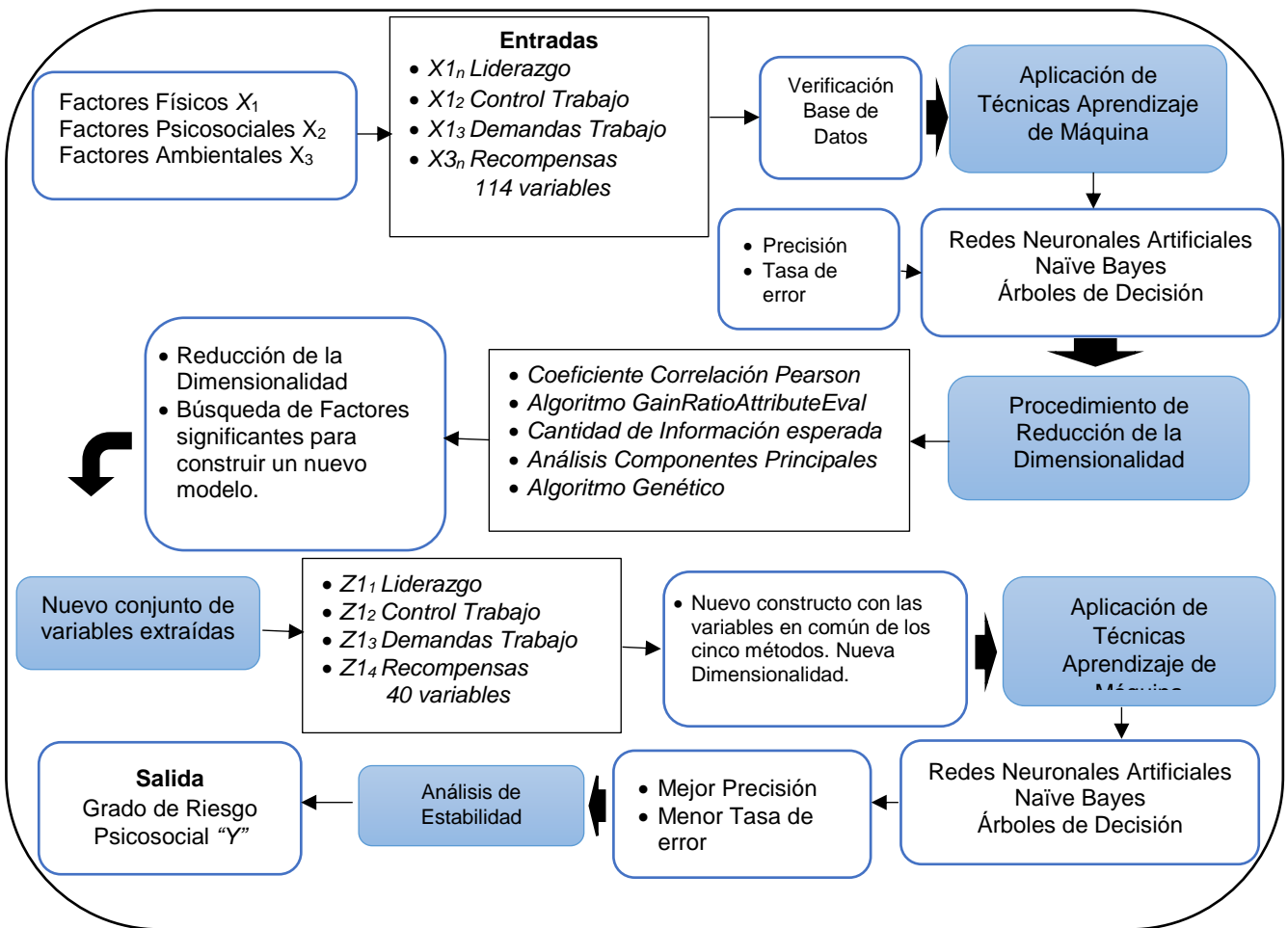


Fig. 1: Visión general de la metodología propuesta.

**Paso 3. Estructuración de Algoritmos Redes Neuronales artificiales (ANN), Naïve Bayes y Árboles de Decisión**

Una Red Neuronal Artificial es un sistema analítico basado en la estructura de Redes neuronales biológicas y su forma de codificar y resolver problemas (Gupta, 2013). En primera instancia se utilizó una Red Neuronal multicapa perceptrón con el algoritmo Back Propagation. La Red Neuronal Artificial inicial comprende tres tipos de capas: la capa de entrada (con 123 variables de entrada), las capas ocultas (compuesta por 12 neuronas) y cuatro capas de salida (Riesgo MuyAlto, Alto, Medio y Bajo).

En segunda instancia, se reemplazó en la capa de entrada el número de variables a introducir (40 variables), dejando igual las capas ocultas y las capas de salida. Cada una de las columnas de los datos, representará los valores  $R_{i,1} \dots R_{i,m}$  que ingresarán a cada uno de los nodos. Posteriormente, se debe definir una función de pesos  $W_{i,1} \dots W_{i,m}$ . Estos pesos serán los valores a ajustar una vez se programe la red (Anderson, 2007, p.222).

$$"F_i(x) = \left( \sum_{j=1}^n W_{i,j} * FR_{i,j} \right)" \tag{1}$$

Dónde para cada uno de los nodos  $N1 \dots Nn$  de la red, se define la función de transferencia que se expone en la Ecuación (1), donde, n es el número de personas evaluadas,  $FR_{i,j}$  es el valor del riesgo de cada persona  $i$  con respecto a cada variable  $j$ ,  $W_{ij}$  es el peso calculado para el índice de riesgo  $i$  por cada persona  $j$ .

