TESIS DE GRADO

INGENIERÍA INDUSTRIAL

ESTUDIO DEL ÍNDICE DE DESARROLLO HUMANO (IDH) MEDIANTE LA APLICACIÓN DE MINERÍA DE DATOS BASADO EN SISTEMAS INTELIGENTES

AUTOR:

ALEJANDRO MARTÍN FERRARI

DIRECTORES DE TESIS:

DRA. PAOLA BRITOS

M. ING. CLAUDIO RANCAN

2009

RESUMEN

La Organización de las Naciones Unidas (ONU) a través del Programa de las Naciones Unidas para el Desarrollo (PNUD), confecciona anualmente el Informe de Desarrollo Humano. El objetivo de dicha publicación es tratar temas como el progreso social, la economía, la igualdad, la participación y libertad, la sostenibilidad y la seguridad humana.

El índice de desarrollo humano (IDH) es el principal enfoque de dicho informe anual. Este indicador tiene en cuenta 3 parámetros: vida larga y saludable, educación y nivel de vida digno. Su principal objetivo es situar a las personas en el centro del proceso de desarrollo en términos de debates económicos, formulación de políticas y promoción.

Este proyecto se basa en la información divulgada por el PNUD anualmente (a través del informe) para realizar un estudio de minería de datos. Por medio de esta herramienta y utilizando la metodología CRISP-DM se pretende realizar un estudio intensivo sobre un indicador de gran relevancia actualmente. Se realizaran distintos experimentos tendientes a la comprensión completa del comportamiento del conjunto de datos, caracterización de la información y su posterior clasificación. Se espera detectar patrones y establecer límites cuantitativos en otras variables de la base de datos en función de su incidencia en el IDH, permitiendo nuevos agrupamientos y clasificaciones para determinar cómo influyen el resto de los indicadores en el índice de desarrollo.

ABSTRACT

The United Nations (UN) through the United Nations Development Programme (UNDP) annually creates the Human Development Report. The objective of this publication is to deal with subjects such as social progress, economy, equality, freedom and human safety.

The human development index (HDI) is the main focus of that annual report. This index takes in account 3 basic dimensions: life expectancy, knowledge and education, and standard of living. Its main objective is to locate people in the center of the development process in terms of economical and political debates, and promotion.

This project is based on the information given by the UNDP annually (through the report) in order to carry out a data mining study. With this tool and using CRISP-DM methodology it intends to conduct an intensive study on the human development index. Different experiments will be made in order to understand the behavior of the database and to characterize and classify the information. It is expected to find patterns in the mass of data and to establish quantitative limits in other variables, allowing the creation of new clusters and classifications to determine the impact that the rest of the variables have on the HDI index.

TABLA DE CONTENIDOS

1. INTRODUCCIÓN

1.1. INTRODUCCIÓN AL PROBLEMA

Según la ONU, el desarrollo humano es el proceso por el que una sociedad mejora las condiciones de vida de sus ciudadanos a través de un incremento de los bienes con los que puede cubrir sus necesidades básicas y complementarias, y de la creación de un entorno en el que se respeten los derechos humanos de todos ellos.

Año a año la Organización de las Naciones Unidas (ONU) confecciona un resumen de un conjunto de indicadores por país y continente. Cuenta con datos geográficos, demográficos, sociales y económicos, entre ellos la superficie del país, densidad de población, esperanza de vida, PBI per cápita, porcentaje de población urbana, etc.

El Índice de Desarrollo Humano (IDH) es una medición por país, elaborada por el Programa de las Naciones Unidas para el Desarrollo (PNUD). Se basa en un indicador social estadístico que va desde 0 hasta 1, compuesto por tres parámetros:

1. Vida larga y saludable (medida según la esperanza de vida al nacer).
2. Educación (medida por la tasa de alfabetización de adultos y la tasa bruta combinada de matriculación en educación primaria, secundaria y superior, así como los años de duración de la educación obligatoria).
3. Nivel de vida digno (medido por el PBI per cápita).

En base al indicador se publica anualmente el ranking de desarrollo humano por país. Se considera IDH Alto cuando es mayor a 0,8; IDH Medio entre 0,5 y 0,8; e IDH Bajo menor a 0,5. En la última publicación (año 2007) Islandia supero a Noruega (quien encabezó la lista los últimos 5 años) como el país con el IDH más alto. Dentro de Latinoamérica, Argentina es el país mejor clasificado (posición 38 a nivel mundial) seguido por Chile (posición 40) y Guatemala es el peor (posición 118). El índice arrojo que los 22 países con IDH Bajo se encuentran todos en el área subsahariana de África, y Sierra Leona es el peor. En estos países, 2 de cada 5 niños no alcanzarán a cumplir 40 años. Cabe destacar que algunos países no se encuentran en el ranking por no contar con datos precisos para la medición.

El objetivo de este proyecto es realizar un análisis intensivo sobre la información aportada por esta base de datos de la ONU. Para esto se realizara un estudio de minería de datos mediante la aplicación de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). La misma fue diseñada para desarrollar proyectos de minería de datos, se describe en términos de un modelo de procesos jerárquico y consta de cuatro niveles de abstracción que van desde lo general a lo mas especifico. Si bien la metodología se desarrollo para llevar adelante grandes proyectos en compañías y empresas, es suficientemente amplia y flexible para poder ser aplicada a proyectos de todo tamaño. Se realizaran distintos experimentos tendientes a la comprensión completa del comportamiento del conjunto de datos, caracterización de la información y su posterior clasificación. Se espera detectar

patrones y establecer límites cuantitativos en otras variables de la base de datos en función de su incidencia en el IDH, permitiendo nuevos agrupamientos y clasificaciones para determinar cómo influyen el resto de los indicadores en el índice de desarrollo.

1.2. DESTINATARIOS

El presente trabajo se encuentra dirigido a las personas involucradas en el desarrollo y confección del informe de desarrollo humano y también a los interesados en conocer la utilidad de la minería de datos y los distintos ámbitos en que puede ser aplicada.

1.3. COMPOSICIÓN DEL PROYECTO

El presente trabajo se encuentra estructurado de la siguiente manera:

Capitulo 1: Introducción

Se da una breve descripción de las razones que llevaron a la realización del proyecto, los destinatarios del mismo y su organización.

Capitulo 2: Estado de la cuestión

Se refleja la situación actual de los desarrollos y estudios realizados sobre los temas que aborda este trabajo. También se realiza una introducción a la minería de datos, la metodología elegida y otros fundamentos teóricos.

Capitulo 3: Descripción del problema

Se presenta el problema que se quiere resolver y los principales motivos de su elección.

Capitulo 4: Solución

Se propone una solución al problema planteado a través del desarrollo de la metodología CRISP-DM.

Capitulo 5: Conclusiones

Se presentan las principales conclusiones del proyecto.

Capitulo 6: Futuras líneas de investigación

Presenta las cuestiones abiertas que desencadena el proyecto posibilitando futuros trabajos sobre las mismas.

Capitulo 7: Referencias

Se enumera la bibliografía utilizada durante la elaboración del proyecto.

Capítulo 9: Anexos

2. ESTADO DE LA CUESTIÓN

En el presente capítulo se desarrollan los temas sobre los que hace foco este proyecto. Se explican los conceptos necesarios para entender el análisis realizado, como el desarrollo humano e IDH, también las herramientas que se utilizan, como la minería de datos, los algoritmos y el software seleccionados, y finalmente se describe la metodología de resolución de problemas que se implementa a lo largo del mismo (CRISP-DM).

2.1. DESARROLLO HUMANO

El Desarrollo Humano comprende la creación de un entorno en el que las personas puedan desarrollar su máximo potencial y llevar adelante una vida productiva y creativa de acuerdo con sus necesidades e intereses. Este enfoque plantea a las personas como la verdadera riqueza de las naciones. Por lo tanto, el desarrollo implica ampliar las oportunidades para que cada persona pueda vivir una vida que valore. Se entiende como la adquisición por parte de los individuos, comunidades e instituciones, de la capacidad de participar efectivamente en la construcción de una civilización mundial que es próspera tanto material como espiritualmente.

Para que existan más oportunidades lo fundamental es desarrollar las capacidades humanas: la diversidad de cosas que las personas pueden hacer o ser en la vida. Las capacidades más esenciales para el desarrollo humano son disfrutar de una vida larga y saludable, haber sido educado, acceder a los recursos necesarios para lograr un nivel de vida digno y poder participar en la vida de la comunidad. Sin estas capacidades, se limita considerablemente la variedad de opciones disponibles y muchas oportunidades en la vida permanecen inaccesibles.

Según Mahbud ul Haq, creador del informe sobre Desarrollo Humano: “El objetivo principal del desarrollo es ampliar las opciones de las personas. En principio, estas opciones pueden ser infinitas y cambiar con el tiempo. A menudo las personas valoran los logros que no se reflejan en forma inmediata, en las cifras de crecimiento o ingresos: mayor acceso al conocimiento, mejores servicios de nutrición y salud, medios de vida más seguros, protección contra el crimen y la violencia física, una adecuada cantidad de tiempo libre, libertades políticas y culturales y un sentido de participación en las actividades comunitarias. El objetivo del desarrollo es crear un ambiente propicio para que la gente disfrute de una vida larga, saludable y creativa”.

El primer objetivo es la libertad del ser humano. Una libertad que es fundamental para desarrollar las capacidades y ejercer los derechos. Las personas deben ser libres para hacer uso de sus alternativas y participar en la toma de decisiones que afectan sus vidas. El desarrollo humano y los derechos humanos se reafirman mutuamente y ayudan a garantizar el bienestar y la dignidad de todas las personas, forjar el respeto propio y el respeto por los demás.

El enfoque de desarrollo humano nació cuando se reconoció la necesidad de un modelo de desarrollo alternativo al existente en la década de 1980 por varias razones, entre las que se incluyen:

- La existencia de evidencia cada vez mayor en contra del convencimiento generalizado, en ese momento, sobre el poder del efecto de goteo de las fuerzas del mercado para propagar los beneficios económicos y erradicar la pobreza.
- Los costos humanos de los Programas de Ajuste Estructural se tornaron más evidentes.
- Las enfermedades sociales (el delito, el debilitamiento del tejido social, el SIDA, la contaminación, etc.) continuaban diseminándose aun frente a un crecimiento económico sólido y sistemático.
- Una ola de democratización a principios de los noventa aumentó las esperanzas en torno a la creación de modelos centrados en las personas.

A partir de 1990, el concepto de desarrollo humano se aplicó a un estudio sistemático de temas mundiales, según se publicó en los Informes anuales sobre Desarrollo Humano patrocinados por el PNUD. Se fundaron las bases conceptuales de un enfoque alternativo y más amplio del desarrollo humano, definido como el proceso de ampliación de las opciones de las personas y mejora de las capacidades humanas y las libertades, para que las personas puedan vivir una vida larga y saludable, tener acceso a la educación y a un nivel de vida digno, y participar en la vida de su comunidad y en las decisiones que afecten sus vidas.

El desarrollo humano siempre ha sido flexible y ha tenido un “final abierto” con respecto a definiciones más específicas. Pueden existir tantas dimensiones del desarrollo humano como modos de ampliar las opciones de las personas. Los parámetros que son claves o prioritarios para el desarrollo humano pueden evolucionar con el tiempo y variar entre los diferentes países y dentro de cada uno de ellos.

Algunos de los temas y asuntos que se consideran de mayor importancia para el desarrollo humano en la actualidad son:

- **El progreso social**: mayor acceso a la educación, mejores servicios de nutrición y salud.
- **La economía**: la importancia del crecimiento económico como medio para reducir las desigualdades y mejorar los niveles de desarrollo humano.
- **La eficiencia** en términos de uso y disponibilidad de los recursos. El desarrollo humano propicia el crecimiento y la productividad, siempre y cuando este crecimiento beneficie de manera directa a las personas pobres, las mujeres y otros grupos marginados.

- **La igualdad** en cuanto al crecimiento económico y otros parámetros del desarrollo humano.
- **La participación y la libertad** , en especial mediante el empoderamiento, la gobernabilidad democrática, la igualdad de géneros, los derechos civiles y políticos y la libertad cultural, particularmente en los grupos marginales definidos por parámetros tales como urbanos/rurales, sexo, edad, religión, origen étnico, parámetros físicos y mentales, etc.
- **La sostenibilidad** para las generaciones futuras, en términos ecológicos, económicos y sociales.
- **La seguridad humana** ante amenazas crónicas de la vida cotidiana tales como el hambre y las discontinuidades repentinas como la desocupación, la hambruna, los conflictos, etc.

Además de establecer principios y asuntos prioritarios, se busca lograr resultados, para lo cual se definen objetivos medibles a cumplir. Durante la cumbre del Milenio del año 2000, 189 países se comprometieron a crear, a nivel nacional y mundial, un entorno propicio para el desarrollo y la eliminación de la pobreza, y así alcanzar objetivos con sus metas específicas para el 2015. Los objetivos planteados son:

1) Erradicar el hambre y la pobreza extrema.
   Meta: Reducir a la mitad, entre 1990 y 2015, el porcentaje de personas cuyos ingresos sean inferiores a un dólar por día.

2) Lograr la educación básica universal.
   Meta: Velar para que en 2015, los niños y niñas de todo el mundo puedan terminar un ciclo completo de enseñanza primaria.

3) Promover la equidad de género y la autonomía de la mujer.
   Meta: Eliminar las desigualdades entre géneros en la enseñanza primaria y secundaria, preferiblemente en 2005, y en todos los niveles de enseñanza antes de finales de 2015.

4) Reducir la mortalidad infantil.
   Meta: Reducir en dos terceras partes, entre 1990 y 2015 la tasa de mortalidad de los niños menores de cinco años.

5) Mejorar la salud sexual y reproductiva.
   Meta: Reducir entre 1990 y 2015, la tasa de mortalidad materna en tres cuartas partes.

6) Combatir el SIDA, la malaria y el dengue.
   Metas: Detener y comenzar a reducir, para 2015, la propagación del VIH/SIDA. Detener y reducir, para 2015, la incidencia de paludismo y otras enfermedades graves.

7) Garantizar la sostenibilidad ambiental.
   Metas: Incorporar los principios del Desarrollo sostenible a las políticas y los programas nacionales. Reducir a la mitad, para 2015, el porcentaje de personas que carezcan de acceso sostenible a agua potable.

8) Fomentar una asociación mundial para el desarrollo.

2.1.1. Programa de las Naciones Unidas para el Desarrollo (PNUD)

El Programa de las Naciones Unidas para el Desarrollo (PNUD) fue creado en 1965, pertenece al sistema de Naciones Unidas y su función es contribuir a la mejora de la calidad de vida de las naciones. Promueve el cambio y conecta a los conocimientos, la experiencia y los recursos necesarios para ayudar a los pueblos a forjar una vida mejor. Está presente en 166 países.

En la Cumbre del Milenio de las Naciones Unidas, celebrada en el año 2000, los líderes del mundo asignaron al desarrollo un papel central dentro del programa mundial mediante los Objetivos de Desarrollo del Milenio. El PNUD utiliza su red mundial para ayudar al sistema de las Naciones Unidas y a sus asociados a despertar una mayor conciencia y verificar los progresos realizados.

Se concentra en ayudar a los países a elaborar y compartir soluciones para los desafíos que plantean las cuestiones siguientes:

- Gobernabilidad democrática: cada día más países tratan de establecer un sistema de gobierno democrático. Su desafío es desarrollar instituciones y procesos que respondan mejor a las necesidades de las personas corrientes, incluyendo a los pobres. El PNUD reúne a las personas dentro de las naciones y en todo el mundo, establece asociaciones y comparte los modos de promover la participación, la responsabilidad y la eficacia en todos los ámbitos. Ayuda a los países a fortalecer sus sistemas electorales y legislativos, a mejorar el acceso a la justicia y a la administración pública y a desarrollar una mayor capacidad para ofrecer los servicios básicos a los que más los necesitan.
  La gobernabilidad democrática es esencial para alcanzar los Objetivos de Desarrollo del Milenio ya que ofrece el ambiente propicio para que ellos se cumplan y, en particular, para eliminar la pobreza. Los servicios esenciales del PNUD para apoyar los procesos nacionales de transiciones democráticas se centran en:
  - Dar asesoría en materia de políticas y apoyo técnico.
  - Fortalecer la capacidad de las instituciones y de las personas.
  - Promover las comunicaciones y la información pública.
  - Promover y mediar en el diálogo.
  - Trabajar con las redes de conocimientos y compartir las buenas prácticas.

- Reducción de la pobreza: con los Objetivos de Desarrollo del Milenio, el mundo está enfrentando las muchas dimensiones del desarrollo humano y entre ellas, la reducción a la mitad de la cantidad de personas que viven en extrema pobreza para el 2015. Los países en desarrollo están trabajando para crear sus propias estrategias nacionales de erradicación de la pobreza basándose en las necesidades y prioridades locales.
  El PNUD promueve estas soluciones nacionales y ayuda a hacerlas efectivas aumentando el acceso a los bienes productivos y a las oportunidades económicas y relacionando los programas en materia de pobreza con las políticas económicas y

financieras internacionales de los países. Al mismo tiempo, el PNUD contribuye con iniciativas que conllevan a la reforma del comercio, al alivio de la deuda y a la orientación de la inversión para dar un mejor apoyo a la reducción nacional de la pobreza y hacer que la globalización beneficie a los pobres. Para ello, el PNUD patrocina proyectos piloto innovadores, contacta a los países con las mejores prácticas y recursos mundiales, promueve el papel de la mujer en el desarrollo y reúne a los gobiernos, a la sociedad civil y a las fuentes externas que ofrecen financiación para coordinar sus esfuerzos.

- Prevención y recuperación de las crisis: a través de su red mundial, el PNUD busca y comparte enfoques innovadores para la prevención de crisis, la alerta temprana y la resolución de conflictos (desastres naturales o conflictos violentos). Las catástrofes naturales que están aumentando tanto en cantidad como en intensidad tienen consecuencias desproporcionadas en los países pobres que no tienen, por lo general, los recursos indicados para mantener acciones de prevención y de atenuación.
  Las acciones llevadas a cabo por el PNUD en los sectores de prevención de crisis y recuperación se benefician de la experiencia de la organización en los campos de intervención correspondientes, incluyendo al apoyo a la gobernabilidad democrática y a la reducción de la pobreza. El PNUD tiene un papel decisivo para ayudar a los países a entrar en una etapa orientada hacia el desarrollo, restableciendo el respeto a la ley y la buena gobernabilidad, desmovilizando a los soldados, reduciendo la cantidad de armas de pequeño calibre, apoyando la lucha contra las minas y ofreciendo medios de subsistencia alternativos para las poblaciones afectadas por la guerra.

- Energía y medio ambiente: estos dos factores son indispensables para el desarrollo sostenible. Los pobres se ven afectados en forma desproporcionada por el deterioro del medio ambiente y la falta de acceso a servicios energéticos limpios y asequibles. Estos problemas son también mundiales, puesto que los cambios climáticos, la disminución de la diversidad biológica y el agotamiento de la capa de ozono no pueden ser resueltos por las naciones individualmente. El PNUD ayuda a los países a fortalecer su capacidad para hacer frente a estos retos en los planos mundial, nacional y comunitario, buscando y compartiendo las mejores prácticas, prestando asesoramiento innovador en materia de políticas y vinculando a los asociados mediante proyectos piloto que ayudan a los pobres a encontrar medios de vida sostenibles.
  Se busca desarrollar la capacidad de los países de administrar el medio ambiente y los recursos naturales, de integrar las dimensiones de medio ambiente y energía en las estrategias de reducción de la pobreza y en los marcos nacionales de desarrollo, y de fortalecer el papel de las comunidades y de las mujeres para promover el desarrollo sostenible. El trabajo del PNUD sobre energía y medio ambiente se centra en seis áreas prioritarias:
  - Marcos y estrategias para un desarrollo sostenible.
  - Gobernabilidad eficaz del agua.

  - Acceso a servicios energéticos sostenibles.
  - Gestión sostenible de la tierra para combatir la desertificación y la degradación de la tierra.
  - Conservación y uso sostenible de la biodiversidad.
  - Política nacional y sectorial y planificación de control de emisiones de ODS y de POP

- SIDA: Para impedir la propagación del HIV y reducir su impacto, los países en desarrollo deben movilizar al gobierno y la sociedad civil en todos los niveles. El PNUD promueve la inclusión prioritaria del HIV/SIDA en la planificación y los presupuestos nacionales; ayuda a consolidar la capacidad nacional para la gestión de iniciativas que incluyan a personas e instituciones que no suelen participar en el sector de la salud pública; fomenta respuestas descentralizadas en apoyo de acciones de nivel comunitario. Dado que el VIH/SIDA es un problema de carácter mundial, el PNUD apoya estos esfuerzos nacionales mediante el aporte de conocimientos, recursos y las mejores prácticas de todo el mundo. El SIDA afecta a las personas en sus etapas de vida más productivas y es particularmente devastador en su manera de incrementar los niveles de pobreza y en como revierte los progresos en materia de desarrollo humano. Con el fin de ayudar a los países a mitigar este impacto sobre el desarrollo humano, se promueven respuestas multi-sectoriales que integran el VIH/SIDA en los planes de desarrollo nacional, programas sectoriales y planes descentralizados. El PNUD apoya la generación de comercio y la legislación en materia de salud y propiedad intelectual en los países para que tengan acceso sostenible a medicamentos de calidad a bajos costos para este virus.

En cada una de estas esferas temáticas, el PNUD propugna la protección de los derechos humanos y especialmente la potenciación de la mujer. Mediante la red mundial, trata de identificar y difundir medios de promover la igualdad de género como una dimensión esencial de asegurar la participación y la responsabilidad política.

El PNUD también realiza una amplia labor de promoción. El Informe sobre Desarrollo Humano anual, encargado por el PNUD, centra el debate mundial sobre cuestiones clave de desarrollo, proporcionando nuevos instrumentos de medición, análisis innovadores y, a menudo, propuestas de política controvertidas.

2.1.2. Índice de Desarrollo Humano (IDH)

El Informe sobre Desarrollo Humano (IDH) fue presentado por primera vez en 1990, con el objetivo único de situar a las personas en el centro del proceso de desarrollo en términos de debates económicos, formulación de políticas y promoción. Este objetivo presenta implicaciones de gran alcance, lograr el desarrollo de las personas y subrayar que la finalidad del desarrollo son las opciones y libertades.

Se trata de un informe independiente que se elabora bajo el mandato del Programa de las Naciones Unidas para el Desarrollo (PNUD) y es el resultado del trabajo de un equipo de académicos destacados, profesionales del desarrollo y miembros de la Oficina encargada del Informe sobre Desarrollo Humano del PNUD. La medición se realiza por país y se basa en un indicador social estadístico compuesto por tres parámetros:

- Vida larga y saludable (medida según la esperanza de vida al nacer).
- Educación (medida por la tasa de alfabetización de adultos y la tasa bruta combinada de matriculación en educación primaria, secundaria y superior, así como los años de duración de la educación obligatoria).
- Nivel de vida digno (medido por el PIB per cápita en dólares).

El PNUD realiza una clasificación de los países teniendo en cuenta su IDH. Se presentan tres grupos en función de los siguientes parámetros:

- Países con desarrollo humano alto: IDH ≥ 0,8 = 70 países.
- Países con desarrollo humano medio: 0,5 ≤ IDH < 0,8 = 85 países.
- Países con desarrollo humano bajo: IDH < 0,5 = 22 países.

El IDH es un índice que surge de las 3 variables mencionadas previamente. Para crear este índice, se establecen valores mínimos y máximos para cada una de las variables (esperanza de vida, educación y PBI) de manera de lograr un índice entre 0 y 1 para cada variable. Finalmente, se calcula el IDH promediando los 3 índices, es decir, se asigna el mismo peso a los 3 índices en el resultado final (33%).

Los valores máximos y mínimos asignados a cada variable son:

- Esperanza de vida: 85 y 25 años
- Educación (ambas tasas): 100% y 0%
- PBI per cápita: 40000 y 100 US$

La fórmula utilizada para calcular el índice de cada variable, expresado entre 0 y 1 es la siguiente:

$$Índice\ de\ la\ variable = \frac{valor - mínimo}{máximo - mínimo}$$

A modo de ejemplo se calculan los índices y el IDH para Italia. Según el informe utilizado, la esperanza de vida al nacer es de 80.3 años, la tasa de alfabetización 98.4%, la tasa de matriculación 90.6% y el PBI per cápita 28529 dólares.

- Cálculo del índice de esperanza de vida (IEV):

$$IEV = \frac{80.3 - 25}{85 - 25} = 0.922$$

- Cálculo del índice de educación (IE): en este caso se pondera con dos tercios al índice de alfabetización (IA) y un tercio índice de matriculación (IM).

$$IA = \frac{98.4 - 0}{100 - 0} = 0.984$$

$$IM = \frac{90.6 - 0}{100 - 0} = 0.906$$

$$IE = \frac{2}{3} IA + \frac{1}{3} IM$$

$$IE = \frac{2}{3} 0.984 + \frac{1}{3} 0.906 = 0.958$$

- Cálculo del índice de PBI (IPIB): en este caso se utiliza el logaritmo para obtener un valor entre 0 y 1.

$$IPBI = \frac{\log(28529) - \log(100)}{\log(40000) - \log(100)} = 0.944$$

- Cálculo del IDH:

$$IDH = \frac{1}{3} IEV + \frac{1}{3} IE + \frac{1}{3} IPBI$$

$$IDH = \frac{1}{3} 0.922 + \frac{1}{3} 0.958 + \frac{1}{3} 0.944 = 0.941$$

En el informe que se utiliza para realizar este proyecto, publicado en 2008, existen 177 países. El de mayor IDH es Islandia con 0.968 y el menor es Sierra Leona con 0.336. En América del Sur, el mayor IDH corresponde a Argentina con 0.869, seguido por Chile con 0.867 y el menor a Bolivia con 0.695.

