6. Determinantes individuales de desempeño en las pruebas de Estado para educación media en Colombia

Jaime Orjuela Viracach'a
Universidad Distrital Francisco José de Caldas
ICFES Bogotá, D.C., Colombia

Resumen

En atención a la convocatoria a estudiantes de maestría y doctorado (ICFES, 2010a), con el fin de desarrollar proyectos de investigación sobre temas relacionados con la calidad de la educación, el presente trabajo se propuso hallar los determinantes individuales de desempeño en las pruebas de Estado para educación media en Colombia. Para ello, se realizó un análisis comparativo de dos técnicas de regresión: la primera, modelo de regresión multinivel (HLM)[1] y la segunda una técnica de aprendizaje computacional usando métodos Kernel (SVM)[2] Se evaluaron y confrontaron los resultados de cada técnica y se exploró en busca de herramientas de software disponibles.

Como resultado del presente trabajo, se espera enriquecer el conocimiento actual de los determinantes individuales de desempeño en los estudiantes colombianos de educación media, así como presentar un caso de estudio entre dos técnicas diferentes de regresión aplicados al tema de interés.

Palabras Claves: *aprendizaje estadístico, minería de datos, evaluación educativa.*

1 Técnica estadística usada en estudios anteriores del mismo tema. En varios estudios se refieren a esta técnica como modelo jerárquico lineal (HLM); en el presente trabajo investigativo se utilizarán indistintamente los dos términos.

2 Máquinas de vectores de soporte. En lo sucesivo se referirá a esta técnica como SVM.

6.1 Introducción

Teniendo en cuenta que uno de los objetivos de la convocatoria GPI-001-2010 propuesta en ICFES (2010a) fue promover el uso, análisis y mejoramiento de la información que este instituto genera como producto de las evaluaciones adelantadas a estudiantes de educación básica, media y superior en Colombia y consultando algunos de los métodos utilizados hasta la fecha en el estudio de factores individuales asociados al desempeño académico, se encontró que todos ellos han recurrido a métodos estadísticos como modelos lineales (Barrientos, 2001; Petra y Wolpin, 2006), análisis de componentes principales (Valero, 2005; Gamboa, 2003), estadística grafica descriptiva (Iregui y Melo, 2006; Lara, 2009) y modelos jerárquicos lineales (Murillo, 1999; PISA, 2009), entre otros. Aunque estos métodos se han utilizado exitosamente como modelos explicativos, presentan algunas desventajas como: en su modo básico y por su naturaleza solo buscan relaciones lineales (Bavik, 1998): en el caso del Análisis de componentes principales, el hecho de que cada componente principal sea una combinación lineal de todos los atributos de entrada, suele complicar la interpretación de resultados (Zou, 2004). Esto implica una restricción para tener en cuenta si se requieren evaluar conjuntos de datos desconocidos. Como contraparte, los métodos de aprendizaje computacional pueden, tratar problemas altamente no lineales, aunque suelen ser más costosos computacionalmente (Huang y Kecman, 2006). No obstante las debilidades y fortalezas de una y otra técnica, ninguna resulta infalible en todas las áreas del conocimiento, y una forma idónea para determinar la más adecuada es realizando un análisis comparativo de técnicas que resuelven la misma tarea desde enfoques diferentes.

Mediante una revisión de los datos disponibles del año 2000 al 2009 para la prueba SABER 11 (cerca de 450.000 registros anuales), descritos por 81 atributos que en su mayoría son numéricos, y teniendo en cuenta las temáticas de investigación propuestas en los términos de referencia de la convocatoria ICFES GPI-001-2010, se concluyó que para permitir la tratabilidad de los datos y los resultados, se requería limitar el conjunto de datos, eligiendo el más reciente y menos numeroso: el primer semestre del año 2009. Por otro lado, se eligió como técnica de análisis estadístico los modelos jerárquicos lineales (HLM) y como técnica de aprendizaje computacional las máquinas de vectores de soporte (SVM).