Tabla 1: Variables de los dominios de factores de riesgo psicosocial intralaboral. Adaptado de Villalobos et al. (2010)

Dominios	Variable
Socio demográficas	Sexo(S1), edad(S2), casado(S3), unión libre(S4), soltero(S5), posgrado(S6), estrato(S7), tipo de vivienda(S8), número de personas que dependen económicamente(S9), antigüedad en el cargo(S10).
Demandas del Trabajo	ruido(d1), frío(d2), calor(d3), airefresco(d4), luz(d5), espacio modo(d6), químicos(d7), esfuerzofísico(d8), herramientascómodos (d9), expuestomicrobios(d10), preocupaaccidentarme(d11), limpioyordenado(d12), tiempoadicional(d13), alcanzaeltiempo(d14), excesotrabajo(d15), esfuerzomental (d16), concentración(d17), memorizarinfo(d18), tomardecisiones difíciles(d19), atenderasuntosalmismotiempo(d20), atencpequeñosdetalles(d21), respondocosasvalor(d22), respondopordinero (d23), respondoseguridadotros (d24), Respondoáreatrabajo(d25), cuidarsaludotraspersonas(d26), órdenescontradictorias (d27), hacercosasinnecesarias(d28), pasarporaltonormasoproced(d29), hacercosasformamáspráctica(d30), horarionoche(d31), hacerpausas(d32), laborardíasdeseanso(d33), casapiensotrabajo(d34), Discutoportrabajo(d35), atendertrabajoencasa (d36), tiempoconfamiliaespoco(d37), atenderasuntopersonal(d38), cambiosdifícultanlabor(d39), Atencientesenojados(d40), Atencientespreocupados(d41), Atencientes tristes (d42), atencpersonas enfermas(d43), atencpersonasnecesitadas(d44), Atenc clientesquemaltratan(d45), demostrarsentimientosdistintosalosmíos (d46), atencsituacionesviolencia(d47), atencsituaciones dolorosas(d48).
Control sobre el Trabajo	Desarrollarmishabilidades(c1), aplicarconocimientos(c2), aprendernuevascosas(c3), asignantrabajocapacidades(c4), decidir cuántotrabajohacerenel día (c5), decidirvelocidadtrabajo (c6), cambiarordenactividades (c7), cambiotrabajobeneficiosos (c8), explicancambios (c9), sugerenciassobrecambios (c10), tienenencuentamisideas (c11), informanfunciones (c12), informandecisionestrabajo (c13), explicanresultadostrabajo (c14), explicanefectotrabajoempresa (c15), explicanobjetivotrabajo (c16), informanorientartrabajo (c17), informanresolverasuntotrabajo (c18), capacitacionestrabajo (c19), capacitaciónútiltrabajo (c20), capacitaciónmejorartrabajo (c21).
Liderazgo y Relaciones Sociales en el Trabajo	instruccionesclaras (I1), jefeayudaorganizartrabajo (I2), jefecuentamispuntosdevista (I3), jefeanimahacermejormitrabajo (I4), jefedistribuyelastareas (I5), jefe comunicainfotrabajo (I6), orientaciónmeayudahacermejortrabajo (I7), jefeayudaprogresartrabajo (I8), jefeayudasentirmebientrabajo (I9), jefeayudasolucionarproblemastrabajo (I10), confiarjefe (I11), jefeescuchaproblemas (I12), jefebrindaapoyocuandonecesito (I13), agradaambiente grupotrabajo (I14), grupotrabajotratanaformarespetuosa (I15), Sientoconfiarcompañerostrabajo (I16), siento gustocompañerostrabajo (I17), grupotrabajoalgunapersonasmaltratan (I18), solucionamosproblemasformarespetuosa (I19), integración grupotrabajo (I20), grupotrabajounido (I21), personashacensentir partedelgrupo (I22), trabajogrupocompañeroscolaboran (I23), fácilponerdeacuerdo grupo (I24), compañerostrabajoayudandificultades (I25), apoyamosunos a otros (I26), compañeros escuchancuandotengoproblemas (I27), informanbientrabajo (I28), informaciónrendimiento clara (I29), evalúantrabajoayudamejorar (I30), informantiemposobre mejorar (I31).
Recompensas	empresacofiantrabajo (r1), empresapagantiemposalario (r2), pagoreciboofrecciónempresa (r3), pagorecibomerezco (r4), trabajo posibilidadesprogresar (r5), personasbientabajoprogresan (r6), empresapreocupabienestartrabajador (r7), trabajoestable (r8), trabajohacesentirbien (r9), Sientoorgullotrabajo (r10), Hablobienempresa (r11).
Grado de Riesgo	Muy Alto, Alto, Medio y Bajo

