

# Uso de Técnicas de Inteligencia Artificial para el Análisis del Impacto de Ambientes Contaminantes en el Índice de Daño Genético Humano

» Jorge Kamlofsky<sup>1</sup>, Vanesa Miana<sup>2</sup>, Elio Prieto Gonzalez<sup>2</sup>

<sup>1</sup>CAETI – <sup>2</sup>CAECIS, Universidad Abierta Interamericana, Argentina  
jorge.kamlofsky@uai.edu.ar, vanesa.miana@uai.edu.ar, elio.prieto@uai.edu.ar

## Resumen

Las técnicas de Inteligencia Artificial (IA) hoy están difundidas en casi todas las disciplinas. En el ámbito de la salud, se las aplica en etapas operacionales de la investigación: sobre bancos de datos se pueden presentar modelos cuya validación se plasma en nuevo conocimiento científico. Sin embargo, en investigaciones específicas, los investigadores deben recopilar sus datos. Estas investigaciones son costosas, por lo que muchas veces, con resultados preliminares basados en pocos datos, se define si se avanza con la investigación o no.

En este trabajo se presenta las tareas que permiten obtener un modelo que permite describir y predecir el impacto en el daño genético evaluado mediante la técnica del ensayo cometa. Este trabajo se basó en el análisis de 54 casos. Se obtuvieron modelos de regresión lineal múltiple previo a un proceso de selección de variables basado en la Teoría de la Información de Shannon (1948). Los modelos obtenidos se evaluaron con el indicador  $R^2$ . Si bien el evaluador obtenido no se encuentra en los niveles recomendables, es suficiente para presentar indicios interesantes.

PALABRAS CLAVE: DATA ANALYSIS, DATA MINING, MINERÍA DE DATOS, ANÁLISIS DE DATOS, DESCUBRIMIENTO DE CONOCIMIENTO, INTELIGENCIA ARTIFICIAL, CONTAMINACIÓN COMET ID BASAL.

## Use of Artificial Intelligence Techniques for the Analysis of the Impact of Polluting Environments on the Human Genetic Damage Index

### Abstract

Artificial Intelligence (AI) techniques are now widespread in almost all disciplines. In the field of health, they are applied in operational stages of research: normally based on databases, models

can be presented whose validation is reflected in new scientific knowledge. However, in specific investigations, researchers must collect their data. These investigations are expensive, so often, with preliminary results based on few data, it is defined whether the investigation is progressed or not.

This paper presents the tasks that allow us to obtain a model that allows us to describe and predict the impact on the genetic damage evaluated by the comet assay technique. This paper was based on the analysis of 54 cases. Multiple linear regression models were obtained prior to a variable selection process based on Shannon's Theory of Information (1948). The obtained models were evaluated with indicator  $R^2$ . Although the evaluator obtained is not at the recommendable levels, it is sufficient to present interesting indications.

---

KEYWORDS: DATA ANALYSIS, DATA MINING, MINERÍA DE DATOS, ANÁLISIS DE DATOS, DESCUBRIMIENTO DE CONOCIMIENTO, INTELIGENCIA ARTIFICIAL, CONTAMINACIÓN COMET ID BASAL.

## 1. Introducción

### 1.1. Trabajos relacionados:

La revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, y transmitir. Gracias al gran progreso en informática y tecnologías relacionadas y la expansión de su uso en diferentes aspectos de la vida, se continúa recogiendo y almacenando en bases de datos gran cantidad de información (Mitra & Acharya, 2005). En los últimos años, el desarrollo informático llega hasta los sistemas domésticos. En la llamada internet de las cosas, es común ver en los hogares gran variedad de dispositivos conectados a internet: teléfonos móviles, heladeras, televisores, equipos de aire acondicionado, cocinas y hornos entre otros (Gustafson & Sheth, 2014), los cuales incrementan exponencialmente la cantidad de dispositivos que obtienen datos de todo tipo, los transmiten y almacenan. La multiplicidad de orígenes de datos junto con el gran volumen generado conforman al término "Big-Data".

Los datos en bruto raramente son beneficiosos directamente. Entre otras causas, su verdadero valor se basa en la habilidad para extraerles información útil para la toma de decisiones (Riquelme, Ruiz & Gilbert, 2006). Un dato o conjunto de datos, solamente pueden ser considerados información cuando se los combinan en respuestas o informes significativos que puedan ayudar a la interpretación de eventos (Dyché, 2000). Es decir, un conjunto de datos de por sí carece de importancia en el aporte de conocimiento. Puede transformarse en información útil o relevante tras su procesamiento y/o análisis, permitiendo sacar a la luz información o evidencias del fenómeno estudiado.

El procesamiento de datos tiene fundamentos puramente matemáticos y usa herramientas informáticas. Se contraponen tanto en sus métodos como en su nomenclatura dos enfoques: Minería de Datos (MD) y Estadísticas (Aluja, 2001).

La MD se define como un *proceso iterativo de extraer patrones predictivos ocultos de grandes conjuntos de datos, usando tanto técnicas de Inteligencia Artificial, como técnicas estadísticas* (Mena, 1999). La Inteligencia Artificial (IA) es un término utilizado para describir el desarrollo de procesos similares a los de la inteligencia humana mediante herramientas informáticas. El termino

fue propuesto en 1956 por John McCarthy (McCarthy & Hayes, 1981). Los procesos considerados son: el aprendizaje, definido como la adquisición de información y de las reglas para su utilización, el razonamiento que mediante el empleo de reglas o algoritmos permite conducir a definiciones o conclusiones parciales o definitivas, y además su capacidad de autocorrección (Mc Corduck, 2004). El término Descubrimiento del Conocimiento (KDD según sus siglas en inglés) a menudo se lo trata como sinónimo de MD. En general está aceptado que MD se refiera a una etapa dentro del proceso KDD que consiste en la aplicación de algoritmos específicos para extraer patrones o modelos de los datos. Hay otras etapas en el proceso KDD: la preparación de los datos, la selección y limpieza de los mismos, la incorporación de conocimiento previo, y la propia interpretación de los resultados de la minería. Estas etapas implementadas de una manera iterativa e interactiva permiten que se les extraiga conocimiento útil a los datos. Hoy es de suma importancia de incluir en la metodología el preproceso de los datos, y la formalización del conocimiento descubierto (Riquelme, Ruiz & Gilbert, 2006).

La Estadística se la define como la rama de la Matemática aplicada a datos observacionales. Trata el estudio de poblaciones, sus variaciones, y los métodos de reducción de datos (Fisher, 1925). Otra definición expresa que es la metodología que permite extraer información a partir de un conjunto de datos expresando la incerteza (Rao, 1989). Se la divide en dos ramas: la Estadística Descriptiva (Análisis de Datos) y la Estadística Inferencial (Modelado de Datos). Mientras que en la Estadística Descriptiva se busca manifestar la información relevante para los problemas planteados partiendo de los datos, en la Estadística Inferencial se estudia decidir entre varias hipótesis a partir de las consecuencias observadas incorporando la aleatoriedad dentro de la decisión (Aluja, 2001).

Muchos de los problemas abordados en Estadísticas son comunes con la IA, pero su abordaje es diferente: mientras que la IA trata de ofrecer soluciones algorítmicas con un costo computacional aceptable, la Estadística busca la generalización de los resultados, esto es, poder inferir los resultados a situaciones más generales que la estudiada (Aluja, 2001).

Las técnicas de MD originalmente se implementaron en mercados como el Financiero, Seguros, Retail, Marketing, donde el análisis de millones de transacciones mediante estas técnicas permitió la obtención de conocimiento nuevo que permitió brindar mejores servicios a sus clientes. Con el evidente éxito luego su aplicación se difundió hacia otras disciplinas. En ámbitos científicos también se plantean nuevos problemas donde la MD se torna imprescindible, como ser las investigaciones surgidas del proyecto Genoma, o las investigaciones realizadas por el CERN con datos provenientes del acelerador de partículas. Estas investigaciones tienen en común la necesidad de tratar con conjuntos de datos complejos y de enorme tamaño, los cuales no pueden tratarse mediante la Estadística Clásica, o bien, los resultados así obtenidos, carecen de riqueza o interés. La MD resulta la opción más adecuada a estos desafíos.

La MD es un campo interdisciplinar con el objetivo general de predecir las salidas y revelar relaciones entre los datos. Para ello se utilizan herramientas automáticas que emplean algoritmos sofisticados para descubrir principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en los data-warehouses u otros repositorios de información, y además, filtran la información necesaria de las grandes bases de datos (Mittra & Acharya, 2005). En el ámbito de la Salud, su implementación es muy amplia: desde el uso de estas técnicas de marketing para mejorar la atención al público en centros de salud pública

(Beerli-Palacio, Santana y Porta, 2008), hasta la investigación bio-médica. Hoy muchas investigaciones médicas se basan en el uso de grandes bancos de datos y bases documentales (Ospina, Reveiz & Cardona, 2005) como MedlinePlus<sup>1</sup> producido por la Biblioteca Nacional de Medicina de EEUU sobre los cuales se pueden implementar las técnicas de MD previamente mencionadas. La variedad y flexibilidad de los algoritmos permite que pueda disponerse de herramientas para tratar investigaciones de características muy diversas.

