# Minería de datos para la predicción en entornos de gestión que manejan datos de alta dimensión: un estudio experimental

Tesis en opción al título de Máster en
Matemática Aplicada e Informática para la Administración

Yoisel Campos Hidalgo

UNIVERSIDAD DE HOLGUÍN
OSCAR LUCERO MOYA


**Universidad de Holguín "Oscar Lucero Moya"**

**Facultad de Informática y Matemática**

# Minería de datos para la predicción en entornos de gestión que manejan datos de alta dimensión: un estudio experimental

**Tesis en opción al título de Máster en Matemática Aplicada e Informática para la Administración**

**Autor:** Ing. Yoisel Campos Hidalgo

**Tutor:** Dr.C. Reyner Pérez Campdesúñer

**Consultante:** Dr.C. Carlos Morell Pérez

Abril de 2012

# Resumen

Los entornos de gestión cuentan con datos generados por sus procesos para la ejecución de sus actividades. Los avances tecnológicos actuales han traído consigo contextos que típicamente trabajan con datos descritos por muchas variables. En particular, existen varios entornos de gestión que basan parte de sus actividades en predicciones realizadas a partir de los datos. La mejora continua de estos entornos de gestión depende significativamente, por tanto, en la mejora de los indicadores que dependen de la exactitud de la predicción. El análisis o minería de esos datos para realizar predicciones es afectado por la gran cantidad de variables presentes, en un fenómeno conocido como "maldición" de la alta dimensión. En la práctica se han identificado tres estrategias generales para lidiar con tal fenómeno. Estas se encuentran a disposición del gestor para la mejora continua de los procesos de su organización, cuando estos incluyen tareas de predicción. Las tres estrategias no han recibido igual atención ni se han desarrollado al mismo ritmo. Existe insuficiente evidencia que permita determinar la factibilidad de la estrategia de predicción mediante ensamblados de modelos, para mejorar los indicadores de gestión en procesos que trabajan con datos de alta dimensión. A través de un estudio experimental, esta investigación aporta la evidencia necesaria para afirmar que la combinación de varios modelos en un ensamblado es la estrategia más prometedora para la mejora continua de los entornos de gestión, en presencia de tareas de predicción a partir de datos de alta dimensión.

# Abstract

Most management environments handle data originated through out their activities. Modern technological innovations have induced the appearance of contexts which tipically work with data described by many variables. In particular, several management environments base most of their activities upon predictions deduced from the data. Hence, the continuous improvement of such environments significantly depends on the improvement of those indicators heavily correlated with the prediction accuracy. Such data mining or analysis for prediction is affected by the high number of variables involved, a phenomenon known in the specialized literature as the “curse” of dimensionality. In practice, three general strategies have been identified to deal with this phenomenon. These strategies are available to every manager to approach the continuous improvement of the organization processes where prediction tasks are present. All three strategies have not received the same attention nor have they developed at the same pace. There is insufficient evidence to determine whether or not the prediction by classifier ensembles is a feasible strategy, towards the improvement of those processes wich handle high dimensional data. By means of an experimental study, this research provides enough evidence to claim that combining multiple models in an ensemble is the most promising strategy for the continuous improvement of management enviroments, when dealing with prediction tasks based on high dimensional data.

# Agradecimientos

Al Dr. Carlos Morell, mentor en mi camino hacia la inteligencia artificial y la investigación científica.

A la Dra. Rosa Isabel Urquiza por su sentido de pertenencia y sus siempre valiosos puntos de vista.

A la MSc. Yaimara Andino por su inapreciable ayuda con el enfoque de gestión basado en procesos.

# Tabla de contenidos

# INTRODUCCIÓN

La toma de decisiones en condiciones de incertidumbre con frecuencia lleva implícita una tarea de estimación o predicción. No siempre las posibles alternativas para tomar la decisión son conocidas en todos sus detalles, de ahí que sea imperativo basar la elección final en estimaciones y aproximaciones de las características del estado futuro al que conduciría determinada decisión.

Una estimación o predicción fiable requiere de un conocimiento que le brinde soporte. Es decir, requiere de un conjunto de leyes, reglas, principios y modelos sobre el dominio donde ocurre la predicción o estimación. Existen contextos en los cuales se ha logrado establecer explícitamente un sistema de conocimientos. Sin embargo, en los contextos que por su novedad se encuentran aún bajo estudio, usualmente el conocimiento no ha sido sistematizado y validado, lo cual afecta directamente la fiabilidad de las predicciones.

En estos casos, el empirismo puede ser una fuente de estimaciones acertadas, siempre y cuando exista alguna experiencia previa acumulada sobre el contexto donde ocurre la toma de decisiones. Dicha experiencia puede ser utilizada para inducir un conocimiento general a partir de casos observados previamente. Este lógicamente será un conocimiento parcializado, debido a que procede solamente de la experiencia con que se cuenta, la cual no tiene que ser necesariamente representativa de la generalidad de situaciones que pueden presentarse en el proceso. No obstante, la predicción a partir de la experiencia previa acumulada permitirá tomar decisiones en correspondencia con éxitos y fracasos anteriores.

Los avances tecnológicos en el tratamiento de la información y en las comunicaciones, han abierto las puertas a la acumulación de gran cantidad de datos operacionales sobre los distintos procesos que tienen lugar en distintos ámbitos del quehacer humano. Estos datos son la fuente principal y más

evidente de experiencia empresarial, científica y técnica acumulada sobre diversos dominios.

Existen determinados contextos que, por su complejidad, incorporan una cantidad tal de variables que los datos que arrojan son prácticamente imposibles de analizar por un gestor humano para realizar predicciones empíricas. El experto humano solo llegará a utilizar e interrelacionar en su mente una parte de las variables en el mejor de los casos, dejando sin explotar gran parte de los datos disponibles [5]. Utilizarlos todos sería o bien demasiado complejo o bien intratable en el tiempo del que dispone para tomar la decisión necesaria. Estos tipos de datos son conocidos en la literatura especializada como datos de “alta dimensión” [3], pues desde el punto de vista algebraico son datos que requieren una gran cantidad de dimensiones para ser representados.

Los avances tecnológicos actuales han traído consigo algunos ejemplos de dominios que típicamente arrojan datos de muy alta dimensión:

- Los estudios genéticos [7-10].
- Los estudios bioquímicos [11-15].
- Los estudios espectrométricos [16-20].
- Varias aplicaciones del reconocimiento de patrones en imágenes y señales, como el reconocimiento facial, de voz y de la escritura a mano [26].

Los datos acumulados por procesos empresariales que trabajan en estos dominios, poseen además otra importante potencialidad: son datos estructurados, almacenados digitalmente, susceptibles de ser procesados por máquinas computadoras para lograr determinado nivel de automatización en el análisis, en la predicción y en la toma de decisiones. La profundidad con la que se aplica dicha automatización varía de contexto en contexto, pues depende en primer lugar de la disponibilidad de datos operacionales almacenados. Una aplicación de cierta popularidad es el diseño e implantación de herramientas de procesamiento analítico en línea (OLAP, por sus siglas en inglés). Estas

persiguen como objetivo facilitar una visión más útil de los datos almacenados, de forma que pueda realizarse un análisis más efectivo y arribar más rápidamente al conocimiento necesario para tomar determinada decisión.

Una aplicación más profunda de la computación al procesamiento de datos es el conjunto de técnicas que se ha dado a conocer bajo el término genérico de "minería" de datos [2, 4]. Estas pretenden ir un paso más allá en el apoyo a la toma de decisiones, encontrando el conocimiento implícito que subyace en los datos con respecto a cierto problema, y poniendo dicho conocimiento a la disposición del gestor humano.

Este conocimiento extraído de los datos es obtenido por lo general en la forma de un modelo, abstracción simplificada del proceso bajo estudio, susceptible de ser empleado en la solución del problema del cual depende la decisión que se tomará. La Inteligencia Artificial es la ciencia de la computación que estudia las formas de representar el conocimiento mediante estructuras computacionales [5]. Ello implica una fuerte intervención de varias técnicas de Inteligencia Artificial en la minería de datos. En el caso particular de la predicción basada en experiencia previa acumulada, la minería de datos constituye una alternativa muy adecuada, ya que permitiría obtener modelos de predicción autónomo y, además, acorde a la experiencia disponible.

El conjunto específico de técnicas capaces de construir un modelo a partir de experiencia previa almacenada de forma estructurada, es objeto de estudio de la disciplina conocida como "aprendizaje automatizado" (*machine learning*) [6]. Estas técnicas "entrenan" un modelo mediante casos conocidos (experiencia previa) para que sea capaz de resolver determinada tarea de estimación, asociación, predicción o clasificación. De este modo se dice que el modelo "aprende" a realizar la tarea para la que fue entrenado.

En un entorno de gestión con pocas variables para tareas como la predicción podrían utilizarse algoritmos y métodos deterministas, que son también susceptibles de ser implementadas en máquinas computadoras. Sin embargo, al lidiar con decenas (o cientos, o miles) de variables la explosión combinatoria

hace que las soluciones deterministas se vuelvan intratables. Por otro lado, los modelos lineales y otros modelos paramétricos que suelen utilizarse en las variantes deterministas, no siempre representan adecuadamente la complejidad de los datos. Estas razones apuntan a que las tareas de predicción para la toma de decisiones, a partir de datos de alta dimensión, requieren un mayor empleo de técnicas estocásticas y no paramétricas en general.

La importancia de realizar un proceso fiable de minería en datos de alta dimensión, ha atraído la atención de la comunidad de investigadores en *machine learning* y disciplinas afines. A partir de la última década del pasado siglo, comenzó a ser evidente que los algoritmos tradicionalmente efectivos para la minería de datos presentaban problemas en su desempeño, al enfrentarse a datos de alta dimensión [21-25]. El estudio de las causas de esta irregularidad puso en evidencia la ocurrencia de determinados fenómenos que han sido bautizados en la literatura como “maldición” de la alta dimensión [3].

El interés inspirado por este tema lo ha convertido en centro del debate en algunos eventos especializados, como “*Knowledge Discovery in Databases*” (KDD) 2001, motivando la organización de competencias dedicadas completamente a la minería de datos de alta dimensión, como incentivo para la investigación en el tema. La primera y más destacada de dichas competencias se organizó durante el evento “*Neural Information Processes and Systems*” (NIPS) 2003 [26], centrada en el estudio de técnicas para reducir la cantidad de dimensiones e intentar aliviar el problema. Recientemente se organizó otra competencia con objetivos similares durante el evento “*Rough Sets and Current Trends in Computing*” (RSCTC) 2010 [27], con abundancia de participantes e interesantes resultados. En estas competencias se han presentado además propuestas de adaptaciones a los algoritmos tradicionales, así como un paradigma diferente para enfrentar la minería de datos de alta dimensión: la construcción de un conjunto de modelos, en lugar de un modelo individual. Esta técnica, conocida como “ensamblado” de modelos, en la cual uno o varios algoritmos de minería de datos entrenan un conjunto de modelos distintos entre

sí para combinar sus decisiones, de forma análoga a un panel de jueces o un jurado, ha ganado atención en la última década [28-31] por sus potencialidades.

Al estudiar las publicaciones relacionadas con la construcción de modelos de predicción a partir de datos de alta dimensión, es posible observar la evolución, durante la última década, del aprendizaje automatizado y demás disciplinas de la Inteligencia Artificial relacionadas con la minería de datos.

No debe soslayarse que toda estimación o predicción trae aparejado un margen de error, que por tanto afecta directamente su fiabilidad para la toma de decisiones. Una preocupación constante de toda organización es la mejora de su gestión. Algunos modelos de calidad, como la familia de normas ISO 9000, sugieren que el enfoque de la gestión por procesos contribuye con esta idea. Los procesos que basan parte de sus actividades en predicciones presentan una interrelación peculiar entre sus indicadores de gestión. Es común que, en entornos de gestión con procesos de predicción, la exactitud (o su complemento, el error) de la predicción constituya uno de los indicadores de gestión principales. La mejora de otros indicadores de importancia puede depender de que las predicciones se realicen con exactitud suficiente.

Esto puede ser especialmente sensible en muchos contextos, donde los errores pueden tener un costo estratégico considerable. Por ejemplo, en los dominios mencionados más arriba puede estar en juego el diagnóstico del padecimiento genético de una persona, o la determinación de cierta acción biológica de una molécula que la convierta en un candidato a fármaco para cierta dolencia, o el descubrimiento de un nuevo yacimiento mineral a partir del análisis de imágenes espectroscópicas satelitales.

La mejora del indicador referente a la exactitud de la predicción, en entornos de gestión que utilizan datos de alta dimensión, ha girado entorno a tres estrategias generales. La primera y más intuitiva es intentar reducir la cantidad de variables que describen a los datos, sin incurrir en una pérdida significativa de información. La segunda es realizar adaptaciones y modificaciones a los métodos, de forma que se comporten de manera más robusta ante datos de alta

dimensión. La tercera estrategia es utilizar ensamblados de varios modelos, cuyas predicciones individuales son combinadas para arrojar una predicción final.

Es necesario destacar que las tres estrategias señaladas no han recibido igual atención ni se han desarrollado al mismo ritmo. La primera (reducción de la cantidad de variables) es objeto de estudio de la selección de rasgos hace varias décadas [32]. Esto la posicionó como la estrategia más explotada desde un inicio, por la experiencia y fundamentos teóricos acumulados. La segunda estrategia está limitada por las características propias de cada algoritmo, provocando que sus resultados sean muy aislados. En cambio, la tercera (utilización de ensamblados como estrategia para la minería de datos de alta dimensión) no cuenta, al menos hasta mediados de 2011, con eventos propios o competencias que sistematicen el estudio de sus efectos y revelen sus potencialidades. En particular, no existe un consenso sobre cuáles algoritmos de minería de datos se benefician al enmarcarlos en un ensamblado, ni tampoco qué características debe tener dicho ensamblado para beneficiar realmente a los algoritmos.

Esta situación revela un **problema cientìfico**: existe insuficiente evidencia que permita determinar la factibilidad de la predicción mediante ensamblados de modelos, como estrategia para mejorar los indicadores de gestión en procesos que trabajan con datos de alta dimensión. El problema descrito motiva la presente investigación, que tiene como **objeto de estudio** la gestión de procesos de minería de datos para la predicción.

Para dar solución al problema se persigue el **objetivo** de evaluar, a través de estudios experimentales, cuánto influye la estrategia de ensamblado de modelos en la mejora del indicador de gestión relacionado con la exactitud de la predicción, durante procesos de minería con datos de alta dimensión.

De modo que la investigación queda enmarcada en el **campo de acción** de la evaluación de las estrategias de minería de datos para la predicción, basada en experiencia previa estructurada en datos de alta dimensión.

El estudio de la situación problèmica descrita, bajo los principios del objetivo definido, suscita varias **preguntas científicas**:

1. ¿Cómo se interrelacionan los indicadores de gestión más relevantes en procesos donde intervienen tareas de predicción?
2. ¿Cuáles son los prerrequisitos y características indispensables para el buen desempeño de un algoritmo de minería de datos a través del aprendizaje automatizado?
3. ¿En qué consisten los fenómenos que afectan la minería en datos de alta dimensión?
4. ¿Cómo enfrentan la “maldición” de la alta dimensión las tres estrategias generales mencionadas?
5. ¿Cómo llevar a cabo un estudio experimental que aporte la evidencia necesaria para resolver el problema?
6. ¿Cuánto influyen las diferentes técnicas de la estrategia ensamblado de modelos en la mejora del indicador de gestión relacionado con la exactitud de la predicción?

Las siguientes **tareas** de investigación han sido definidas para responderlas y darle cumplimiento al objetivo:

1. Analizar, con un enfoque basado en procesos, la gestión de la minería en datos de alta dimensión.
2. Estudiar y resumir los fundamentos teóricos del aprendizaje automatizado para la construcción de modelos de predicción.
3. Sistematizar los fenómenos asociados a la “maldición” de la alta dimensión y las estrategias para enfrentarla.
4. Diseñar un estudio experimental comparativo entre las estrategias más prometedoras para enfrentar este tipo de minería de datos, en correspondencia con los estándares actuales para experimentación en *machine learning*.

5. Llevar a cabo los experimentos diseñados.
6. Realizar un análisis crítico sobre los resultados de la experimentación, que aporte la evidencia necesaria para darle solución al problema científico planteado.

Entre los **métodos** teóricos de investigación utilizados se cuenta el analítico-sintético, tanto en la concepción de la investigación como para el análisis de los fundamentos teóricos y los resultados finales. También el enfoque sistémico y la modelación, para la concepción del proceso de descubrimiento del conocimiento, el análisis de la gestión con enfoque de procesos y el diseño de un marco de aplicación de la estrategia de ensamblados. Entre los métodos empíricos hay un fuerte componente de experimentación, así como métodos estadísticos para la interpretación de los resultados experimentales.

Como **resultado** de esta investigación se ofrece la caracterización de las técnicas que muestran el mejor aporte a la optimización de los indicadores de gestión en procesos de minería con datos de alta dimensión, así como el diseño de un marco para el entrenamiento de un ensamblado de modelos de predicción, susceptible de ser utilizado en un problema real específico.

En el **Capítulo 1** se encuentran recogidos en detalle los paradigmas, técnicas, algoritmos, fenómenos y estrategias que sirven como fundamentos teóricos a esta investigación, respondiendo las preguntas científicas 1, 2, 3 y 4 y dando cumplimiento a las tareas 1, 2 y 3 de la investigación. El **Capítulo 2** detalla el diseño del estudio experimental realizado, los datos y algoritmos empleados, así como las técnicas estadísticas empleadas para la interpretación de los resultados. El análisis de dichos resultados conduce a la formulación de las conclusiones generales de la investigación y las recomendaciones que el autor y tutor consideraron pertinentes.

# CAPÍTULO 1
# Fundamentos teóricos

En el presente capítulo son sintetizadas las características de los datos de alta dimensión y cómo estas influyen en las tareas de predicción. Los principios del descubrimiento de conocimiento y su relación con la gestión de procesos de minería de datos, específicamente para tareas de predicción, son también detallados. Esta relación es ilustrada mediante la descripción de un entorno de gestión representativo y sus indicadores, con un enfoque de gestión basada en procesos.

El propósito de la minería de datos es descubrir y representar el conocimiento implícito subyacente en los datos. La Inteligencia Artificial es la ciencia que estudia las formas de reproducir, en máquinas computadoras, procesos y actividades inherentes del intelecto humano, como el aprendizaje, la inferencia y la representación del conocimiento. Esta es la razón fundamental por la cual las disciplinas que abordan la minería de datos hacen uso directo de técnicas y paradigmas de la Inteligencia Artificial.

Las particularidades de los datos de alta dimensión son también detalladas, especialmente las razones que dificultan la minería convencional en estos datos. Se profundiza además en las direcciones que han tomado las investigaciones para enfrentar la adquisición de conocimiento a partir de tales datos, a pesar de la “maldición” de la alta dimensión. Estas estrategias no han recibido igual atención en cuanto a su aplicación en entornos de gestión que predicen a partir de datos de alta dimensión. Su estudio arroja perspectivas esperanzadoras para la mejora continua en los indicadores de gestión de los procesos de minería de datos. Estas perspectivas son exploradas a fondo en el Capítulo 2 del presente informe.

## 1.1. Entornos de gestión con datos de alta dimensión

Para realizar predicciones, con frecuencia es necesario contar con determinado conocimiento; esto es: algunas leyes, reglas o principios que sirvan de base a la acción de predecir. Cuando dicho conocimiento no está disponible explícitamente, pero se cuenta con un cúmulo de datos sobre predicciones anteriores, se impone entonces la modelación del conocimiento implícito que subyace en dichos datos.

