**TAREAS Y TÉCNICAS DE LA MINERÍA DE DATOS**

**Integrantes:**

Fredy Ricardo Cortés Ramírez  
1193443881

Karen Lisbeth Gelvez Lesmes  
1090526367

**Universidad de Pamplona.**

**Facultad de ingenierías y arquitectura.**

**Programa de ingeniería de sistemas.**

**Electiva de Ingeniería I**

**Villa del Rosario.**

**2019-1**

**Contenido**

[INTRODUCCIÓN 3](#_Toc12727521)

[PLANTEAMIENTO DEL PROBLEMA 3](#_Toc12727522)

[JUSTIFICACIÓN 4](#_Toc12727523)

[OBJETIVOS 4](#_Toc12727524)

[Objetivo General 4](#_Toc12727525)

[Objetivos Específicos 4](#_Toc12727526)

[MARCO TEÓRICO 5](#_Toc12727527)

[DESARROLLO 7](#_Toc12727528)

[**Tareas Predictivas** 7](#_Toc12727529)

[**Tareas Descriptivas** 8](#_Toc12727530)

[**Árboles de Decisión** 8](#_Toc12727531)

[**Agrupamiento** 10](#_Toc12727532)

[**Regresión Lineal** 11](#_Toc12727533)

[**Redes Neuronales** 12](#_Toc12727534)

[**Modelos estadísticos** 14](#_Toc12727535)

[**Técnica K-vecinos** 16](#_Toc12727536)

[CONCLUSIONES 17](#_Toc12727537)

[REFERENCIAS 18](#_Toc12727538)

Índice de Figuras

[Figura 1. Ejemplo de Árbol de Decisión en una Transacción 9](#_Toc12727544)

[Figura 2. Técnica de agrupamiento de datos 11](#_Toc12727545)

[Figura 3. Regresión de una variable dependiente y una variable independiente 12](#_Toc12727546)

[Figura 4. Estructura de una red neuronal artificial 14](#_Toc12727547)

[Figura 5. Ejemplo de Medidas de dispersión de datos 15](#_Toc12727548)

[Figura 6. Ejemplo de Probabilidad al sacar una pelota roja 15](#_Toc12727549)

[Figura 7. Ejemplo de la Técnica K-vecinos 16](#_Toc12727550)

# INTRODUCCIÓN

El uso de minería de datos en el entorno académico es casi o probablemente ignorado. La realidad es que las empresas o entornos de producción invierten lo necesario para poder utilizar y ejecutar las técnicas que subyacen de la minería de datos. Por tanto en este documento se realizará una exploración a cerca de las tareas y técnicas que abarca la minería de datos, esto con el fin de conocer a profundidad los alcances y beneficios que pueden llegar a tener al momento de implementarse en el entorno académico.

# 

# PLANTEAMIENTO DEL PROBLEMA

El uso de la Minería de datos en las empresas o entornos de producción es tan demandado que dichas empresas ofrecen lo necesario por obtener y ejecutar las técnicas que trae consigo la minería de datos. Esto porque son tantos los beneficios que pueden llegar a generar el uso de estas técnicas en la misma producción o comercialización de los productos que fabrican estas empresas, ya sea predicciones de ventas, análisis de calidad de producción, etc. Todo esto con el fin de buscar éxito o mejoramiento del producto a ofrecer.

Todo esto conlleva a hacernos una pregunta. ¿Por qué la minería de datos no es un tema que se abarque demasiado en el entorno académico?

# JUSTIFICACIÓN

Se debe conocer los alcances y beneficios que brinda el uso de las técnicas y tareas que abarca la minería de datos en los entornos académicos. Esto para dejar atrás la inseguridad que se tiene al momento de arraigar estos temas y empezar a comprender lo necesario que resulta implementar estas técnicas en proyectos reales de producción masiva.

# OBJETIVOS

## Objetivo General

Conocer las tareas, técnicas y aplicaciones de la Minería de Datos obteniendo información en páginas de internet.