Estos son estudios de diseño retrospectivo: los datos existen con anterioridad a la planificación del estudio (Castillo, 2011). Son económicos y convenientes, ya que en general los datos son de calidad y se los dispone en gran cantidad, lo cual es apropiado para su análisis, y además, muchas veces se los dispone gratuitamente. En estudios prospectivos, la recolección de datos se realiza luego de la planificación del estudio. En éstos, es ventajoso que se tiene control de los aparatos y de las mediciones, pero las principales desventajas radican en su elevado costo, y el largo tiempo de recolección. Las encuestas son muy usadas en estudios prospectivos y observacionales (no experimentales). Los sujetos son observados una única vez: gracias a ello, a cada sujeto se le trata de extraer la mayor cantidad de información.

En estudios exploratorios, prospectivos y observacionales mediante el uso de encuestas, sucede entonces, que los conjuntos de datos suelen ser desproporcionalmente anchos y poco profundos: con gran cantidad de variables, y poca cantidad de casos. Este escenario hace que sea complicado obtener modelos robustos, tanto por la gran cantidad de variables como por la escasez de casos, pero muchas veces suficiente para obtener indicios que permitan confirmar o no una línea de investigación.

En este trabajo se implementan algunas técnicas de IA sobre una encuesta hecha a un pequeño número de pacientes residentes en zonas periféricas de la Provincia de Buenos Aires (Argentina) con la intención de obtener conocimiento preliminar, que relacione a la obesidad, la contaminación y toxicidad ambiental con el daño genético humano.

### 1.2. *Objetivos del Trabajo*

El objetivo principal de este trabajo es presentar un método para obtener nueva información de carácter preliminar en forma de modelos matemáticos, respecto al daño en el material genético humano a causa de contaminantes tanto del ambiente como de los hábitos alimenticios y tóxicos de los encuestados.

Un objetivo subyacente es el de compartir la experiencia de un trabajo interdisciplinario entre investigadores del sector médico y del sector de las tecnologías de la información.

### 1.3. *Motivación y Alcance*

#### *Motivación*

Cuando se tratan problemas de análisis de datos, se pone énfasis en la cantidad de datos: se pretende que la cantidad de datos experimentales sea tan grande como se pueda. En este caso, a partir de una cantidad muy pequeña de datos, y con gran desarrollo en anchura, se pretende obtener modelos que brinden indicios para el estudio de nuevas hipótesis.

---

<sup>1</sup> <http://www.nlm.nih.gov/medlineplus/>

## Alcance

En este trabajo se presenta el método para la obtención de modelos de regresión lineal múltiple con su correspondiente evaluador. En este trabajo no se presentan explícitamente los modelos, los cuales se presentarán en un trabajo específico.

### 1.4. Contenido y Organización de este Trabajo

La sección segunda presenta el Estado del Arte acerca de las técnicas de inteligencia artificial y de la Genotoxicidad y el estudio de la enfermedad neoplásica. La tercera sección presenta las características del modelo experimental que se utiliza en este trabajo. La cuarta sección presenta la implementación de los algoritmos utilizados y los resultados obtenidos. Luego se presentan las conclusiones y los trabajos futuros.

## 2. Marco Teórico

### 2.1. Descubrimiento de Conocimiento Nuevo a partir de un Conjuntos de Datos

#### Presentación

Frecuentemente se utilizan como sinónimos: MD y Análisis de Datos que suele hacer hincapié en las técnicas de análisis estadístico. Otro término muy utilizado es el de extracción o descubrimiento del conocimiento (KDD: siglas del término en inglés: Knowledge Discovery in Databases). En muchas ocasiones, ambos términos se utilizan indistintamente. Últimamente, se acepta que KDD se refiere a un proceso que consta de una serie de fases, mientras que MD es solo una de esas fases.

En (Fayyad, Piatetski & Smyth, 1996) se define KDD como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos. De esta definición, las propiedades deseables del KDD quedan resumidas:

*Válido:* Se refiere a que los patrones deben seguir siendo precisos para datos nuevos, y no solo para los datos que han sido usados para su obtención.

*Novedoso:* Que aporte algo desconocido tanto para el sistema como para el usuario.

*Potencialmente útil:* Los usuarios deben recibir algún tipo de beneficio a partir de la información obtenida.

*Comprensible:* La información comprensible proporciona conocimiento.

La Figura 1 (obtenida de Hernandez Orallo et al, 2004) presenta un esquema basado en la definición previamente presentada, donde se muestra el proceso de KDD: descubrimiento del conocimiento a partir de un conjunto de datos.

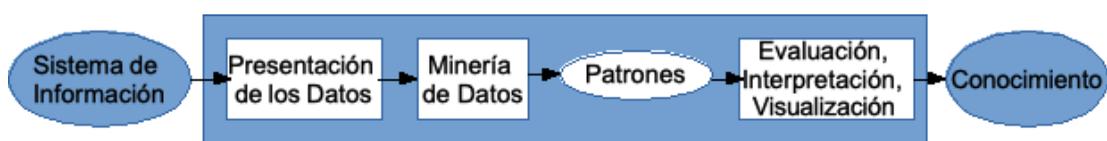


Fig. 1: Etapas en el Proceso de KDD

Así, los sistemas de KDD permiten la selección, extracción, limpieza, transformación y proyección de los datos, analizar los datos para extraer patrones y modelos adecuados, evaluar e interpretar los patrones para convertirlos en conocimiento, resolviendo conflictos con conocimiento previo. Todo esto clarifica la relación entre KDD y MD: mientras KDD es un proceso global, MD se refiere a la aplicación de los algoritmos y métodos para la obtención de patrones y modelos.

Durante el proceso KDD pueden identificarse las siguientes fases (Hernandez Orallo et al, 2004), las cuales se describen a continuación: Pre-procesamiento de datos, MD, evaluación de los modelos y difusión, uso y monitoreo de resultados. La fase de MD tiene como objetivo analizar los datos para extraerles conocimiento, en forma de relaciones, patrones, reglas, etc. (Hernandez Orallo et al, 2004). Las tareas propias de la fase de MD pueden ser descriptivas, por ejemplo: descubrir patrones interesantes o relaciones que describen los datos, o predictivas, por ejemplo: clasificar nuevos datos basándose en datos anteriormente disponibles.

Los modelos y patrones obtenidos de la etapa de MD deben ser medidos y evaluados. Y cada algoritmo presenta indicadores que así lo permiten. Finalmente, los resultados deben ser interpretados y contextualizados. El nuevo conocimiento debe ser entonces, presentado y difundido (Hernandez Orallo et al, 2004).

### *El Pre-Procesamiento de la Información*

Rara vez los datos obtenidos en bruto se encuentran en el formato adecuado, sin errores, completos y con etiquetas y codificado adecuadamente. Se transforman los datos para darles un formato común. Frecuentemente se los recopila en bancos de datos donde se consigue unificar toda la información. Se debe encontrar y tratar datos faltantes, erróneos y anomalías. A esta etapa se la llama también Extracción-Transformación-Limpieza o etapa ETL (según sus siglas). Con esto se obtiene lo que se conoce como “vista minable” que es el conjunto de datos listo para aplicarle algoritmos de MD.

El pre-procesamiento de los conjuntos de datos es un proceso esencial de KDD. De hecho en la práctica, habitualmente esta tarea ocupa más tiempo que el minado de datos (De Jonge & Van Der Loo, 2013). Las tareas más importantes en el pre-procesamiento de la información son:

*Formateo de datos:* Para el correcto tratamiento por parte de los algoritmos de MD es necesario que tengan el formato adecuado. Por ejemplo, los números deben tener formato numérico y no de string.

*Detección de errores de carga o inconsistencias:* Con funciones sencillas es posible identificar muchos de ellos. Por ejemplo, en una persona, se detecta un error si Edad (en años) > 150, o si Altura (en metros) > 2,50. Otros pueden requerir condiciones más complejas. Un ejemplo: un dato inconsistente: Embarazo='SI' y Sexo='Masculino'.

*Detección de Datos Vacíos:* Muchos algoritmos pueden soportar datos vacíos sin problema, pero otros no. Más allá de soportarlos o no, un dato vacío puede quitar mucha información valiosa y generar ruido. Es importante considerar como se los trata. Y frente a ello, la sugerencia es consultar a los investigadores para obrar en consecuencia.

*Outliers o valores anómalos:* Dependiendo el tipo de estudio que se está llevando a cabo, el tratamiento de outliers puede ser crítico. Mientras que en algunas aplicaciones los valores anómalos

generan ruido y es bueno eliminarlos, en otras, su detección es primordial, por ejemplo, en los sistemas de detección de fraudes. Detectar outliers puede lograrse fácilmente usando funciones estadísticas. Por ejemplo, en R:<sup>2</sup> `boxplot.stats(x)$out`.

*Normalización:* La reducción de escala para crear un rango más pequeño suele ser recomendado. En análisis multivariable, si no se normaliza las variables, pueden causar errores por el peso que la variable que trae consigo.