En la ejecución cotidiana de sus actividades, las organizaciones de la era de la información utilizan datos provenientes de diferentes procesos. Las características de estos datos influyen, en ocasiones de forma decisiva, en su utilización para la toma de decisiones y la mejora continua de la gestión.

Existen determinados contextos que, por su complejidad, incorporan una cantidad tal de variables que los datos que arrojan son prácticamente imposibles de analizar por un gestor humano para realizar predicciones empíricas. Estos tipos de datos son conocidos en la literatura especializada como datos de “alta dimensión” [3], pues desde el punto de vista algebraico son datos que requieren una gran cantidad de dimensiones para ser representados. Los avances tecnológicos actuales han traído consigo algunos ejemplos de dominios que típicamente arrojan datos de muy alta dimensión:

- Los estudios genéticos [7-10], donde el grado de expresión de cada gen en una secuencia de ADN es cuantificado y considerado una variable. La cantidad de genes incluidos varía de estudio en estudio, desde unos pocos miles (2 o 3 mil) hasta 50 o 60 mil genes. Estos datos cuentan con la complejidad adicional de registrar pocos casos, por el elevado costo asociado a la secuenciación genética. Por lo general, la cantidad de casos registrados oscila alrededor de 200.
- Los estudios bioquímicos [11-15], basados en relaciones cuantitativas entre la estructura química y la actividad biológica de las moléculas (QSAR por sus siglas en inglés). El objetivo de dichos estudios es

aprender a predecir la influencia de las sustancias químicas sobre determinada función biológica, lo cual es la base para el descubrimiento molecular y la innovación de fármacos. Las moléculas son representadas mediante una multitud de descriptores químicos estructurales, espaciales, electrostáticos, lipídicos, entre otros, de manera que una molécula puede ser descrita por una cantidad de variables que usualmente oscila entre alrededor de mil y alrededor de 100 mil. Al igual que los estudios genéticos, la obtención de una muestra de moléculas resulta altamente costosa en tiempo y en recursos, lo cual ocasiona que estos conjuntos de datos cuenten solo con unos pocos cientos de casos.

- Los estudios espectrométricos [16-20]; en los cuales se realiza el análisis espectral de determinadas sustancias, o de la superficie terrestre desde un satélite. Persiguen el objetivo de detectar la presencia de determinados elementos químicos que revelen la presencia de una enfermedad, de un yacimiento mineral, de un determinado tipo de cultivo o de potenciales agentes contaminantes en la atmósfera, el suelo o el agua, entre otros. El análisis espectral se compone de las mediciones de la intensidad de la luz en varias longitudes de onda, considerándose cada longitud de onda como una variable. En dependencia del tipo de estudio, la cantidad de variables oscila entre alrededor de 200 y alrededor de 50 mil.
- Varias aplicaciones del reconocimiento de patrones en imágenes y señales, como el reconocimiento facial, de voz y de la escritura a mano.

Estos son contextos en los cuales la predicción es una tarea constante, con un rol fundamental en la consecución de sus objetivos. El análisis de los datos para tareas de predicción requiere la identificación de patrones y generalidades, que conllevan a sistematizar un cúmulo de conocimientos sobre el entorno de gestión. Las ciencias de la computación estudian este proceso bajo el nombre de “descubrimiento de conocimiento en datos”.

### 1.1.1. Descubrimiento de conocimiento en datos

Para realizar predicciones, con frecuencia es necesario contar con determinado conocimiento; esto es: algunas leyes, reglas o principios que sirvan de base a la acción de predecir. Cuando dicho conocimiento no está disponible explícitamente, pero se cuenta con un cúmulo de datos sobre predicciones anteriores, se impone entonces la modelación del conocimiento implícito que subyace en dichos datos.

Este proceso ha sido estudiado y sistematizado en los últimos lustros mediante el "descubrimiento de conocimiento en bases de datos" (*Knowledge Discovery in Databases*, KDD) [33]. El término fue acuñado alrededor de 1989 y desde entonces ha adquirido mucha atención.

KDD ha sido definido por sus principales promotores (Fayyad y Piatetsky-Shapiro) como:

"el proceso no trivial de identificar, en los datos, patrones válidos, novedosos, comprensibles y potencialmente útiles" [2].

Entiéndase por "patrón" un modelo aplicable a los datos, una descripción general del conjunto de datos disponibles. "No trivial" significa que alguna búsqueda o inferencia está implicada en el descubrimiento, o sea, no se trata del cálculo de cantidades predefinidas, como por ejemplo la media aritmética.

Por otro lado, el término "proceso" implica que KDD está compuesto por varios pasos. Ha existido debate en la comunidad científica alrededor del tema de cuáles pasos componen el proceso, aunque generalmente las propuestas abordan los mismos puntos. En la Figura 1 se muestran distintas representaciones del proceso. Los principales autores ofrecen un marco amplio que consta de 9 pasos:

1. Comprensión del dominio de donde provienen los datos y definición de la meta del proceso.
2. Creación del conjunto de datos objeto de exploración

3. Limpieza y pre-procesamiento de los datos
4. Reducción y proyección de los datos
5. Especificación del método de minería a utilizar (clasificación, regresión, agrupamiento, asociación)
6. Elección del algoritmo de minería de datos que se empleará
7. Realización de minería en los datos, buscando o construyendo patrones de interés
8. Interpretación de los patrones obtenidos: representación del conocimiento
9. Consolidación del conocimiento descubierto, incorporándolo en un sistema que haga uso de este para el objetivo que se definió al inicio del proceso

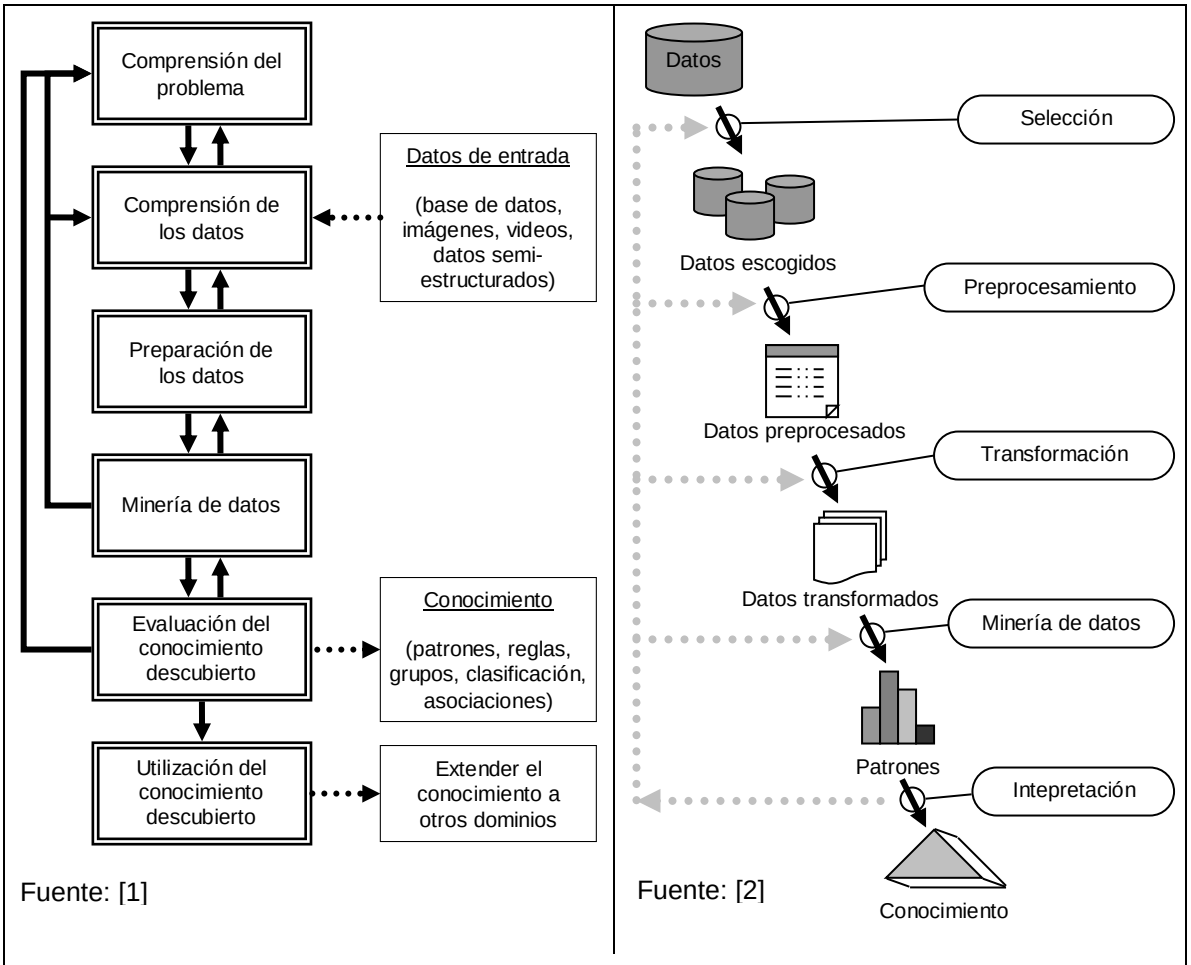


**Figura 1**. Distintas representaciones del proceso KDD según sus autores.

En cuanto al uso del término “minería de datos”, también ha existido debate. Ha sido utilizado ampliamente para referirse al proceso de KDD completo, cuando realmente es solo un paso del proceso. Durante la “minería” ocurre la búsqueda e inferencia de patrones y conocimiento en los datos. Sin embargo, el resto de los pasos puede tener una importancia crítica en la calidad de la minería en sí misma, así como del conocimiento descubierto. Al respecto de este debate, los autores de KDD han ofrecido [2] algunas ideas generales para conciliar los diferentes enfoques, explicando que el paso de minería de datos es el que ha recibido la mayor atención, siendo incluso objeto de estudio de otras disciplinas, como el aprendizaje automatizado.

En el caso particular de los datos de alta dimensión, el descubrimiento de conocimientos se ve afectado por un conjunto de fenómenos que se detallan a continuación.

### 1.1.2. La “maldición” de la alta dimensión

En sus inicios, el estudio de algoritmos para la estimación y la predicción tenían en cuenta una cantidad de variables que resultaba intuitiva a la comprensión humana. Modelos de 2 y 3 variables son fáciles de representar y analizar en el espacio natural de 3 dimensiones, asignando una dimensión del espacio a cada variable y representando los objetos o casos como puntos en dicho espacio. Sin embargo 4, 5 o más variables no resultan ni tan intuitivas, ni tan fáciles de representar en el espacio natural. De hecho, en fechas tan tempranas como 1948, un modelo de 4 o 5 variables era considerado “de alta dimensión”[57], pues **su representación requiere de un espacio vectorial con más dimensiones de las que es posible analizar intuitivamente**.

Más tarde, hacia 1961[58], comenzó a utilizarse la frase de “maldición de la alta dimensión”, refiriéndose al efecto que ejerce, sobre el proceso de búsqueda, el aumento en la cantidad de dimensiones del espacio de representación. Especialmente, la búsqueda exhaustiva se vuelve rápidamente intratable con el crecimiento de la cantidad de dimensiones, al necesitarse $2^d$ evaluaciones para optimizar una función de $\boldsymbol{d}$ variables binarias.

Este efecto se ve acentuado con la aparición, en los últimos lustros, de dominios que arrojan conjuntos de datos descritos por miles de variables [28]. La minería de estos datos modernos se convierte, entonces, en ideal caldo de cultivo para algoritmos basados en búsqueda heurística, computación blanda y métodos no paramétricos. Además, se han convertido en una buena razón para revisar los algoritmos tradicionalmente efectivos, adaptarlos, hacerlos más robustos e incluso para diseñar nuevos esquemas, nuevos algoritmos [23, 25, 59, 60].

El estudio de los efectos provocados por el crecimiento en la cantidad de variables que describen los datos, ha revelado la ocurrencia de ciertos fenómenos en los espacios de alta dimensión. El conjunto de dichos fenómenos ha sido recogido [3] también bajo la nomenclatura de “maldición” de la alta dimensión y se sintetizan a continuación:

- La concentración de las distancias. Las métricas tradicionales utilizadas para mediciones de distancia entre dos puntos en un espacio vectorial, por ejemplo la norma Euclidiana, arrojan mediciones menos intuitivas cuando son utilizadas en espacios de cientos o miles de dimensiones. A medida que crece la cantidad de dimensiones, crecen también los órdenes de magnitud de las distancias medidas. Para dimensiones muy altas, todas las mediciones de distancia entre dos puntos cualesquiera de un conjunto de datos son tan grandes (o pequeñas) en magnitud, que la noción de “cerca” o “lejos” se distorsiona. Ocurre que la distancia de un punto a su vecino más próximo se asemeja a la distancia al punto más lejano. Por ello se dice que las distancias entre puntos vecinos se concentran alrededor de un valor, mostrando una varianza muy pequeña. Lógicamente, una métrica de distancia que arroja siempre aproximadamente el mismo valor, sean cuales sean los puntos, es inútil como criterio de similitud entre objetos de un conjunto de datos, afectando a muchos algoritmos geométricos y basados en la búsqueda de vecinos.

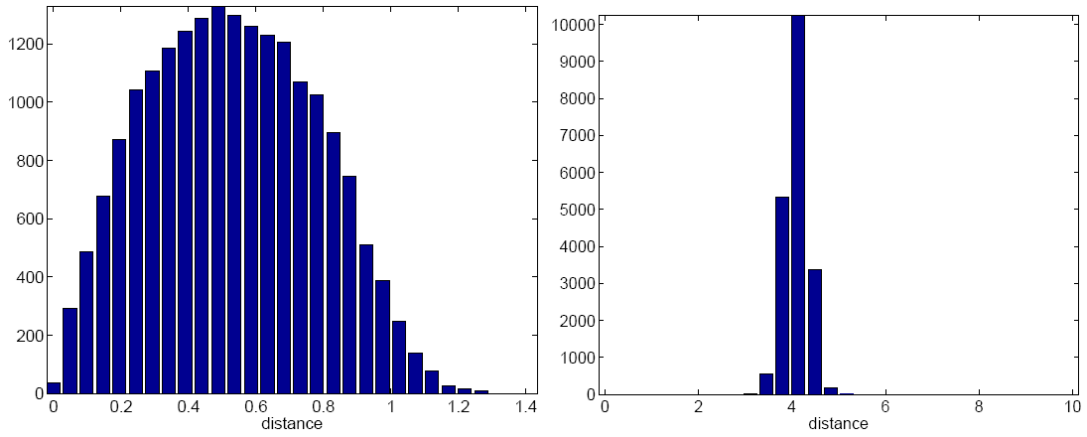


**Figura 6**. Histograma de distancias en espacios de datos con 2 (izquierda) y 100 (derecha) dimensiones, poblados con 250 puntos distribuidos uniformemente. Fuente: [3].

- El fenómeno del espacio vacío. Con el crecimiento de la cantidad de dimensiones, se necesitan muchos más puntos de datos para cubrir de forma densa y uniforme una región del hiper-espacio. Si en un espacio de 2 dimensiones se considera que 100 puntos de datos son suficientes para cubrir densamente una unidad cuadrada, al pasar a 3 dimensiones se necesitarán 1000 puntos. En un espacio de ***n*** dimensiones se necesitarán $10^n$ puntos para cubrir densamente la misma región. Sin embargo, por lo general no se cuenta con tantos datos, por tanto la región bajo estudio estará mucho menos densamente cubierta. Para miles de dimensiones, el espacio parecerá prácticamente vacío, a pesar de que se cuente con muchos casos en el conjunto de datos, pues no serán suficientes de ninguna manera. Esto afecta severamente la estimación de la distribución

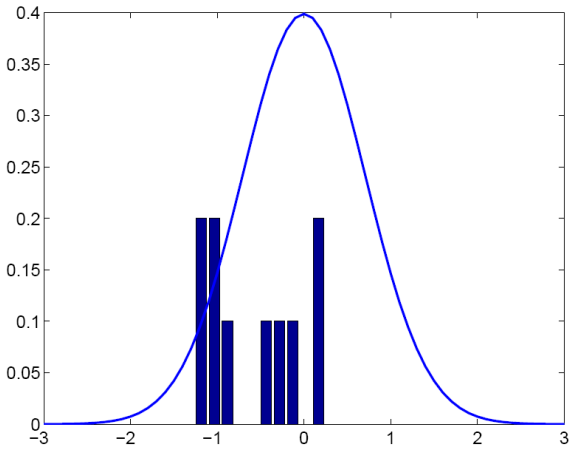


**Figura 7**. Aproximación de la función Gaussiana de distribución probabilística mediante un histograma, en presencia de insuficientes puntos. Fuente: [3].

probabilística de los datos, lo cual es directa o indirectamente la base de muchas técnicas de minería de datos, como las estadísticas y las que aproximan funciones no lineales basándose en *kernels* como la distribución normal, también conocida como distribución Gaussiana.

- El fenómeno de Hughes. Bautizado con el nombre de Greg Hughes, el primero en describirlo [61], este fenómeno consiste en una variación drástica en la eficacia de un algoritmo de predicción al aumentar la cantidad de dimensiones. Entrenar un modelo con muy pocas variables puede resultar en un desempeño pobre, pues podrían resultar insuficientes para discriminar entre las clases del problema. A medida que se añaden variables, alimentando el modelo con más información, la exactitud se incrementa. Pero al continuar adicionando variables y, por tanto, incrementando la cantidad de dimensiones, la exactitud alcanzará un máximo y caerá abruptamente, pues las nuevas variables solo traerán ruido y redundancia al entrenamiento, ante lo cual prácticamente todas las técnicas son sensibles.

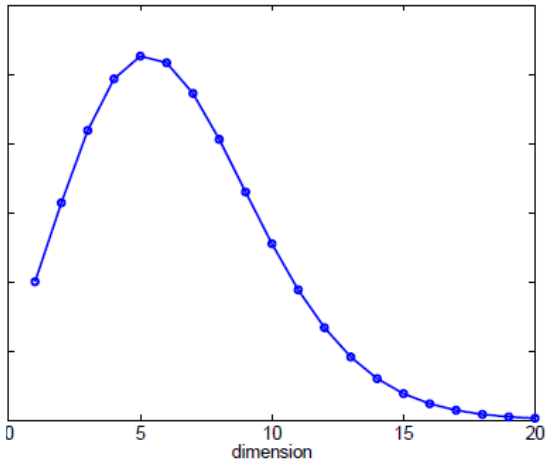


**Figura 8**. La exactitud de un modelo de predicción varía con la cantidad de variables, manifestándose en un gráfico típico. Fuente: [3].

El análisis a profundidad de estos fenómenos y sus efectos sobre el aprendizaje automático de modelos para la predicción, ha incidido en el curso que han seguido las investigaciones en los últimos lustros. Un estudio del estado del arte sobre aprendizaje supervisado con datos de alta dimensión revela la aparición de tres estrategias generales para enfrentar el problema.

## 1.2. Gestión con enfoque de procesos a partir de datos de alta dimensión

En el epígrafe anterior se relacionaron varios contextos que se encuentran entre las actividades tecnológicas que arrojan datos de alta dimensión. Este tipo de organizaciones se enfrenta al descubrimiento de conocimiento en sus datos para realizar predicciones. La calidad de sus predicciones puede influir en todo un conjunto de indicadores de gestión y por tanto merece ser estudiada con más profundidad.

La familia de normas ISO 9000 para los “Sistemas de Gestión de la Calidad”, de la *International Standards Organization*, se sustenta en varios principios, que se presentan como pilares para implantar modelos de gestión orientados a obtener buenos resultados de manera eficaz y eficiente.