Durante los últimos años, los estudios cuantitativos en educación han utilizado frecuentemente los modelos estadísticos de análisis multinivel (Goldstein, 1999), también llamados modelos jerárquicos lineales (Bryk y Raudenbush, 1992), debido a que toman en cuenta los efectos potenciales que surgen de la forma como los estudiantes se asignan a los colegios. En algunos países, por ejemplo, el estatus socioeconómico de un estudiante determina en gran parte el tipo del colegio al cual él o ella asistirán y existe una estrecha relación entre el estatus socioeconómico de éste y los demás estudiantes dentro del colegio. En contraste, otros países o sistemas tienen colegios en los que su estudiantado se circunscribe en una amplia variedad de estatus socio-económicos, pero, dentro del colegio, hay diferenciación de los

cursos en que se asignan, lo cual afecta la varianza dentro del colegio. En ese sentido, por tanto, habrá un impacto en la variable de salida[3] que depende del estudiante, de los colegios, de la varianza dentro de cada colegio y de la varianza entre colegios.

Aparte de los modelos estadísticos, el aprendizaje computacional a partir de los datos experimentales se perfila como una alternativa viable a los modelos formales para el análisis y extracción de conocimiento. Este aprendizaje se relaciona con una nueva área de las ciencias de la computación llamada *soft computing.* No existe una definición[4] estricta del soft computing, pues engloba diversas técnicas computacionales para solucionar problemas con información incompleta, con cierto grado de incertidumbre y/o de naturaleza estocástica que tienen en común que los métodos de solución imitan el aprendizaje humano. Algunas de las técnicas más usadas en esta área son: las redes neuronales *(NN)*[5], las máquinas de vectores de soporte *(SVM)*[6] y los sistemas de lógica difusa *(FL)*[7]. Para su implementación, las redes neuronales y las máquinas de vectores de soporte se apoyan en estructuras matemáticas para el aprendizaje, mientras que los sistemas de lógica difusa se basan en estructuras humanas de clasificación. Otras técnicas como algoritmos genéticos evolutivos, razonamiento probabilístico, teorías fractales y del caos también pueden incluirse dentro del área del soft computing.

En ese orden de ideas, hay dos objetivos principales que pretende lograr esta investigación: (1) enriquecer el conocimiento actual de los atributos individuales más relevantes que relacionan al estudiantado con su desempeño en los resultados de las pruebas de Estado de educación media; (2) enriquecer las técnicas utilizadas para la determinación de esos atributos, aportando y/o reforzando los resultados de estudios ya realizados, sí como proporcionar nuevas herramientas metodológicas para su análisis.

La investigación se estructura así: una presentación detallada de los datos y de las exclusiones, consideraciones y justificación de los datos tratados. Luego se exponen respectivamente los dos modelos utilizados, HLM y SVM, realizando en cada caso un análisis de resultados. Finalmente se muestra el análisis comparativo de las dos técnicas así como las conclusiones y el trabajo futuro propuesto.

3 Puntaje obtenido en las áreas de Matemáticas, Ciencias y Lenguaje.
4 Ni tampoco una traducción estricta. En este estudio se utilizarán indistintamente los términos soft computing y aprendizaje computacional.
5 Neural Networks.
6 Support Vector Machines.
7 Fuzzy Logic.

6.2 Presentación de los datos

Los datos utilizados en la presente investigación fueron provistos por el ICFES y corresponden a los resultados de la aplicación de la prueba de Estado para educación media (SABER11) del primer semestre del 2009 (69.740 registros); aunque existen resultados en trece áreas[8], se analizan dos de las que en el criterio de los expertos (Figel, 2009) evalúan competencias básicas claves para el aprendizaje: Matemáticas y Lenguaje.

Hay 35 variables predictoras relacionadas con cuatro (4) grupos de características[9]: dos (2) con las características del colegio, 21 relacionadas con las características socioeconómicas, tres (3) relacionadas con las características del estudiante y cinco (5) con las características del grupo familiar. Dos variables adicionales se tuvieron en cuenta: el índice de nivel socioeconómico del estudiante (INSE) y la clasificación socioeconómica del colegio (CSE) provistas también por el ICFES (2010b).