### *Minería de Datos*

MD es un conjunto de algoritmos matemático / computacionales que permiten extraer información no trivial y potencialmente útil que reside implícitamente en los datos analizados. Es también, el área de estudio que le otorga a las computadoras la habilidad de aprender sin ser explícitamente programadas. MD tiene como objetivo analizar los datos para extraerles conocimiento, en forma de relaciones, patrones, reglas, etc.

La MD se usa en un sin número de implementaciones de uso cotidiano, sin que uno sea consciente de ello: cuando se usa un buscador como Google o Yahoo, cuando un antivirus monitorea la navegación en Internet, cuando se solicita un crédito bancario o se contrata un seguro, aplicaciones de reconocimiento de voz o visión artificial. Entes gubernamentales y/o empresas de seguridad o de tarjetas de crédito usan MD cuando monitorean patrones y comportamiento atentos a comportamientos anómalos, entre otros.

### *Tipos de Modelos de MD*

La minería de datos tiene como objetivo analizar los datos para extraer conocimiento en forma de modelos surgidos de los datos analizados. En la práctica, los modelos pueden ser del tipo predictivos y descriptivos:

*Modelos Predictivos:* pretenden estimar valores futuros o desconocidos de variables de interés que denominaremos “variables objetivo o independientes” usando otras variables o campos de las bases de datos. Por ejemplo, un modelo predictivo puede ser aquel que permite estimar la demanda de un nuevo producto en función de la publicidad realizada.

*Modelos Descriptivos:* identifican patrones que explican los datos. Es decir, sirven para explorar las propiedades de los datos examinados. Por ejemplo: una agencia de viajes desea identificar grupos de personas con gustos similares con el objetivo de realizar diferentes ofertas para cada grupo.

Algunas tareas de la MD que producen modelos predictivos son la clasificación y la regresión, mientras que las tareas que producen modelos descriptivos son el agrupamiento, las reglas de asociación y el análisis correlacional. Cada tarea puede realizarse usando diferentes algoritmos o técnicas de MD por ejemplo, arboles de decisión, redes neuronales, redes bayesianas, etc. (Hernandez Orallo et al, 2004).

Los algoritmos de MD también pueden clasificarse en función de su necesidad o no de supervisión:

*Modelos Supervisados:* Existe una variable a predecir (normalmente llamada “objetivo” o “target”). Se conoce el resultado hacia donde se quiere llegar, pero no el camino de cómo conseguirlo. Por ejemplo: Modelos de OCR en scanners.

*Modelos No Supervisados:* No existe ninguna variable a predecir. Se conoce el camino para llegar y los datos históricos pero no se conoce el resultado futuro. Por ejemplo: predicción de la demanda, predicción de la Bolsa de Comercio.

*Modelos Semi-Supervisados:* Muchas veces se conoce la etiqueta (es decir: el valor que toma la variable target a predecir) de algunos casos en el conjunto de datos, pero no de otros. Generalmente se presentan muchos casos sin etiquetar, y pocos casos con etiquetas. De esta manera es posible clasificar aquellos casos sin etiquetas en base a aquellos pocos que se encuentran etiquetados.

### *Tipos de Algoritmos de Minería de Datos*

A continuación se presentan los principales tipos de algoritmos de MD:

*Cluster Analysis:* Segmenta elementos que son similares en algún sentido. Se aplica ampliamente en Marketing y Empresas, que desean segmentar el comportamiento de sus clientes y productos en el mercado y realizar marketing directo y personalizado. Son modelos descriptivos.

*Clasificación:* Cada registro posee un atributo de clase que se obtiene a partir de los otros. El objetivo es predecir la clase de nuevos registros. Son modelos predictivos.

*Regresión:* Consiste en aprender una función lineal que asigna a cada instancia un valor real. El objetivo en este caso es minimizar el error entre el valor predicho y el valor real. Son modelos predictivos.

*Reglas de asociación:* Son métodos para descubrir nuevas relaciones no explícitas entre variables de una gran base de datos. Son muy utilizadas en supermercados y en publicidad. Son modelos descriptivos.

*Árboles de Decisión:* Se construyen diagramas lógicos en formas de árbol, leyéndose de arriba hacia abajo, hasta llegar a una decisión (hojas). Son modelos de predictivos.

*Redes Neuronales:* Son modelos que imitan el funcionamiento del cerebro humano. Permite modelar problemas complejos en los que puede haber interacciones no lineal entre las variables. Son modelos predictivos. La Figura 2 muestra un modelo de red neuronal.

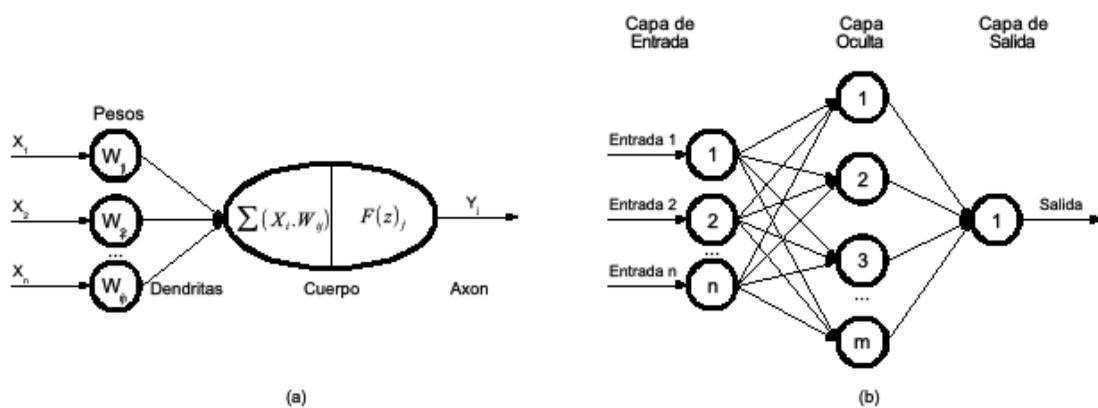


Fig. 2: (a) El modelo básico de Neurona Artificial o Perceptron. (b) Modelo de Red Neuronal Artificial.

**Inferencia Bayesiana:** Generan modelos ampliamente utilizados en la Teoría de la Decisión. En base a evidencias (nuestros datos) intenta predecir que una hipótesis sea cierta. Por ejemplo si alguien va a sobrevivir o no, de un determinado tumor, si un cliente va a comprar un producto o no. Se basa el Teorema de Bayes. Son modelos predictivos.

**Regresión Logística:** Es un modelo ampliamente utilizada en modelos de Scoring, para el análisis de riesgos. Algunos sectores que lo utilizan: Seguros, Financiero (otorgamiento de créditos). Genera modelos de predicción basados en la probabilidad que suceda un evento, ajustando los datos a la función logística. Son modelos predictivos.

Pueden obtenerse más detalles en (Hernandez Orallo et al, 2004) entre otros textos.

### Validación, Evaluación e Interpretación de Modelos

Para entrenar y probar (o validar) un modelo se divide al conjunto de datos en dos: un conjunto de entrenamiento y un conjunto de prueba. Esto es necesario para garantizar que la validación de la precisión del modelo sea una medida independiente. Mientras que el conjunto de entrenamiento se lo usa para la construcción del modelo, el conjunto de prueba se lo usa para validar al modelo obtenido.

**Validación simple:** Se reserva un porcentaje del conjunto de datos como conjunto de prueba. El porcentaje suele ser entre 5% y 50% del conjunto original, separado aleatoriamente.

**Validación cruzada:** Se usa en casos donde la cantidad de datos es muy pequeña (y no se desea perder datos para prueba). Los datos se dividen aleatoriamente en dos conjuntos. Con el primero se elabora el modelo y se lo prueba en el segundo conjunto. Luego se elabora otro modelo con el segundo conjunto y se lo valida contra el primer conjunto. Una variante es la validación cruzada con  $n$  pliegues donde  $n$  es la cantidad de conjuntos en los que se divide a la base de datos.

**Bootstrapping:** Se usa cuando la cantidad de datos es pequeña. Consiste en construir inicialmente un modelo con todos los datos iniciales. Luego se crean numerosos conjuntos de datos haciendo un muestreo de los datos originales con reemplazo. Luego, por cada conjunto se elabora un modelo y se lo prueba contra el resto. El error final estimado se calcula promediando los errores.

Según el tipo de algoritmo, las medidas de evaluación de los modelos varían:

*Clasificación:* Se evalúa la calidad de los patrones encontrados con respecto a su precisión predictiva. Ello se obtiene dividiendo la cantidad de los resultados correctamente clasificados sobre el total. Esto puede obtenerse de las matrices de confusión.

*Reglas de asociación:* Se usan los indicadores de cobertura y confianza. La cobertura se refiere a la cantidad de instancias a las que la regla aplica y predice correctamente. La confianza se refiere a la proporción de instancias a las que la regla predice correctamente. Es decir, la cobertura dividida el número de instancias a las que se puede aplicar la regla.

*Regresión:* La manera habitual de evaluar una regresión es mediante el error cuadrático medio:  $R^2$ . Normalmente en regresiones lineales, con  $R^2 > 0,90$  se acepta el modelo.