## Objetivos Específicos

* Buscar en diferentes sitio de internet sobre lo que abarca la Minería de Datos
* Clasificar los temas, técnicas y aplicaciones encontradas en diferentes fuentes.
* Comprender los resultados obtenidos de la búsqueda y clasificación.

# MARCO TEÓRICO

Los **datos** representan un fragmento de una cantidad, medida, descripción o palabra, los cuales son agrupados o **clasificados de una determinada manera para generar de información.** En forma general, los datos solo sirven después de ser procesados según una intención y relevancia. Las bases de datos, por ejemplo, agrupan los datos en estructuras lógicas y sistemáticas para luego ser analizadas o procesadas según un propósito.

La **minería de datos** es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.

Para poder obtener resultados o patrones esperados, existen varias **técnicas**, las cuales se aplican a los grandes volúmenes de conjunto de datos que se poseen. Dichas técnicas son algoritmos y provienen de la inteligencia artificial y la estadística.

Las **tareas** de Minería de Datos integran toda la información que se necesita para iniciar un proceso de preparación y calcular el modelo de minería. Dicha información consiste en los valores de minería y definición de los datos de entrada.

Un proceso común de Minería de Datos suele tener cuatro partes principales:

1. **Determinación de los objetivos:** A partir de las necesidades del cliente y con orientación del especialista en data mining.
2. **Preprocesamiento de datos:** Es la selección y transformación de las bases de datos.
3. **Determinación del modelo:** Comienza con el análisis estadísticos de los datos, y después se implementa un primer boceto del diseño de la visualización. En esta etapa pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. **Análisis de los resultados:** Se verifica la coherencia de los resultados y se los compara con los obtenidos por estadística y visualización gráfica. También aquí, el cliente opina si se trata de resultados novedosos, útiles y si sirve a sus objetivos iniciales.

La minería de datos puede ser útil en tipos específicos de misiones comerciales. Por ejemplo: determinando la efectividad de un programa de Marketing, manejo del ciclo de vida de un cliente, mostrando escenarios posibles en un proceso de negocio y procesando datos para luego combinarlos en partes para mejor visualización.

# DESARROLLO

A continuación se describen las diferentes tareas en la minería de datos.

### **Tareas Predictivas**

Utiliza algunas de las variables para predecir los valores futuros desconocidos de la misma variable o bien de otras variables.

Existen dos tipos de Tareas Predictivas:

* Clasificación: el sistema de minería de datos dará a cada objeto o dato que se encuentre una clase. Esta tarea asocia o empareja datos a grupos predefinidos y encuentran modelos que describen y distinguen conceptos para predicciones futuras. Es quizá la tarea más conocida y popular en la minería de datos.
* Regresión: se usa para predecir los valores ausentes de una variable, basándose en su relación con otras variables del conjunto de datos. Existe regresión lineal, regresión no lineal, regresión logística, regresión logarítmica, regresión univariada, regresión multivariada, etcétera.

### **Tareas Descriptivas**

Busca patrones humano-interpretables que describen los datos. Existen tres tipos de Tareas Descriptivas:

* Agrupamiento: esta tarea busca organizar los datos en clases. Divide los datos en grupos, de tal forma que los grupos capturan la estructura natural de los datos. Divide datos sin etiqueta de grupos de tal forma que datos que pertenecen al mismo grupo son similares y datos que pertenecen a diferentes grupos son diferentes.
* Reglas de Asociación: dichas reglas describe una relación de asociación entre los elementos de un conjunto de datos relevantes.
* Secuenciación: esta tareas es usada para descubrir secuencias de patrones en los datos, estos patrones son similares a los encontrados con reglas de asociación pero tales relaciones son basadas en el tiempo.

A continuación se describen las diferentes técnicas en la minería de datos.

