REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (35). Páginas 100-117.
Junio de 2023. ISSN: 1886-516X. D.L: SE-2927-06.
www.upo.es/revistas/index.php/RevMetCuant/article/view/5377

Características del hogar y pobreza: una aplicación de las máquinas de soporte vectorial

RAHMER, BRUNO DE JESÚS
Fundación Universitaria Tecnológico Comfenalco (Colombia)*
Correo electrónico: brunodejesus.2509@gmail.com

GARZÓN SAÉNZ, HERNANDO*
Correo electrónico: hgarzons2019@gmail.com

ORTIZ PIEDRAHITA, GUSTAVO*
Correo electrónico: gustavoaop@gmail.com

SOLANA GARZÓN, JOSÉ*
Correo electrónico: ingjosemsolanag@gmail.com

RESUMEN

El uso de técnicas cuantitativas para la clasificación de segmentos poblacionales es una fase crítica para evaluar sus condiciones de materiales de existencia, información que sirve como input para los procesos de planificación de estrategias dirigidas a paliar la pobreza y la intervención discrecional de tales grupos, bajo los criterios de racionalidad económica e instrumental. En este artículo se construye un modelo de máquinas de soporte vectorial, entendido éste como un algoritmo de aprendizaje supervisado que proporciona un clasificador lineal no probabilístico con un superlativo nivel de precisión. De este modo, se segmenta una muestra de núcleos familiares residentes en Cartagena de Indias, en función de ciertas variables económicas y sociodemográficas. La obtención de los resultados analíticos refrenda el hecho de que los factores con mayor poder de discriminación entre los agentes económicos son el estatus laboral, la accesibilidad a servicios públicos y la renta percibida por los núcleos familiares. Por otra parte, se corrobora que las condiciones de vecindario y la recepción de transferencias monetarias corrientes tienen un poder clasificatorio reducido.

Palabras clave: algoritmo de aprendizaje; hogares censales; máquinas de soporte vectorial; métodos de clasificación; pobreza.
Clasificación JEL: C00, M00.
MSC2010: 62P20, 68T01, 62P25.

Artículo recibido el 16 de octubre de 2020 y aceptado el 29 de agosto de 2022

Household characteristics and poverty: an application of support vector machines

ABSTRACT

The use of quantitative techniques for the classification of population segments is a critical phase to evaluate their conditions. This information will serve as input for planning strategies to alleviate poverty. In this article, we present a model of vector support machines. Consequently, a sample of families residing in Cartagena de Indias is segmented, based on certain economic and sociodemographic variables. Analytical results confirm that most important factors are employment status, accessibility to public services and familiar income. In addition, it is corroborated that neighborhood conditions and monetary transfers have a low discriminatory power.

Keywords: learning algorithm; household data; support vector machines; classification methods; poverty.
JEL classification: C00, M00.
MSC2010: 62P20, 68T01, 62P25.

1. Introducción

La condición de pobreza tiene como implicación la carestía de recursos para la satisfacción de necesidades biopsicosociales elementales. La privación refleja diferentes grupos de necesidades humanas insatisfechas, por lo que su delimitación conceptual permite dilucidar los factores causales más relevantes, así como también, la determinación de la naturaleza agregativa de las distintas tipologías de privaciones. No es una realidad vedada que un amplio segmento de núcleos familiares y colectivos en el espacio geográfico colombiano, afronta una escabrosa problemática de pobreza abyecta y escasez de bienes y/o servicios esenciales. Por tanto, el curso de acción que ha de ser acometido para sortear tales circunstancias indeseables, es el diagnóstico de las condiciones materiales de tales grupos poblacionales. Asimismo, es menester la vertebración de estrategias tendentes a reducir la probabilidad de recaída en la deprivación absoluta y el detrimento de las condiciones materiales de los agentes económicos. Tales operaciones coordinadas implican el concurso activo de las fuerzas mercantiles y los órganos de planificación central.

En virtud de lo esgrimido anteriormente se deriva el interés por caracterizar la dinámica de los agentes económicos en estado de vulnerabilidad y la comprensión de los factores configurantes de su estatus socioeconómico. Tradicionalmente la medición y el análisis de la pobreza se ha fundamentado en la evaluación de variables de orden económico como los ingresos primarios, tasa de consumo o ahorro. Sin embargo, éstos no son los únicos indicadores apropiados para explicar el nivel de vida, en tanto que el bienestar material no es reductible a un fenómeno monetario. Por ello es una faena de gran complejidad capturar los rasgos distintivos de las poblaciones subsumidas en estado de deprivación, máxime cuando múltiples dimensiones distales y de orden estructural, que se hallan inextricablemente entrelazadas, inciden simultáneamente en el status socioeconómico de los agentes.

El objetivo ulterior de este paper es examinar exhaustivamente la incidencia de variables sociodemográficas y económicas en el status de una muestra significativa de núcleos familiares que residen en la ciudad de Cartagena de Indias (Colombia) y presentan condiciones vitales deterioradas, relativamente heterogéneas. La segmentación de los agentes económicos se realiza a partir de la inclusión de predictores cuantitativos y categóricos. Para tal efecto, se propone la construcción de un modelo de máquinas de soporte vectorial, que es especialmente proficiente para el abordaje de problemáticas investigativas como la aquí comentada (Cuentas, Peñabaena-Niebles & Gar, 2017; Jara, Giral & Martínez, 2016). Las máquinas de vectores de soporte pertenecen a una compilación de métodos kernel que han de ser entendidos como rutinas de aprendizaje reunidos bajo la cubierta de la minería de datos. Así, provisto un conjunto de datos distribuidos en las dos clases señaladas anteriormente, el modelo SVM propuesto busca hallar un hiperplano, de suerte que, una fracción mayoritaria de puntos pertenecientes a una clase, se localicen en el mismo lado, además de hacer máxima la distancia de tales clases al hiperplano definido.

Puede afirmarse taxativamente, que el tópico de la pobreza reviste especial importancia en el contexto de la economía del desarrollo, amén de ser una tendencia crecientemente activa en la literatura especializada (Mohamoud, Kirby & Ehrenthal, 2019; Roos, Wall-Wieler & Boram, 2019). La relevancia de esta investigación estriba en su capacidad para proporcionar una plataforma teórica expedita para la caracterización de poblaciones en función de factores variopintos como los de orden social, económico y geográfico, mediante el uso de una técnica de clasificación avanzada.

2. Metodología

La investigación empírica se despliega en un conjunto de hogares de la ciudad de Cartagena de Indias a fin de evaluar su status socioeconómico. La muestra objeto de investigación está conformada por 122 familias que residen en sectores geográficos periféricos como las Localidades 1 y 2. El proceso de selección muestral se realiza de forma estratificada de modo que la totalidad de casos seleccionadas para el estudio se segmenta en subpoblaciones para escoger aleatoriamente a las entidades finales de

los distintos estratos, proporcionalmente. De esta manera, se posibilita la identificación de conglomerados o macro categorías altamente homogéneas que exhiben cierto grado de heterogeneidad intergrupal. En este caso particular, la respuesta solo admite dos estados, a saber: pobreza monetaria/ no pobreza monetaria. Para construir la regla discriminante se definen variables cuantitativas y categóricas de índoles económica y demográfica.