*Clustering:* La cohesión y separación entre grupos se puede formalizar utilizando la distancia media al centro del grupo de los miembros de un grupo y la distancia entre grupos respectivamente.

*Otros modelos:* En otros modelos, se usa la precisión predictiva (similar a lo explicado para Clasificación), la cual la entregan los modelos.

A pesar de todas las medidas presentadas precedentemente, en muchos casos hay que evaluar e interpretar al modelo en función del contexto donde se va a utilizar. Por ejemplo, en el ámbito de la salud pública, en el hecho de enviar o no una ambulancia, no son aceptables niveles de confianza de (por ejemplo) 90%. En todo caso, pueden ser aceptables niveles de (por ejemplo) 99,5%. Los indicadores son aceptados o no teniendo en cuenta el contexto. Pueden aceptarse modelos con niveles de confianza a medida, según el caso.

La Teoría de la Información de Shannon y su Aplicación en la Reducción de la Dimensionalidad de los Sets de Datos.

Una de las tareas a la que se enfrenta el analista de datos es seleccionar aquellos atributos que mejor describen a la variable objetivo. Esto es, elegir aquellos atributos que aportan más información a la variable explicada (García Serrano, 2014). Una solución sencilla puede implementarse mediante el uso del concepto de entropía en el contexto de la teoría de la información (Shannon, 1948): la entropía de la información es una medida del desorden de un conjunto de datos y puede definirse:

$$H(X) = - \sum_i^n p_i \cdot \log_2(p_i)$$

donde  $p_i$  es la probabilidad relativa de aparición de la propiedad  $i$  en el conjunto de datos. El valor aquí calculado es la medida de desorden de la variable explicada (variable "Objetivo" o "Target") con un valor que va de 0 a 1. Un valor de 0 indica orden total, mientras que cuanto más cerca está de 1, mayor desorden. El concepto de entropía condicional es la entropía generada fijando de antemano el valor de una segunda variable. Es entonces, la entropía de la variable Target condicionada al valor de la variable condicionada. Se calcula:

$$\sum_i^n p_x \cdot H(X|Y=y)$$


---


$$H(X|Y) = \sum_i p_x \cdot H(X|Y=y)$$

donde  $p_x$  es la probabilidad relativa de aparición de la propiedad  $x$  en el conjunto de los datos en los que  $Y = y$  y  $H(X|Y = y)$  es la entropía de la propiedad  $x$  en el conjunto de datos donde  $Y = y$ . Es decir, la entropía condicional mide cuánta información aporta una variable a la variable explicada. De ese modo, puede analizarse la entropía de la variable explicada condicionada a todas las variables. Así se tiene una medida de aporte de cada variable, y fácilmente puede eliminarse a aquellas cuyo aporte sea poco significativo.

Puede definirse la Ganancia de Información a causa de una variable: nos dice cómo se reduce la entropía cuando añadimos la nueva variable. Se calcula:

$$GI(H|Y) = H(X) - H(X|Y)$$

Mientras menor sea el valor de  $GI$ , menor es el aporte de información de la variable condicional, por lo que pueden descartarse sin temor a perder capacidad predictiva.

### *Trabajos Recientes y Tendencias en IA*

Las tecnologías de IA se están aplicando en múltiples disciplinas: En (Pedersen, 2018) se presenta una patente de un sistema experto para automóviles basado en IA que avisa y permite tomar el control en caso de manejo peligroso. Un problema recurrente a todas es el de la reducción de la dimensionalidad. En (Hinton & Salakhutdinov, 2006) se lo trata mediante el uso de redes neuronales. Una línea de desarrollo de IA es el de las arquitecturas Deep Learning. En (Bengio, 2009) se introduce el concepto de arquitecturas Deep Learning. En (Glorot, 2011) se presenta un enfoque de Deep Learning al tratamiento y análisis de sentimientos presentes en textos extraídos de blogs, redes sociales y demás sitios con opinión de usuarios. En (Lv, Duan, Kang, Li & Wang, 2015) se propone un sistema de predicción de tráfico basado en un enfoque de Deep Learning. En (Yang, 2015) se presenta un sistema de detección de rostros basados en esta línea, con resultados destacables.

Fuera de los desarrollos en IA, en un trabajo crítico, (Natale & Ballatore, 2017) partiendo de los artículos publicados en dos revistas científicas dedicadas a la ingeniería y a la tecnología se extrajeron patrones dominantes en la construcción del mito de la IA. En (Brynjolfsson, Rock & Syverson, 2018) se presenta una interesante paradoja: mientras que muchos sistemas que implementan IA han logrado mejorar el desempeño en comparación con el humano, la productividad media de la economía se redujo a la mitad y se presentan posibles causas.

### *2.2. Genotoxicidad y Oncoepidemiología*

#### *Genética: Nociones Básicas*

El genotipo es el conjunto de genes de un individuo, es la expresión funcional del genoma que son los genes del organismo o de la especie y que puede comprenderse como una totalidad variante, pero una totalidad en fin de cuentas. El genotipo se concibe de manera reduccionista al referirlo a una característica del sujeto, de ahí que podemos hablar del genotipo asociado a la celiaquía, al

cáncer hereditario no polipósico o a la susceptibilidad a la obesidad. Por otra parte, el fenotipo (lo externo o *phaenomenon*) es el resultado de las interacciones entre el genotipo y el ambiente Jones, Pembrey, Golding & Herrick, 2005).

Es una definición clásica de Genética, aparentemente simple: Genes + Ambiente = Fenotipo, sin embargo, no describe lo que realmente ocurre: el fenotipo es el resultado de las interacciones entre los efectos de los genes en un sistema biológico y los efectos de los determinantes ambientales entre sí y con el sistema biológico. Reducido, a un conjunto limitado de genes que en la realidad esta interconectado mediante las proteínas codificadas por estos con las proteínas codificadas por otros genes del individuo, aunque no sean consideradas dentro del genotipo en cuestión (Lehner, 2007). Un ejemplo son las variantes clínicas en la Celiaquía que pueden ser consecuencia de genes modificadores no HLA, aunque cuando se habla de genotipo asociado a la Celiaquía, la mayoría de las veces solamente se consideran los genes HLA DQ2 y DQ8 (Coleman, Quinn, Ryan, Conroy, Trimble, Mahmud, Kennedy, Corvin, Morris, Donohoe, O'Morain, MacMathuna, Byrnes, Kiat, Trynka, Wijmenga, Kelleher, Ennis, Anney & McManus, 2016).

### *Genética y Cancer*

El cáncer puede ser considerado como un fenotipo cuyas principales características son las que expresan las interacciones entre las mutaciones en diversos genes supresores y oncogenes con el ambiente, que en principio causó la mutación desreguladora (Dingli, Chalub, Santos, Van Segbroeck & Pacheco, 2009). Es un fenotipo en el que se observa la proliferación descontrolada, la indiferenciación, los cambios en la expresión de receptores hormonales, la angiogénesis y la capacidad de metastizar entre otras (Gupta, Kim, Prasad & Aggarwal, 2010).

El cáncer (o los cánceres), en sus diferentes localizaciones, sus variables histológicas y su respuesta a la terapia onco específica, es la expresión de fenotipos. La incidencia y prevalencia de los cánceres, guarda relación con agentes ambientales y cambios del entorno metabólico individual, tales como los relacionados con la obesidad (Hu et al, 2015).

En tal sentido la cooperación de los carcinógenos ambientales y su efecto sobre individuos de una población heterogénea en cuanto a sus genotipos, convierten la tarea de definir aquellas exposiciones influyentes sobre la incidencia de cánceres específicos, en algo muy difícil que ha obligado a generalizar las relaciones carcinógeno – cáncer, aunque no siempre las condiciones intervinientes son similares (Domingo & Nadal, 2017).

### *Epidemiología del Cáncer*

La prevención del cáncer requiere de la elucidación de cuáles son las interacciones que determinan el aumento de la incidencia en poblaciones específicas en hábitats específicos. Los estudios en los que se definen las redes de control bioquímico, las variaciones en la expresión genética y sus interacciones son necesariamente reduccionistas, puesto que es imposible investigar todos los genes candidatos, con todos los carcinógenos sospechados, sin dejar de lado, aquellos alimentos o conductas que funcionan como anti carcinógenos, tal es el caso de las frutas y el ejercicio físico.

