¿CAMBIARÁ EL PARADIGMA?

La Minería de Datos en la Ciencia Económica.

ANDRES FLOREZ LLANO
TRABAJO DE GRADO-UNIVERSIDAD EAFIT

Tabla de Contenido

1. INTRODUCCIÓN

La era de los datos está cambiando la forma en la que entendemos los procesos económicos y sociales. La Innovación Dirigida por Datos propone un nuevo reto para los Economistas y Diseñadores de Política Pública, pues es nuestro deber entender como la *datafication*[1] está transformando nuestros modelos mentales y posiblemente nuestros paradigmas. En una apuesta por cambiar los marcos normativos y metodológicos, se espera que la era del Big Data empodere a los individuos y a las instituciones en la toma de decisiones más informadas y de manera más oportuna (OECD, 2015).

Los últimos 20 años han venido acompañados de un progresivo crecimiento en el flujo de datos; por el lado de la oferta, el perfeccionamiento de las tecnologías de captación y la democratización de la tecnología, generaron en 2015, 2.5 billones de gigabytes por día según informe de la (OECD, 2015); mientras que por el lado de la demanda (OECD, 2015) se estima que el mercado se encuentra alrededor de los 17 mil millones de dólares, con un crecimiento promedio del 40% anual. Las magnitudes del mercado conjunto a la incertidumbre de los verdaderos retornos, han creado técnicas que proponen un nuevo foco a través del cual entender la economía.

Según la (OECD, 2015)[2], la ciencia económica se enfrenta a un nuevo reto. Y junto a ello, se alza la duda de como sabremos adoptar las nuevas propuestas de un mundo Dirigido por la Innovación de los Datos; y entender si las posturas tan disruptivas en las técnicas modernas contrarían con nuestra epistemología, tanto así, como para no solo no ser adoptadas sino incluso abolidas.

En esta medida, el presente trabajo brinda un marco inicial a lo que consideramos será la discusión de los próximos años en la ciencia económica, debido a los aportes de la minería de datos. Como economistas debemos desde la teoría, analizar cuál será el impacto de la minería de datos en nuestra ciencia, desde lo aplicado, preguntarnos si cambiará la forma en la que hacemos y resolvemos nuestras preguntas, y desde lo metodológico, identificar y aplicar las correctas herramientas y recomendaciones que pone a nuestra disposición el análisis del comportamiento humano a través de la Minería de Datos.

Buscando la mayor objetividad posible para dar respuesta a la pregunta de investigación, el presente trabajo se apoyará en tres medios. La primera sección refleja las posturas, las diferencias y las discusiones dentro de la ciencia económica entorno a la minería de datos. En la segunda etapa se expone el enfoque para la aplicación, donde se analizan los trabajos aplicados de minería de datos en la economía y se comparan frente a las metodologías estándar de la ciencia económica. En tercer lugar, se estimará un modelo de predicción utilizando la metodología CRISP-DM y las bases del (Minnesota Population Center, 2017) buscando comparar los resultados de un proceso metodológico clásico de la ciencia

[1] Definido como la corriente tecnológica enfocada en convertir todos los aspectos del comportamiento humano en datos.

[2] Segun la (OECD, 2015) "Seizing these benefits poses a formidable challenge for policymakers. In the years ahead, the pivot to a data-driven world will have important implications for policy ranging for privacy, consumer policy, competition, taxation, innovation, and specially jobs and skills."

económica con el trabajo de (Medina & Posso, 2011) frente a los resultados otorgados por un árbol de decisión. Por ultimo, en la Cuarta Etapa se concluye integrando los aportes de las secciones previas con el fin de dar respuesta a ¿Cuáles pueden ser los cambios en la ciencia cconómica dados los aportes de la Minería de Datos?.

2. ENFOQUE TEORICO

Aunque en la ciencia económica el termino minería de datos hacía referencia únicamente al proceso metodológico enfocado en encontrar las variables significativas para ser usadas en el modelo, en la actualidad la minería de datos se ha consolidado como un campo interdisciplinar independiente; que combina diferentes enfoques, tecnicas y algoritmos para reflejar los patrones que subyacen en los datos. La siguiente definición de minería de datos brindada por (Han, Kamber, & Pei, 2012) :

> *"Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically."*

Es la definición más interdisciplinar, y más apropiada académicamente; no solo porque incorpora el impacto de los grandes volúmenes de datos que se manejan en la actualidad, sino también, porque la sitúa como una etapa y no una metodología especifica dentro del proceso de KDD (Knowledge Discovery from Data)[3].

Los objetivos centrales de la minería de datos pueden resumirse en dos grandes grupos: **Describir** y **Predecir** (Han, Kamber, & Pei, 2012), (North, 2012), (Herrera & Vasquez, 2006); lo cual podría asemejarse a procesos de estadística descriptiva e inferencial, pero debido a que la información para la toma de decisiones que proveen los algoritmos de descripción y predicción de la minería de datos, poseen una característica de hipótesis, a saber, con respecto al número de grupos o clúster que se desean, a la variable endógena por la cual se opta o al número máximo de nodos esperado; dichos algoritmos van más allá de la estadística descriptiva clásica y proponen ciertas restricciones en los resultados entregados.

Un análisis más profundo desarrollado por (Herrera & Vasquez, 2006) concluye que son 4 las tareas principales de la minería de datos[4]:

Clasificación: Conociendo las características del individuo o grupo, se examinan los atributos para identificar la categoría o la clase para ser asignados. Algunos ejemplos típicos son: Clasificar en buenos o malos clientes, categorizar los individuos según sus gustos, entre otros.

3 Descrita como una secuencia iterativa de pasos compuesta por: Limpieza, Integración, Selección, Transformación, Minado, Evaluación de Patrones y Presentación según (Han, Kamber, & Pei, 2012).

4 Estos 4 grupos continúan inmersos en los 2 grupos anteriores (Describir y Predecir): **Descriptivos:** Asociación y Agrupamiento, **Predictivos**: Clasificación y Pronostico, sin embargo, brindan mayor claridad respecto al objetivo de los algoritmos. Dichas tareas también se encuentran dentro del grupo de Aprendizaje Supervisado, Semi-Supervisado y No supervisado o los dos grupos inicialmente descritos, y es que cabe aclarar que más allá del uso de la estadística según (Han, Kamber, & Pei, 2012) la minería de datos se apoya en un amplio conjunto de tecnologías como el Machine Learning, los Sistemas de Bases de Datos y Data WareHousing, e Information Retrieval o recuperación de información para Text Mining.

Pronóstico: El principal propósito es estimar una función de la forma $F_{(x,\delta)}$ donde x representa las variables y δ los parámetros de dicha función, usando el conjunto de variables de la muestra, para pronosticar el comportamiento de los datos para la variable objetivo Y. Entre los ejemplos más comunes se encuentran: La predicción de las compras futuras del individuo o grupo, las posibles actitudes que tomarán los individuos según sus atributos previos, entre otros.

Segmentación: En dicha tarea se busca agrupar la muestra o el conjunto de datos, en subgrupos o clústeres homogéneos, a diferencia de la clasificación estos no requieren conocer la variable endógena u objetivo. Sus principales casos de uso describen y generan posibles nichos de mercado usando los vecinos más cercanos al número de segmentaciones declarado.

Asociación: El principal objetivo de dichos algoritmos es identificar los elementos comunes o con mayor asociación en un ambiente declarado. Se han utilizado en gran medida para identificar las ventas cruzadas, o canastas de bienes de alta probabilidad.

Ahora bien, la discusión teórica de la minería de datos dentro de la ciencia económica se suele centrar en los problemas de predicción[5] y el impacto que puede tener el uso de dichas metodologías dentro de la epistemología de la ciencia debido al tratamiento que se le da al proceso generador de datos. Estos se pueden remontar a (Lovell, 1983), (Hoover K. D., 1995), (Hoover & Perez, 1999), (Mayer, 2000) ó (Breiman, 2001) y en ellos se puede evidenciar el tratamiento de la minería de datos como una aproximación al modelo de la LSE (London School of Economics) o el enfoque "*general-to-specific"* [6]*.*

El trabajo de (Lovell, 1983) presenta la minería de datos como mucho más que una etapa o un proceso, llegando a catalogarla incluso como un "*paradigma de investigación",* referenciando que los investigadores que trabajan las técnicas de la minería de datos suelen tener una metodología propia, e incluso, un modelo mental diferente para explicar el por qué, el cómo y las conclusiones de sus investigaciones, en principal medida, debido a las justificaciones en la elecciones de las variables explicativas.

(Sapra, 2014) define la minería de datos como un creciente y bien definido campo interdisciplinar que encuentra en algunos economistas declarados detractores (Lovell, 1983), (Breiman, 2001), (Feelders), (Mayer, 2000) debido a los supuestos que realizan los economistas con respecto al comportamiento humano, a la metodología estadística que los separa, y al objetivo final de cada enfoque. (Feelders) se pregunta por la aplicabilidad y contribución de la minería de datos en el análisis económico y concluye que la principal diferencia en los enfoques, depende de si la evaluación exploratoria será guiada por la teoría o guiada por los datos.

Como lo hace notar (Breiman, 2001) las metodologías clásicas de la ciencia económica difieren cultural y epistemológicamente de las aplicadas en la minería de datos debido a

[5] Esto se debe principalmente al propósito metodológico que busca dicho grupo dentro del análisis de datos y a su histórica ponderación frente a las otras metodologías, Según (Mayer, 2000) "that is where the problem usually arises in economics. Partly because of this my definition of data mining is narrower than what philosophers sometimes call 'peeking', 'snooping' or 'hunting with a shotgun'.