El ejercicio de clasificación se segmenta en dos fases bien distinguidas: la fase de aprendizaje automático y la fase de reconocimiento (González, Barrientos & Toa, 2017). En la primera se selecciona el conjunto de datos de entrenamiento, se extraen los atributos y características del espacio de entrada y se entrena el clasificador. Para efectos de validación, esto es, para el ajuste de los hiperparámetros contenidos en el modelo se dispone del método *k-fold* que divide a los subconjuntos de prueba/entrenamiento en grupos equivalentes en tamaño. La lógica seguida para el particionamiento de los datos se explicita a continuación: una porción mayoritaria de los casos se destina a la muestra de entrenamiento, otro subconjunto menor en tamaño de los registros comprende la muestra de prueba y la cantidad de registros remanentes corresponden a la muestra de reserva. Tras el proceso de aprendizaje se obtiene un conjunto de parámetros que definen la función discriminante, estableciéndose entonces una frontera definida entre clases o regiones.

Dado que el problema abordado no es linealmente separable, el Kernel seleccionado para la construcción del modelo de soporte vectorial es de tipo polinomial. En la fase de reconocimiento, el modelo del clasificador entrenado permite asignar a los nuevos datos de entrada una de las clases, a saber, abandono o permanencia, según la similitud de sus características. Una vez obtenido el modelo final se reportan estadísticos descriptivos, la matriz de confusión y otros estadísticos derivados del proceso de clasificación de los agentes económicos. Posteriormente se reportan ciertos indicadores que permiten evaluar la capacidad discriminante del modelo y compararlo con otros.

3. Generalidades de las máquinas de soporte vectorial

Los modelos de soporte vectorial (SVM) son sistemas de aprendizaje automatizado cuyos procesos de entrenamiento están controlados por un agente externo y están controlados por un agente externo y están propiamente relacionados con problemas de clasificación binaria o multiclase mediante la inducción de un separador lineal ya sea en el espacio original de los ejemplos de entrada, si éstos son separables o cuasi-separables o en un espacio transformado (denominado espacio de características) si los ejemplos no son separables linealmente en el espacio original. El principio inductivo asociado a esta tipología de modelo busca la minimización de riesgo estructural (Saidi, Fnaiech & Ben Ali, 2015), esto es, la reducción de la probabilidad de clasificación errónea de nuevos ejemplos. La lógica interna que rige estos modelos discriminantes es simple, pues a partir de una cantidad limitada de patrones de aprendizaje y etiquetas de clase se entrena una SVM con el objeto de que ésta aprenda una superficie de decisión apropiada. Las SVMs representan en un eje de coordenadas los vectores de entrenamiento, maximizando la distancia entre las muestras más cercanas de las distintas clases. Así, las entradas nuevas eventualmente introducidas se colocan sobre el mismo eje y en función de la proximidad de los grupos antes separados, serán clasificadas en una u otra clase (Sánchez, 2015). El grueso del desarrollo teórico sobre clasificación supervisada puede ser entendido de un modo levemente restrictivo y simplista, como un problema en el que se tiene a disposición dos clases con nula pérdida de generalidad en el que se procura separarlas a través de la inducción de separadores en espacios de alta dimensionalidad.

Sea entonces un conjunto de datos de entrenamiento que comprende pares de atributos-etiquetas $\{(x_1, y_1) \dots . (x_2, y_2) \dots (x_n, y_n)\}$ en el que $x_i \in \mathbb{R}^d \; e \; y_i \in \{-1,1\}$ para cada $i = 1,2,3 \dots . n$. El modelo de máquinas de soporte vectorial, desde una perspectiva estrictamente pragmática, halla un hiperplano de separación óptima expresado por $(\omega * x) + b = 0$ donde $\omega, x \in \mathbb{R}^d, b \in \mathbb{R}$ de entre otros alternativos, que proporciona mayor separación entre él y los casos disponibles (Suchacka,

Skolimowska-Kulig & Potempa, 2015). Para patrones linealmente separables la función discriminante queda expresada como:

$$d(\omega, x, \beta)\frac{1}{\|w\|} \rightarrow |\omega \cdot x + \beta| = 1$$

Así, se tiene que el margen de separación canónico en el que los patrones de entrenamiento más próximos al plano poseen distancia normalizada $d(\omega, x, \beta) = 1$ y para los patrones restantes $d(\omega, x, \beta) > 1$ (Jiménez & Rengifo, 2010). Una propiedad deducible inmediatamente a partir de la definición teórica de hiperplano de separación óptimo es que éste equidista del caso más próximo a cada clase configurada. Desde una óptica algorítmica y meramente procedimental, el problema de optimización del margen geométrico representa un problema de optimización cuadrática con restricciones lineales que puede ser resuelto mediante la utilización de técnicas estándares de programación no lineal en el cual, la cantidad de variables es equivalente al cúmulo de datos dispuestos para el entrenamiento (Carmona, 2014). La propiedad de convexidad exigida para su abordaje confirma la existencia de una solución única. Adicionalmente, es preciso señalar que si tal problema se transforma apelando al principio de dualidad adquiere la estructura:

$$\min \frac{\|\omega\|^2}{2}$$

sujeto a

$$y_i(x_1 * \omega + \beta) - 1 \geq 0, \qquad i = 1,2,3 \ldots . n$$

El problema de optimización propuesto se resuelve hallando el punto de silla del Langragiano Primal:

$$\mathcal{L}(\omega, \beta, \alpha) = \frac{1}{2}\langle \omega, \omega \rangle - \sum_{i=1}^{m} \alpha_1 [y_i(\langle \omega, x_i \rangle + \beta) - 1]$$

Para hallar el punto de ensilladura $(\omega_0, \beta_0, \alpha_0)$ se minimiza la función $\mathcal{L}(\omega, \beta, \alpha)$ respecto a ω y β y se maximiza respecto $\alpha_i > 0$ que representa una solución en el espacio primal. Al establecer el gradiente igual a cero se obtienen los mínimos locales:

$$\nabla L(\omega^*, \beta^*, \alpha) = \begin{bmatrix} \omega^* - \sum_{i=1}^{m} \alpha_i y_i x_i \\ \omega^* - \sum_{i=1}^{m} \alpha_i y_i \end{bmatrix} = 0$$

La solución (ω^*, β^*) satisface la condición complementaria de Kuhn-Tucker expresada como:

$$\alpha_1 [y_i(\langle \omega^*, x_i \rangle + \beta^*) - 1] = 0 \quad \forall i$$