Según uno de los Principios de la Gestión de la Calidad recogidos en la norma ISO 9000:2000, “un resultado se alcanza más eficientemente cuando las actividades y los recursos relacionados se gestionan como un proceso” [54]. Este principio es precisamente el **enfoque basado en procesos**. Bajo este enfoque, una organización acomete la mejora continua de su gestión, a partir de la mejora de sus procesos. La propia norma define un proceso como “un conjunto de actividades mutuamente relacionadas o que interactúan, las cuales transforman elementos de entrada en resultados”.

El modelo EFQM (*European Foundation for Quality Management*) para la Excelencia Empresarial también incluye entre sus conceptos fundamentales la “Gestión por procesos y hechos” [54], considerando que permite a las organizaciones “actuar de manera más efectiva cuando todas sus actividades interrelacionadas se comprenden y se gestionan de manera sistemática”. Este modelo a su vez concibe un proceso como una “secuencia de actividades que van añadiendo valor mientras se produce un determinado producto o servicio a partir de determinadas aportaciones”.

Ambas visiones describen la existencia de transformaciones, las cuales son medidas con indicadores de gestión y manipuladas a partir de determinadas variables. La familia de normas ISO 9000 establece cuatro pasos generales para enfocar la gestión por procesos:

1. La identificación y secuencia de los procesos
2. La descripcion de cada uno de los procesos
3. El seguimiento y la medición para conocer los resultados que obtienen
4. La mejora de los procesos con base en el seguimiento y medición realizados.

Los entornos de gestión que manejan tareas de predicción como parte de sus procesos, encuentran en el análisis de sus datos un factor decisivo para la mejora de sus procesos. La importancia de la minería de datos para la mejora de los indicadores de gestión, y por tanto de los procesos, puede estudiarse mejor mediante un ejemplo.

### 1.2.1. Ejemplo: diseño de fármacos bajo el enfoque de gestión basado en procesos

Los estudios bioquímicos se encuentran entre las actividades tecnológicas que arrojan datos de alta dimensión. En particular, el descubrimiento de fármacos es un tipo de descubrimiento molecular, aplicado a estudios bioquímicos con el fin de obtener nuevos fármacos.

Se afirma que es una actividad cuya gestión implica un enorme costo en tiempo, esfuerzo y recursos, pues por cada fármaco que es aprobado para su uso médico, fueron probados al menos otros 10 en humanos, al menos otros 100 fueron considerados como candidatos viables y al menos otros 100 000 fueron sintetizados y verificados [53].

Si bien el descubrimiento de fármacos es una disciplina bien establecida desde la primera mitad del siglo XX, con la introducción de las máquinas computadoras esta actividad científica se ha transformado. En la actualidad, las organizaciones

de la industria farmacológica a nivel mundial incorporan o contratan a grupos de **descubrimiento de fármacos asistido por ordenador**, mejor conocidos por sus siglas en inglés: grupos CADD (*Computer Assisted Drug Design*).

Según expertos, un grupo CADD se dedica al "descubrimiento y desarrollo de nuevas entidades químicas de demostrada utilidad para la atenuación o curación de una enfermedad" [53]. Esta, además de ser muy costosa, es una actividad de gran complejidad técnica y organizativa. Involucra varios procesos llevados a cabo por equipos multidisciplinarios. Entre los especialistas que participan se encuentran médicos, microbiólogos, químicos, bioquímicos, biofísicos, farmacólogos, inmunólogos, parasitólogos, sicólogos, toxicólogos, enzimólogos, cristalografistas, espectroscopistas, informáticos y estadísticos. Todos cooperan en un entorno organizacional para la puesta en uso de nuevos fármacos, a través de tres fases fundamentales:

1. Descubrimiento y optimización de nuevas entidades químicas como candidatos farmacológicos.
2. Desarrollo clínico de candidatos farmacológicos.
3. Supervisión post-marketing de los resultados de aplicación limitada del nuevo fármaco.

Los grupos CADD son esencialmente organizaciones y como tal son gestionados. El **enfoque basado en procesos** puede aplicarse al desempeño de un grupo CADD.

En el descubrimiento de fármacos pueden identificarse 5 procesos clave:

1. Descubrimiento de nuevos candidatos farmacológicos
2. Optimización de candidatos
3. Desarrollo preclínico de candidatos
4. Desarrollo clínico de candidatos
5. Supervisión post-marketing

Así como otros procesos estratégicos (Planeación estratégica, Investigación) y de soporte (Síntesis molecular, Sistemas de Información).

El proceso de descubrimiento de nuevos candidatos estudia constantemente compuestos químicos que podrían tener algún efecto sobre determinada actividad biológica. Los compuestos son representados, descritos, clasificados y escogidos, para ser presentados como posibles bio-activos positivos.

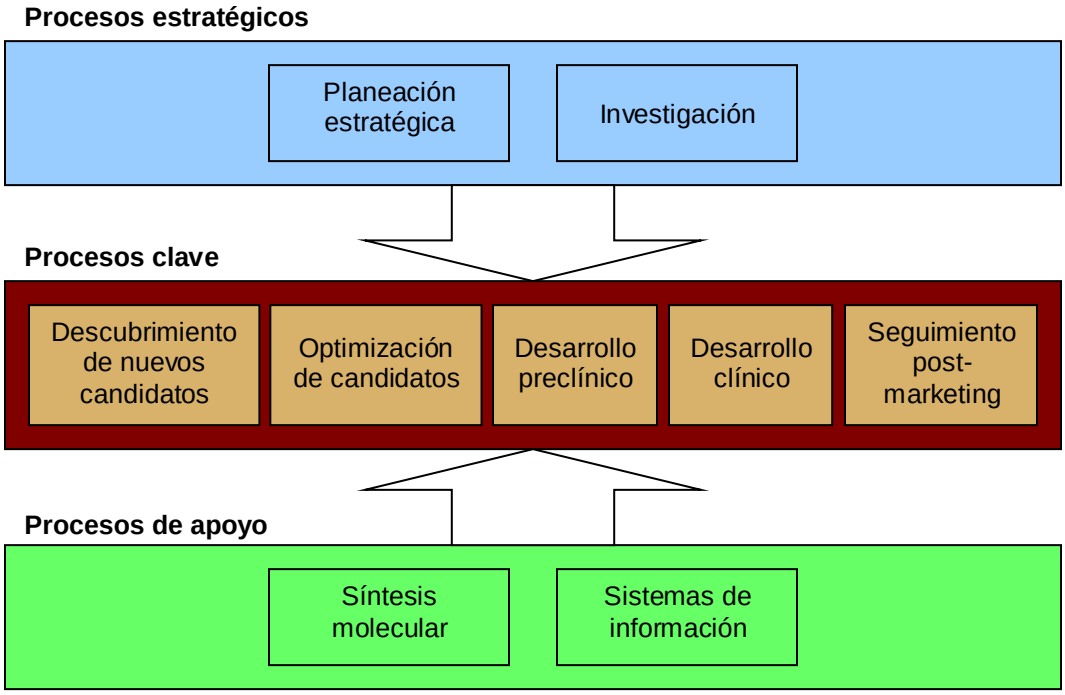


**Figura 4**. Mapa principal de procesos de un grupo CADD.

Durante la optimización de candidatos se valoran las posibilidades reales de sintetizar los compuestos moleculares prometedores y se construyen modelos de relación estructura-actividad (QSAR, *Quantitative Structure-Activity Relationship*), que permiten determinar con mayor confianza si un candidato es activo (positivo) o inactivo (negativo). Además podrían ayudar a descubrir o identificar más fácilmente buenos candidatos en el futuro. Como parte de estas tareas comienzan las primeras pruebas en laboratorio.

Los candidatos farmacológicos validados en los procesos anteriores son sometidos a desarrollo preclínico. Durante este se realizan pruebas de estabilidad, seguridad, toxicidad y otras, comenzándose los experimentos en tejidos vivos y en animales vivos. Este proceso exige la síntesis masiva de los compuestos moleculares a analizar, incurriendo ya en un gran costo.

Algunos candidatos pueden sobrepasar con éxito los procesos anteriores y ser considerados para su desarrollo clínico. Continúa la síntesis molecular de compuestos, con pruebas a pequeña escala sobre humanos sanos y enfermos, todo respaldado por un marketing dirigido a obtener posibles sujetos de prueba.

Por último, la supervisión post-marketing recolecta y analiza los resultados de las pruebas clínicas, arrojando conclusiones sobre la viabilidad de los compuestos químicos estudiados.

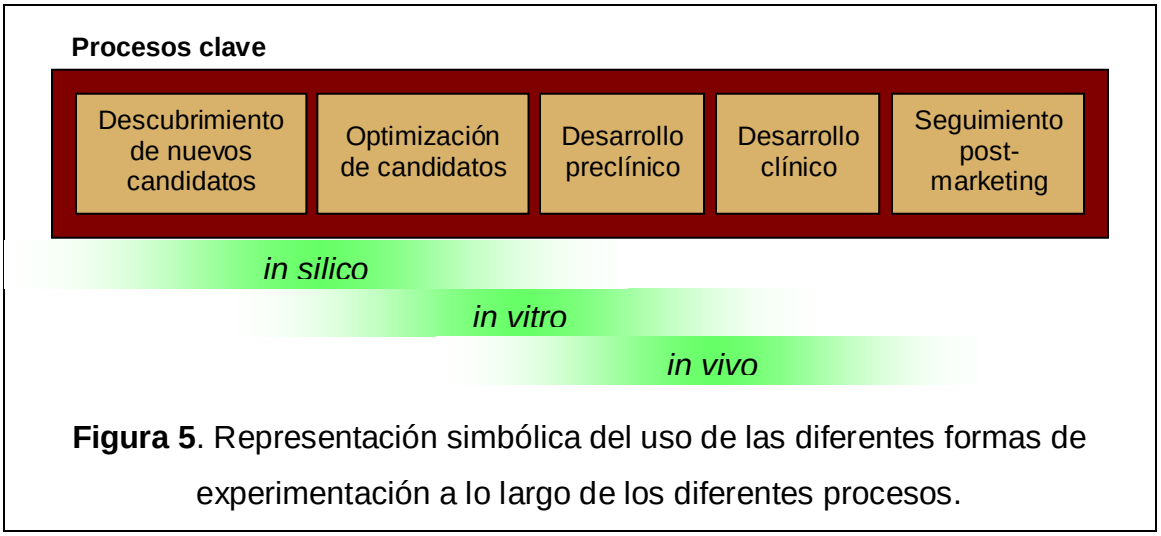


**Figura 5**. Representación simbólica del uso de las diferentes formas de experimentación a lo largo de los diferentes procesos.

A lo largo de estos procesos, el grupo CADD hace un uso extenso e intenso de la experimentación. Se distinguen 3 marcos experimentales fundamentales [55]:

1. La experimentación *in vitro*. Proviene de la alocución latina “en el vidrio” y se refiere a los estudios experimentales que tienen lugar en recipientes y ambientes controlados de laboratorio. Por lo general, estudian la interacción bioquímica de diferentes compuestos, esperando encontrar pistas de cómo podrían interactuar en un organismo vivo. Este tipo de experimentación es utilizada con mayor frecuencia en los procesos de optimización de candidatos y desarrollo preclínico. Suele demorar de días a semanas en ofrecer resultados verificables.

2. La experimentación *in vivo*. Tiene raíz en la alocución latina “en el ser vivo”, se trata del marco experimental que estudia fenómenos biológicos que se efectúan y observan en órganos y organismos vivos. Es el tipo de

experimentación más utilizada durante el proceso de desarrollo clínico de candidatos farmacológicos, aunque también se emplea en el desarrollo preclínico. Pretende someter a organismos vivos a la acción de los compuestos químicos bajo análisis, para estudiar su interacción. Suele demorar de semanas a meses, o incluso años, en ofrecer resultados verificables.

3. La experimentación *in silico*. Proviene del latín y se interpreta como “en el silicio”, refiriéndose al metal semiconductor que constituye el principio de funcionamiento de las máquinas computadoras actuales. Lógicamente, es un marco experimental que nació con el desarrollo de la computación y se basa fundamentalmente en la simulación, la modelación simbólica y el análisis de datos. Interviene sobre todo en los procesos de descubrimiento y optimización de los candidatos farmacológicos, donde la descripción, clasificación, modelación y predicción de las interacciones estructura-actividad de las moléculas se beneficia enormemente de las capacidades de cálculo y almacenamiento de las computadoras. Estos beneficios conllevan a que la duración de la experimentación *in silico* sea solo de unas horas a algunos días.

Las pruebas *in silico* tienen una utilidad particular en la construcción de modelos QSAR, que ocurre en la optimización de candidatos. Usualmente, estos estudios se basan en procedimientos estadísticos y matemáticos, por lo general de carácter determinista. Es crucial la repercusión que tiene la obtención de buenos modelos en este proceso, por su utilidad en la identificación de potenciales bio-activos positivos durante el proceso de descubrimiento de nuevos candidatos. Los indicadores fundamentales de este proceso se detallan en la Tabla 2.

La obtención de buenos modelos QSAR implicaría una mejor selección de potenciales candidatos durante la primera fase. El desempeño de dichos modelos se refleja directamente en el indicador **exactitud**, que ha sido descrito en las ecuaciones (5) y (6), en la sección 1.2.1.

**Tabla 2**. Principales indicadores del proceso “Optimización de candidatos”.

| **Indicador** | **Expresión** | **Interpretación** |
|---|---|---|
| Exactitud del modelo QSAR ($acc$)<br>$acc \in [0,1]$<br>Objetivo: maximizar | $\frac{\textit{candidatos correctamente clasificados}}{\textit{total de candidatos}}$ | Proporción de aciertos del modelo QSAR. |
| Costo de síntesis molecular ($C$)<br>Objetivo: minimizar | $\sum_i^n \sum_j^k c_{ij}$ | Costo total de la síntesis de los $n$ compuestos en los $k$ experimentos. |
| Tiempo de validación ($T$)<br>Objetivo: minimizar | $\frac{\sum_i^n \sum_j^k t_{ij}}{n+k}$ | Tiempo medio de validación de los $n$ compuestos en los $k$ experimentos. |
| Especificidad ($Sp$)<br>$Sp \in [0,1]$<br>Objetivo: maximizar | $\frac{\textit{negativos correctamente identificados}}{\textit{total de negativos}}$ | Proporción en la cual la experimentación confirma la predicción negativa del modelo QSAR. |
| Sensibilidad ($Se$)<br>$Se \in [0,1]$<br>Objetivo: maximizar | $\frac{\textit{positivos correctamente identificados}}{\textit{total de positivos}}$ | Proporción en la cual la experimentación confirma la predicción positiva del modelo QSAR. |
| *Area Under ROC Curve* ($AURC$)<br>$AURC \in [0,1]$<br>Objetivo: maximizar | *Aproximación por método del polígono convexo* | Balance entre especificidad y sensibilidad |

Este indicador de gestión es una referencia comúnmente aceptada sobre la eficacia de la experimentación *in silico*, la cual ejerce su influencia en otros indicadores (véase la **Tabla 2**). Por ejemplo, en el indicador tiempo de validación experimental, pues se contaría con un modelo de predicción que reduce considerablemente los cálculos y estimaciones.

Además influye en la mejora del indicador costo de síntesis más adelante, puesto que se reduce la probabilidad de que un compuesto inviable se sintetice en vano. Como alternativa a la exactitud, suelen emplearse los indicadores especificidad, sensibilidad y AURC.

| | |
|---|---|
| **Proceso**: Optimización de candidatos | **Propietario**: Jefe de química analítica |
| **Misión**: Validar la actividad biológica de los candidatos farmacológicos mediante pruebas *in vitro* e *in silico*, construyendo modelos de relación estructura-actividad. | |
| **Alcance**<br>Empieza: Con el descubrimiento y cribado (selección) de nuevos candidatos farmacológicos.<br>Incluye: Síntesis molecular a pequeña escala, cálculo/medición de propiedades físicas, construcción de modelos de relación estructura-actividad (QSAR).<br>Termina: Con la validación de seguridad y efectividad, a través de pruebas toxicológicas sobre cultivos *in vitro*. | |
| **Entradas**: Posibles candidatos farmacológicos. Descriptores moleculares<br>**Proveedores**: Proceso de descubrimiento de nuevos candidatos. Ciencias afines | |
| **Salidas**: Candidatos farmacológicos validados en cuanto a seguridad y efectividad. Modelos de relación estructura-actividad.<br>**Clientes**: Ensayos clínicos y preclínicos. Síntesis molecular. | |
| **Variables de control:**<br>• Capacidad de síntesis molecular<br>• Plazo de validación para ensayo clínico<br>• Modelo QSAR | **Indicadores:**<br>• *acc*: Exactitud del modelo QSAR<br>• *T*: Tiempo de validación experimental<br>• *C*: Costo de síntesis molecular<br>• *Sp*: Especificidad<br>• *Se*: Sensibilidad<br>• *AURC*: *Area Under ROC Curve* |

**Figura 5**. Ficha del proceso “Obtención de nuevos candidatos”

Es precisamente en la obtención de buenos modelos QSAR donde se manipulan datos de alta dimensión. Esto también ocurre durante empleo de dichos modelos

en la predicción de la actividad biológica de un nuevo compuesto, lo cual ocurre constantemente en el proceso de descubrimiento.

### 1.2.2. Impacto de la minería de datos en la gestión de procesos de predicción.

Las pruebas *in silico* tradicionales en estudios QSAR se auxilian sobre todo de métodos de análisis estadístico. Estos métodos parametrizados con frecuencia carecen de complejidad suficiente para representar la naturaleza real de los datos con los que trabajan.

En esencia, los modelos de predicción QSAR son construidos a partir de la experiencia almacenada en bases de datos bioquímicas. Este marco es ideal para la aplicación de minería de datos. La diversidad y complejidad de estas técnicas, la mayoría de las cuales no presentan restricciones en cuanto a la representación de los datos, conduce a excelentes resultados en estudios QSAR, como ha sido demostrado durante la última década y continúa demostrándose [56].

Retomando el análisis de la sección anterior, puede concluirse que mejores técnicas de minería de datos influirán drásticamente en los indicadores de gestión del descubrimiento de fármacos:

- tiempo de validación experimental, reduciendo el riesgo de fallo en la experimentación *in vitro*, que (recuérdese) suele tomar hasta meses para ser completada. Téngase en cuenta además el impacto posterior en la fiabilidad de la experimentación *in vivo*, que puede tomar hasta años normalmente.
- costo de la síntesis molecular, reduciendo el uso de reactivos, equipamiento y procedimientos químicos especiales, pues se reduce el riesgo de síntesis de compuestos inviables.

Este ejemplo ilustra el potencial de la minería de datos para la mejora continua de los indicadores de gestión, en procesos que se basan en predicciones a partir de datos de alta dimensión. Generalizando, mejores técnicas de minería de datos contrinuirán con la mejora en los indicadores de gestión de estos

procesos. La Inteligencia Artificial estudia el problema de la minería de datos en una disciplina conocida como aprendizaje automatizado.

### 1.2.3. Aprendizaje automatizado

De acuerdo con varios autores especializados en el tema [34],[35] el aprendizaje automatizado (más conocido por su nombre en inglés: *machine learning*) persigue entrenar un modelo a partir de datos disponibles. El término "aprendizaje" debe interpretarse en este entorno como "adquisición de nuevo conocimiento". Desde un punto de vista externo, se establece que un modelo "aprende" cuando exhibe una mayor habilidad en la ejecución de la tarea para la cual fue entrenado. Este enfoque apunta claramente al carácter cuantificable del proceso de aprendizaje, partiendo de uno o varios indicadores de desempeño que permitan medir la "habilidad" en cuestión.

Las bases teóricas sobre las que ha sido fundamentado el aprendizaje automatizado parten de concepciones filosóficas, sicológicas y de la lógica matemática [34]. Esto pone de manifiesto su carácter interdisciplinario y la fuerte influencia de la inteligencia artificial en sus métodos. El paradigma que implementa es el del **aprendizaje inductivo**, cuya cualidad distintiva es la adquisición de nuevo conocimiento mediante la obtención de leyes generales a partir de la observación de casos particulares; esto es, la construcción de un modelo abstracto a partir de casos que describen el dominio.