Tras la creación del primer informe, se han desarrollado cuatro nuevos índices compuestos de desarrollo humano: el índice de desarrollo humano; el índice de desarrollo humano orientado a un género; el índice de empoderamiento de la mujer; y el índice de pobreza humana. Cada uno de los informes se concentra también en un tema muy específico del debate actual sobre

el desarrollo y proporciona análisis de vanguardia y recomendaciones en materia de política. Los mensajes de los informes — y las herramientas para implementarlos — han sido adoptados por pueblos de diversas partes del mundo, como se puede comprobar por la publicación de informes sobre desarrollo humano a nivel nacional en más de 140 países y su traducción a más de 12 idiomas.

Existen a su vez informes sobre desarrollo humano regionales, nacionales y subnacionales que abordan la temática del desarrollo humano desde una perspectiva regional y nacional, y son elaborados e impulsados por equipos regionales y nacionales. Estos equipos aportan datos y análisis al Informe Mundial, al mismo tiempo que se nutren de ellos. Hasta el momento se han elaborado más de 600 informes regionales y subnacionales, en más de 140 países.

Los informes nacionales sitúan el desarrollo humano en el primer plano de la agenda política nacional. Son herramientas de análisis político que reflejan las prioridades de la gente, fortalecen las capacidades de los países, generan el compromiso de colaboradores nacionales, identifican desigualdades y miden el progreso. Como instrumentos de medición del progreso humano y como desencadenantes de acciones para el cambio, los informes regionales promueven alianzas regionales para influenciar el cambio y abordar cuestiones relacionadas con los derechos humanos, la pobreza, la educación, la reforma económica, el VIH/SIDA y la globalización, desde la mirada propia de cada región.

Por su carácter de herramientas de promoción diseñadas para atraer a un público vasto, los Informes de Desarrollo Humano pueden suscitar debates públicos y fomentar iniciativas de apoyo para la acción y el cambio. Asimismo, han colaborado en la integración de las percepciones y prioridades de la gente y han servido como fuente de opinión alternativa en materia de políticas, para la planificación del desarrollo a través de diversas temáticas.

2.2. INTRODUCCIÓN A LA MINERÍA DE DATOS

El creciente volumen de datos en todas las áreas de aplicación humana demanda nuevas y poderosas técnicas de transformación de los datos en conocimiento útil. Se puede decir que la minería de datos es un avance en dicha cuestión. Busca generar información similar a la que podría producir un experto humano, que además satisfaga el Principio de Comprensibilidad (utilizar lenguaje adecuado a la temática del trabajo) [Britos et al, 2005].

La minería de datos (data mining) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación.

En cuanto a su aplicación, las herramientas de Data Mining predicen futuras tendencias y comportamientos, permitiendo en los negocios tomar decisiones proactivas y conducidas por un conocimiento acabado de la información (knowledge-driven). Los análisis prospectivos automatizados ofrecidos por un producto así van más allá de los eventos pasados provistos por herramientas retrospectivas típicas de sistemas de soporte de decisión. Las herramientas de Data Mining pueden responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para poder ser resueltas y que los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus expectativas.

Un proceso típico de minería de datos consta de los siguientes pasos generales (figura 2.1):

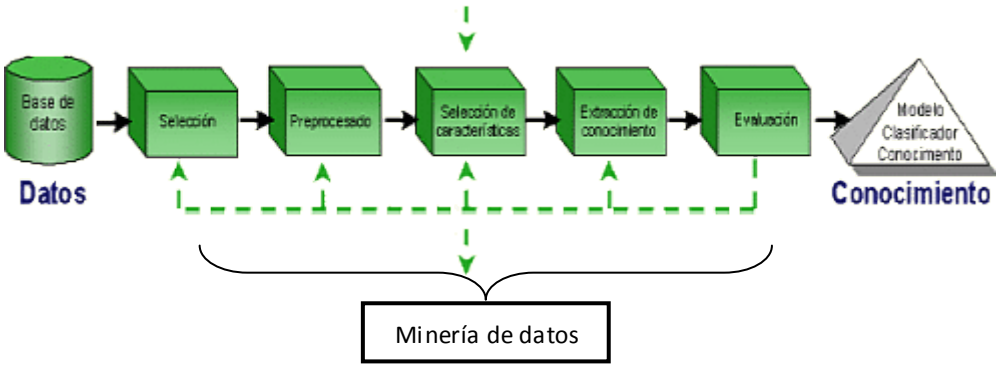


Figura 2.1 Pasos de un proceso típico de data mining.

1) Selección del conjunto de datos, tanto en lo que se refiere a las variables dependientes, como a las variables objetivo, como posiblemente al muestreo de los registros disponibles.
2) Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

3) Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.
4) Seleccionar y aplicar la técnica de minería de datos, se construye el modelo predictivo, de clasificación o segmentación.
5) Evaluar los resultados contrastándolos con un conjunto de datos previamente reservado para validar la generalidad del modelo.

2.2.1. Algoritmos de minería de datos

Las técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos (más o menos sofisticados) que se aplican sobre un conjunto de datos para obtener resultados.

El algoritmo de minería de datos es el mecanismo que crea un modelo de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos y luego busca patrones y tendencias específicos. El algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería de datos. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas.

El modelo de minería de datos que crea un algoritmo puede tomar diversas formas, incluyendo:

- Un conjunto de reglas que describen cómo se agrupan los productos en una transacción.
- Un árbol de decisión que predice si un cliente determinado comprará un producto.
- Un mode lo matemático que predice las ventas.
- Un conjunto de clústeres que describe cómo se relacionan los casos de un conjunto de datos.

Los algoritmos se dividen en dos categorías principales:

- Supervisados o predictivos: predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida. Ejemplos de esta categoría son: arboles de decisión, inducción neuronal, series temporales, etc.
- No supervisados o de descubrimiento del conocimiento: con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas. Ejemplos de esta categoría son: clustering, patrones secuenciales, detección de desvíos, etc.

2.2.1.1. Algoritmo de clasificación

Los algoritmos de clasificación utilizan un procedimiento por el cual ítems individuales son puestos en grupos basándose en información cuantitativa, una o más características inherentes a los ítems y también basándose en un set de entrenamiento/aprendizaje de ítems previamente etiquetados.

El problema en términos formales puede describirse de la siguiente manera:

Dada cierta información de entrenamiento/aprendizaje $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ se produce un clasificador $h: X \to Y$ que mapea cualquier objeto $x \in X$ hacia su etiqueta verdadera $y \in Y$ definida por una función desconocida $g: X \to Y$. Por ejemplo, si el problema es filtrar el correo spam (correo "basura"), x_i es alguna representación de un email e y es "Spam" o "No Spam".

Estos algoritmos predicen una o más variables discretas, basándose en otros atributos del conjunto de datos. El método se conoce como supervisado debido a que, para el conjunto de entrenamiento, se conoce la clase a la cual pertenece cada ítem del mismo y se le indica al modelo si la clasificación que realiza es correcta que predicen una o más variables discretas, basándose en otros atributos del conjunto de datos.

De los distintos algoritmos de clasificación existentes, en este proyecto se utilizan en particular los árboles decisión. Se trata de un modelo de predicción utilizado en el ámbito de la inteligencia artificial. Dada una base de datos se construyen diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Un árbol de decisión tiene entradas, que pueden ser un objeto o una situación descripta por medio de un conjunto de atributos, y a partir de esto genera una salida. Los valores que pueden tomar las entradas y salidas pueden ser discretos o continuos. Los valores discretos son más utilizados por simplicidad, cuando se utilizan valores discretos en las funciones de una aplicación se denomina clasificación y cuando se utilizan los continuos se denomina regresión.

Un árbol de decisión lleva a cabo un test a medida que este se recorre hacia las hojas para alcanzar así una decisión. El árbol de decisión suele contener nodos internos, nodos de probabilidad, nodos hojas y arcos. Un nodo interno contiene un test sobre algún valor de una de las propiedades. Un nodo de probabilidad indica que debe ocurrir un evento aleatorio de acuerdo a la naturaleza del problema. Un nodo hoja representa el valor que devolverá el árbol de decisión y finalmente las ramas brindan los posibles caminos que se tienen de acuerdo a la decisión tomada.

En particular, se utiliza el algoritmo de clasificación C4.5. Fue desarrollado por Ross Quinlan en 1993, es una extensión del ID3 (otro algoritmo de clasificación), que acaba con muchas de sus limitaciones. Permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos A<=N y otra para A>N. Además, los árboles son menos frondosos porque cada hoja no cubre una clase en particular sino una distribución de clases, lo cual los hace menos profundo. El C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente, según la estrategia de profundidad-primero (depth-first) [Servente &García Martínez, 2002]. Antes de cada partición de datos, el algoritmo considera todas las pruebas posibles que pueden dividir el

conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información o en la mayor proporción de ganancia de información.

Las principales ventajas del C4.5 en comparación al ID3 son:

- Evitar sobreajuste de los datos.
- Determinar que tan profundo debe crecer el árbol de decisión.
- Reducir errores en la poda.
- Condicionar la Post-Poda.
- Manejar atributos continuos.
- Escoger un rango de medida apropiado.
- Manejar datos de entrenamiento con valores faltantes.
- Manejar de atributos con diferentes valores.
- Mejorar la eficiencia computacional.

2.2.1.2. Algoritmo de clustering

Los algoritmos de clustering agrupan objetos en grupos (llamados clústeres) de modo tal que los objetos dentro de un mismo clúster sean más similares entre sí que a los objetos de otros clústeres. Este algoritmo es no supervisado ya que no se conoce de antemano las distintas clases del conjunto de datos de entrenamiento. El proceso identifica áreas densamente pobladas de acuerdo a alguna medida de distancia, en un conjunto de datos multidimensional. Se basa en maximizar la similitud de las instancias en cada clúster y minimizar la similitud entre clústeres.

En función de la forma en que realizan la partición de los datos, los principales tipos de clustering son:

- Jerárquico: encuentra clústeres sucesivos usando los clústeres que estableció previamente.
- Particional: encuentra todos los clústeres de una sola vez y puede utilizarse como un algoritmo de división al utilizar clustering jerárquico.
- Basados en densidad: se utilizan para descubrir clústeres con forma arbitraria. El clúster es una región en la cual la densidad de información de los ítems supera un cierto umbral.
- Bi-Clustering o Co-Clustering: son métodos en los cuales no solo los ítems son clusterizados sino también las características de los ítems.

Para el proyecto, se selecciona un método particional llamado K-Means. Su forma estándar fue propuesta por primera vez en 1957 por Stuart Lloyd aunque recién fue publicada en 1982.

Es un método de análisis que apunta a particionar *n* observaciones en *k* clústeres de manera que cada observación pertenezca al clúster con media más cercana. Dado un set de observaciones $(x_1, x_2, \ldots, x_n)$ donde cada observación es un vector d-dimensional, el

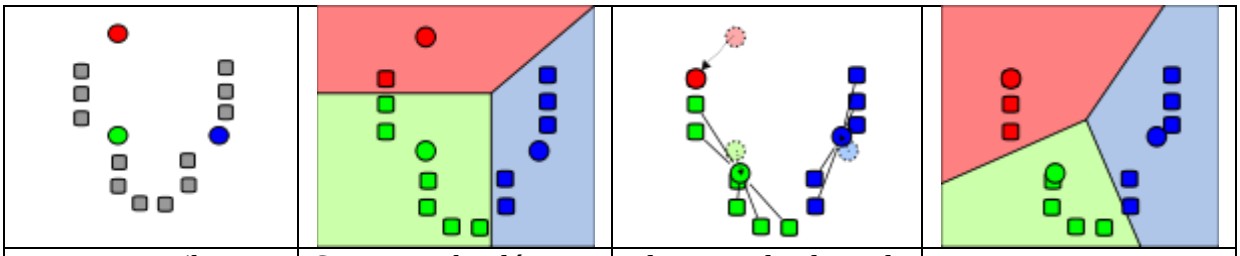
algoritmo particiona el set en k clústeres $(k < n), S = \{S_1, S_2, \ldots, S_k\}$ de manera de minimizar la suma de cuadrados dentro del clúster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

Donde μ_i es la media de S_i.

Los pasos del algoritmo son (figuras 2.2 a 2.5):

1. Elegir el numero de clústeres, *k*.
2. Generar *k* clústeres de forma aleatoria y determinar sus centros o directamente generar *k* puntos al azar como los centros de los clústeres.
3. Asignar cada ítem al clúster con centro más cercano.
4. Recalcular el centro de los clústeres.
5. Repetir los pasos 3 y 4 hasta que se alcance el criterio de convergencia seleccionado.



Los centros (*k=3* en este caso) son seleccionados al azar del conjunto de datos del dataset.	Se crean *k* clústeres asociando cada ítem con su centro más próximo.	El centroide de cada uno de los *k* clústeres se convierte en el nuevo centro.	Los 2 pasos anteriores son repetidos hasta alcanzar la convergencia.

Figuras 2.2 a 2.5 Pasos del algoritmo K-Means

La desventaja de este algoritmo es que no siempre se consigue el mismo resultado con distintas corridas, ya que depende de las semillas aleatorias iniciales. Por otra parte, minimiza la varianza intra-clúster pero no asegura que el resultado tenga la menor varianza global. Sus principales ventajas son su simplicidad y velocidad que permiten correrlo en datasets grandes.

2.2.2. Herramienta de minería de datos

A la hora de realizar un proyecto de minería de datos es crucial la selección de una herramienta de software que resulte apropiada. Existen múltiples cuestiones a considerar, algunas de ellas son: costo de la aplicación, actualizaciones y revisiones frecuentes, facilidad de uso, existencia de tutoriales o archivos de ayuda, tipo de salidas a obtener, disponibilidad de los algoritmos necesarios, etc. A su vez existen otros factores que inciden en la decisión, como es el conocer o haber utilizado la herramienta, lo que permite ahorrar gran cantidad de tiempo; también se debe estar seguro que las salidas obtenidas tienen el formato y la calidad esperada.

En este proyecto en particular, se utilizará una herramienta denominada Tanagra. Esto se debe principalmente a la recomendación del experto en minería de datos, quien consideró que resultaba apta para los objetivos planteados en el proyecto. Se trata de una aplicación gratuita y de uso libre que cuenta con una gran cantidad de algoritmos disponibles y en particular con aquellos seleccionados en la etapa previa. A su vez, es una herramienta relativamente nueva y cuenta con actualizaciones periódicas lo que asegura su correcto funcionamiento (la última versión fue publicada el 15 de abril de 2009).

Las principales ventajas del software seleccionado son:

- Software gratuito, apuntado al ámbito académico y de investigación.
- "Open Source Project", lo que permite que las personas especializadas puedan agregar sus propios algoritmos para resolver cuestiones específicas y mejorar el programa.
- Conforme a las normas actuales de programas de minería de datos, lo que hace que su interfaz gráfica y modo de uso sea conocido por quienes utilizan otros programas similares.
- Bajos requerimientos y utilización de la computadora por parte del programa (los requerimientos de hardware se ven afectados por el tamaño de la base de datos).

A su vez, posee ciertas desventajas a la hora de compararlo con otras herramientas:

- Falta de un tutorial completo (solo cuenta de algunos archivos de ayuda para cuestiones especificas).
- Difícil de usar y poco intuitivo si no se conoce el tema.
- No posee (como la mayoría del software comercial) acceso a datawarehouses, databases y otras herramientas.

2.3. METODOLOGÍA CRISP-DM

Gran parte del éxito de un proyecto, se basa en su forma de trabajo. En los proyecto de minería de datos al igual que en los proyectos de cualquier cosa es necesario seguir una forma de trabajo, es decir una metodología.

La metodología CRISP-DM consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos.

A nivel más general, el proceso está organizado en seis fases, estando cada fase a su vez estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas. Así, si en el segundo nivel se tiene la tarea general "limpieza de datos", en el tercer nivel se explican las tareas que tienen que desarrollarse para un caso específico, como por ejemplo, "limpieza de datos numéricos", o "limpieza de datos categóricos". El cuarto nivel, recoge el conjunto de acciones, decisiones y resultados sobre el proyecto de data mining específico.

La metodología CRISP-DM estructura el ciclo de vida de un proyecto de Data Mining en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto (figura 2.6).

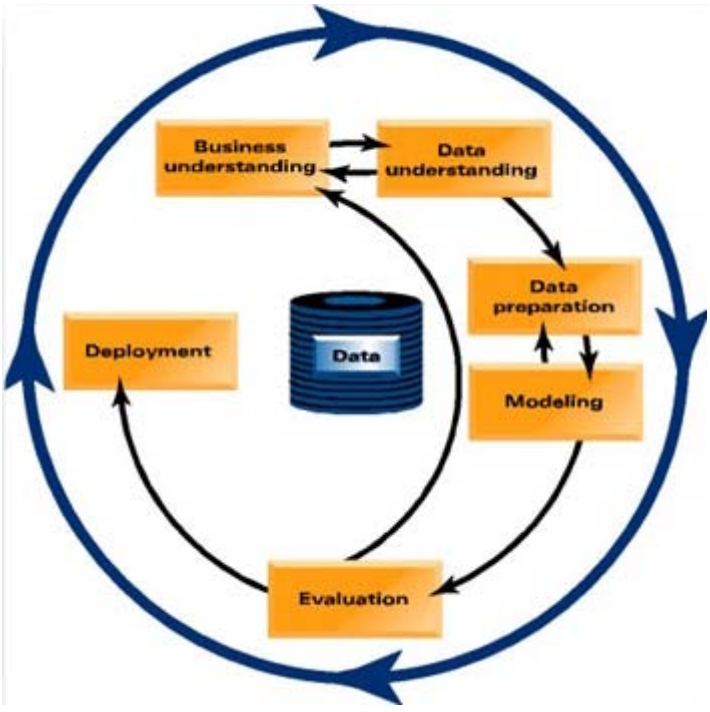


Figura 2.6 Ciclo del proyecto según la metodología CRISP-DM

I. *Comprensión del negocio*

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

II. *Comprensión de los datos*

La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

III. *Preparación de datos*

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos en las herramientas de modelado) de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

IV. *Modelado*

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

V. *Evaluación*

En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos.

Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida.

VI. *Desarrollo*

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo. Dependiendo de los requerimientos, la fase de

desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

La figura 2.7 presenta las fases del proceso acompañadas por tareas genéricas y las salidas.

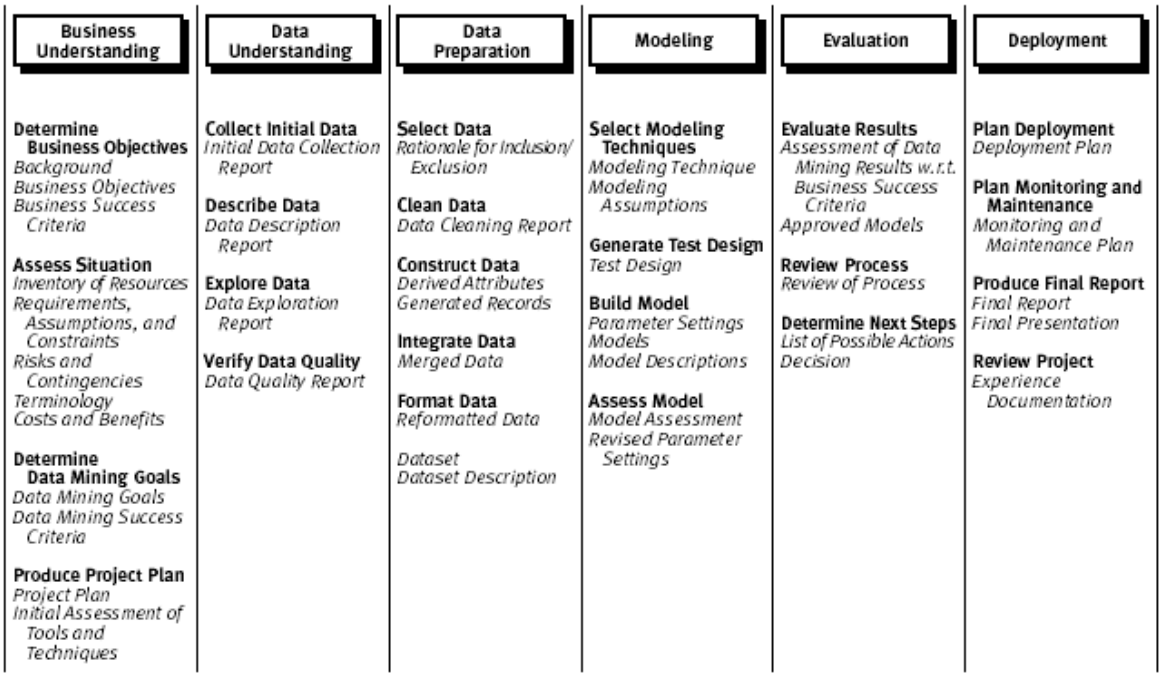


Figura 2.7 Fases del proceso, tareas genéricas (en negrita) y salidas (en itálica)

3. DESCRIPCIÓN DEL PROBLEMA

Cada año la ONU, a través del Programa de las Naciones Unidas para el Desarrollo (PNUD) se ocupa de recolectar gran cantidad de información e indicadores de los distintos países en los que tiene presencia. Esta información es analizada y compilada en informes que luego se publican y atraen gran interés a nivel mundial.

Con el paso del tiempo, las nuevas tecnologías, la especialización, los nuevos escenarios y eventos a nivel mundial y el mayor interés en estas cuestiones resultan en un número creciente de variables e indicadores. A su vez, el mundo mantiene su constante cambio, los limites y jurisdicciones de los países varían, se desatan nuevas guerras, catástrofes obligan a grandes masas de población a migrar a nuevas localizaciones, logrando un escenario para nada estático. Es decir que se conjugan dos factores, una masa de datos que crece en cantidad y que varía rápidamente con el paso del tiempo, que resultan en una situación cada vez más compleja y en un esfuerzo cada vez mayor para contrarrestarla.

Como con muchos otros indicadores y análisis estadísticos, sus resultados son aceptados y tenidos en cuenta por un alto porcentaje de la población de cada país, sin preguntarse sobre la veracidad o precisión de los mismos. En los últimos años, el índice de desarrollo humano ha cobrado gran relevancia en el mundo, sobretodo en países en vías de desarrollo, presentándose como un indicador de las buenas prácticas y de una correcta administración, lo que ha despertado interés en sus gobernantes por la evolución del mismo. Un claro reflejo son los más de 600 informes regionales que se han elaborado últimamente en más de 140 países.

Realizar un análisis estadístico tradicional sobre una masa de datos tan grande y con tantas variables involucradas sería casi imposible si se tiene en cuenta el alcance del proyecto en cuestión (tanto en tiempo, como costos, necesidad de personal, disponibilidad de herramientas y equipos informáticos, etc.) y a su vez requeriría un gran estudio y entendimiento previo para poder formular hipótesis a priori. Es por esto que la minería de datos se presenta como una opción alternativa interesante, permitiendo realizar un manejo correcto de la masa de datos y sin la necesidad de ser un profesional del área en cuestión. Por medio de esta herramienta se pretende realizar un estudio intensivo sobre un indicador de gran relevancia actualmente. La intención es determinar la correlación entre el IDH y las variables que lo definen, desglosando la incidencia individual de cada de una de ellas. Por otra parte, aprovechar las variables secundarias, que no se utilizan en el cálculo del IDH, para determinar su correlación, incidencia en el desarrollo humano y utilizarlas para verificar que los datos se encuentren alineados en los países.

Las clasificaciones (IDH Alto, Medio y Bajo) prefijadas por el PNUD también son un tema de interés. En muchos casos el valor de IDH no es tenido en cuenta y solamente se considera su clasificación, lo que hace que los limites que definen estos grupos, sean de gran importancia. Es por esto que resulta interesante agregar las variables secundarias al estudio, ya que al aportar mayor detalle que las primarias (o directas) pueden ser útiles para analizar si la clasificación realizada representa la realidad.

Dada la utilidad y relevancia del índice, se desea analizarlo, verificar que no existan países con datos inconsistentes, agregar más variables al análisis y determinar su relevancia, armar grupos de países y compararlos con la clasificación fijada por el PNUD para estudiar sus límites.

4. SOLUCIÓN

Una vez planteados tanto el estado de la cuestión, como el problema que se requiere resolver, se continúa con la solución del problema.

Para buscar una solución, aplicar las técnicas descriptas y utilizar las herramientas seleccionadas, se lleva a cabo la metodología CRISP-DM que aporta el marco necesario para desarrollar el proceso de manera ordenada.