Si se reemplazan estas condiciones en el Langrangiano original se obtiene que:

$$\max \mathcal{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i x_j, \rangle$$

sujeto a:

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

En este caso existe un a tal que $\alpha_i^* = 0$ para todo x_i que satisface $y_i(\omega^{*'} x_i + \beta^*) > 1$ y $\alpha_i > 0$ siempre que $y_i(\omega^{*'} x_i + \beta^*) = 1$. Los vectores soporte son los vectores x_i del conjunto de entrenamiento que proporcionan un multiplicador $\alpha_i > 0$ y son los más próximos a la cota de decisión. Es evidente que el problema planteado bajo esta formulación puede ser resuelto mediante el uso de técnicas ortodoxas de programación no lineal y garantizando simultáneamente el alcance de cierta economía computacional. Se tiene entonces que el valor de β^* puede ser obtenido de manera inmediata utilizando las restricciones del primal, promediando a partir de los vectores soportes. Por otro lado, ω^* puede ser expresado como una combinación lineal de los vectores de entrada. De este modo, se formula la expresión del sesgo $\hat{\beta}$ y ω^* quedan respetivamente como:

$$\beta^* = \frac{1}{2}\left[\min_{y_t=+1}(\langle \omega^*, x_1 \rangle) + \max_{y_t=-1}(\langle \omega^*, x_1 \rangle)\right]$$

$$\omega^* = \sum_{i=1}^{m} \alpha_i y_i \, x_i$$

Y el clasificador o función de decisión $f(x)$ para el hiperplano se construye como:

$$f(x) = sign(\sum_{i=1}^{m} \alpha^*{}_i y_i \langle x, x_i \rangle + \beta^*)$$

Ahora bien, para el caso no separable, las restricciones de separabilidad son relajadas mediante la introducción de una penalización por clasificaciones erróneas $\xi(i = 1,2,3 \ldots . n)$ de modo que comparece un problema de minimización del tipo:

$$\min \frac{\|\omega\|^2}{2} + \mathbb{C} \sum_{i=1}^{m} \xi_i^k$$

sujeto a

$$y_i(x_1 * \omega + \beta) - 1 + \xi_i \geq 0 \quad , \quad i = 1,2, \ldots . n \quad y \quad \xi_i \geq 0$$

siendo el $\mathbb{C}$ una constante regularización determinada vía validación cruzada. Un incremento de este valor permite la obtención de un margen más estricto, enfatizando en minimizar el número de clasificaciones erróneas; una disminución de este valor, por el contrario, hace permisible la ocurrencia de más infracciones en el proceso clasificatorio (Awad & Khanna, 2015). Como consecuencia de esta reformulación el Langrangiano primal se modifica:

$$\mathcal{L}(\omega, \beta, \alpha) = \frac{1}{2}\langle \omega, \omega \rangle + \mathbb{C} \sum_{i=1}^{m} \xi_i^k - \sum_{i=1}^{m} \alpha_1 \left[y_i(\langle \omega, x_i \rangle + \beta) - 1 + \xi_i^k \right] - \sum_{i=1}^{m} \delta_i \xi_i^k$$

siendo δ_i los multiplicadores de Lagrange que hacen $\xi_i^k > 0$. Luego entonces, la formulación del problema en su forma dual resultaría siendo una obviedad:

$$\max \mathcal{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i x_j, \rangle$$

sujeto a:

$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq \mathbb{C}$$

4. Resultados y discusión

Este epígrafe está segmentado en cuatro subsecciones bien distinguidas. En la primera parte se explica de modo pormenorizado el proceso de pretratamiento y depuración de la información recabada. Asimismo, se reportan algunos estadísticos descriptivos (de tendencia central, dispersión y localización) asociadas a las variables explicativas que son de naturaleza continua. Para aquellas variables categóricas se ha explicitado el procedimiento de codificación, consistente en la traslación de información de tipo alfanumérico a numérico, lo que facilita ostensiblemente la interpretación de los resultados estimados. De forma inmediata se explica la lógica interna del particionamiento de los patrones de entrada destinados para muestra de reserva, de validación y de entrenamiento del modelo de soporte vectorial. Subsecuentemente se describen las propiedades matemáticas del kernel seleccionado, así como también, el procedimiento de ajuste de los hiperparámetros del modelo, denominado *validación cruzada.* En la última fase, se relaciona una batería de estadísticos útiles para diagnosticar el perfomance del modelo obtenido.

4.1. Caracterización de variables y datos de entrada

En este caso, se seleccionarán de forma intencional a 227 familias de composición heterogénea pertenecientes a los estratos 1 y 2. La cobertura geográfica del estudio abarca varias localidades de la zona suroriental y suroccidental de Cartagena de Indias.

El tipo de muestreo es bietápico con estratificación en las unidades de primera etapa. Las unidades de muestreo de primera etapa son los barrios de estrato 1 y 2 de la zona suroccidental y suroriental y las de segunda etapa son las familias residentes en estos espacios geográficos. Acótese que en el interior de las familias no se efectúa ningún proceso de submuestreo. Se hipotetiza que las variables predictoras insertas en el estudio explicarían el estatus socioeconómico de la muestra de familias seleccionadas. En ese orden de ideas se segmentan las propiedades mensurables en dos tipos bien distinguidos: variables independientes escaladas métricamente y de tipo continuo y variables de naturaleza cualitativa que son relevantes para efectos discriminativos.

Para las variables de naturaleza cuantitativa el cambio de escala se realiza con el objeto de compactar el espectro de valores admisibles y eliminar la independencia de las variables respecto a las unidades de medida. El ajuste se hace bajo un esquema de normalización tradicional, de modo que los datos quedan expresados puntuaciones estándares. Para el caso de las variables categóricas se ha utilizado un sistema de codificación dummy, asignando coeficientes numéricos a las modalidades observadas en la base de datos construida. Tal procedimiento, mejora patentemente el tratamiento de estas variables para su posterior inclusión en el modelo a estimar. No obstante, la identificación de las categorías precisa de un elevado grado de subjetividad que puede repercutir en la sensibilidad del modelo estimados. Así, en la Tabla 1 se registra la descripción para las variables cuantitativas y las codificaciones respectivas para las variables categóricas. Debido a que existen múltiples variables de entrada el tiempo de convergencia del algoritmo de aprendizaje será relativamente alto.

Tabla 1. Caracterización de Variables de entrada.

