Artículo Original
Revista Varianza - Año 2021 Nro. 18
ISSN 2789-3529 versión en línea
Páginas: 21-33

APLICACIÓN DE MACHINE LEARNING SIN SUPERVISIÓN

MACHINE LEARNING APPLICATION WITHOUT SUPERVISION

Juan Carlos Flores López[1]
Instituto de Estadística Teórica y Aplicada-UMSA, La Paz -Bolivia
✉ *caarloslopez1@gmail.com*

Artículo recibido: 2021-08-16
Artículo aceptado: 2021-09-12

RESUMEN

El presente artículo, tiene como objetivo principal el desarrollo de la aplicación de *machine learning* no supervisado. La aplicación de esta metodología se realiza considerando la Encuesta Sociodemográfica del Departamento de La Paz, realizada en el año 2015. La base de datos considerada tiene datos de migración, salud, educación, empleo, ingresos, agropecuaria, vivienda, etc. De esta se considera los 75 municipios del Departamento de La Paz y los indicadores educativos, empleo, demográficos y vivienda y dentro de esta se consideran: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

Los resultados muestran que los municipios Santiago de Huata, y Tito Yupanqui, muestran similares características respecto a los indicadores: tasas de participación, relación de masculinidad y tasa de alfabetismo. Otro *cluster* definido por la customización son los municipios Humanata, Alcapata y Ayata y son parecidos en distribución de hogares según disponibilidad de dormitorios por persona y tasa de alfabetismo. Se concluye que la customización es la mejor forma de clasificación por la forma en que se presenta en forma mucho más clara que otras formas de clasificar consideradas en el estudio.

Palabras clave: *Aprendizaje no supervisado, clasificación, dendograma*

ABSTRACT

The main objective of this article is the development of the unsupervised machine learning application. The application of this methodology is carried out considering the Sociodemographic Survey of the Department of La Paz, carried out in 2015. The database considered has data on migration, health, education, employment, income, agriculture, housing, etc. Of this, the 75 municipalities of the Department of La Paz and the educational, employment, demographic and housing indicators are considered and within this are considered: the literacy rate, participation rate, distribution of households according to availability of bedrooms and masculinity ratio.

The results show that the municipalities of Santiago de Huata and Tito Yupanqui show similar characteristics regarding the indicators: participation rates, masculinity ratio and literacy rate. Another cluster defined by the customization are the municipalities Humanata, Alcapata and Ayata and they are similar in distribution of households according to availability of bedrooms per person and literacy rate. It is concluded that customization is the best form of classification due to the way it is presented in a much clearer way than other forms of classification considered in the study.

Keywords: *Unsupervised learning, classification, dendrogram*

[1] M. Sc. en Estadística, M.Sc. en Educación Superior, Dr.c. en Educación Superior e Investigación Transdisciplinar, investigador del Instituto de Estadística Teórica y Aplicada de la carrera de Estadística, tutor de varias tesis de pregrado. Docente CEPIES (Centro Psicopedagógico y de Investigación en Educación Superior). https://orcid.org/0000-0002-5522-1949

INTRODUCCIÓN

El aprendizaje estadístico es un conjunto de herramientas para comprender los datos. Estas herramientas pueden clasificarse como supervisadas o no supervisadas. El aprendizaje estadístico supervisado implica construir un modelo estadístico para predecir, o estimar, una salida basada en una o más entradas. Este tipo de problemas ocurren en diferentes campos por ejemplo en la economía, educación, negocios, medicina, astrofísica, política pública, etc. Con el aprendizaje estadístico no supervisado, hay entradas, pero sin salida de supervisión; sin embargo, podemos aprender relaciones y estructura de tales datos.

OBJETIVO

Como objetivo principal del presente artículo mostrar los resultados de la aplicación del aprendizaje estadístico *machine learning* (aprendizaje automático) considerando el aprendizaje sin supervisión, de la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad de los 75 municipios de la ciudad de La Paz.

METODOLOGÍA

La aplicación de esta metodología se realiza considerando la encuesta sociodemográfica del departamento de La Paz, realizada en el año 2015. La base de datos considerada tiene datos de migración, salud, educación, empleo, ingresos, agropecuaria, vivienda, etc.

Machine learning se interpreta como aprendizaje automático y está estrechamente relacionado muy a menudo, con estadística computacional; una disciplina que también se especializa en hacer predicciones. El aprendizaje automático se emplea en una variedad de disciplinas científicas.

El aprendizaje sin supervisión, es aquel que solo proporciona datos de salida, sin ninguna entrada. El objetivo es descubrir una "estructura interesante" en los datos; esto a veces se llama descubrimiento del conocimiento. A diferencia del aprendizaje supervisado, no se indica cuál es el resultado deseado para cada entrada. En cambio, se formaliza la tarea como una de estimación de densidad, es decir, queremos construir modelos de la forma $P(X_i \mid \theta)$. Hay dos diferencias con respecto al caso supervisado. Primero, se escribe $P(X_i \mid \theta)$ en vez de $P(y_i \mid X_i, \theta)$; es decir, el aprendizaje supervisado es una estimación de densidad condicional, mientras que el aprendizaje no supervisado es una estimación de densidad incondicional. Segundo, X_i es un vector de características, por lo que se necesita crear modelos de probabilidad multivariados. Por el contrario, en el aprendizaje supervisado y_i generalmente es solo una variable que se está tratando de predecir, esto significa que, para la mayoría de los problemas de aprendizaje supervisado, se puede utilizar modelos de probabilidad univariados (con parámetros dependientes de la entrada), lo que simplifica significativamente el problema (Murphy, 2012).