**Naïve Bayes**

Naïve Bayes es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales (Perikos et al., 2013 p.257). En el clasificador Bayer (la función Classify) ((Perikos et al., 2013 p.257) se define como:

$$p(X_1=x_1, \dots, X_n=x_n | C=c) = \prod_{i=1}^n p(X_i=x_i | C=c) \tag{2}$$

Dónde: En el caso de que las n variables predictoras  $X_1, \dots, X_{113}$  sean continuas, se tiene que el paradigma Naïve Bayes se convierte en buscar el valor de la variable C, que denotamos por c, que maximiza la probabilidad a posteriori de la variable C, dada la evidencia expresada como una instanciación de las variables  $X_1, \dots, X_{113}$ , esto es,  $X = (X_1, \dots, X_{113})$ . Por tanto, en el paradigma Naïve Bayes, la búsqueda de la

predicción más probable,  $c^*$ , una vez conocidos los valores de las variables  $(X_1, \dots, X_{113})$  de un determinado individuo, ((Perikos et al., 2013 p.257) se reduce a:

$$"c^* = \arg_c \max \rho (C=c | X_1= x_1, \dots, X_n= x_n)" \tag{3}$$

$$" = \arg_c \max \rho (C=c) \prod_{i=1}^n \rho (X_i= x_i | C=c)" \tag{4}$$

*Árboles de Decisión*

Los datos de entrenamiento son un sistema  $S = s_1, s_2, \dots, s_{113}$  de muestras ya clasificados. Cada ejemplo  $s_i = \{x_1, x_2, \dots, x_{113}\}$  es un vector donde  $x_1, x_2, \dots, x_{113}$  representan los atributos o características. Los datos de entrenamiento son aumentados con un vector  $C = \{c_1, c_2, \dots, c_{113}\}$  donde  $c_1, c_2, \dots, c_{113}$  representan la clase a la que pertenece cada dato (Quinlan, J. R. 1993). En nuestro caso el modelo referido tomo las 113 variables como predictores y una variable de clase de respuesta.

*Paso 4. Aplicación algoritmos para reducción de la dimensionalidad*

Se aplica la reducción de la dimensionalidad de los datos con las técnicas escogidas.

*Coefficiente de Correlación de Pearson*

En estadística, el coeficiente de correlación, a veces también llamado coeficiente de correlación cruzada, es una cantidad que da la calidad de un ajuste por mínimos cuadrados a los datos originales. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables. (Ott et al., 2016, p.588).

$$"r_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{S_{xx}S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}" \tag{5}$$

Donde, generalmente, el  $r_{x,y}$  de una correlación es un número positivo, y tiende a aumentar cuando  $x$  aumenta,  $r_{x,y}$  es negativo si  $y$  tiende a disminuir a medida que  $x$  aumenta, y  $r_{x,y}$  es cero si no hay relación entre cambios en  $x$  y cambios en  $y$  o si hay una relación no lineal tal que los patrones de aumento y disminución en  $y$  (a medida que  $x$  aumenta) se cancelan el uno otro.

*Cantidad de información esperada*

La medida basada en la cantidad esperada de información que proporciona el atributo se obtiene calculando la entropía mediante la siguiente expresión (Tuya et al., 2007, p.117):

$$"Entropia = I (P(V_1), \dots, P(V_n)) = \sum_{i=1}^n - P(V_i) \log_n P(V_i)" \tag{6}$$

Siendo  $P(V_i)$  la probabilidad de la clase  $i$ ésima y  $n$  el número de clases posibles. La entropía es 1 cuando la probabilidad de las clases es igual (Tuya et al., 2007, p.117). Dónde los valores de las clases se transforman a cero y uno, según el grado de riesgo (MuyAlto, Alto =1; Medio y Bajo=0).

$P(V_i)$  = Clasificación MuyAlto y Alto,  $n$  = Clasificación Medio y Bajo

*Algoritmo Gain Ratio Attribute Eval*

Evalúa el valor de un atributo midiendo la relación de ganancia con respecto a la clase (Hall, 1999).

$$"GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)." \tag{7}$$

Dónde:

$H(Class) = X_{114}$  (MuyAlto, Alto, Medio y Bajo);  $H(Attribute) = X_{1...X_{113}}$

*Análisis de componentes principales*

Los grandes conjuntos de datos son cada vez más comunes y son a menudo difíciles de interpretar. El análisis de componentes principales (PCA, por sus siglas en inglés) es una técnica para reducir la dimensionalidad de tales conjuntos de datos, incrementando la interpretabilidad, pero al mismo tiempo minimizando la pérdida de información (Jolliffe et al., 2016). Para aplicar el análisis de componentes principales en nuestra metodología utilizamos el método basado en la matriz de correlación, cuando los datos no son dimensionalmente homogéneos o el orden de magnitud de cada variable aleatoria medida no es el mismo. El método inicia de la matriz de correlación, considerando que tenemos el valor de cada una de las  $X$  variables aleatorias  $F_j$ . Para cada uno de los  $n$  individuos se toman el valor de estas variables y se escribe el conjunto de datos en forma de matriz (Jolliffe, 2002, p.25):

$$" \widetilde{\alpha}_k^T \widetilde{\alpha}_k = \lambda_k, \tag{8}$$

Donde se puede observar que cada conjunto:

$$" k=1, 2, \dots, p, " \tag{9}$$

Es considerado una muestra aleatoria para la variable  $F_j$ . A partir de los  $X * n$  datos que corresponden a las  $X$  variables aleatorias, se construye la matriz de correlación muestral, que se define por (Jolliffe et al., 2002, p.25):

$$\frac{\lambda_k \alpha_{jk}}{[\text{var}(x_j) \text{var}(z_k)]^{1/2}} = \lambda_k^{1/2} \alpha_{kj} " \tag{10}$$

Dónde:

$\text{cov}(X_{(i)}, Z_k) = \text{Covarianza de } (F_1, F_{123});$

$\sqrt{\text{var } X_j \text{var } Z_k} = \text{Raíz cuadrada de las Varianzas de } (F_1, F_{123}).$

Dado que la matriz de correlación es simétrica, entonces resulta diagonalizable y sus valores propios  $\lambda_i$  vienen dados por (Jolliffe et al., 2002, p.26):

$$" \sum_{j=1}^p R_{j:q}^2 " \tag{11}$$

*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.

$$\text{Fitness}_{G,\text{Riesgo}} = \min \left\{ \sum_{i=1}^N P_{ij} \right\} \tag{12}$$

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:

$$\text{Individuo}_i = \{ \text{conjunto de 40 variables de las 123} \}, ( i=1,2,3,\dots,10) \tag{13}$$

*Paso 5. Evaluación con las técnicas de aprendizaje de máquina*

Se realizó nuevamente la evaluación a los datos con las 40 variables seleccionadas con los tres algoritmos de aprendizaje de máquina (Redes neuronales artificiales, Naïve Bayes y árboles de decisión), para obtener los resultados de precisión y tasa de error de cada uno.