Tomando en cuenta que la prueba de Estado es de carácter obligatorio para todos las personas que terminan su ciclo de formación media, se distinguen varios grupos bien diferenciados: el primero corresponde al tipo de jornada: por un lado, quienes asisten a jornadas mañana, tarde o completa, que son jóvenes de no más de 20 años, generalmente dependientes de sus padres y, por otro lado, quienes asisten a jornadas nocturnas, sabatinas y/o dominicales que asisten a programas de validación del bachillerato que por ley deben ser mayores de edad al momento de presentar la prueba de validación (MEN, 2009) y por tanto en condiciones sociales, familiares y culturales diferentes a las de jornadas diurnas ordinarias. El segundo criterio diferenciador distingue el sector del colegio, es decir, si es público o privado en Iregui y Melo (2006) se hace una evaluación y análisis de la eficiencia de la educación en Colombia comparando colegios públicos y privados, y se llega a la conclusión de que el sector del colegio es un factor diferenciador. No obstante las diferencias expuestas, se realizó el análisis tanto para el conjunto de datos completo como para los grupos identificados, con el fin de corroborar las hipótesis propuestas.

Con base en el análisis de los datos se puede evidenciar que gran parte (85%) de los estudiantes de educación media asisten en la edad esperada (siendo menores de edad). También se aprecia que la oferta educativa pública a nivel nacional es comparativamente mayor que la privada, situación que se evidencia ampliamente en la zona rural; no obstante, en la mayoría de ciudades capitales la cifra puede ser inversa.

8 Biología, Ciencias Sociales, Filosofía, Física, Geografía, Historia, Lenguaje, Matemáticas, Química, Profundización, Idioma Extranjero y Área Interdisciplinar (que puede ser una de las siguientes: Medio Ambiente, Violencia y Sociedad o Medios de Comunicación y Cultura).

9 En el presente trabajo se utilizan indistintamente los términos característica, atributo y variable.

En general y como era de esperarse, los colegios públicos atienden a estudiantes de menores recursos; sin embargo, los colegios privados de jornadas diferentes a las diurnas ordinarias no se alejan demasiado de la media. En términos de tamaño existen más colegios pequeños que grandes; llama la atención la gran cantidad de estudiantes en algunas instituciones de validación del bachillerato en razón de su auspicio para las pruebas de Estado.

En promedio, los estudiantes que asisten a colegios privados en jornada diurna pagan ocho (8) veces más que quienes asisten a colegios públicos, y cuatro (4) veces más en el caso de jornadas nocturna, sabatina y dominical. En términos generales, la distribución de género es homogénea con sutiles diferencias en la jornada diurna. Por otro lado, la mayoría de estudiantes que asisten a colegios públicos se circunscriben en colegios de categoría media, baja e inferior.

6.3 Análisis estadístico multinivel

En los últimos años, los modelos multinivel se han utilizado ampliamente en el análisis de datos en la educación, principalmente porque los modelos clásicos de regresión no tienen en cuenta los efectos potenciales que surgen de la forma como los estudiantes están distribuidos en los colegios. Dos tipos de índices son relevantes en el análisis multinivel: (1) los coeficientes de regresión, denotados como parámetros fijos del modelo; (2) los residuos mostrados por las varianzas estimadas, denotadas como parámetros aleatorios del modelo.

6.3.1 Método

El primer paso recomendado en el análisis de regresión multinivel es la descomposición de la varianza de la variable de respuesta en diferentes niveles, primero sin considerar covariables[10], luego con variables a nivel del colegio y por último con variables tanto a nivel de colegio como de estudiante. En el caso de estudio, la varianza del estudiante será descompuesta en dos componentes: la varianza dentro de cada colegio y la varianza entre colegios. Para la determinación del modelo jerárquico se propone el siguiente modelo:

$$y = X\beta + Zu + \epsilon \qquad \textbf{(1)}$$

Donde:

y es un vector de respuesta $n \times 1$ para cada prueba de interés (Matemáticas o Lenguaje).
X es una matriz $n \times p$ que contiene los regresores de efecto fijo.
β es un vector $p \times 1$ de parámetros de efectos fijos.
Z es una matriz $n \times q$ de regresores de efecto aleatorio.
u es un vector $q \times 1$ de efectos aleatorios.
ϵ es un vector $n \times 1$ de errores.

10 Se utilizan indistintamente el término covariable, variable independiente y/o variable predictoria.

En términos más textuales, se podría afirmar que el resultado obtenido en la prueba de Matemáticas o Lenguaje para cada estudiante que asiste a la escuela es igual al promedio general de los colegios de la prueba deseada más dos efectos aleatorios (varianzas), uno debido a la relación entre colegios otro debido al desempeñó dentro de cada colegio.