El entrenamiento para el aprendizaje ocurre a partir de experiencia acumulada sobre el dominio en una base de datos. Por lo general, dichos datos deben suministrarse al algoritmo de aprendizaje estructurados en forma de "casos conocidos". Es decir, la experiencia empresarial, tecnológica o científica que recoge los datos operacionales debe ser homogeneizada y tabulada, de acuerdo con un conjunto específico de variables. Dichas variables son determinadas en los primeros pasos del proceso de KDD y son rasgos o atributos propios del dominio de la tarea de aprendizaje. De este modo, cada registro almacenado es un caso, una ocurrencia, un suceso independiente recogido en la experiencia previa [36]. Como resultado, se obtiene un conjunto de datos estructurados, que

reciben la denominación de “datos de aprendizaje” o “de entrenamiento”, pues estos serán suministrados a un algoritmo de aprendizaje para entrenar un modelo.

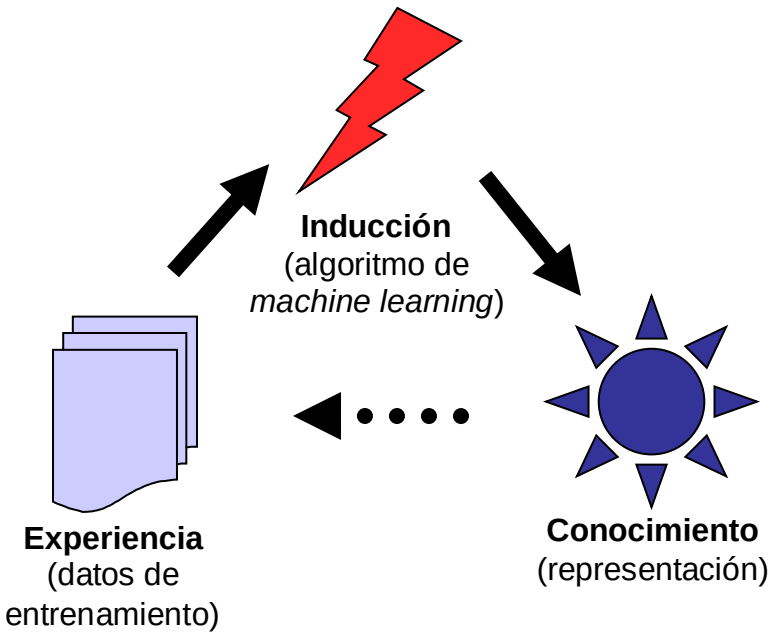


**Figura 2**. El paradigma de aprendizaje inductivo automatizado.

De acuerdo con dicha estructura, cada una de las variables del dominio puede asumir un rol determinado en el aprendizaje. En dependencia del problema o tarea de aprendizaje a la que se enfrenten, las técnicas de *machine learning* se clasifican en dos categorías generales [6]: aprendizaje supervisado o aprendizaje no supervisado. El tipo de tarea de aprendizaje es definido en el paso 5 del proceso de KDD y es crucial para poder obtener conocimiento útil e interpretable.

Las tareas en las cuales al menos una de las variables del dominio será adoptada como variable objetivo del aprendizaje, pueden ser tratadas con técnicas de **aprendizaje supervisado**. Los principales tipos de tareas de aprendizaje supervisado son:

- Clasificación. Pretende construir un modelo que sea capaz de clasificar un objeto dado, asignándole una de las categorías definidas previamente. Se realiza cuando la variable objetivo es nominal u ordinal, desde el punto

de vista estadístico. Por ejemplo, dados los signos y síntomas de un paciente, clasificarlo en “diabético” o “no diabético”.

- Estimación numérica. Se asocia comúnmente a la técnica estadística de regresión. Persigue la construcción de un modelo que sea capaz de estimar el valor numérico de la variable objetivo, ante la presencia de determinados valores de las variables descriptoras. Un ejemplo es la estimación de cuántos kilogramos de vegetales se podrá obtener de una parcela, conociendo las características del suelo y del cultivo, el régimen de precipitaciones, la fertilización, etc.

Por otro lado, cuando todas las variables son igualmente relevantes en el aprendizaje y ninguna es objetivo particular, la tarea en cuestión es el **aprendizaje no supervisado**. Su propósito es entrenar un modelo que identifique las interrelaciones existentes entre los elementos de datos. Los principales tipos de tareas del aprendizaje no supervisado son:

- Agrupamiento (*clustering*). Construye un modelo que identifica y describe la forma en que se relacionan los casos en la base de datos. Esta relación es representada mediante un conjunto de grupos (también conocidos como *clusters*) en los que están distribuidos y organizados los casos de la base de datos. Con frecuencia, el agrupamiento se utiliza como antesala de la clasificación, ya que al identificar los grupos existentes, acto seguido es factible asignarle una etiqueta a cada grupo, con lo cual es posible realizar un aprendizaje supervisado. Fundamentalmente por esta razón, el agrupamiento en ocasiones es llamado “clasificación no supervisada”.
- Asociación. Persigue identificar las interrelaciones que se manifiestan entre las variables del dominio, construyendo un modelo que describe dichas interrelaciones. Usualmente, el conocimiento obtenido es representado con un conjunto de reglas que expresan las asociaciones descubiertas.

En general, de un algoritmo de *machine learning* se espera obtener como resultado un conjunto de reglas, leyes o patrones, como forma de representación

del conocimiento descubierto en los datos. Sin embargo, el análisis de los tipos de aprendizaje y sus objetivos permite concluir que la predicción a partir de experiencia previa almacenada en bases de datos, es objeto de estudio del aprendizaje supervisado.

### Evaluación del aprendizaje supervisado

Para la evaluación del desempeño de un modelo de clasificación, luego de su entrenamiento este se somete a una fase de validación. En esta fase, al modelo se le suministran casos de validación cuya variable objetivo tiene valor conocido. El propósito es que el modelo emita una predicción para poder comparar la clase predicha con la clase real a la que pertenece el caso de validación en cuestión. El índice de desempeño más utilizado [36] es la proporción de errores que comete el modelo, al probarlo con datos de validación. Cuando el modelo predice correctamente la clase de un caso dado, se cuenta como un éxito, en caso contrario se cuenta como un fracaso. O sea:

$$error = \frac{e}{N}, \tag{1}$$

siendo $N$ la cantidad de casos de validación suministrados al modelo y $e$ la cantidad de casos clasificados incorrectamente.

Los resultados de la validación pueden ser representados de forma más general mediante una matriz de confusión, donde son tabuladas las predicciones realizadas a favor de cada clase, contra las clases reales del problema.

**Tabla 1**. Ejemplo de matriz de confusión en un problema de 3 clases.

| | | **Clase predicha** | | |
|---|---|---|---|---|
| | | **A** | **B** | **C** |
| **Clase real** | **A** | 23 | 11 | 0 |
| | **B** | 6 | 49 | 0 |
| | **C** | 0 | 4 | 10 |

En la Tabla 1 se ilustra un ejemplo de matriz de confusión $M$, como podría obtenerse a partir de la validación de un modelo de clasificación, en un problema de 3 clases o categorías. $M$ es una matriz cuadrada de orden $c$, siendo $c$ la cantidad de clases del problema. Cada valor $M_{ij}$ de la matriz representa la cantidad de casos pertenecientes a la clase $i$ que fueron clasificados como clase $j$. Puede afirmarse intuitivamente que un clasificador es perfecto cuando:

$$M_{ij} = 0, \forall i, j \in [1,...,c], \forall i \neq j \tag{2}$$

O, expresado de otra manera, cuando:

$$\sum_i M_{ii} = N, \forall i \in [1,...,c], \tag{3}$$

ya que, en definitiva:

$$\sum_i \sum_j M_{ij} = N, \forall i, j \in [1,...,c] \tag{4}$$

Lo cual conduce a otra expresión para evaluar el desempeño de un modelo, la cual no es más que el complemento del error: el **índice de exactitud**, calculado a partir de su matriz de confusión:

$$exactitud = \frac{\sum_i M_{ii}}{\sum_i \sum_j M_{ij}}, \forall i, j \in [1,...,c] \tag{5}$$

Sin embargo, los datos provenientes de los contextos científicos y técnicos en la actualidad, usualmente no presentan la misma cantidad de casos representando a todas las clases, lo cual es conocido en la literatura como “datos con clases desequilibradas”[27]. Por ello se emplea una generalización de (5), conocida como índice de exactitud balanceado[37]:

$$exactitud = \frac{1}{c} \sum_i \frac{M_{ii}}{\sum_j M_{ij}}, \forall i, j \in [1,...,c] \tag{6}$$

Los datos de validación del modelo son obtenidos de la misma forma que los datos con que el modelo fue entrenado, durante las etapas iniciales del proceso de KDD. Una estrategia puede ser utilizar los mismos datos de entrenamiento como datos de validación. No obstante, el desempeño medido de esta manera

no es un estimado fiable sobre la capacidad generalizadora del modelo [38]. Una alta capacidad generalizadora es una característica deseable en cualquier modelo de predicción, pues esta se refiere al desempeño que tendrá el modelo ante casos que no le fueron suministrados en el entrenamiento, casos nuevos y desconocidos.

Para estimar la capacidad generalizadora del modelo, la estrategia más lógica es entrenarlo con unos datos y validarlo con otros distintos. Por lo general, el conjunto de datos estructurados obtenidos durante el proceso de KDD es dividido en dos subconjuntos: datos de entrenamiento y datos de validación.

Un esquema de validación ampliamente utilizado[ 39], por su máximo aprovechamiento de los datos disponibles y su robustez estadística, es la validación cruzada (*cross-validation*). Este esquema divide los datos disponibles en ***k*** subconjuntos o pliegos (*folds*), donde los valores del dominio de la variable objetivo están equitativamente representados en cada pliego. Entonces se realizan ***k*** iteraciones de validación, empleando sucesivamente en cada iteración cada uno de los pliegos como datos de validación y el resto como datos de entrenamiento. De esta forma, se obtienen ***k*** valores de desempeño del modelo, calculados utilizando alguno de los índices detallados anteriormente (o algún otro). A estos valores se les determina la media aritmética, la cual es tomada como valor final del desempeño del modelo en el conjunto de datos analizados.

Un tipo particular de validación cruzada es aquella que se realiza en ***N*** pliegos, siendo ***N*** la cantidad de casos con que cuenta el conjunto de datos. Este tipo de esquema se conoce como validación cruzada “dejando uno fuera” (*leave-one-out cross-validation*, LOOCV), debido a que cada iteración se realiza utilizando sucesivamente un caso para validación y el resto de los casos para entrenamiento. Como también es un esquema de validación cruzada, al terminar todas las iteraciones se calcula igualmente la media aritmética para determinar el valor final del desempeño del algoritmo sobre el conjunto de datos.

Estos esquemas de validación también son aplicables al aprendizaje no supervisado, aunque utilizando otros índices de desempeño apropiados.

### Algoritmos de aprendizaje supervisado

A un algoritmo de aprendizaje le es suministrado un conjunto de datos de entrenamiento. En este caso, al algoritmo se le supervisa al especificarse la variable objetivo en ese conjunto de datos, pues de esta forma el algoritmo podrá disponer, durante el entrenamiento, del valor esperado correspondiente a cada caso del dominio.

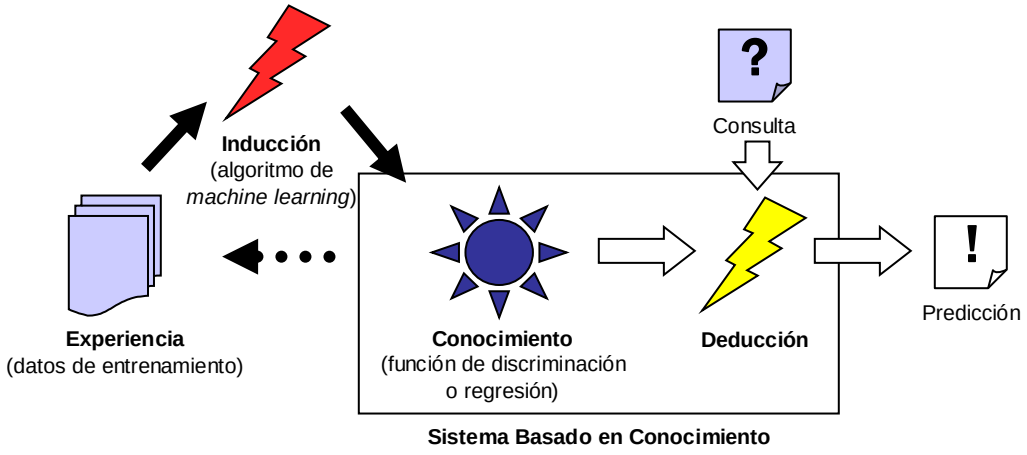


**Figura 3**. Aprendizaje supervisado para la predicción.

El epígrafe anterior destacaba dos tipos de tareas en el aprendizaje supervisado: la clasificación y la estimación numérica. La diferencia entre ambas radica en la naturaleza de la variable objetivo: depende de su dominio.

Cuando el dominio de la variable objetivo es un conjunto infinito de valores numéricos, el aprendizaje radica en aproximar una función que relacione los valores de entrada con la salida esperada [40]. En el campo de la estadística matemática, los análisis de regresión y de tendencia persiguen este mismo objetivo, por lo cual la estimación numérica es conocida también en *machine learning* como regresión.

Si el dominio de la variable objetivo es un conjunto finito de categorías predefinidas, entonces el aprendizaje consiste en encontrar o aproximar los criterios que permiten diferenciar las distintas categorías en el conjunto de datos [40]. La clasificación ha sido estudiada también como un tipo de aproximación de funciones. A diferencia de la regresión, en que se aproximan funciones de

tendencia, en la clasificación se aproximan una o varias funciones de discriminación, que permitan separar los elementos de datos de las diferentes clases o categorías.

Diversas técnicas [41],[42] han sido desarrolladas para la aproximación de funciones, tanto para la regresión como para la discriminación. La forma en que estas técnicas o algoritmos operan está fuertemente influida por la forma en que representan el conocimiento descubierto. Esto ocasiona que, en rasgos generales, se destaquen varias familias de algoritmos que operan de forma similar. Varias de estas familias han recibido mucha atención a lo largo del desarrollo del *machine learning* como disciplina [6, 34-36, 40-42], concentrando a su alrededor la mayor parte de los estudios en esta rama.

Los **árboles de decisión** seleccionan de forma iterativa y jerárquica las variables del conjunto de entrenamiento que son más relevantes para la tarea en cuestión, formando un árbol. Como resultado, se obtiene una estructura en cuyas hojas radican las predicciones. Los algoritmos emblemáticos de esta familia son CART (*Classification And Regression Tree*) [43], ID3 [44] y C4.5 [45]. Su gran popularidad radica fundamentalmente en su facilidad de interpretación por parte de una mente humana. El árbol de decisión es un modelo que permite realizar inferencias con gran facilidad, así como derivar intuitivamente un conjunto de reglas de decisión.

Las **reglas de decisión** son estructuras en la forma:

“si [precondiciones], entonces [conclusiones]”.

Durante la ejecución de un algoritmo de construcción de reglas se generan varias de ellas, a partir de las relaciones existentes entre las variables del conjunto de entrenamiento. Luego las reglas generadas son combinadas y depuradas para simplificar el conocimiento descubierto y, con frecuencia, son utilizadas para la predicción en forma de lista de decisión; es decir, en un orden específico: si las precondiciones de la primera regla no se cumplen se evalúa la segunda y así sucesivamente. En la línea de algoritmos que construyen reglas para la predicción se destaca RIPPER[46]. A pesar de su utilidad predictiva para

la clasificación y su facilidad de interpretación, los problemas de estimación numérica son mejor tratados al explotar las propiedades estadísticas, geométricas y topológicas de los datos.

Los **modelos estadísticos** se basan fundamentalmente en la teoría de las probabilidades. En un marco general son estimadas las funciones de distribución probabilística de los valores de cada variable del conjunto de datos, condicionadas por los valores de la variable objetivo. Luego, aplicando el teorema de probabilidad condicional de Bayes, es calculada la probabilidad a posteriori de que determinado objeto de clase desconocida pertenezca a una u otra clase. Por lo general, para la estimación se asume que las variables del problema son igualmente importantes y mutuamente independientes, lo cual ocurre rara vez en la práctica. Sin embargo, esta suposición “ingenua” puede sorprender por su eficacia, como demuestran estudios tempranos con el uso del algoritmo Naïve Bayes [47]. No obstante su simplicidad y efectividad para problemas de clasificación, no ofrece las mismas ventajas para la estimación numérica, en la cual han tenido más éxito los modelos geométricos, como la regresión lineal.

Los **modelos lineales** y sus generalizaciones no lineales fueron estudiados a fondo en sus inicios también por la estadística, en el método de los mínimos cuadrados. En la práctica, estos algoritmos aproximan de forma iterativa una función lineal de tendencia o de discriminación, rectificando en cada iteración sobre la base del margen de error obtenido. La simplicidad y poder representativo de una función lineal hacen muy atractivos estos métodos. No obstante, han sido introducidas extensiones cuadráticas, polinómicas y de otra índole también con éxito. Desde el punto de vista geométrico, los algoritmos exitosos de árboles, como el C4.5, se basan en funciones lineales anidadas. En la actualidad, uno de los algoritmos de mejores resultados en la minería de datos se basa en modelos lineales, mejorado utilizando funciones *kernel*: la máquina de vectores de soporte o SVM (del inglés *Support Vector Machine*)[48].

Las **redes neuronales** implementan una forma diferente de generalizar los modelos lineales y las funciones *kernel*. Inspiradas en el funcionamiento del sistema nervioso, son estructuras fuertemente interconectadas, donde cada elemento aproxima funciones y la combinación de varios elementos permite aproximar funciones mucho más complejas. Esta estructura le permite ser un modelo de predicción muy efectivo, pero de difícil interpretación. Este tipo de estructura ha recibido mucha atención desde la introducción de la neurona tipo "Perceptrón" de Rosenblatt en la década de 1950. Ha sido una de las técnicas más estudiadas, sobre todo el Perceptrón Multi-Capa o MLP (del inglés *Multi-Layer Perceptron*) [49], aunque no aparece recogido entre los 10 mejores algoritmos para minería de datos, según expertos mundiales [50], lo cual se debe fundamentalmente al elevado costo computacional de su entrenamiento y su baja utilidad para explicar el conocimiento descubierto. No obstante, con el tiempo ha surgido una vertiente de sistemas basados en conocimiento, conocidos como sistemas neuro-borrosos, que combinan la efectividad de las redes neuronales con el poder expresivo de la lógica borrosa.

El **aprendizaje basado en casos** es un paradigma diferente: en lugar de construir un modelo como abstracción del conocimiento subyacente en los datos, utiliza directamente los datos de entrenamiento para realizar predicciones, sin modelo intermedio. Para hacerlo, esta familia de algoritmos se basa en una regla simple: objetos similares deben tener valores similares para la variable objetivo. Para predecir dicho valor en presencia de un caso desconocido se realiza una búsqueda de aquellos casos de entrenamiento que más se le asemejan. Esta operación es conocida como "búsqueda de vecinos más cercanos" o NN-search (del inglés *nearest neighbor search*). Por tanto, el proceso de entrenamiento de este tipo de algoritmos se concentra, fundamentalmente, en determinar la cantidad ***k*** de vecinos que resulta más adecuado considerar para la predicción. Este es el formato general de un emblemático algoritmo basado en casos: k-NN[ 51]. La versatilidad de este método lo ha hecho objeto de muchos estudios, mejoras y generalizaciones. La cuestión de cuantificar la "similitud" entre los objetos es de máxima importancia

para estos algoritmos, pues se basan en la búsqueda de vecinos similares. Tradicionalmente se ha utilizado como criterio de similitud la distancia entre los objetos, representados en un espacio vectorial, pero también se han explorado muchas otras maneras, exploración alentada por la diversificación de los tipos de datos que aparece con frecuencia en muchos dominios.