Enfrentar el problema de las interacciones entre los conductos de variables que caracterizan a los individuos (hereditarias, daño genético, metabólicas, etc.) y las variables ambientales (nutricionales, ocupacionales, hábitat, hábitos tóxicos, etc.) es una empresa que reviste mucha complejidad y que debe ser abordada con un enfoque dual. Debe continuar el estudio en profundidad de cada

uno de los aspectos biológicos que están incluidos en cada uno de los aspectos mencionados. Ej. Hereditarios: genes de respuesta al estrés oxidativo, genes de enzimas del complejo de citocromo p450, genes de reparación del ADN, entre otros. Daño genético: es el proceso neoplásico de las lesiones de simple y doble cadena del ADN, mutaciones inducidas en genes de reparación, inestabilidad cromosómica, por nombrar algunos. Mientras que en los estudios ambientales pueden mencionarse como ejemplos factores nutricionales, el consumo de ácidos grasos omega 3 y 6, el consumo de antioxidantes, de fibras, de carne roja. El conjunto de metabolitos que surgen de estos alimentos o Metaboloma, las exposiciones ocupacionales a hidrocarburos, bifenilos policlorados, benceno, argón, agroquímicos que generan agresiones como resultado de sus acciones individuales o de sus mezclas (Torres-Bugarín, Fernandez-García, Torres-Mendoza, Zavala-Aguirre, Nava-Zavala, Zamora-Perez, 2009; Damasceno, Sinzato, Bueno, Dallaqua, Lima, Calderon, Rudge & Campos, 2013; Davies & Albeck, 2018; Easmond, 2017).

### *La Genotoxicidad y el Estudio de la Enfermedad Neoplásica*

Los genes se pueden secuenciar, determinar sus variantes asociadas al desarrollo neoplásico, rastrear su presencia en las poblaciones, definir su capacidad de producir cáncer al mutar, saber si es necesaria una o más mutaciones y que proteínas alteradas aparecen cuando se producen los cambios genéticos. En el ambiente pueden detectarse aquellos agentes capaces de producir aberraciones cromosómicas, micro cromosomas, mutaciones puntuales. También es posible evaluar a la obesidad como generadora de un entorno pro carcinogénico, los efectos de la inflamación crónica asociada a esa enfermedad, las alteraciones vinculadas a las adipocitoquinas, el estrés oxidativo, las proteínas de estrés (Brinke & Buchinger, 2017). Esos participantes: genes, ambiente, entorno metabólico, sistema inmune son el terreno donde se desarrollan los cánceres y están originando un área de investigación que llevaría *form bench to prevention*, no *from bench to bed*, puesto que su propósito es descubrir aquellos factores ambientales e individuos que al vincularse pueden llevar al cáncer (Umbuzeiro, Heringa & Zeiger, 2017; Afanasieva & Sivolob, 2018).

Es decir que, como complemento de los estudios reduccionistas es imperativo tomar lo que la realidad nos aporta como efectos de las combinaciones de lo individual y lo ambiental. La complejidad de las interacciones genes ambiente, sus dimensiones espacio temporales orientan a pensar en la necesidad de hacer comparaciones multivariadas, que por sus dimensiones combinatorias puedan abarcar una gran número de variables intervinientes. El objetivo: definir cuáles son las combinaciones que, respondiendo a los conocimientos actuales sobre la transformación maligna, puedan ser las responsables mayores de la presencia de cáncer en las poblaciones en estudio.

### *2.3. Acerca de la Colaboración Interdisciplinaria*

El proyecto “Hábitos Tóxicos y Contaminación Ambiental y su efecto sobre la salud genética” está radicado en el CAECIS, y se encuentra dirigido por el médico genetista Dr. Elio Prieto Gonzalez. Los investigadores profesionales de la salud que colaboran en el proyecto son: la Lic. Vanesa Miana y la Lic. Paola Aldegani. Colaboran también estudiantes becarios.

El matemático Lic. Jorge Kamlofsky fue contactado para colaborar en el proyecto con el análisis de datos. Su incorporación reorientó el enfoque del análisis de datos hacia las técnicas de IA. Pero también colaboró con el diseño de un nuevo esquema de recopilación de datos: el desarrollo de una aplicación web para las encuestas, para lo cual se incorporó al proyecto, al Ingeniero en Sistemas Informáticos Claudio Milio. Tanto Jorge Kamlofsky como Claudio Milio son

investigadores del CAETI. La colaboración de los investigadores del CAETI ayuda a hacer más eficiente el manejo de datos y la implementación de algoritmos para la obtención de resultados promisorios a partir de la recopilación de datos.

### 3. La Experiencia

#### 3.1. Introducción

Los estudios epidemiológicos se basan en el análisis de los efectos de muchos genes y factores ambientales en la aparición del cáncer (Grasgruber, Hrazdira, Sebera & Kalina, 2018). Al enfoque tradicional epidemiológico, se ha sumado el de la Inteligencia Artificial que dispone de las herramientas para el análisis de una gran cantidad de datos y que puede determinar cuáles son las variables que más se asocian al estado de paciente con enfermedad neoplásica (Liu, Chen, Chen & Jia, 2009). La identificación de relaciones entre factores dentro de un conjunto de información mediante MD, puede aplicarse al estudio de las interacciones de elementos del ambiente con el ADN y al de los factores asociados al desarrollo del cáncer.

El presente estudio tiene como objetivo evaluar los efectos de la obesidad, los hábitos tóxicos y los ambientes contaminantes sobre los niveles de daño en el ADN determinado por electroforesis alcalina de células aisladas en gel de agarosa o Ensayo Cometa.

#### 3.2. Los Datos Experimentales

##### *Diseño Experimental*

Consiste en una encuesta realizada a 54 mujeres. La encuesta fue realizada a mujeres adultas (desde 22 años a 67 años), residentes en la zona, y se atienden en la Sala de atención primaria. Las respuestas se volcaron en un archivo de hoja de cálculos. La cantidad total de variables es de 159. La variable objetivo o “target” es el índice de daño genético.

##### *Tiempo y Ubicación de la Encuesta*

Los datos se recolectaron en la Ciudad de La Plata, Provincia de Buenos Aires, Argentina. El relevamiento fue realizado los días 22, 23, 29 y 30 de Septiembre de 2014 y 3 de Marzo de 2015.

##### *Responsables de la Obtención de los Datos*

Los datos fueron obtenidos en ocasión de las consultas médicas de pacientes a la sala y bajo su consentimiento. Los profesionales de la salud a cargo de la recolección de los datos son el Doctor Elio Prieto Gonzalez y la Lic. Vanesa Miana.

##### *Las Planillas*

Los datos relevados se encuentran en una planilla de cálculo que contiene las siguientes hojas: “Datos de la Encuesta”, “Datos y Antecedentes”, “Frecuencia consumo alimentos”, “Hábitos tóxicos”, “Indicadores antropométricos” y “Microambiente”. Cada una de ellas se las presenta a continuación.

- » *Datos de la Encuesta*: Esta planilla contiene las fechas de realización del relevamiento y su validación, por cada paciente, indicando las iniciales de los doctores encuestadores (EP: Dr. Elio Prieto Gonzalez, VM: Lic. Vanesa Miana). Las columnas corresponden con la identificación del paciente cuya ID se mantiene en el resto de las planillas.

- » *Datos y Antecedentes*: Contiene datos de cada paciente, entre otros: nombre, edad (en el momento de la encuesta), dirección de residencia, etc. Los antecedentes contienen datos de antecedentes patológicos personales, familiares, y datos de la vivienda.
- » *Frecuencia de consumo de alimentos*: Contiene las frecuencias de consumos de carnes, frutas y hortalizas, pescados y mariscos, pastas, lácteos, harinas, infusiones y demás, así como la forma de cocción.
- » *Hábitos tóxicos*: Se presenta hábitos tóxicos del encuestado, como ser consumo de alcohol y tabaquismo.
- » *Indicadores antropométricos*: Se incluyen datos de mediciones hechas en la consulta: altura, peso, presión arterial y otros junto con una puntuación por su impacto en el daño genético.
- » *Microambiente*: Contiene datos de contaminación ambiental: cercanía a basurales, quemas, reciclado de residuos, entre otros.

### 3.3. Equipamiento Usado

La algoritmia se programó en Python 2,7. El computador usado es una notebook HP Pavillion con procesador AMD A10 con 12Mb de Ram y 4 cores. El Sistema operativo usado es Kali Linux: una distribución Linux basada en núcleo Debian.

## 4. Proceso de Descubrimiento del Conocimiento

En esta sección, dados los datos recibidos tal como se los presentó previamente, se los transformará en conocimiento útil aplicando cada una de las fases del KDD.

### 4.1. Pre-Procesamiento (ETL): Armado de las Vistas Minables

Los datos experimentales en bruto, tal como se recibieron deben ser procesados para su posterior tratamiento por parte de los algoritmos. Puede observarse que las planillas así recibidas requieren de muchas tareas de pre-procesamiento, algunas no descritas previamente. Por ejemplo, puede observarse que el desarrollo de todas las planillas se da en anchura. Es decir, los casos están en columnas mientras que los atributos están en filas. A simple vista, también puede observarse que se hay datos incompletos, los formatos de datos no son iguales, entre otros.

Las tablas recibidas tienen desarrollo en anchura. Es decir, los atributos están en filas y los casos se presentan en columnas. Para el proceso de MD se requiere invertir este orden. Finalmente se requiere la normalización y discretización de datos, según la necesidad de cada algoritmo. En la Figura 3 se muestra una porción de la Tabla “Frecuencia de consumos de alimentos”.