6.3.2 Presentación de resultados por el método jerárquico lineal

Del modelo propuesto en (1) se obtuvieron los siguientes efectos fijos y aleatorios para el modelo sin covariables:

Efecto fijo (intersección): 45, 01

Varianza entre colegios: 57, 74

Varianza dentro de los colegios: 89, 32

El estadístico que representa el grado de variabilidad existente entre colegios en comparación con la variabilidad dentro de cada colegio, se denomina coeficiente de correlación intraclase y se define como:

$$\rho = \frac{varianza\ entre\ colegios}{varianza\ total} = 0,39 \qquad \textbf{(2)}$$

Donde un valor de ρ cercano a 0 indica que **no** hay heterogeneidad entre colegios; valores superiores y hasta un máximo hipotético de 1 indican que existe heterogeneidad entre colegios. En este caso el valor obtenido en (2) confirma la necesidad de utilizar un modelo multinivel para el análisis de los datos del ICFES.

Con base en los resultados obtenidos, se concluye que se esperan mejores resultados en las pruebas de Lenguaje que en las de Matemáticas. También se espera que en promedio haya mejores resultados en los colegios públicos que en los privados, situación justificada teniendo en cuenta la mayor disparidad entre colegios privados comparados con sus homólogos públicos que poseen un mayor control estatal. En los colegios públicos, el factor no académico que más influye en los resultados de las pruebas de Estado es el índice de nivel socioeconómico del estudiante, mientras que en los colegios privados pesa más el nivel socioeconómico del colegio.

Con referencia a las diferencias entre géneros, se puede afirmar que los hombres tienden a obtener mejores puntajes que las mujeres principalmente en Matemáticas. Al respecto Gaviria (2001) expone cómo la brecha entre hombres y mujeres en las pruebas de aptitud

escolástica han sido fuente de polémica en Estados Unidos, pues aunque los hombres tienden a obtener mejores puntajes en las pruebas para educación media, las mujeres presentan mejores puntajes en la universidad. Por tanto, no se puede calificar al género como factor determinante de desempeño.

6.4 Aplicación de Regresión mediante maquinas de vectores de soporte

Las máquinas de vectores de soporte se han utilizado ampliamente en la resolución de problemas de clasificación; no obstante, pueden aplicarse también en tareas de regresión (aproximación funcional) (Smola y Vapnik, 1997). Para ello se tienen l datos de entrenamiento a partir de los cuales se pretende aprender la relación entrada-salida $f(x)$.

Para lo cual se define un conjunto de datos de entrenamiento $D=\{[x(i),y(i)]\epsilon R^n \times R, i=1,...,l\}$ que está formado por l pares $(x1,y1),(x2,y2),...,(xl,yl)$, donde las entradas x son vectores n-dimensionales $x \epsilon R^n$ y la respuesta al sistema son los valores continuos $y \epsilon R$ que en este caso corresponden al resultado de la prueba de Matemáticas o Lenguaje. Dados los parámetros $C > 0$ y $\varrho > 0$, la forma estándar de los vectores de soporte es:

$$min_{\omega,b,\varepsilon,\varepsilon*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\varepsilon_i + C\sum_{I=1}^{L}\varepsilon_i* \qquad \textbf{(3)}$$

$$sujeto\ a:\ \omega^T\phi(x_i)+b - y_i \leq \epsilon + \varepsilon_i,$$

$$y-\omega^T\phi(x_i) - b \leq \epsilon + \varepsilon_i*,$$

$$\varepsilon_i,\varepsilon_i* \geq 0, i = 1,...,l$$

Donde:

$\phi(x_i)$ mapea x_i en un espacio dimensional mayor.
$C > 0$ es el parámetro de regularización.