Como ya se ha destacado en epígrafes anteriores, el proceso de descubrimiento de conocimiento en datos implica realizar búsqueda e inferencia, elementos que, de una manera u otra, están presentes en las familias de algoritmos aquí descritas. La evolución lógica de estos algoritmos, al menos en el plano experimental, ha estado condicionada desde sus inicios por los datos y problemas disponibles. A lo largo de décadas de estudio, y sobre todo tras la revolución de las tecnologías de la información y las comunicaciones hacia finales del siglo XX, los datos disponibles han aumentado en complejidad y volumen.

La cantidad de variables que describen a un conjunto de datos influye poderosamente en los procesos de búsqueda en inferencia que tienen lugar durante la minería de datos. A raíz del incremento en la complejidad y volumen de los datos disponibles, se ha estudiado la sensibilidad de los algoritmos históricamente más eficientes. Como resultado, se han obtenido comportamientos no esperados, encontrándose los algoritmos bajo los efectos de lo que se ha llamado la “maldición” de la alta dimensión [52].

## 1.3. Estrategias para enfrentar los procesos de predicción a partir de datos de alta dimensión

El aprendizaje con datos de alta dimensión ha sido enfrentado en ocasiones indirectamente. No obstante, en algunos dominios ha adquirido importancia y, por tanto, se le ha prestado mayor atención. Para tareas de predicción algunas propuestas han mostrado resultados prometedores, que son aplicables a la mejora de los indicadores de gestión en procesos basados en la predicción.

En general, en la literatura especializada se observan tres amplias estrategias para lidiar con los datos de alta dimensión: la primera y más evidente es reducir la cantidad de dimensiones de los datos, prescindiendo de las variables menos relevantes para la tarea en cuestión; la segunda es adaptar las técnicas existentes para que se desempeñen mejor en ambientes de alta dimensión; la tercera, combinar varios modelos entrenados para la misma tarea en lugar de emplear solo uno en la predicción.

### 1.3.1. Reducción de la dimensionalidad

En espacios de muchas dimensiones, formados por un gran número (cientos, miles) de variables, la existencia de dimensiones prácticamente coincidentes es altamente probable. Dichas dimensiones corresponden a variables altamente correlacionadas, las cuales tributan prácticamente la misma información. Además, a causa del espacio vacío, existirán dimensiones prácticamente vacías, no ocupadas por ninguna, o casi ninguna proyección de los datos. Estas dimensiones tienen por tanto poca participación en la distribución real de los puntos de datos, correspondiendo a variables con bajo poder discriminatorio.

Por otro lado, el fenómeno de Hughes sugiere que un modelo de aprendizaje supervisado debe ser entrenado utilizando la cantidad de variables para la cual su exactitud alcanza el máximo.

En estas condiciones se impone la búsqueda de una cantidad menor de variables, con alta relevancia, buen poder discriminatorio y baja redundancia. Para lograrlo, las técnicas desarrolladas se agrupan en dos grandes familias: selección de rasgos y construcción de rasgos. La selección persigue encontrar el subconjunto de las variables más útiles, entre todas las originales. La construcción crea nuevas variables a partir de combinaciones de las originales, con el fin de obtener un conjunto menos numeroso de variables útiles.

Ya sea mediante selección o mediante extracción de rasgos, la reducción de la dimensión del problema es una estrategia pertinente en la situación descrita. Un análisis de la competencia desarrollada en el evento NIPS 2003 [26] sobre este

tema, permite distinguir que las técnicas de selección son más utilizadas que las de extracción, debido a razones de costo computacional.

La selección de rasgos es usualmente tratada como un problema de optimización multi-objetivo, donde se desea obtener el menor conjunto posible de variables, que maximice la cantidad de información útil y la exactitud del modelo entrenado con las variables seleccionadas. Inicialmente enfocada con estrategias de búsqueda exhaustiva, la mencionada "maldición" de la alta dimensión la ha transformado en una búsqueda intratable. Obsérvese que, en un conjunto de $\boldsymbol{m}$ variables, habría que evaluar $2^m$ posibles subconjuntos para determinar el mejor de ellos. Es por ello que la búsqueda parcial, heurística, se ha encargado de este problema en los últimos lustros [37].

En cuanto a la evaluación de la calidad de cada subconjunto, se han explotado fundamentalmente dos alternativas [62]: la externa, entrenar un modelo y tomar su desempeño como criterio de calidad del subconjunto utilizado; o la interna, estimar la utilidad del subconjunto para la tarea en cuestión mediante algún análisis de los datos en sí mismos. La alternativa externa es conocida en la literatura como *wrapper* (en español: envoltura, el algoritmo de selección envuelve el entrenamiento y evaluación de un modelo) y la interna como *filter* (en español: filtro, el algoritmo de selección estima la utilidad de cada subconjunto a considerar y luego filtra los más relevantes).

Por su parte, las conclusiones del mencionado NIPS 2003[63] afirman que las técnicas de selección más exitosas para datos de alta dimensión resultan ser las basadas en filtros uni-variable. Un filtro uni-variable no evalúa subconjuntos de variables, sino cada una de ellas por separado. Luego las ordena en un escalafón (*ranking*) y filtra utilizando un umbral de aceptación.

Para determinar la relevancia (utilidad) de cada variable son empleados diversos criterios. Algunos trabajos [3] dan crédito a criterios estadísticos, como la prueba Gamma y la correlación lineal. Otros trabajos [64], [65] coinciden en emplear exitosamente filtros para la selección. Pero estos trabajos se basan en criterios de relevancia tomados de la teoría de la información fundada por Claude

Shannon, como la incertidumbre simétrica y la información mutua condicional, ambas basadas en la entropía de las variables. Su principal argumento es que estos criterios son capaces de revelar dependencias no lineales entre los rasgos de decisión y los descriptores. También pueden encontrarse otros estudios [66-71] basados en diferentes algoritmos de filtrado y criterios de relevancia. Puede apreciarse que las métricas basadas en entropía muestran sistemáticamente un buen desempeño entre las diferentes propuestas.

Algunas técnicas de mayor complejidad [16, 72-75] persiguen encontrar los subconjuntos de variables que representan mejor cada una de las categorías o clases en las que están distribuidos los datos. De esta forma, cada clase queda relacionada a uno o varios de estos subconjuntos. Este último enfoque, sin embargo, no es adecuado para el contexto bajo estudio en esta investigación, donde se cuenta con relativamente pocos casos, por tratarse de espacios con miles de dimensiones. Además, con frecuencia las clases están fuertemente desbalanceadas. Por otra parte, no serían aplicables a problemas de estimación numérica, en los cuales no existen clases de objetos. No obstante, realizándole adaptaciones o en combinación con otras estrategias, como la utilización de multi-clasificadores (abordados más adelante en este epígrafe y en el siguiente), podrían ser adecuados para este contexto.

En cuanto a la construcción de rasgos, aunque no es la práctica más generalizada, existen determinados dominios en los que se ha aplicado exitosamente. Ejemplos de estos dominios son el reconocimiento facial [76], tratamiento de video [24], clasificación de sonidos [77] y el análisis de datos espectrales [78], [17]. Como puede verse, en la mayoría de los ejemplos se trata de algún tipo de señal. Esto permite que el proceso de construcción de rasgos se beneficie directamente de las técnicas, métricas y conocimientos propios del tratamiento digital de señales. Además, cabe destacar que la cantidad de rasgos originales utilizados para la extracción, rara vez excedió los pocos cientos, cifra pequeña comparado con los miles de variables de que constan los ejemplos mencionados.

### 1.3.2. Adaptación de los métodos, algoritmos y métricas

Utilizar las causas conocidas de la “maldición” de la alta dimensión para mejorar los algoritmos existentes, de manera que se comporten más robustamente, es otra estrategia pertinente.

Las métricas de distancia son ampliamente utilizadas como criterio de disimilitud. Por ello, el fenómeno de concentración de las distancias ha recibido mucha atención. Se ha estudiado, tanto teórica como empíricamente, la influencia del fenómeno sobre las técnicas basadas en funciones *kernel* [22], como las máquinas de vectores de apoyo (SVM por sus siglas en inglés). También sobre técnicas basadas en búsqueda de vecinos más cercanos [3], como el clasificador *k*-NN y otras como las basadas en centroides [79].

En general, se recomienda el uso de métricas de distancia alternativas a la euclidiana para rasgos numéricos. En especial se recomiendan, como robustas ante este fenómeno, otras métricas de la familia Minkowski ($\ell_p$), para valores de $p \leq 1$. No obstante, no todos los estudios mencionados ofrecen resultados experimentales con datos descritos por miles de rasgos, sino solamente en el orden de los cientos.

Por otro lado, en la búsqueda de vecinos más cercanos se ha estudiado [25], [80] la aparición de objetos que son vecinos de muchos otros en el conjunto, como consecuencia de la concentración de las distancias. A estos objetos se les ha denominado "puntos centrales" y se ha demostrado que pueden afectar el rendimiento de los algoritmos de aprendizaje supervisado basados en distancias, como *k*-NN y SVM, así como algoritmos no supervisados como el prestigioso *k-means* de agrupamiento. Por ello, los autores del referenciado trabajo recomiendan eliminar los puntos centrales perjudiciales o ponderar la participación de los *k* vecinos más cercanos, atendiendo a cuán perjudicial resulta su "centralidad". En el contexto que sirve de objeto de estudio a la presente investigación, la variante de eliminar los objetos perjudiciales no será recomendable en muchos casos, teniendo en cuenta la escasez de ejemplares en varios dominios anteriormente mencionados. Eliminar objetos de datos

agravará el problema del espacio vacío, por ello la alternativa de ponderar la participación de los *k* vecinos luce más adecuada.

El fenómeno del espacio vacío también ha sido objeto de estudio, sobre todo por su incidencia en las estimaciones de densidad probabilística. En dicho sentido, también se han realizado propuestas de nuevas técnicas de estimación de densidad, como la utilización de un *kernel* Gaussiano generalizado[3] y los árboles de difusión basados en procesos Dirichlet [81], [82], esta última con algún éxito en competencias organizadas [83].

### 1.3.3. Combinar varios modelos en un ensamblado

En la experiencia humana, un comité de expertos en varias aristas de un mismo tema suele tomar decisiones más fiables y seguras que cualquiera de los expertos por sí solo. En algunos casos, una parte de los expertos tendrá más certeza sobre la decisión a tomar, porque el caso en cuestión pertenece a su dominio de experticia. En otros casos, serán otros expertos en el comité los que puedan decidir con más seguridad. El consenso entre todos, a la larga, garantizará un proceso balanceado y más fiable.

Esta motivación ha sido extrapolada al aprendizaje automatizado, teniendo en cuenta que un modelo entrenado mediante un algoritmo de aprendizaje, cumple una función análoga a un experto en el dominio en que fue entrenado. Extendiendo la analogía, un conjunto de modelos entrenados para la misma tarea se comportará como un comité de expertos en ese dominio. Solo resta combinar de alguna manera las decisiones de los modelos del conjunto y podrá obtenerse un consenso.

Esta idea ha sido explotada en *machine learning*, inicialmente para complementar el funcionamiento de varias redes neuronales; más adelante, para aprovechar las ventajas de diferentes algoritmos a la vez. Esto ha dado lugar al estudio de los multi-clasificadores, en el caso de tareas de clasificación o, de forma general, el estudio de los “ensamblados” de modelos.

Uno de sus más dedicados investigadores, Thomas G. Dietterich, define [84] un ensamblado como: un conjunto de modelos cuyas decisiones son combinadas de algún modo para responder a nuevas consultas. Esta combinación se ha realizado tradicionalmente, para las tareas de clasificación, mediante un voto mayoritario. En dicha técnica, cada modelo del ensamblado vota por una de las clases del problema, luego de recoger todos los votos, el nuevo caso es etiquetado con la clase que más votos obtuvo. Otra técnica es “promediar” las predicciones de todos los modelos para calcular la decisión final del ensamblado.

Ha sido demostrado [84-86] que, en una tarea de aprendizaje supervisado, la combinación de un ensamblado de modelos puede superar significativamente el rendimiento de cualquiera de los modelos individuales del conjunto. Para ello debe garantizarse que cada modelo del ensamblado tenga un buen desempeño individual y, además, que los modelos no fallen de forma correlacionada entre unos y otros. Es decir, debe contarse con exactitud y diversidad en el ensamblado. En el aprendizaje supervisado se considera [40] que un modelo tiene buen desempeño cuando su margen de error es menor que el error en el que se incurriría al determinar al azar el valor de la variable objetivo. Por su parte, la diversidad en un ensamblado equivale, en la experiencia humana, a que los miembros del comité sean expertos en distintas facetas del dominio del problema. Un ensamblado en el que todos los modelos coinciden siempre en sus predicciones, carece de sentido. Este punto crucial ha motivado que la generación de diversidad a la hora de construir un ensamblado haya sido objeto de muchas investigaciones.

En una de sus obras [86], Ludmila Kuncheva sintetiza los niveles que componen la construcción de un ensamblado, ilustrando de forma muy intuitiva los diferentes puntos donde el investigador puede incidir para generar diversidad.

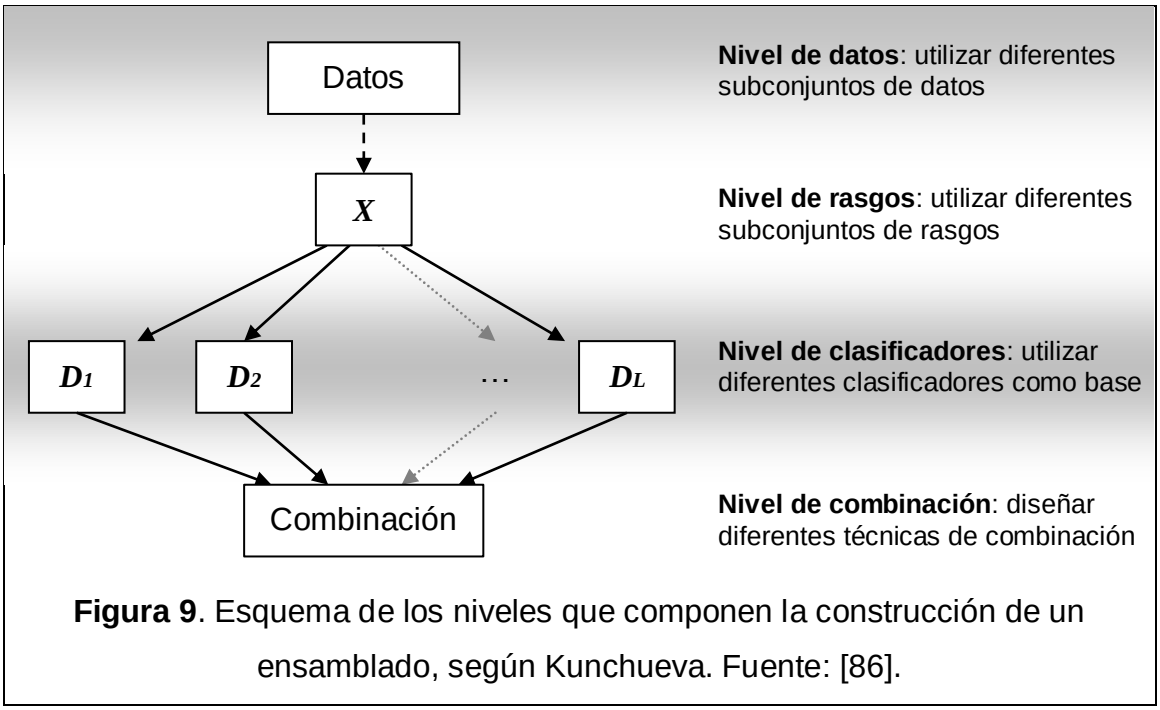


**Figura 9**. Esquema de los niveles que componen la construcción de un ensamblado, según Kunchueva. Fuente: [86].

En estos 4 niveles: datos, rasgos, modelos y combinación, se puede enmarcar gran parte de los esfuerzos investigativos que han tenido lugar en las últimas dos décadas alrededor del tema de los ensamblados para *machine learning*.

En el nivel de datos se trabaja con los casos de entrenamiento obtenidos de una base de datos original, durante las primeras etapas del proceso de KDD. Para generar diversidad, algunas técnicas entrenan cada modelo del ensamblado con un subconjunto distinto de casos de entrenamiento. Para ello se efectúan sub-muestreos sucesivos a la muestra original de los datos.

Apoyándose en el sub-muestreo, las técnicas de "*bagging*" [87] y "*boosting*" [88] han sido tradicionalmente las más exitosas para la construcción de ensamblados. Sin embargo, al enfrentarse con datos de alta dimensión el sub-muestreo, lejos de beneficiar la construcción del ensamblado, la perjudica. Ocurre que mediante sub-muestreo de los datos (que ignora la cantidad de variables), no solo permanecen los efectos de la concentración de las distancias y del fenómeno de Hughes, sino que además se acentúa el efecto del espacio vacío[*], al entrenarse cada modelo con menos casos que los pocos con los que

[*] Para más detalles sobre los mencionados fenómenos, remítase al epígrafe 1.3.

ya se cuenta. Como consecuencia, para datos de alta dimensión por lo general no se trabaja en el nivel de datos.

En dichas circunstancias son más apropiadas las técnicas que descomponen el conjunto de variables en subconjuntos, trabajando en el nivel de rasgos. Cada uno de los subconjuntos de variables obtenidos servirá para entrenar uno de los modelos individuales del ensamblado. Como consecuencia, los datos de entrenamiento originales serán proyectados en sub-espacios de menor cantidad de dimensiones que el original, siendo cada proyección diferente a las demás. De esta forma se logra simultáneamente descomponer el problema en sub-problemas, disminuir la dimensión de cada sub-problema y generar diversidad en los datos de entrenamiento para cada modelo individual.

Entre las técnicas de descomposición más utilizadas se encuentran las aleatorias, las cuales fueron las primeras que recibieron especial atención para el entrenamiento de ensamblados con datos de alta dimensión. La descomposición aleatoria resulta especialmente atractiva, pues es necesario recordar que existen $2^m$ posibles subconjuntos de variables, lo cual hace que la búsqueda exhaustiva e incluso algunas heurísticas sean impracticables. Un trabajo que marcó esta área del aprendizaje automatizado fue *Random Subspace Method* (Método de los sub-espacios aleatorios) [60]. Este método realiza una descomposición aleatoria con solapamiento, es decir, no se restringe la asignación de una variable a un solo subconjunto. Su simplicidad y conveniencia han hecho de este método un punto de referencia para otras técnicas.

Se ha experimentado con la descomposición aleatoria sin solapamiento (particiones aleatorias) [28], con buenos resultados. También Leo Breiman, uno de los principales investigadores en el área de árboles de predicción, introdujo aleatoriedad en uno de sus algoritmos, para generar no uno, sino un conjunto de árboles, en lo que llamó “bosque” aleatorio (*Random Forest*)[59], técnica que también se ha convertido en un *benchmark* para las investigaciones en este campo.