TABLA DE PREDICTORES

VARIABLE RESPUESTA	CATEGORÍA	DESCRIPCIÓN	VALORES ADMISIBLES
Estatus socioeconómico de la familia	Multidimensional	Este estado de pobreza del infante está representado por una variable dicotómica que identifica al infante como pobre multidimensionalmente o no	0: Estado de pobreza multidimensional 1: No se encuentra en estado de pobreza multidimensional

CATEGORÍA	VARIABLE	DESCRIPCIÓN	VALORES ADMISIBLES
VARIABLES ECONÓMICAS Y SOCIODEMOGRÁFICAS	Renta familiar	Variable cuantitativa. Representa el nivel de renta percibido por núcleo por núcleo familiar	$(0, P)$
	Años de escolaridad	Variable cuantitativa. Indica el tiempo de escolaridad los cabezas de familia, medidos en años	$(0, P)$
	Transferencias monetarias	Variable cuantitativa. Es una variable agregada que recoge las transferencias monetarias condicionadas y no condicionadas, receptada por núcleo familiar	$(0, P)$
	Estatus laboral	Variable cualitativa. Indica el tipo de empleabilidad en que se encuentran los jefes familiares	0: Empleado formal 1: Empleado informal 2: No empleado
	Accesibilidad a servicios públicos	Variable categórica. Indica la posibilidad del núcleo familiar para acceder a servicios públicos esenciales y complementarios	0: Acceso a todo al elenco de servicios públicos 1: Acceso a servicios públicos esenciales 2: Acceso deficiente a servicios públicos
	Condiciones de vecindario	Variable categórica. Indica la frecuencia de ocurrencia de hechos violentos en el entorno donde subyacen los hogares	0: No ocurrencia de hechos violentos en el último año 1: Ocurrencia de cinco hechos violentos, como máximo 2: Ocurrencia de más de cinco hechos violentos, como máximo

Fuente: Elaboración propia.

En la Tabla 2 se realiza un análisis exploratorio de la totalidad de variables identificadas en el proceso de recolección de información relevante para los propósitos investigativos. Una vez definidas las variables de estudio, se procede a efectuar un análisis descriptivo de las mismas.

Para variables numéricas, en las que puede existir una cantidad considerable de valores observados de naturaleza distinta se ha de optar por un método de análisis distinto. En primera instancia, se reportan los estadísticos de resumen para cada variable. Se reportan, además, las medidas de dispersión que describen el grado de lejanía de las observaciones con respecto a la medida de la tendencia central. Igualmente, se registran medidas de la tendencia central proporcionan un guarismo que resume la distribución media de una variable.

Nótese que la renta media de los hogares no pobres, corregida por tamaño familiar, es superior a la renta familiar media de las familias en estado de deprivación. La desviación estándar en ambos segmentos poblacionales es bastante similar. Aunque el promedio de las transferencias monetarias receptadas en los hogares pobres y no pobres difiere, la mediana de ambos grupos es similar. Nótese igualmente, que el tiempo de escolaridad medio de los jefes de núcleo familiar que no se hayan en estado de deprivación es aproximadamente dos veces mayor que en sus contrapartes que se hayan en

estado de pobreza. También es particularmente interesante que el tiempo de escolaridad más frecuente de los jefes familiares de estos hogares, es de 1,15 años.

Tabla 2. Análisis descriptivo de variables cuantitativa.

	No pobreza / Pobreza									
	0					1				
	Media	Mediana	Moda	Rango	Desviación estándar	Media	Mediana	Moda	Rango	Desviación estándar
Renta Inicial (Miles de pesos)	834,65	854,80	417,50	776,08	31,71	610,63	621,03	422,31	475,78	34,27
Transferencias monetarias (Miles de pesos)	152,50	227,79	281,17	194,45	33,79	271,19	280,58	300,00	475,48	29,54
Tiempo de escolaridad	4,297	5,05	5,47	3,10	8,30	2,04	3,65	1,15	5,13	17,58

Fuente: Elaboración propia.

En la Tabla 3 se muestra un análisis descriptivo de las variables cualitativas consideradas en el estudio. Percíbase que las características más comunes de los grupos de hogares pobres pueden ser resumidas del siguiente modo. El 71,4% de los jefes de hogar se halla en estado de informalidad y sólo el 7,9% de los jefes de estos grupos están vinculados formalmente a un empleo directo. Asimismo, es perceptible que el 92,5% de los hogares ha reportado frecuentemente hechos delictivos. Un 54% de los hogares tiene acceso al elenco de servicios esenciales ofertados en el distrito, mientras que un 22,2% no acceden a una gama de servicios no esenciales y esenciales.

Tabla 3. Análisis descriptivo de variables categóricas.

		0		**1**	
		Recuento	**% de N totales de columna**	**Recuento**	**% de N totales de columna**
Estatus Laboral	0	40	59,7%	5	7,9%
	1	21	31,3%	45	71,4%
	2	6	9,0%	13	20,6%
Condiciones del vecindario	0	23	36,5%	2	3,0%
	1	28	44,4%	3	4,5%
	2	12	19,0%	62	92,5%
Accesibilidad a servicios públicos	0	47	70,1%	34	54,0%
	1	13	19,4%	15	23,8%
	2	7	10,4%	14	22,2%

Fuente: Elaboración propia.

Por otro lado, los hogares que no se encuentran en estado de pobreza multidimensional tienen las siguientes características. El 59,7% de los jefes de hogar se encuentra vinculado a empleo formal y sólo el 9% de los encuestados reporta estar desempleado. El 36,5% de los hogares no ha reportado

hechos violentos en su vecindad y el 19% de los hogares en este estado reportan una cantidad considerable de hechos de esta naturaleza. Mientras tanto, el 70,1% de estos hogares tienen acceso total a la gama de servicios ofertados en el distrito.

4.2. Creación de la partición de datos

La mecánica interna del particionamiento de los datos es la siguiente: el 60% de los casos se destinan a la muestra de entrenamiento; el 30% de los registros comprenden la muestra de prueba. Esta última partición se crea con el objeto de ejecutar un seguimiento. Los registros remanentes conciernen a la muestra de reserva. Los conjuntos de validación y test están constituidos por un conjunto de registros inmediatamente posteriores a los datos que conforman la muestra destinada para el entrenamiento. El algoritmo de aprendizaje crea un modelo tentativo a partir de una fracción muestral reducida antes de trasladarlo a la totalidad de registros destinados para el entrenamiento. Posteriormente, éste se actualiza gradualmente con base en los outputs del ciclo. Este procedimiento se aplica en reiteradas ocasiones hasta alcanzar la convergencia en una cantidad máxima permisible de vectores soporte.

4.3. Selección de Kernels

La función Kernel que se define como un producto escalar de dos puntos bajo el mapeo ϕ y puede ser definida formalmente del siguiente modo:

Sea un X un conjunto no vacío. Si se tiene una función $K: X \times X \to \mathbb{K}$ es simétrica y semidefinida positiva existe un espacio de Hilbert H y un mapa $\phi: X \to H$ tal que para todo $x, x' \in X$ y se tiene:

$$k\langle x, x'\rangle \coloneqq \langle \phi(x), \phi(x')\rangle$$

siendo ϕ un mapa de características y H a un espacio de características de k. Para el problema que nos ocupa se utiliza un Kernel de tipo Polinomial que adquiere la forma $k\langle z, z'\rangle \coloneqq (\langle z, z'\rangle + c)^m$ donde $z, z' \in \mathbb{C}^d$ siendo $m > 0$, $d \geq 1$ enteros y $c > 0$ u número real. La calibración de los parámetros se ejecuta de modo automático.