Agrupamiento (*clustering*)

El agrupamiento es una de las técnicas más utilizadas para el análisis exploratorio de datos. En todas las disciplinas, desde las ciencias sociales hasta la biología y la informática, las personas intentan tener una primera intuición sobre sus datos identificando grupos significativos entre los puntos de datos. Por ejemplo, los biólogos

computacionales agrupan genes sobre la base de similitudes en su expresión en diferentes experimentos; los minoristas agrupan a los clientes, en función de sus perfiles de clientes, para fines de marketing dirigido; y los astrónomos agrupan estrellas en función de su proximidad espacial. (Shalev, 2014).

El primer punto que se debe aclarar es, naturalmente, ¿qué es la agrupación? Intuitivamente, la agrupación es la tarea de agrupar un conjunto de objetos de manera que los objetos similares terminen en el mismo grupo y los objetos diferentes se separen en grupos diferentes. Claramente, esta descripción es bastante imprecisa y posiblemente ambigua. Sorprendentemente, no está nada claro cómo llegar a una definición más rigurosa.

Hay varias fuentes para esta dificultad. Un problema básico es que los dos objetivos mencionados en la declaración anterior pueden en muchos casos contradecirse. Matemáticamente hablando, la similitud (o proximidad) no es una relación transitiva, mientras que el *cluster* compartido es una relación de equivalencia y, en particular, es una relación transitiva. Más concretamente, puede darse el caso de que haya una larga secuencia de objetos, $x_1, \ldots, x_m$ de manera que cada x_i es muy similar a sus dos vecinos, x_{i-1} y x_{i+1}, pero x_1 y x_m son muy diferentes. Si deseamos asegurarnos de que cada vez que dos elementos sean similares compartan el mismo grupo, entonces debemos colocar todos los elementos de la secuencia en el mismo grupo. Sin embargo, en ese caso, terminamos con elementos diferentes (x_1 y x_m) que comparten un *cluster*, lo que viola el segundo requisito (Shalev, 2014).

Un modelo de agrupación

Las tareas de agrupación pueden variar en términos del tipo de entrada que tienen y el tipo de resultado que se espera que calculen. Para concretar, nos centraremos en la siguiente configuración común:

Entrada

Un conjunto de elementos χ, y una función de distancia sobre él. Es decir, una función $d: \chi \times \chi \rightarrow R_+$ que es simétrica, satisface $d(x,x)=0$ para todo $x \in \chi$ y, a menudo, también satisface la desigualdad del triángulo. Alternativamente, la función podría ser una función de similitud $s: \chi \times \chi \rightarrow [0,1]$ que es simétrica y satisface $s(x, x) = 1$ para todo $x \in \chi$. Además, algunos algoritmos de agrupación también requieren un parámetro de entrada k (que determina el número de agrupaciones requeridas).

Salida

Una partición del dominio establece χ en subconjuntos. Es decir, $C = (C_1, \ldots, C_k)$ donde $\bigcup_{i=1}^{k} C_i = x$ y para todo i ≠ j, $C_1 \cap C_k = \emptyset$. En algunas situaciones, la agrupación es "blanda", es decir, la partición de χ en los diferentes grupos es probabilística donde la salida es una función que asigna a cada punto de dominio, $x \in \chi$, un vector $(p_1(x), \ldots, p_k(x))$, donde $p_1(x) = P[x \in C_i]$ es la probabilidad de que x pertenezca al grupo C_i. Otra salida posible es un dendograma de agrupación (del griego dendron = árbol, gramma = dibujo), que es un árbol jerárquico de subconjuntos de dominio, que tiene los conjuntos únicos en sus hojas y el dominio completo como raíz (Shalev, 2014).

RESULTADOS

Para la aplicación de *machine learning* no supervisado, se toma en cuenta la información de la Encuesta Sociodemográfica

del Departamento de La Paz, realizada en el año 2015. La base de datos considerada tiene datos de migración, salud, educación, empleo, ingresos, agropecuaria, vivienda, etc.

De esta base de datos se considera los 75 municipios y los indicadores educativos, empleo, demográficos y vivienda y dentro de esta se consideraron: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

La tasa de alfabetismo se determina bajo el siguiente criterio:

$$Tasa\ de\ alfabetismo = \frac{Población\ del\ grupo\ edad\ i\ que\ sabe\ leer\ y\ escribir}{Población\ total\ del\ grupo\ edad\ i}\%$$

Se define como la magnitud relativa de la población que sabe leer y escribir. La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio), sexo y grupos de edad. Y cuyo código en la base de datos es ED040.

La tasa de participación fue determinada bajo el siguiente criterio:

$$Tasa\ global\ de\ participación = \frac{Población\ económicamente\ activa}{Población\ en\ edad\ de\ trabajar}\%$$

Se define como el porcentaje de la población en edad de trabajar que forma parte de la población económicamente activa.

La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio), sexo. Y cuyo código en la base de datos es EI0102.

Índice de disponibilidad de dormitorios, es la distribución de hogares según disponibilidad de dormitorios por persona, dado por:

$$Tasa\ de\ disponibilidad\ de\ dormitorios = \frac{Nº\ de\ miembros\ del\ hogar}{Nº\ de\ dormitorios\ existentes}\%$$

La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio). Cuyo código en la base de datos es VV0101.