Las técnicas aleatorias disfrutan de la ventaja de un menor costo computacional. No obstante, otras técnicas de descomposición más complejas [29, 56, 89-91] están siendo estudiadas en años recientes para enfrentar el aprendizaje con datos de alta dimensión. Estas técnicas explotan mediante heurísticas una desventaja de la descomposición aleatoria: variables que en conjunto pueden resultar muy útiles para la predicción no son asignadas al mismo sub-espacio, pues la distribución se realiza al azar. Las heurísticas van dirigidas fundamentalmente a extender los principios y fundamentos de la selección de rasgos, convirtiéndolos en algoritmos de descomposición del conjunto de variables. Los trabajos de Lior Rokach y su equipo en esta dirección [29, 89, 90] han arrojado resultados alentadores, aunque sus técnicas funcionan con un tipo específico de árbol.

En el nivel de modelos se entrenan los miembros del ensamblado con los datos y variables escogidos en los niveles anteriores. Una manera adicional de generar diversidad es utilizar diferentes algoritmos de aprendizaje para entrenar cada modelo. No obstante, para datos de alta dimensión, sobre todo desde la publicación del *Random Subspace Method* y de los *Random Forests*, la simplicidad impulsa a utilizar el mismo algoritmo de aprendizaje para todos los modelos, generando la diversidad en el nivel de rasgos.

Independientemente de la forma en que se genere el ensamblado de modelos, existen varias estrategias para combinar los resultados de los modelos individuales, lo cual ocurre en el nivel de combinación. De forma general, las técnicas de combinación se pueden enmarcar en dos categorías: fusión y selección. La fusión utiliza los resultados de todos los modelos del ensamblado para generar una salida, mientras la selección escoge un modelo individual para que tome la decisión final. Las técnicas del voto mayoritario y de la media aritmética, mencionadas en párrafos anteriores, son ejemplos típicos de fusión y han sido las técnicas de referencia en los estudios sobre este tema. Se han propuesto extensiones a estas técnicas, como el voto ponderado y la media ponderada, y se ha trabajado en otras, como la combinación bayesiana. La obra

de Kuncheva [86] ofrece un estudio a profundidad sobre las principales técnicas de combinación.

La selección ha tomado diversas formas, siendo la más estudiada inicialmente la selección dinámica de un único modelo del ensamblado para realizar la predicción en nombre de todos (*Dinamic Classifier Selection*) [92]. Sin embargo, en los últimos años ha adquirido auge la selección de un subconjunto de modelos, para entre ellos realizar la predicción a nombre de todos (*Dinamic Classifier Ensemble Selection*) [30, 93]. Otra de las formas de selección que han mostrado buenos resultados es la selección “en cascada” (según Kuncheva) o “por delegación” (según sus autores) [94]. Esta técnica encarga al primer modelo del ensamblado la responsabilidad de arrojar un resultado, si se estima que el primer modelo no podrá realizar una predicción con suficiente certeza, su tarea es “delegada” al segundo modelo y así en lo sucesivo.

Se han propuesto diferentes formas sofisticadas de fusión ponderada [95], [96] y de selección mediante criterios de exactitud local de los modelos [97, 98]. Incluso existe una propuesta de combinación que escoge dinámicamente cuál es la estrategia más conveniente para cada caso: si fusión o selección [99].

Al estudiar el estado del arte, luego de recorrer de forma sintética los cuatro niveles de la construcción de un ensamblado según Kuncheva, se destaca la insuficiente publicación de resultados en el nivel de modelos. En especial, pocos trabajos abordan el por qué de la selección de uno u otro algoritmo de aprendizaje para entrenar los modelos del ensamblado. Una brecha interesante en este campo resulta el estudio del comportamiento de los diferentes algoritmos de aprendizaje al ser utilizados, como base, en la construcción de ensamblados generados por descomposición del conjunto de variables, específicamente cuando se trata de datos de alta dimensión. Cada familia de algoritmos de clasificación, por ejemplo, es sensible a distintas formas de complejidad de los datos [100]: la complejidad de las fronteras, la ambigüedad de las clases y la densidad de los datos.

La gestión de un proceso de KDD para crear modelos de predicción se beneficiaría, en gran medida, con alguna evidencia que apoye la elección de algunos algoritmos sobre otros, especialmente cuando los datos disponibles son de alta dimensión. Dicha evidencia podría ser aportada por estudios experimentales que profundicen en el impacto de los distintos algoritmos de aprendizaje sobre los indicadores de gestión, especialmente la exactitud, bajo la prometedora (y aún no completamente explorada) estrategia de los ensamblados.

## 1.4. Conclusiones parciales del capítulo

- Los fenómenos asociados a los datos de la alta dimensión ofrecen una explicación satisfactoria a las irregularidades en el desempeño de los algoritmos de aprendizaje supervisado, afectando por tanto los indicadores de gestión que dependen de la exactitud de la predicción.
- El empleo de técnicas de aprendizaje automatizado podría influir significativamente en la mejora del indicador de gestión **exactitud**, influyendo así de forma global en la gestión de procesos que emplean minería en datos de alta dimensión, debido a la estrecha interrelación entre este indicador y otros de importancia.
- El éxito parcial de las estrategias de selección de rasgos y adaptaciones mencionadas indica que los datos de alta dimensión, con frecuencia, contienen variables redundantes o poco informativas. Esta es una característica distintiva de contextos que aún se encuentran bajo estudio, de los que se cuenta con un escaso conocimiento explícito. Ello permite inferir que estas estrategias presentan limitaciones en cuanto a la complejidad de los datos que manejan.
- Las potencialidades estadísticas, computacionales y de representación asociadas con el empleo de ensamblados de modelos de predicción, para la mejora de los procesos de minería de datos de alta dimensión, se encuentran en los inicios de su explotación. Esto lo demuestra la escasez de propuestas especialmente diseñadas para lidiar con tales tipos de datos, imponiéndose un estudio más profundo de esta estrategia.

# CAPÍTULO 2
# Estudio experimental

En el capítulo anterior fue introducido el proceso de KDD, así como las principales herramientas y algoritmos para la minería de datos con el propósito de crear modelos de predicción. También fueron abordadas las particularidades de la gestión de procesos que incorporan minería de datos de alta dimensión, los principales indicadores de gestión que comprende, los principales fenómenos que la afectan y las estrategias generales que se han seguido para enfrentar dichos fenómenos.

El presente capítulo aborda en detalle el diseño del estudio experimental realizado. Su finalidad es aportar evidencia que permita determinar cuánto influye la estrategia de ensamblado de modelos en la mejora del indicador de gestión relacionado con la exactitud de la predicción, durante el aprendizaje supervisado con datos de alta dimensión. De este modo, el análisis de los resultados permitirá arribar a algunas conclusiones sobre la potencialidad del uso de ensamblados como estrategia para la gestión de este tipo de tarea.

El diseño del estudio experimental comprende la elección del indicador de desempeño utilizado (exactitud de la predicción), el esquema de experimentación, la elección de los algoritmos de control o referencia, la descripción de los datos utilizados, así como el procesamiento estadístico mediante los *tests* de Friedman, Nemenyi y Wilcoxon, imprescindibles para la interpretación de los resultados.

También se incluyen en el presente capítulo: el diseño del marco empleado para la experimentación con ensamblados de modelos y su comparación con los algoritmos de referencia; una relación detallada de los resultados experimentales y el correspondiente análisis crítico.

## 2.1. Diseño del estudio experimental

Como complemento al estudio del estado del arte, esta investigación ha proyectado un estudio experimental para explorar, con mayor profundidad, las potencialidades de la estrategia de los ensamblados de modelos, específicamente los multi-clasificadores. Tomando como base las técnicas de descomposición y combinación más utilizadas y menos complejas, el estudio experimental se propone **determinar cómo se comporta la exactitud de los mejores algoritmos de aprendizaje supervisado al ser enmarcados en un multi-clasificador entrenado con datos de alta dimensión, como estrategia para construir un modelo de predicción**.

El marco para la experimentación con multi-clasificadores seguirá las siguientes pautas:

- Todos los casos del nivel de datos se le suministrarán a los modelos individuales del ensamblado.
- En el nivel de rasgos se empleará una descomposición aleatoria con solapamiento, según establece el *Random Subspace Method* [60].
- En el nivel de modelos se entrenarán 10 modelos clasificadores, cantidad escogida teniendo en cuenta estudios anteriores [101] y por razones de eficiencia. Cada clasificador será entrenado con un subconjunto de variables diferente, generado en el nivel de rasgos. Los 10 modelos serán entrenados utilizando el mismo algoritmo de aprendizaje.
- En la combinación de las decisiones individuales de los modelos se empleará el voto mayoritario tal y como lo describen Dietterich [84] y Kuncheva [86].

Este marco permitirá comparar el desempeño de distintos algoritmos de aprendizaje empotrados en el enfoque de un multi-clasificador.

Como grupo de control o de referencia en la experimentación, se emplearán los mismos algoritmos de clasificación, pero generando solo un modelo con cada uno, sin utilizar el enfoque del ensamblado.

Para la ejecución de los experimentos se ha utilizado Weka [36], herramienta *open source* desarrollada por investigadores de la Universidad de Waikato, Nueva Zelanda y enriquecida por una vasta comunidad de investigadores a nivel mundial. Implementada en lenguaje de programación Java, ha sido concebida para estudios de minería de datos, aprendizaje automatizado, pre-procesamiento de datos y otros temas relacionados. La gran variedad de algoritmos que incluye, así como la amigable interfaz de usuario, que incluye un módulo de experimentación, la han convertido en una herramienta muy adecuada para este tipo de estudios experimentales.

Los estudios experimentales en *machine learning* publicados en los últimos años siguen las pautas trazadas por Dietterich en uno de sus trabajos [39]. En este se propone un esquema de experimentación, basado en validación cruzada que garantiza una mayor robustez estadística. Su propuesta consiste en realizar, con cada conjunto de datos, 5 repeticiones de una validación cruzada[*] en 2 pliegos generados de forma distinta en cada repetición y, por último, utilizar la media aritmética del desempeño en las 10 validaciones como valor final del desempeño del método con el conjunto de datos en cuestión. Este esquema es referenciado en la literatura como “5x2cv” y será el empleado en el presente estudio experimental.

Para el análisis de los resultados, se emplearán las recomendaciones [102] ofrecidas a la comunidad de investigadores en *machine learning* por Janez Demšar, en un trabajo que se ha convertido en uno de los más referenciados en éste campo de investigación. Demšar recomienda utilizar el esquema 5x2cv, anteriormente descrito, para estudios experimentales donde se comparan varias técnicas o métodos con varios conjuntos de datos. Los índices de desempeño arrojados por cada técnica con cada conjunto de datos durante la validación serán utilizados como variables para la realización de pruebas de hipótesis, bajo la hipótesis nula de que no existen diferencias significativas en el desempeño de las técnicas comparadas.

[*] Remítase al Epígrafe 1.2.3. para más detalles sobre la validación cruzada.

Como primera prueba estadística, Demšar recomienda un *test* de K muestras relacionadas, como el test de Friedman. En caso de encontrarse diferencias significativas, indica la aplicación de pruebas *post-hoc*, como el test de Nemenyi o el de Bonferroni-Dunn, para encontrar subgrupos de técnicas cuyo desempeño es significativamente diferente. En este punto, aconseja fuertemente las pruebas post-hoc en lugar de pruebas apareadas, o sea, de 2 muestras relacionadas. Solo después de detectados los grupos de comportamiento significativamente diferente, se señala que podría ser útil la exploración más profunda con el empleo de pruebas apareadas, como el test de Wilcoxon. En general Demšar sugiere el empleo de técnicas no paramétricas, evitando la verificación de los supuestos necesarios para aplicar pruebas paramétricas. Es común que los datos provenientes de entornos de gestión reales no cumplan dichos supuestos, por lo cual las recomendaciones de Demšar son tenidas en cuenta en este tipo de estudios.

Para efectuar las pruebas estadísticas de Friedman y Wilcoxon se ha empleado el asistente estadístico SPSS en su versión 15.0 para el sistema operativo Microsoft Windows. Para la prueba *post-hoc* de Nemenyi se han utilizado los valores del ranking generado por la prueba de Friedman en el asistente SPSS.

Teniendo en cuenta el marco definido al inicio de este epígrafe, así como las recomendaciones realizadas por investigadores líderes en este campo, se ha realizado un conjunto de experimentos con 14 conjuntos de datos de alta dimensión extraídos de diferentes dominios, utilizados para entrenar 14 técnicas diferentes de clasificación. Tanto los conjuntos de datos como los clasificadores serán presentados y detallados en epígrafes subsiguientes. Finalmente se resumirán los resultados experimentales y su análisis.

## 2.2. Algoritmos utilizados

En su obra "*The Top Ten Algorithms in Data Mining*"[50], un colectivo de destacados investigadores en aprendizaje automatizado y KDD realiza un levantamiento y descripción detallada de los algoritmos más exitosos en varias

formas de minería de datos. Entre los 10 señalados, 6 de ellos son algoritmos de aprendizaje supervisado, estos son:

- C4.5
- Support Vector Machines (SVM)
- AdaBoost + C4.5
- k-NN
- Naïves Bayes (NB)
- CART

Estos algoritmos han sido escogidos como grupo de referencia para el presente estudio experimental. Además se incluye en este grupo "*Simple Logistic Regression*" [103], una técnica de discriminación no paramétrica basada en funciones lineales, que se ha reportado [28] con éxito en el aprendizaje con datos de alta dimensión. Con ello suman 7 los algoritmos de referencia. En el caso particular de k-NN, el valor óptimo de $k$ para cada conjunto de datos fue estimado mediante LOOCV (validación cruzada dejando uno fuera)* sobre los datos de entrenamiento, escogiéndose el mejor valor para $k$ que cumpla:

$$k \in \mathrm{N} : 1 \leq k \leq 15 \tag{9}$$

Para la exploración de sus potencialidades en el marco de un multi-clasificador, se han utilizado los 7 ensamblados siguientes:

- Random Subspace Method (RSM) + C4.5
- RSM + SVM
- RSM + Naïves Bayes
- RSM + k-NN
- RSM + Logistic Regression (LR)
- Random Forest (RF)
- RF + CART

De esta forma, el estudio cuenta con un total de 14 algoritmos. Todos se encuentran implementados en la herramienta Weka y han sido ejecutados utilizando los valores predefinidos para todos sus parámetros. El Epígrafe 1.2.2 ofrece detalles sobre las familias a las que pertenecen los algoritmos y describe cada uno de ellos. En cuanto a los ensamblados, aparecen descritos y detallados en el Epígrafe 1.4.3. Como detalle adicional es necesario señalar que

* Remítase al Epígrafe 1.2.3. para más detalles sobre LOOCV.

el algoritmo AdaBoost, recogido en el grupo de referencia, es un multi-clasificador que implementa la técnica *boosting*, descrita también en el último epígrafe mencionado.

## 2.3. Conjuntos de datos utilizados

Usualmente en estudios experimentales en *machine learning*, las técnicas bajo estudio son sometidas a validación utilizando diferentes conjuntos de datos [102]. Ello persigue el propósito de obtener una estimación más fiable sobre el poder generalizador de cada modelo entrenado. Precisamente uno de los aportes realizados por varios eventos centrados en el aprendizaje supervisado con datos de alta dimensión, es que ponen a la disposición de los investigadores varios conjuntos de datos apropiados para estudios de este tipo. De esta forma se convierten en *benchmarks*, con los cuales la comunidad de machine learning tiene una base de datos común para probar sus técnicas y comparar los resultados.

En NIPS 2003 [26, 63] se realizó una competencia donde se publicaron 4 conjuntos de datos provenientes de dominios como la espectrometría, el OCR*, la minería de textos y la bioquímica. Fueron etiquetados con los nombres “Arcene”, “Gissete”, “Dexter”, “Dorothea”, respectivamente. Además se publicó “Madelon”, un conjunto de datos generados artificialmente con determinadas características afines al objetivo de la competencia. Cada uno de los conjuntos de datos se publicó dividido en 3 sobconjuntos: datos de entrenamiento, datos de validación y datos de prueba, para un total de 15 conjuntos de datos.

Para el presente estudio no fueron considerados los conjuntos de datos “Dexter” y “Madelon”, pues sus características no coinciden con la cantidad de variables y la complejidad que resultan de interés. Entre los restantes, fueron escogidos o construidos 3 conjuntos de datos, que se detallan en la Tabla 3. En la tabla, **N** representa la cantidad de casos, **m** la cantidad de variables y **c** la cantidad de

* *Optical Character Recognition*, reconocimiento de caracteres ópticos. Se aplica con frecuencia al reconocimiento de la escritura a mano.

clases. Las clases están igualmente representadas en cada conjunto de datos y equilibradas en cuanto a la cantidad de casos.

**Tabla 3**. Conjuntos de datos obtenidos de NIPS 2003.

| Conjunto de datos | N | m | c |
|---|---|---|---|
| Arcene | 200 | 3217 | 2 |
| Dorothea | 350 | 5699 | 2 |
| Gisette | 1000 | 979 | 2 |

En RSCTC 2010 [27] se realizaron dos competencias de aprendizaje supervisado: una básica y una avanzada. Para cada variante se publicaron 6 conjuntos de datos, divididos en datos de entrenamiento y datos de validación. Estos fueron extraídos de diferentes estudios genéticos y bioquímicos. Para el presente estudio fueron utilizados los 6 conjuntos de datos de la variante básica, uniendo los de entrenamiento y los de validación, al no disponer de los datos de la variante avanzada. Los conjuntos resultantes se detallan en la Tabla 4.

**Tabla 4**. Conjuntos de datos obtenidos de RSCTC 2010.

| Conjunto de datos | N | m | c |
|---|---|---|---|
| E-GEOD-10334 | 246 | 17897 | 2 |
| E-GEOD-5406 | 209 | 1178 | 3 |
| E-GEOD-13425 | 189 | 4342 | 5 |
| E-GEOD-13904 | 225 | 3828 | 5 |
| E-GEOD-4290 | 178 | 20315 | 4 |
| E-GEOD-9635 | 184 | 12696 | 5 |

Estos conjuntos de datos presentan un fuerte desequilibrio en sus clases, lo que añade complejidad al estudio.

Por último, los organizadores de RSCTC 2010 hacen referencia a los datos originales que utilizaron para crear ambas competencias. Estos se encuentran

en 5 conjuntos de datos, que los organizadores del evento dividieron en múltiples formas y a los cuales realizaron varios pre-procesamientos para hacerlos indistinguibles a ojos de los competidores. Por ello, el presente estudio emplea también los conjuntos de datos originales, diferentes a los empleados en las competencias. Éstos se detallan en la Tabla 5.

**Tabla 5**. Conjuntos de datos originales sin modificar para RSCTC 2010.

| Conjunto de datos | N | m | c |
|---|---|---|---|
| BurkittLymphoma | 220 | 8844 | 3 |
| HepatitisC | 383 | 23772 | 4 |
| MouseType | 123 | 11944 | 7 |
| OvarianTumor | 214 | 12233 | 3 |
| VariousCancer | 283 | 11907 | 10 |

Los 14 conjuntos de datos descritos presentan características deseables para el presente estudio, especialmente la gran cantidad de variables y muchos menos casos que variables (N<<m), excepto en los datos de “Gisette”, que se incluyó en el estudio de todas maneras para mayor diversidad. Existe además variedad en cuanto a la cantidad de clases y al equilibro de las clases en cada conjunto de datos.

Antes de someterlos a la experimentación, todos los conjuntos de datos fueron pre-procesados mediante una técnica de selección de rasgos, que redujo sus respectivas cantidades de variables a las cifras que se muestran en las tablas anteriores. Este pre-procesamiento persigue reducir la complejidad del aprendizaje y por tanto el costo computacional de la experimentación. Además teniendo en cuenta que se trata de datos procedentes de entornos reales, la selección de rasgos es una alternativa más recomendable que la construcción de variables, tal como indican los resultados de NIPS 2003 [26]. La construcción de variables implica mayor complejidad computacional y resta homogeneidad al estudio, ya que no todos los conjuntos de datos cumplen con las características

necesarias para aplicar las mismas técnicas. La selección aplicada consistió en un filtro uni-variable basado en incertidumbre simétrica*, que eliminó las variables poco informativas.