### **Árboles de Decisión**

Los árboles de decisión son un modelo de predicción que provee de una herramienta de clasificación muy potente, que dada una gran cantidad de datos, éste realiza diagramas de construcciones lógicas permitiendo representar y categorizar una serie de condiciones. Su uso en el manejo de datos la hace muy popular por las posibilidades que brinda y la facilidad en la compresión de sus resultados por cualquier usuario, permitiendo:

* Segmentación: estableciendo los grupos más importantes para clasificar un cierto ítem.
* Clasificación: asignar ítems a uno de los grupos en la que está particionada una población.
* Predicción: establecer reglas para poder hacer predicciones ante distintos eventos.
* Reducción de la dimensión de los datos: identificar cuáles son los datos importantes para realizar los modelos de un fenómeno.
* Identificación-interrelación: identificar las variables y relaciones importantes para los grupos identificados a partir del análisis de datos.
* Recodificación: establecer criterios cualitativos perdiendo la menor cantidad de información posible.

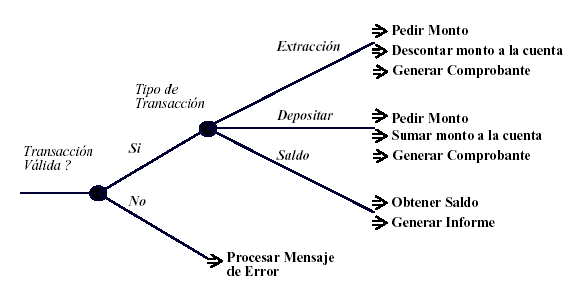


Figura 1. Ejemplo de Árbol de Decisión en una Transacción

### **Agrupamiento**

Este es un proceso de agrupar datos de acuerdo a un criterio especificado. Estos criterios suelen ser generalmente de semejanza o distancia.

Dividir datos sin etiqueta en grupos (clusters) de tal forma que los datos que pertenecen al mismo grupo son similar, y datos que pertenecen a diferentes grupos son diferentes.

Los algoritmos más utilizados en esta técnica se presentan a continuación.

* **Algoritmo K-Means**
* Agrupamiento particional.
* Cada cluster está asociado con un centroide (valor de la media del cluster).
* Cada punto es asignado al cluster más cercano al centroide.
* El número de clusters “K” debe ser especificado.
* El algoritmo básico es muy simple:

1. Seleccionar K puntos como los centroides iniciales.
2. **Repetir.**
3. Desde K clusters asignar todos los puntos al centroide más cercano.
4. Recalcular el centroide de cada cluster.
5. **Hasta** El centroide no cambia

* **Algoritmo K-Medoids**
* K-medoid es una técnica clásica de particionado de grupos que divide los datos conformados por *n* objetos en *k* grupos (con *k* conocido de antemano).
* Es más robusto ante el ruido y a partes aisladas que k-means porque minimiza una suma de disimilaridades (entre pares de puntos) en vez de una suma de distancias euclidianas cuadradas.