*Paso 6. Análisis de estabilidad*

Con el fin de analizar la estabilidad del sistema en los resultados obtenidos, un análisis de varianza usando la ecuación (14) debe ser realizado con el fin de establecer si existen diferencias significativas en los resultados. En este análisis,  $T_i$  representa la respuesta a las variables;  $T_i$  es el efecto causado por el tratamiento  $i$ -ésimo y  $\epsilon_i$ , el  $i$ -ésimo error experimental. La información recolectada debe cumplir los requerimientos de independencia y normalidad requeridos. El análisis de varianza es realizado bajo un intervalo de confianza del 99.5% (Rodríguez 2014).

$$"Y_i = \mu + T_i + \epsilon_i" \quad (14)$$

**RESULTADOS**

Primero la organización de los datasets y comprobación de la información, luego la aplicación de técnicas de Aprendizaje de Máquina, posterior reducción de la dimensionalidad, de nuevo evaluación del nuevo conjunto de datos con técnicas de aprendizaje de máquina y por último el análisis de estabilidad.

*Paso 1 y 2. Datasets y comprobación de la información*

Se consideraron 114 variables, con valores cuantitativos y cualitativos y 5340 registros de muestra correspondientes a la aplicación de la batería de riesgos psicosocial (Ministerio de la Protección Social, 2010) en docentes de colegios públicos del área metropolitana de una ciudad colombiana, la aplicación de la evaluación se hizo de manera directa a cada uno de las personas, la que contenía 113 variables predictoras y 1 variable de clase.

*Paso 3. Aplicación de técnicas de Aprendizaje de Máquina*

Luego de aplicar en primera instancia los algoritmos Redes Neuronales Artificiales, Naïve Bayes y Árboles de Decisión 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.

% Entrenamiento	% Validación	% Clasificación / Tasa Error Redes Neuronales Artificiales	% Clasificación / Tasa Error Naïve Bayes	% Clasificación / Tasa Error Árboles de Decisión
10	90	76 / 24	74 / 26	74 / 26
20	80	77 / 23	74.5 / 25.5	75 / 25
40	60	78.5 / 21.5	75 / 25	75 / 25
50	50	78.5 / 21.5	75 / 25	76 / 24
70	30	78.5 / 21.5	75 / 25	76 / 24
80	20	79 / 21	75 / 25	78 / 22
	Promedio Simple	78 / 22.08	75 / 25.25	76 / 24.33

Los resultados muestran en este caso que el porcentaje de clasificación y la tasa de error de las redes neuronales artificiales es mayor y menor respectivamente, en comparación a los algoritmos de Naïve Bayes y árboles de decisión, con un 78% de correcta clasificación y una tasa de error del 22% frente al 75% y 25% de Naïve Bayes y de 76% y 24% de los árboles de decisión.

*Paso 4. Reducción de la dimensionalidad*

Debido a los bajos resultados obtenidos por cada algoritmo en la clasificación, se optó por reducir la dimensionalidad de los datos. Para realizar la reducción de la dimensionalidad de los datos se utilizaron cinco

métodos, la técnica estadística de coeficiente de correlación de Pearson, la cantidad de información esperada que proporcionan los atributos, el algoritmo Gain Ratio Attribute Eval, el análisis de componentes principales y algoritmo genético, como se muestra en la tabla 3, tabla 4, tabla 5 y tabla 6, de cada uno de los métodos se extrajeron las variables en común para conformar una nueva organización de variables (40), como se muestra en la tabla 7, teniendo como base estadístico el coeficiente de correlación de Pearson. Con esta nueva organización de los datos (40 variables) se probaron de nuevo los tres algoritmos para definir el grado de predicción y la tasa de error.

*Coefficiente de Correlación de Pearson*

Los valores de la matriz compuesta por los 5340 datos y las 113 variables del conjunto de datos se ingresan a la matriz de correlación para definir la correlación entre cada variable y el grado de riesgo psicosocial, con el fin de priorizar las variables que serán seleccionadas para conformar el nuevo conjunto de variables al ser consolidadas con las variables de los demás métodos de reducción, obteniendo como resultado las 40 primeras variables que se muestran en la tabla 3.

Tabla 3: Variables seleccionadas con el coeficiente de relación de Pearson.

