# RESUMEN DE TECNICAS

**Marcelo Adrian Lopez Peña 1803561**

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

Minería de Datos

**Facultad de Ciencias Físico Matemático**
**02/10/2020**

En la Minería de Datos las tareas se pueden dividir en dos categorías

- **Descriptivas:** Estas nos ayudan a descubrir las características mas importantes de las bases de datos, estas son Clustering, Reglas de Asociación, Detección de Outliers y Visualización de Datos
- **Predictivas:** estas sirven para predecir los valores de un atributo en particular basándose en los resultados recolectados de otros atributos, estas son Regresión, Clasificación, Patrones de Secuencia y Predicción.

# Clustering – Equipo 4

Una técnica utilizada en la minería de datos, su proceso consiste en la división de los datos en grupos de objetos similares, esta técnica es la más utilizada en algoritmos matemáticos se encargan de agrupar objetos.

Esta se puede dividir en dos conceptos

- **Cluster**: colección de objetos de datos similares entre si dentro del mismo grupo
- **Análisis de cluster**: Dado un conjunto de puntos de datos trata de entender su estructura, encontrar similitudes entre los datos.

**Aplicaciones**

- Estudio de terremotos
- Planificación de la ciudad
- Marketing
- Aseguradoras
- Uso del suelo

**Métodos de agrupación**

- Asignación jerárquica frente a punto
- Datos numéricos o Simbólicos
- Determinística vs Probabilística
- Exclusivo vs Superpuesto
- Jerárquico vs Plano
- De arriba abajo y De abajo a arriba

Algoritmos de clustering

Simple k-means: para utilizarlo debemos tener definido el número de clusters que se desean obtener

1. Asumir de forma aleatoria los centros de cada cluster, el algoritmo hará los siguientes pasos

- Determina coordenadas del centroide
- Determina la distancia de cada objeto a los centroides
- Agrupa los objetos basados en la menor distancia

2. Quedaran agrupados los clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo.

- **X-means**: variante mejorada del K-means, se le define un limite inferior K-min (Número mínimo) y un limite superior K-max (número máximo) el algoritmo es capaz de obtener el número óptimo de clusters.

- **Cobweb:** aprendizaje incremental, realiza agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol donde las hojas representan los segmentos y el nodo rai engloba por completo el conjunto de datos.

## Reglas de Asociación – Equipo 1

Esta es una búsqueda de patrones frecuentes o estructurals causales entre conjuntos de elementos u objetos en bases de datos de transacciones, u otros repositorios de información.

Este es un método para la generación de los elementos que aparecen con mayor frecuencia se utiliza el Principio Priori, este reduce el número de candidatos. Este algoritmo fue uno de los primeros en ser desarrollados y se compone de dos etapas:

1.- Identificar los ítems sets que ocurren con mayor frecuencia

2.- Convertir esos ítems sets frecuentes en reglas de asociación

**Aplicaciones**

- Análisis de datos de la banca
- Cross-marketing
- Diseño de catálogos

**Objetivo**: el objetivo de la minería de reglas de asociación es encontrar todas las reglas teniendo el Umbral mínimo de soporte y el Umbral mínimo de confianza.

# Detección de Outliers – Equipo 2

Estudio el comportamiento de los valores extremos que difieren del patrón general de una muestra

### Valores Atípicos

Son valores diferentes a los que observamos en el mismo grupo de datos, los datos atípicos ocasionados por

- Errores de entrada y procedimiento
- Acontecimientos extraordinarios
- Valores extremos

**Técnicas de detección para los valores atípicos** (los cuales se pueden eliminar o sustituir)

- Prueba de GRUBBS
- Prueba DIXON
- Prueba de TUKEY
- Análisis de valores
- Regresión Simple

### Aplicación de Outliers

- Detección de fraudes financieros
- Tecnología Informática y Telecomunicaciones
- Nutrición y Salud
- Negocios

# Visualización de Datos – Equipo 7

Esta representa los datos en un formato ilustrados. Esto nos proporciona una manera accesible de comprender los datos de manera visual o gráfica.

### Tipos de Visualización de Datos

- Gráficos
- Mapas
- Infografías
- Cuadros de mando

**Aplicaciones**

- Comprender la Información
- Identifica Relaciones y Patrones
- Identificar Tendencias Emergentes

# Regresión – Equipo 3

Modelo matemático para determinar el grado de dependencia entre una o mas variables, es decir conocer si existe relación entre ellas, existen dos tipos:

**Regresión lineal:** cuando una variable independiente ejerce influencia sobre otra variable dependiente

**Regresión lineal múltiple:** cuando dos o mas variables independientes influyen sobre una variable dependiente

El objetivo es analizar los datos de un conjunto, para predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

Permite examinar la relación entre dos o más variables para así identificar cuales son las de mayor impacto.

**Tipos de Variables**

- **Variable Dependiente:** La variable que se intenta predecir
- **Variable Independiente:** Es el factor que influye en tu variable dependiente.

**Factores Arrojados**

R representa el coeficiente de correlación y $R^2$ representa el coeficiente de determinación.

# Clasificación – Equipo 8

Consiste en el ordenamiento por clases tomando en cuenta las características de los elementos que contiene.

Existen distintos métodos de clasificación:

- **Análisis discriminante**: es utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos
- **Reglas de clasificación**: buscan términos no clasificados de forma periódica, si se encuentra una coincidencia se agrega a los datos de clasificación.
- **Arboles de decisión**: método analítico que a través de una representación esquemática ayuda a decidir

- **Redes neuronales artificiales**: también conocidas como sistemas conexionistas, es utilizado para mandar múltiples señales.

**Características de los métodos de clasificación**

- Precisión en la predicción
- Eficiencia
- Robustez
- Escalabilidad
- Interpretabilidad

# Patrones de Secuencia – Equipo 6

Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Estos eventos se enlazan con el aspo del tiempo. Sus reglas expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos.

**Características**

- Orden
- Objetivo
- Tamaño de secuencia es cantidad de elementos
- Longitud de secuencia es cantidad de ítems
- Soporte de secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S
- Secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

**Aplicaciones**

- Medicina
- Análisis de mercado
- Web

**Ventajas**

- Flexibilidad en comportamiento de parámetros
- Eficiencia de cálculos

**Desventajas**

- Utilización de valores adecuados para parámetros
- Sesgado por los primeros patrones

# Predicción – Equipo 5

Utilizada para la proyección de tipos de datos, para predecir resultados de eventos.

### Aplicaciones

- Banca
- Clima
- Deportes
- Inmobiliaria

### Técnicas de predicción

- Modelos estadísticos simples como regresión
- Estadísticas no lineales como series de potencias.
- Redes neuronales, RBF, etc

### Características de la predicción

- Los valores son generalmente continuos
- Las predicciones sueles ser sobre el futuro
- Variables independientes corresponden a los atributos ya conocidos
- Variables de respuesta corresponden a lo que queremos saber