Índice de masculinidad fue determinado bajo el siguiente criterio:

$$Índice\ de\ masculinidad = \frac{Población\ masculina}{Población\ femenina} \times 100$$

Se define la proporción de hombres frente a las mujeres.

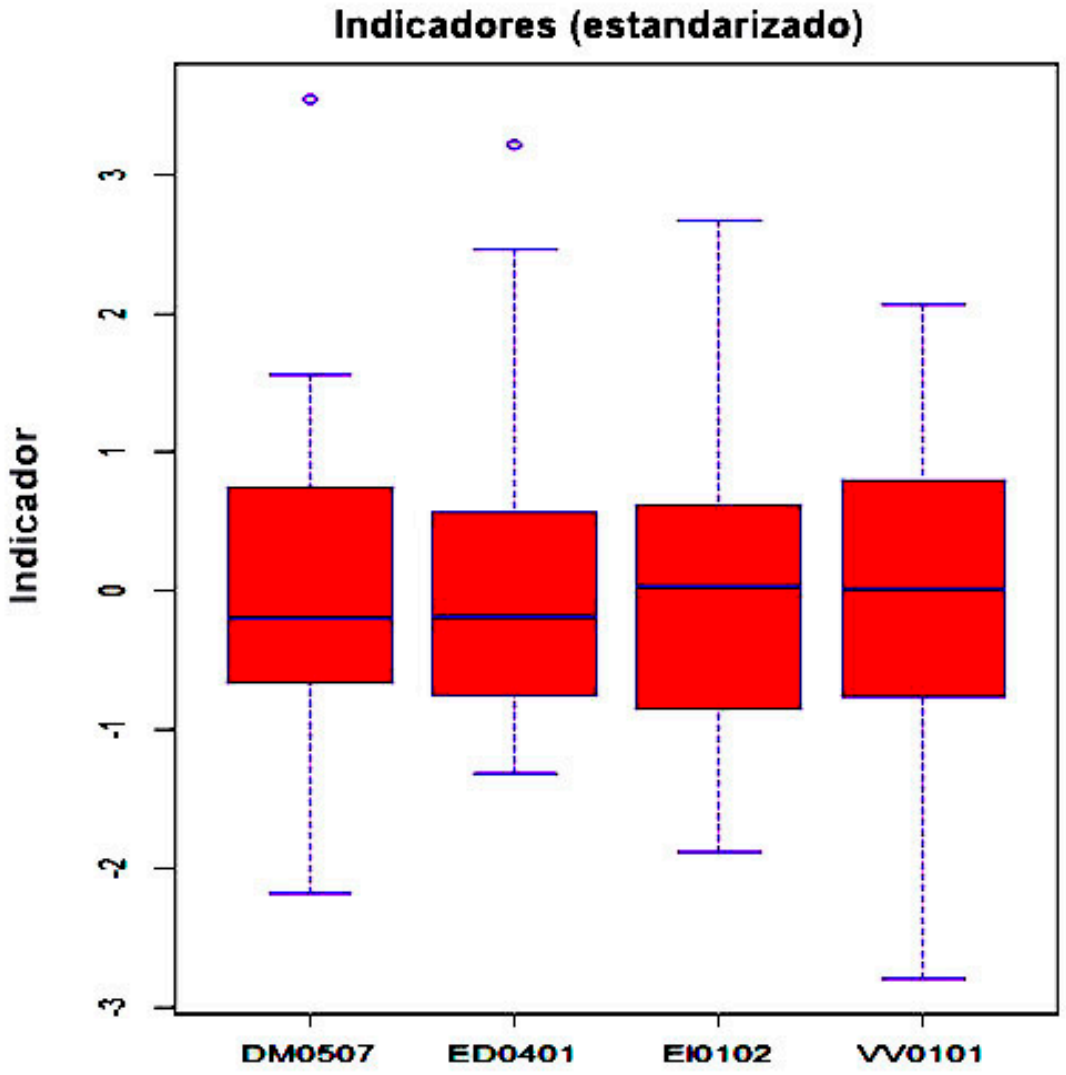
La desagregación considerada en el estudio fue: la región, municipio, área (capital y resto de municipio). Cuyo código en la base de datos es DM0507.

A continuación, se presenta los resultados obtenidos de la investigación que fueron los siguientes:

En la Figura No. 1 se observa los indicadores en forma estandarizada, esta estandarización se realizó con el fin de realizar la aplicación de *machine learning* no supervisada.