4.4. Validación cruzada del modelo y entrenamiento

La validación cruzada es un procedimiento elemental para ajustar los hiperparámetros del modelo de soporte vectorial a través del uso de múltiples conjuntos de prueba/entrenamiento partiendo de los datos disponibles (Wainer & Cawley, 2017) y sortear el indeseado fenómeno del sobreajuste. La validación cruzada *k-fold* divide equitativamente el conjunto de datos en k subconjuntos. Un subconjunto se usa como el conjunto de prueba ℸ, mientras que el resto $(k - 1)$ de subconjuntos forman el conjunto de entrenamiento X (Wen et al., 2017). Para este caso, el proceso de validación cruzada se aplica en 5 iteraciones con cada uno de los posibles subconjuntos de datos de prueba. Los resultados para las k repeticiones posteriormente son promediados para producir una única medida de estabilidad que permita evaluar la precisión del modelo de soporte vectorial para datos no observados.

El proceso de aprendizaje es bi-etápico. En la fase inicial, el algoritmo de entrenamiento inicia una búsqueda exhaustiva de un estimado de la constante de regularización-simbolizada con el grafema ℂ- para lograr una precisión clasificatoria colosal. Una vez calculado el valor de ℂ, en la fase subsiguiente, este valor es seleccionado para entrenar el modelo utilizando íntegramente la partición reservada para tal ejercicio.

4.5. Obtención del modelo final y evaluación

En el cuadro resumen reportado en la Tabla 4 se listan las especificaciones del modelo SVM, incluida la cantidad de vectores de soporte y sus ponderaciones respectivas, la función kernel utilizada y los hiperparámetros, en conjunto el valor de la constante de regularización ℂ que es igual a 6 para el caso

presente. También se reportan los resultados de la validación cruzada y las estadísticas de clasificación para el entrenamiento y las pruebas. La cantidad de vectores soporte es igual a 12 y tales participan de forma directa en la definición del hiperplano de separación óptimo. No existe una cantidad extensiva de vectores soporte y por ello no existe una pérdida en la parsimonia del modelo.

La aplicación de un principio inductivo, cuyo objeto no es otro sino el de minimizar del riesgo estructural le atribuye al modelo formulado una superlativa capacidad de generalización que permite generar predicciones con relativo acierto, tal y como se percibirá a la postre. Obsérvese que la acuracidad del proceso de validación cruzada es superior al 93,75%.

Tabla 4. Resumen del modelo.

Tipo de SVM: Clasificatorio con $\mathbb{C} >$ =4,00		
Tipo de Kernel: Lineal		
Número de Vectores Soporte: 8		
Precisión de la validación cruzada (%) = 90,75%		
Vectores Soporte por categoría:	Pobreza: 4	No pobreza: 4

PESOS	**VECTORES SOPORTES**					
	Renta Familiar	Tiempo de escolaridad	Transferencias monetarias	Estatus laboral	Accesibilidad a servicios públicos	Accesibilidad a oferta institucional
4,00	3,20	5	2,22	1,02	1,3	1,1
5,23	4,60	3	2,36	0,34	1,6	1,2
2,84	3,10	4	2,12	0,25	1,9	1,3
1,23	3,10	2	1,67	0,23	2,1	0,3
-4,10	4,05	6	1,43	1,76	0,6	0,4
-3,76	2,00	7	2,59	1,23	0,3	0,2
-0,96	3,10	9	1,71	1,50	0,2	0,6
-2,54	2,00	2	1,23	2,48	0,6	1,1

Fuente: Elaboración propia.

La matriz de confusión que se presenta en la Tabla 5 es una matriz $M = (m_{ij})$ que consta de entradas $m_{ij} = \#\{x: x$ se predice como perteneciente una clase i pero que pertenece en realidad a la clase j y es útil para corroborar que los errores cometidos en el proceso clasificatorio no están concentrados en alguna clase (Markowetz, Edler & Vingron, 2003). A la luz de los resultados presentados es notorio que una fracción casi inescrutable de clasificaciones son erróneas.

Tabla 5. Resultados de la Matriz de Confusión.

Distribución de la muestra (%)	**Entrenamiento:** 75%	**Prueba:** 20%	
Precisión de la clasificación (%)	**Entrenamiento:** 96,7875%	**Prueba:** 88.889%	**Global:** 95,122%

Fuente: Elaboración propia.

En la Tabla 5 se registra una hoja de cálculo resultado del análisis *What if*?" en la que se identifican los comportamientos de los núcleos familiares frente a variaciones en los valores admitidos por cada una de los predictores dispuestos para la confección del modelo sucedería si se realizaran cambios en un caso de predictor particular. Para efectos de simplificación, sólo se reportan los resultados de 12 casos predichos para los que se verifica la respuesta del modelo.

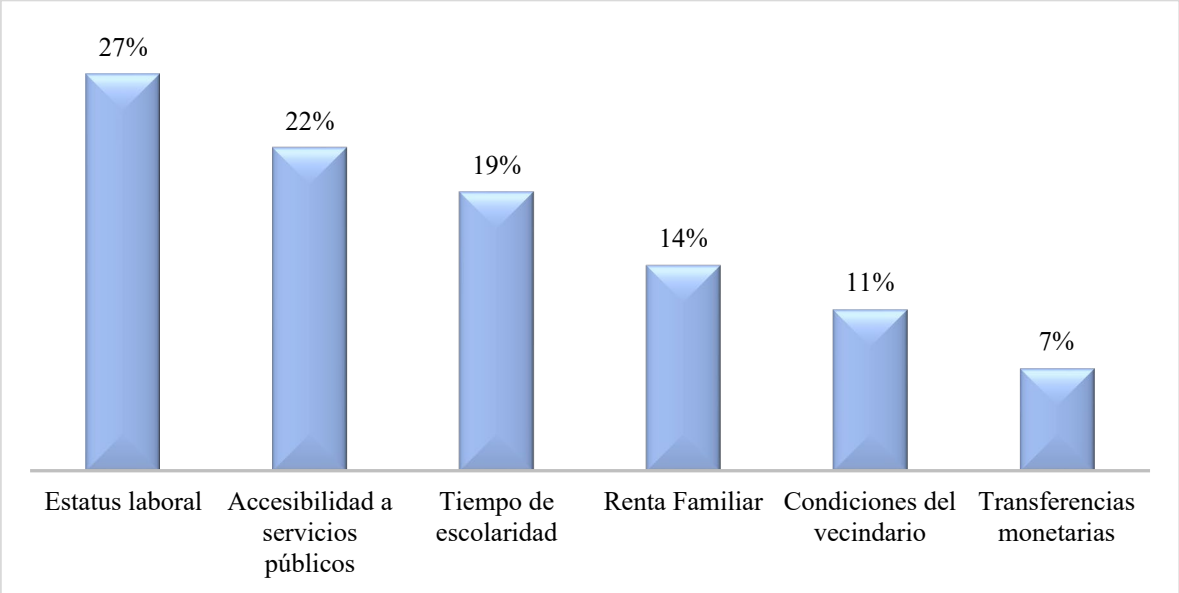
Tabla 6. Predicciones generadas.