Numero de Caso	Puntos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Frecuencia de consumo de carnes	7	X		X	X	X	X		X	X	X	X	X										
Frecuencia de consumo de pollo	4							X															
Diariamente:	SI	7	X		X	X	X	X	X	X				X			X		X			x	
	NO	3		X				X										X				x	
Entre 1 a 2 veces por semana	2																					x	
Entre 2 y 3 veces por semana	3		X				X				X	X		X	X		X		X		x		x
Tipo de coccion:																							
	Frita	12	X	XX	X		XX	X				XX	X	X	X	X	X	X	X	x	x		
	Cocida	4	X	XX	X	X	X	X	XX	XX	XX	XX	XX	X	XX	X	XX	X	x	x	x	x	
	Horno	4																					
	Parrilla	10	X	X	X	X				X											x	x	

Fig. 3: Porción de la Tabla de Datos: Frecuencia de consumos de alimentos

### Eliminación y combinación de atributos

Una primer tarea para cambiar el formato de todas las planillas, consiste en seleccionar los atributos que serán de interés y eliminar aquellos no útiles o no computables (por ejemplo: “Domicilio”), combinar, fusionar y cambiar nombres. Ello se realiza manualmente.

### Verificación de Formato Adecuado de Datos

Hay planillas que contienen solamente entradas numéricas (enteros), otras que tienen sólo entradas binarias (“X”) o ambas. La gran mayoría de las entradas son binarias. En las planillas ello se manifiesta con una “X”. Sin embargo, hay entradas que tienen “x”, “XX”, “XXX” u otras versiones o textos que pretenden acentuar o bien presentan detalles que amplían el “SI” manifestado con “X”, tal como puede observarse en la Figura 3.

Esta tarea puede realizarse simplemente con una fórmula en la planilla de cálculos<sup>3</sup> como sigue:

$$=SI(ESBLANCO(Celda);" ";SI(ESNÚMERO(Celda);Celda;"X"))$$

Esto es: si la celda en cuestión es un número o bien está en blanco, ello se mantiene. En caso contrario, se pone “X”.

### Completado de Datos

En las planillas que tienen entradas numéricas, la falta de un dato es visualmente evidente. Detectada la ausencia, se consultó a los investigadores para que completen los mismos y se obtuvo la respuesta correspondiente.

### Cálculo de puntuaciones

En varias planillas se requiere la asignación de puntuación o ponderación numérica a respuestas binarias en diferentes atributos en las hojas: “Frecuencia consumo alimentos”, “Hábitos tóxicos” y “microambiente”, para analizar numéricamente su impacto en el índice de daño Comet ID Basal. Los puntos a cada respuesta fueron definidos por los investigadores. Se requiere asignar la puntuación si dicho atributo se presentó en la respuesta del encuestado o no. Esto se puede realizar con una simple función en planilla de cálculo:

3 En este trabajo se usó Libre Office: <https://es.libreoffice.org/>

=SI(Celda="x";CeldaPuntosSi;CeldaPuntosNo)

### *Consistencia Lógica*

Hay respuestas que están en función de otras. Por lo tanto, si algunas faltan, se pierde consistencia en los datos. La detección de estos casos se realiza con funciones lógicas de planilla de cálculos. Se verificó la consistencia datos numéricos mediante comparaciones varias y análisis de outliers. Los casos detectados fueron informados a los investigadores y corregidos o completado por ellos.

### *Armado de Dataset básico*

Se armó un banco de datos concentrando la información de todas las planillas. En la etapa de limpieza, hubo variables que evidentemente carecen de interés, por lo que fueron eliminadas (por ejemplo: "Domicilio", o "Número de Caso"). Otras se fusionaron entre sí. Luego se eliminan las filas en blanco. De ese modo, la cantidad de filas (los atributos) de cada planilla se reduce notoriamente. La cantidad de variables que serán usadas se reduce notoriamente: de 159 variables originales (159 filas en todas las planillas) a 64 variables: 30 numéricas y 34 discretas. De éstas, como la variable Target es el Índice de Daño "Comet\_Id\_Basal", habrá una variable Target (u Objetivo) nominal y otra numérica, según que modelo se utilice.

Se eliminó el atributo "Número de Caso", y además, se eliminaron los casos #23 y #41 que no contenían datos. Para la discretización de las variables nominales se usó el criterio brindado por los investigadores. Los subtotales y totales fueron eliminados debido a que no son independientes, sino que son dependientes.

Debido a que hay modelos que trabajan con datos discretos y otros que trabajan con datos numéricos, se dividió el set de datos en dos: uno conteniendo solo los datos discretos y otro conteniendo los datos numéricos. De guardaron los el sets de datos traspuestos en formato CSV con los nombres: "dataset\_valor.csv" y dataset\_disc.csv".

## *4.2. Minería de Datos*

### *Resumen*

Sobre los datos procesados se va a implementar el algoritmo de Regresión múltiple. Para ello, se importa desde Python la librería de R para Python: "rpy2". Así, se pide a R que obtenga un modelo de regresión lineal múltiple sobre una combinación de variables presentada por Python junto con su evaluador  $R^2$  y se los almacena en una lista. Python presentará a R todas las combinaciones de variables posibles y se seleccionará aquellos modelos obtenidos con mayor valor en el evaluador  $R^2$ .

Previo a ello, sobre los sets de datos mencionados, se trata el problema de "Reducción de Dimensionalidad" utilizando una implementación de la Teoría de la Información de Shannon (Shannon, 1948).

### *El Problema de la Reducción de la Dimensionalidad*

Según lo presentado en el apartado anterior, se puede ver que hay 34 variables discretas y 30 variables numéricas. Descontando a la variable target, se puede decir que el set de datos discretos posee 33 variables predictoras y una variable objetivo o target, mientras que el set de datos numéricos posee 29 variables predictoras y una variable objetivo o target.

Se evaluará cada una de todas las combinaciones posibles de los conjuntos de datos. La cantidad posible de combinaciones, puede obtenerse usando herramientas del cálculo combinatorio. La cantidad de combinaciones de las  $n$  variables tomadas de a  $m$  son:

$$\binom{n}{m} = \frac{n!}{m!(n-m)!} \quad (\text{ecuación 1})$$

$$Q = \frac{\sum_{m=1}^n n!}{m!(n-m)!} = \binom{n}{1} + \binom{n}{2} + \dots + \binom{n}{n} \quad (\text{ecuación 2})$$

Siendo  $Q$  la cantidad posible de combinaciones,  $n = 33$  la cantidad de predictores y  $m$  la cantidad de elementos de  $n$  que se usan, finalmente  $Q$  será;

$$Q = \frac{\sum_{m=1}^{33} 33!}{m!(33-m)!} = \binom{33}{1} + \binom{33}{2} + \dots + \binom{33}{33} = 33 + \binom{33}{2} + \binom{33}{3} + \dots + 1 = 33 + \frac{\sum_{m=2}^{33} 33!}{m!(33-m)!} + 1$$

El cálculo de  $Q$  usando consola de R:

```
> Q = 33 + 528 + 5456 + 40920 + 237336 + 1107568 + 4272048 + 3884156 + 38567100 + 92561040 +
193536720 + 354817320 + 573166440 + 818809200 + 1037158320 + 1166803110 + 1166803110 +
1037158320 + 818809200 + 573166440 + 354817320 + 193536720 + 92561040 + 38567100 + 3884156
+ 4272048 + 1107568 + 237336 + 40920 + 5456 + 528 + 33 + 1
```

```
> Q
```

```
[1] 8569934591
```

Siendo  $Q \approx 8,57 \times 10^9$ , con la computadora usada, se estima que obtener todos los modelos para todas las posibles combinaciones insume aproximadamente  $t \approx 4,2 \times 10^6$  seg  $\approx 1,13 \times 10^4$  hs  $\approx 486$  días  $\approx 1,33$  años, lo que obviamente no es aceptable para este análisis.

Análogamente, puede repetirse el cálculo para el set de datos de variables numéricas, lo cual da un número algo menor, pero igualmente de alto nivel de complejidad. Todo esto expone la necesidad de tratar el problema conocido como “reducción de dimensionalidad” a niveles aceptables, para obtener modelos en tiempos razonables y con tamaños que permitan entender el conocimiento descubierto.

*Análisis de Correlación:* Una forma de reducir la cantidad posible de modelos a analizar consiste en el armado de modelos solo con las variables predictoras que pudieran tener fuerte correlación con la variable objetivo. Mediante un análisis de correlación entre las variables numéricas y la variable objetivo, puede saberse si un predictor incide fuertemente sobre o no el objetivo. Con valores de correlación próximos a  $|1|$  se asume fuerte incidencia, mientras que con valores de correlación próximos a  $|0|$ , la incidencia es débil.

Deberán calcularse las correlaciones entre todas las distintas variables con la variable Objetivo (var\_24). En R:

```
cor_i_24 <- cor(var_i,var_24)
```

Del análisis se concluyó que ninguna variable está correlacionada con la variable “target”.

*Reducción de la Dimensionalidad Mediante la Implementación de la Teoría de la Información (Shannon):* La cantidad de variables aún no permiten que puedan realizarse análisis de este conjunto de datos en forma eficiente. Se requiere una fuerte reducción de la cantidad de variables. Para ello puede utilizarse la Teoría de la Información: por cada variable se calcula la Ganancia de Información, es decir, cuánto aporta cada variable a la variable Target. El presente desarrollo se basó en un ejemplo presentado en (García Serrano, 2014). En la Tabla 1 se muestra la ejecución y salida por pantalla de la Ganancia de Información.