4.1. APLICACIÓN DE CRISP-DM

Como se mencionó en el capítulo 2 Estado de la Cuestión, se procede a aplicar la metodología CRISP-DM con las siguientes fases:

- Fase I: Comprensión del negocio
- Fase II: Comprensión de los datos
- Fase III: Preparación de los datos
- Fase IV: Modelado
- Fase V: Evaluación

4.1.1. Fase I: Comprensión del negocio

4.1.1.1. Determinar los objetivos del negocio

El Índice de Desarrollo Humano es una medición por país, elaborada por el Programa de las Naciones Unidas para el Desarrollo (PNUD). En la Cumbre del Milenio de las Naciones Unidas, celebrada en 2000, los líderes del mundo asignaron al desarrollo un papel central dentro del programa mundial mediante los Objetivos de Desarrollo del Milenio, que establecen metas claras para reducir la pobreza, la enfermedad, el analfabetismo, la degradación del medio ambiente y la discriminación contra la mujer para el año 2015. Presente en 166 países, el PNUD utiliza su red mundial para ayudar al sistema de las Naciones Unidas y a sus asociados a despertar una mayor conciencia y verificar los progresos realizados, a la vez que conecta a los países con los conocimientos y los recursos necesarios para lograr estos objetivos.

Se concentra principalmente en ayudar a los países a elaborar y compartir soluciones para los desafíos que plantean las cuestiones siguientes:

- Gobernabilidad democrática.
- Reducción de la pobreza.
- Prevención y recuperación de las crisis.
- Energía y medio ambiente.

- SIDA (HIV).

El PNUD también realiza una amplia labor de promoción. El Informe sobre Desarrollo Humano anual, encargado por el PNUD, centra el debate mundial sobre cuestiones clave de desarrollo, proporcionando nuevos instrumentos de medición, análisis innovadores y, a menudo, propuestas de política controvertidas. El equipo independiente de expertos que elabora el Informe utiliza la contribución de una red mundial de personalidades destacadas del sector académico, el gobierno y la sociedad civil que aportan datos, ideas y las mejores prácticas. Los países en desarrollo y sus asociados internacionales utilizan el Informe para calibrar los resultados y configurar nuevas políticas.

El objetivo del proyecto es realizar un estudio intensivo sobre este indicador (IDH), teniendo en cuenta las variables con que se calcula como también otras variables indirectas para detectar patrones, armar grupos de comportamiento y establecer la influencia de variables no-directas en el mismo.

Los criterios de éxito del proyecto consisten en:

- Descubrir indicadores que demuestren como influyen las variables no-directas en el IDH.
- Determinar la correcta correlación entre el IDH y las variables que lo definen directamente.
- Establecer límites cuantitativos a las variables no-directas que permitan describir los patrones de comportamiento de estos grupos.
- Lograr predecir correctamente el IDH de países en que este indicador sea conocido (contrastando el predicho contra el real).

4.1.1.2. Evaluación de la situación

El PNUD cuenta con vastos recursos alrededor del mundo (está presente en 166 países) y también con la colaboración de profesionales independientes que sin pertenecer a la organización, realizan aportes para contribuir a su labor. Este proyecto utiliza la información divulgada por el PNUD así como también información adicional presentada por la ONU pero se realiza de manera independiente. Los recursos humanos asociados al mismo son:

- Tutor del Proyecto: es quien tiene la experiencia en este tipo de proyectos, el experto en la explotación de información y supervisor de las tareas a realizarse.
- Líder del proyecto: es quien conduce el proyecto de explotación de información y realiza las tareas para cumplir los objetivos.

Los requisitos identificados para avanzar con el proyecto son los siguientes:

- Acordar el alcance del proyecto.
- Profundizar los conocimientos sobre la herramienta de minería de datos a utilizar.

- Contar con la información y las bases de datos en etapas tempranas para comenzar su estudio y realizar pruebas.
- Realizar reuniones con los especialistas de la escuela de postrado de ingeniería de software del ITBA con el objetivo de analizar los datos y realizar una evaluación de las técnicas de minería de datos más convenientes.

A su vez, los principales supuestos que se realizan son:

- El IDH sigue una perfecta correlación con las variables que lo definen.
- Una vez identificada la relevancia de las variables no-directas en el IDH, se podrán utilizar para estimarlo en los países donde no está calculado.

Las principales restricciones que se enfrentan son:

- Cierta información no es divulgada por algunos países, con lo cual los valores en la base de datos son estimaciones de los mismos y el PNUD los utiliza para calcular su IDH.
- Hay países donde no se divulga ninguna de las variables necesarias para determinar el IDH y como las mismas no están estimadas, no existe valor de IDH para los mismos.
- Existen en la base de datos campos inconsistentes, con valores erróneos o vacíos que afectan al análisis de los resultados.
- Cierta información específica y detallada, que podría enriquecer aun más el análisis, se encuentra restringida a profesionales, no es de acceso libre por lo cual no puede ser incluida en el proyecto.

Las expectativas a la hora de llevar a cabo este proyecto son:

- Profundizar el conocimiento sobre las técnicas de minería de datos y las herramientas de software que se utilizan para realizarlos.
- Llevar a cabo un proyecto de minería de datos de manera integral, de principio a fin, que permita adquirir el know-how necesario para replicarlo a otros ámbitos y utilidades profesionales.
- Profundizar los conocimientos sociológicos, macroeconómicos y políticos a nivel global, y los requerimientos que existen a lo largo del mundo en materia de desarrollo.

4.1.1.3. Terminología

Esta sección tiene como finalidad ayudar a quienes no conocen el ámbito en el cual se sitúa el proyecto a familiarizarse con los términos más importantes (tabla 4.1).

INFORME

DEFINICIONES, ACRONIMOS, ABREVIATURAS			
Termino	**Descripción**	**Tipo**	**Referencia**
CRISP-DM	*Cross Industry Standard Process for Data Mining.*	*Abreviatura*	- *Página web de CRISP-DM*
IDH	*Índice de Desarrollo Humano.*	*Abreviatura*	- *Pagina web del PNUD*
ONU	*Organización de las Naciones Unidas.*	*Abreviatura*	- *Pagina web de la ONU*
PBI	*Producto Bruto Interno.*	*Abreviatura*	- *Pagina web de la ONU*
PNUD	*Programa de las Naciones Unidas para el Desarrollo*	*Abreviatura*	- *Página web del PNUD*
Variables indirectas o secundarias	*Aquellas variables de la base de datos que no se utilizan para calcular el IDH (ej.: superficie, población, porcentaje de actividad económica, etc.).*	*Definición*	

Tabla 4.1 Informe de definiciones, acrónimos y abreviaturas

4.1.1.4. Determinar el objetivo asociado a la minería de datos

Los objetivos de la minería de datos se relacionan con las búsquedas a realizar en los datos para cumplir con los objetivos del proyecto. Las pautas definidas son:

- Identificar la incidencia en el IDH de las variables que lo definen.
- Detectar patrones de incidencia de las variables no-directas sobre el IDH.
- Identificar grupos de comportamiento de países según su IDH en función de variables no-directas, y las reglas de causalidad que lo conforman.
- Predecir el IDH de países donde no esté calculado (por falta de alguna de las variables directas) a través de los patrones detectados en las variables no-directas.

Los supuestos asociados a dichos requisitos son:

- Cuanto mayor sea la esperanza de vida, mayor será el IDH.
- Cuando mayor sea el PBI, mayor será el IDH.
- A mayor índice de alfabetización, mayor será el IDH.
- Cuanto mayor sea el porcentaje de población urbana, mayor será el IDH.
- Cuanto menor sea el porcentaje de actividad económica primaria, mayor será el IDH.
- A mayor porcentaje de población mayor de 65 años, mayor IDH.

- A menor porcentaje de población menor de 15 años, mayor IDH.
- A menor fertilidad, mayor IDH.
- Cuantos más médicos, gasto en salud, cantidad de celulares y conexiones a internet, mayor IDH.
- Existen grupos de países (resultantes de subdivisiones geográficas de continentes) con niveles de IDH muy similares.
- Los grupos de países con igual IDH pueden ser definidos a través de las variables no-directas.
- A través de los patrones de comportamiento identificados en las variables no-directas, se podrá estimar el IDH en países donde no haya sido calculado.

4.1.1.5. Riesgos y contingencias

En esta instancia se establecen los riegos del proyecto y sus contingencias en proyectos de explotación de información. Es importante identificar los posibles riegos del proyecto, para que en caso de ocurrir alguno de ellos se puedan ejecutar las acciones planteadas para minimizar el efecto negativo que pudiesen provocar.

- *Riesgo 1:* Existen ciertos datos faltantes y falta obtener un atributo para todos los países (índice de alfabetización).

  Estrategia de la mitigación: Los datos faltantes se buscarán en versiones anteriores de la base de datos para ver si existían o si nunca estuvieron presentes. Se intentará conseguir el atributo faltante en la página web de la ONU.

  Acción de contingencia: Se reemplazaran los datos faltantes por alguna variable (valor cero o negativo) que los distinga de los que si existen para dejarlos fuera del análisis.
  Si el atributo no puede conseguirse, se dejara de lado y se analizara el IDH teniendo en cuenta las otras 2 variables que lo definen.

- *Riesgo 2:* El atributo “Índice de alfabetización” no se encuentra en la base de datos actual sino que debe ser obtenido de alguna otra base de datos de la ONU.

  Estrategia de la mitigación: Una vez conseguido el atributo, se agregaran estos datos a la base de datos actual, intentando que el resultado sea homogéneo para todos los países.

  Acción de contingencia: En caso de que el atributo no exista para todos los países (datos incompletos) se procederá de la misma manera que en el riesgo anterior, utilizando algún valor que deje estos campos fuera del análisis.

4.1.1.6. Realizar el plan del proyecto

La metodología CRISP-DM se estructura en 6 fases. A continuación se describe el plan para llevar a cabo el proyecto, con una duración aproximada de 5-6 meses.

Fase I - Comprensión del negocio:

1. Búsqueda y análisis de información relacionada al PNUD y el IDH.
2. Definir objetivos de negocio y de minería de datos.
3. Evaluar riesgos y criterios de éxito.

Fase II - Comprensión de los datos:

4. Adquirir la base de datos con la información pertinente.
5. Recolección y búsqueda de datos adicionales para estudiar vínculos entre los datos y completar información faltante.
6. Analizar detalles y características de los datos.
7. Verificar la calidad de los datos.

Fase III - Preparación de los datos:

8. Dar formato y preparar los datos (armado del *dataset*).
9. Seleccionar atributos y filtros.

Fase IV - Modelado:

10. Aplicar las técnicas y herramientas previamente seleccionadas.
11. Generar los modelos y compararlos.

Fase V - Evaluación:

12. Evaluar el modelo obtenido.
13. Validar el modelo con el especialista en minería de datos.

Fase VI - Implementación:

14. Generar el informe final y presentar el modelo obtenido.

La información obtenida en esta fase fue realizada a través de un proceso de educción propuesto en [Britos, P. *et al*, 2008], el mismo puede verse desarrollado en el Anexo A.

4.1.2. Fase II: Comprensión de los datos

4.1.2.1. Recolección de datos iniciales

Para llevar a cabo el proyecto se utiliza una base de datos publicada en un informe anual del Programa de las Naciones Unidas para el Desarrollo (PNUD). En este caso se utiliza la última versión, correspondiente al periodo 2007/2008. Esta base de datos cuenta con numerosas tablas que hacen foco en distintos temas, de las cuales se utilizan 6 que repercuten directamente en el tema en desarrollo.

El informe se encuentra en formato "Portable Document Format" (pdf). Las tablas se extraen del mismo como archivos de texto (txt) para luego convertirse en tablas de Microsoft Excel. Este proceso insume bastante tiempo ya que los archivos de texto poseen caracteres adicionales que dificultan la tarea de encolumnarlos en tablas y que luego deben ser limpiados manualmente para dejar en formato homogéneo y prolijo los datos.

Las 5 tablas que conforman la base de datos a ser utilizada son:

- IDH: esta tabla posee los datos relacionados directamente con el Índice de Desarrollo Humano, como son la esperanza de vida al nacer, el índice de alfabetización, el PBI per cápita, los índices de estas 3 variables para determinar el IDH, el IDH y el ranking por países.
- Demografía: para cada uno de los países rankeados, esta tabla informa el continente, población total (en 3 años distintos), tasa de crecimiento poblacional, porcentaje de población urbana, porcentaje de población menor de 16 años y mayor de 65, densidad poblacional y porcentaje de actividad económica.
- Salud: esta tabla refiere a la situación en que se encuentra la salud de la población de cada país. Las variables que contiene son: gasto (público, privado y per cápita) en salud, cantidad de niños de 1 año inmunizados contra tuberculosis y sarampión, porcentaje de mujeres casadas entre 15 y 49 años, porcentaje de nacimientos atendidos por profesionales y cantidad de médicos.
- Crisis: esta tabla hace enfoque en las crisis y riesgos de salud a los que están expuestos los ciudadanos. Contiene datos como prevalencia de SIDA, uso de preservativo, medidas contra la malaria, casos de tuberculosis y porcentaje de adultos fumadores.
- Technology: esta tabla refiere a la penetración de las distintas tecnologías en la población. Los datos con que cuenta son: penetración del teléfono fijo, celular, internet, patentes registradas, regalías por patentes recibidas por persona, gasto en investigación y desarrollo y cantidad de personas dedicadas a la investigación.

4.1.2.2. Descripción de los datos iniciales

Se procede a describir la estructura y los atributos con que cuenta cada una de las tablas incluidas (tablas 4.2 a 4.6).

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
HDI	Valor de IDH	Numerico
Life expect	Esperanza de vida al nacer	Numerico
Literacy	Tasa de alfabetización	Numerico
Education	Tasa de enrolamiento a educación primaria, secundaria y terciaria	Numerico
GDP	PBI per cápita	Numerico
Life index	Indice de esperanza de vida para calculo del IDH	Numerico
Education index	Indice de alfabetizacion para calculo del IDH	Numerico
GDP index	Indice de PBI para calculo del IDH	Numerico
GDP rank - HDI rank	Ranking de PBI menos Ranking de IDH	Numerico

Tabla 4.2 Tabla IDH

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
Continent	Continente al que pertence el país	Texto
Total Pop	Población total (años 1975, 2005 y 2015)	Numerico
Growth	Tasa anual de crecimiento poblacional (periodo 1975-2005 y 2005-2015)	Numerico
Urban Pop	Porcentaje de población urbana	Numerico
Pop < 15 years	Población menor de 15 años	Numerico
Pop > 65 years	Población mayor de 65 años	Numerico
Fertility rate	Tasa de fertilidad	Numerico
Density	Densidad poblacional	Numerico
Economical activity	Porcentaje de actividad primaria, secundaria y terciaria	Numerico

Tabla 4.3 Tabla Demografía

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
Health expenditure	Gasto en salud publico, privado y per capita	Numerico
Immunized	Porcentaje de niños de 1 año inmunizados contra tuberculosis y sarampion	Numerico
Diarrohoea	Porcentaje de niños con diarrea que reciben rehidratación oral y alimentación continua	Numerico
Contracep	Porcentaje de mujeres entre 15-49 años casadas	Numerico
Births skilled	Porcentaje de nacimientos atendidos por profesionales	Numerico
Physicians	Cantidad de médicos	Numerico

Tabla 4.4 Tabla Salud

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
HIV prevalence	Porcentaje de personas entre 15-49 años infectadas	Numerico
Condom	Porcentaje de uso de preservativo en relaciones sexuales	Numerico
Antimalarial	Medidas utilizadas para tratar personas infectadas de malaria	Numerico
Tuberculosis	Cantidad de casos de tuberculosis existentes, detectados y curados	Numerico
Smoking	Porcenaje de adultos que fuman	Numerico

Tabla 4.5 Tabla Crisis

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
Telephone	Cantidad de telefonos fijos cada 1000 personas	Numerico
Cellular	Cantidad de celulares cada 1000 personas	Numerico
Internet	Cantidad de usuarios de internet cada 1000 personas	Numerico
Patents	Cantidad de patentes otorgadas cada millon de personas	Numerico
Royalties	Regalias recibidas por patentamiento por persona	Numerico
R&D	Gasto en investigación y desarrollo	Numerico
Researchers	Cantidad de personas dedicadas a la investigación	Numerico

Tabla 4.6 Tabla Tecnología

Tanto el campo “HDI Rank” como “Country” se encuentra presente en las 5 tablas y sirve como nexo entre ellas.

4.1.2.3. Dataset seleccionado

Dataset es el conjunto de datos que se utilizará para realizar la minería de datos. El dataset presentado a continuación es preliminar. Cuenta con los datos que resultan de mayor interés y que muestran mayor completitud para todos los países. En una instancia posterior, podrían agregarse otros atributos que demuestren importancia, crearse nuevos atributos (por ejemplo discretizando la información existente) o realizarse modificaciones y correcciones sobre los datos actuales (tabla 4.7).

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
HDI	Valor de IDH	Numerico
Life expect	Esperanza de vida al nacer	Numerico
Literacy	Tasa de alfabetización	Numerico
Education	Tasa de enrolamiento a educación primaria, secundaria y terciaria	Numerico
GDP	PBI per cápita	Numerico
Life index	Indice de esperanza de vida para calculo del IDH	Numerico
Education index	Indice de alfabetizacion para calculo del IDH	Numerico
GDP index	Indice de PBI para calculo del IDH	Numerico
Continent	Continente al que pertence el país	Texto
Total Pop	Población total (años 1975, 2005 y 2015)	Numerico
Urban Pop	Porcentaje de población urbana	Numerico
Pop < 15 years	Población menor de 15 años	Numerico
Pop > 65 years	Población mayor de 65 años	Numerico
Fertility rate	Tasa de fertilidad	Numerico
Density	Densidad poblacional	Numerico
Economical activity	Porcentaje de actividad primaria, secundaria y terciaria	Numerico
Health expenditure	Gasto en salud publico, privado y per capita	Numerico
Physicians	Cantidad de médicos	Numerico
HIV prevalence	Porcentaje de personas entre 15-49 años infectadas	Numerico
Cellular	Cantidad de celulares cada 1000 personas	Numerico
Internet	Cantidad de usuarios de internet cada 1000 personas	Numerico

Tabla 4.7 Dataset preliminar

4.1.2.4. Exploración de los datos

Esta instancia tiene como finalidad lograr un mayor entendimiento preliminar de los datos. Antes de utilizar la herramienta de minería de datos es importante conocer la información que se va a utilizar, lo cual permite comprender los resultados obtenidos con la herramienta, detectar patrones desconocidos y realizar correcciones o ajustes si algo resulta fuera de lugar.

Si bien este proceso también permite analizar la calidad de la información, detectar valores faltantes y erróneos, las correcciones a los mismos se realizan en una etapa posterior.

- IDH: a continuación se presenta la cantidad de países con IDH alto, medio y bajo, así como también un mapa del desarrollo a nivel mundial. Se ve la clara acumulación en el continente africano de países de bajo desarrollo humano (figura 4.1 y 4.2).

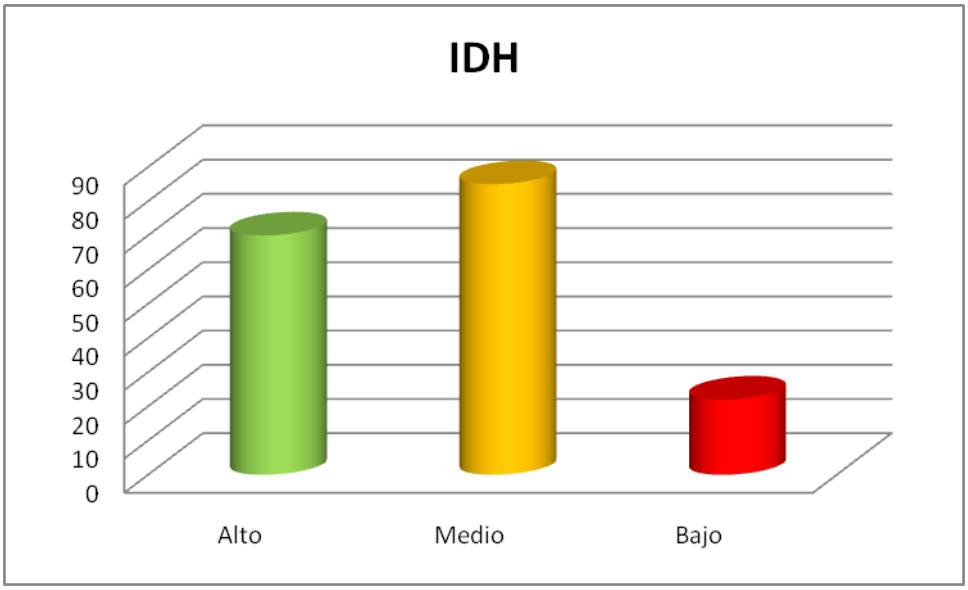


Figura 4.1 Cantidad de países según nivel de IDH

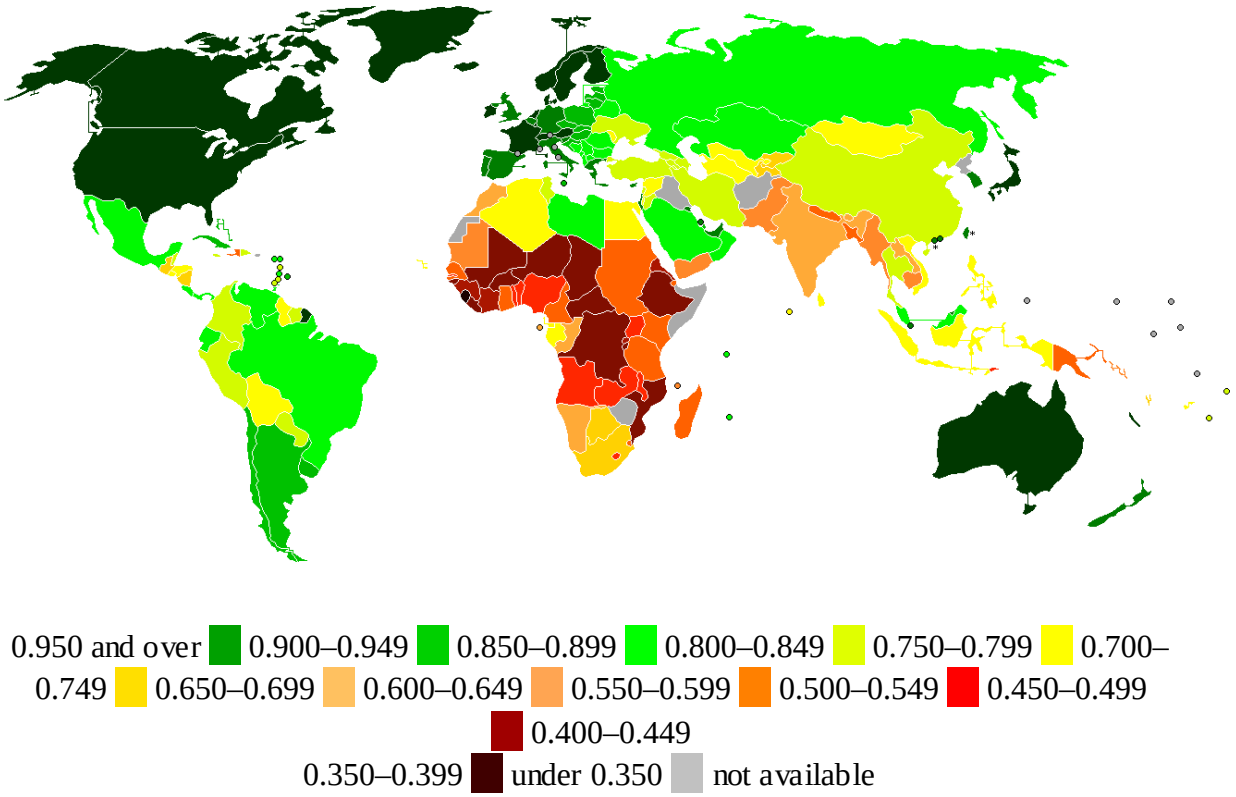


Figura 4.2 Mapa mundial del desarrollo humano

- Expectativa de vida: de los 177 países considerados, ninguno excede los 90 años y solo 11 (6%) superan los 80 años. Los 18 países con expectativa menor a 50 años, pertenecen a África (figura 4.3).

Figura 4.3 Expectativa de vida al nacer

- Educación: esta variable informa, en porcentaje, la tasa de enrolamiento a educación primaria, secundaria y terciaria. Existen 5 países que carecen de este dato y se trataran en el apartado de limpieza de datos para que no afecten el análisis. De los 23 países con enrolamiento inferior al 50%, 19 pertenecen al continente africano (figura 4.4).