El problema dual es entonces:

$$min_{\alpha,\alpha*} \frac{1}{2}(\alpha - \alpha*)^T\ Q(\alpha - \alpha*) + \epsilon\sum_{i=1}^{l}(\alpha + \alpha*) + \sum_{i=1}^{l} y_i(\alpha - \alpha*) \qquad \textbf{(4)}$$

$$sujeto\ a:\ e^T(\alpha - \alpha*) = 0$$

$$0 \leq \alpha_i, \alpha_i* \leq C, i = 1,...,l$$

Donde $Q_{ij} = K(x_i, X_j) \equiv \phi(x_i)^T \phi(x_j)$. Después de resolver el problema de la ecuación (4), la función de aproximación es:

$$\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) \; K(x_i, x) + b \qquad \textbf{(5)}$$

El paquete LIBSVM entrega como parámetro de salida del modelo $(\alpha^* - \alpha)$. Dos medidas de desempeño utilizadas para evaluar la regresión son el error cuadrático medio (MSE^{11}) y el coeficiente de correlación r_2.

$$MSE = \frac{1}{l} \sum_{i=1}^{l} (f(x_i) - y_i)^2 \qquad \textbf{(7)}$$

$$r^2 = \frac{(l \sum_{i=1}^{l} f(x_i) y_i - \sum_{i=1}^{l} f(x_i) \sum_{i=1}^{l} y_i)^2}{(l \sum_{i=1}^{l} f(x_i)^2 - (l \sum_{i=1}^{l} f(x_i))^2) \; (l \sum_{i=1}^{l} y_i - (l \sum_{i=1}^{l} y_i)^2)} \qquad \textbf{(8)}$$

6.4.1 Método

Una máquina de vector de soporte se puede ver como una caja negra que aprende a partir de los datos de aprendizaje, con base en un kernel determinado. En la fase de entrenamiento observa cada entrada xi respecto a su correspondiente salida yi y estima los parámetros w (pesos) y de esta forma mapea el comportamiento del sistema $y = f(x,w)$. Se aclara que la máquina no solo trata de interpolar las parejas entrada/salida, sino también busca una función de aproximación que generalice adecuadamente el comportamiento del sistema. Después del entrenamiento, en la fase de prueba, la salida de la máquina $o = fa(x,w)$ se espera que sea un buen estimador de la verdadera respuesta del sistema y (Huang y Kecman, 2006).

En la implementación de la regresión por máquinas de soporte se probaron varios paquetes de software: el propuesto por Kecman, (2001), el propuesto por Parella (2007) y el que finalmente se eligió para trabajar con los datos del ICFES, que propusieron Chang y Lin (2011), llamado *Libsvm*, por ser el más reciente y robusto en cuanto a cantidad de datos de entrada. Las tres herramientas se encuentran en internet y son de libre uso y distribución. La dirección web del paquete Libsvm es http://www.csie.ntu.edu.tw/cjlin/libsvm/. Libsvm es actualmente uno de los software para máquinas de vectores de soporte más utilizados.

11 Por sus siglas en inglés Mean Square Error.

6.4.2 Presentación e interpretación de resultados mediante máquinas de vectores de soporte

Se trató sin éxito de descubrir las relaciones individuales generadas en el modelo, por lo cual se mostrará la salida de la máquina de aprendizaje en términos de la bondad de ajuste del modelo. Inicialmente se hizo un muestreo aleatorio del 50% del conjunto de datos para el entrenamiento y 50 % para la prueba, considerando los mismos conjuntos de datos utilizados en la implementación de HLM. Después se realizaron dos iteraciones más cambiando la relación entre conjunto de entrenamiento y prueba así: 20%/80% y 80 %/20%.

Con base en (**9**) se calculó r^2, y se encontró que existe una mayor y directa correlación entre el nivel socioeconómico del colegio y de los estudiantes en colegios privados que en públicos de las jornadas diurnas ordinarias. Caso contrario ocurre en las jornadas nocturnas, donde la diferencia entre públicos y privados no es muy marcada y además no está tan correlacionada con el desempeño escolar.

También puede concluirse que hay una mayor relación del estatus socioeconómico con el desempeño escolar para las pruebas de Matemáticas en comparación con las pruebas de Lenguaje. Evaluando el error cuadrático medio se concluye un comportamiento similar en las pruebas de Lenguaje que en las pruebas de Matemáticas.

En cuanto al desempeño del algoritmo, se puede aprecia un comportamiento relativamente parejo frente a la cantidad de casos de aprendizaje y prueba, es decir, aún cambiando drásticamente la cantidad de casos para aprendizaje y prueba, las máquinas de vectores de soporte muestran estabilidad frente al error cuadrático medio y al coeficiente de correlación.