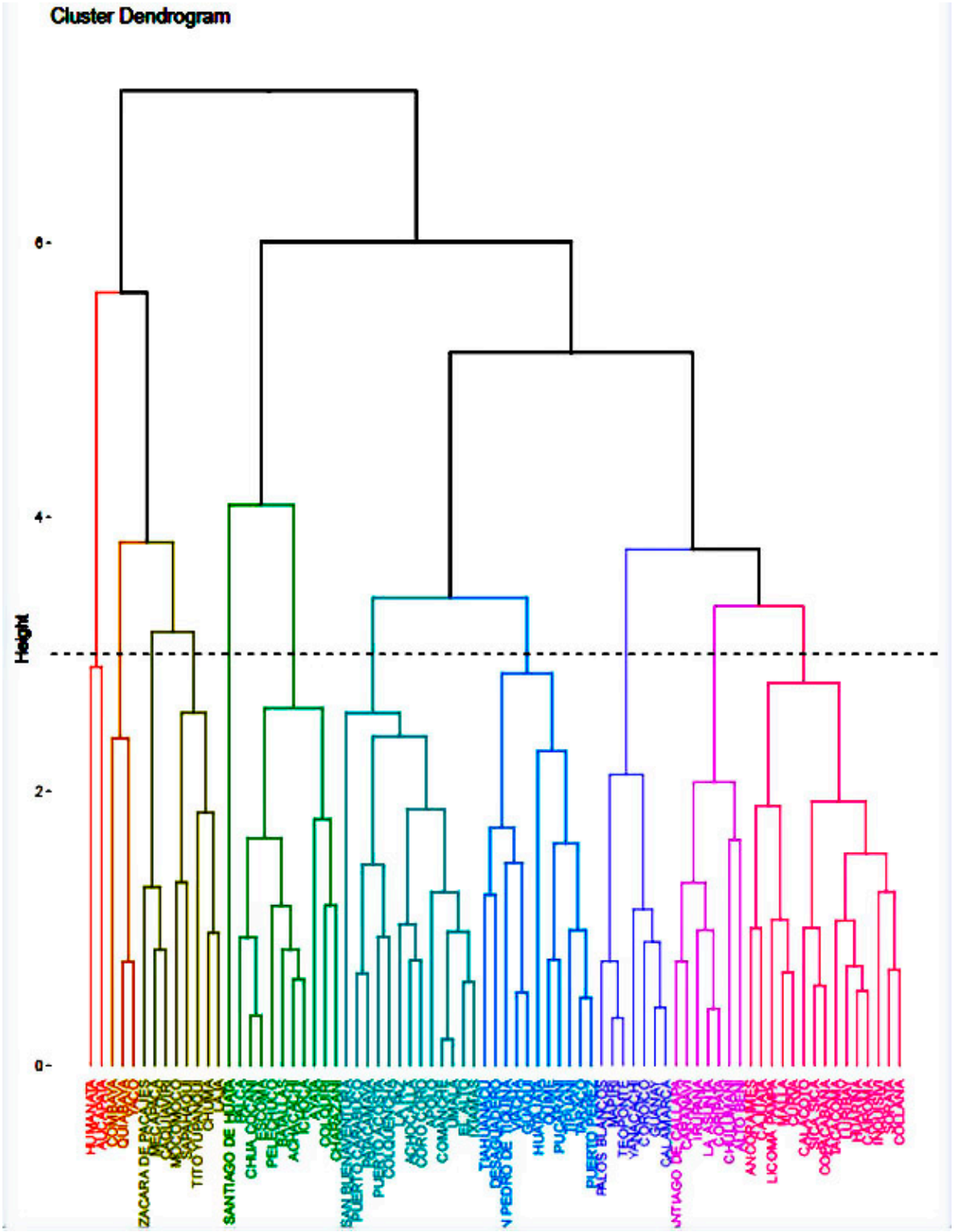
La aplicación de agrupación jerárquica que es una alternativa a los métodos de agrupación de *clusters* de particiones que no requiere que se pre-especifique el número de *clusters*, y en el presente estudio, se muestra en la Figura No. 2.

Figura No. 1
Indicadores de estudio, estandarizado



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Figura No. 2
Agrupaciones K-*medoids clustering*

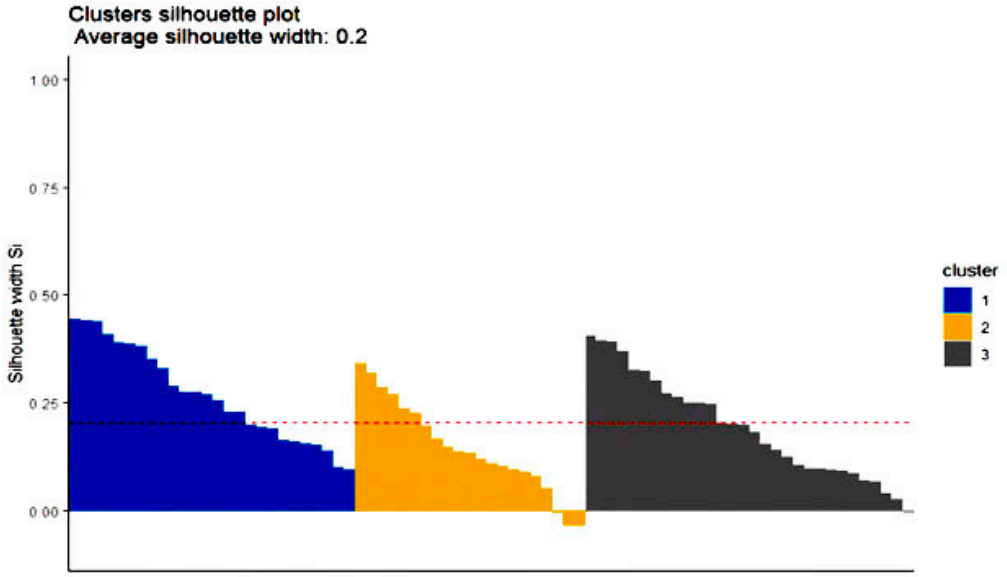


Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Para la validación interna de los *clusters*, se considera la homogeneidad (también llamada *compactness* o cohesión) sea lo mayor posible, a la vez es necesario la separación entre *clusters*. Cuantificar estar dos características es una forma de evaluar como de bueno es el resultado obtenido.

En esta investigación se utilizó el índice de *silhouette widht*. Cuyo resultado se muestra en la Figura No. 3.