Nombre de la Variable	Coefficiente Correlación	Nombre de la Variable	Coefficiente Correlación	Nombre de la Variable	Coefficiente Correlación
nperdepecon	0,998	expresltdtrab	0.531	explobjtrabj	0,529
casado	0,698	hcfopract	0.531	trabcasa	0,529
trabestab	0,542	atenperenferm	0.531	demossentdist	0,529
trabatenviolen	0,541	perhcbntrabprogr	0.531	aprenderncs	0,528
espcomd	0,537	sgrotros	0,530	mgtrabeuni	0,528
sugsobrecambios	0,537	sotrabemp	0,530	jaspqsrqsprab	0,527
japrugtrab	0,536	jctirt	0,530	hienmigrprab	0,527
instrclaras	0.536	evalayudmej	0,530	herramcom	0,527
atenclmtris	0.536	clarfcnes	0,530	atencpeqdetll	0,527
calor	0.534	respvalor	0,530	jbacn	0,527
excesotrabaj	0.534	discportrab	0,530	hacerpausas	0,527
jefeanimtrab	0.532	trabatenstrisdol	0,530	sqpccomtrab	0,527
ecemtrab	0.532	infoorntrabj	0,530		
labdiasdesc	0.532	explefctrab	0,529		

*Cantidad de información esperada (Entropía)*

A partir del conjunto de datos que se tiene (5340 datos y 123 variables) se determina cuáles son los atributos que más influyen en la clasificación, la cual se obtiene calculando la entropía de los datos, los cuales se obtienen de correlacionar cada una de las variables con el grado de riesgo, el cual ha sido transformado en 1 (Riesgo MuyAlto y Alto) y 0 (Riesgo Medio y Bajo), de donde se obtienen las 40 variables, de donde saldrán las variables en común con los demás métodos para consolidar un nuevo conjunto de datos, como se muestra en la tabla 4.

Tabla 4: Cantidad de información esperada (Entropía)

Nombre de la Variable	Coefficiente Entropía	Nombre de la Variable	Coefficiente Entropía	Nombre de la Variable	Coefficiente Entropía
calor	0.918	cambordactiv	0.515	sugsobrecambios	0.409
memoriz	0.918	tencuentideassug	0.515	explobjtrabj	0.409
atencpeqdetll	0.918	caputiltrabj	0.515	jctirt	0.409
sgrotros	0.918	ecslpdfrespt	0.515	ctqrtrabglcc	0.409
trabcasa	0.918	eptmsal	0.515	hccosinnecs	0.384
jefeanimtrab	0.918	esfmental	0.458	demossentdist	0.384



Tabla 4 (continuación)

Atenperenferm	0.918	hcfpract	0.458	casado	0.323
excesotrabaj	0.658	jefeptvista	0.458	tomadecdif	0.323
expresltdtrab	0.658	trabatenviolen	0.412	respvalor	0.323
evalayudmej	0.658	nperdepecon	0.409	labdiasdesc	0.323
espcomd	0.550	limpioyord	0.409	aprenderncs	0.323
concentr	0.550	desrrllhab	0.409	tomarparausas	0.323
atnsaludotrper	0.515	asigrtrabcapcd	0.409		
pasnormoproc	0.515	dectrabdia	0.409		

*Algoritmo Gain Ratio Attribute Eval*

La evaluación del valor de los atributos de los 5340 datos con respecto a la relación de ganancia de la clase de cada atributo permite a través de este algoritmo conocer el coeficiente de ganancia y redimensionar el tamaño de las variables a evaluar, seleccionando las variables con mayor peso para ser tenidas en cuenta en la consolidación del conjunto de datos nuevo, como se muestra en la tabla 5.

Tabla 5: Algoritmo GainRatioAttributeEval. (Relación de ganancia de la clase).

Nombre de la Variable	Relación de Ganancia	Nombre de la Variable	Relación de Ganancia	Nombre de la Variable	Relación de Ganancia
nperdepecon	0.992	infoorntrabj	0.164	hienmigrprtrab	0.162
casado	0.784	ecemtrab	0.164	pararauntpers	0.162
trabestab	0.166	empbienestrab	0.164	expresltdtrab	0.162
espcomd	0.166	aprenderncs	0.164	quimicos	0.162
sugsobrecambios	0.165	labdiasdesc	0.164	explefctrab	0.162
jctirt	0.165	ordcontra	0.163	veltrabj	0.162
herramcom	0.165	explobjtrabj	0.163	atencImtris	0.162
calor	0.165	capmejortrab	0.163	clarfcnes	0.162
trabatenviolen	0.165	jasbientrab	0.163	hcfpract	0.162
sotrabemp	0.165	evalayudmej	0.163	mgtrabeuni	0.161
jbacn	0.165	perhcbntrabprogr	0.163	hacerparausas	0.161
sgrotros	0.165	atenperenferm	0.163	jaspqsrqsptrab	0.161
demossentdist	0.165	jefeanimtrab	0.163		
instrclaras	0.164	ecslpdfrespt	0.163		

*Análisis de componentes principales*

La construcción de una matriz de correlación basado en la división de las covarianzas de las 123 variables entre la raíz cuadrada de las varianzas de los mismos datos nos permite obtener unos valores que muestran que variables están más ligadas al grado de riesgo y nos permite trabajar con las 40 nuevas variables, como se muestra en la tabla 6.

Tabla 6: Análisis de componentes principales.

