# Unidad 8

Minería de Datos

# Introducción

# Aprendizaje Automático

- Es una de las técnicas utilizadas en la Minería de Datos para extraer información en su forma natural y presentarla de manera comprensible

# Minería de Datos

- La Minería de Datos a grandes rasgos es la extracción de información a través de encontrar patrones en los datos
- Se define como el proceso de descubrir patrones en los datos, ya sea de manera automática o semi automática

# Fundamentos de la Minería de Datos

- La Minería de Datos ha surgido a partir del desarrollo e integración de otras ramas, entre las que se encuentran:
  - Estadística
  - Inteligencia Artificial
  - Aprendizaje Máquina / Automatizado (*Machine Learning*)

# Datos

- Los datos son almacenados y la búsqueda es realizada de manera automática por un sistema de tal manera que busca identificarlos, validarlos y en su momento usarlos para predecir comportamientos futuros

# Patrones

- Los patrones que se lleguen a encontrar deberán ser útiles y deberán servir para un cierto fin, predicciones no triviales de nuevos datos
- Si bien los patrones pueden verse como cajas negras, también es posible analizarlos y entender el por qué de las predicciones realizadas

# Patrones Estructurados

- La obtención de patrones estructurados se dará a través del análisis de datos estructurados
- Los datos estructurados son aquellos con una organización (estructura) bien definida, por ejemplo los que están almacenados en una Base de Datos

# Ejemplo

- En base a ésta información, ¿será posible decir si un futuro cliente sería bueno o malo?

| id | Edad | Hijos | Salario | Buen Cliente |
|---|---|---|---|---|
| 1 | joven | no | alto | si |
| 2 | joven | no | medio | no |
| 3 | joven | si | medio | no |
| 4 | mayor | si | bajo | si |
| 5 | mayor | si | alto | si |
| 6 | joven | si | alto | si |

# Minería de Datos

# Tareas de la Minería de Datos

- Las tareas en Minería de Datos pueden clasificarse en predictivas o descriptivas
  - Las tareas **Predictivas** estiman valores futuros o desconocidos de variables de interés usando otras variables
  - Las tareas **Descriptivas** identifican patrones que explican o resumen los datos, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos

# Clasificación de las Tareas

- Tareas Predictivas: Clasificación y Regresión.
- Tareas Descriptivas: Agrupamiento (clustering), asociación, asociación secuencial, correlaciones

# Clasificación

- Cada instancia (o registros de una base de datos) pertenece a una clase, la cuál se indica mediante el valor de un atributo que se denomina clase de la instancia

- El atributo puede tomar varios valores discretos, cada uno perteneciente a una clase, el resto de los atributos se utilizan para predecir la clase

- El objetivo es predecir la clase de nuevas instancias de las que se desconoce la misma

# Ejemplo

- Una Universidad además del examen de admisión quiere considerar el historial académico de sus candidatos
- A la vez ha seguido el comportamiento de sus alumnos actuales y ha relacionado ésta información con la del historial académico

# Regresión

- Consiste en aprender una función real que asigna a cada instancia un valor real
- La principal diferencia respecto a la Clasificación es que el valor a predecir es numérico

# Ejemplo

- Se quiere predecir el comportamiento del clima a partir de estadísticas pasadas

# Clasificación vs Regresión

- La principal diferencia entre Clasificación y Regresión que la Regresión busca predecir datos no vistos
- La Clasificación busca clasificar información ya conocida a partir de información ya vista

# Agrupamiento

- También conocido como Clustering o Segmentación
- Permite obtener grupos a partir de los datos
- Esto generaría una etiqueta (clase) que después se utilizaría para la clasificación
- Los datos son agrupados buscando maximizar la similitud entre los elementos y minimizar la similitud entre grupos
- Los objetos de un mismo grupo son similares entre sí y distintos de los objetos de otro grupo

# Ejemplo

- El departamento de Recursos Humanos busca agrupar a los empleados en base a ciertas características de tal forma que pueda comprender mejor su comportamiento

# Asociación

- Ofrece un comportamiento muy parecido a las Correlaciones
- Identifica relaciones no explícitas entre atributos categóricos.
- El planteamiento más común es del estilo "*si el atributo X toma el valor d, entonces el atributoY toma el valor b*"
- Dos elementos relacionados con la Asociación son la Cobertura y la Precisión
  - Cobertura. Es el porcentaje de transacciones que contienen elementos del lado izquierdo
  - Precisión. Medida de cuántas veces puede ser verdadera una regla

# Ejemplo

- Un análisis muy claro de una asociación se da en las tiendas, por ejemplo, si un cliente compra Refrescos, es muy posible que compre Botanas

$$\{\text{refresco}\} \Rightarrow \{\text{botanas}\}$$

# Cobertura y Precisión

- Cobertura. Es el porcentaje de transacciones que contienen elementos del lado izquierdo {refrescos} y del lado derecho {botanas}
  - Si se tienen 100 ventas y en 10 de ellas se compraron botanas y refrescos, la cobertura es de 10%
  - Si se tienen 100 ventas, pero solo 50 compraron refrescos y de esas 50 solo 10 compraron botanas, entonces hay una precisión del 20%

# Asociación Secuencial

- Permite determinar patrones secuenciales basados en secuencias temporales de acciones
- Contrario a las reglas de Asociación, se basan en el tiempo
- Se tienen dos secuencias de acciones que contrario a una Asociación normal, se dan en distintos momentos de tiempo
- A la primera secuencia se le conoce como **Predictor** de la segunda
- La Precisión se define como la probabilidad de que cuando ocurra una acción, tiempo después ocurrirá otra

# Ejemplo

- Considerar una tienda de videojuegos, el evento **Predictor** sería la compra de una {consola}
- Tiempo después de que se compra una consola, se esperaría que el cliente pudiera comprar {juegos} o {accesorios}
- También podría darse un evento Predictor si un cliente compra un {juego} y tiempo después sale una {secuela}

# Patrones de Tiempo

- Es un tipo especial de secuencia de eventos que son todos del mismo tipo
- Se utilizan para descubrir patrones y secuencias en ciertos periodos de tiempo

# Correlaciones

- Examina el grado de similitud de los valores de dos variables numéricas
- La fórmula estándar para medir la correlación lineal es el coeficiente de correlación *r*, que tiene valores entre -1 y 1
  - Si *r* es > 0, cuando una variable crezca o decrezca, la otra tendrá un comportamiento similar
  - Si *r* es < 0, cuando una variable crezca o decrezca, la otra tendrá un comportamiento opuesto

# Ejemplo

- Una tienda buscaría encontrar la correlación entre el aumento de ventas o la disminución de las mismas
- Un ejemplo sería el aumento de ventas ante la presencia de ofertas o descuentos (Correlación Positiva)
- Por el contrario, podría haber factores que presenten una Correlación Negativa, en este caso se podría considerar las estaciones en la disminución de venta de cierto tipo de ropa