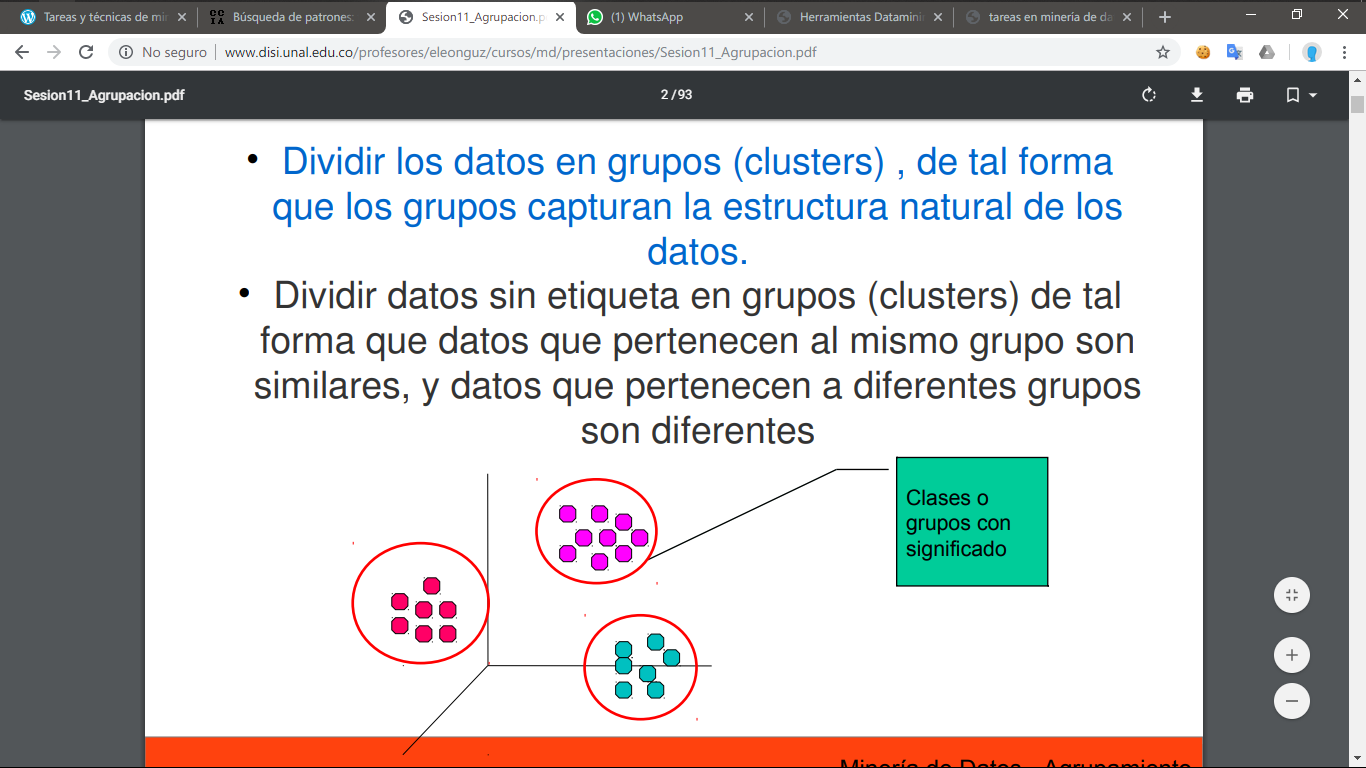


Figura 2. Técnica de agrupamiento de datos

### **Regresión Lineal**

Esta técnica se utiliza principalmente en la minería de datos, ya que con ella podemos relacionar los datos.

Este es un modelo matemático utilizado para aproximar la relación de la dependencia entre una variable dependiente y otras variables independientes en un término aleatorio.

Esta técnica de minería de datos es aplicada en las líneas de tendencia, lo cual nos dice si un conjunto de datos en específico han aumentado o decrecido en un periodo determinado, dependiendo de la curvatura o la independencia lineal de los datos con respecto al periodo es como se correlaciona a la extracción de datos.

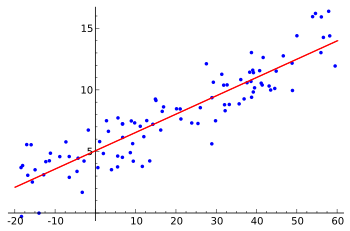


Figura 3. Regresión de una variable dependiente y una variable independiente

### **Redes Neuronales**

Son un conjunto de elementos de procesamiento de información altamente interconectados, que son capaces de aprender con la información con que se alimentan.

Puede aplicarse a gran número de problemas que puede ser tanto complejos reales como teóricos sofisticados, como por ejemplo.

* Reconocimiento de imágenes.
* Reconocimiento de voz.
* Análisis y filtrado de señales.
* Clasificación.
* Discriminación.
* Análisis financiero.
* Predicción Dinámica.
* Entre otros.

Se puede definir una red neuronal artificial como un sistema inteligente capaz de **aprender** y **generalizar**.