Renta Familiar	Tiempo de escolaridad	Transferencias monetarias	Estatus laboral	Accesibilidad a servicios públicos	Accesibilidad a oferta institucional	VARIABLE RESPUESTA
4,01	4	4,28	0	1	0	P(✓)
5,02	6	2,52	0	1	1	P (✓)
9,8	8	1,91	0	0	1	P (✓)
4,7	9	2,01	0	0	0	P (✓)
5	11	1,80	1	1	1	No P (•)
6	3	1,20	0	0	0	P (✓)
1,3	6	2,10	1	1	1	No P (✓)

Fuente: Elaboración propia.

En la Ilustración 1 se muestra la importancia relativa para clasificar entre grupos socioeconómicos. Nótese que las tres variables con mayor poder discriminatorio son, en orden decreciente, Estatus laboral, Accesibilidad a servicios públicos y Tiempo de escolaridad con 27%, 22% y 19%. respectivamente. La variable con menor relevancia corresponde a la tasa de consumo familiar, que sólo tiene una importancia marginal de siete puntos porcentuales.

Ilustración 1. Importancia relativa de cada variable.



Fuente: Elaboración propia.

En la Tabla 7 se registran algunos estadísticos utilizados para la diagnosis de los modelos de soporte vectorial. Nótese que las métricas consignadas no difieren significativamente en ambas submuestras (entrenamiento y validación). Así, por ejemplo, se percibe que tanto la sensibilidad como especificidad del modelo es próxima al 100%. En consecuencia, la probabilidad de clasificación de la fracción de verdaderos positivos y verdaderos negativos es sumamente alta y la tasa de falsos de positivos (que representa la tasa de casos negativos que el modelo detecta como positivos pobres) es casi inescrutable.

Tabla 7. Estadísticos de desempeño del modelo.

Estadístico	Conjunto de Entrenamiento	Conjunto de Validación
Especificidad	0,93	0,94
Precisión	0,91	0,92
Exhaustividad	0,90	0,87
F-score	0,91	0,90
FPR	0,89	0,86
Índice Kappa de Cohen	0,88	0,87

Fuente: Elaboración propia.

El modelo provee una buena acuracidad incluso para la predicción de clases minoritarias. Debido a que la medida de exhaustividad es próxima a 1 se evidencia que una fracción de instancias relevantes son plenamente recuperadas. Las medidas de exhaustividad y precisión son resumidas en una única métrica de rendimiento rotulada como F-Score que se expresa como la media armónica ponderada de aquellas. Por otro lado, el índice Kappa de Cohen muestra que la calidad de acuerdo es sustancial y que, por tanto, el performance del modelo respecto a uno que simplemente "vaticina" azarosamente la pertenencia a clases según sus frecuencias, es sensiblemente superior.

Es necesario utilizar variopintos indicadores cuantitativos que estimen la discrepancia entre valores prospectados y los valores observados. El objetivo en este aparte será efectuar un análisis comparativo del poder discriminante entre distintos modelos. En consecuencia, se reportan los siguientes indicadores: Se registra el valor del RMSE, que corresponde a la raíz cuadrada del error medio. También se muestra el MSE, que cuantifica el error cuadrado promedio de las predicciones y viene dado por la fórmula $\mathrm{MSE} = \frac{1}{\mathrm{T}}\left[\sum_{t=1}^{\mathrm{T}}(\mathrm{Y_{ts}} - \mathrm{Y_{ta}})^2\right]$. Igualmente se presenta el error absoluto medio, expresado en términos porcentuales (MAPE). Este indicador es mucho más intuitivo que el RMSE, en tanto que no implica una estimación de la media para medir la magnitud del error. Está dado por la siguiente fórmula: $\mathrm{MAPE} = \frac{1}{\mathrm{T}}\left[\sum_{t=1}^{\mathrm{T}}\frac{\mathrm{Y_{ts}}-\mathrm{Y_{ta}}}{\mathrm{Y_{ta}}}\right] * 100\%$. Finalmente se muestra la función de entropía cruzada, que es una métrica dada por la siguiente expresión $\mathrm{H}(p,q) = -\sum_x p(x)\log q(x)$. Esta función de costo es una medida de precisión para variables categóricas que describe la pérdida entre pares de distribuciones de probabilidad.

La Tabla 8 consigna un análisis comparativo entre diversas funciones de costo para otros modelos de clasificación. Nótese que el modelo de SVM presenta una raíz cuadrática del error medio y un error absoluto medio inferior a los demás modelos. Sin embargo, la regresión logística ponderada tiene un MAPE inferior y el modelo de perceptrón multicapa la entropía cruzada más baja.

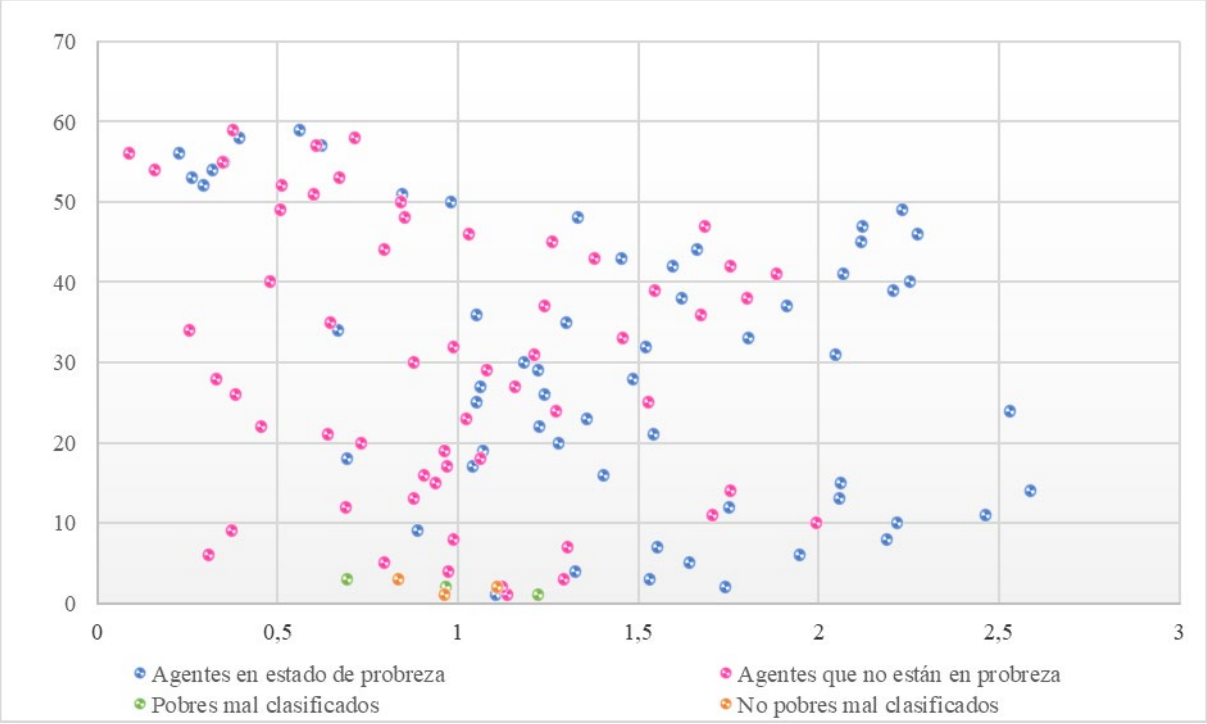
Tabla 8. Análisis comparativo entre modelos de clasificación.