6. 5 Discusión general y conclusiones

Se calcularon los efectos fijos y aleatorios mediante la regresión multinivel a los conjuntos de aprendizaje, para luego aplicar esos efectos a los datos de prueba y hallar el error cuadrático medio y el coeficiente de correlación utilizando las ecuaciones (**8**) y (**9**).

El equipo de cómputo utilizado para ejecutar tanto los algoritmos estadísticos como los computacionales fue un servidor HP-Proliant ML110-G6 con un procesador intel Xeon Quad-Core, 4 GB de memoria RAM y 500GB de almacenamiento en disco duro, con sistema operativo Windows Server Enterprise 2008 Service Pack 1. En el momento de ejecutar SPSS o Libsvm, no se ejecutó ningún otro programa.

Se pudo apreciar un comportamiento muy cercano en los dos algoritmos (SVR y HLM), bueno en ambos casos, con mayor estabilidad al cambio en los conjuntos de datos para

SVR, mientras que en términos de desempeño computacional el HLM fue mejor. En cuanto a resultados, se nota la gran expresividad de los modelos estadísticos versus la gran complejidad de establecer la forma como en SVR se obtiene la relevancia de los atributos. Teniendo en cuenta los objetivos de la presente investigación, se concluye que existe una relación directa entre algunas características no académicas del estudiante y sus competencias medidas en la prueba de Estado al finalizar su formación media; y que esas características (sociales y económicas) difieren notoriamente entre colegios públicos y privados, y entre jornadas diurnas ordinarias y otras jornadas.

Tanto en el modelo estadístico como en el computacional se esperan mejores resultados en las pruebas de Lenguaje que en las de Matemáticas. También se espera que en promedio haya mejores resultados en los colegios públicos que en los privados, situación que se justifica teniendo en cuenta la mayor disparidad entre colegios privados comparados con sus homólogos públicos que poseen un mayor control estatal. En los colegios públicos, el factor no académico que más influye en los resultados de las pruebas de Estado es el índice de nivel socioeconómico del estudiante, mientras que en los colegios privados pesa más el nivel socioeconómico del colegio.
Tomando en cuenta que debido a la exhaustividad del trabajo sólo se incluyó el conjunto de datos correspondiente a un semestre del 2009, como trabajo futuro se podría hacer extensivo a los demás años (2000-2011).

El segundo gran objetivo propuesto al inicio de la investigación era comparar un método estadístico (HLM) con uno computacional (SVR). La conclusión de esta comparación fue una notable superioridad del método estadístico frente al computacional, primero por su expresividad, pues la interpretación de los modelos es más sencilla y en segunda instancia por su desempeño computacional, pues en algunos casos superó en diez (10) veces el tiempo de procesamiento que requirió el método de máquinas de vectores de soporte. En contraposición, el método computacional demostró ser más robusto a cambios drásticos en el tamaño del conjunto de datos, y es posible hacer buenas aproximaciones incluso en el aprendizaje con el 20 % del conjunto de datos.

No obstante los resultados obtenidos, se deben tener en cuenta algunas consideraciones de comparabilidad de los modelos: una máquina de vector de soporte tiene más relación funcional con una regresión lineal que con una regresión multinivel, dado que tanto la regresión lineal como la regresión por máquinas de vectores de soporte poseen sólo una componente de error; los modelos jerárquicos descomponen el efecto aleatorio del modelo en varios niveles, permitiendo una estructura de varianza/covarianza compleja. Por otro lado, los modelos de regresión se utilizan como herramientas descriptivas y predictivas, sin embargo, algunos modelos se orientan a una u otra tarea, como en los modelos jerárquicos que en su mayoría se utilizan como herramienta descriptiva, labor que no realizan las máquinas de vectores de soporte que están totalmente focalizadas a la tarea predictiva. En ese sentido, la comparación realizada en este trabajo puede resultar injusta para uno u otro modelo.

En vista de que en el análisis de los datos se fraccionaron los datos en seis (6) grupos diferenciados por jornada y tipo de colegio (públicos y privados), se propone como trabajo futuro elaborar la regresión multinivel incluyendo todos los datos con variables binarias discriminatorias para esos grupos, además de las interacciones entre esas variables y varias componentes de error para cada nivel.