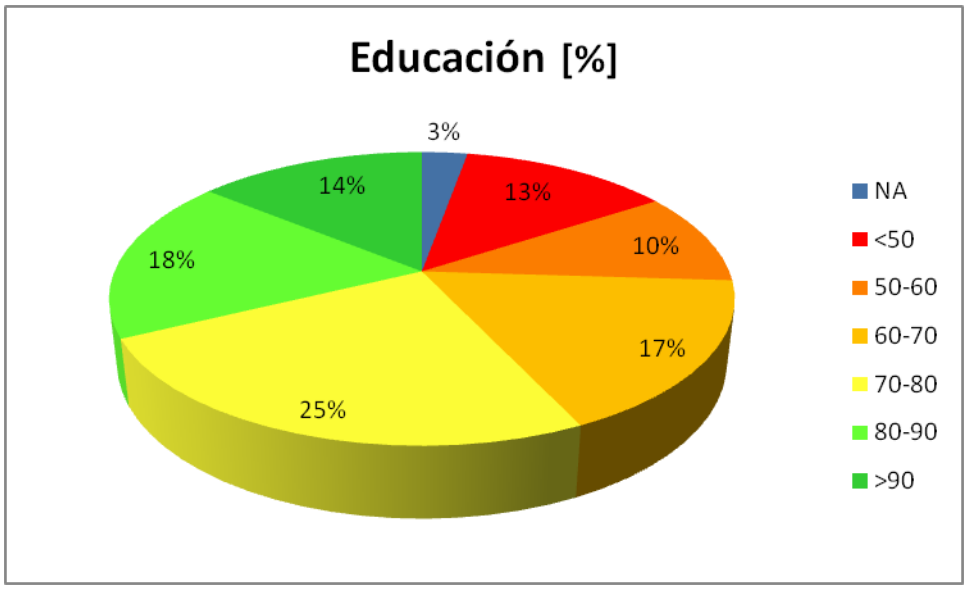


Figura 4.4 Tasa de enrolamiento a educación primaria, secundaria y terciaria

- PBI: el PBI per cápita (en inglés GDP) es la tercera variable que define el IDH. En este caso se puede ver nuevamente que los países de bajo ingreso se acumulan en África (8 de los 9 con PBI per cápita menor a 1000 dólares) (figura 4.5 y 4.6).

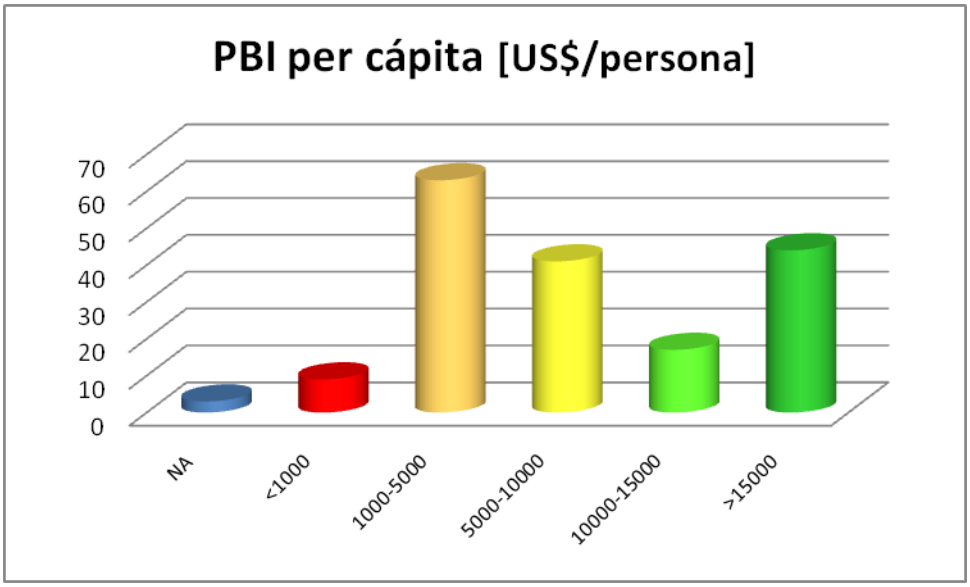


Figura 4.5 PBI per cápita

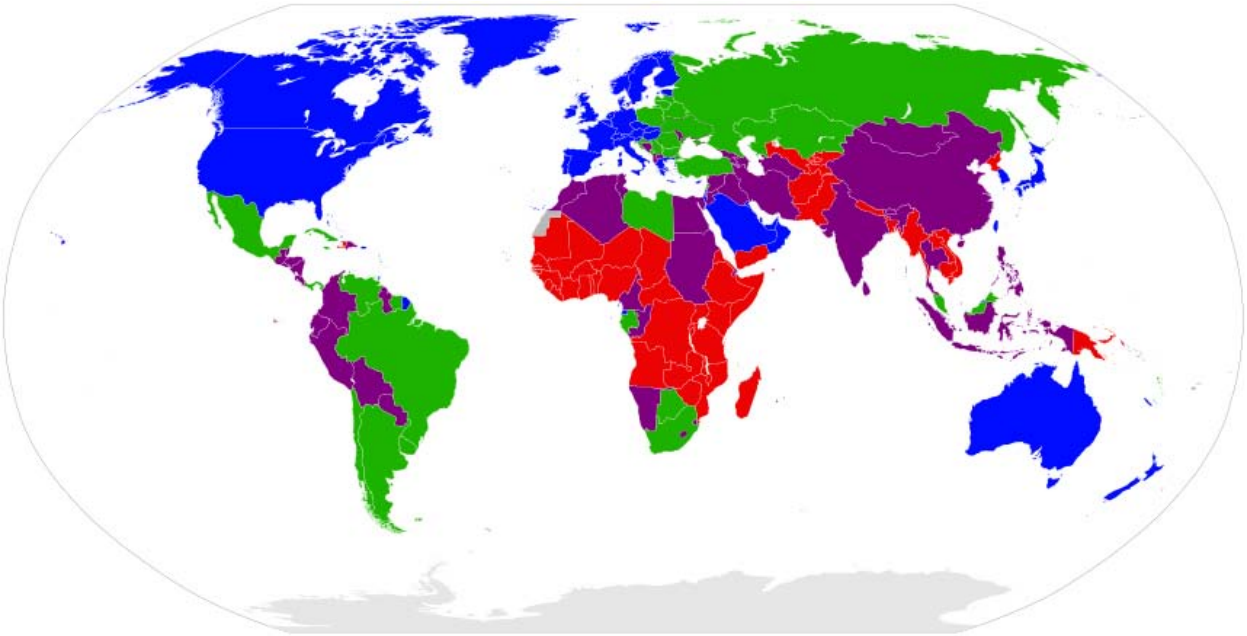


Figura 4.6 Mapa mundial del ingreso per cápita

- Porcentaje de población urbana: en el siguiente gráfico se puede ver la cantidad de países en cada intervalo. Como se puede ver, solo 5 países poseen menos de 15% de población urbana, y 13 poseen más del 90% (figura 4.7).

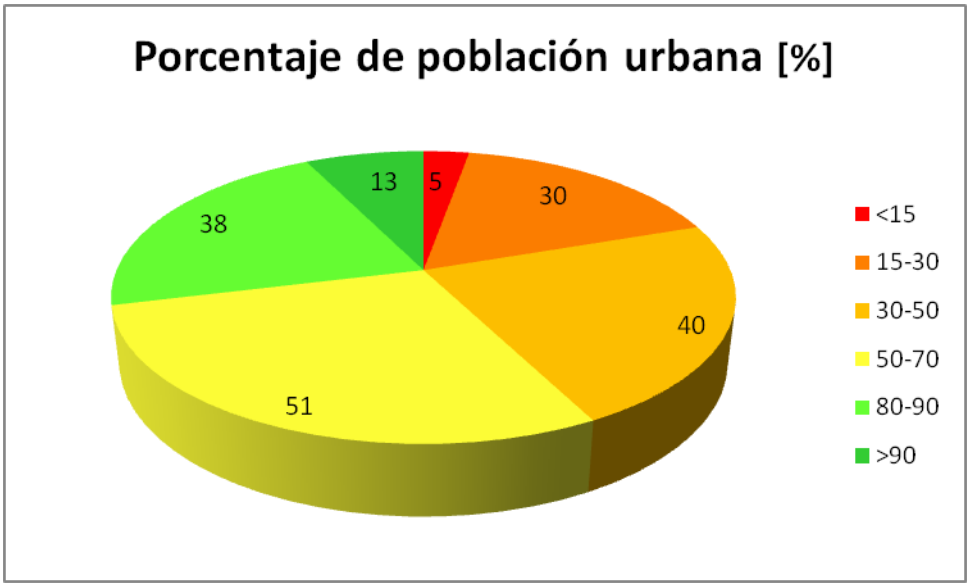


Figura 4.7 Porcentaje de población urbana

- Fertilidad: esta variable indica la cantidad de nacimientos promedio por mujer. El único país (en realidad es una región administrativa especial) cuya tasa de fertilidad es menor a 1 es Hong Kong. De los 31 países con más de 5 nacimientos por mujer, 28 pertenecen a África (figura 4.8).

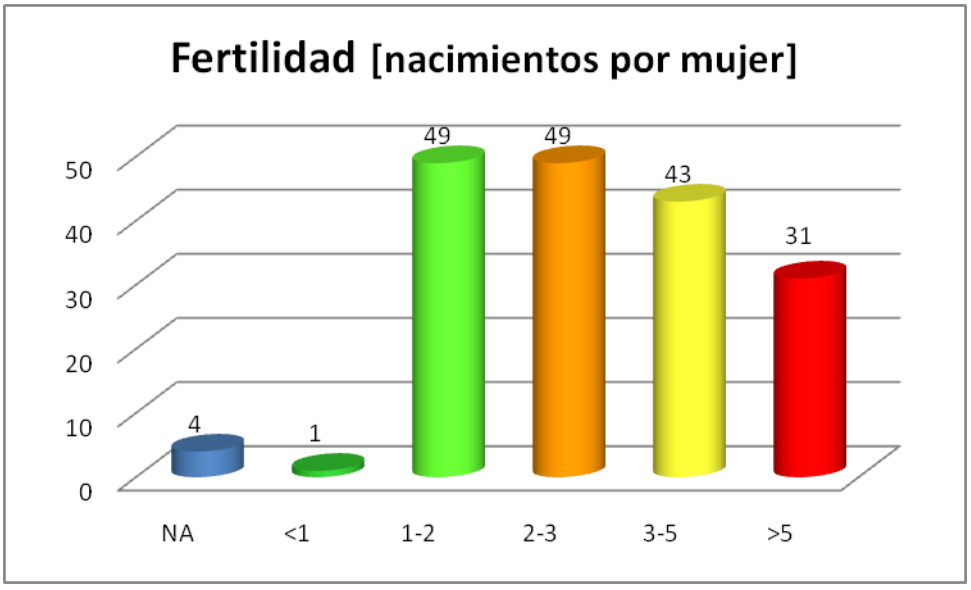


Figura 4.8 Fertilidad

- Gasto público en salud: la variable en análisis es “Health Exp_D” que refiere al gasto que realiza cada país destinado a la salud de la población y está representado en función al porcentaje del PBI (por ejemplo: el valor “<1” significa que ese país gasta menos del 1% de su PBI en salud). Existen 9 países con gasto inferior al 1%, 3 pertenecen a África y 6 a Asia. En el otro extremo, 7 países gastan más del 8%, 4 son de Europa, 2 de África y 1 de Oceanía (figura 4.9).

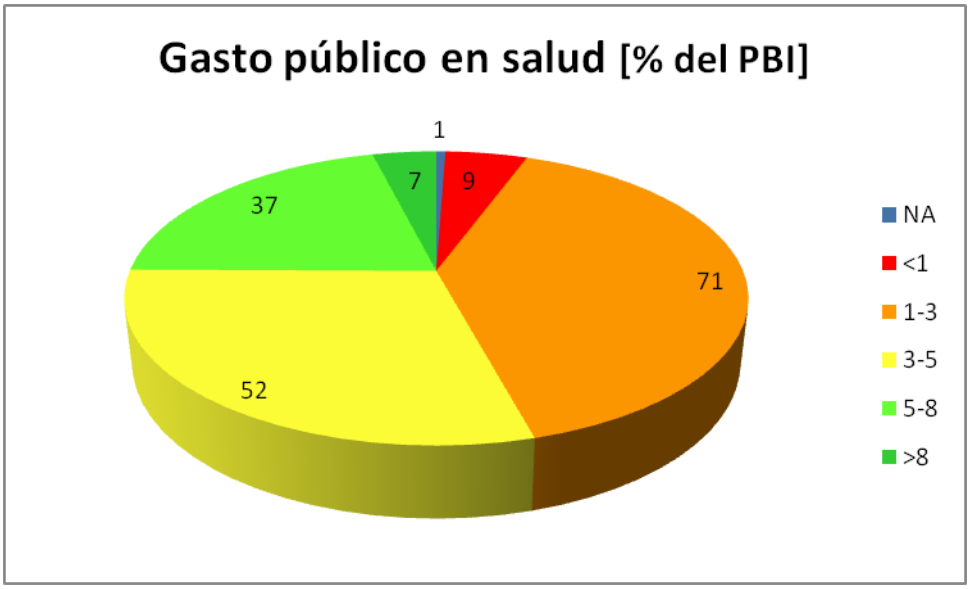


Figura 4.9 Gasto público en salud

- Cantidad de médicos: la variable "Physicians" representa la cantidad de médicos cada 1000 habitantes en cada país. En el siguiente gráfico se muestra una línea de tendencia, donde, dejando de lado los casos puntuales, se puede ver como a medida que decrece el índice de desarrollo, la cantidad de médicos es menor. Existen 3 países en África con 2 médicos cada 1000 personas (figura 4.10).

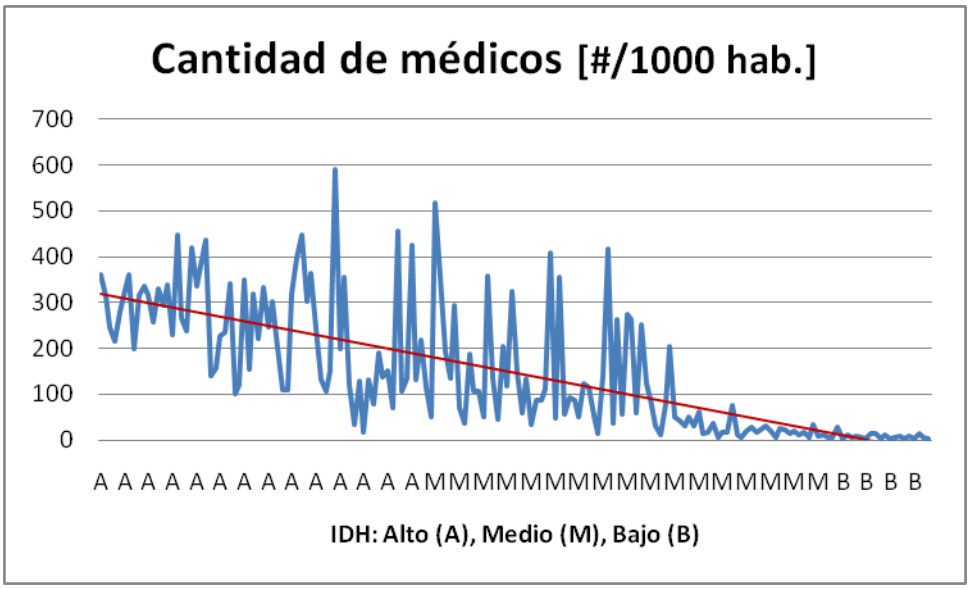


Figura 4.10 Cantidad de médicos por cada 1000 habitantes

- HIV: esta variable representa el porcentaje de habitantes del país infectados con HIV. Se puede ver que existen 48 países donde los infectados exceden el 1.5% de la

población. En el gráfico de torta se muestra la cantidad de países por continente que exceden el 1.5% de infectados (figura 4.11 y 4.12).

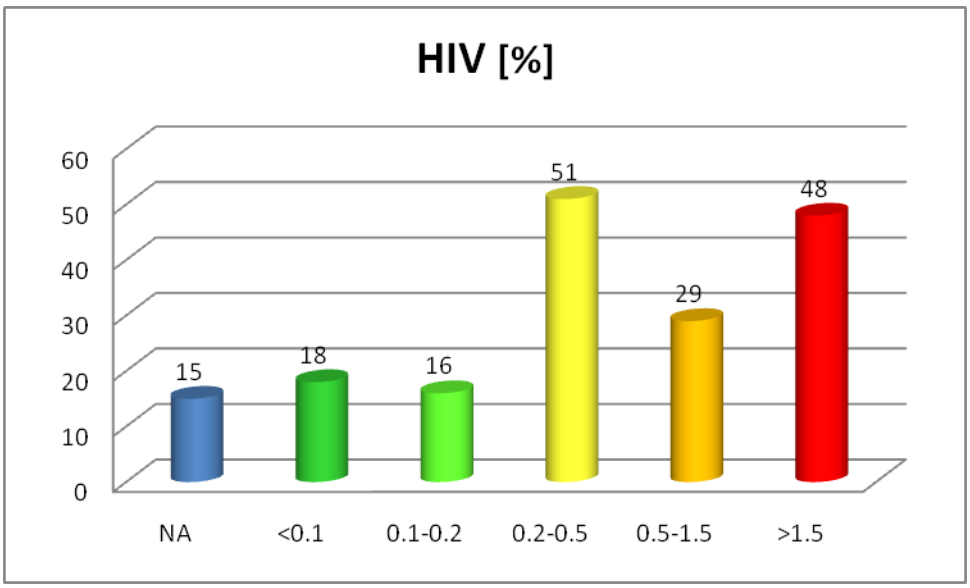


Figura 4.11 Porcentaje de personas infectadas con HIV

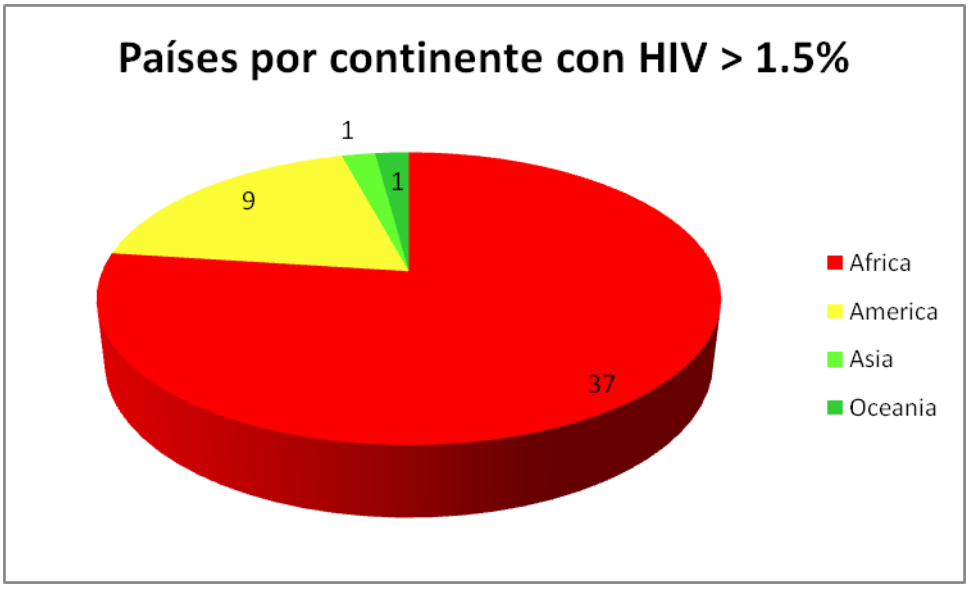


Figura 4.12 Cantidad de países por continente con más de 1.5% de infectados

- Celular: representación de la variable "Cellular" que informa la cantidad de suscripciones a servicios de telefonía celular cada 1000 habitantes. De los 38 países que exceden los 800 celulares por cada 1000 personas, hay 17 de ellos que exceden los 1000 (es decir, más de uno por persona) (figura 4.13).

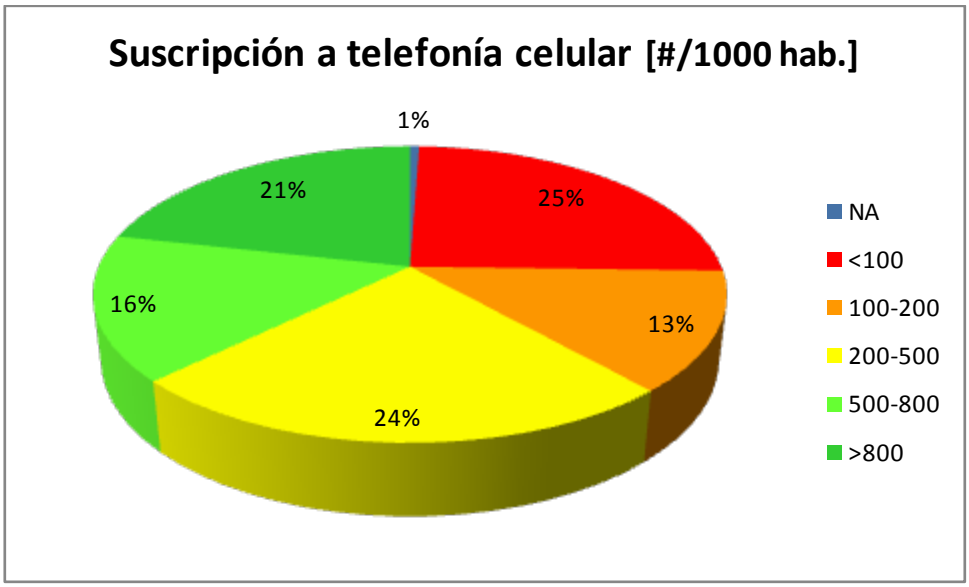


Figura 4.13 Cantidad de suscripciones a telefonía celular

4.1.2.5. Verificación de calidad de datos

Durante la exploración de los datos fue posible detectar ciertas inconsistencias en la masa de datos. Si bien los datos en si son confiables por la fuente de la cual provienen, existen ciertos datos que no están disponibles ya sea porque no han sido medidos o porque esos países eligen no divulgar esa información lo que resulta en campos vacios. Durante la selección del Dataset se tuvo en cuenta este fenómeno en una primera aproximación, eligiendo variables con alta completitud de datos en la medida en que fue posible, sin embargo, algunas variables son de suma importancia para el análisis y han sido incorporadas mas allá de estas faltantes.

La variable "Literacy" que representa la tasa de alfabetización y junto con la tasa de enrolamiento define el índice "Educación" del IDH, presenta 21% de campos vacios (38 en total), lo cual es demasiado alto.

La variable "HIV" posee 15 campos vacios, el 8% del total y la variable "Economical Activity" posee 13 campos vacios, el 7% del total.

Existen otras variables como "Education", "GDP", "Pop < 15 years", "Pop > 65 years", "Fertility", etc. donde también hay campos vacios, pero en el peor de los casos hay menos de 3% faltante.

Por otra parte, existen valores de ciertas variables que no son mediciones concretas sino estimaciones realizadas por el PNUD con la finalidad de calcular el IDH para ese país. En estos casos el organismo asigna un valor a dicha variable lo que permite realizar el cálculo del IDH pero no se trata de la medición sino una estimación del mismo. Por ejemplo se

asigna el valor máximo (99%) a la variable “Literacy” a la hora de calcular el índice de educación para Islandia.

En la fase siguiente, “Preparación de los datos” se analiza estos inconvenientes encontrados y se aplican medidas para solucionarlos.

4.1.3. Fase III: Preparación de los datos

4.1.3.1. Preparación del Dataset

La versión definitiva del Dataset cuenta con 177 registros, los países tenidos en cuenta, y 23 atributos para cada uno de ellos. El detalle de los campos se presenta a continuación (tabla 4.8).

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
HDI	Valor de IDH	Numerico
Life expect	Esperanza de vida al nacer	Numerico
Literacy	Tasa de alfabetización	Numerico
Education	Tasa de enrolamiento a educación primaria, secundaria y terciaria	Numerico
GDP	PBI per cápita	Numerico
Life index	Indice de esperanza de vida para calculo del IDH	Numerico
Education index	Indice de alfabetizacion para calculo del IDH	Numerico
GDP index	Indice de PBI para calculo del IDH	Numerico
Continent	Continente al que pertence el país	Texto
Total Pop	Población total (años 1975, 2005 y 2015)	Numerico
Urban Pop	Porcentaje de población urbana	Numerico
Pop < 15 years	Población menor de 15 años	Numerico
Pop > 65 years	Población mayor de 65 años	Numerico
Fertility rate	Tasa de fertilidad	Numerico
Density	Densidad poblacional	Numerico
Economical activity	Porcentaje de actividad primaria, secundaria y terciaria	Numerico
Health expenditure	Gasto en salud publico, privado y per capita	Numerico
Physicians	Cantidad de médicos	Numerico
HIV prevalence	Porcentaje de personas entre 15-49 años infectadas	Numerico
Cellular	Cantidad de celulares cada 1000 personas	Numerico
Internet	Cantidad de usuarios de internet cada 1000 personas	Numerico

Tabla 4.8 Dataset

4.1.3.2. Selección de los datos

Las tablas seleccionadas del informe del PNUD cuentan con un total de 46 atributos, de estos, se seleccionan 23 que conforman el Dataset. Se trata de los datos con mayor impacto en el estudio a realizar así como también los que presentan un buen nivel de completitud. Es posible que durante la Fase de Modelado se agreguen otros atributos si los elegidos no satisfacen las necesidades o no resultan convincentes. A su vez, se procederá en la "Construcción de datos" a crear nuevos atributos a partir de los existentes, discretizando los mismos.

Los campos incluidos en el análisis son los siguientes:

HDI Rank: campo numérico que informa la posición en el ranking de países según el IDH.