[6] A grandes rasgos dicha metodología supone que existe un modelo suficientemente complejo que puede explicar el funcionamiento de la economía como un todo y que todo modelo que de una manera simplificada demuestre simularlo será en definitiva una versión mejorada de dicho modelo.

que el tratamiento con respecto al proceso generador de datos en el primer caso, subyace en los supuestos del comportamiento del individuo bajo la teoría económica, mientras que en el caso dicho proceso generador se desconoce y son los algoritmos los que reflejan dicho comportamiento.

(Breiman, 2001) resalta como puntos débiles del primer enfoque la posibilidad del investigador para asumir el proceso generador de datos o su distribución, las deficiencias en los análisis debido a que el foco se centra en el modelo y no en el problema, y fundamentalmente expone como las pruebas de bondad de ajuste y los análisis de los residuales no son aplicables a menos de que exista una combinación lineal de los datos, llevándolo a concluir con respecto al Data Modeling Culture que (Breiman, 2001, pág. 203) "The question of how well the model fits the data is of secondary importance compared to the construction of an ingenious stochastic model."

En el caso de (Hoover K. D., 1995) su revisión de literatura sobre las metodologías de la minería de datos en la ciencia económica, demuestra que existe un interés teórico en mostrar el vínculo entre la adopción de dicha metodología y la epistemología misma de la ciencia, y en general, en demostrar como dichas metodologías aportan a la ciencia en el mismo sentido que pueden hacerlo la Organización industrial[7].

El trabajo realizado por (Hoover & Perez, 2000) resume 3 posiciones desde la ciencia económica para con la minería de datos: la evasión, el uso por obligación o su visión como etapa esencial, resumen el gran espectro partidario que existe en la ciencia económica frente a dicho campo interdisciplinar. En el primer aspecto -la evasión- se resalta la inaplicabilidad de la minería de datos en la ciencia económica a menos de que se ajuste la inferencia estadística que esta metodología provee, en la segunda -el uso por obligación- se la presenta como una etapa inevitable cuyos únicos resultados de interés son aquellos que sobrevivieron de manera "*Darwiniana*" a las diferentes especificaciones estadísticas y por ultimo -su visión como etapa esencial- bajo la cual los verdaderos descubrimientos económicos con garantía epistemológica se basan en la aplicabilidad de una minería de datos responsable, esto debido a que en definitiva el interés de la ciencia económica no es probar a través de los modelos los supuestos económicos sino como en la matemática o la física descubrir verdades escondidas en el comportamiento expresado a través de los datos.

Diferenciando entre el *economista analítico* que se basa en los resultados de un lenguaje formal (lógico o matemático) para desarrollar su ciencia, y el *economista sintético* que se basan en experiencias sensoriales para construir su ciencia, Hoover concluye que lo empírico-metodológico es en sí mismo inevitable y que en el mismo sentido en que los economistas y los politólogos están conectados a través de un *mutatis mutandis*[8] el futuro de la ciencia llevará a los economistas a crear esta relación con sus pares metodológicos.

[7] Como lo muestran (Aiginger, Mueller, & Weiss, 1998), la Organización Industrial se ha logrado consolidar como un campo de investigación independiente, diferenciado de los enfoques clásicos de la microeconomía, al brindarle las herramientas necesarias al investigador para tratar mercados con presencia de economías a escala, diferenciación de productos, alta concentración y en general, con características diferentes a los supuestos de la microeconomía clásica.

[8] "Mutatis mutandis es una frase en latín que significa 'cambiando lo que se debía cambiar'. Informalmente el término debe entenderse como "de manera análoga haciendo los cambios necesarios". Este término se utiliza frecuentemente en leyes y en economía. Implica que el lector debe prestar atención a las diferencias entre el argumento actual y uno pasado, aunque sean análogos." (Wikipedia, 2017)

En definitiva, la discusión teórica con respecto a aplicación de la minería de datos en la ciencia económica se centra en la garantía epistemológica que se puede certificar a través del uso de la misma. El desarrollo de pruebas estadísticas que garanticen la robustez de los modelos, el desarrollo de instrumentos independientes que mejoren la capacidad de interpretar la realidad económica y los supuestos económicos en la distribución o comportamiento de los datos, han sido la respuesta desde la ciencia económica para garantizar el uso *correcto* de la minería de datos. Sin embargo, (Hoover & Perez, 2000) resaltan que el problema epistemológico no radica en el desarrollo de estas, sino más bien, en la pregunta: Si algunos o todos los instrumentos anteriores son realizados. ¿sería esto una verdadera conclusión para la ciencia económica? o solo responde a los resultados de un contexto especifico y no a verdades más allá de los datos.

3. ENFOQUE APLICADO

Para comenzar este segmento es necesario aclarar que para el objetivo de la investigación existe una diferencia en el uso de la minería de datos en la ciencia económica y en el sector real. La aplicabilidad de la minería de datos dentro de la ciencia económica hace alusión al control de las variables explicativas, a la capacidad predictiva y al diseño del modelo (basado en datos o en algoritmos); mientras que su aplicabilidad en el sector real se puede resumir mejor en la investigación desarrollada por (Padhy, Mishra, & Panigrahi, 2012) donde se recopilan las principales industrias de aplicabilidad de la minería de datos, como lo son:

- Sector salud.
- Sector educación.
- Mercadeo.
- Manufactura.
- CRM (customer relationship management).
- Análisis de Lenguaje (Languaje Research).
- Análisis Web.
- Sector financiero.
- Planeación e infraestructura (principalmente en los últimos años).

Respecto a las investigaciones desarrolladas utilizando metodologías de minería de datos, (Liao, Chu, & Hsiao, 2012) exploran y analizan 216 artículos en 158 revistas académicas multidisciplinares entre el 2000 y el 2011; logrando clasificar en nueve categorías las DMT (Data Mining Techniques)[9] y obteniendo importantes conclusiones de su revisión con respecto al futuro de dichas técnicas.

Para comenzar, concluyen que las DMT tenderán a ser más “Expertise-Oriented”; resaltando la importancia en los antecedentes, el conocimiento y la experiencia del investigador con datos del tratamiento. Según (Han, Kamber, & Pei, 2012) dadas las técnicas tan diversas, el área de aplicación será orientado por los intereses del investigador, lo cual podría generar a futuro *investigaciones de autor*[10] que supondría un problema para con la minería de datos dentro de un contexto de ciencia económica, es decir, entiéndase el concepto como investigaciones perfiladas según los conocimientos que posee el autor

[9] “Neural networks, Algorithm architecture, dynamic prediction-based, Analysis of systems architecture, Intelligence agent systems, Modeling, knowledge-based systems, System optimization and Information systems “

en la industria en la que este se encuentra y su propuesta personal. El problema radica como lo resume (Foucault, 1969, pág. 29) en que:

"*el reexamen del texto de Galileo bien puede cambiar el conocimiento que tenemos de la historia de la mecánica, nunca puede cambiar a la mecánica misma. En cambio, el reexamen de los textos de Freud modifica al mismo psicoanálisis, y los de Marx, el marxismo*".

En el mismo sentido, si las investigaciones en la ciencia económica apoyados en DMT continúan su tendencia a ser orientadas por los expertos y no en la refrendación de la epistemología de la ciencia, nos encontraremos en un espacio donde las conclusiones aun en un mismo tema se contraríen.

En segundo lugar, acentúan la tendencia de los DMT a ser más "Problem-centered", es decir, dado que dichas técnicas resuelven problemas específicos según el objetivo del algoritmo mismo, (Liao, Chu, & Hsiao, 2012) suponen que, a futuro; las DMT se irán perfilando en el problema organizacional mismo, en diferencia al objetivo central de la economía aplicada, donde el encontrar verdades que cimientan la ciencia como un todo es la consigna. En tercer y último lugar, como se había declarado previamente, encuentran un margen entre el sector industrial o de contexto económico y las investigaciones académicas para campos específicos o en nuestro caso la ciencia económica, respecto a ello, (Liao, Chu, & Hsiao, 2012) citan:

"It is suggested that different social science methodologies, such as psychology, cognitive science and human behavior might use DMT as an alternative methodology. Integration of qualitative, quantitative and scientific methods and the integration of studies of DMT methodologies will increase understanding of the subject ".

(Liao, Chu, & Hsiao, 2012) esperan que dicho margen se solucione con la multidisciplinariedad de las DMT conjunto al perfeccionamiento de los nuevos aprendizajes otorgados por estas dentro de las investigaciones aplicadas a cada ciencia. Concluyendo que las DMT, podrán también retroalimentarse de las otras disciplinas y así incrementar su capacidad aplicativa.

Ahora bien, más allá de ahondar en los casos aplicados de la minería de datos por sector económico o técnica aplicada, el interés de la presente sección es mostrar la aplicabilidad de la minería de datos en la ciencia económica con trabajos como los de (Lovell, 1983), (Breiman, 2001) y (Varian, 2014). En ellos se reflejan tanto las restricciones de aplicabilidad de dichas metodologías, como las ventajas en la investigación aplicada frente a las metodologías clásicas de la ciencia económica.