Nombre de la Variable PCA	Relación de Ganancia	Nombre de la Variable PCA	Relación de Ganancia	Nombre de la Variable PCA	Relación de Ganancia
excesotrabaj	0.992	atencpeqdetill	0.731	pararauntpers	0.445
elpagomerez	0.992	concentr	0.731	aprenderncs	0.445
hienmigrprtrab	0.833	mgtrabeuni	0.731	memoriz	0.445
explccambios	0.830	explefctrab	0.731	prepaccidnt	0.445
pctiempoflia	0.825	infodecstrabj	0.731	ecemtrab	0.445

Tabla 6 (continuación)

Trabcasa	0.823	cambordactiv	0.532	evalayudmej	0.162
hacerpausas	0.822	tomardiasdesc	0.532	ecslpdfrespt	0.162
herramcom	0.822	tomadecdif	0.532	jectptrab	0.162
sotrabemp	0.822	tiempoact	0.532	casapienstrab	0.162
trabestab	0.820	calor	0.532	labdiasdesc	0.161
infoesclara	0.820	frio	0.532	luz	0.161
jbacn	0.820	antigue	0.530	edad	0.161
confjefe	0.820	pgeqmolaemp	0.530	atenperenferm	0.161
japrugtrab	0.820	islqdmemtrab	0.530	atencImtris	0.161
instrclaras	0.820	jasbientrab	0.530		
infoorntrabj	0.820	cambdfclabor	0.530		

**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. El método de entrenamiento es el de k vecinos más cercanos y para entrenar se usa solo el 60% de las muestras, logrando obtener un 93% de eficiencia en la selección para clasificar, como se muestra en la tabla 7.

Tabla 7: Algoritmo genético.

Nombre Variable AG	Nombre Variable AG	Nombre Variable AG	Nombre Variable AG	Nombre Variable AG	Nombre Variable AG	Nombre Variable AG
pararauntpers	jefeayudorgtr	gtrabtrfresp	emtrablpnauao	hrrionoct	explccambios	tomardiasdesc
trabcasa	jbacn	antigue	nperdepecon	atenperenferm	respreatrb	trabatenviolen
elpagomerez	airefres	atnsaludotrper	infotiempsomej	acdtrabescpr	espcomd	jefedistareas
expmicrobios	trabestab	soltero	infodecstrabj	desrllhab	ulibre	explefctrab
ojmahmtrab	excesotrabaj	sotrabemp	instrclaras	ordcontra	atenclenoj	esfmental
atencImtris	capmejortrab	cambbenf	perhcbntrabprogr	atencvrasunt		

**Selección de variables en común de los cinco métodos.**

La selección de las variables con que se redujo la dimensionalidad de los datos, se realizó teniendo en cuenta las variables que figuraban en común en los cinco métodos, como se puede observar en la tabla 8.

**Paso 5. Evaluación del nuevo conjunto de datos con técnicas de aprendizaje de máquina**

Se realizaron 20 réplicas con cada algoritmo y se seleccionaron los 10 mejores resultados para ser expuestos. Los resultados obtenidos se exponen en la Tabla 9. 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 tres algoritmos, redes neuronales artificiales, Naïve Bayes y árboles de decisión para obtener una precisión respectivamente de 93%, 86% y 90% en la predicción del grado de riesgo psicosocial y una tasa de error del 7.25%, 14.21% y 10% comparado con los datos de la base de datos, como se ve en la tabla 10.

Tabla 8: Nueva Selección de Variables definida de los cuatro métodos.

Variables Seleccionadas	Variables Seleccionadas	Variables Seleccionadas	Variables Seleccionadas	Variables Seleccionadas
nperdepecon	jefeanimtrab	evalayudmej	atencpeqdetll	calor
casado	ecemtrab	infoorntrabj	jbacn	excesotrabaj

Tabla 8 (continuación)

Trabestab	labdiasdesc	explefctrab	hacerpausas	sotrabemp
trabatenviolen	expresltdtrab	explobjtrabj	pararauntpers	jctirt
espcomd	hcforpract	trabcasa	memoriz	hienmigrprab
sugsobrecambios	atenperenferm	aprenderncs	elpagomerez	herramcom
instrclaras	perhcbntrabprogr	mgtrabeuni	pctiempoflia	tiempoadct
atencImtris	sgrotros	jaspqsprqsprab	confjefe	antigue

Tabla 9: Replicas con cada algoritmo y su promedio general.

Algoritmos	Replicas										Promedio General
ANN	92,2	91,5	92,8	92,4	92,5	93,75	94	93	93,7	93,8	92,965
Naïve Bayes	86,12	86,25	87	85,45	86,55	85,93	85,13	84,58	86,9	86,9	86,094
Arboles de Decisión	88	89,5	90	90	89	89,8	90	90	92	93	90,05

Tabla 10: Comparación entre algoritmos de clasificación

% Entrenamiento	% Validación	% Clasificación / Tasa Error Redes Neuronales Artificiales	% Clasificación / Tasa Error Naïve Bayes	% Clasificación / Tasa Error Árboles de Decisión
10	90	92 / 8	84 / 16	89 / 11
20	80	92.5 / 7.5	85.7 / 14.3	88 / 12
40	60	93 / 7	86.5 / 13.5	88 / 12
50	50	93 / 7	86.5 / 15.3	91 / 9
70	30	93 / 7	86 / 14	92 / 8
80	20	93 / 7	86 / 14	92 / 8
	Promedio Simple	93 / 7.25	86 / 14.21	90 / 10

**Paso 6. Análisis de estabilidad**

Tomando como referencia las soluciones encontradas en el paso 9 (Tabla 9), se realizó el análisis de varianza, bajo un intervalo de confianza del 99.5%. Dado que en la Tabla 11, 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 11: Análisis de varianza (Anova).