Country: campo de texto que informa el nombre del país en cuestión.

HDI: campo numérico que contiene el valor de IDH para ese país.

Life expect: campo numérico que informa, en años, la expectativa de vida al nacer en ese país.

Literacy: campo numérico que indica la tasa de alfabetización. Informa el porcentaje de personas mayores de 15 años consideradas alfabetos.

Education: campo numérico que indica la tasa combinada de enrolamiento a educación primaria, secundaria y terciaria. Se presenta como el porcentaje de cada país.

GDP: campo numérico que indica el PBI per cápita de dicho país (en inglés "Gross domestic product", GDP), en dólares por persona.

Life index: campo numérico que indica el coeficiente de esperanza de vida para calcular el IDH de ese país (el IDH se calcula con el coeficiente de esperanza de vida, el de educación y el de PBI).

Education index: campo numérico que indica el coeficiente de educación/alfabetización, para calcular el IDH de ese país (el IDH se calcula con el coeficiente de esperanza de vida, el de educación y el de PBI).

GDP index: campo numérico que indica el coeficiente de PBI per cápita para calcular el IDH de ese país (el IDH se calcula con el coeficiente de esperanza de vida, el de educación y el de PBI).

Continent: campo de texto que informa el continente al que pertenece el país.

Total Pop: campo numérico que indica la población total del país en millones de personas.

Urban Pop: campo numérico que indica el porcentaje de población del país que vive en urbes.

Pop < 15 years: campo numérico que indica el porcentaje de la población que es menor de 15 años.

Pop > 65 years: campo numérico que indica el porcentaje de la población que es mayor de 65 años.

Fertility rate: campo numérico que indica la tasa de fertilidad de dicho país. Esta expresado en cantidad de nacimientos por mujer.

Density: campo numérico que indica la densidad poblacional del país, expresado en cantidad de personas por kilometro cuadrado.

Economical activity: campo numérico que indica cómo se distribuye la actividad económica del país. Indica el porcentaje de la actividad económica que es primaria, secundaria y terciaria.

Health expenditure: campo numérico que expresa el gasto público que realiza el país en materia de salud. Esta expresado como porcentaje del PBI de dicho país.

Physicians: campo numérico que indica la cantidad de médicos que existen en el país, expresado como cantidad de médicos cada 1000 habitantes.

HIV prevalence: campo numérico que informa el porcentaje de personas entre 15 y 49 años infectadas de HIV en dicho país.

Cellular: campo numérico que indica la cantidad de suscripciones a servicios de telefonía celular por cada 1000 personas.

Internet: campo numérico que indica la cantidad de personas con acceso a internet por cada 1000 personas.

4.1.3.3. Limpieza de datos

Tal como se verificó en la etapa de "Verificación de calidad de datos" existen ciertos atributos con cantidad importante de valores faltantes, así como otros con cantidades mínimas.

En primera instancia, se utilizaron las estimaciones del PNUD para completar aquellos campos que contaban con una estimación. Si bien no se trata de una medición real, se asume la confiabilidad de la fuente y por eso se utilizan las estimaciones para completar estos valores faltantes.

Este proceso soluciona los 38 campos vacios de la variable "Literacy" (la más crucial ya que representan un 21% del total), los 5 campos vacios de la variable "Education" y los 3 campos vacios de la variable "GDP". Por lo tanto, estas 3 variables, que se utilizan directamente en el cálculo del IDH, quedan sin valores faltantes.

En el resto de las variables los campos vacios no pueden ser completados ya que no existe información ni estimaciones confiables para hacerlo. De todas formas, el porcentaje de completitud en estos casos es suficientemente alto para conducir el análisis.

Cuando las variables son continuas, se reemplazara el valor nulo por "-1" para que el atributo pueda seguir siendo considerado como continuo y si la variable se discretiza, se utilizara el valor "NA" para el atributo discreto.

4.1.3.4. Construcción e integración de datos

En esta instancia se procede a construir los datos que conforman el Dataset definitivo, el que se utilizará en la etapa de Modelado.

Como la mayoría de los atributos presentes en el Dataset seleccionado se tratan de variables continuas, se procede a discretizar las mismas seleccionando intervalos de valores. Este paso sirve para ayudar a la herramienta de minería de datos ya que algunas variables tienen datos muy dispersos y de esta manera se puede tener una visión más rápida del comportamiento de los mismos.

Como primer paso se realiza un paneo general de la masa de datos para determinar su máximo y mínimo. Luego se analiza con mayor profundidad para determinar la variabilidad del conjunto, esto permite elegir los puntos de corte de los intervalos, intentando que no sean demasiado grandes (lo que causaría perdida sustancial de información) ni demasiado chicos (con lo cual la discretización no tendría sentido). A continuación se presentan 2 ejemplos de los valores que pueden tomar las variables discretizadas (tabla 4.9 y 4.10).

IDH	
Rango Continuo	Valor Discreto
IDH >0.8	Alto
0.5< IDH <0.8	Medio
0.5> IDH	Bajo

Tabla 4.9 Discretización de la variable “IDH”

Education	
Rango Continuo	Valor Discreto
Edu >90	>90
80< Edu <90	80-90
70< Edu <80	70-80
60< Edu <70	60-70
50< Edu <60	50-60
50> Edu	<50

Tabla 4.10 Discretización de la variable “Education”

El Dataset resultante luego de agregar las variables discretas, se presenta a continuación (tabla 4.11).

Atributo	Descripción	Tipo
HDI rank	Ranking por paises segun su IDH	Numerico
Country	Pais	Texto
HDI	Valor de IDH	Numerico
HDI_D	Variable "HDI" discreta	Texto
Life expect	Esperanza de vida al nacer	Numerico
Life expect_D	Variable "Life expect" discreta	Texto
Literacy	Tasa de alfabetización	Numerico
Education	Tasa de enrolamiento a educación primaria, secundaria y terciaria	Numerico
Education_D	Variable "Education" discreta	Texto
GDP	PBI per cápita	Numerico
GDP_D	Variable "GDP" discreta	Texto
Life index	Indice de esperanza de vida para calculo del IDH	Numerico
Education index	Indice de alfabetizacion para calculo del IDH	Numerico
GDP index	Indice de PBI para calculo del IDH	Numerico
Continent	Continente al que pertence el país	Texto
Total Pop	Población total (años 1975, 2005 y 2015)	Numerico
Total Pop_D	Variable "Total Pop" discreta	Texto
Urban Pop	Porcentaje de población urbana	Numerico
Urban Pop_D	Variable "Urban Pop" discreta	Texto
Pop < 15 years	Población menor de 15 años	Numerico
Pop < 15 years_D	Variable "Pop < 15 years" discreta	Texto
Pop > 65 years	Población mayor de 65 años	Numerico
Pop > 65 years_D	Variable "Pop > 65 years" discreta	Texto
Fertility rate	Tasa de fertilidad	Numerico
Fertility rate_D	Variable "Fertility rate" discreta	Texto
Density	Densidad poblacional	Numerico
Density_D	Variable "Density" discreta	Texto
Economical activity	Porcentaje de actividad primaria, secundaria y terciaria	Numerico
Economical activity_D	Variable "Economical activity" discreta	Texto
Health expenditure	Gasto en salud publico, privado y per capita	Numerico
Health expenditure_D	Variable "Health expenditure" discreta	Texto
Physicians	Cantidad de médicos	Numerico
Physicians_D	Variable "Physicians" discreta	Texto
HIV prevalence	Porcentaje de personas entre 15-49 años infectadas	Numerico
HIV prevalence_D	Variable "HIV prevalence" discreta	Texto
Cellular	Cantidad de celulares cada 1000 personas	Numerico
Cellular_D	Variable "Celular" discreta	Texto
Internet	Cantidad de usuarios de internet cada 1000 personas	Numerico
Internet_D	Variable "Internet" discreta	Texto

Tabla 4.11 Dataset definitivo

4.1.4. Fase IV: Modelado

En esta etapa del proyecto se seleccionan y aplican las técnicas de modelado sobre el dataset que ha sido armado, configurado y limpiado en la fase anterior.

El desarrollo de esta fase está directamente vinculado a la herramienta de software seleccionada para llevar a cabo el estudio de minería de datos. Para esto, se ha elegido el programa llamado "Tanagra" cuyas características principales fueron descriptas en la sección 2.2.2 Herramienta de minería de datos.

En el desarrollo de esta fase se explicarán las técnicas a utilizar y el modo de hacerlo (al menos a grandes rasgos) en el software utilizado y se presentarán las salidas obtenidas junto con su posterior análisis.

4.1.4.1. Selección de la técnica de modelado

Las técnicas de modelado a ser utilizadas dependen de los objetivos que se han definido previamente. El Índice de Desarrollo Humano es una variable continua, sin embargo, el PNUD designa rangos a este índice generando un agrupamiento de los países en 3 categorías posibles (IDH Alto, Medio o Bajo). En principio se desea analizar la correlación entre el IDH y las variables que lo definen. Para realizar esto, se selecciona el algoritmo "C4.5".

Este algoritmo se utiliza para generar arboles de decisión. Los arboles generados se pueden utilizar para clasificar, tratándose de un clasificador estático. Se utiliza como un modelo predictivo que realiza mapeos sobre observaciones de una variable para generar conclusiones sobre la variable objetivo. En este caso en particular, se analiza el comportamiento de las 3 variables que definen el IDH para obtener conclusiones sobre la variable objetivo (el IDH discreto) y de esta forma verificar la correlación existente.

En segunda instancia, se desea analizar distintos agrupamientos de los países, teniendo en cuenta las variables directas como las no directas. Para esta finalidad se selecciona el algoritmo "K-Means".

Este algoritmo de clustering es el más utilizado en la actualidad por su facilidad y eficiencia. Se trata de un método de análisis de clústeres que apunta a particionar "n" observaciones en "k" clústeres teniendo en cuenta el comportamiento de la masa de datos. En particular, se busca realizar agrupamientos alternativos al que realiza el PNUD y compararlos.

Para analizar la conformación de los grupos se agrega un algoritmo denominado "Group Characterization" el cual realiza comparaciones estadísticas para caracterizar los grupos. También se desea comparar las reglas que los estructuran, de modo que utiliza en esta instancia el algoritmo C4.5 para obtener arboles de decisión que expliquen el comportamiento de los datos dentro de cada clúster.

Otro objetivo planteado es el análisis del comportamiento de las variables no directas con el IDH, es decir, establecer si existen reglas que puedan demostrar si existe relación entre ellas. Nuevamente, el algoritmo C4.5 se utilizara con esta finalidad.

Finalmente se utilizaran las reglas definidas y los clústeres creados con la finalidad de predecir el rango de IDH de los países sin tener en cuenta las variables directas, lo que busca demostrar la validez de las mismas y la alta incidencia de las variables no directas.

Además de los algoritmos mencionados, existen herramientas secundarias en la aplicación que también serán utilizadas, ya sea para configurar el algoritmo, para exportar datos al Excel, para graficar o modificar la forma de presentación de resultados.

4.1.4.2. Construcción del modelo y resultados experimentales

A continuación se presentan los modelos utilizados, se explica el funcionamiento de los mismos y los resultados obtenidos junto con su análisis pertinente. A su vez se detallarán las configuraciones realizadas y los parámetros utilizados en la herramienta.

4.1.4.2.1. IDH – Variables directas

Como se menciona en los objetivos del proyecto, en primera instancia se pretende analizar la incidencia en el IDH de las variables que lo definen: esperanza de vida, tasa de alfabetización y PBI per cápita. El algoritmo seleccionado para realizarlo es el C4.5.

Se comienza por cargar el dataset preparado en la Fase III al Tanagra, el cual informa que se trata de una base con 39 atributos y 177 ejemplos (los países incluidos). Como la cantidad de atributos y ejemplos es la correcta (es decir, no falta ninguno), se puede asegurar la homogeneidad en la forma de los datos ya que de no ser así, el programa hubiese cargado solo la información en formato correcto y la cantidad de datos sería inferior a la esperada.

A continuación, se agrega en el programa una función denominada “Define Status” que permite seleccionar los atributos de entrada (“Input”) y atributos objetivo (“Target”) que se utilizaran con el próximo algoritmo.

Luego, de seleccionar los atributos, se elige el algoritmo requerido, en este caso “Supervised Learning (C4.5)”, el cual debe ser configurado teniendo en cuenta el número mínimo de hojas posibles (las hojas de un árbol de decisión son los conjuntos ya clasificados de ejemplos) y el nivel de confianza deseado.

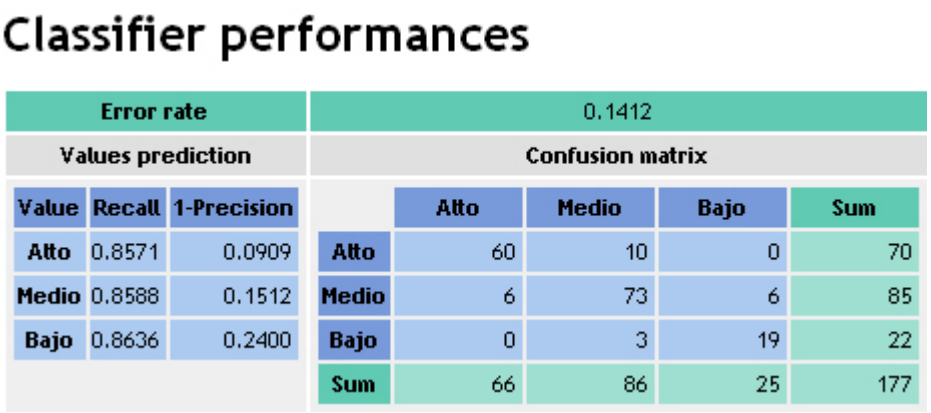
Se realizaran 4 análisis individuales, uno para cada variable directa contra el IDH y uno más teniendo en cuenta las 3 variables directas contra el IDH. Esto es para analizar la incidencia de cada variable individualmente, compararlas y también tener en cuenta la incidencia de las variables como conjunto.

- IDH – Esperanza de vida

Define Status: Input: “Life Expect” Target: “HDI_D”

Supervised Learning (C4.5): Min size of leaves: 2 Confidence level: 0.25

Los resultados obtenidos y la matriz de confusión son (figura 4.14):



Classifier performances

Error rate	0.1412

Values prediction

Value	Recall	1-Precision
Alto	0.8571	0.0909
Medio	0.8588	0.1512
Bajo	0.8636	0.2400

Confusion matrix

	Alto	Medio	Bajo	Sum
Alto	60	10	0	70
Medio	6	73	6	85
Bajo	0	3	19	22
Sum	66	86	25	177

Figura 4.14 Esperanza de vida e IDH

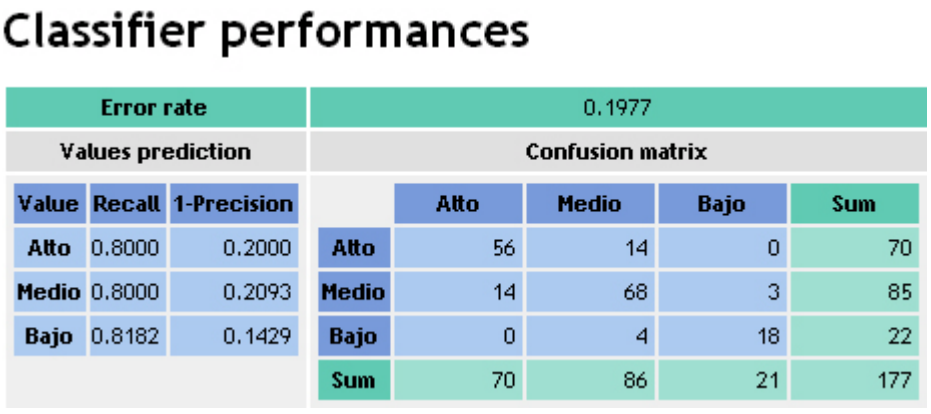
Como se puede observar, la esperanza de vida por si sola sirve para clasificar el IDH con un error del 14%. En la matriz de confusión, cuantos más elementos estén sobre la diagonal, menor el error cometido (si los totales estuvieran sobre la diagonal, el error cometido sería nulo), es decir, más ejemplos clasificados correctamente bajo el conjunto de reglas generado. En este caso, se puede ver que se clasifican 60 ejemplos de IDH Alto correctamente de un total de 70, 73 de un total de 85 IDH Medio y 19 de un total de 22 IDH Bajo. A su vez, 25 ejemplos en total, son clasificados incorrectamente.

- IDH – Tasa de alfabetización

Define Status: Input: “Education” Target: “HDI_D”

Supervised Learning (C4.5): Min size of leaves: 2 Confidence level: 0.25

Los resultados obtenidos y la matriz de confusión son (figura 4.15):



Classifier performances

Error rate	0.1977

Values prediction

Value	Recall	1-Precision
Alto	0.8000	0.2000
Medio	0.8000	0.2093
Bajo	0.8182	0.1429

Confusion matrix

	Alto	Medio	Bajo	Sum
Alto	56	14	0	70
Medio	14	68	3	85
Bajo	0	4	18	22
Sum	70	86	21	177

Figura 4.15 Tasa de alfabetización e IDH

En este caso, si se predice el IDH teniendo en cuenta solamente el índice de alfabetización se comente un error del 20%, es decir que, aproximadamente 1 de cada 5 países se clasifican mal. En este caso, la matriz de confusión muestra que se han clasificado menor cantidad de ejemplos bien para las 3 categorías de IDH que con la

esperanza de vida, lo que también se ve reflejado en que el error en este caso es 6% mayor.

- IDH – PBI per cápita

Define Status: Input: "GDP" Target: "HDI_D"

Supervised Learning (C4.5): Min size of leaves: 2 Confidence level: 0.25

Los resultados obtenidos y la matriz de confusión son (figura 4.16):

Classifier performances

Error rate	0.0847

Values prediction

Value	Recall	1-Precision
Alto	0.8714	0.0000
Medio	0.9765	0.1354
Bajo	0.8182	0.1000

Confusion matrix

	Alto	Medio	Bajo	Sum
Alto	61	9	0	70
Medio	0	83	2	85
Bajo	0	4	18	22
Sum	61	96	20	177

Figura 4.16 PBI per cápita e IDH

Aquí se puede observar la predicción del IDH al tener en cuenta únicamente el PBI per cápita. En este caso, el error cometido es del 8%. Este resultado es claramente mejor a los dos casos anteriores, siendo un 6% menor que con la esperanza de vida y un 12% menor que con la tasa de alfabetización. La matriz de confusión permite llegar a la misma conclusión, donde se ve que 61 instancias de IDH Alto se calificaron correctamente (de un total de 70), 83 de IDH medio (de un total de 85) y 18 de IDH bajo (de un total de 22), y se calificaron un total de 15 incorrectamente.

Estos resultados muestran que los 3 atributos poseen una clara relación con el IDH así como también la preponderancia del PBI per cápita, lo que se ve reflejado en que se puede predecir el IDH de los 177 países teniendo en cuenta solamente el PBI y equivocarse en solo 15 de ellos.

- IDH – Esperanza de vida, tasa de alfabetización y PBI per cápita

Define Status: Input: "Life Expect", "Education", "GDP" Target: "HDI_D"

Supervised Learning (C4.5): Min size of leaves: 2 Confidence level: 0.25

Los resultados obtenidos y la matriz de confusión son (figura 4.17):

Classifier performances

Error rate	0.0339

Values prediction

Value	Recall	1-Precision
Alto	0.9857	0.0417
Medio	0.9529	0.0241
Bajo	0.9545	0.0455

Confusion matrix

	Alto	Medio	Bajo	Sum
Alto	69	1	0	70
Medio	3	81	1	85
Bajo	0	1	21	22
Sum	72	83	22	177

Figura 4.17 Esperanza de vida, tasa de alfabetización, PBI per cápita e IDH

En este caso, teniendo en cuenta las 3 variables en conjunto, el error cometido es el menor de todos, siendo del 3%. Esto era esperable ya que son estas 3 variables de las que deriva el cálculo del IDH, con lo cual, si los resultados fuesen distintos, indicarían un error o problema con el método en uso. Al obtener un error menor que en los otros casos, también podemos ver que las 3 variables aportan información a la predicción.

Para profundizar el análisis y obtener una visión más completa de cómo se realiza esta clasificación se utiliza el árbol de decisión construido (figura 4.18):

Decision tree

- GDP < 8854.5000
  - Life expect < 56.7500
    - GDP < 1232.0000 then HDI_D = **Bajo** (100.00 % of 16 examples)
    - GDP >= 1232.0000
      - Education < 55.5000 then HDI_D = **Bajo** (62.50 % of 8 examples)
      - Education >= 55.5000 then HDI_D = **Medio** (100.00 % of 8 examples)
  - Life expect >= 56.7500
    - Education < 87.0000
      - Life expect < 72.6500 then HDI_D = **Medio** (98.41 % of 63 examples)
      - Life expect >= 72.6500
        - GDP < 6869.5000 then HDI_D = **Medio** (87.50 % of 8 examples)
        - GDP >= 6869.5000 then HDI_D = **Alto** (66.67 % of 6 examples)
    - Education >= 87.0000 then HDI_D = **Alto** (75.00 % of 4 examples)
- GDP >= 8854.5000
  - GDP < 12443.5000
    - GDP < 10996.0000 then HDI_D = **Alto** (100.00 % of 8 examples)
    - GDP >= 10996.0000 then HDI_D = **Medio** (66.67 % of 3 examples)
  - GDP >= 12443.5000 then HDI_D = **Alto** (100.00 % of 53 examples)

Figura 4.18 Árbol de decisión

El árbol de decisión arroja un total de 10 hojas. Se puede ver, por ejemplo, que en el 100% de 53 casos, los países con PBI per cápita mayor a 12443.5 dólares tienen IDH Alto. También es posible apreciar que la incidencia de la esperanza de vida y sobretodo de la tasa de alfabetización en la generación de estas reglas, es menor que la incidencia

del PBI per cápita. Esto resulta coherente con los resultados obtenidos al analizar las variables por separado.

El PNUD calcula el IDH teniendo en cuenta 3 índices (uno para cada variable) y asigna finalmente un ponderación del 33.3% a cada uno de estos índices para obtener el IDH. Es decir que, a simple vista, el peso ponderado de las 3 variables es el mismo, al menos matemáticamente hablando. Sin embargo, aunque cada índice tenga la misma incidencia en el resultado final, no se debe perder de vista que estos índices están calculados en función de las 3 variables (al compararlas con valores definidos como máximos para las mismas), y estas variables pueden relacionarse o tener cierta dependencia una en otra.

Estos resultados permiten suponer que la esperanza de vida y la tasa de alfabetización son en gran parte, consecuencia del PBI per cápita, lo que demuestra porque esta última variable por sí sola, permite explicar en mayor medida que las otras dos, los conjuntos del IDH. Cuando se deja de mirar los datos como tales, el resultado es coherente: aquellos países donde los ciudadanos tienen mayores ingresos anuales, muestran una esperanza de vida mayor (resultado de una mayor inversión en salud, higiene, etc.) y una tasa de alfabetización mayor (a causa de mayor inversión en educación pública, etc.).

Otra cuestión que surge al analizar las salidas obtenidas es la utilidad del índice. Si los resultados obtenidos teniendo en cuenta únicamente el PBI son tan cercanos y permiten explicar con un error pequeño el IDH, tal vez podría directamente utilizarse el PBI como índice para rankear los países, lo que resultaría en mayor simplicidad a la hora de realizar los cálculos y el análisis posterior y también se ahorraría mucho tiempo y dinero en la recolección de información (se necesitaría 1 dato por país en vez de 3).

Más allá de esto, existen casos donde un mayor PBI no significa necesariamente una tasa de alfabetización mejor, lo que permite analizar la eficiencia a la hora de invertir el PBI que tiene cada país. Sería pertinente efectuar un análisis comparando el aporte de información que realiza el índice por sobre el PBI únicamente contra el esfuerzo necesario para calcular y recolectar estos datos anualmente (este análisis escapa a los objetivos y a la disponibilidad de información de este proyecto).

4.1.4.2.2. Clustering

En esta instancia el objetivo es identificar grupos de comportamiento de países en función de las variables no directas y las reglas de causalidad que los conforman. A su vez, se utiliza la misma herramienta para armar grupos de comportamiento en función de las variables directas para luego compararlos con los grupos armados por el PNUD (IDH alto, medio y bajo), estableciendo similitudes y diferencias.

El algoritmo para armar los clústeres es K-Means. Nuevamente se utiliza la función “Define Status” para establecer las variables de entrada y objetivo. El algoritmo debe ser configurado teniendo en cuenta el número de clústeres deseado, cantidad máxima de iteraciones, cantidad

de corridas de prueba y otros parámetros como tipo de generador de semilla, promedio computacional, etc. En particular se utilizarán 3 clústeres para que sean comparables con los del IDH.

Una vez que se corrió el algoritmo, se agrega una función más, "Export Dataset" la cual sirve para exportar el dataset con el nuevo atributo generado (el clúster al que pertenece cada país) a un archivo para luego ser utilizado en el Excel. Una vez que se agrega el atributo al dataset, este se carga nuevamente en Tanagra para tener disponible este nuevo atributo y poder realizar análisis adicionales sobre el mismo.