La selección de las variables explicativas de un modelo a través de la minería de datos y su discrepancia frente a la ciencia económica[11]. Suponen un tema de tratamiento en el trabajo de (Lovell, 1983); en él, Lowell demuestra como el incremento en el conjunto de regresores o variables explicativas reduce ampliamente la probabilidad de reportar correctamente cuando la hipótesis nula de no significancia es correcta, o lo que es igual, el incremento en el conjunto de datos utilizados en la regresión distorsiona el verdadero valor

[11] Donde la selección previa del conjunto de variables explicativas se basa en modelos teóricos y no únicamente en la relativa significancia estadística o como lo denomina (Lovell, 1983) en "well-defined a priori considerations".

de Alpha o de significancia estadística bajo el cual se debería probar la significancia de las variables[12], demostrando que valores altos de R^2 y significativos niveles del t-estadístico pueden ser producidos simplemente por el incremento en el número de regresores; y en definitiva, llevar al analista a engañosas conclusiones, tanto para sus pruebas de hipótesis como a los supuestos del proceso generador de datos.

En la misma línea se encuentra el trabajo de (Breiman, 2001), quien refleja la división en la creación de los modelos de lo que él denomina las 2 culturas estadísticas. El enfoque ***The Data Modeling Culture (DMC),*** donde asumimos que el proceso generador de datos es estocástico y estimamos las variables dependientes como una función de las variables independientes, de los parámetros y de un ruido blanco; en contraposición a ***The Algorithmic Modeling Culture (AMC),*** donde se desconoce el proceso generador de datos y a través de un algoritmo se encuentra el patrón que pronosticará los valores de la variable endógena.

Entre las diferencias que presentan ambas culturas de modelamiento no solo están los supuesto sobre el proceso generador de datos, también, la validación de los modelos son una importante diferencia aplicada. En el caso del DMC los test de bondad de ajuste y el análisis de los residuales son los que otorgan testimonio de un correcto modelado, por el contrario, en el AMC es la precisión predictiva la garante en la validación del modelo.

Según (Breiman, 2001) lo anterior lleva en definitiva a las DMC a realizar el análisis focalizado en el modelo y no en el problema. y es que la asunción bajo la cual una correcta representación de la naturaleza de los datos será generada por un modelo paramétrico basado en las medidas de bondad de ajuste, no es suficiente garante aplicativo, para la toma de decisiones. (Breiman, 2001) refleja que las ventajas aplicadas de los AMC frente a los DMC pasan por la validación cruzada de los modelos, por su precisión predictiva y por su manejo de la multidimensionalidad, (Breiman, 2001) concluye respecto al uso de los DMC y AMC:

"*Approaching problems by looking for a data model imposes an a priori straight jacket that restricts the ability of statisticians to deal with a wide range of statistical problems. The best available solution to a data problem might be a data model; then again it might be an algorithmic model. The data and the problem guide the solution. To solve a wider range of data problems, a larger set of tools is needed."*

(Varian, 2014) realiza una comparación de las clásicas metodologías de estimación económica como los modelos de regresión lineal, modelos Logit o Probit, o modelos de series de tiempo frente a metodologías tradicionales de la minería de datos como Decision Trees, Random Forest y LASSO. Las diferencias aplicadas encontradas por Varian resaltan la poca precisión que tienen los modelos de la economía clásica para trabajar en contextos "out-of-sample", es decir, aun cuando los test de bondad de ajuste, las pruebas de hipótesis o las comprobaciones en los términos de error reflejan la veracidad del modelo clásico; para (Varian, 2014) la verdadera pérdida de confianza en el modelo, se refleja en la incapacidad de este para trabajar con predicciones fuera de la muestra.

El fenómeno descrito anteriormente hace referencia en la minería de datos a los problemas de sobreajuste o "overfitting problems". Las propuestas aplicadas desde la minería de datos como el manejo de la complejidad a través de la *regularización,* los cross-data-validation o

[12] Para mayor claridad se anexa la tabla del trabajo de (Lovell, 1983).

las sub-muestras de entrenamiento y prueba, son herramientas aplicadas propuestas desde la minería de datos hacia la econometría clásica y podrían convertirse en garantes de aplicación por su funcionalidad predictiva fuera de las muestras.

La primera comparación realizada por (Varian, 2014) es entre un modelo Logit y un modelo de árboles de decisión. Dicho modelo busca establecer la probabilidad de supervivencia en el Titanic usando como variables explicativas la edad y la clase en el tiquete de compra. Como lo demuestra Varian la regresión Logistica es incapaz de reflejar la significancia de la edad en la probabilidad de supervivencia[13] debido a que dicha variable resulta determinante solo en las observaciones de los extremos, es decir, la edad mostró poseer un peso en la probabilidad de supervivencia para los niños (menores a 16 años) y los adultos mayores (mayores a 80). Concluye entonces Varian que las ventajas aplicativas en este caso pasan por la capacidad de reflejar los patrones escondidos en los datos de una manera simple y con mayor sustento que la justificación del conocimiento del investigador sobre los datos o sus supuestos.

En su segunda comparación Varian demuestra el caso aplicado en una investigación para el préstamo de hipoteca. Donde la variable de raza demuestra ser altamente significativa en el modelo Logit y relativamente no significativa bajo un árbol de decisión, es de notar que inclusive si se extrae la variable del modelo la capacidad predictiva del mismo no se ve afectada, y en definitiva como concluye (Varian, 2014) aun cuando dicho resultado pueda ser por la colinealidad entre la variable “raze” y demás variables usadas en el modelo; es como mínimo notoriamente interesante, que tanto su presencia como su ausencia en términos de la validación aplicativa según la perspectiva de la minería de datos, lo anterior estructure un buen modelo.

Por último, (Varian, 2014) realiza un importante ejercicio respecto a la selección de las variables utilizadas en un modelo de predicción del crecimiento económico comparando 4 modelos[14]. En ellos encuentra que para los primeros 5 predictores los modelos concuerdan en su significancia, sin embargo, la ponderación de las variables a partir de allí contrasta entre modelos[15], y es que (Varian, 2014) respecto a la discusión entre la causalidad que brindan los modelos de la ciencia económica versus la capacidad predictiva de los modelos de la minería de datos concluye que; debido a que la extracción del verdadero impacto en la variación de alguna de las variables predictivas implica la realización de un experimento (natural o diseñado) el uso de un modelo predictivo con mayor capacidad en los términos de la minería de datos permitiría mejorar la capacidad de extracción del efecto cambiante cuando este ocurra[16].

[13] Es sabido que para el tratamiento de la Edad se suelen usar Dummies por los segmentos para tratar la no linealidad en los impactos sobre la variable endógena, sin embargo, bajo los modelos CART es más fácil evidenciar la significancia de posibles segmentos en las explicativas.

[14] Los modelos utilizados son Bayesian model averaging, CDF(0),LASSO y Spike-and-Slab de los trabajos realizados por Ley and Steel (2009), Sala-i-Martín (1997) y Hendry and Krolzig (2004) respectivamente.

[15] Para mayor claridad se anexó la tabla de (Varian, 2014).

[16] Permitiendo incluso a lo que en mi opinión será un gran descubrimiento para la ciencia económica y es la posibilidad de extraer los impactos de una variable de intervención sin la necesidad de un grupo de control.

4. ENFOQUE METODOLOGICO

4.1. EL MODELO CRISP-DM

La minería de datos va más allá de un conjunto de herramientas de descripción o predicción para la solución de problemas analíticos apoyados en grandes bases de datos. La sistematización de los procesos para encontrar verdadero conocimiento en la toma de decisiones acertadas es fundamental desde lo metodológico; Según (Herrera & Vasquez, 2006): El SEMMA (Simple, Explore, Modify, Model, Asses) y el CRIPS-DM (Cross Industry Standar Process for Data Mining) son los marcos metodológicos de mayor aplicación respecto a la minería de datos. El primero, desarrollado por una empresa líder en el sector de análisis (SAS) para los procesos dentro de su organización. La segunda, propuesta por un consorcio de empresas europeas incluyendo la alemana AG y la inglesa SPSS.

La encuesta realizada por (Piatetsky, 2014) con la pregunta " *What main methodology are you using for your analytics, data mining, or data science projects?* " muestra como *CRISP-DM* no solo es la metodología con mayor acogida con un 43%, frente a la segunda más utilizada *My Own* con un 27.5%, o *SEMMA* en tercer lugar con un 8.5%. además de ello, resalta su prevalencia comparado con la encuesta de 2007. (Wirth & Hipp, 2000) prueban y aplican la metodología demostrando su buena estructura y flexibilidad para aplicaciones independientes del sector y la tecnología utilizadas; resaltando su capacidad comunicativa y documental para el manejo del componente interdisciplinar de la minería de datos.

El CRISP-DM puede entenderse como una propuesta para la gestión del ciclo de vida de un proyecto analítico, Según (Wirth & Hipp, 2000) compuesto por 6 fases y 4 niveles de abstracción desde lo general a lo especifico[17], cabe aclara que para los objetivos del documento no se abordaran todos los procesos, las fases de Evaluación y Despliegue, son unas limitantes debido a los tiempos y recursos.

La explicación resumida de cada fase se presenta a continuación:

ETAPA DENTRO DEL CRISP-DM	DEFINICION U OBJETIVO CENTRAL
Comprensión del negocio o comprensión Organizacional	Esta fase inicial se centra en conocer los objetivos y requerimientos del proyecto, traducir el problema de investigación en un problema definido dentro de la minería de datos.
Comprensión de los Datos	Recoge tanto la parte de la recolección y organización de los datos como los primeros descubrimientos de conocimiento dentro de estos, el descubrimiento de posibles sub-grupos o hipótesis escondidas.
Preparación de los Datos	Cubre todas las actividades para organizar el grupo final de datos con los cuales se trabajará, en el formato o diseño requerido por el modelo a trabajar.

[17] (Fases, Tareas específicas, Tareas especializadas, Instancias de los procesos).