## 2.4. Resultados del estudio experimental

Luego de ejecutar 5x2cv de cada uno de los 14 algoritmos con cada uno de los 14 conjuntos de datos, se calcularon los desempeños de los algoritmos, medidos utilizando la exactitud balanceada, según la ecuación (6). Los resultados se muestran en las tablas 6 y 7.

**Tabla 6**. Exactitud balanceada de los algoritmos de referencia.

| | **SVM** | **NB** | **kNN** | **AdaB.** | **C4.5** | **CART** | **LR** |
|---|---|---|---|---|---|---|---|
| E-GEOD-10334 | **0.8709** | 0.8510 | 0.8032 | 0.8016 | 0.7856 | 0.7518 | 0.8420 |
| E-GEOD-5406 | **0.7886** | 0.6819 | 0.6822 | 0.6719 | 0.6309 | 0.5480 | 0.7107 |
| E-GEOD-13425 | **0.9292** | 0.7685 | 0.9115 | 0.8626 | 0.8118 | 0.7819 | 0.9029 |
| E-GEOD-13904 | **0.5464** | 0.5279 | 0.4576 | 0.4517 | 0.3847 | 0.2481 | 0.4942 |
| E-GEOD-4290 | **0.6587** | 0.6460 | 0.6520 | 0.6192 | 0.5578 | 0.5555 | 0.6445 |
| E-GEOD-9635 | 0.6241 | **0.6481** | 0.5699 | 0.5174 | 0.4334 | 0.3452 | 0.5783 |
| Arcene | **0.8433** | 0.7030 | 0.8174 | 0.7894 | 0.6946 | 0.6726 | 0.7989 |
| Gisette | **0.9556** | 0.9364 | 0.9358 | 0.9378 | 0.8988 | 0.8926 | 0.9476 |
| Dorothea | 0.6935 | **0.8188** | 0.5206 | 0.7342 | 0.7228 | 0.7803 | 0.7709 |
| BurkittLymphoma | **0.8673** | 0.8348 | 0.7625 | 0.7963 | 0.7040 | 0.7322 | 0.8304 |
| HepatitisC | **0.8907** | 0.8308 | 0.8568 | 0.8094 | 0.6988 | 0.7287 | 0.8807 |
| MouseType | **0.7414** | 0.5757 | 0.6012 | 0.5990 | 0.4968 | 0.4516 | 0.7118 |
| OvarianTumor | **0.7492** | 0.7347 | 0.6797 | 0.5760 | 0.5557 | 0.3805 | 0.6942 |
| VariousCancer | 0.8468 | 0.7370 | **0.8600** | 0.7735 | 0.6352 | 0.6025 | 0.8369 |

* Remítase al Epigrafe 1.3.1. para más detalles sobre selección de rasgos.

Los valores resaltados en negrita representan el mejor desempeño obtenido para cada conjunto de datos. Obsérvese cómo SVM supera a los demás algoritmos de referencia en casi todos los experimentos. Sin embargo, al enmarcarlos en un ensamblado se observa cómo LR se incorpora a la competencia.

**Tabla 7**. Exactitud balanceada de los algoritmos ensamblados.

| | **RSM+ SVM** | **RSM+ NB** | **RSM+ kNN** | **RF** | **RSM+ C4.5** | **RF+ CART** | **RSM+ LR** |
|---|---|---|---|---|---|---|---|
| E-GEOD-10334 | **0.8714** | 0.8494 | 0.8120 | 0.8299 | 0.8050 | 0.8255 | 0.8619 |
| E-GEOD-5406 | **0.7947** | 0.6841 | 0.6971 | 0.6730 | 0.6933 | 0.6464 | 0.7623 |
| E-GEOD-13425 | 0.9273 | 0.7544 | 0.9139 | 0.8404 | 0.8704 | 0.8258 | **0.9274** |
| E-GEOD-13904 | **0.5406** | 0.5272 | 0.4707 | 0.4406 | 0.4290 | 0.4251 | 0.5146 |
| E-GEOD-4290 | 0.6616 | 0.6339 | 0.6320 | 0.6216 | 0.5999 | 0.6101 | **0.6713** |
| E-GEOD-9635 | **0.6213** | 0.6149 | 0.5437 | 0.5400 | 0.4903 | 0.5064 | 0.5930 |
| Arcene | **0.8465** | 0.7037 | 0.8354 | 0.7835 | 0.7373 | 0.7644 | 0.8170 |
| Gisette | 0.9560 | 0.9404 | 0.9332 | 0.9350 | 0.9304 | 0.9264 | **0.9572** |
| Dorothea | 0.6961 | **0.8217** | 0.5088 | 0.6854 | 0.7263 | 0.6305 | 0.7188 |
| BurkittLymphoma | **0.8673** | 0.8222 | 0.7725 | 0.8075 | 0.7724 | 0.7659 | 0.8658 |
| HepatitisC | 0.8896 | 0.8180 | 0.8800 | 0.8424 | 0.7419 | 0.8420 | **0.8970** |
| MouseType | 0.7458 | 0.5637 | 0.5973 | 0.5938 | 0.5847 | 0.5826 | **0.7508** |
| OvarianTumor | 0.7490 | **0.7503** | 0.6888 | 0.5700 | 0.5651 | 0.5436 | 0.7309 |
| VariousCancer | 0.8458 | 0.6912 | **0.8682** | 0.6974 | 0.7323 | 0.7079 | 0.8448 |

Tomando como base estos resultados, se ha realizado un *test* de Friedman para verificar la existencia de diferencias significativas en el desempeño de los 14 algoritmos. Como parte de dicha prueba estadística se calcula el ranking promedio de cada técnica, quedando ordenados como se muestra en la Tabla 8. Además obsérvese en la Figura 10 el comportamiento de las 14 técnicas,

ordenadas según la media de sus exactitudes balanceadas sobre los 14 conjuntos de datos. Se muestra también en la gráfica el 100% de la desviación típica en su desempeño.

**Tabla 8**. Ranking promedio según el *test* de Friedman.

| Algoritmos | Ranking promedio |
|---|---|
| RSM+SVM | 1.678571 |
| SVM | 1.678571 |
| RSM+LR | 2.428571 |
| LR | 4.214286 |
| NaiveBayes | 5.571429 |
| RSM+kNN | 5.928571 |
| RSM+NB | 6.071429 |
| kNN | 6.214286 |
| AdaBoost | 7.5 |
| RF | 7.928571 |
| RSM+C4.5 | 8.714286 |
| RF+CART | 9.642857 |
| C4.5 | 11.5 |
| CART | 11.92857 |

**Tabla 9**. Resultados del *test* de Friedman ($\alpha$=0.05).

| Estadísticos de contraste | valor |
|---|---|
| N | 14 |
| Chi-cuadrado | 113.6485 |
| Significación asintótica | $3.59 \cdot 10^{-18}$ |

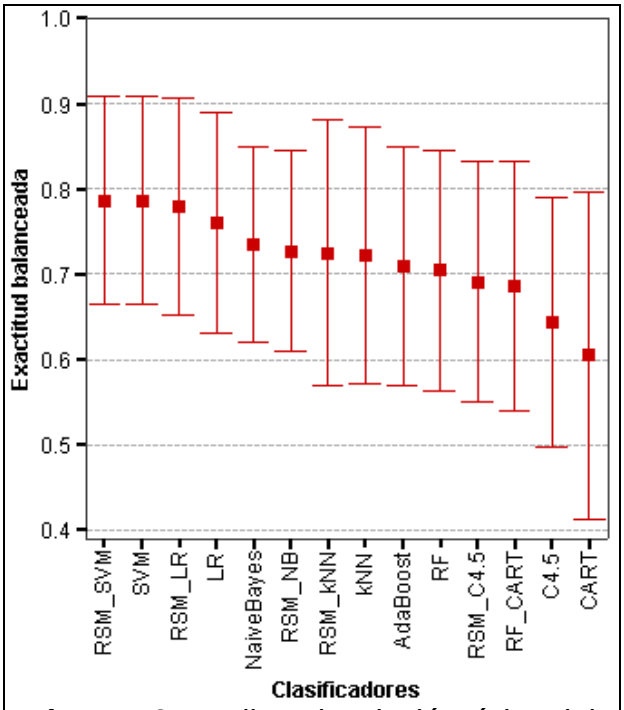


**Figura 10.** Media y desviación típica del desempeño de los algoritmos.

Como puede observarse en la Figura 10, el ordenamiento a partir de la media aritmética coincide casi exactamente con el ordenamiento por ranking. La única excepción la hacen los ensamblados de k-NN y de Naïve Bayes, un caso aislado, que aparecen intercambiados en los ordenamientos. Esto se debe a que el RSM+NB fue el más exacto para “Dorothea”, con una diferencia mucho mayor

sobre el RSM+kNN que la diferencia de RSM+kNN sobre RSM+NB en “VariousCancer”. Ya mediante un análisis a simple vista se destaca el buen desempeño y estabilidad de los tres primeros algoritmos del ranking, así como el pobre desempeño y la inestabilidad de los algoritmos basados en árboles.

Como puede apreciarse en la Tabla 9, el *test* de Friedman con un nivel de significación $\alpha$=0.05 arroja un valor de significación asintótica mucho menor que 0.05, lo cual permite rechazar la hipótesis nula. Los resultados evidencian, por tanto, que existen diferencias significativas en el desempeño de los algoritmos comparados.

Para detectar tales diferencias se ha realizado un *test* de Nemenyi con los valores del ranking promedio. Este procedimiento establece que el desempeño de dos algoritmos es significativamente diferente, si sus respectivos valores del ranking difieren en al menos el valor de cierta diferencia crítica (*critical difference*, CD) [102].

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}}\ , \tag{7}$$

donde el valor crítico $q_\alpha$ para $\alpha$=0.05 y 14 variables, correspondientes a los 14 algoritmos bajo comparación, es $q_{0.05}$=3.354 [104]. Al sustituir en (7) se obtiene:

$$CD = 3.354\sqrt{\frac{14(14+1)}{6\cdot 14}} = 5.303 \tag{8}$$

A partir del valor calculado, los resultados del *test* de Nemenyi pueden ser ilustrados mediante un diagrama de diferencias críticas (Figura 11). Los algoritmos cuya diferencia en ranking no excede la diferencia crítica no son considerados significativamente diferentes, y son unidos por una barra horizontal.

Pueden observarse en el diagrama dos grupos bien definidos de algoritmos, donde SVM y su ensamblado encabezan el grupo de mejor desempeño, mientras que los árboles de decisión y sus ensamblados están en el grupo de peor desempeño.

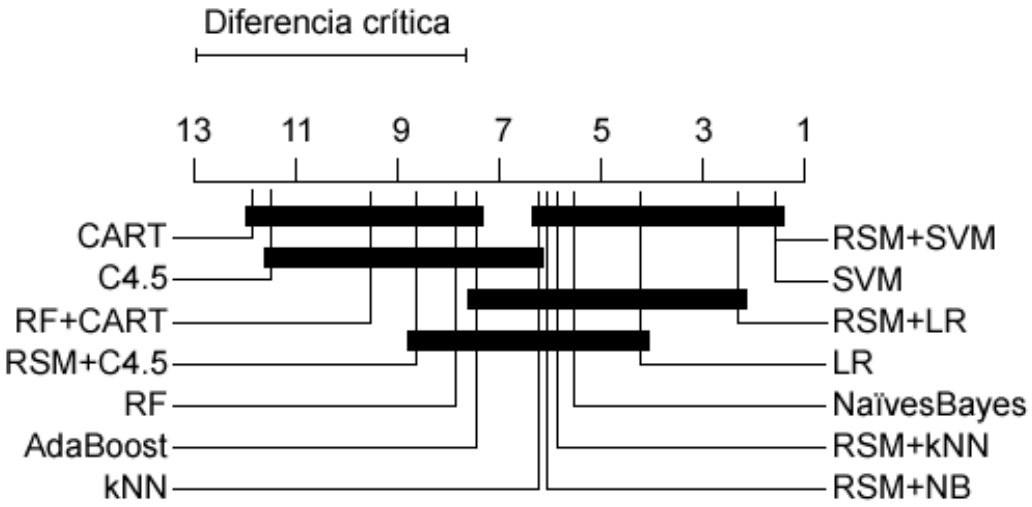


**Figura 11**. Comparación de todos los clasificadores contra los demás según la prueba de Nemenyi. Los grupos de clasificadores conectados no son significativamente diferentes.

Por último, para responder a la interrogante original de cómo se comportan los algoritmos de minería de datos cuando se enmarcan en un ensamblado, se ha procedido a explorar dentro de los grupos identificados por la prueba de Nemenyi. Se ha realizado un conjunto de 6 pruebas apareadas de Wilcoxon, donde se comparan 6 de los algoritmos con sus ensamblados.

En la Tabla 10 se muestran los valores de significación asintótica (p) de cada *test*, con $\alpha$=0.05. También se recoge en la tabla, como dato complementario para el análisis, la cantidad de veces que un algoritmo le gana al otro.

Las pruebas realizadas permiten rechazar las tres primeras hipótesis nulas, no así las restantes tres. En otras palabras, los ensamblados de árboles CART y C4.5 se comportan de forma significativamente superior a sus respectivos algoritmos de referencia, al igual que ocurre entre el ensamblado de *Simple Logistic Regression* y su algoritmo de referencia. Por otro lado, los algoritmos Naïve Bayes, k-NN y SVM no se desempeñan de forma significativamente diferente al enmarcarlos en multi-clasificadores.

**Tabla 10**. Resultado de las 6 pruebas apareadas de Wilcoxon ($\alpha$=0.05).

| Algoritmos | N | Significación asintótica (p) |
|---|---|---|
| RSM+C4.5 > C4.5<br>RSM+C4.5 < C4.5 | 14<br>0 | **0.000982** |
| RF+CART > CART<br>RF+CART < CART | 13<br>1 | **0.009181** |
| RSM+LR > SLR<br>RSM+LR < SLR | 13<br>1 | **0.015654** |
| RSM+NB > NaiveBayes<br>RSM+NB < NaiveBayes | 5<br>9 | 0.124043 |
| RSM+kNN > kNN<br>RSM+kNN < kNN | 9<br>5 | 0.396726 |
| RSM+SVM > SVM<br>RSM+SVM < SVM | 7<br>6 | 0.506746 |

## 2.5. Análisis de los resultados experimentales

El análisis estadístico de los resultados permite concluir que existe un potencial en algunos algoritmos de aprendizaje que merece ser explotado. En primer lugar, debe destacarse el buen desempeño de las máquinas de vectores de apoyo (SVM) en todos los conjuntos de datos utilizados. Ello es consistente con la descripción de dicho algoritmo y su diseño, del cual se dice que es robusto ante la alta dimensión y que no requiere gran cantidad de ejemplares para una buena capacidad de generalización [50].

Si bien es cierto que SVM no mejoró significativamente al enmarcarlo en un ensamblado, obsérvese que prácticamente todos los ensamblados superaron (significativamente o no) el desempeño de sus equivalentes de referencia, incluyendo a SVM, lo que puede apreciarse tanto en el ranking promedio, como en el gráfico de exactitud media, como en las pruebas de Wilcoxon. La

excepción es Naïve Bayes, que superó al correspondiente ensamblado, aunque no significativamente.

Una explicación podría ser que Naïve Bayes se beneficia al tener en cuenta todas las variables relevantes posibles y por tanto se perjudica ante una descomposición del conjunto de variables. Se ha afirmado [37, 50] que Naïve Bayes se beneficia con la selección de rasgos, pero téngase en cuenta que los ensamblados construidos en la experimentación fueron generados mediante una descomposición aleatoria del conjunto de rasgos. La aleatoriedad trae aparejada la posibilidad de que la descomposición obtenida esté lejos de ser óptima. En cualquier caso, debe reconocerse igualmente la robustez de Naïve Bayes, sin adaptaciones ni ensamblados, para trabajar con datos de alta dimensión. Ya ha sido demostrado [83] que el aprendizaje bayesiano, con una cuidadosa estimación de la densidad probabilística de las clases, puede ser una técnica exitosa.

El desempeño del clasificador k-NN, aunque mejoró al enmarcarlo en un ensamblado, no constituyó una mejora significativa. Mostró igualmente que la descomposición aleatoria del conjunto de variables puede estar limitando sus potencialidades. Teniendo en cuenta que, al proyectar el conjunto de datos hacia un espacio de menor cantidad de dimensiones que el original, pueden cambiar completamente las relaciones topológicas entre los objetos, la proyección debe ser cuidadosamente escogida.

Los algoritmos basados en árboles mostraron una gran sensibilidad a la complejidad de los conjuntos de datos y su alta dimensión. Debe recordarse que CART y C4.5 se encuentran entre los mejores algoritmos de aprendizaje supervisado para la minería de datos. Sin embargo, su pobre desempeño demuestra que los fenómenos asociados a la “maldición” de la alta dimensión ejercen una fuerte influencia sobre determinados métodos.

No obstante, todos los *test* demostraron que los algoritmos basados en árboles se benefician significativamente al servir de base para la construcción de un ensamblado. Esta potencialidad es la que ha impulsado el desarrollo de técnicas

de ensamblado basadas en árboles, como *Random Forest* y varias otras [28, 90]. El pobre desempeño de *Random Forest* en el presente estudio parece indicar que requiere un ajuste de sus parámetros para un mejor funcionamiento. Por ejemplo, es posible que requiera una mayor cantidad de modelos en el ensamblado, en lugar de solo 10, pero en tal caso debe verificarse si este incremento en la complejidad del modelo y en el consumo de recursos computacionales está justificado. Los experimentos muestran que varios otros algoritmos se comportaron significativamente mejor que *Random Forest*, utilizando solo 10 clasificadores.

Un análisis similar podría efectuarse sobre AdaBoost, que por lo general reporta sus mejores resultados asintóticos cuando construye entre 10 y 25 clasificadores [84, 101]. Pero sobre este algoritmo influye además el fenómeno del espacio vacío a causa del sub-muestreo del conjunto de datos, como fue detallado en el Epígrafe 1.4.3.

Por último deben destacarse como muy prometedores los resultados del algoritmo de discriminación *Simple Logistic Regression*. Obsérvese que, entre aquellos algoritmos que mejoraron significativamente al ensamblarlos (CART, C4.5 y LR), solo LR compite en el grupo de mejor desempeño. Es decir, no solo mejora al enmarcarlo en un ensamblado, sino que muestra un desempeño que no difiere significativamente de SVM y su ensamblado.

No obstante, los resultados sugieren que un diseño más cuidadoso de un ensamblado de LR o SVM, con un mejor ajuste de los parámetros, con una descomposición más adecuada, podría superar incluso al SVM de referencia. Igualmente podría esperarse una mejoría aún mayor en los ensamblados de Naïve Bayes o *k*-NN.

Existen como mínimo tres perspectivas prometedoras en lo que respecta al aprendizaje mediante multi-clasificadores con datos de alta dimensión. En primer lugar, la utilización de algoritmos de descomposición no aleatorios, orientados a obtener mejores proyecciones. En segundo lugar, el empleo de algoritmos de aprendizaje supervisado especialmente adaptados para funcionar en

condiciones de alta dimensión y pocos casos, como SVM, LR y Naïve Bayes. Por último, la aplicación de técnicas de combinación más sofisticadas que el voto mayoritario, entrenadas para desempeñarse en coordinación con el algoritmo de descomposición empleado.