Con los nuevos atributos cargados, se utiliza el algoritmo C4.5 para armar arboles de decisión que expliquen las reglas que determinan la formación de los clústeres.

- Clúster con variables directas

Define Status: Input: "Life Expect", "Education", "GDP" Target: -

K-Means: Nº of clusters: 3 Max iterations: 15 Nº of trials: 5

Los resultados obtenidos con esta configuración son los siguientes (figura 4.19):

Cluster size and WSS

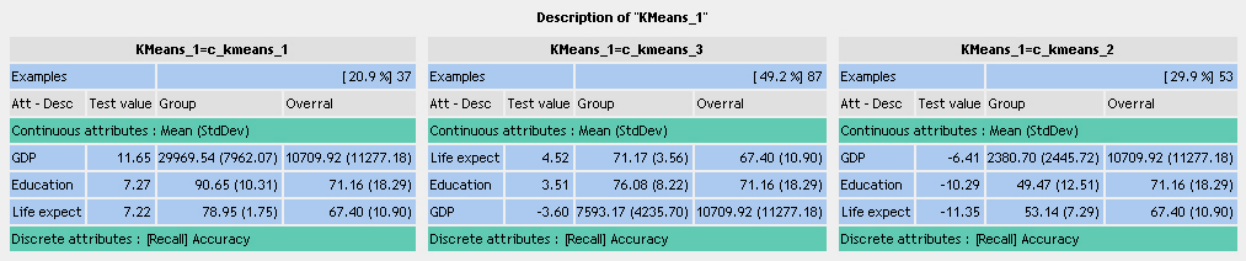
Clusters	3		
Cluster	**Description**	**Size**	**WSS**
cluster n°1	c_kmeans_1	37	30.1093
cluster n°2	c_kmeans_2	53	50.0049
cluster n°3	c_kmeans_3	87	38.6204

R-Square for each attempt

Number of trials	5
Trial	**R-square**
1	0.776222
2	0.775822
3	0.775822
4	0.776394
5	0.775822

Figura 4.19 Clústeres con variables directas

Se puede ver que con 4 pruebas se obtiene el mayor R^2 (que da información sobre la bondad de ajuste del modelo). Para un análisis más profundo de los grupos, se utiliza el algoritmo Group Characterization, cuyos resultados son (figura 4.20):



Description of "KMeans_1"

KMeans_1=c_kmeans_1			
Examples			[20.9 %] 37
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
GDP	11.65	29969.54 (7962.07)	10709.92 (11277.18)
Education	7.27	90.65 (10.31)	71.16 (18.29)
Life expect	7.22	78.95 (1.75)	67.40 (10.90)
Discrete attributes : [Recall] Accuracy			

KMeans_1=c_kmeans_3			
Examples			[49.2 %] 87
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Life expect	4.52	71.17 (3.56)	67.40 (10.90)
Education	3.51	76.08 (8.22)	71.16 (18.29)
GDP	-3.60	7593.17 (4235.70)	10709.92 (11277.18)
Discrete attributes : [Recall] Accuracy			

KMeans_1=c_kmeans_2			
Examples			[29.9 %] 53
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
GDP	-6.41	2380.70 (2445.72)	10709.92 (11277.18)
Education	-10.29	49.47 (12.51)	71.16 (18.29)
Life expect	-11.35	53.14 (7.29)	67.40 (10.90)
Discrete attributes : [Recall] Accuracy			

Figura 4.20 Caracterización grupal de clústeres con variables directas

La visualización de estos resultados permite entender rápidamente la composición de los clústeres. En el clúster 1, los países poseen una media de PBI de casi 30000 dólares con un desvío aproximado de 8000 dólares, es decir, el rango va desde los 22000 a los 38000 dólares, el extremo superior de esta variable. En el caso de la educación, la media es superior al 90% con un desvió del 10%, el rango va desde el 80% al 100%, nuevamente el extremo superior. La esperanza de vida repite este comportamiento, con una media de 79 años y un desvió de 2 años aproximadamente, su rango va desde 77 hasta 81 años. El comportamiento en cuanto a las 3 variables es el mismo, reuniendo a los países con los mejores indicadores, comparable con el clúster de IDH Alto establecido por el PNUD. Sin embargo, la cantidad de individuos en el grupo es diferente, incluye 37 países cuando el IDH Alto incluye 70.

En el clúster 3, los países poseen un rango de PBI que va desde 11800 dólares hasta 3400 dólares aproximadamente, se trata de valores medios para esta variable. En el caso de la educación, el rango va desde 84% hasta 68%, también valores medios. La esperanza de vida también posee valores medios, el rango es 67-75 años. El comportamiento de este clúster es comparable al del IDH Medio, y también lo es en cantidad, incluye 87 países contra 85 en el IDH Medio.

El clúster 2 posee un rango de PBI 0-4800 dólares aproximadamente, los valores inferiores de la variable. El intervalo de la educación es 38-62%, los mínimos también de la variable. En cuanto a la esperanza de vida, el rango también incluye los valores inferiores de la variable, 46-60 años. Nuevamente, el comportamiento en comparación con los otros dos clústeres se asemeja al IDH Bajo del PNUD. El clúster incluye 53 países contra los 22 del IDH Bajo, lo cual muestra una diferencia sustancial.

Para una mejor comprensión de las reglas que definen los clústeres, se utiliza el algoritmo C4.5 para construir un árbol de decisión y analizar su matriz de confusión (figura 4.21). Se configura el algoritmo de la siguiente manera:

Supervised Learning (C4.5): Min size of leaves: 4 Confidence level: 0.25

Classifier performances

Error rate	0.0056

Values prediction

Value	Recall	1-Precision
c_kmeans_1	0.9730	0.0000
c_kmeans_3	1.0000	0.0114
c_kmeans_2	1.0000	0.0000

Confusion matrix

	c_kmeans_1	c_kmeans_3	c_kmeans_2	Sum
c_kmeans_1	36	1	0	37
c_kmeans_3	0	87	0	87
c_kmeans_2	0	0	53	53
Sum	36	88	53	177

Figura 4.21 Matriz de confusión de clústeres con variables directas

Como se puede ver, el error utilizando un tamaño mínimo de 4 hojas para el árbol es de 0.56%. Solo se comete un error, al clasificar una instancia en el clúster 3 cuando en realidad pertenece al clúster 1. De todas formas, la clasificación es casi perfecta. El árbol de decisión se presenta a continuación (figura 4.22):

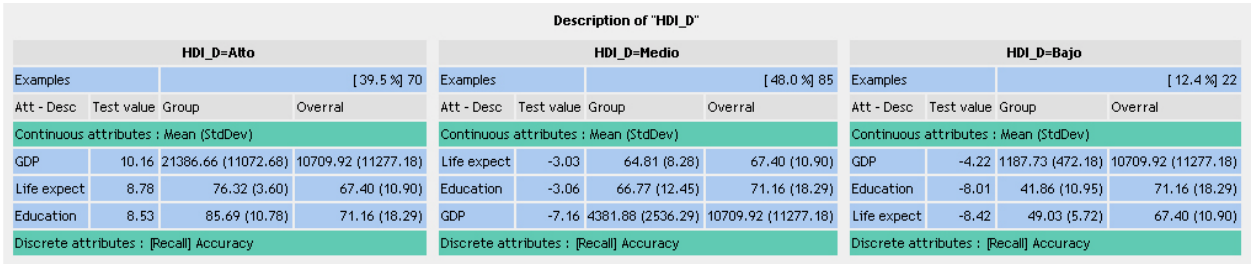
Decision tree

- GDP < 18784.5000
  - Life expect < 64.8000
    - Education < 63.4500 then KMeans_1 = **c_kmeans_2** (100.00 % of 48 examples)
    - Education >= 63.4500
      - Life expect < 57.9500 then KMeans_1 = **c_kmeans_2** (100.00 % of 5 examples)
      - Life expect >= 57.9500 then KMeans_1 = **c_kmeans_3** (100.00 % of 4 examples)
  - Life expect >= 64.8000 then KMeans_1 = **c_kmeans_3** (98.81 % of 84 examples)
- GDP >= 18784.5000 then KMeans_1 = **c_kmeans_1** (100.00 % of 36 examples)

Figura 4.22 Árbol de decisión de clústeres con variables directas

Las reglas que definen estos clústeres muestran los mismos patrones definidos al analizar la caracterización grupal de los mismos. El clúster 1 agrupa los países con los mejores valores de PBI, esperanza de vida y educación; el clúster 2 aquellos con valores intermedios; y el clúster 3 agrupa los países con los peores indicadores. Se puede ver que el PBI per cápita permite clasificar por completo al clúster 1 (como sucede con el IDH Alto), mostrándose nuevamente como un indicador predominante.

Al comparar los intervalos de estos clústeres (definidos por las reglas) con los que presentan los clústeres de IDH del PNUD se ven diferencias, lo mismo que sucede al comparar los tamaños de los grupos. Para comparar directamente los clústeres de uno y otro, se realiza una caracterización grupal de los clústeres del IDH con el algoritmo "Group Characterization" (figura 4.23).



Description of "HDI_D"

HDI_D=Alto			
Examples			[39.5 %] 70
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
GDP	10.16	21386.66 (11072.68)	10709.92 (11277.18)
Life expect	8.78	76.32 (3.60)	67.40 (10.90)
Education	8.53	85.69 (10.78)	71.16 (18.29)
Discrete attributes : [Recall] Accuracy			

HDI_D=Medio			
Examples			[48.0 %] 85
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Life expect	-3.03	64.81 (8.28)	67.40 (10.90)
Education	-3.06	66.77 (12.45)	71.16 (18.29)
GDP	-7.16	4381.88 (2536.29)	10709.92 (11277.18)
Discrete attributes : [Recall] Accuracy			

HDI_D=Bajo			
Examples			[12.4 %] 22
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
GDP	-4.22	1187.73 (472.18)	10709.92 (11277.18)
Education	-8.01	41.86 (10.95)	71.16 (18.29)
Life expect	-8.42	49.03 (5.72)	67.40 (10.90)
Discrete attributes : [Recall] Accuracy			

Figura 4.23 Caracterización grupal de clústeres del IDH

La comparación de los puntos medios y los desvíos se presenta en la siguiente tabla (tabla 4.24):

	GDP (USD)		Life Expect (años)		Education (%)		Cantidad de paises
	Media	Desvío	Media	Desvío	Media	Desvío	
Cluster 1	30000	8000	79	2	90	10	37
IDH Alto	21000	11000	76	4	86	11	70
Cluster 3	7600	4200	71	4	76	8	87
IDH Medio	4400	2500	65	8	67	12	85
Cluster 2	2400	2400	53	7	50	12	53
IDH Bajo	1200	500	50	6	42	11	22

Tabla 4.24 Comparación de clústeres

- Al comparar el IDH Alto con el clúster 1 vemos que las 3 variables poseen en el clúster 1 una media más extrema (más cercana al límite superior) y un desvío menor, estos dos fenómenos explican que este clúster este menos poblado que el IDH Alto, ya que en valores extremos de la variable se encuentran menor cantidad de países y al tener un desvío más chico, abarca un intervalo menor, donde se encuentran menos países.
- Entre el clúster 2 y el IDH Bajo el fenómeno es similar pero invertido. En este caso, las 3 variables presentan para el IDH Bajo un valor más extremo (más cercano al límite inferior) y un desvío menor, lo que resulta en que la población sea casi la mitad que la del clúster 2.
- Entre el clúster 3 y el IDH Medio, no existe un comportamiento uniforme para las 3 variables. El clúster 3 presenta un PBI más alto pero mayor desvío, y esperanza de vida y educación más altos pero con desvíos menores. La única diferencia es que el clúster 3 presenta un corrimiento con respecto al IDH Medio, abarcando valores más altos de las 3 variables. Pero al no existir un comportamiento uniforme en las 3 variables y al tratarse de valores medios (donde no existen densidades de datos muy dispares entre un intervalo y el próximo), la cantidad de individuos en los dos clústeres es similar y no hay causas aparentes para que no lo sean.

Como se puede ver, los dos agrupamientos (el generado con la herramienta y el del PNUD) coinciden en cuanto al comportamiento de los países que los componen, esto es:

un conjunto de países con los mejores valores para las 3 variables, otro conjunto con los valores medios y el último con los peores valores en las 3 variables. Y los dos agrupamientos difieren en la amplitud y posición relativa de los clústeres.

Resulta interesante efectuar un análisis para determinar que agrupamiento representa mejor el comportamiento o la categorización que se pretende dar a los países que lo componen. Es decir, el PNUD presenta un IDH Alto (70 países) y Medio (85 países) muy poblados, y un IDH Bajo (22 países) poco poblado. Como se trata de puntos de cortes fijados (al decir que un país con IDH>0.8 es "Alto"), es pertinente comprobar que esos puntos de corte representen realmente la realidad, es decir, que esos 22 países con índice de desarrollo humano bajo sean los únicos que no satisfacen las necesidades de los ciudadanos (en materias como salud, higiene, etc.) y que no existan otros países (de IDH Medio) que deberían (por no cumplir muchas de estas necesidades) ser considerados de IDH Bajo y por lo tanto el punto de corte de los grupos ser distinto. Este fenómeno se analiza al realizar los clústeres con variables indirectas y al analizar la incidencia de las variables indirectas en el IDH.

- Clúster con variables indirectas

A continuación se construyen clústeres con las variables indirectas, aquellas que no se utilizan para el cálculo del IDH. Nuevamente se utilizan 3 clústeres para que sean comparables con los del IDH y los construidos con las variables directas. Las variables seleccionadas para servir de input a la herramienta de clusterización son 9. No se utilizan todas debido a que algunas de ellas no tienen un comportamiento definido cuando se analiza su incidencia en el desarrollo humano (esto se explica en mayor detalle en la fase "4.2.3 IDH - variables indirectas") o se comportan de manera casi idéntica a otra variable si incluida. La configuración del algoritmo K-Means es la siguiente:

Define Status: Input: "Urban pop", "Pop < 15 years", "Pop > 65 years", "Fertility", "Primary", "Health exp", Physicians", "Cellular", "Internet"
Target: -

K-Means: Nº of clusters: 3 Max iterations: 15 Nº of trials: 5

Los resultados obtenidos con esta configuración son los siguientes (figura 4.25):

Cluster size and WSS

Clusters	3		
Cluster	**Description**	**Size**	**WSS**
cluster n°1	c_kmeans_1	42	126.6452
cluster n°2	c_kmeans_2	72	243.2903
cluster n°3	c_kmeans_3	63	269.5740

R-Square for each attempt

Number of trials	5
Trial	**R-square**
1	0.597498
2	0.598550
3	0.595588
4	0.598550
5	0.598550

Figura 4.25 Clústeres con variables indirectas

Se puede ver que con 2 pruebas se obtiene el mayor R^2. Para un análisis más profundo de los grupos, se utiliza el algoritmo Group Characterization, cuyos resultados son (figura 4.26):

Description of "KMeans_6"

KMeans_6=c_kmeans_1				KMeans_6=c_kmeans_3				KMeans_6=c_kmeans_2			
Examples			[23.7 %] 42	Examples			[35.6 %] 63	Examples			[40.7 %] 72
Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
Pop > 65 years	10.90	14.57 (2.78)	7.09 (5.07)	Urban Pop	4.12	64.27 (19.28)	54.39 (23.63)	Pop < 15 years	10.14	40.34 (5.36)	29.82 (11.39)
Internet	10.86	480.17 (172.14)	179.24 (205.11)	Physicians	1.45	165.60 (123.41)	145.14 (139.56)	Fertility	10.06	4.64 (1.39)	3.04 (1.75)
Cellular	10.12	934.12 (203.43)	429.68 (368.68)	Cellular	0.80	459.59 (235.36)	429.68 (368.68)	Primary	8.72	28.48 (15.71)	16.33 (15.30)
Physicians	8.28	301.36 (95.54)	145.14 (139.56)	Internet	-1.63	145.35 (104.69)	179.24 (205.11)	Health Exp	-4.70	2.72 (1.96)	3.62 (2.11)
Health Exp	7.37	5.72 (1.93)	3.62 (2.11)	Health Exp	-1.73	3.25 (1.31)	3.62 (2.11)	Pop > 65 years	-7.28	3.73 (1.12)	7.09 (5.07)
Urban Pop	6.38	74.78 (13.13)	54.39 (23.63)	Pop > 65 years	-2.22	5.95 (3.77)	7.09 (5.07)	Internet	-7.81	33.35 (37.80)	179.24 (205.11)
Primary	-5.89	4.15 (4.00)	16.33 (15.30)	Pop < 15 years	-3.17	26.16 (9.06)	29.82 (11.39)	Physicians	-8.58	36.11 (53.40)	145.14 (139.56)
Fertility	-6.41	1.52 (0.35)	3.04 (1.75)	Primary	-3.71	10.57 (8.10)	16.33 (15.30)	Urban Pop	-9.54	33.86 (14.16)	54.39 (23.63)
Pop < 15 years	-8.14	17.30 (2.81)	29.82 (11.39)	Fertility	-4.62	2.22 (1.08)	3.04 (1.75)	Cellular	-9.55	109.25 (96.07)	429.68 (368.68)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

Figura 4.26 Caracterización grupal de clústeres con variables indirectas

La visualización de estos resultados permite entender la composición de los clústeres. Se denominarán de aquí en adelante clúster 1*, clúster 2* y clúster 3*, para diferenciarlos de los anteriores.

En el clúster 1*, los países poseen una media de 15% de población mayor de 65 años (la media de todos los países es 7%), 480 personas (cada 1000) poseen acceso a internet (la media de todos los países es 180), más de 300 médicos (cada 100000 personas), el gasto en salud pública supera el 5% del PBI (la media poblacional es 3.6% del PBI), la

fertilidad es de 1.5 nacimientos por mujer (contra 3 de media entre todos los países). En los países más desarrollados se espera por ejemplo: menor fertilidad, mayor acceso a internet, menor porcentaje de actividad primaria, mayor porcentaje de población urbana, mayor cantidad de médicos, mayor gasto en salud, etc. Teniendo en cuenta el comportamiento esperado de estas variables en función de su contribución al desarrollo de un país, y los valores encontrados en este clúster, podemos decir que agrupa aquellos países con mayor índice de desarrollo humano, con mejor calidad de vida. Por lo tanto, podríamos comparar el comportamiento de este grupo (más allá de que las variables que lo formen no sean las mismas) con el de IDH Alto. El clúster es menos poblado (42 países contra 70 de IDH Alto).

En el clúster 3* encontramos una media de 165 médicos (la media poblacional es 145), 145 personas con acceso a internet (la media es 180), casi 6% de población de más de 65 años, el gasto en salud pública es de 3.25% del PBI y la fertilidad 2.2 nacimientos por mujer. Al comparar los valores medios de las variables en el grupo con las medias poblacionales podemos comprobar que son muy próximos, lo que denota que este clúster agrupa a los países con comportamiento intermedio. Si se tienen en cuenta nuevamente los valores esperados de estas variables podemos asumir que se trata de países comparables con aquellos de IDH Medio. En este caso, el clúster agrupa 63 países contra 85 de IDH Medio.

El clúster 2* posee una media de 33 personas con acceso a internet, 36 médicos, 3.7% de población mayor de 65 años, fertilidad de 4.6 nacimientos por mujer y un gasto en salud pública equivalente a 2.7% del PBI. En todas las variables el comportamiento se repite, reflejando los valores inferiores esperados para las mismas (teniendo en cuenta su influencia en el desarrollo), lo que define a este grupo como aquellos países con peor desarrollo humano y calidad de vida, comparable con aquellos de IDH Bajo. El clúster 2* agrupa 72 países, mas de 3 veces los incluidos en el IDH Bajo (22 países).

Para analizar las reglas que llevan a la formación de estos clústeres y ver en mayor detalle la incidencia de cada variable, se utiliza el algoritmo C4.5 para construir el árbol de decisión. La configuración es la siguiente:

Supervised Learning (C4.5): Min size of leaves: 4 Confidence level: 0.25

A continuación se presenta la matriz de confusión asociada (figura 4.27):

Classifier performances

Error rate	0.0339
Values prediction	Confusion matrix

Value	Recall	1-Precision
c_kmeans_1	0.9524	0.0000
c_kmeans_3	0.9683	0.0615
c_kmeans_2	0.9722	0.0278

	c_kmeans_1	c_kmeans_3	c_kmeans_2	Sum
c_kmeans_1	40	2	0	42
c_kmeans_3	0	61	2	63
c_kmeans_2	0	2	70	72
Sum	40	65	72	177

Figura 4.27 Matriz de confusión de clústeres con variables indirectas

El error asociado al realizar la predicción es del 3.4%. Se clasifican mal solo 6 instancias de las 177 totales, lo que refleja un comportamiento adecuado. A continuación se presenta el árbol de decisión construido (figura 4.28):

Decision tree

- Pop > 65 years < 8.4500
  - Cellular < 406.0000
    - Urban Pop < 59.4000
      - Pop < 15 years < 32.3500
        - Urban Pop < 45.8000 then KMeans_6 = **c_kmeans_2** (75.00 % of 8 examples)
        - Urban Pop >= 45.8000 then KMeans_6 = **c_kmeans_3** (100.00 % of 5 examples)
      - Pop < 15 years >= 32.3500 then KMeans_6 = **c_kmeans_2** (100.00 % of 64 examples)
    - Urban Pop >= 59.4000 then KMeans_6 = **c_kmeans_3** (81.82 % of 11 examples)
  - Cellular >= 406.0000 then KMeans_6 = **c_kmeans_3** (100.00 % of 37 examples)
- Pop > 65 years >= 8.4500
  - Cellular < 646.0000 then KMeans_6 = **c_kmeans_3** (83.33 % of 12 examples)
  - Cellular >= 646.0000 then KMeans_6 = **c_kmeans_1** (100.00 % of 40 examples)

Figura 4.28 Árbol de decisión de clústeres con variables indirectas

Las reglas coinciden con el comportamiento analizado al caracterizar los grupos, por ejemplo: el clúster 1* se conforma con aquellos países con más de 8.45% de población mayor de 65 años y con más de 646 celulares cada 1000 personas; el clúster 2* muestra los peores valores (en cuanto a desarrollo), población urbana menor al 46.8%, más de 32% de población menor de 15 años; y el clúster 3* nuevamente, valores intermedios.

Más allá de los valores puntuales, lo importante es que el comportamiento es el descripto anteriormente, claramente definido por un clúster de desarrollo alto, otro de medio y otro de bajo. También es interesante comprobar que la clasificación se realiza con un error del 3.4%, es decir, se clasifican solo 6 instancias mal, teniendo en cuenta únicamente 4 variables de las 9 incluidas en la herramienta. Como los clústeres claramente reflejan el nivel de desarrollo de los países que incluyen, este fenómeno refleja que ciertas variables tienen una mayor incidencia o preponderancia que otras sobre el desarrollo.

Si bien estos clústeres no se realizaron con las mismas variables de entrada, reflejan 3 niveles definidos de desarrollo. Como se mencionó al construir los clústeres en función de las variables directas, se desea analizar si el agrupamiento que realiza el PNUD es apropiado al categorizar un país como de desarrollo alto, medio o bajo o si los límites que definen las agrupaciones, deberían ser distintos. A continuación se presenta una tabla con la cantidad de países que incluye cada clúster (tabla 4.12):

Desarrollo Alto			Desarrollo Medio			Desarrollo Bajo		
IDH Alto	Cluster 1	Cluster 1*	IDH Medio	Cluster 3	Cluster 3*	IDH Bajo	Cluster 2	Cluster 2*
70	37	42	85	87	63	22	53	72

Tabla 4.12 Cantidad de países por clúster

En principio se ve que el clúster 1 y el clúster 1*, son similares en tamaño, pero sobretodo, que ambos son significativamente más chicos que el IDH Alto. En cuanto a desarrollo medio, el clúster 3* se aleja del clúster 3 e IDH Medio (que presentan un comportamiento similar) pero en menor medida. En el extremo inferior, el clúster 2* es mayor al clúster 2 y mucho mayor al IDH Bajo. Si bien los tamaños no son idénticos, el comportamiento que reflejan ambas clusterizaciones es similar, reduciendo el clúster superior y aumentando el inferior. Ambos pasan de una situación donde el clúster alto era significativamente mayor al clúster bajo, a una situación donde el clúster bajo es ahora, mayor al alto. Si tenemos en cuenta que los niveles de los clústeres denotan desarrollo, ambas clusterizaciones "elevan" los requerimientos mínimos para decir que un país es de alto desarrollo humano, lo que genera un corrimiento y una población mucho mayor en el estrato menor.

Al existir en los clústeres con variables indirectas, un estrato inferior mucho más poblado, también refleja que muchos países poseen valores no deseables para estos indicadores. A su vez, hay que tener en cuenta que las variables indirectas se encuentran correlacionadas con las directas (la esperanza de vida tiene alta relación con la inversión en salud, con la cantidad de médicos, etc.), de alguna forma, dan mayor información sobre las mismas. Resulta llamativo que existan países de IDH alto o medio, con expectativas de vida muy altas, donde el gasto en salud o la cantidad de médicos es muy baja, es decir, el indicador mayor (en este caso la expectativa de vida) tiene un valor que al compararse con indicadores relacionados (gasto en salud, médicos, etc.) no resulta convincente.