Fuente Variación	SS	df	MS	F	Prob>F
Columnas	238.45	2	119.22	105.17	-1.8E-16
Error	30.61	27	1.134		
Total	269.067	29			

**DISCUSIÓN**

Algunos trabajos actuales están utilizando algoritmos de machine learning como Naïve Bayes en la aplicación como clasificador de texto para observar la opinión de los usuarios con respecto a un tema (Jiang et al.,2016), para el reconocimiento de emociones en los escritos (Perikos et al.,2016) se están utilizando como línea para el desarrollo de análisis de minería de texto, para profundizar en la gran cantidad de información que se genera en las redes sociales, lo cual, dista de la aplicación que se le dio a estos algoritmos en este trabajo, pero que dan una idea de su utilidad en otros campos de la seguridad y la salud en el trabajo. Otros autores, utilizan las técnicas de programación genética para mejorar la clasificación en el reconocimiento de imágenes (Liang et al.,2017) y desarrollan técnicas con heurísticas para mejorar la eficiencia de la cadena de abastecimiento basado en la programación genética (Diabat et al.,2016), este trabajo optimiza la reducción de la dimensionalidad de los datos para mejorar la eficiencia en la clasificación. Algunos autores utilizan los algoritmos genéticos para realizar

modelos que sirvan para la programación de la producción mejorando los tiempos de operación obteniendo muy buenos resultados (Coca et al., 2016), en este artículo, se utilizó algoritmo genético para hallar el óptimo global con respecto a las variables más eficientes para clasificar el grado de riesgo psicosocial.

La aplicación de redes bayesianas para el cálculo de probabilidad de sufrir estrés asociado a problemas de salud en un estudio que se llevó a cabo en España teniendo como base la V encuesta europea de condiciones de trabajo, destacan que la probabilidad más alta de sufrir niveles de estrés elevados la tienen cuando se introduce la variable del insomnio en el modelo. Es decir, si no se es capaz de conciliar el sueño regularmente, esas probabilidades de sufrir niveles de estrés elevados ascienden a un 15,58% (Cárdenas, 2016), con lo que se demuestra la utilidad de las redes bayesianas para identificar sintomatología de estrés laboral. También se tiene que con un sistema de clasificación Bayesiano para la identificación temprana de la diabetes Mellitus con base en el análisis de algunas variables tales como número de embarazos, presión arterial diastólica, espesor cutáneo del tríceps, índice de masa corporal, herencia y edad, se logró el 98,46% de precisión en la predicción incorporando la variable insulina en suero, demostrando que los algoritmos bayesianos son capaces de trabajar adecuadamente en la identificación de variables fisiológicas (Castrillón et al., 2017).

Un trabajo donde se utilizaron técnicas de Redes Neuronales Artificiales, Redes Bayesianas y Maquinas de Soporte Vectorial para la predicción del diagnóstico de cáncer de seno, con resultados de 65%, 51% y 71% respectivamente (Kourou et al., 2015), lo que nos permite inferir que desde el punto de vista de diagnósticos médicos o clínicos o en áreas de la salud, los algoritmos utilizados en este trabajo presentan unos buenos resultados. Trabajos dedicados a la minería de datos evaluando riesgo psicosocial en población de docentes de colegios públicos (Mosquera et al., 2016) muestran gran adaptabilidad y precisión con los resultados obtenidos y al compararlos con este trabajo, los investigadores dan un paso importante en la utilización de algoritmos de machine learning para la predicción de riesgo psicosocial como herramienta para la prevención de riesgos laborales. Técnicas de redes neuronales también se han utilizado para predecir el precio del oro a nivel internacional donde se toman cuatro variables económicas para lograr esta predicción (Villada et al., 2016), con buenos resultados, esto muestra la gran utilidad de las redes neuronales en diferentes ámbitos, y se puede apreciar como con mayor cantidad de variables en este trabajo, se logran muy buenos resultados en la predicción del grado de riesgo psicosocial. Los árboles de decisión utilizados para el mejoramiento de procesos para identificar y reducir los tiempos muertos de una línea de producción (Garcés et al., 2017), también contribuyen en los dominios Control sobre el trabajo y Liderazgo y relaciones sociales en el trabajo, viendo como la variable experticia de rol influye en el mejoramiento de las condiciones de productividad y de comportamiento ético dentro de las organizaciones. Para futuras líneas de investigación, es necesario recopilar datos cuantitativos usando métodos fisiológicos (como ECG, EEG, EMG y EDA; Caywood et al., 2017) psicológicos y epidemiológicos para generalizar las relaciones entre los factores de riesgo que afectan el estrés laboral con el fin de obtener una mejor precisión en el uso de los algoritmos de machine learning, se puede desarrollar un método, modelo o metodología, que considere mejorar la correcta clasificación de instancias y reduzca aún más la dimensionalidad de los datos epidemiológicos existentes logrando altos niveles de clasificación (accuracy).

## CONCLUSIONES

Con base en los resultados encontrados y su discusión se concluye que: 1) Los cuatro grandes factores de riesgo, generales, físicos, ambientales y psicológicos, constituyen un tema complejo dentro de la organización laboral, pues afectan el desempeño de los trabajadores. 2) La alta dimensionalidad de encuestas, evaluaciones y baterías afectan la eficacia entre las relaciones de los factores estadísticamente. 3) Las técnicas de machine learning ha probado ser una herramienta eficaz para el análisis de estos factores teniendo una dimensionalidad baja y permitiendo obtener muy buenos resultados en cuanto a las instancias correctamente clasificadas y mostrando bajas tasas de error. 4) Los resultados de la predicción experimentales en la reducción individual de la dimensionalidad con cada método presentó una mejora en cuanto a la experimentación inicial, lo que conlleva a realizar la experimentación con los cinco métodos para reducir la dimensionalidad, sin afectar completamente todos los factores que se evalúan desde la seguridad y salud en el trabajo, teniendo en cuenta la legislación en materia ocupacional actual del país. 5) Se recomienda la utilización de algoritmos de machine learning para seguir realizando experimentaciones con datos epidemiológicos de riesgo psicosocial que permitan a los especialistas tener herramientas e instrumentos que faciliten la labor de los sistemas de gestión de seguridad y salud en el trabajo y mejoren la toma de decisiones a nivel gerencial para el mejoramiento continuo de las organizaciones.