Modelación	Aplicar las DMT más acordes con el problema a solucionar.
Evaluación	Esta etapa comprende el cumplimiento de los objetivos, la revisión del proceso por los pares y el listado de posibles acciones.
Despliegue	En esta etapa se generan los reportes, se hacen las modificaciones recomendadas por el análisis y se evalúa si implementación con respecto al objetivo deseado.

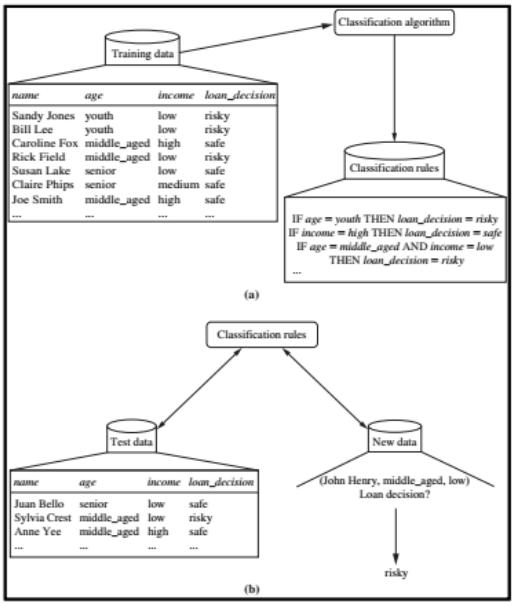
Para esta sección se estimará un modelo de predicción utilizando la metodología CRISP-DM y las bases del (Minnesota Population Center, 2017) buscando comparar los resultados de un proceso metodológico clásico de la ciencia económica con el trabajo de (Medina & Posso, 2011) frente a los resultados otorgados por un árbol de decisión técnica clásica de la minería de datos.

4.2. ARBOLES DE DECISIÓN, ARBOLES DE CLASIFICACIÓN Y ARBOLES ALEATORIOS

El uso de clasificadores es una de las metodologías más utilizadas en la minería de datos. El objetivo central es predecir los atributos o variables explicativas para un modelo en el cual se conoce la variable dependiente y esta es categórica. Generalmente se puede resumir en 2 etapas el proceso de clasificación: la construcción del modelo basado en la muestra de entrenamiento (*Learning Step*), y la prueba de precisión del modelo para clasificar nuevos datos (*Classification Step*).

Dentro del *Learning Step* el algoritmo busca la relación entre los valores de los vectores de las variables explicativas y el clasificador o variable dependiente categórica, las tuplas[18] generadas por dicho proceso se convierten en la muestra de entrenamiento y luego estas son utilizadas en el *Classification Step* para probar la precisión del clasificador en un nuevo conjunto de datos. El grafico utilizado por (Han, Kamber, & Pei, 2012) resume dicho proceso y lo presenta de una forma intuitiva:

[18] "*Una tupla es una lista ordenada de elementos. Una n-tupla es una secuencia (o lista ordenada) de n elementos, siendo n un número natural (entero no-negativo) ...Las tuplas suelen anotarse listando sus elementos entre paréntesis "(,)", separados por comas. Por ejemplo, {(2,7,4,1,7)} denota una 5-tupla. Las tuplas se emplean para describir objetos matemáticos que tienen estructura; es decir, que son capaces de ser descompuestos en un cierto número de componentes. Por ejemplo, un grafo dirigido se puede definir como una tupla de (V, E), donde V es el conjunto de nodos y E es el subconjunto de V × V que denota las aristas del grafo."* (Wikipedia, 2017)



Esto nos lleva a los Arboles de Decisión, una representación gráfica de los algoritmos de clasificación previamente descritos cuyos nodos representan las pruebas sobre los atributos, las raíces los resultados de dicha prueba y las hojas su relación con la variable de clasificación; para su aplicación se crea un camino entre el nodo inicial y las hojas a través de un algoritmo de inducción o diseño *top-down* con los 3 parámetros:

- **D** = La partición de los datos, que en caso inicial es el conjunto total de la muestra.
- **Attribute_List** = La lista de atributos que describe las tuplas.
- **Attribute_Selection_Method** = Proceso Heurístico para la selección del atributo que mejor discrimina la tupla dada según el clasificador.

La información sobre el funcionamiento del algoritmo y las diferentes clases de métodos de selección pueden ser encontradas en (Han, Kamber, & Pei, 2012).

Dentro de los métodos para la mejora en el rendimiento de los árboles de decisión fuera de la muestra, se puede optar por una "*poda"*, por técnicas para control de los problemas de sobremedida, por test de significancia, por la validación cruzada de modelos, por una matriz de confusión, entre otros.

En nuestro caso utilizaremos un Classification Tree y un método de ensamblaje Random Forest en el cual existe una colección N de árboles de decisión, generados individualmente a través de un proceso aleatorio en la selección de los nodos o atributos y cuyo modelo final se determina bajo el criterio del "más repetitivo". Específicamente se utilizará el paquete de R -randomForest- donde el algoritmo para la construcción de los árboles de decisión se basa en la metodología CART.

4.3. DESARROLLO METODOLOGICO

El objetivo central de la presente etapa es poder concluir metodológicamente en la aplicación de técnicas de minería de datos, frente a las metodologías clásicas de la ciencia económica[19], por ello en el ejercicio de las etapas del CRISP-DM la profundización no es exhaustiva y por el contrario se presentan los hallazgos de ambas partes (usando el CRISP-DM y el trabajo de (Medina & Posso, 2011) para comparar los resultados obtenidos por un modelo Logit con Boosting y un Decission Tree (posteriormente un Random Forest.).

4.3.1. COMPRENSIÓN ORGANIZACIONAL

El objetivo central del trabajo desarrollado por (Medina & Posso, 2011) busca encontrar los principales determinantes para la decisión de migración de los colombianos en los años 1990-2000-2005. La etapa de comprensión organizacional busca delimitar el objetivo en la investigación y formular la mejor herramienta o algoritmo de minería de datos según el caso a tratar, para ello la ciencia económica se basa en la evaluación exhaustiva del marco teórico y el estado del arte respecto al problema, a diferencia de la metodología CRISP-DM cuyo medio, es la conceptualización organizacional y la problemática misma.

(Medina & Posso, 2011) encuentran que para el 2005 aproximadamente 3 millones de colombianos residían en Estados Unidos, una cifra relevante si comparamos con los 2.5 millones que residen en Medellín. Además de su magnitud, Medina y Posso definen los principales problemas de migración en 2 frentes; la necesidad de los gobiernos locales por conocer las causas en la decisión de migración (más específicamente comprender los causales en el proceso de *fuga de cerebros*), y el conocimiento individual que puedan obtener las personas que desean viajar a Estados Unidos respecto a una eventual evaluación en la decisión de permanencia.

En el análisis de los antecedentes en las teorías de movilización de capital humano, (Medina & Posso, 2011) presentan los principales supuestos que con el tiempo se han desarrollado en la ciencia económica. Para comenzar, la escuela Neo-Clásica de economía donde se ha definido el fenómeno de migración como un proyecto de vida de tiempo indefinido, por lo cual se entendería que el proceso de retorno solo se daría, si el proceso de migración es fallido. Otra posición y asunción es la realizada por la NELM (New Economy on Labor Migration) quienes defienden que dicha decisión de retorno es una estrategia lógica predefinida, por último, se presenta la posición desde los Analistas Estructurales, los cuales propone el contexto como el principal determinante, incluyendo las características del país de origen y del país destino representados tanto en los factores sociales como institucionales.

Desde la perspectiva de la minería de datos, el objetivo de la presente etapa indaga en la estructura organizacional y el objetivo de la pregunta de investigación; logrando centrar las preguntas en un perfil de minería de datos que permita la clasificación de la pregunta central dentro de las 4 grandes tareas de la minería de datos. Realizando dicha tarea sobre el

[19] La ciencia económica no tiene un framework metodológico especifico, sin embargo, en términos generales las investigaciones en la ciencia económica poseen la misma estructura lo cual nos permite comparar con la sistematización del CRISP-DM.

trabajo de (Medina & Posso, 2011), hallamos que el principal objetivo de encontrar los principales determinantes en la decisión de migración de los Colombianos en Estados Unidos para los años de 1990,2000 y 2005 podría ser llevado a un problema de clasificación, puesto que en este caso, además de conocerse la variable endógena a modelar (Probabilidad de Migración), se conocen las características que definen a ambos grupos, permitiéndonos identificar los principales atributos para la asignación dentro de las clases.

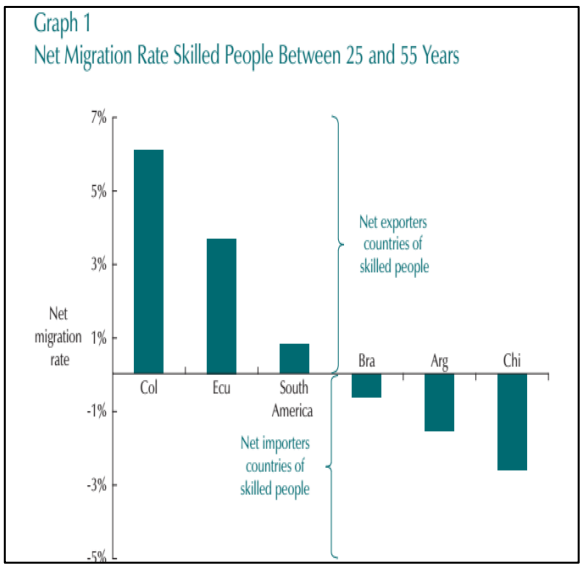
Como lo podemos apreciar, la diferencia central en la etapa de comprensión organizacional entre la metodología de la minería de datos y la ciencia económica subyace en que para la ciencia económica, son las bases teóricas y las investigaciones previas, quienes sustentan y cimientan los siguientes pasos, mientras que, en el caso de la minería de datos específicamente en el CRISP-DM, es el problema quien orienta las futuras decisiones, confirmando los hallazgos previamente presentados en este documento.