Tabla 1: Salida de Pantalla de la Ganancia de la Información

```
Python 2.7.10 (default, Oct 14 2015, 16:09:02)
[GCC 5.2.1 20151010] on linux2
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
>>> import teoriaDeLaInformacion as ti
>>> gi=ti.obtenerGananciaInformacion('dataset_valor.csv',24)
>>> print str(gi)
[[0, 'Edad', 4.0354], [1, 'Anos_de_residencia', 4.27792], [2, 'Reproductividad_ cantHijos', 2.11740], [3, 'Consumo_Carne', 3.73137], [4, 'Consumo_Pescado', 1.02496], [5, 'Consumo_Vegetales', 1.86022], [6, 'Consumo_Pastas', 1.79545], [7, 'Consumo_Arroz', 0.96276], [8, 'Consumo_Lacteos', 2.06573], [9, 'Consumo_Infusiones', 1.75058], [10, 'Consumo_Vino', 1.45191], [11, 'Consumo_Endulzantes', 0.73150], [12, 'Consumo_Harinas', 1.61902], [13, 'Tabaquismo_activo', 2.40985], [14, 'Tabaquismo_pasivo', 3.29121], [15, 'Consumo_Alcohol', 1.56825], [16, 'Peso', 5.09380], [17, 'Talla', 3.75186], [18, 'IMC', 5.06938], [19, 'Porc_grasa_corporal', 4.99095], [20, 'circ_abdominal', 4.56919], [21, 'tension_arterial_max', 2.55768], [22, 'tension_arterial_min', 2.08566], [23, 'glucemia', 4.59880], [24, 'Comet_id_basal', 5.22625], [25, 'Se_reciclan_residuos', 1.91096], [26, 'Se_queman_neumaticos', 1.11369], [27, 'Se_queman_hojas_maderas_restos_de_comida', 1.02837], [28, 'Se_queman_botellas_de_plastico', 1.02837], [29, 'Hay_olores_fuertes', 3.03169]]
>>>
```

Tabla 2: Variables seleccionadas luego de la aplicación de la Teoría de la Información.

Id	Nombre	Información	Id	Nombre	Información
28	'Comet_id_basal'	1.06398	24	'Comet_id_basal'	5.22625
1	'Anos_de_residencia'	0.20214	16	'Peso'	5.09380
8	'Nivel_Educacion'	0.13475	18	'IMC'	5.06939
0	'Edad'	0.12608	19	'Porc_grasa_corporal'	4.99096
23	'IMC'	0.12252	23	'glucemia'	4.59880
2	'Nacionalidad'	0.08809	20	'circ_abdominal'	4.56920
25	'circ_abdominal'	0.07976	1	'Anos_de_residencia'	4.27793
27	'Glucemia'	0.07433	0	'Edad'	4.03549
26	'tension_arterial'	0.06787	17	'Talla'	3.75187
20	'Tabaquismo_activo'	0.05648	3	'Consumo_Carne'	3.73137
17	'Consume_Vino_frecuentemente'	0.05615	14	'Tabaquismo_pasivo'	3.29121
24	'Porc_grasa_corporal'	0.05397	29	'Hay_olores_fuertes'	3.03169
7	'Cocina_a_Gas'	0.05397	21	'tension_arterial_max'	2.55768
13	'Consume_Pastas_Frecuentemente'	0.05137	13	'Tabaquismo_activo'	2.40985
29	'Se_reciclan_residuos'	0.05021	22	'tension_arterial_min'	2.11740

Set de datos: dataset\_disc.csv

Set de datos: dataset\_valor.csv

Los resultados presentados en la Tabla 1 muestran el aporte de información de cada una de las variables predictoras a la variable Target. Entonces, ordenando los resultados en forma decreciente puede tenerse un ranking de los aportes de información. Así, en algún lugar conveniente puede realizarse el corte que permita definir la dimensionalidad del conjunto de datos a analizar. El criterio elegido consiste en aceptar la cantidad de variables que aportan más del 5% de la información, y sobre esas variables se aplicará los diferentes modelos aquí mencionados. La Tabla 2 presenta las 14 variables seleccionadas según este criterio.

*Cantidad de Combinaciones Posibles:* Una vez definidas las variables que finalmente serán utilizadas, los modelos serán alimentados por todas las combinaciones posibles de las variables. Con ello, se eliminarán sesgos, y solamente se seleccionarán modelos que cumplan con los niveles de evaluación requeridos. La cantidad total de combinaciones se obtiene a partir de la ecuación 2.

Calculado usando la consola de R será:

```
> Q=14+91+364+1001+2002+3003+3432+3003+2002+1001+364+91+14+1
> Q
[1] 16383
```

### Obtención de Modelos

Mediante el algoritmo Regresión Lineal Múltiple se propone un modelo en el que la variable target se obtenga como una función lineal de las otras variables. En este caso, se calcula con R las regresiones múltiples a partir del conjunto de las 16383 combinaciones posibles preparadas con Python: Por cada combinación, se obtiene el modelo de regresión múltiple junto con el valor de  $R^2$ . Se selecciona el mejor modelo y el más sencillo.

### 4.3. Resultados

La Tabla 3 presenta el reporte de los modelos obtenidos.

Tabla 3: Reporte de las Regresiones Lineales Múltiples

```
Python 2.7.10 (default, Oct 14 2015, 16:09:02)
[GCC 5.2.1 20151010] on linux2
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
+-----+
| REPORTE DE REGRESIONES LINEALES MULTIPLES:
| Cantidad de Regresiones múltiples calculadas: 16383
| Cantidad de variables seleccionadas: 14
| Cantidad de modelos elegidos: 38
| El mejor modelo: ('[24, 16, 18, 19, 20, 23, 1, 0, 17, 3, 14, 29, 21, 13, 22]', '0.31042')
| El modelo más sencillo: ('[24, 1, 0, 17, 3, 29, 21]', '0.26937')
| Demora: 0:02:24.341189
+-----+
>>>
```

El resultado incluye la siguiente información: la cantidad de modelos calculados (16383), la cantidad de variables seleccionadas (14), la cantidad de modelos elegidos (38), y por el mejor modelo y por el modelo más simple: la selección de las variables usadas y el valor de  $R^2$  obtenido. Mientras que en el mejor modelo  $R^2 = 0,31042$  (o bien  $R = 0,55715$ ), en el más simple:  $R^2 = 0,26937$  (o bien  $R = 0,519$ ). El tiempo insumido que fue de 2:24,341189 minutos.

## Conclusiones

Mediante técnicas de IA se han obtenido modelos en forma de ecuaciones que permiten predecir el valor del índice de daño genético en función de las variables que más información aportan, partiendo de una cantidad muy reducida de datos.

A partir de los modelos, se obtuvieron indicios acerca de cuáles son los factores que inciden en el índice de daño genético. Una relación que expresa mejor la realidad que la que se obtiene haciendo evaluaciones que no incluyen todas las combinaciones posibles de variables.

Si bien el valor del evaluador  $R^2$  no permite que los resultados obtenidos puedan ser concluyentes, los mismos aportan indicios interesantes.

Finalmente, se mostró la importancia del aporte de las tecnologías de la información y sus métodos en la investigación médica logrado gracias a la colaboración inter-disciplinaria.

## Trabajos Futuros

Un trabajo posterior al presente está en curso: se desarrolló una encuesta en ambiente web con una base de datos que evita errores de ingreso de datos, y permitire reducir los tiempos tanto en el proceso ETL como en la implementación de los algoritmos de IA. Con una cantidad de datos notoriamente mayor, los resultados que se obtengan, pueden ser de interés, y concluyentes.