El método estadístico multinivel, además de ser un buen método de regresión, también permite hacer disertaciones acerca de la relevancia de cada atributo así como de la forma como se agrupan jerárquicamente los datos. El método computacional, por su parte, resulta un buen método de regresión en el que no se requiere demasiado conocimiento de los datos, se puede tratar con conjunto de datos pequeños y con alto nivel de ruido. Como trabajo futuro se propone buscar otros métodos computacionales que superen los inconvenientes encontrados en este trabajo frente al método estadístico, principalmente el que se refiere a la agrupación recursiva de datos en varios niveles.

Bibliografía

- **Barrientos, Gaviria** (2001, noviembre). *Determinantes de la calidad de la educación en Colombia.* Archivos de Economía , 1 (159), 88.

- **Bavik, R. B.** (1998). *Multiscale pca with application to multivariate statistical process monitoring.* Department of Chemical Engineering, 1 , 10.

- **Bryk, A., y Raudenbush, S.** (1992). *Hierarchical linear models in social and behavioral research: Applications and data analysis methods.* Newbury Park, CA.: Sage Publications.

- **Chang, C.C., y Lin, C.J.** (2011). *LIBSVM: A library for support vector machines.* ACM Transactions on Intelligent Systems and Technology , 2 , 27:1–27:27. (Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm)

- **Figel, J.** (2009, septiembre-octubre). *Competencias clave para el aprendizaje permanente.* Al Tablero, 1 (52), 9-11.

- **Gamboa, L.** (2003, diciembre). *La teoría del valor agregado: una aproximación a la calidad de la educación en Colombia.* Revista de Economía de la Universidad del Rosario ,1 ,95.

- **Gaviria, A.** (2001). *Determinantes de la calidad de la educación en Colombia.* Archivos de Economía , 1 , 19.

- **Goldstein, H.** (1999). *Multilevel statistical models.* Bristol: Bristol University.

- **Huang, I. K., y Kecman, V.** (2006). *Kernel based algortihms for mining huge data sets.* Vol. 17. Polish Academy of Sciences. Poland: Springer. ICFES. (2010a). Convocatoria a estudiantes de maestría y doctorado (Términos de Referencia). ICFES.

- **ICFES.** (2010b, junio). *Metodología de construcción del índice de nivel socioeconómico de los estudiantes (INSE) y de la clasificación socioeconómica (CSE) de los colegios.* www.icfes.gov.co , 1 , 13.

- **Iregui L.A., y Melo, J. R.** (2006, febrero). *Evaluación y análisis de eficiencia de la educación en Colombia.* Revista Banco de la República , 1 , 108.

- **Kecman, V.** (2001). *Learning and soft computing: Support vector machines, neural networks and fuzzy logic models. Massachusetts:* MIT Press.

- **Lara, M.** (2009). *Asociación entre la efectividad de la funcionalidad familiar en las familias de los estudiantes de la Facultad de Enfermería de la Fundación Universitaria Sanitas y el rendimiento académico.* Tesis de Master no publicada, Universidad Nacional de Colombia.

- **Ministerio de Educación Nacional** (2009, febrero). *Decreto 299 del 2009: Por el cual se reglamentan algunos aspectos relacionados con la validación del bachillerato en un solo examen. Bogotá:* MEN.

- **Murillo, J.** (1999). *Los modelos jerárquicos lineales aplicados a la investigación sobre eficacia escolar.* Revista de Investigación Educativa , 2 (17), 6.

- **Parella, F.** (2007). *Online support vector regression.* Department of Information Science, 1, 50.

- **Petra, T., y Wolpin, K.** (2006, November). *The production of cognitive achievement in children: Home, school and racial test score gaps.* University of Pennsylvania , 1 , 70.

- **PISA.** (2009). *PISA, Data Analysis Manual SPSS® (Second ed.).* Secretary-General of OECD: Organisation for economic co-operation and development.

- **Smola, Vapnik** (1997). *Support vector regression machine.* Advances in Neuronal Information Processing Systems , 1 , 10.

- **Valero, Q. W.** (2005). *Propuesta para la elaboración de un índice de calidad de las instituciones educativas privadas en Bogotá.* Tesis de Master no publicada, Universidad Nacional de Colombia.

- **Zou, H. T.**(2004). *Sparce principal component analysis.* Standford University , 1 , 1