Figura No. 3
Cluster silhoutte



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

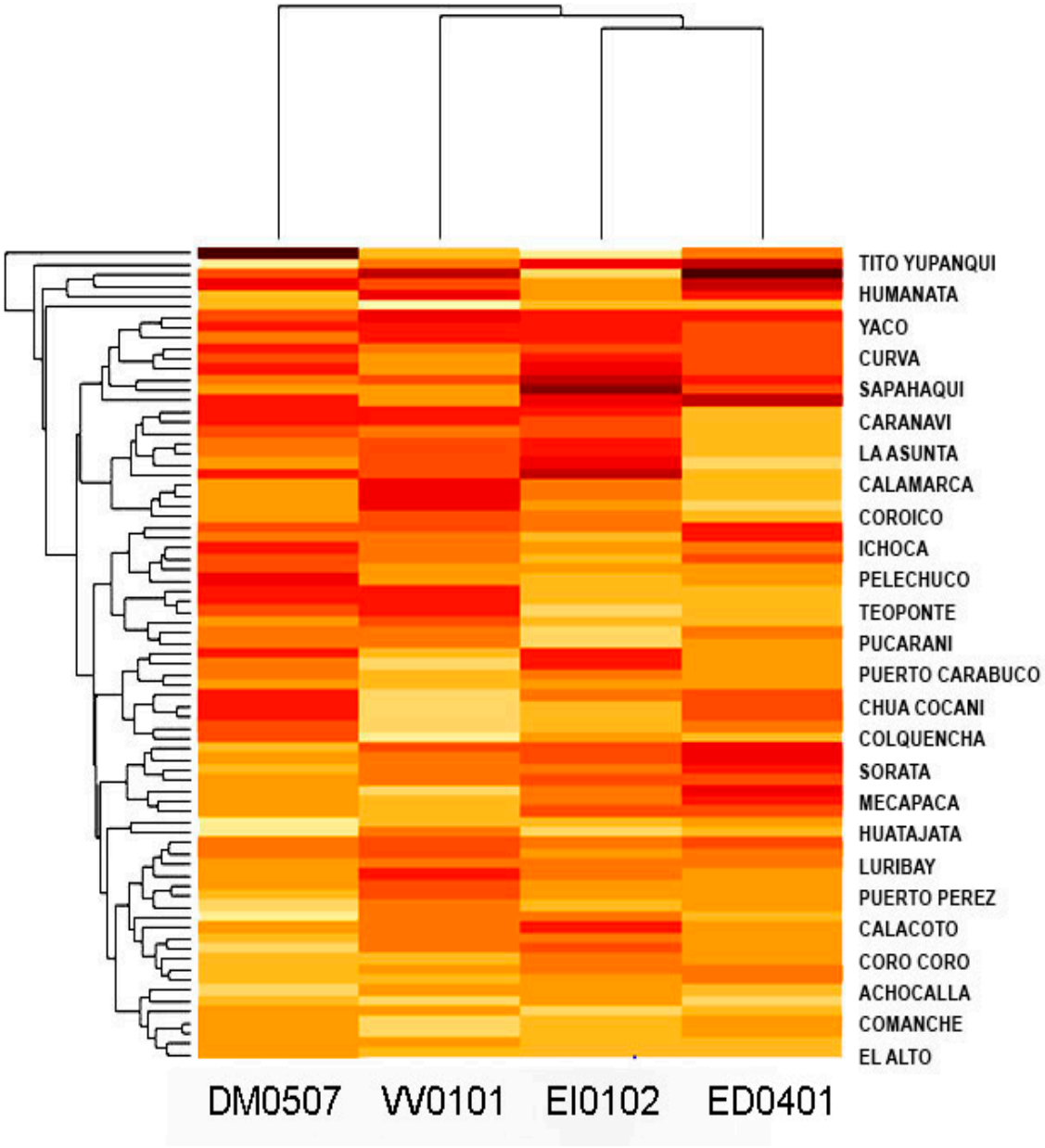
El *cluster* 2 muestra cierta dificultad por tener algunos valores negativos. Lo que implica que esas observaciones podrían tener no clasificadas correctamente.

A continuación, se presenta un *heatmaps* (mapa de calor). Los *heatmaps* son el resultado obtenido al representar una matriz de valores en la que, en lugar de números, se muestra un gradiente de color proporcional al valor de cada variable en cada posición.

La combinación de un dendograma con un *heatmap* permite ordenar por semejanza las filas y o columnas de la matriz, a la vez que se muestra con un código de colores el valor de las variables. Se consigue así representar más información que con un simple dendograma y se facilita la identificación visual de posibles patrones característicos de cada *cluster*.

En este estudio, se muestra un ejemplo de mapa de calor, de los datos de la encuesta sociodemográfica donde se considera los 75 municipios y los indicadores educativos, empleo, demográficos y vivienda y dentro de esta se consideraron: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad. Como se muestra en la Figura No. 4.