MODELO	RMSE	MSE	MAPE	ENTROPÍA CRUZADA
Máquina de Soporte Vectorial	**3,630**	**3,271**	3,123	0,756
Regresión Logística	5,739	4,695	5,420	0,920
Regresión Logística Ponderada	6,557	6,634	**3,853**	0,943
Random Forest	4,335	5,291	4,394	0,944
Árbol de decisión	7,587	5,613	5,453	0,037
Perceptrón Multicapa	3,359	3,991	3,934	**0,772**

Fuente: Elaboración propia.

En la Ilustración 2 se reproduce la máquina de soporte vectorial para casos históricos a fin de evaluar la eficiencia global del modelo construido en la fase de entrenamiento. Percíbase que la cantidad de entidades que mantienen su relación comercial o que abandonan es prácticamente marginal, en tanto que representan, en conjunto, el 4% de las clasificaciones efectuadas. Se ratifica, por tanto, la hipótesis que la eficiencia clasificatoria es significativamente alta.

Ilustración 2. Evaluación de la máquina de soporte Vectorial a partir de datos históricos.



Fuente: Elaboración propia.

A la postre se presenta una breve discusión sobre los resultados alcanzados en el proceso investigativo a la luz del estado del arte.

Para explicar la fenomenología de la pobreza se ha aducido que su emergencia y persistencia obedece a factores individuales, estructurales o fatalistas (Bastias, Cañadas, Sosa & Moya, 2019). El primer grupo de factores designa aquellas características de la individualidad que serían potenciales

determinantes de la pobreza. La segunda visión atribuye la pobreza a factores macro como, por ejemplo, sistemas institucionales, efecto agregado de las acciones sociales, entre otros. Por último, las atribuciones de tipo fatalistas, se refieren a variables que superan la esfera individual y social.

En la presente investigación se pretende evaluar la importancia de los factores causales y demás determinantes de la pobreza en la ciudad de Cartagena. Es un hecho claro observar la predominancia de variables de orden estructural, de naturaleza socioeconómica e individual.

Al analizar la situación laboral, es obvio que la probabilidad de caída en la pobreza es significativamente mayor cuando el agente económico se halla en estado de desempleo que cuando está ocupado. Ciertas ramas de la actividad económica suelen tener un "efecto protector" para la pobreza, pero requieren de una intensiva actividad intelectual, que se correlaciona positivamente con mayor tiempo de escolaridad y la consecuente adquisición de competencias laborales con valor agregado diferencial. Esto es particularmente cierto en el caso presente donde el factor "Nivel de escolaridad" tiene un potencial discriminatorio considerable. El impacto de cada nivel educativo alcanzado en la reducción de la pobreza ya ha sido corroborado en la literatura especializada. Adicionalmente, las circunstancias contextuales y el modelo productivo en ciertas localizaciones geográficas que están basadas en el sector primario con demanda de mano de obra menos calificada, con menores niveles de escolaridad y, por lo tanto, con menores remuneraciones (Arias, Sánchez & Agüero, 2018). Esto es particularmente cierto para este caso investigativo.

También es preciso mencionar que a la luz de los resultados obtenidos la baja importancia marginal de la variable "Transferencias monetarias" obedece a que la recepción de tales no permite segmentar las poblaciones en los estratos socioeconómicos previamente establecidos. Sobre esta cuestión puede aducirse que la justificación de esta interferencia estatal en las decisiones acerca de la inversión en capital humano debe depender, virtualmente, del éxito constatado de su aplicación (García, 2017). Su éxito puede ser cuantificado en: la mejora de las condiciones sociosanitarias, el rendimiento académico, el aumento del gasto en alimentos y la disminución de la probabilidad de desnutrición infantil, el aumento del gasto en vestimenta, etc. En relación con esto, debe cuestionarse el verdadero alcance causal del "efecto de la condicionalidad" sobre el estatus socioeconómico de los agentes si las cantidades absolutas son al menos similares en ambos segmentos poblacionales considerados.

Asimismo, es preciso mencionar que las modificaciones en la tasa de crecimiento del ingreso y la pobreza relativa o absoluta ocurren conjuntamente y son causadas por factores que se modifican en simultáneo e interactúan entre ellos. Por ejemplo, se ha corroborado que la tasa de crecimiento del ingreso total es dependiente de las tasas de crecimiento de los ingresos laborales y de la participación relativa de tales fuentes sobre el ingreso total. Sin embargo, la composición por fuentes de los ingresos de los hogares puede diferir en las distintas zonas geográficas. Ciertamente, a distintas tasas de crecimiento entre fuentes de ingreso la desigualdad en la distribución del ingreso será más apuntalada (Bracco, Gasparini & Tornarolli, 2019). Sin embargo, estas distinciones no han sido tomadas en cuenta para el presente estudio. En concordancia con lo expuesto en la literatura especializada es notorio que el ingreso familiar es un factor ponderante para segmentar los grupos de agentes económicos.

En este caso, se hace evidente que el capital intelectual (medido por el nivel de escolaridad) está asociado con la reducción de los niveles de pobreza. Asimismo, la prevalencia de la pobreza afecta el stock de este tipo de capital (L.Harrison & Montgomery, 2019). Es notorio que las comunidades con niveles más altos de capital intelectual tienden a tener tasas de pobreza más bajas y que la pobreza puede representar barreras para la formación de este tipo de capital, siendo esto particularmente cierto para los segmentos poblacionales más vulnerables. Tal aserción, coincide con los resultados reportados, en tanto que la formación escolar es una variable discriminante de relevancia moderada.