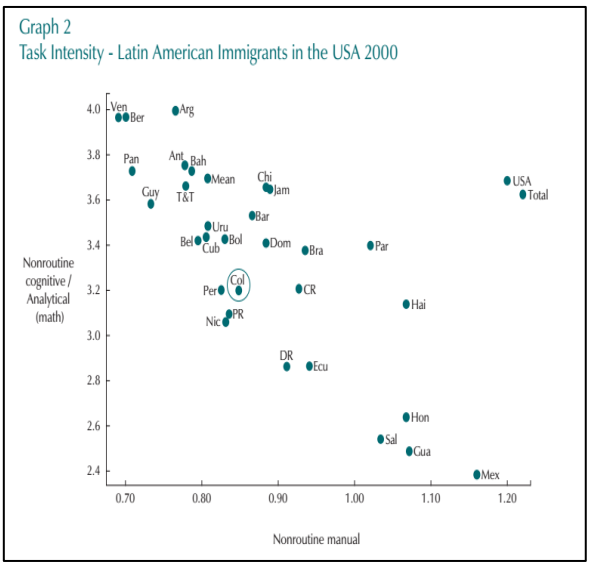
4.3.2. COMPRENSIÓN DE LOS DATOS

El objetivo de esta etapa es brindar una comprensión general respecto al proceso de recolección y a las estadísticas descriptivas de las variables de tratamiento. Respecto a (Medina & Posso, 2011) la fuente principal de los datos fue el IPUMS-International para las muestras de Estados Unidos 1990-2000-2005, una muestra general de la estructura de los datos y sus estadísticas descriptivas son presentadas en el código de los anexos[20].

Los hallazgos presentados por (Medina & Posso, 2011) reflejan la metodología general asumida en la ciencia económica, presentando la estadística descriptiva inicial y levantando las primeras hipótesis sobre los comportamientos y correlaciones en los datos. La estadística descriptiva inicial muestra a Colombia como un exportador neto de capital humano, reflejando la problemática de la *fuga de cerebros* basado en la selección positiva explicada por Medina y Posso. En el segundo gráfico, se presenta el nivel de intensidad del trabajo desarrollado, lo cual permite apreciar que los colombianos en promedio, se encuentran desarrollando trabajos en su mayoría analíticamente intensivos, pero por debajo de la media latinoamericana.

[20] Tanto la extracción específica para el IPUMS-International, como la transformación de las variables difieren del trabajo realizado por (Medina & Posso, 2011) por lo cual pueden existir diferencias en el número de variables y el tratamiento de las mismas, sin embargo, se intentó replicar el dataset en la mayor medida posible.





Fuente: (Medina & Posso 2011)

Siguiendo con el ejercicio de profundización en las estadísticas de los datos encontramos que el conocimiento adquirido al realizar esta etapa bajo el enfoque general de la ciencia económica es muy significativo. Más, si tenemos en cuenta que aún nos encontramos en una fase exploratoria y, sin embargo, los hallazgos reflejan hipótesis escondidas. Por ejemplo, evaluando el contexto, encontramos que alrededor del 70% de los encuestados entre las tres muestras no se encontraba viviendo con su madre biológica, oscilando entre un 57.6% y 80.97% entre las muestras; sumado a esto, aproximadamente 73.81% de los encuestados viven en un hogar con entre 2 a 6 personas, más específicamente el 43.23% conviven con otras 4 a 6 personas. Otra estadística significativa muestra que el 73.36% de los encuestados en el 2000 llevaban más de 5 años desde su migración.

Por el contrario, el ejercicio metodológico de la presente etapa desde el modelo CRISP-DM, busca previamente la identificación de las características de los datos, las particularidades de las fuentes (Minnesota Population Center, 2017) y la tipología y significado de los mismos, en definitiva, busca centrar los esfuerzos en la comprensión organizacional de los datos desde las fuentes primarias, logrando establecer la selección en bruto y sus beneficios o inconsistencias para el ejercicio metodológico, por ejemplo la importante ausencia de la posible pregunta de clasificación para la muestra de 2005.[21]

En definitiva, en la presente etapa, aunque existen ciertas similitudes entre la ciencia económica y el CRISP-DM, las diferencias importantes resaltan en la incorporación del diccionario de datos en el CRISP-DM. Etapa que por lo general no es abordada en la ciencia económica y puede ser un determinante en las criticas al poco detalle en el tratamiento de los datos por parte de la ciencia económica; más allá de ello, no se encuentran diferencias significativas, e incluso, las conclusiones en ambas metodologías llevan a un punto común,

[21] Para el presente trabajo la variable endógena se elaboró desde el atributo de la fuente (Migration5), modificado los niveles de la variable categórica.

encontrar las relaciones a priori entre las variables y contextualizar al investigador y al lector sobre la muestra.

4.3.3. PREPARACION DE LOS DATOS

El objetivo de esta etapa es crear las submuestras y corregir o modificar los tipos de los datos. Lo anterior, con el fin de adecuar los datos de la mejor manera posible para corregir problemas de sesgos o colinealidad, e incluso de capacidad computacional o desempeño para los algoritmos a utilizar. En este sentido el presente trabajo muestra 2 grandes ejemplos; para comenzar, (Medina & Posso, 2011) demuestran cómo puede existir un proceso de contaminación en el análisis que incluso lleva a la subestimación en los efectos de las variables explicativas, al existir una interacción entre submuestras no diferenciadas y cuya solución modifica la función de estimación de los regresores[22].

Por otro lado, desde la metodología CRISP-DM por la forma en la que funcionan los algoritmos de clasificación de los árboles de decisión explicados previamente, las variables categóricas con un alto número de niveles o clases, impiden el objetivo de simplicidad en la analítica del algoritmo y su ejecución; por lo que se hace necesario generar nuevos niveles o categorías para las variables categóricas y convertir a variables numéricas ciertos atributos.

Las conclusiones en la implementación de esta etapa muestran la importancia de incorporar un espacio para la explicación detallada en la manipulación de la muestra, tanto para garantizar la eficiencia de los modelos como dar respuesta a la crítica metodológica en la manipulación de los datos, cabe resaltar que dicho ejercicio de modificación de las variables exógenas para la mejora del desempeño del algoritmo puede crear sesgos analíticos para el caso de los árboles de decisión, sin embargos estos solo disminuyen el nivel de detalle analítico mas no los resultados generales. En conclusión podemos apreciar que, aunque el garante epistemológico desde la ciencia económica busque la profundización en los elementos diferenciadores del modelo econométrico, la validez de los modelos siempre quedará inconclusa si no existe un espacio para el tratamiento y preparación de los datos.

4.3.4. MODELADO

El objetivo central de esta etapa dentro del CRISP-DM es mostrar los resultados de la técnica de minería de datos aplicada. Dentro de nuestro contexto, el objetivo central es comparar los resultados otorgados por una técnica clásica de la ciencia económica[23] frente a una metodología clásica de la minería de datos, específicamente compararemos los resultados obtenidos por una estimación a través de un modelo Logit frente a los resultados otorgados por una estimación a través de un árbol de decisión.

(Perlich, Provost, & Simonoff, 2003) elaboran un ejercicio de pruebas de comparación entre los modelos Logit y la inducción por arboles de decisión específicamente el enfoque de C4.5, encontrando 4 significativas diferencias relacionadas con: su desempeño en la

[22] La explicación detallada del proceso de solución del sesgo de contaminación se puede encontrar en (Medina & Posso, 2011) y en el trabajo citado por ambos en Heckman and Robb (1985).
[23] La técnica implementada por (Medina & Posso, 2011) fue un modelo de regresión lineal a través de unos MCO y posteriormente una estimación para modelos Logit y Probit.

clasificación, su desempeño según el tamaño de la muestra, su efectividad en la selección de las variables explicativas y su capacidad para el manejo del ruido. Para ello, Perlich et al utilizan como medida, el área bajo la curva de aprendizaje de ambos modelos, incluyendo sus posibles variaciones y mejoramientos de ensamble. (Perlich, Provost, & Simonoff, 2003) concluyen que no solo el desempeño es comparable entre modelos sino también entre especificaciones de ensamble y control de los sesgos. Para concluir, otorgan una guía respecto al uso de uno u otro modelo e incluyen la propuesta de modelos híbridos como una opción viable.

"*In particular, consider the following strategy.*

1. Run C4.5-PET with the maximally feasible training-set size. For example, use all of the data available or all that will run well in main memory. C4.5 typically is a very fast induction alternative (cf., Lim, Loh, and Shih, 2000).

2. If the resultant AUR is high (0.85 or greater) continue to explore tree-based (or other nonparametric) options (for example, BPET, or methods that can deal with more data than can fit in main memory (Provost and Kolluri, 1999)).

3. If the resultant AUR is low, try logistic regression.

An alternative strategy is to build a hybrid model. …where tree building takes the probability estimation from a logistic regression model as an additional input variable."

Los hallazgos presentados por (Medina & Posso, 2011) pasan primero por una exhaustiva explicación de las correcciones del modelo a utilizar; lo anterior debido a un problema de sesgo de contaminación, que no permite identificar claramente la variable endógena entre las muestras. Para los objetivos del presente trabajo, es este un caso ejemplificador de los tecnicismos sujetos a críticas en la mayoría de trabajos dentro de la ciencia económica. Como se ha declarado previamente existe un propósito claro dentro de la forma clásica de hacer econometría que se concentra en convencer al lector de que tanto los supuestos como la especificidad del modelo son los más adecuados para analizar tal o cual caso.