Figura No. 4
heatmap



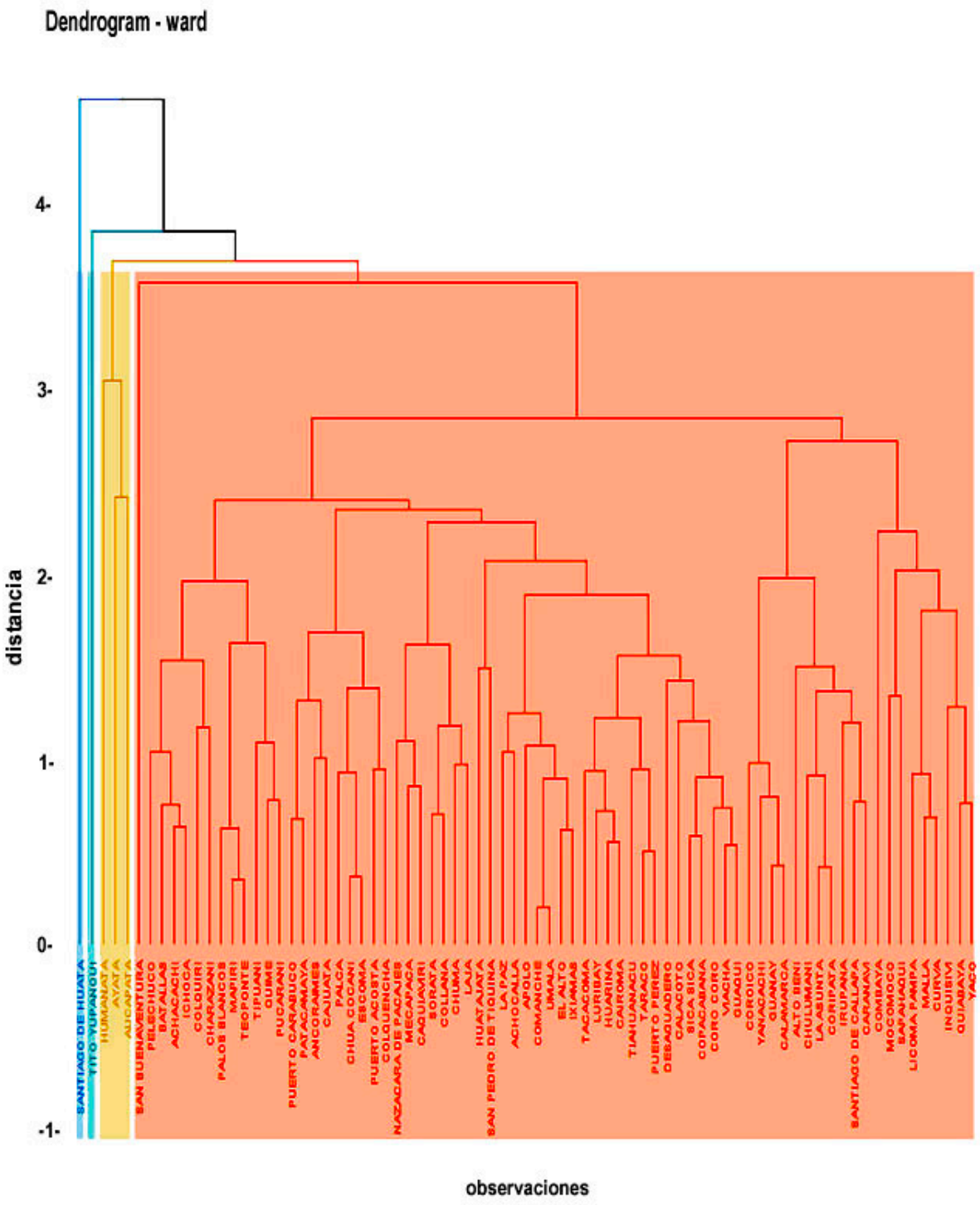
Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Es una forma de presentación a través de heatmap (*stats*) estos datos pueden ser comparables.

Prosiguiendo con la presentación de resultados, consideramos la customización de dendogramas que significa que toda observación pertenece al menos a uno de los k *cluster*. Que refiere a modificar algo de acuerdo a las preferencias personales.

Puede decirse, por lo tanto, que customizar un objeto es lo mismo que personalizarlo. Como se muestra en la Figura No. 5.