También es particularmente cierto que un aumento de los ingresos suele tener un efecto considerable en el acceso a los servicios públicos que usufructúan los hogares. Esta tendencia puede apuntalarse si es posible mejorar el acceso a la educación y el acceso a los servicios de salud, agua y saneamiento, sobre todo en zonas urbanas que se hayan en estado de deprivación (Sanogo, 2019). Tal

veredicto conclusivo, que circula en el ámbito de lo normativo, puede estar en concordancia con los resultados obtenidos en esta investigación.

4. Consideraciones finales

En este paper se propuso la formulación y estructuración de un modelo cuantitativo avanzado para solucionar un problema de naturaleza no algorítmica como es la caracterización y perfilación socioeconómica. De este modo, es posible capturar las dinámicas socioeconómicas que exhiben los segmentos poblacionales en estado de pobreza. En este caso, se construyó un modelo automatizado que emula la conducta inteligente de sistemas biológicos y que posee una superlativa capacidad de aprendizaje. Tal técnica de modelamiento es expedita para la extracción de insight en cúmulos masivos de información, imbuida de excepcional complejidad. La intuición subyacente del modelo construido para el caso presente es la búsqueda de un hiperplano que proporcione la separación máxima entre dos clases no linealmente separables como son el grupo de familias en estado de pobreza y el grupo de familias que no se hallan en este estado. Ambos grupos conviven en la ciudad de Cartagena de Indias y poseen características distintivas que confieren cierto grado de heterogeneidad intergrupal. Entre los hallazgos principales se anotan, por ejemplo: que el estatus laboral, la accesibilidad a servicios públicos esenciales y la renta percibida por los núcleos familiares son predictores significativos del estatus socioeconómico. Mientras tanto, las condiciones del vecindario y la receptación de transferencias monetarias corrientes parecen no tener un potencial discriminatorio considerable.

El carácter representativo de este caso de estudio concreto permite extrapolar los elementos centrales del diseño metodológico a otros casos empíricos. Si embargo, la generalización analítica de los resultados obtenidos debe realizarse con cautela dado que las particularidades del contexto y la diacrónica evolución de los factores que marcan el sendero dinámico del status socioeconómico de los agentes este espacio geográfico puede afectar la validez de los resultados obtenidos.

En última instancia, se exhorta a orientar esfuerzos en otras líneas de investigación y en la construcción de estudios de casos en los que se identifiquen otros predictores de la pobreza multidimensional y monetaria en sectores geográficos distintos, así como también la evaluación de la plausibilidad de otras técnicas de machine learning e informática industrial reseñadas en la literatura especializada en aras de realizar, de modo verosímil, perfilaciones del status socioeconómico de ciertos grupos poblacionales.

Referencias

Arias, R., Sánchez, L., & Agüero, O. (2018). Impacto de la educación sobre la pobreza en regiones de planificación de Costa Rica. *Revista Estudios del Desarrollo Social: Cuba y América Latina , 6*(1), 1-21.

Awad, M., & Khanna, R. (2015). Support Vector Machines for Classification: Theories, Concepts, and Applications for Engineers and System Designers. En M. Awad, & R. Khanna, *Efficient Learning Machines* (págs. 39-66). Apress, Berkeley, CA.

Bastias, F., Cañadas, B., Sosa, V., & Moya, M. J. (2019). Explicaciones sobre el origen de la pobreza según área de formación profesional. *Propósitos y Representaciones, 7*(2), 107-120. https://doi.org/10.20511/pyr2019.v7n2.282

Bracco, J., Gasparini, L., & Tornarolli, L. (mayo de 2019). *Explorando los Cambios de la Pobreza en Argentina: 2003-2015*. Centro de Estudios Distributivos Laborales y Sociales.

http://sedici.unlp.edu.ar/bitstream/handle/10915/75173/Documento_completo.pdf-PDFA.pdf?sequence=1&isAllowed=y

Carmona, E. (11 de Julio de 2014). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. Recuperado el 17 de Marzo de 2018 de http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf

Cuentas, S., Peñabaena-Niebles, R., & Gar, E. (2017). Support vector machine in statistical process monitoring: a methodological and analytical review. *The International Journal of Advanced Manufacturing Technology, 91*(1-4), 485-500.

García, F. (2017). Responsabilidad y legitimidad en las transferencias monetarias condicionadas. *Diánoia, 62*(79), 193-216.

González, R., Barrientos, A., & Toa, M. (2017). Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial. *Revista Iberoamericana de Automática e Informática Industrial, 14*(4), 394-405.

Jara, J., Giral, D., & Martínez, F. (2016). Implementación de algoritmos basados en máquinas de soporte vectorial (SVM) para sistemas eléctricos: revisión de tema. *Revista Tecnura, 20*(48), 149-170.

Jiménez, L., & Rengifo, P. (2010). Al interior de una máquina de soporte vectorial. *Revista de Ciencias, 14*, 73-85.

L. Roos, L., Wall-Wieler, E., & Boram Lee, J. (2019). Poverty and Early Childhood Outcomes. *Pediatrics, 143*(6). https://doi.org/10.1542/peds.2018-3426

L.Harrison, J., & Montgomery, C. A. (2019). A spatial, simultaneous model of social capital and poverty. *Journal of Behavioral and Experimental Economics, 78*, 83-192. https://doi.org/10.1016/j.socec.2018.09.001

Markowetz, F., Edler, L., & Vingron, M. (2003). Support Vector Machines for Protein Fold Class Prediction. *Biometrical Journal, 45*(3), 377-389.

Mohamoud, Y., Kirby, R., & Ehrenthal, D. P. (2019). Poverty, urban-rural classification and term infant mortality: a population-based multilevel analysis. *BMC Pregnancy Childbirth, 19*(40), 1-11. https://doi.org/10.1186/s12884-019-2190-1

Saidi, L., Fnaiech, F., & Ben Ali, J. (2015). Application of higher order spectral features and support vector machines for bearing faults classification. *ISA Transactions, 54*, 193-206.

Sánchez, N. (2015). Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario. *ODEON*(9), 113-172.

Sanogo, T. (2019). Does fiscal decentralization enhance citizens' access to public services and reduce poverty? Evidence from Côte d'Ivoire municipalities in a conflict setting. *World Development, 113*, 204-221. https://doi.org/10.1016/j.worlddev.2018.09.008

Suchacka, G., Skolimowska-Kulig, M., & Potempa, A. (2015). Classification Of E-Customer Sessions Based On Support Vector Machine. *ECMS*, 1-7.

Wainer, J., & Cawley, G. (2017). Empirical Evaluation of Resampling Procedures for Optimising SVM Hyperparameters. *Journal of Machine Learning Research, 18*(15), 1-35.

Wen, Z., Li, B., Ramamohanarao, K., Chen, J., Chen, Y., & Zhang, R. (2017). Improving Efficiency of SVM k-Fold Cross-Validation by Alpha Seeding. *AAAI*, 2768-2774.