Está formada por unidades de procesamiento que reciben el nombre de **neuronas** o **nodos**. Estas a su vez están organizadas en grupos que se llaman **capas**.

Cada neurona está conectada con otras a través de unos enlaces. En estos enlaces el valor de salida de la neurona anterior es multiplicado por un valor de peso. Estos pesos en los enlaces pueden incrementar o inhibir el estado de activación de las neuronas adyacentes. Del mismo modo, a la salida de la neurona, puede existir una función limitadora o umbral, que modifica el valor resultado o impone un límite que se debe sobrepasar antes de propagarse a otra neurona. Esta función se conoce como [función de activación](https://es.wikipedia.org/wiki/Funci%C3%B3n_de_activaci%C3%B3n).

El objetivo de la red neuronal es resolver los problemas de la misma manera que el cerebro humano, aunque las redes neuronales son más abstractas. Las redes neuronales actuales suelen contener desde unos miles a unos pocos millones de unidades neuronales.

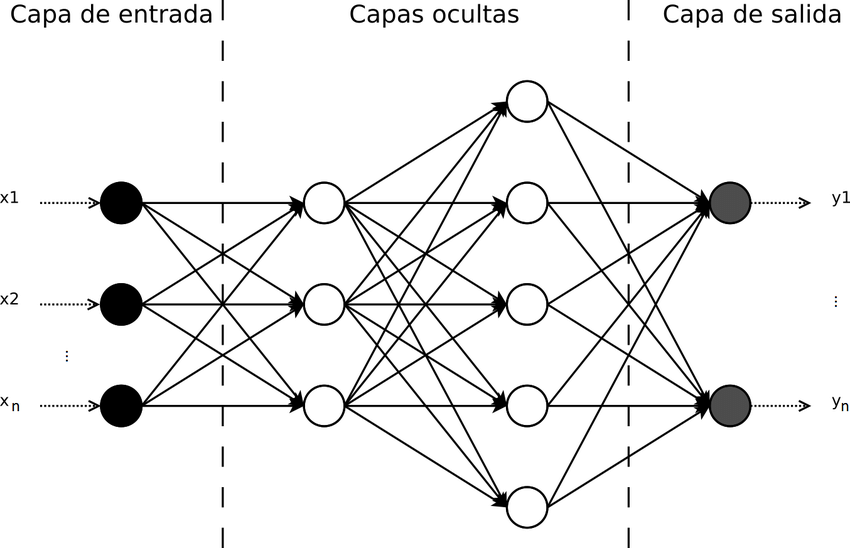


Figura 4. Estructura de una red neuronal artificial

### **Modelos estadísticos**

Los modelos estadísticos son usados en la toma de decisiones, ya que es casi absoluta su aplicabilidad en la minería de datos. Se tienen datos crudos que se procesan con un fin específico. La estadística aquí ayuda a corroborar las hipótesis planteadas o encontrar patrones, debido a que hay gran cantidad de fórmulas y técnicas específicas que se pueden aplicar usando los datos crudos obtenidos, para luego procesarlos matemáticamente y así tomar la decisión u obtener el aprendizaje esperado o necesitado.

Algunos ejemplos de modelos estadísticos son:

* Medidas de dispersión: son parámetros estadísticos que indican como se alejan los datos respecto de la media aritmética. Sirven como indicador de la variabilidad de los datos.