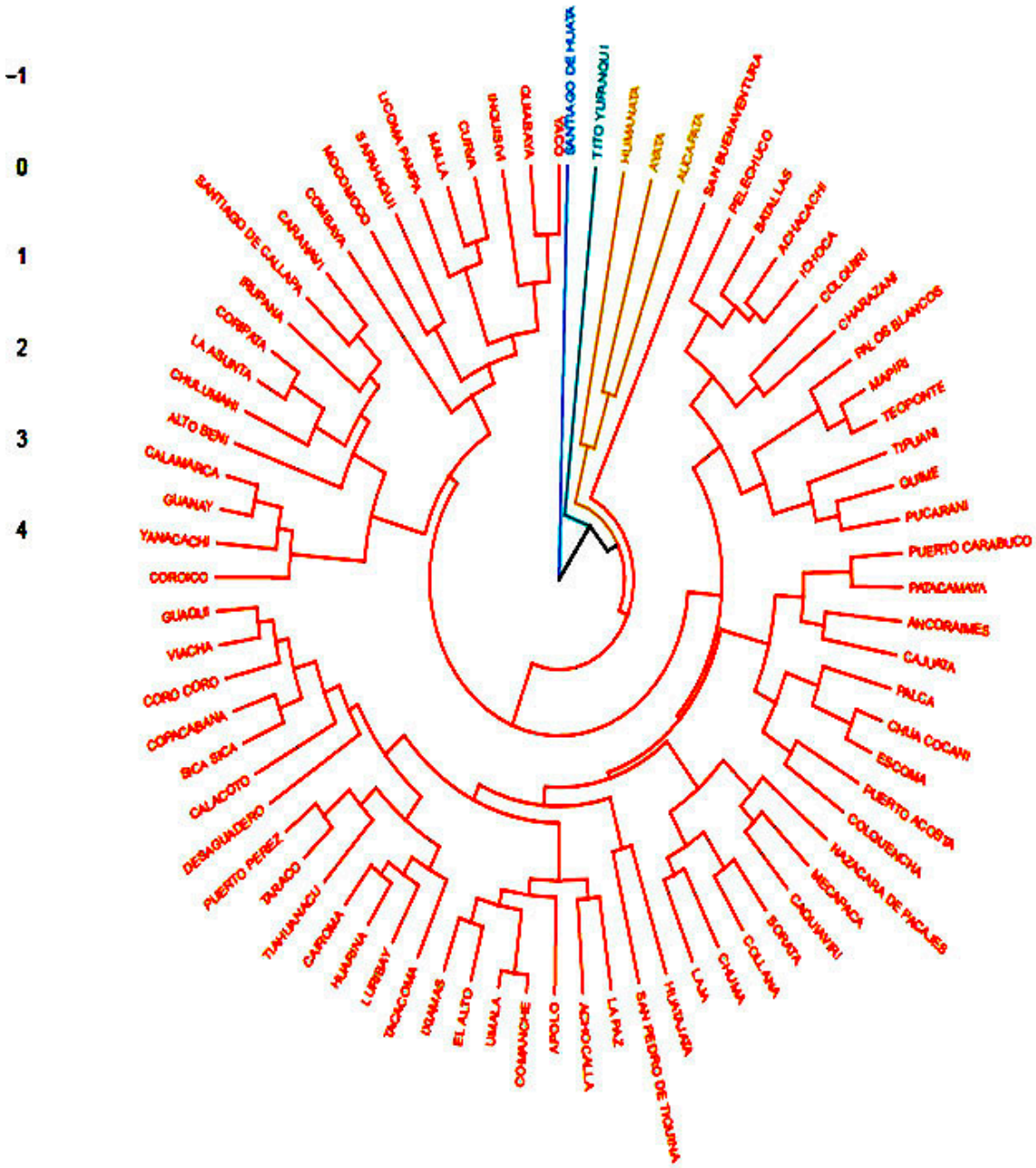
Figura No. 5
Dendograma customizado



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Existen varias formas de presentar, por ejemplo, dendograma circular como se muestra en la Figura No. 6.

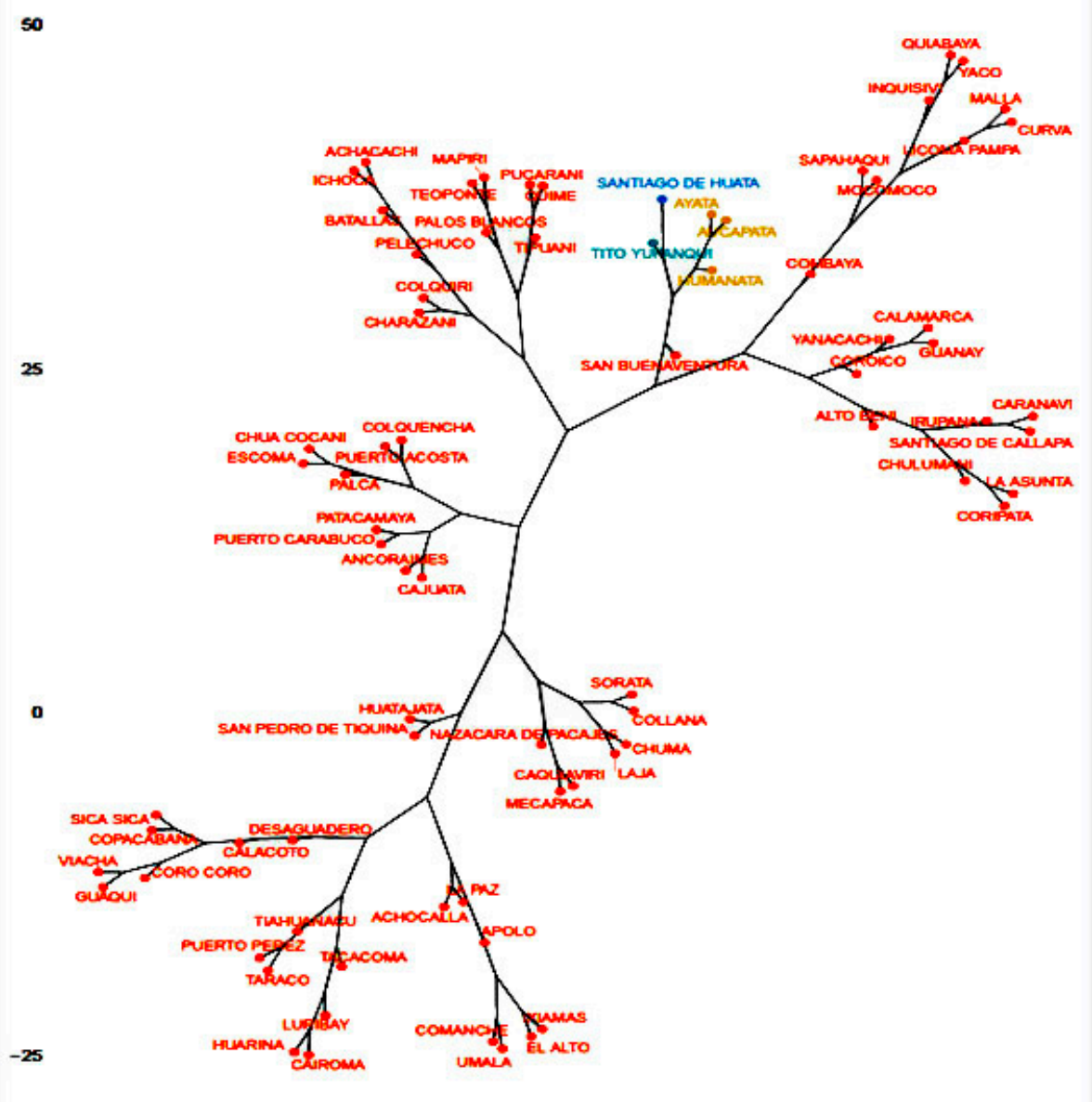
Figura No. 6
Dendograma circular (customizado)