A modo de ejemplo, los siguientes países: Albania, Belice, Paraguay, Sri Lanka, Vietnam, Siria (existen más), están catalogados como de IDH Alto o Medio. Sin embargo, el clúster de las variables indirectas, los posiciona en el estrato de desarrollo bajo. Al analizarlos más profundamente, se corrobora que todos poseen el indicador de esperanza de vida de sus ciudadanos muy alto, casi igual o mayor a lo requerido para ser de IDH Alto. Sin embargo, en variables como gasto en salud, cantidad de médicos, fertilidad, estos países ni siquiera llegan al promedio poblacional, al valor medio. Es decir que cuentan con un indicador en el estrato superior pero en otros indicadores altamente relacionados, no llegan siquiera al valor medio.

4.1.4.2.3. IDH - Variables indirectas

En esta instancia se pretende identificar la relación existente entre las variables indirectas y el IDH. Se analiza cada una de las variables y su comportamiento en función del IDH para determinar la correlación o vínculo existente, cuan fuerte es dicha correlación y también para agrupar aquellas variables con mayor influencia.

El algoritmo seleccionado es el C4.5. Las variables a analizar son: continente, población, población urbana, población menor de 15 años, población mayor de 65 años, fertilidad, densidad, porcentaje de actividad primaria, gasto público en salud, cantidad de médicos, personas infectadas de HIV, cantidad de celulares y cantidad de personas con acceso a internet.

Para cada variable se utiliza el algoritmo para determinar el error al intentar clasificar los países según su IDH y las reglas que permiten la clasificación.

- IDH – Población urbana

Define Status: Input: "Urban Pop" Target: "HDI_D"

Supervised Learning (C4.5): Min size of leaves: 3 Confidence level: 0.25

Los resultados obtenidos y la matriz de confusión son (figura 4.29):

Classifier performances

Error rate	0.2429
Values prediction	**Confusion matrix**

Value	Recall	1-Precision
Alto	0.7857	0.1667
Medio	0.8353	0.2828
Bajo	0.3636	0.3333

	Alto	Medio	Bajo	Sum
Alto	55	14	1	70
Medio	11	71	3	85
Bajo	0	14	8	22
Sum	66	99	12	177

Figura 4.29 Matriz de confusión, IDH – Población urbana

El error cometido al realizar la clasificación es del 24%, se clasifican 43 instancias incorrectamente. Más allá de que el error sea alto, esto no implica que no existan reglas que resulten interesantes. Para analizarlas, se presenta el árbol de decisión con el que se realiza la clasificación (figura 4.30):

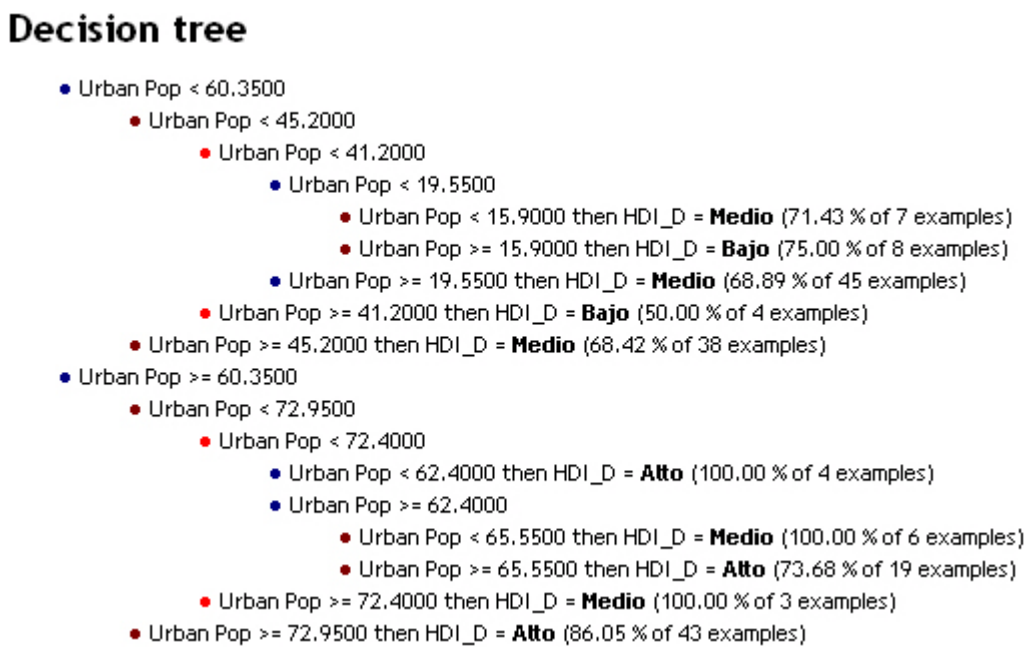


Figura 4.30 Árbol de decisión, IDH – Población urbana

Al analizar el árbol, se ve por ejemplo, que solo 8 instancias se clasifican en el clúster IDH Bajo, cuando en realidad existen 22 países en el mismo. Más allá de los errores, existen reglas que si son precisas, aun cuando la variable muestra un 24% de error. Por ejemplo, 86% de 43 ejemplos se clasifican correctamente en el grupo IDH Alto con la última regla (aquellos de población urbana mayor a 72.9%).

Como presentar la matriz de confusión de cada una de las 13 variables y el conjunto de reglas que da lugar a las mismas resulta muy extenso, se confecciona una sola tabla con los errores de cada una de las variables, para que su lectura sea más simple. En la etapa "4.2.4 Predicción del IDH" se presentan las reglas más significativas de las distintas variables.

La siguiente tabla (tabla 4.13) presentan los errores de cada variable al analizarse con el IDH:

Variable	Error
Continente	33%
Población	35%
Pob. Urbana	24%
Pob<15 años	18%
Pob>65 años	19%
Fertilidad	20%
Densidad	40%
Act. Primaria	25%
Salud	29%
Médicos	22%
HIV	38%
Celulares	16%
Internet	19%

Tabla 4.13 Errores de IDH – Variables indirectas

Como se puede ver, ninguna variable por si sola presenta un error menor al 10%, como si sucedía con el PBI, que individualmente poseía un error del 8%. Por su parte, la variable tasa de alfabetización individualmente, generaba un error del 20%, el cual es mejorado por 4 variables indirectas.

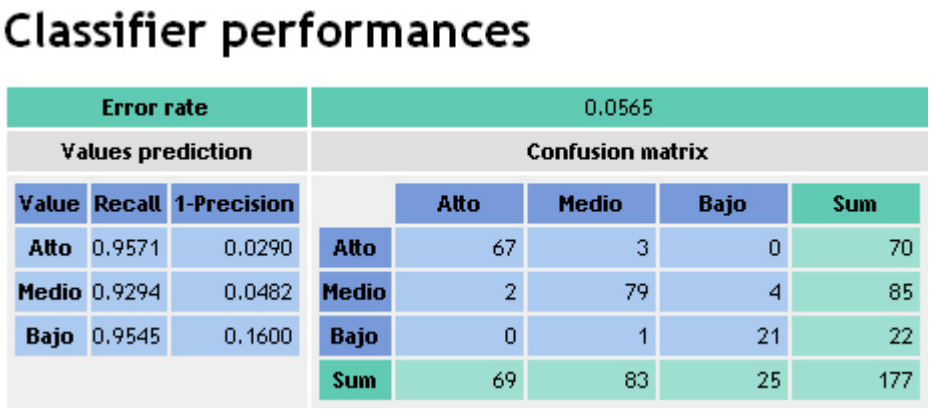
A su vez, existen variables como Población, Densidad y HIV que poseen un error muy alto (superior a 35%) y que al mirarse en detalle, no presentan un comportamiento definido en relación al IDH. Por lo tanto, estas variables se excluyen al armar clústeres.

De la misma forma que se realizó con las variables directas, ahora se analiza la incidencia de las variables indirectas no individualmente, sino tomando varias de ellas juntas. Luego de probar con muchas configuraciones distintas, se obtienen algunas cuyos resultados son más que aceptables. A continuación se presenta la configuración del algoritmo y la matriz de confusión asociada a una de estas configuraciones (figura 4.31):

Define Status: Input: “Pop < 15 years”, “Pop > 65 years”, “Primary”, “Fertility”

Target: “HDI_D”

Supervised Learning (C4.5): Min size of leaves: 3 Confidence level: 0.25



Classifier performances

Error rate	0.0565
Values prediction	Confusion matrix

Value	Recall	1-Precision
Alto	0.9571	0.0290
Medio	0.9294	0.0482
Bajo	0.9545	0.1600

	Alto	Medio	Bajo	Sum
Alto	67	3	0	70
Medio	2	79	4	85
Bajo	0	1	21	22
Sum	69	83	25	177

Figura 4.31 Matriz de confusión de agrupación de variables indirectas

Esta agrupación incluye las variables: población menor de 15 años, población mayor de 65 años, porcentaje de actividad primaria y fertilidad. Como se puede ver, el error asociado es de 5.6%, resulta cercano al obtenido con las 3 variables directas (3.4%) y solo 10 instancias se clasifican incorrectamente.

Es posible verificar mediante estos resultados la incidencia y el poder de clasificación de las variables indirectas en el IDH. Algunas variables mostraron individualmente, errores inferiores al de variables directas, y al analizarlas en conjunto se obtuvo una predicción precisa con un error levemente superior que al utilizar las 3 variables directas en conjunto. También fue posible encontrar variables que no poseen un comportamiento definido en función del desarrollo.

4.1.4.2.4. Predicción del IDH

El último objetivo planteado pretende predecir el clúster de IDH al que pertenece un cierto país, sin tener en cuenta las variables directas, únicamente con las variables indirectas. Esto, en caso de ser posible, demuestra la incidencia de estas variables y el poder de clasificación que reside en ellas.

Para llevar a cabo el objetivo se utilizan las reglas provenientes de los árboles de decisión resultantes de comparar las variables indirectas con el IDH. Como no todas las reglas resultan útiles, se debe hacer una selección de las mismas. Se tendrán en cuenta solo aquellas reglas que sirven para aislar los países en ambos extremos (IDH Alto e IDH Bajo), ya que en el estrato medio, los comportamientos no son tan definidos y en caso de tener un país cuyo IDH se desconoce, es más simple ver si cumple los requisitos para ser categorizado como Alto o si no cumple los requisitos mínimos y por lo tanto es categorizado como Bajo. En caso de no cumplir los requisitos para ser de IDH Alto pero tampoco pertenecer al IDH Bajo, se categorizaran como IDH Medio.

Se confecciona una tabla que agrupa todas las reglas utilizadas. Se trata de un total de 17 reglas, 10 que se utilizan para categorizar un país en el clúster Alto y 7 para el clúster Bajo (tabla 4.14).

Variable	Regla		Precisión
Continent	Continent in [Europe] -->	IDH Alto	95%
Continent	Continent in [Africa] -->	IDH Bajo	43%
Urban Pop	Urban Pop >= 72.95 -->	IDH Alto	86%
Urban Pop	Urban Pop < 19.55 -->	IDH Bajo	75%
Pop<15 years	Pop<15 < 28.1 -->	IDH Alto	82%
Pop<15 years	Pop<15 > 42.65 -->	IDH Bajo	74%
Pop>65 years	Pop>65 > 8.05 -->	IDH Alto	93%
Fertility	Fertility < 2.05 -->	IDH Alto	82%
Fertility	Fertility > 5.65 -->	IDH Bajo	84%
Primary	Primary < 9.05 -->	IDH Alto	79%
Primary	Primary >= 53.7 -->	IDH Bajo	75%
Health	Health > 4.45 -->	IDH Alto	77%
Physicians	Physicians > 276.5 -->	IDH Alto	79%
Physicians	Physicians < 3.5 -->	IDH Bajo	100%
Cellular	Cellular > 609 -->	IDH Alto	96%
Internet	Internet > 170.5 -->	IDH Alto	89%
Internet	Internet < 12 -->	IDH Bajo	61%

Tabla 4.14 Reglas seleccionadas

A partir de esta tabla, se comparan las variables de cada país contra las reglas. Para un dado país, se asigna un "+1" a cada regla de IDH Alto que cumpla, y un "-1" a cada regla de IDH Bajo que cumpla. Luego, se hace la sumatoria. Si un país, por ejemplo, cumple con 5 reglas de IDH Alto y con 1 regla de IDH Bajo, su total será: 5 + (-1) = 4. Finalmente se procede a ajustar el total requerido para catalogar a un país como IDH Alto o Bajo, intentando que el error cometido sea el menor posible. El ajuste final determina que se requiere que un país tenga un total>=4 para ser de IDH Alto o un total<=-2 para ser de IDH Bajo, si no cumple ninguna de las dos, resulta de IDH Medio.

Esta configuración da como resultado la siguiente clasificación (tabla 4.15):

IDH	Población	Incorrectos	Error
Alto	70	12	17%
Medio	85	12	14%
Bajo	22	1	5%
TOTAL	177	25	14%

Tabla 4.15 Predicción de IDH

Esta tabla muestra la comparación entre la predicción del IDH con el sistema de reglas, con el IDH que el PNUD asigna a ese país, si ambos coinciden se considera correcta la predicción y si no coinciden, se considera un error. Como se puede ver, las reglas y el ajuste utilizado

logran clasificar 152 instancias correctamente de un total de 177, es decir, se comete un error del 14%. En el clúster IDH Bajo, el error desciende a 5%.

Hay que destacar que las reglas incluidas tienen como finalidad mostrar únicamente los comportamientos extremos de las variables, reflejar los límites presentes en cada variable que contribuyen a determinar que el comportamiento de ese país corresponde a uno de IDH Alto o Bajo, de manera de ser aplicable a otros países, no presentes en esta masa de datos. A continuación se presenta la herramienta de predicción con algunos ejemplos (tabla 4.16):

Country	HDI_D	Reglas																HDI Predicción
		Contin		Urban		Pop<15		Pop>65	Fertility		Primary		Health	Physic		Cellular	Internet		Resultado	
		Alto	Bajo	Alto	Bajo	Alto	Bajo	Alto	Alto	Bajo	Alto	Bajo	Alto	Alto	Bajo	Alto	Alto	Bajo	Alto + Bajo	
Sweden	Alto	1		1		1		1	1		1		1	1		1	1		10	Alto
Switzerland	Alto	1		1		1		1	1		1		1	1		1	1		10	Alto
Japan	Alto					1		1	1		1		1			1	1		7	Alto
Netherlands	Alto	1		1		1		1	1		1		1	1		1	1		10	Alto
Turkey	Medio												1				1		2	Medio
Peru	Medio										1								1	Medio
Ecuador	Medio																		0	Medio
Philippines	Medio																		0	Medio
Congo (Democr	Bajo		-1				-1			-1		-1						-1	-5	Bajo
Ethiopia	Bajo		-1		-1		-1			-1					-1			-1	-6	Bajo
Chad	Bajo		-1				-1			-1								-1	-4	Bajo

Tabla 4.16 Herramienta de predicción

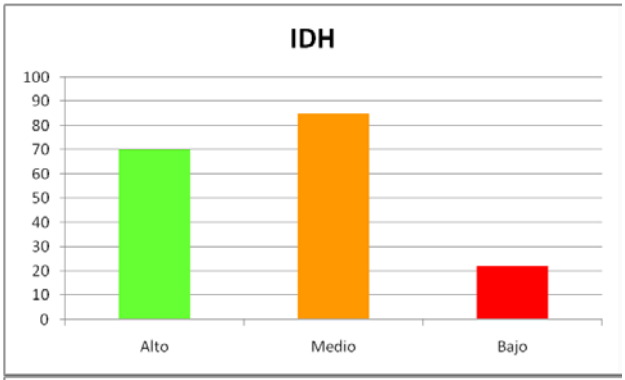
Los resultados obtenidos con la herramienta y la utilización de las reglas seleccionadas son muy satisfactorios. Se clasificaron 152 países correctamente de un total de 177, sin la utilización de las 3 variables con las que se calcula el IDH. Esto es un claro reflejo de la información disponible en las variables indirectas, de la correlación que tienen con las directas y del poder de predicción obtenido a través de las mismas.

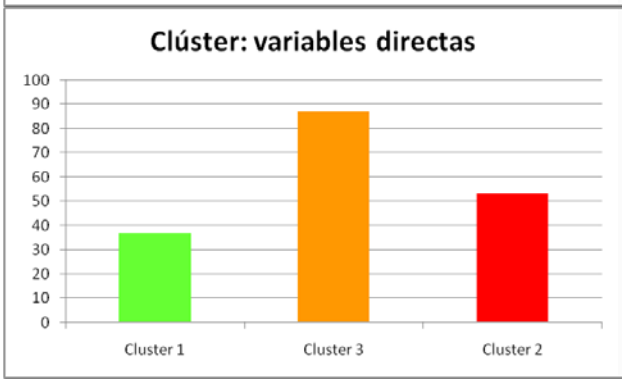
4.1.4.3. Evaluación del modelo

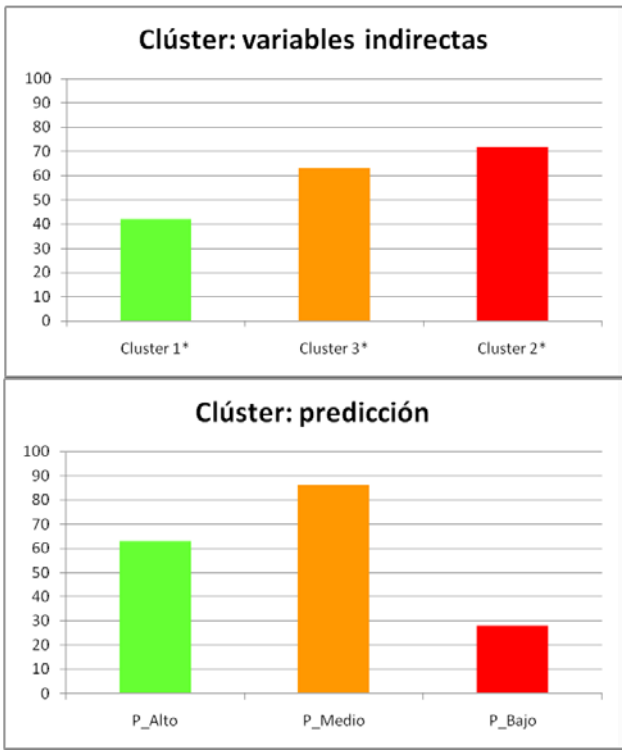
Los resultados obtenidos a lo largo de las distintas experimentaciones han sido consistentes.

En primera instancia se evaluó la incidencia y correlación de las variables directas con el IDH. Las 3 variables mostraron capacidad de clasificación individualmente, lo que demuestra su correlación con el IDH, pero más importante aún, al utilizar las 3 conjuntamente, la clasificación realizada fue casi perfecta. Esto corrobora que las 3 variables aportan información (más allá que alguna predomine por sobre las otras).

En segunda instancia se construyeron clústeres con las variables directas y luego con las indirectas. Al analizar las variables directas, el comportamiento de las 3 fue el esperado, pero el tamaño de los clústeres fue distinto al de la clasificación del IDH. Con las variables indirectas, fue posible determinar que nuevamente los 3 clústeres se construían siguiendo una escala relativa al desarrollo de los países que incluían, donde el comportamiento de todas las variables indirectas fue uniforme (definiendo claramente un clúster de desarrollo alto, uno de medio y otro de bajo). En esta ocasión, el tamaño de los clústeres también fue distinto al del PNUD lo que permite pensar que tal vez los puntos de corte establecidos no sean los que mejor reflejen la categorización que se pretende dar. A continuación se presentan los tamaños poblacionales de estas 3 agrupaciones y también una cuarta agrupación, la resultante de la predicción con las reglas seleccionadas (figuras 4.32 a 4.35).







Figuras 4.32 a 4.35 Tamaños poblacionales de los clústeres

Como se puede ver las 3 nuevas agrupaciones obtenidas coinciden en 2 efectos con respecto al IDH: el 1º clúster, aquel que agrupa países de mayor desarrollo, es menos poblado que el IDH Alto, y el 3º clúster, aquel que agrupa países con el menor desarrollo, es más poblado que el IDH Bajo. Esto se refleja tanto con variables directas, como con indirectas y también con las reglas de decisión. Los requerimientos para que un país sea de desarrollo alto son más altos, resultando en menos países en este grupo, y la cantidad de países cuyos indicadores están por debajo de la línea mínima es mayor.

En tercera instancia se analiza la correlación entre las variables indirectas y el IDH, logrando verificar que esta correlación es alta, en algunos casos mayor que con alguna de las variables directas. Al agrupar de distintas maneras estas variables, se consiguen clasificaciones del IDH con un error similar al conseguido con las variables directas, lo que denota su incidencia y utilidad.

Por último se realiza una selección de reglas obtenidas a partir de los árboles de decisión de las variables indirectas. Se utilizan para crear una herramienta de clasificación de los países, es decir, determinar el si el IDH de un país es Alto, Medio o Bajo, por medio de variables secundarias. La herramienta, luego de ser configurada muestra un poder de clasificación muy satisfactorio, logrando una predicción correcta en el 86% de los casos.

4.1.5. Fase V: Evaluación

Mediante la aplicación de los algoritmos seleccionados se ha logrado cumplimentar los objetivos planteados y se han encontrado patrones inesperados en los datos que dieron lugar a análisis más profundos y a nuevas cuestiones que posibilitan futuras líneas de investigación.

Previo al desarrollo del proyecto no se tenían conocimientos profundos sobre el tema, simplemente intuiciones sobre el comportamiento de algunas variables involucradas y de los países analizados. Mediante la minería de datos se realizó un estudio intensivo de los datos, logrando determinar el comportamiento de las variables, su influencia en los indicadores seleccionados y el desarrollo de hipótesis para explicar los resultados inesperados.

Las conclusiones principales y conocimientos adquiridos a lo largo del proceso son:

- Se corroboró la relación entre las variables directas y el IDH, tanto individual como grupalmente, mostrando una correlación mucho mayor al tratarlas grupalmente. A su vez, se determinó la preponderancia del PBI por sobre las otras dos variables, mostrando un poder de clasificación individual comparable al de las tres variables juntas. Este fenómeno permite plantear cuestiones como la utilidad del índice o su conveniencia al tener en cuenta el mayor esfuerzo que representa su confección anual comparada con el de una única variable.
- Al analizar la relación entre las variables indirectas y el IDH se encuentran casos de variables con un poder de clasificación muy alto (mejor inclusive que variables directas individualmente), es decir, una correlación muy grande con el índice. También existen casos de variables que muestran un comportamiento aleatorio, no ligado de manera directa al índice. A su vez, se logra crear conjuntos de estas variables y obtener un poder de clasificación del IDH casi tan alto como con las 3 variables directas, lo que demuestra su alta correlación y el aporte de información que cada una de ellas realiza.
- Se obtiene gran cantidad de reglas para las variables indirectas, mediante las cuales se crea un conjunto de reglas precisas, que delimitan los puntos de corte para cada una de estas variables al clasificar el índice de desarrollo humano.
- Mediante el conjunto de reglas mencionado, se desarrolla una herramienta cuya finalidad es la predicción del clúster de IDH al que pertenece un país sin la utilización de ninguna variable directa. Una vez configurada, la herramienta clasifica correctamente 152 países, de un total de 177, es decir, el 86%. Este resultado resalta la utilidad de estas variables secundarias, y su capacidad de realizar un aporte más detallado de información y aun obtener una clasificación certera.
- Mediante la herramienta de agrupamiento se obtienen 3 clústeres teniendo en cuenta las variables indirectas. Los 3 grupos están claramente diferenciados, en cada uno de ellos todas las variables incluidas se comportan de la misma manera en función de su incidencia en el desarrollo. Se obtiene un grupo, de alto desarrollo, donde las 9 variables secundarias se comportan igual (representando el estrato superior de su categoría en función de su incidencia en el desarrollo), por ejemplo: la mayor cantidad

de médicos, el mayor gasto en salud, la menor fertilidad, la mayor cantidad de celulares, etc. También se obtiene un grupo de desarrollo medio y uno de desarrollo bajo, que agrupa a los países con los peores indicadores en todas las categorías.

- Esta agrupación plantea dos cuestiones:
    - La primera es la existencia de países que siendo clasificados como IDH Alto por el PNUD o los mejores de IDH Medio, tienen un comportamiento pésimo en estas variables secundarias (no llegando siquiera a las medias poblacionales de las mismas). Si tenemos en cuenta que estas variables de alguna forma explican o aportan mayor cantidad de información de cómo se componen las variables directas, estos comportamientos resultan incoherentes y despiertan dudas sobre la veracidad de las mismas.
    - La segunda cuestión es la diferencia de cantidad de población al comparar estos clústeres con los agrupamientos del PNUD.

      Para corroborarlo, se realizan clústeres con las variables directas y este fenómeno se repite. En ambos casos se presenta un grupo de alto desarrollo menos poblado que el IDH Alto y un grupo de bajo desarrollo más poblado que el IDH Bajo. Estos fenómenos permiten suponer que los puntos de corte del IDH seleccionados por el PNUD no sean los que mejor representen el comportamiento de todos los países que contienen. Es decir que, en realidad, tal vez los requerimientos para que un país sea de alto desarrollo sean más altos (y por lo tanto el punto de corte para IDH Alto tenga que ser superior a 0.8). Esto resultaría en menos países que cumplan los requerimientos, y que por lo tanto, existen muchos más países en condiciones pésimas y que deben ser tratados como de bajo desarrollo (y por lo tanto el punto de corte para IDH Bajo ser superior a 0.5).