## Referencias

- » Afanasieva K, Sivolob A. (2018). "Physical principles and new applications of comet assay". *Biophys Chem.*;238:1-7. doi: 10.1016/j.bpc.2018.04.003. Epub 2018 Apr 20.
- » Aluja, Tomàs (2001). "La minería de datos, entre la estadística y la inteligencia artificial." *Qüestió: quaderns d'estadística i investigació operativa* 25.3: 479-498.
- » Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1), 1-127.
- » Beerli-Palacio, A; Santana, J. M., y Porta, M. (2008). "El marketing como herramienta para incrementar la eficacia de los planes de salud pública". *Informe SESPAS 2008. Gaceta Sanitaria* 22, pp. 27-36
- » Brinke A, Buchinger S. (2017). "Toxicogenomics in Environmental Science". *Adv Biochem Eng Biotechnol.*;157:159-186. Doi: 10.1007/10\_2016\_15.
- » Brynjolfsson, E., Rock, D., & Syverson, C. (2018). "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics". In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.
- » Castillo, Andres. (2011). *El diseño de Investigación*. Disponible en Web: <https://www.monografias.com/trabajos101/disenos-investigacion/disenos-investigacion.shtml> (consulta: 30-11-2018)
- » Coleman C, Quinn EM, Ryan AW, Conroy J, Trimble V, Mahmud N, Kennedy N, Corvin AP, Morris DW, Donohoe G, O'Morain C, MacMathuna P, Byrnes V, Kiat C, Trynka G, Wijmenga C, Kelleher D, Ennis S, Anney RJ, McManus R. (2016). "Common polygenic variation in coeliac disease and confirmation of ZNF335 and NIFA as disease susceptibility loci". *Eur J Hum Genet.* 24(2):291-7. doi: 10.1038/ejhg.2015.87. Epub 2015 Apr 29. PubMed PMID: 25920553; PubMed Central PMCID: PMC4717209.
- » Damasceno DC, Sinzato YK, Bueno A, Dallaqua B, Lima PH, Calderon IM, Rudge MV, Campos KE. (2013) Metabolic profile and genotoxicity in obese rats exposed to cigarette smoke. *Obesity (Silver Spring)*. (8):1596-601. doi: 10.1002/oby.20152. Epub 2013 May 13
- » Davies AE, Albeck JG. (2018). "Microenvironmental Signals and Biochemical Information" Processing: Cooperative Determinants of Intratumoral Plasticity and Heterogeneity. *Front Cell Dev Biol.* 20;6:44. doi: 10.3389/fcell.2018.00044. eCollection 2018. Review. PubMed PMID: 29732370; PubMed Central PMCID: PMC5921997.
- » De Jonge, Edwin; Van Der Loo, Mark. (2013). "An Introduction to Data Cleaning with R". *Statistics Netherlands*, The Hague/Heerlen.
- » De Luis D. A., Aller R., Conde, R. Izaola O., de la Fuente B., Gonzalez Sagrado, M. Primo D. Ruiz Mambrilla M. (2012) Relación del polimorfismo rs9939609 del gen FTO con factores de riesgo cardiovascular y niveles de adipocitoquinas en pacientes con obesidad mórbida. *Nutr Hosp.* 2012;27(4):1184-1189
- » Dingli D, Chalub FA, Santos FC, Van Segbroeck S, Pacheco JM. (2009). "Cancer phenotype as the outcome of an evolutionary game between normal and malignant cells". *Br J Cancer*;101(7):1130-6. doi: 10.1038/sj.bjc.6605288. Epub 2009 Sep 1. PubMed PMID: 19724279; PubMed Central PMCID: PMC2768082.
- » Domingo JL, Nadal M. (2017). Carcinogenicity of consumption of red meat and processed meat: A review of scientific news since the IARC decision". *Food Chem Toxicol.*105:256-261. doi: 10.1016/j.fct.2017.04.028. Epub 2017 Apr 24. Review. PubMed PMID: 28450127.
- » Dyche, Jill. (2000). "E-Data: Turning data into information with data warehousing". *Addison-Wesley Professional*.
- » Eastmond DA. (2017). "Recommendations for the evaluation of complex genetic toxicity data sets when assessing carcinogenic risks to humans". *Environ Mol Mutagen.* ;58(5):380-385. doi: 10.1002/em.22078. Epub 2017 Mar 7. PubMed PMID: 28266084.

- » Fayyad, U.; Piatetski-Shapiro, G.; Smyth, P. (1996) "From Data Mining to Knowledge Discovery: an Overview". *Advances in Knowledge Discovery and Data Mining*, pp-1-34, AAAI/MIT Press.
- » Fisher, R. A. (1925) "Statistical Methods, Experimental Design and Scientific Inference". *Oxford Science Publications*.
- » García Serrano, Alberto (2014). "Selección de atributos relevantes usando la entropía de Shannon". Disponible en Web: <http://www.inteligenciapredictiva.com/2014/06/seleccion-de-atributos-relevantes-entropia.html> (consulta: 07-08-2016).
- » Glorot, X., Bordes, A., & Bengio, Y. (2011). "Domain adaptation for large-scale sentiment classification: A deep learning approach". In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 513-520).
- » Grasgruber P, Hrazdira E, Sebera M, Kalina T. (2018). "Cancer Incidence in Europe: An Ecological Analysis of Nutritional and Other Environmental Factors". *Front. Oncol*, 13 June. Disponible en Web: <https://doi.org/10.3389/fonc.2018.00151> (consulta: 30-11-2018)
- » Gupta SC, Kim JH, Prasad S, Aggarwal BB. (2010). "Regulation of survival, proliferation, invasion, angiogenesis, and metastasis of tumor cells through modulation of inflammatory pathways by nutraceuticals". *Cancer Metastasis Rev.*;29(3):405-34. doi: 10.1007/s10555-010-9235-2. Review. PubMed PMID: 20737283
- » Gustafson, S., and Sheth, A. (2014). "Web of Things". *Computing Now* 7.3.
- » Hernández Orallo, J; Ramírez Quintana, M, y Ferri Ramirez, C. (2004). "Introducción a la Minería de Datos". *Editorial Pearson Educación SA, Madrid*.
- » Hu Z, Brooks SA, Dormoy V, Hsu CW, Hsu HY, Lin LT, Massfelder T, Rathmell WK, Xia M, Al-Mulla F, Al-Temaimi R, Amedei A, Brown DG, Prudhomme KR, Colacci A, Hamid RA, Mondello C, Raju J, Ryan EP, Woodrick J, Scovassi AI, Singh N, Vaccari M, Roy R, Forte S, Memeo L, Salem HK, Lowe L, Jensen L, Bisson WH, Kleinstreuer N. (2015). "Assessing the carcinogenic potential of low-dose exposures to chemical mixtures in the environment: focus on the cancer hallmark of tumor angiogenesis". *Carcinogenesis*. 36 Suppl 1:S184-202. doi: 10.1093/carcin/bgv036. Review. PubMed PMID: 26106137; PubMed Central PMCID: PMC4492067.
- » Jones R, Pembrey M, Golding J, Herrick D (2005). "The search for genotype/phenotype associations and the phenome scan". *Paediatr Perinat Epidemiol*. 19(4):264-75. Review. PubMed PMID: 15958149, 2005.
- » Lehner B. (2007). "Modelling genotype-phenotype relationships and human disease with genetic interaction networks". *J Exp Biol*; 210(Pt 9):1559-66. Review. PubMed PMID: 17449820.
- » Liu Z, Chen D, Chen X, Jia H. (2009). "Computational Data Mining in Cancer Bioinformatics and Cancer Epidemiology". *Journal of Biomedicine and Biotechnology*, 582697. Disponible en Web: <http://doi.org/10.1155/2009/582697> (consulta: 30-11-2018).
- » Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F. Y. (2015). "Traffic flow prediction with big data: A deep learning approach". *IEEE Trans. Intelligent Transportation Systems*, 16(2), 865-873.
- » Mc Carthy, John, and Hayes, Patrick (1981). "Some philosophical problems from the standpoint of artificial intelligence". *Readings in artificial intelligence*. 431-450.
- » Mc Corduck, P. (2004). "Machines Who Think. A personal inquiry into the history and prospects of artificial intelligence". *New York: A K Peters/CRC Press*.
- » Mena, Jesus. (1999). "Data mining your website". *Digital Press*.
- » Mitra, Sushmita, and Tinku Acharya. (2005). "Data mining: multimedia, soft computing, and bioinformatics". *John Wiley & Sons*.
- » Ospina, E; Reveiz Herault, L y Cardona, A. (2005). "Uso de bases de datos bibliográficas por investigadores biomédicos latinoamericanos hispanoparlantes: estudio transversal". *Revista Panamericana de Salud Pública*, vol no 4. 17, pp. 230-236.

- » Pedersen, R. D. (2018). "Motor vehicle artificial intelligence expert system dangerous driving warning and control system and method" *U.S. Patent No. 9,919,648*. Washington, DC: U.S. Patent and Trademark Office.
- » Rao, C. R. *Statistics and Truth*. CSIR, New Delhi, 1989.
- » Riquelme, José, Ruiz, Roberto y Gilbert, Karina. (2006). "Minería de datos: Conceptos y tendencias." *Revista Iberoamericana de Inteligencia Artificial* 10.29: 11-18.
- » Shannon, Claude Elwood (1948). "A mathematical theory of communication." *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1, pp. 3-55.
- » Shih, Stephanie. (2011). "Random Forests for Classification Trees and Categorical Dependent Variables: an informal Quick Start R Guide". *Stanford University | University of California, Berkeley*.
- » Torres-Bugarín O, Fernández-García A, Torres-Mendoza BM, Zavala-Aguirre JL, Nava-Zavala A, Zamora-Perez AL(2009) Genetic profile of overweight and obese school-age children, *Toxicological & Environmental Chemistry*, 91:4, 789-795, DOI: 10.1080/02772240802404966
- » Umbuzeiro GA, Heringa M, Zeiger E. (2017). "In Vitro Genotoxicity Testing: Significance and Use in Environmental Monitoring". *Adv Biochem Eng Biotechnol.*; 157:59-80. Doi: 10.1007/10\_2015\_5018
- » Yang, S., Luo, P., Loy, C. C., & Tang, X. (2015). "From facial parts responses to face detection: A deep learning approach". In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3676-3684).