## AGRADECIMIENTOS

Los más sinceros agradecimientos a la Universidad Nacional de Colombia, sede Manizales, por su apoyo en la realización del proyecto que actualmente adelanta el autor del artículo, de la Tesis Doctoral perteneciente al programa de Doctorado en Ingeniería - Industria y Organizaciones cuyos resultados parciales se ilustran en el presente trabajo.

**REFERENCIAS**

- Anderson, J. A., *Redes Neuronales*, 1ª Ed., Alfaomega Grupo Editor, México, D.F. (2007)
- 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)
- Caywood, M.S., D.M. Roberts y otros tres autores, Gaussian Process Regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks, doi: 10.3389/fnhum.2016.00647, *Frontiers in Human Neuroscience*, 10, 647 (2017)
- Coca, G.A., O.D. Castrillón y S. Ruiz-Herrera, Programación de un Sistema de Fabricación tipo "Job Shop" bajo un Enfoque de Sostenibilidad, doi: 10.4067/S0718-07642016000600005, Inf. Tecnol., 27(6), 31-52 (2016)
- Diabat, A. y R. Deskoors, A hybrid genetic algorithm based heuristic for an integrated supply chain problem, *Journal of Manufacturing Systems*, 38, 172-180 (2016)
- Garcés, D.A. y O.D. Castrillón, Diseño de una Técnica Inteligente para Identificar y Reducir los Tiempos Muertos en un Sistema de Producción, doi: 10.4067/S0718-07642017000300017, Inf. Tecnol., 28(3), 157-170 (2017)
- Gupta, N., Artificial neural network, ISSN 2225-0603, *Network and Complex Systems*, 3(1), 24-28 (2013)
- Hall, M. A., *Correlation-based Feature Subset Selection for Machine Learning*, Hamilton, New Zealand (1999)
- Jiang, L., C. Li., S. Wang y L. Zhang, Deep feature weighting for naive Bayes and its application to text classification, *Eng. Appl. Artif. Intell.*, 52, 26-39 (2016)
- Jolliffe, I.T. y J. Cadima, Principal component analysis: a review and recent developments, doi: 10.1098/rsta.2015.0202, *Philosophical transactions, Series A, Mathematical, physical, and engineering sciences*, 374(2065), 20150202 (2016)
- Jolliffe, I.T., *Principal Component Analysis*, 2ª Ed., Springer (2002)
- Jordan, M. I. y T. M. Mitchell, *Machine learning: Trends, perspectives, and prospects*, *Science*, 349(6245), 255-260 (2015)
- Kourou, K., T.P. Exarchos y otros tres autores, Machine learning applications in cancer prognosis and prediction, *Computational and Structural Biotechnology Journal*, 13, 8-17 (2015)
- Langenhan, M.K., S. Leka y A. Jain, Psychosocial risks: is risk management strategic enough in business and policy making? *Saf Health Work*, 4, 87-94 (2013)
- Leka, S., Psychosocial risk management: calamity or opportunity? *Occupational Medicine*, 66(2), 1, 89-91 (2016)
- Leka, S., W. Van Wassenhove y A. Jain, Is psychosocial risk prevention possible? Deconstructing common presumptions, *Safety Sci.*, 71 (1), 61-67 (2015)
- Liang, Y., M. Zhang y W.N. Browne, Image feature selection using genetic programming for figure-ground segmentation, *Eng. Appl. Artif. Intell.*, 62, 96-108 (2017)
- 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)
- Ott, L. y M. Longnecker, *An introduction to statistical methods and data analysis*, 7ª Ed., Belmont, CA, Brooks/Cole Cengage Learning, USA (2016)
- Patil, T.R. y S.S. Sherekar, Performance analysis of Naive Bayes and J48 classification algorithm for data classification, *International Journal of Computer Science and Applications*, 6(2), 256-261 (2013)
- Perikos, I. e I. Hatzilygeroudis, Recognizing emotions in text using ensemble of classifiers, *Eng. Appl. Artif. Intell.*, 51, 191-201 (2016)
- Quinlan, J.R., *C4. 5: programs for machine learning*, Elsevier, Burlington, Canada (1993)
- Stajic, J., R. Stone, G. Chin y B. Wible, Rise of the Machines, *Science*, 349(6245), 248-249 (2015)
- Tuya, J., R. I. Ramos y C.J. Dolado, *Técnicas cuantitativas para la gestión en la ingeniería del software*, 1ª Ed., Netbiblo, La Coruña, España (2007)
- Villalobos, G., A.M. Vargas y otros tres autores, *Batería de instrumentos para la evaluación de factores de riesgo psicosocial*, Bogotá, Ministerio de la Protección Social (2010)
- Villada, F., N. Muñoz y E. García-Quintero, Redes Neuronales Artificiales Aplicadas a la Predicción del Precio del Oro, doi: 10.4067/S0718-07642016000500016, Inf. Tecnol., 27(5), 143-150 (2016)
- Wang, J., Pearson correlation coefficient. In *Encyclopedia of Systems Biology*, Springer, New York, p. 1671 (2013)
- Whitley, D., A genetic algorithm tutorial, *Statistics and Computing*, 4(2), 65-85 (1994)