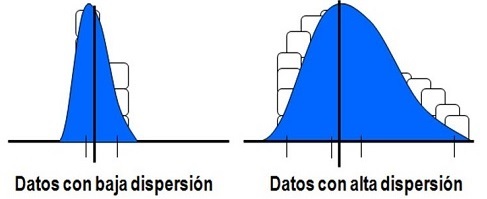


Figura 5. Ejemplo de Medidas de dispersión de datos

* Teoría de probabilidad: **es una herramienta matemática que establece un conjunto de reglas o principios útiles para calcular la ocurrencia o no ocurrencia de fenómenos aleatorios y procesos estocásticos.**

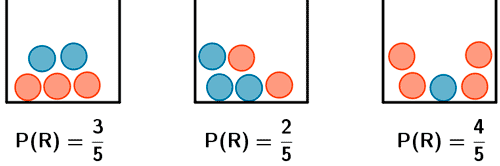


Figura 6. Ejemplo de Probabilidad al sacar una pelota roja

### **Técnica K-vecinos**

La técnica de los K-vecinos es considerada como un método de clasificación fácil de entender y fácil de implementar. Dicha técnica se centra en buscar datos con tendencias y características semejantes, que permitan obtener información relevante. El algoritmo de esta técnica, clasifica cada dato nuevo en el grupo que corresponda, según tenga k vecinos más cerca de un grupo o de otro. Es decir, calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenecer.

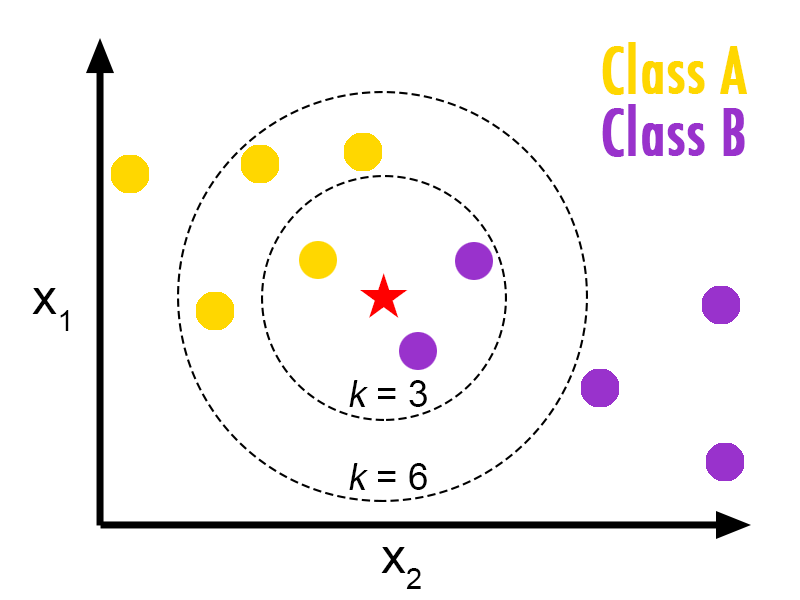


Figura 7. Ejemplo de la Técnica K-vecinos

# CONCLUSIONES

* El uso de la minería de datos en el entorno académico puede ser tan provechoso para los estudiantes, ya que nos prepara para un entorno laboral demandante en aspectos de datos masivos, con los cuales podemos aplicar las técnicas estudiadas para el tratamiento o análisis de estos.
* Con la minería de datos no simplemente obtenemos cantidades de datos valiosos, también ayuda a empresas o entornos de producción a tener un control de calidad y comercialización de su producto de venta, a partir de las técnicas que subyacen de la minería de datos, las cuales permiten realizar una serie de acciones beneficiosas como pueden ser predicciones de ventas que al final ayudan a estos entornos de producción a mejorar integralmente y posteriormente perfeccionar el modelo de comercialización. Por lo cual podemos concluir que la minería de datos es una parte fundamental y clave en la inteligencia de negocio.

# REFERENCIAS

[1] Concepto de datos. Disponible en:

<https://www.significados.com/datos/>

[2] Concepto de Minería de datos. Disponible en:

<https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017>

[3] Concepto de Árbol de decisión. Disponible en:

<https://es.wikipedia.org/wiki/Minería_de_datos#Técnicas_de_minería_de_datos>

<http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion11_Agrupacion.pdf>

[4] Conceptos Básicos sobre la Minería de Datos. Disponible en: <http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentaci%C3%B3n_-_Conceptos_B%C3%A1sicos.41132532.pdf>