Específicamente (Medina & Posso, 2011) presentan la metodología apropiada para corregir los problemas de sesgo, lo que lleva a modificar los parámetros del estimador de los impactos de las variables explicativas con el fin de modelar correctamente a través de un modelo Logit los factores determinantes en la probabilidad de estadía de los colombianos en Estados Unidos. Los resultados del modelo final muestran claramente una significancia respecto al nivel educativo en aquellos colombianos que deciden quedarse en Estados Unidos y en definitiva se acepta la hipótesis bajo la cual el proceso migratorio de colombianos en Estados Unidos es un proceso de selección positiva, donde solo aquellos que tienen características educativas por encima del nivel medio son quienes terminan asentándose en E.U.

Otro hallazgo presentado por (Medina & Posso, 2011) refleja la importancia de ser mujer en la decisión de permanencia para el análisis entre 1990-2000 incrementando la probabilidad de la estadía, exceptuando los casos en los cuales se encontrase un menor de 10 años en el hogar o una persona mayor a 60. Sin embargo, aunque se propone que

dicho patrón puede reflejar una característica estructural en el proceso, el ejercicio desarrollado para los factores determinantes en el periodo 2000-2005 contraria dichos resultados[24].

El ejercicio de modelado desde la tecnica de árboles de decisión[25] para armonizar con los objetivos del presente documento se dividió en 3 etapas. En la primera, las variables utilizadas para los regresores fueron las mismas que mostraron significancia en el ejercicio de (Medina & Posso, 2011) para las muestras de 1990-2000, en dicho ejercicio se puede apreciar que en ambos modelos las variables de edad, nivel educativo y número de hijos en el hogar son significativas, sin embargo, en el modelo del árbol de decisión la variable de sexo no muestra ser significativa para la muestra de 1990-2000 asemejando el ejercicio de (Medina & Posso, 2011) para 2000-2005.

En la segunda etapa se utilizó un modelo sobre todas las variables de la muestra, dejando al algoritmo ser quien definiese las variables significativas a través de su proceso de clasificación; en este se encontraron nuevas variables significativas, tal es el caso de la presencia en el mismo hogar por parte del individuo en años pasados, lo que puede aproximarse al hallazgo de (Medina & Posso, 2011) bajo el cual el tiempo de presencia en Estados Unidos puede ser un factor determinante. además, el estado de ciudadanía mostró ser una nueva variable significativa que no había sido contemplada anteriormente, a ello se suman las variables de ingreso, de tipo de trabajo, la presencia de un hijo de edad adulta en el hogar y la asistencia a la escuela por parte de un miembro en la familia lo que aporta un descubrimiento de nuevo conocimiento no basado en los antecedentes teóricos o en los supuestos.

Por último, en la tercera etapa se utilizó el método de ensamble de árboles aleatorios o Random Forest explicado previamente. Los resultados muestran con alto nivel de significancia a los ingresos, a la edad, al hogar de migración, a la región, al tamaño de la familia y a la categoría de la ciudadanía; mejorando las medidas de precisión de un 17.01% de error en la clasificación del modelo árbol de decisión usando las variables propuestas por (Medina & Posso, 2011), a un 16.82% utilizando la metodología de Classification Tree sobre todas las variables de la muestra; por último, se estimó el modelo con una metodología de ensamble Random Forest; el cual mejora el desempeño al optimizar sus parámetros, pasando a un 16.12% de error en la clasificación.

La conclusión en la comparación de los modelos metodológicos de la ciencia económica frente a las metodologías clásicas de la minería de datos, nos permiten, apoyados en el trabajo de (Perlich, Provost, & Simonoff, 2003); concluir que la implementación de uno u otro no necesariamente es excluyente, por el contrario, ambas metodologías se han enriquecido por las propuestas de su contraparte, por ejemplo, el uso de metodologías de ensamblaje como el Bootstrap en el caso de (Medina & Posso, 2011) son reflejo de que la

[24] (Medina & Posso, 2011) atribuyen dicho cambio en los determinantes al periodo de estudio haciendo alusión a la crisis vivida en 1999 en Colombia.

[25] Tanto la representación gráfica de los árboles de decisión, como las medidas de validación, se presentan en los Anexos y quedan en el código adjunto.

ciencia económica también puede aprender de las nuevas técnicas de ensamble, o de la misma manera. el caso citado en (Perlich, Provost, & Simonoff, 2003) del California Housing data set, donde las variables para la construcción del árbol de clasificación fueron pre-propuestas por un modelo Logit; ambos reflejan bondades tomadas de parte y parte para enriquecer el fundamento técnico de uno o las ausencias teóricas del otro.

Los hallazgos específicos en nuestro caso de estudio nos permiten verificar que pueden existir variables significativas que han sido pasadas por alto al no haber estado sustentadas específicamente en los modelos teóricos, y que en la misma medida en la que los hallazgos de la organización industrial han alimentado nuestra ciencia, estas nuevas herramientas proponen un nuevo foco desde lo empírico para analizar lo teórico.

5. CONCLUSIONES

En esta sección se brindarán las conclusiones a la pregunta de investigación integrando las conclusiones de las 3 etapas anteriores. El presente documento buscó proporcionar al lector una perspectiva holística de los cambios que pueden generar en la ciencia económica las nuevas metodologías de la minería de datos, en la integración de las etapas se descubrió que el impacto de la minería de datos en la epistemología de la ciencia puede estar siendo subestimado.

El avance en las características técnicas que ha desarrollado la econometría puede estar guiando a la ciencia económica a enfocar sus esfuerzos en la especificidad de los modelos y no en las verdades en las decisiones de los individuos oculto en los datos. Asegurando que el garante de la filosofía de la ciencia se halla en las pruebas de bondad de ajuste de los modelos o en la capacidad técnica del investigador en sustentar un comportamiento a-priori de los datos, basado en supuestos teóricos o evidencias empíricas fuertemente contextualizadas y focalizadas, respecto a ello, la minería de datos plantea un nuevo interrogante, demostrando que es posible encontrar nuevas verdades dejando de lado los supuestos en el comportamiento humano y de los datos.

Sumado a ello, las nuevas propuestas de validación en los modelos econométricos que se encuentran en estas nuevas técnicas, tales como la medición de la capacidad predictiva, el desempeño fuera de la muestra e incluso las pruebas sobre los datos de entrenamiento y validación, aportarán al saber-hacer de nuestra ciencia modificando nuestra forma de plantear las conclusiones, es decir, podremos ampliar nuestro horizonte resolutivo dentro de la misma creación del modelo.

Teniendo en cuenta lo anterior, es importante resaltar que la aplicabilidad y metodología propias de la minería de datos brindarán aportes sustanciales al que hacer de la ciencia. Son una alternativa que se suma al conjunto de herramientas técnicas e incluso metodológicas, donde la incorporación de aspectos específicos como el diccionario de datos, el tratamiento de los mismos o la evaluación segmentada, servirán de faros para contextualizar a los pares y responder a las críticas de la metodología actual, donde la manipulación de los datos es una opción que preocupa e incluso resta validez a los correctos modelos.

En definitiva, Podemos concluir que las discusiones propuestas desde la creciente adopción de la minería de datos en la ciencia económica pasarán por la contrastación de las técnicas, el desarrollo de nuevas medidas para la validación de los modelos, la creación de modelos híbridos y una contante retroalimentación entre las disciplinas. donde los costos en nuestros pivotes técnicos serán altamente superados por nuestros beneficios teóricos aplicados y metodológicos.

6. BIBLIOGRAFÍA

(NCR), P. C., (SPSS), J. C., (NCR), R. K., (SPSS), T. K., (DaimlerChrysler), T. R., (SPSS), C. S., & (DaimlerChrysler), R. W. (2000). *CRISP-DM 1.0 step-by-step data mining guide.* SPSS Inc.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 199-231.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 199-231.

Feelders, A. (s.f.). Data Mining in Economic Science.

Foucault, M. (1969). ¿Qué es un Autor? *Bulletin de la Société française de philosophie*, 73-104.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann.

Herrera, M. A., & Vasquez, J. D. (2006). *ESTUDIO SOBRE EL ESTADO DE LAS SOLUCIONES ICT Y DE LOS CASOS PRÁCTICOS DE APLICACIÓN DE LA MINERÍA DE DATOS A NIVEL MUNDIAL EN AL MENOS 5 CASOS REPRESENTATIVOS.* Medellin: UNIVERSIDAD EAFIT.

Hoover, K. D. (1995). In Defense of Data Mining: Some Preliminary Thoughts. En K. D. Hoover, & S. Sheffrin, *Monetarism and the Methodology of Economics- Essays in Honor to Thomas Mayer.* Edward Elgar Publishing.

Hoover, K. D. (1995). WHY DOES METHODOLOGY MATTER FOR ECONOMICS? *The Economic Journal*, 715-734.

Hoover, K. D., & Perez, S. J. (1999). Data mining reconsidered: encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, 167-191.

Hoover, K. D., & Perez, S. J. (2000). Three attitudes towards data mining. *Journal of Economic Methodology*, 195-210.

Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications*, 11303-11311.

Lovell, M. C. (1983). DATA MINING. *The Review of Economics and Statistics*, 1-12.

Mayer, T. (2000). Data mining: a reconsideration. *Journal of Economic Methodology*, 183-194.