## 2.6. Conclusiones parciales del capítulo

- El empleo de ensamblados como estrategia demuestra ser viable para la mejora de los indicadores de gestión en procesos de minería con datos de alta dimensión, al superar en casi todos los casos (excepto Naïve Bayes) el desempeño de los algoritmos de referencia (**Tabla 8**), aunque no siempre significativamente.
- La estrategia de ensamblados podría ser especialmente adecuada en la gestión de procesos donde resulte conveniente el empleo de árboles de decisión. Los *tests* estadísticos muestran (**Tabla 10**) una mejora significativa en el indicador de gestión utilizado (exactitud) cuando se emplean algoritmos basados en árboles, enmarcados en un ensamblado.
- Los algoritmos basados en discriminación lineal, como SVM y LR, mostraron gran robustez al enfrentarse con datos de alta dimensión, especialmente al entrenarlos en un ensamblado. Por su efectividad y simplicidad, constituyen los mejores candidatos para el diseño futuro de un ensamblado con descomposición no aleatoria y combinación más sofisticada, que podría exhibir un desempeño aún mejor que los algoritmos del presente estudio.
- Naïve Bayes y k-NN no se benefician con una descomposición aleatoria del conjunto de variables, pero con otro tipo de descomposición podrían beneficiarse significativamente, en especial k-NN.
- *Random Forest* y AdaBoost no se beneficiaron con el tamaño escogido para el ensamblado. Es incierto si merecen o no un estudio aparte para confirmar que el aumento en complejidad se ve justificado por un aumento significativo en la exactitud. En el caso de AdaBoost, se confirma además su alta sensibilidad al fenómeno del espacio vacío.

# CONCLUSIONES

- El empleo de ensamblados como estrategia demuestra ser viable para la mejora de los indicadores de gestión en procesos de minería con datos de alta dimensión, especialmente la exactitud de la predicción. Esto, unido a las potencialidades de esta estrategia que apenas comienzan a ser explotadas, convierte al uso de ensamblados en la estrategia más promisoria.
- La estrategia de ensamblados podría ser especialmente adecuada en la gestión de procesos donde resulte conveniente el empleo de árboles de decisión. No obstante, los algoritmos basados en árboles demostraron ser altamente sensibles a la “maldición” de la alta dimensión, por lo cual su elección sobre otros algoritmos, cuando se trabaja con datos de alta dimensión provenientes de procesos tecnológicos reales, debe ser cuidadosamente estudiada.
- El objetivo planteado para esta investigación fue alcancado, como avalan las conclusiones parciales del Capítulo 2, así como estas conclusiones generales.
- La naturaleza de las adaptaciones propuestas a los algoritmos parece indicar que debe continuarse estudiándose, con más profundidad, el efecto que ejerce la “maldición” de la alta dimensión sobre la forma en que los algoritmos representan y manipulan los datos y los modelos.

# RECOMENDACIONES

- Aplicar los algoritmos señalados como los mejores a un problema empresarial o tecnológico real, con el fin de estudiar el impacto real sobre otros indicadores de gestión, así como implementar un sistema basado en conocimiento que sirva de apoyo a la toma de decisiones.
- Extender el estudio experimental realizado en la presente investigación, enriqueciéndolo con otros entornos de gestión, otros conjuntos de datos y con otras configuraciones de los algoritmos y los ensamblados.
- Realizar un estudio experimental comparativo entre la estrategia de ensamblado y las adaptaciones propuestas en los últimos años a los algoritmos SVM y k-NN.

# Referencias Bibliográficas


1. K.J. Cios, W. Pedrycz, R.W. Swiniarski, and L.A. Kurgan, *Data Mining. A Knowledge Discovery Approach*, Springer, 2007.
2. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “Knowledge Discovery and Data Mining: Towards a Unifying Framework,” *KDD-96 Proceedings*.
3. D. François, “High-dimensional data analysis: optimal metrics and feature selection,” Université Catholique de Louvain. Faculté des Sciences Apliquées, 2007.
4. L. Wang, and X. Fu, *Data Mining with Computational Intelligence*, Springer-Verlag, 2005.
5. A. Konar, *Artificial Intelligence and Soft Computing. Behavioral and Cognitive Modeling of the Human Brain*, CRC Press LLC, 2000.
6. N.J. Nilsson ed., Introduction to Machine Learning, Department of Computer Science. Stanford University, 1997.
7. U. Alon, et al., “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” *Proc. Nat. Acad. Sci.*, vol. 96, 1999, pp. 6745-6750.
8. T.R. Golub, D.K. Slonim, and P. Tamayo, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, 1999, pp. 531-537.
9. H.E. Parkinson, et al., “Arrayexpress update - from an archive of functional genomics experiments to the atlas of gene expression,” *Nucleic Acids Research*, vol. 37, 2009, pp. 868-872.
10. W. Reik, and J. Walter, “Genomic imprinting: parental influence on the genome,” *Natur. Rev. Genetics*, vol. 2, no. 1, 2001, pp. 21–32.
11. R. Blair, et al., “Estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands,” *Toxicological Sci.*, vol. 54, 2000, pp. 138-153.

12. W.S. Branham, et al., “Phytoestrogens and mycoestrogens bind to the rat uterine estrogen receptor,” *J. Nutrition*, vol. 132, 2002, pp. 658-664.
13. A.K. Debnath, G. Debnath, A.J. Shusterman, and C. Hansch, “ A QSAR investigation of the role of hydrophobicity in regulating mutagenicity in the Ames test: 1. Mutagenicity of aromatic and heteroaromatic amines in Salmonella typhimurium TA98 and TA100,” *Environmental and Molecular Mutagenesis*, vol. 19, 1992, pp. 37-52.
14. C. Glende, H. Schmitt, L. Erdinger, G. Engelhardt, and G. Boche, “Transformation of mutagenic aromatic amines into non-mutagenic species by alkyl substituents. Part I: Alkylation ortho to the amino function,” *Mutation Research*, vol. 498, 2001, pp. 19-37.
15. J. Weston, F. Perez-Cruz, O. Bousquet, O. Chapelle, A. Elisseeff, and B. Schoelkopf, “Feature Selection and Transduction for Prediction of Molecular Bioactivity for Drug Design,” *Bioinformatics*, 2002.
16. H. Bagan, W. Takeuchi, Y. Yamagata, X. Wang, and Y. Yasuoka, “Extended Averaged Learning Subspace Method for Hyperspectral Data Classification,” *Sensors*, vol. 9, 2009, pp. 4247-4271.
17. B.-C. Kuo, and K.-Y. Chang, “Regularized Feature Extractions and Support Vector Machines for Hyperspectral Image Data Classification,” *LNAI*, vol. 3681, 2005.
18. D. Landgrebe, and M.M. Dundar, “Toward an Optimal Supervised Classifier for the Analysis of Hyperspectral Data,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 1, 2004, pp. 271-277.
19. E.F. Petricoin, et al., “Use of proteomic patterns in serum to identify ovarian cancer,” *The Lancet*, vol. 359 no. 1, 2002.
20. E.F. Petricoin, et al., “Serum proteomic patterns for detection of prostate cancer,” *Journal of the NCI*, vol. 94, no. 20, 2002.
21. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When Is "Nearest Neighbor" Meaningful?,” *Lecture Notes in Computer Science*, vol. 1540, 1999, pp. 217-235.

22. P.F. Evangelista, M.J. Embrechts, and K. Szymanski Boleslaw, “Applied Soft Computing Technologies: The Challenge of Complexity,” A. Abraham, B.d. Baets, M. Koppen, and B. Nickolay eds., Springer Verlag, 2006.
23. D. François, V. Wertz, and M. Verleysen, “The concentration of fractional distances,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, 2007, pp. 873-886.
24. Y. Liu, Y. Liu, and K.C.C. Chan, “Dimensionality reduction for heterogeneous dataset in rushes editing,” *Pattern Recognition*, vol. 42, no. 2, 2009, pp. 229-242.
25. M. Radovanović, A. Nanopoulos, and M. Ivanović, “Nearest Neighbors in High-Dimensional Data: The Emergence and Influence of Hubs,” *Proceedings of the 26th International Conference on Machine Learning*.
26. I. Guyon, S.R. Gunn, A. Ben-Hur, and G. Dror, “Design and Analysis of the NIPS2003 Challenge,” *StudFuzz*, vol. 207, 2006, pp. 237–263.
27. M. Wojnarski, A. Janusz, S.N. Hung, and others, “RSCTC'2010 Discovery Challenge: Mining DNA Microarray Data for Medical Diagnosis and Treatment,” *Lecture Notes in Artificial Intelligence*, vol. 6086, 2010, pp. 4-19.
28. H. Ahn, H. Moon, M.J. Fazzari, N. Lim, J.J. Chen, and R.L. Kodell, “Classification by ensembles from random partitions of high-dimensional data,” *Computational Statistics & Data Analysis*, vol. 51, 2007, p. 6166.
29. L. Rokach, “Genetic algorithm-based feature set partitioning for classiffication problems,” *Pattern Recognition*, vol. 41, 2008, p. 1676-1700.
30. L. Rokach, “Collective-agreement-based pruning of ensembles,” *Computational Statistics and Data Analysis*, vol. 53, 2009, pp. 1015-1026.
31. S. Sun, “An Improved Random Subspace Method and Its Application to EEG Signal Classification,” *LNCS*, vol. 4472, 2007, pp. 103-112.
32. K. Kira, and L. Rendell, “The feature selection problem: traditional methods and new algorithm,” *Proceedings of AAAI'92*.
33. G. Piatetsky-Shapiro, “Knowledge Discovery in Real Databases,” *AI Magazine*, 1991.

34. T.M. Mitchell ed., Machine Learning, McGraw-Hill Science, 1997.
35. K. Fukunaga ed., Introduction to Statistical Pattern Pattern Recognition. Second Edition, Morgan Kaufmann, 1990.
36. I.H. Witten, and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, 2005.
37. I. Guyon, and A. Elisseeff, *Feature extraction: Foundations and Applications*, Springer-Verlag, 2006.
38. G.C. Cawley, and N.L.C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Journal of Machine Learning Research*, vol. 11, 2010, pp. 2079-2107.
39. T.G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural Computation*, vol. 10, 1998, pp. 1895-1924.
40. R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley, 2001.
41. A.R. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, 2002.
42. D. Michie ed., Machine Learning, Neural and Statistical Classification, Elsevier, 1994.
43. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.
44. J.R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, 1986., pp. 81–106.
45. J.R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
46. W. Cohen, "Fast Efficient Rule Induction.," *Twelfth International Conference on Machine Learning*, pp. 115-123.
47. D.M. Titterington, G.D. Murray, L.S. Murray, D.J. Spiegelhalter, A.M. Skene, J.D.F. Habbema, and G.J. Gelpke, "Comparison of discrimination techniques applied to a complex data set of head injured patients," *Journal of the Royal Statistical Society*, vol. 144, 1981, pp. 145-175.
48. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.

49. D.E. Rumelhart, G.E. Hinton, and R.J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, 1986, pp. 533-536.
50. X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” *Knowl Inf Syst*, vol. 14, 2008, pp. 1-37.
51. D.W. Aha, D. Kibler, and M.K. Albert, “Instance-based learning algorithms,” *Machine Learning*, vol. 6, no. 1, 1991, pp. 37–66.
52. C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
53. P. Gund, G. Maggiora, and J. Snyder eds., Gidebook on Molecular Modeling in Drug Design, Academic Press, 1996.
54. J. Beltrán, M.A.C. Calvo, R.C. Pérez, M.A.R. Zapata, and F.T. Panchón, *Guía para una gestión basada en procesos*, Instituto Andaluz de Tecnología, 2002.
55. J. Jaworska, N. Nikolova-Jeliazkova, and T. Aldenberg, “QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review,” *ATLA*, vol. 33, 2005, pp. 445-459.
56. V.A. Huynh-Thu, L. Wehenkel, and P. Geurts, “Exploiting tree-based variable importances to selectively identify relevant variables,” *JMLR W&P*, vol. 4, 2008, pp. 60-73.
57. C. Rao, “The utilisation of multiple measurements in problems of biological classification,” *Journal of the Royal Statistical Society*, vol. 10, 1948, pp. 159–203.
58. R. Bellman, *Adaptive control processes: A guided tour*, Princeton University Press, 1961.
59. L. Breiman, “Random forest,” *Machine Learning*, vol. 45, no. 1, 2001, pp. 5-32.
60. T.K. Ho, “The random subspace method for constructing decision forests,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, 1998, pp. 832-844.

61. G. Hughes, “On the mean accuracy of statistical pattern recognizers,” *Information Theory, IEEE Transactions on*, vol. 14, no. 1, 1968, pp. 55-63.
62. I. Guyon, and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *Journal of Machine Learning Research*, vol. 3, no. 26, 2003.
63. I. Guyon, S.R. Gunn, A. Ben-Hur, and G. Dror, “Result Analysis of the NIPS 2003 Feature Selection Challenge,” *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*.
64. L. Yu, and H. Liu, “Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution,” *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, AAAI Press, pp. 856-863.
65. F. Fleuret, “Fast Binary Feature Selection with Conditional Mutual Information,” *Journal of Machine Learning Research*, vol. 5, 2004, pp. 1531-1555.
66. Y.W. Chen, and C.J. Lin, “Combining SVMs with various feature selection strategies,” *Feature extraction, foundations and applications*, Springer, 2006.
67. W. Duch, T. Wieczorek, J. Biesiada, and M. Blachnik, “Comparison of feature ranking methods based on information entropy,” *IEEE International Joint Conference on Neural Networks. Proceedings.*, pp. 1415-1419.
68. V. Lemaire, and F. Clerot, “An input variable importance definition based on empirical data probability and its use in variable selection,” *Proceedings IEEE International Joint Conference on Neural Networks.*, pp. 1375 - 1380.
69. P. Liu, N. Wu, J. Zhu, J. Yin, and W. Zhang, “A Unified Strategy of Feature Selection,” *Lecture Notes in Artificial Intelligence*, vol. 4093, 2006, pp. 457-464.
70. R. Ruiz, J.C. Riquelme, and J.S. Aguilar-Ruiz, “Heuristic Search over a Ranking for Feature Selection,” *Computational Intelligence and Bioinspired Systems*, 2005, pp. 742-749.

71. L. Song, A. Smola, A. Gretton, K.M. Borgwardt, and J. Bedo, “Supervised feature selection via dependence estimation,” *ICML '07: Proceedings of the 24th international conference on Machine learning*, ACM, pp. 823-830.
72. T. Balachander, and R. Kothari, “Introducing Locality and Softness in Subspace Classification,” *Pattern Analysis & Applications*, vol. 2, 1999, pp. 53-58.
73. J. Laaksonen, “Subspace Methods in Recognition of Handwritten Digits,” Helsinki University of Technology, 1997.
74. M. Prakash, and M.N. Murty, “Extended subspace methods of pattern recognition,” *Pattern Recognition Letters*, vol. 17, no. 11, 1996, pp. 1131-1139.
75. S. Watanabe, P.F. Lambert, C.A. Kulikowski, J.L. Buxton, and R. Walker, “Evaluation and selection of variables in pattern recognition,” *Computer and Information Sciences*, Academic Press, pp. 91-122.
76. M.-H. Yang, “Discriminant Isometric Mapping for Face Recognition,” *ICVS*, pp. 470-480.
77. K.-S. Park, S.-H. Oh, W.-J. Yoon, and K.-K. Lee, “A Robust Approach to Content-Based Musical Genre Classification and Retrieval Using Multi-feature Clustering,” *ASIAN*, pp. 212-222.
78. S. Tadjudin, and D. Landgrebe, “Classification of High Dimensional Data with Limited Training Samples,” School of Electrical and Computer Engineering. Purdue University., 1998.
79. P. Hall, D.M. Titterington, and J.-H. Xue, “Median-Based Classifiers for High-Dimensional Data,” *Journal of the American Statistical Association*, vol. 104, no. 488, 2009, pp. 1597-1608.
80. M. Radovanović, A. Nanopoulos, and M. Ivanović, “Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data,” *Journal of Machine Learning Research*, vol. 11, 2010, pp. 2487-2531.
81. S. Yoon, “Regularizing covariance estimation by quantized eigenvalues and its application to image classification,” *38th Conference on Signals, Systems and Computers, 2004.*, pp. 1687 - 1689.

82. R.M. Neal, “Density Modeling and Clustering Using Dirichlet Diffusion Trees,” *Bayesian Statistics*, vol. 7, 2003, pp. 619-629.
83. R.M. Neal, and J. Zhang, “High Dimensional Classification with Bayesian Neural Networks and Dirichlet Diffusion Trees,” *StudFuzz*, vol. 207, 2006, pp. 265-296.
84. T.G. Dietterich, “Ensemble Methods in Machine Learning,” *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, Springer, pp. 1-15.
85. R.A. Berk, “An Introduction to Ensemble Methods for Data Analysis,” *Sociological Methods & Research*, vol. 34, no. 3, 2006, pp. 263-295.
86. L.I. Kuncheva, *Combining Pattern Classifiers. Methods and algorithms*, Wiley-Interscience, 2004.
87. L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, 1996, pp. 123-140.
88. Y. Freund, and R. Schapire, “Experiments with a new boosting algorithm,” *Proceedings of the Thirteenth International Conference on Machine Learning*, pp. 148-156.
89. O. Maimon, and L. Rokach, *Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications*, World Scientific, 2005.
90. L. Rokach, and O. Maimon, “Feature set decomposition for decision trees,” *Intelligent Data Analysis*, vol. 9, 2005, pp. 131-158.
91. K.M.O. Vale, F.G. Dias, A.M.P. Canuto, and M.C.P. Souto, “A Class-Based Feature Selection Method for Ensemble Systems,” *Eighth International Conference on Hybrid Intelligent Systems*.
92. G. Giacinto, and F. Roli, “Adaptive selection of image classifiers,” *Lecture Notes in Computer Science*, vol. 1310, 1997, pp. 38-45.
93. J. Xiao, C. He, X. Jiang, and D. Liu, “A dynamic classifier ensemble selection approach for noise data,” *Information Sciences*, vol. 180 2010, pp. 3402-3421.
94. C. Ferri, P. Flach, and J. Hernández-Orallo, “Delegating classifiers,” *Proceedings of the 21st International Conference on Machine Learning*.

95. C.-H. Chuang, B.-C. Kuo, and H.-P. Wang, “Fuzzy Fusion Method for Combining Small Number of Classifiers in Hyperspectral Image Classification,” *8th International Conference on Intelligent Systems Design and Applications*.
96. B. Zhang, and S.N. Srihari, “Class-Wise Multi-Classifier Combination Based on Dempster-Shafer Theory,” *Proceedings of the 7th International Conference On Control, Automation, Robotics And Vision (ICARV 2002)*.
97. L. Didaci, and G. Giacinto, “Dynamic Classifier Selection by Adaptive k-Nearest-Neighbourhood Rule,” *LNCS*, vol. 3077, 2004, pp. 174-183.
98. L.I. Kuncheva, “Clustering-and-selection model for classifier combination,” *Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, pp. 185-188.
99. L. Kuncheva, “Switching Between Selection and Fusion in Combining Classifiers: An Experimental,” *IEEE Transactions on Systems, Man, and Cybernetics. Part B: Cybernetics*, vol. 32, 2002.
100. T.K. Ho, and M. Basu, “Complexity Measures of Supervised Classification Problems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, 2002, pp. 289–300.
101. D. Opitz, and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” *Journal of Artifcial Intelligence Research*, vol. 11, 1999, pp. 169-198.
102. J. Demšar, “Statistical Comparisons of Classifiers over Multiple Data Sets,” *Journal of Machine Learning Research*, vol. 7, 2006, pp. 1-30.
103. M. Sumner, E. Frank, and M. Hall, “Speeding up Logistic Model Tree Induction,” *9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 675-683.
104. 25/octubre/2010, “Critical values for Nemenyi test,” http://nikolaos.kourentzes.com/Nemenyi_critvals.pdf.