5. CONCLUSIONES

A través del presente proyecto se comprueba la utilidad y el valor que agrega la aplicación de sistemas inteligentes como la minería de datos al estudio del índice de desarrollo humano (IDH) confeccionado por la ONU.

Mediante la minería de datos fue posible realizar un análisis sobre el indicador incluyendo gran cantidad de variables adicionales lo que resulto en una base de datos relativamente grande pero por sobretodo compleja, por la gran cantidad de relaciones que incluye. Por otra parte, debido a que no se requiere formular una hipótesis a priori sobre los datos, no es necesario ser un experto del tema o depender de uno. El interés y estudio previo en el tema permitió llevar a cabo el proceso, contrastar las suposiciones que se tenían sobre el mismo y analizar los resultados.

En particular es importante destacar la utilidad del proyecto por la posibilidad de ser replicado. El haber llevado a cabo un proyecto de minería de datos de principio a fin permitió adquirir el know-how necesario para volver a hacerlo. Más allá del tema que se elija, se aprende a seguir la metodología seleccionada, a utilizar las herramientas y a analizar los resultados. Es decir que el proyecto resulta interesante tanto por los resultados puntuales del mismo como por la posibilidad de repetir el proceso a lo largo de la vida profesional en cualquier empresa o en el ámbito personal, cada vez que se cuente con una masa de datos de la que se desee obtener mayor conocimiento.

A través del análisis realizado se logró corroborar la relación entre el IDH y las variables que lo definen. Cuando se las tuvo en cuenta individualmente se determinó la preponderancia del PBI por sobre las otras dos, mostrando una capacidad de clasificación y una correlación mucho mayor. Este fenómeno despierta dudas sobre la utilidad del índice frente a la utilización de una sola variable (el PBI). Si bien las otras dos variables agregan información tendiente al desarrollo, son más difíciles de medir y por sobretodo, de corroborar su veracidad. Tal vez sea mejor utilizar esa sola variable antes que un índice que puede ser tergiversado por los países más controversiales (como algunos de América latina, etc.), donde justamente sería muy útil (si es confiable).

A la hora de agregar variables secundarias para profundizar el análisis, se encontró un grupo que no presenta una clara relación o implicancia en el desarrollo humano y otro con una correlación e incidencia muy importante. De esta manera se obtuvieron grupos de variables capaces de clasificar el IDH con una potencia comparable a las variables directas.

Por medio de las variables secundarias se armó un conjunto de reglas que define la incidencia de las mismas en el desarrollo humano. Luego se diseño y configuro una herramienta que utiliza este conjunto de reglas, para clasificar a los países según su IDH, logrando una clasificación correcta en el 86% de los casos, sin la necesidad de utilizar las variables directas. Este fenómeno refleja la utilidad de estas variables y la información que aportan despertando mayor interés en las mismas.

A la hora de realizar agrupamiento de los datos con las variables secundarias, se logran 3 grupos con comportamientos distintos, claramente definidos. Uno que refleja un comportamiento comparable a alto desarrollo, otro de medio y el ultimo de bajo, corroborando la tendencia esperada en estas variables. Por otra parte estos grupos se comparan con los establecidos por el PNUD, encontrando diferencias significativas en la cantidad de población en los grupos de desarrollo alto y bajo, lo que despierta dudas sobre los limites prefijados y su correcta representación de la realidad. El comportamiento se repite al utilizar otras formas de agrupamiento. Este fenómeno permite pensar que los límites reales para que un país tenga desarrollo alto son más difíciles de conseguir y por lo tanto son en realidad menos países los que logran esta categoría. A su vez, existen más países que no cumplen los requisitos mínimos para considerarse de desarrollo medio, cuyos indicadores son pésimos y por lo tanto resulta en un grupo con IDH bajo más poblado que el oficial.

Estas técnicas también despiertan dudas sobre los datos oficiales presentados por algunos países. Se encuentran países que siendo clasificados por el PNUD como de IDH Alto, pertenecen a clústeres que denotan desarrollo bajo (en función de las variables secundarias). El posterior análisis de los datos de estos países refleja que la variable directa tiene un valor excelente pero las variables secundarias que están altamente correlacionadas con la directa, muestran valores pésimos. En un indicador que aumenta su repercusión a nivel mundial y que es cada vez más importante para los gobernantes, estos comportamientos hacen dudar sobre la veracidad o precisión de los datos.

6. FUTURAS LINEAS DE INVESTIGACIÓN

La primera cuestión que surge del proyecto y despierta interés en un mayor análisis es la utilidad del IDH en comparación al PBI por sí solo. Este hecho surge de que el índice presenta una muy alta correlación con el PBI (mucho mayor que las otras variables). Debido al tiempo, costo y esfuerzo que conlleva la medición y obtención de estas 3 variables a nivel mundial, es importante corroborar que valga la pena hacerlo. Si bien el índice aporta mayor información que el PBI y las dos variables adicionales contribuyen en lo que se define como desarrollo humano, también se trata de variables más difíciles de medir y de corroborar su veracidad por parte de terceros. Resultaría interesante comparar el aporte de información que hacen estas variables (y la posibilidad de tergiversarlas) por medio del índice, contra el esfuerzo y gasto que resulta de su recolección, medición y el posterior cálculo del índice, etc.

La segunda cuestión que abre nuevas líneas de investigación son los límites prefijados para la clasificación del IDH (el punto de corte que define el paso entre IDH alto, medio o bajo). Las técnicas de agrupamiento con variables directas e indirectas presentaron diferencias significativas en este aspecto, despertando dudas sobre la conformación del grupo de desarrollo alto y el de bajo, es decir, sobre los límites que lo definen. Sería interesante analizar más profundamente el tema, para determinar si los requisitos necesarios para que un país tenga desarrollo medio no deberían ser más altos (resultando en más países en el estrato inferior) y lo mismo para el desarrollo alto (resultando en menos países en el estrato superior).

En tercera instancia, sería interesante realizar una búsqueda de información sobre las variables secundarias en países que actualmente no se encuentran en el informe de desarrollo humano. Si bien algunos países no divulgan esta información (variables directas) y por eso el índice no está calculado, tal vez pueda conseguirse información sobre las variables secundarias y utilizarla para estimar el índice de esos países.

Finalmente, se propone utilizar esta técnica para analizar otros indicadores. Ya sean los informes de desarrollo regionales, indicadores ajenos al desarrollo, o cualquier medición que se realice en el país. Además de permitir un mejor entendimiento de los mismos, es posible agregar variables relacionadas y utilizar el estudio como una manera de auditar la veracidad de los mismos.

7. REFERENCIAS

Bases de Datos (ONU), 2009. Organización de las Naciones Unidas. Disponible en http://www.un.org/spanish/databases/databases.htm, página vigente al 30/04/2009

Índice de Desarrollo Humano, 2009. Wikipedia. Disponible en http://es.wikipedia.org/wiki/Desarrollo_humano, página vigente al 30/04/2009

Metodologías para la realización de Proyectos de Data Mining, 2004. Estadistico.com. Disponible en http://www.estadistico.com/arts.html?20040426, página vigente al 30/04/2008

Process for Data Mining. CRISP 1.0 Process and User Guide. Disponible en http://www.crisp-dm.org/CRISPWP-0800.pdf, página vigente al 30/04/2009

Ranking de Desarrollo Humano, 2007. Revista Punto Suspensivo. Disponible en http://revistapuntosuspensivo.wordpress.com/2007/12/03/ranking-de-desarrollo-humano-onu/, página vigente al 13/09/2008

[Britos, P. et al, 2008]. Britos, P., Dieste, O., García-Martínez, R. 2008. Requirements Elicitation in Data Mining for Business Intelligence Projects. En Advances in Information Systems Research, Education and Practice. David Avison, George M. Kasper, Barbara Pernici, Isabel Ramos, Dewald Roode Eds. (Boston: Springer), IFIP International Federation for Information, 274: 139–150

Geoffrey Hinton, Terrence J. Sejnowski 1999. Unsupervised Learning and Map Formation: Foundations of Neural Computation, MIT Press. ISBN 0-262-58168-X

List of countries and dependencies by population density, 2009. Wikipedia. Disponible en http://en.wikipedia.org/wiki/List_of_countries_by_population_density, página vigente al 30/04/2009

C4.5 Algorithm, 2007. Wikipedia. Disponible en http://en.wikipedia.org/wiki/C4.5, página vigente al 30/04/2009.

Decision tree learning, 2007. Wikipedia. Disponible en http://en.wikipedia.org/wiki/Decision_tree_learning, página vigente al 30/04/2009

Acerca del PNUD. Programa de las Naciones Unidas para el Desarrollo. Disponible en http://www.undp.org/spanish/about/, página vigente al 30/04/2009

El concepto de desarrollo humano, Programa de las Naciones Unidas para el Desarrollo. Disponible en http://hdr.undp.org/es/desarrollohumano/, página vigente al 30/04/2009

Britos, P., Hossian, A., García-Martínez, R. y Sierra, E., 2005. Minería de Datos Basada en Sistemas Inteligentes. Editorial Nueva Librería. Buenos Aires. ISBN 987-1104-30-8.

Manfred Max Neef, 2001. Desarrollo a Escala Humana. Editorial Cepaur. Medellín. ISBN 997-4420-05-2

Algoritmos de minería de datos, 2009. Microsoft. Disponible en http://msdn.microsoft.com/es-es/library/ms175595.aspx, página vigente al 30/04/2009

Statistical classification, 2007. Wikipedia. Disponible en http://en.wikipedia.org/wiki/Statistical_classification, página vigente al 30/04/2009

Servente, M.; García-Martínez, R., 2002. Algoritmos TDIDT Aplicados a la Minería Inteligente. http://www.fi.uba.ar/laboratorios/lsi/R-ITBA-26-datamining.pdf, página vigente al 30/04/2009

López Takeyas, B, 2005. Algoritmo C4.5. Instituto Tecnológico de Nuevo Laredo. http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf, página vigente al 30/04/2009

TANAGRA Project, 2004. Ricco Rakotomalala . Disponible en http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html, página vigente al 30/04/2009

El modelo de referencia CRISP-DM, 2007. Dataprix. Disponible en http://www.dataprix.com/el-modelo-de-referencia-crisp-dm, página vigente al 30/04/2009

8. ANEXOS

Anexo A: Educción de requisitos

1. Fase 1 "Entender el dominio del proyecto"

En la tabla A1 se presentan los conceptos educidos sobre la terminología utilizada en el caso.

DEFINICIONES, ACRÓNIMOS Y ABREVIATURAS

INFORME DEFINICIONES, ACRONIMOS, ABREVIATURAS			
Analista:	*Alejandro M. Ferrari*	*Fecha*	*04.03.2009*
ID#	*Estudio del IDH*		
Termino	**Descripción**	**Tipo**	**Referencia**
CRISP-DM	*Cross Industry Standard Process for Data Mining.*	*Abreviatura*	- *Página web de CRISP-DM*
IDH	*Índice de Desarrollo Humano.*	*Abreviatura*	- *Página web del PNUD*
ONU	*Organización de las Naciones Unidas.*	*Abreviatura*	- *Página web de la ONU*
PBI	*Producto Bruto Interno.*	*Abreviatura*	
Variables indirectas o secundarias	*Aquellas variables de la base de datos que no se utilizan para calcular el IDH (ej.: superficie, población, porcentaje de actividad económica, etc.).*	*Definición*	

Tabla A.1: Informe resumen de la educción de definiciones, acrónimos y abreviaturas.

2. Fase 2 "Conocer los datos involucrados en el dominio del proyecto"

En las tablas A.2 a A.16 se presentan los conceptos educidos sobre los datos involucrados el proyecto.

OBJETIVO DEL REQUISITO

INFORME OBJETIVO DEL REQUISITO 1			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Objetivo	*Identificar la incidencia en el IDH de las variables que lo definen.*		

Tabla A.1: Objetivos del requisito 1

INFORME OBJETIVO DEL REQUISITO 2			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Objetivo	*Detectar patrones de incidencia de las variables no-directas sobre el IDH.*		

Tabla A.3: Objetivos del requisito 2

INFORME OBJETIVO DEL REQUISITO 3			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Objetivo	*Identificar grupos de comportamiento de países según su IDH en función de variables indirectas, y las reglas de causalidad que lo conforman.*		

Tabla A.4: Objetivos del requisito 3

INFORME OBJETIVO DEL REQUISITO 4			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Objetivo	*Predecir el IDH de países donde no esté calculado (por falta de alguna de las variables directas) a través de los patrones detectados en las variables no-directas.*		

Tabla A.5: Objetivos del requisito 4

ORIGEN DE LA INFORMACIÓN DEL REQUISITO

INFORME ORIGEN DE LA INFORMACION DEL REQUISITO 1 y 2			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		

Origen de la información	**Tipo**	**Responsable**	**Referencia**
Datos de Países	*Base de Datos*	*ONU*	*Página web del PNUD*

Tabla A.6: Origen de la información del requisito

SUPUESTOS DEL REQUISITO

INFORME SUPUESTOS DEL REQUISITO 1			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		

Supuesto	***Descripción***	***Referencia***
Supuesto 1	*Cuanto mayor sea la esperanza de vida, mayor será el IDH.*	*Inferencia resultante de la lectura de la web de ONU*
Supuesto 2	*Cuando mayor sea el PBI, mayor será el IDH.*	*Inferencia resultante de la lectura de la web de ONU*
Supuesto 3	*A mayor índice de alfabetización, mayor será el IDH.*	*Inferencia resultante de la lectura de la web de ONU*

Tabla A.7: Supuestos del requisito 1

INFORME SUPUESTOS DEL REQUISITO 2			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		

Supuesto	***Descripción***	***Referencia***
Supuesto 4	*Cuanto mayor sea el porcentaje de población urbana, mayor será el IDH.*	*Inferencia resultante de la lectura de la web de ONU*
Supuesto 5	*Cuanto menor sea el porcentaje de actividad económica primaria, mayor será el IDH.*	*Inferencia resultante de la lectura de la web de ONU*

Tabla A.8: Supuestos del requisito 2

INFORME SUPUESTOS DEL REQUISITO 3			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*

ID#	*Estudio del IDH*	
Supuesto	***Descripción***	***Referencia***
Supuesto 5	*Existen grupos de países (resultantes de subdivisiones geográficas de continentes) con niveles de IDH muy similares.*	*Inferencia resultante de la lectura de la web de ONU*
Supuesto 6	*Los grupos de países con igual IDH pueden ser definidos a través de las variables no-directas.*	*Inferencia resultante de la lectura de la web de ONU*

Tabla A.9: Supuestos del requisito 3

INFORME SUPUESTOS DEL REQUISITO 4			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Supuesto	***Descripción***		***Referencia***
Supuesto 7	*A través de los patrones de comportamiento identificados en las variables no-directas, se podrá estimar el IDH en países donde no haya sido calculado.*		*Inferencia resultante de la lectura de la web de ONU*

Tabla A.10: Supuestos del requisito 4

ATRIBUTOS INVOLUCRADOS EN EL REQUISITO

INFORME ATRIBUTOS INVOLUCRADOS EN EL REQUISITO 1			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		

Atributo	***Fuente***	***Referencia***
PBI	*Base de datos*	*Pagina web de la ONU*
Esperanza de Vida	*Base de datos*	*Pagina web de la ONU*
Índice de Alfabetización	*Base de datos*	*Pagina web de la ONU*

Tabla A11: Atributos involucrados del requisito

INFORME ATRIBUTOS INVOLUCRADOS EN EL REQUISITO 2, 3 y 4			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		

Atributo	***Fuente***	***Referencia***

Continente	*Base de datos*	*Pagina web de la ONU*
Superficie	*Base de datos*	*Pagina web de la ONU*
Población	*Base de datos*	*Pagina web de la ONU*
Densidad de Población	*Base de Datos*	*Pagina web de la ONU*
Porcentaje de Población Urbana	*Base de Datos*	*Pagina web de la ONU*
Porcentaje de Actividad Económica	*Base de Datos*	*Pagina web de la ONU*

Tabla A.12: Atributos involucrados del requisito

3. Fase 3 "Comprender los objetivos del proyecto"

En las tablas A.13 a A.16 se presentan los conceptos educidos sobre los objetivos del proyecto.

OBJETIVO DEL PROYECTO

INFORME **OBJETIVOS DEL PROYECTO**				
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*	
ID#	*Estudio del IDH*			
Objetivo	***Descripción***			***Referencia***
Objetivo 1	*Realizar un estudio de minería de datos sobre el IDH, teniendo en cuenta las variables con que se calcula como también otras variables indirectas para detectar patrones, armar grupos de comportamiento y establecer la influencia de variables no-directas.*			*Inferencia resultante de la lectura de la web de ONU*

Tabla A.13: Objetivos del proyecto

CRITERIOS DE ÉXITO

INFORME **CRITERIOS DE EXITO DEL PROYECTO**				
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*	
ID#	*Estudio del IDH*			
Criterio	**Descripción**	**Objetivos asociados**		**Referencias**
Criterio 1	*Descubrir indicadores que demuestren como influyen las*	- *Objetivo 1*		Inferencia resultante

	variables no-directas en el IDH.		de la lectura de la web de ONU
Criterio 2	*Determinar la correcta correlación entre el IDH y las variables que lo definen directamente.*	- *Objetivo 1*	Inferencia resultante de la lectura de la web de ONU
Criterio 3	*Establecer límites cuantitativos a las variables no-directas que permitan describir los patrones de comportamiento de estos grupos.*	- *Objetivo 1*	Inferencia resultante de la lectura de la web de ONU
Criterio 4	*Lograr predecir correctamente el IDH de países en que este indicador sea conocido (contrastando el predicho contra el real).*	- *Objetivo 1*	Inferencia resultante de la lectura de la web de ONU

Tabla A14: Criterios de éxito del proyecto

EXPECTATIVAS DEL PROYECTO

INFORME EXPECTATIVAS DEL PROYECTO			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Expectativa	**Descripción**	**Objetivos asociados**	**Referencias**
Expectativa 1	*Determinar el IDH de aquellos países en que no pueda ser calculado, para rankearlos internacionalmente.*	- *Objetivo 1*	Inferencia resultante de la lectura de la web de ONU

Tabla A.15: Expectativas del proyecto

SUPOSICIONES DEL PROYECTO

INFORME SUPOSICIONES DEL PROYECTO			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		
Suposición	**Descripción**	**Objetivos asociados**	**Referencias**
Suposición 1	*Una vez identificada la relevancia de las variables no-directas en el IDH, se podrán utilizar para estimarlo en los países donde no está calculado.*	- *Objetivo 1*	Inferencia resultante de la lectura de la web de ONU

Tabla A.16: Suposiciones

4. Fase 4 "Identificar a los recursos humanos involucrados"

En la tabla A.17 se presenta los recursos involucrados en el proyecto.

RECURSOS HUMANOS INVOLUCRADOS

INFORME RECURSOS HUMANOS INVOLUCRADOS					
Analista:	*Alejandro M. Ferrari*		Fecha:	*04.03.2009*	
ID#	*Estudio del IDH*				
Posición	**Perfil de la posición**	**Pertenece a:**	**Coordenadas**		
			Apellido y Nombre	**E-mail**	**TE**
Tutor del Proyecto	Es quien tiene la experiencia en este tipo de proyectos, el experto en la explotación de información y supervisor de las tareas a realizarse.	ITBA	*Britos Paola V.*	*paobritos@gmail.com*	*63934800 int. 5841*
Líder del proyecto	Es quien conduce el proyecto de explotación de información y realiza las tareas para cumplir los objetivos.	ITBA	*Ferrari Alejandro M.*	ferrarialejandro@hotmail.com	*4823-2368*

Tabla A.17: Informe resumen de la educción de recursos humanos involucrados

5. Identificación de Riesgos y Contingencias para proyectos de explotación de información

En las tablas A.18 a A.20 se presentan los riesgos y contingencias involucrados en el proyecto.

IDENTIFICACION DE RIESGOS Y CONTIGENCIAS

Identificación de riesgos

INFORME IDENTIFICACION DE RIESGOS			
Analista:	*Alejandro M. Ferrari*	Fecha:	*04.03.2009*
ID#	*Estudio del IDH*		

	Riesgo	Severidad	Frecuencia	Tipo de Riesgo	Anexo
	▪ SI ▪ NO	▪ catastrófica ▪ crítica ▪ seria ▪ menor ▪ sin importancia	▪ frecuente ▪ probable ▪ ocasional ▪ remota ▪ improbable	▪ intolerable ▪ alta ▪ media ▪ baja ▪ sin importancia	Numero de Anexo en el que van aclaraciones
A. Ingeniería del proceso					
1. REQUERIMIENTOS					
a. La estabilidad ¿Las necesidades del negocio están cambiando constantemente?	No	---	---	---	---
b. La integridad ¿Los requisitos están especificados de forma incompleta?	No	---	---	---	---
c. La claridad ¿Los requisitos son inciertos o poco claros en su interpretación?	No	---	---	---	---
d. La validez ¿Los datos a utilizar llevarán a que el cliente cumpla con los objetivos propuestos?	Si	---	---	---	---
e. La viabilidad ¿Hay requisitos que son técnicamente difíciles llevar a cabo?	No	---	---	---	---
f. Precedente ¿Los requisitos nunca especifican algo hecho antes o más allá de la experiencia de personal?	No	---	---	---	---
2. DATOS					
a. La calidad					

¿Hay algún problema con en la completitud de los datos?	Si	Menor	*Probable*	*Media*	---
b. La cantidad ¿Existen problemas de cantidad de datos (pocos datos representativos)?	No	---	---	---	---
c. La performance ¿Existen problemas en el hardware o software a utilizar para la explotación de datos?	No	---	---	---	---
d. La validación ¿Los datos tomados como validación se encuentran en condiciones completas?	Si	---	---	---	---
e. La ubicación ¿Se encuentran los datos integrados en un mismo lugar o hay que unificarlos?	No	Menor	*Probable*	*Media*	---

B. Las restricciones

1. LOS RECURSOS

a. El equipo de trabajo ¿El personal es inexperto, le falta conocimiento del dominio o habilidades?	No	---	---	---	---
b. El presupuesto ¿Existe el riesgo de discontinuidades en el flujo de fondos del proyecto?	No	---	---	---	---
c. Los recursos técnicos ¿Los medios son inadecuados para construir y entregar lo planificado?	No	---	---	---	---
d. Inestabilidad de personal ¿El personal involucrado en el proyecto es inestable?	No	---	---	---	---

2. CONTRACTUALES

a. El tipo de contrato ¿El tipo del contrato es una fuente de riesgo al proyecto?	No	---	---	---	---
b. Las restricciones ¿El contrato incluye alguna restricción impropia?	No	---	---	---	---
c. Las dependencias ¿El programa tiene alguna dependencia crítica con proyecto externos o internos?	No	---	---	---	---

Tabla A.18: Plantilla de identificación de potenciales riesgos del proyecto de explotación de información

Identificación de contingencias

INFORME IDENTIFICACION DE CONTIGENCIAS			
Analista:	*Alejandro M. Ferrari*	Fecha informe:	*04.03.2009*
ID proyecto#	*Estudio del IDH*		
ANEXO:	*1*	Fecha identificación:	*25.02.2009*
EXPERTO:	*Paola V. Britos*	ID riesgo#:	*Completitud de datos*
Descripción del riesgo:	*Existen ciertos datos faltantes y falta obtener un atributo para todos los países (índice de alfabetización).*		
Estrategia de la mitigación:	*Los datos faltantes se buscarán en versiones anteriores de la base de datos para ver si existían o si nunca estuvieron presentes.* *Se intentará conseguir el atributo faltante en la página web de la ONU.*		
Acción de contingencia:	*Se reemplazaran los datos faltantes por alguna variable (valor cero o negativo) que los distinga de los que si existen para dejarlos fuera del análisis.* *Si el atributo no puede conseguirse, se dejara de lado y se analizara el IDH teniendo en cuenta las otras 2 variables que lo definen.*		
ID# de riesgos que impacta:	*Ubicación de datos*		

Tabla A.19 : Identificación de contingencias para el riesgo 1

INFORME IDENTIFICACION DE CONTIGENCIAS			
Analista:	*Alejandro M. Ferrari*	Fecha informe:	*04.03.2009*
ID proyecto#	*Estudio del IDH*		
ANEXO:	*2*	Fecha identificación:	*25.02.2009*
EXPERTO:	*Paola V. Britos*	ID riesgo#:	*Ubicación de datos*
Descripción del riesgo:	*El atributo "Índice de alfabetización" no se encuentra en la base de datos actual sino que debe ser obtenido de alguna otra base de datos de la ONU.*		
Estrategia de la mitigación:	*Una vez conseguido el atributo, se agregaran estos datos a la base de datos actual, intentando que el resultado sea homogéneo para todos los países.*		
Acción de contingencia:	*En caso de que el atributo no exista para todos los países (datos incompletos) se procederá de la misma manera que en el riesgo anterior, utilizando algún valor que deje estos campos fuera del análisis.*		
ID# de riesgos que impacta:	---		

Tabla A.20: Identificación de contingencias para el riesgo 2