Medina, C. A., & Posso, C. M. (2011). Inmigrantes Colombianos en Estados Unidos: Educacion,clasificacion laboral y decision de retornar. (B. d. Colombia, Ed.) *Ensayos sobre Politica Economica*, Vol 29, Num 65.

Minnesota Population Center. (2017). Integrated Public Use Microdata Series, International: Version 6.5 [dataset]. Minneapolis: University of Minnesota. doi:http://doi.org/10.18128/D020.V6.5.

North, M. (2012). *Data Mining for the Masses.* Global Text Project Book.

OECD. (2015). *Data-Driven Innovation Big Data for Growth and Well-Being.* Paris: OECD Publishing. Obtenido de http://www.oecd-ilibrary.org/science-and-technology/data-driven-innovation_9789264229358-en

Padhy, N., Mishra, D. P., & Panigrahi, R. (2012). The Survey of Data Mining Applications And Feature Scope. *International Journal of Computer Science*, Vol.2, No.3.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. *Journal of Machine Learning Research*, 211-255.

Piatetsky, G. (10 de 2014). *http://www.kdnuggets.com/.* Obtenido de KDnuggets: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Sapra, S. (2014). A Useful Role for Data Mining in Economics. *Business and Economics Journal*.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 3-28.

Wikipedia. (04 de 03 de 2017). *es.wikipedia.org*. Obtenido de www.wikipedia.org: https://es.wikipedia.org/wiki/Tupla

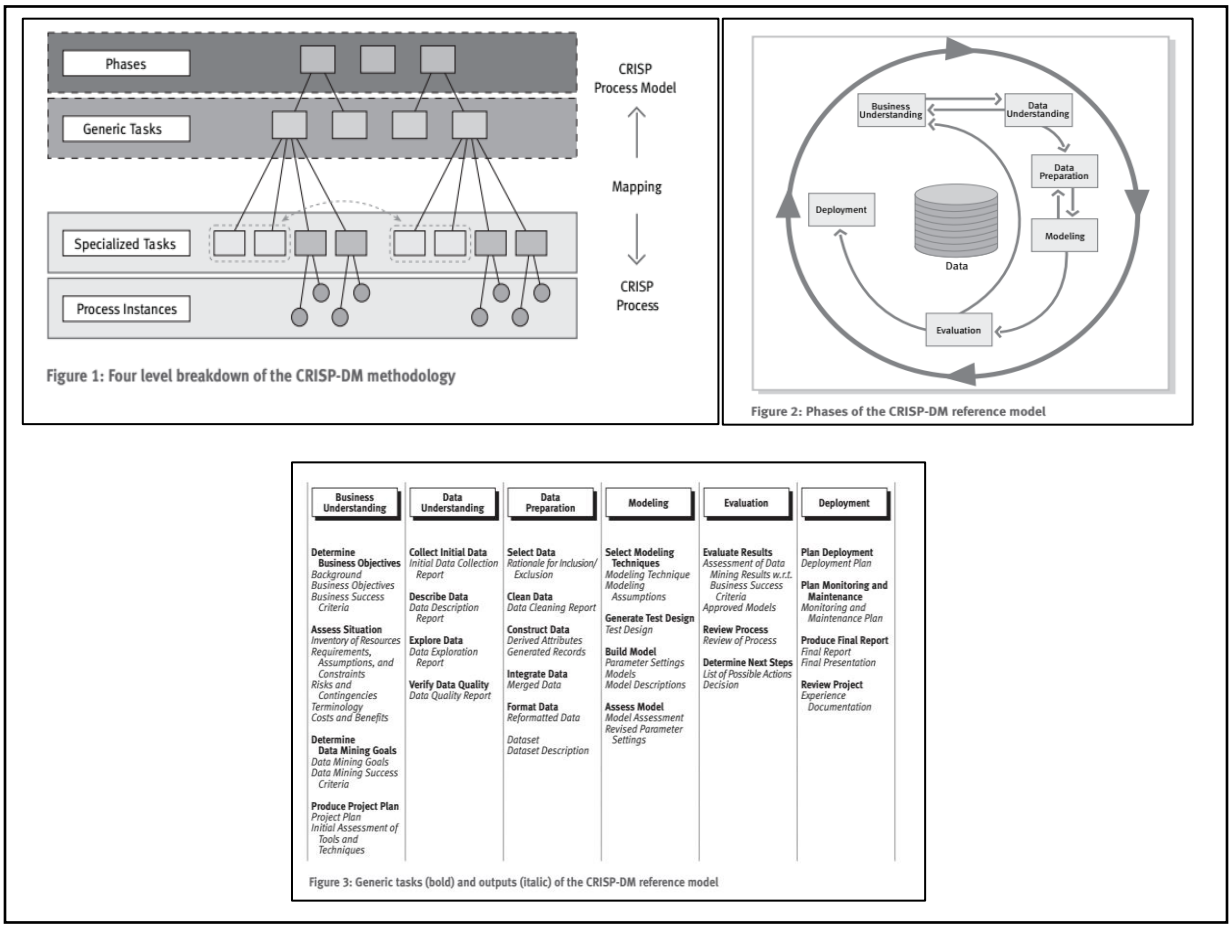
Wikipedia. (07 de 04 de 2017). *Wikipedia*. Obtenido de www.wikipedia.com: https://es.wikipedia.org/wiki/Mutatis_mutandis

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.* In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining.

7. ANEXOS

7.1. CRISP-DM:

Para una mayor comprensión de los componentes y fases se presentan los gráficos que resumen la metodología CRISP-DM según el trabajo original de ((NCR) et al, 2000):



Fuente: (NCR et al, 2000)

7.2. CODIGO Y REPRESENTACIONES:

A continuación, se presentan las representaciones graficas de los modelos estimados y se detalle el código utilizado en la estimación usando el complemento de Rmarkdown para Rstudio.