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

También se presenta los resultados de la investigación a través de dendograma en forma de árbol filogenético, como se muestra en la Figura No. 7.

Figura No. 7
Dendograma en forma de árbol filogenético



Fuente: Encuesta Sociodemográfica del Departamento de La Paz-UMSA-IETA, elaboración propia.

Las Figuras No. 5, 6 y 7 son formas distintas de presentar los resultados estudiados de los datos de la Encuesta Sociodemográfica del Departamento de La Paz, realizada en el año 2015, de los indicadores tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

CONCLUSIONES Y DISCUSIÓN

En el presente estudio se llega las siguientes conclusiones:

- La metodología *machine learning* es muy importante porque proporciona información que aporta en la descripción, el análisis y su posterior toma de decisiones.

- El presente estudio tomó en cuenta los datos de la encuesta sociodemográfica realizado el 2015 considerando los 75 municipios del departamento de La Paz y los indicadores educativos, empleo, demográficos y vivienda, dentro de esta se consideraron: la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

- Se debe notar que, la metodología *machine learning* es tipo supervisado y no supervisado, en el presente estudio se implementó estudio de *machine learning* no supervisado.

- Inicialmente tomó en cuenta al análisis *cluster* clásico el cual proporcionó la clasificación de agrupamiento de municipios similares respecto a la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad. Que no necesariamente fue muy claro por tener muchos municipios. La clasificación fue bastante confusa y dificultosa para su análisis.

- Se concluye que la customización es la mejor forma de clasificación por la forma en que se presenta en forma mucho más clara que las otras formas de clasificar.

- Existe la alternativa de clasificación como el dendograma circular, el dendograma en forma de árbol filogenético. Estas alternativas de clasificación son muy buenas porque permiten realizar las interpretaciones más claras y más contundentes.

- La clasificación de los municipios respecto al alfabetismo, tasa de participación, relación de masculinidad y la distribución de hogares según disponibilidad de dormitorios, dio como resultado, lo siguiente:

Los municipios Santiago de Huata, y Tito Yupanqui, muestran similares características respecto a los indicadores: tasas de participación, relación de masculinidad y tasa de alfabetismo.

Otro *cluster* definido por la customización son los municipios Humanata, Alcapata y Ayata y son parecidos en distribución de hogares según disponibilidad de dormitorios por persona, tasa de alfabetismo.

El resto de los *clusters* contiene a todos los restantes municipios cuya clasificación indica que tienen una similitud respecto a la tasa de alfabetismo, tasa de participación, distribución de hogares según disponibilidad de dormitorios y relación de masculinidad.

Recomendaciones

- Se sugiere seguir estudiando *machine learning* no supervisado con otras alternativas para enriquecer este tipo de estudios y metodologías innovadoras.

- Seguir estudiando *machine learning* supervisado las cuales permitirán enriquecer este tipo de estudios y metodologías innovadoras.

- Como esta metodología está involucrada con *big data*, minería de datos, se sugiere seguir profundizando con las herramientas sugeridas.

- Se recomienda el manejo de *Python* para la ampliación de *machine learning*.

REFERENCIAS BIBLIOGRÁFICAS

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis (2° ed.) Springer-Verlag, New York. Cambridge, MA, 1997. Mit Press.

Cristianini, N. and Shawe- Taylor, J. (2000). An Introduction to Support Vector Machines

Harig, A.L. y Fausch, K.D.(2002). Minimum habitat requirements for establishing translocated cutthroat trout populations. Ecol. Appl.12 (2): pp. 535-551.

Murphy, K.P (2012). Introduction to Support Vector Machines

Nagelkerke, N.J. (1991). A note on a general definition of the coefficient of determination. Biometrika 78: pp. 691-692.

Osuna E., Freund, R. and Girosi, F. 2007. "An Improved Training Algorithm for Support.

Platt, J.C. (1997). "Fast Training of Support Vector Machines Using Sequential Minimum.

Río, M. del; Bravo, F.; Pando, V.; Sanz, G. & Sierra, R.(2004). Influence of individual tree and stand attributes in stem straightness in Pinus pinaster Ait. Stands. Ann. Sci. For.61(2): pp. 141-148.

Shalev, S. (2014). Understanding Machine Learning: From Theory to Algorithms.

Smola, A.J. and Schölkopf, B. (2004). "A tutorial on Support Vector Regression," Neuro COLT2.

Vaptnik, V. Golowich, S. and Smola, A. (1996). "Support vector method for function".

Vaptnik, V. N.(1995). The nature of Statical Learning Theory, New York: Wiley.