TrabajoADyRF.R

Andres Florez Llano

Mon May 15 19:36:22 2017

```
#
install.packages("foreign", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/Usuario/Documents/R/win-library/3.3'
## (as 'lib' is unspecified)

## package 'foreign' successfully unpacked and MD5 sums checked

## The downloaded binary packages are in
##  C:\Users\Usuario\AppData\Local\Temp\RtmpO6c3ya\downloaded_packages

library("dplyr")

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library("ggplot2")
library("party")

## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

## Loading required package: strucchange

## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

library("randomForest")

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine

library("foreign")
setwd("C:/Users/Usuario/Desktop/Andres Florez/Trabajo de Grado/Metodologia/Datos")
##=== Carga y Limpieza las variables===##
UScensus <- read.dta(("USCENSUS.dta"), convert.factors = TRUE, convert.dates = TRUE,missing.type = FALSE)
UScensus$age <- as.integer(UScensus$age)
UScensus$momloc <- as.factor(UScensus$momloc)
UScensus$famsize <- as.numeric(UScensus$famsize)
UScensus[] <- lapply(UScensus,function(x) if(is.factor(x)) factor(x) else x)
UScensus90 <- UScensus[UScensus$sample== "United States 1990",]
UScensus00 <- UScensus[UScensus$sample== "United States 2000",]
UScensus05 <- UScensus[UScensus$sample== "United States 2005",]
UScensusFV90 <- UScensus90[,c("year","serial","geo1_us","regnus","pernum","perwt","momloc","poploc","sploc","parrule","sprule","polymal","famsize","nchild","nchlt5","eldch","yngch","age","age2","sex","marst","birthyr","chborn","citizen","yrimm","yrsimm","yrsimm2","school","edattain","empstat","occisco","indgen","classwk","inctot","incearn","incwage","incwel","migrate5","migctry5","migyrs2","mighouse","disabled","diswork")]
UScensusFV90$momloc <- ifelse(UScensusFV90$momloc == 0,yes = "No Madre en Casa",no = "Madre en Casa" )
UScensusFV90$momloc <- as.factor(UScensusFV90$momloc)
UScensusFV90$poploc <- ifelse(UScensusFV90$poploc == 0,yes = "No Padre en Casa",no = "Padre en Casa" )
UScensusFV90$poploc <- as.factor(UScensusFV90$poploc)
```

```
UScensusFV90$sploc <- ifelse(UScensusFV90$sploc == 0,yes = "No Conyuge en
Casa",no = "Conyuge en Casa" )
UScensusFV90$sploc <- as.factor(UScensusFV90$sploc)
UScensusFV90$nchild <- as.numeric(UScensusFV90$nchild)
UScensusFV90$nchlt5 <- as.numeric(UScensusFV90$nchlt5)
UScensusFV90$eldch <- ifelse(UScensusFV90$eldch == "50 or older",yes= "Hi
jo Mayor a 50",no=ifelse(UScensusFV90$eldch == "No own child in household
",yes= "No Hijo en casa",no= "Hijo Menor a 50"))
UScensusFV90$eldch <- as.factor(UScensusFV90$eldch)
UScensusFV90$sex <- as.factor(ifelse(UScensusFV90$sex == "Male",yes = "Ho
mbre",no = "Mujer"))
UScensusFV90$marst <- as.factor(ifelse(UScensusFV90$marst == "Single/neve
r married",yes= "Soltero",no=ifelse(UScensusFV90$marst == "Married/in uni
on", yes="Casado/Union Libre",no=ifelse(UScensusFV90$marst == "Separated/
divorced/spouse absent",yes= "Separado,Divorciado",no="Viudo"))))
UScensusFV90$chborn <- as.factor(ifelse(UScensusFV90$chborn=="No children
",yes= "No hijo",no=ifelse(UScensusFV90$chborn=="1 child",yes="1 Hijo",no
=ifelse(UScensusFV90$chborn=="2 children",yes="2 Hijos",no=ifelse(UScensu
sFV90$chborn=="NIU (not in universe)",yes="No universo",no="Mas de 2"))))
)
UScensus00$yrimm <- as.numeric(levels(UScensus00$yrimm))[UScensus00$yrimm
]
UScensusFV90$edattain<- as.factor(ifelse(UScensusFV90$edattain == "Less t
han primary completed" | UScensusFV90$edattain == "Primary completed",yes
= "Menor a Secundaria",no= ifelse(UScensusFV90$edattain== "Secondary comp
leted",yes= "Secundaria",no=ifelse(UScensusFV90$edattain =="University co
mpleted", yes= "Mayor a Secundaria",no="NIU (not in universe)"))))
UScensusFV90$inctot <- as.numeric(ifelse(UScensusFV90$inctot== 9999998 |
UScensusFV90$inctot== -9999 |UScensusFV90$inctot== -2005 | UScensusFV90$i
nctot== -1862 | UScensusFV90$inctot==-200,yes= "NA",no=ifelse(UScensusFV9
0$inctot == 9999999,yes= "NIU (not in universe)",no= UScensusFV90$inctot)
))

## Warning: NAs introducidos por coerción

UScensusFV90$migrate5 <- as.factor(ifelse(UScensusFV90$migrate5=="Same ma
jor administrative unit" | UScensusFV90$migrate5=="Different major admini
strative unit",yes = "Permanecio",no = "Volvio"))
UScensusFV90$diswork <- as.factor(ifelse(UScensusFV90$diswork == "No disa
bility that affects work", yes="No discapacidad",no=ifelse(UScensusFV90$d
iswork=="Disability causes difficulty or limits work",yes="Discapacidad",
no=ifelse(UScensusFV90$diswork == "Disability prevents work",yes = "Disca
pacidad previene trabajo",no="NIU (not in universe)"))))
## Preparacion de datos (problemas de correlacion y de valores nulos o ni
veles > 5)
UScensusFV90 <- UScensusFV90[UScensusFV90$edattain != "NIU (not in univer
se)",]
UScensusFV90 <- UScensusFV90[UScensusFV90$age >= 20 & UScensusFV90$age<=8
0,]
UScensusFV90 <- UScensusFV90[UScensusFV90$classwk != "NIU (not in univers
```

```
e)",]
UScensusFV90$classwk <- factor(UScensusFV90$classwk)
UScensusFV90$edattain <- factor(UScensusFV90$edattain)
UScensusFV90$edattain <- as.factor(ifelse(UScensusFV90$edattain == "Secun
daria" | UScensusFV90$edattain == "Mayor a Secundaria",yes=">= Secundaria
",no= "< Secundaria"))
UScensusFV90 <- subset(UScensusFV90,select=c(-year,-serial,-geo1_us,-pern
um,-perwt,-parrule,-sprule,-polymal,-yngch,-age2,-birthyr,-yrsimm,-occisc
o,-indgen,-incearn,-incwage,-incwel,-migctry5,-migyrs2,-disabled,-yrimm,-
yrsimm2))
UScensusFV90$diswork <- factor(UScensusFV90$diswork)
UScensusFV90$school <- factor(UScensusFV90$school)
UScensusFV90$empstat <- factor(UScensusFV90$empstat)
## Arboles de Decision ##
set.seed(1234)
pd <- sample(2,nrow(UScensusFV90),replace=TRUE,prob=c(0.8,0.2))
train <- UScensusFV90[pd==1,]
valid <- UScensusFV90[pd==2,]
# Primer Arbol de clasificacion- variables explicativas basadas en (Medin
a y Posso)
tree <- ctree( migrate5 ~ edattain + age + sex + nchlt5,data = train)
plot(tree)

predporc <- predict(tree,valid,type = "prob")
tab <-table(predict(tree),train$migrate5) # tabla de prediccion.
tab

##              Permanecio Volvio
##   Permanecio       6573   1348
##   Volvio              0      0

1-sum(diag(tab))/sum(tab)

## [1] 0.1701805

# Segundo Arbol incluyendo todas las variables finales de la muestra.
tree1 <- ctree( migrate5 ~ .,data = train)
plot(tree1)

tab1 <- table(predict(tree1),train$migrate5)
tab1

##
##              Permanecio Volvio
##   Permanecio       6143    903
##   Volvio            430    445

1-sum(diag(tab1))/sum(tab1)

## [1] 0.1682868
```

```r
# Tercer Modelo con el Random Forest.
rf <- randomForest(migrate5 ~ .,data=train,mtry=7,ntree=200,na.action = n
a.exclude)
rf$confusion
```

```
##            Permanecio Volvio class.error
## Permanecio       6253    313  0.04766981
## Volvio           1043    305  0.77373887
```

```r
1-sum(diag(rf$confusion))/sum(rf$confusion)
```

```
## [1] 0.1714279
```

```r
importance(rf)
```

```
##          MeanDecreaseGini
## regnus          180.77715
## momloc           28.17483
## poploc           16.49521
## sploc            33.75115
## famsize         170.22907
## nchild           62.17827
## nchlt5           36.89436
## eldch            16.60304
## age             377.17240
## sex              30.15988
## marst            74.71823
## chborn           92.15181
## citizen         117.02835
## school           49.51997
## edattain         51.00849
## empstat          66.14221
## classwk          36.41578
## inctot          470.16053
## mighouse        208.13139
## diswork          17.77693
```

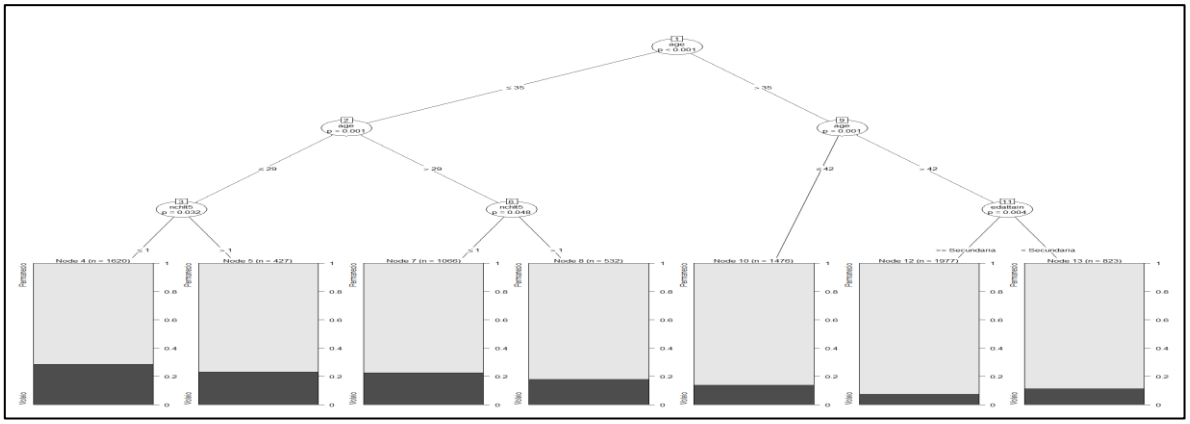
```r
varImpPlot(rf)
train <- na.exclude(train)
temp <- subset(train,select=c(-migrate5))
bestmtry <- tuneRF(temp,train$migrate5,ntreeTry = 200,stepFactor = 1.5,im
prove = 0.01)
rf <- randomForest(migrate5 ~ .,data=train,mtry=3,ntree=200,na.action = n
a.exclude)
rf$confusion
```

```
##            Permanecio Volvio class.error
## Permanecio       6476     90  0.01370698
## Volvio           1191    157  0.88353116
```

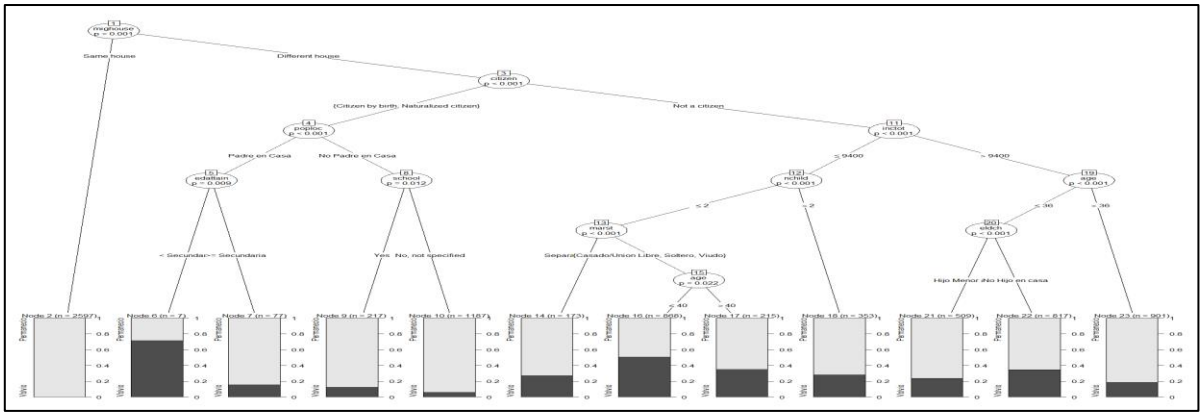
```r
1-sum(diag(rf$confusion))/sum(rf$confusion)
```

```
## [1] 0.1619601
```

Classification Tree sobre las variables propuestas por (Medina & Posso, 2011):



Classification Tree sobre el conjunto total de variables en la muestra:



Representación grafica de la importancia de variables con el método de ensamble de RandomForest: