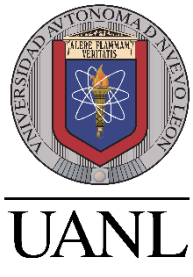

UNIVERSIDAD AUTONOMA DE NUEVO LEÓN

Facultad De Ciencias Físico Matemáticas

MINERIA DE DATOS

Milton Humberto Zuñiga Cedilo 1863305

G003

Las tareas de la minería de datos se dividen en dos categorías:

- **Descriptivas**: nos ayudan a descubrir las características mas importantes de las bases de datos.
- **Predictivas:** predecir el valor de un atributo en particular basándose en los resultados recolectados de otros atributos.

Descriptivas	Predictivas
Clustering	Regresión
Reglas de asociación	Clasificación
Detección de outliers	Patrones de secuencia
Visualización	Predicción

Analizaremos cada tarea de manera individual.
En cuanto a las tareas predictivas podemos concluir lo siguiente:

Clustering

También conocido como agrupamiento, el proceso consiste en la división de los datos en grupos de objetos similares.
Las técnicas de clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos y una vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las distintas clases.

Conceptos básicos:

- **Cluster**: Colección de objetos de datos similares entre si dentro del mismo grupo, disimilar a los objetos en otros grupos.
- **Análisis de cluster**: Dado un conjunto de puntos de datos tratar de entender su estructura, encontrar similitudes entre los datos de acuerdo con las características encontradas.

Aplicaciones:

- Estudios de terremotos.
- Aseguradoras.
- Uso del suelo.
- Marketing.

Métodos de agrupación:

- Asignación jerárquica frente a un punto.
- Datos numéricos y/o simbólicos.
- Determinística vs probabilística.
- Excesivo 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, siguiendo los siguientes pasos:
    1. Se asume de forma aleatoria los centros para cada cluster, el algoritmo hará los tres pasos siguientes:
        1. Determina las coordenadas del centroide.
        2. Determina la distancia de cada objeto a los centroides.
        3. 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:** es una variante mejorada del K-means, se le define un limite inferior K-min (número mínimo de clusters) y un limite superior K-max (número máximo de clusters) el algoritmo es capaz de obtener el numero óptimo de clusters.
- **Cobweb:** pertenece a la familia de algoritmos jerárquicos, durante la ejecución del algoritmo se forma un árbol (árbol de clasificación) donde las hojas representan los segmentos y el nodo la raíz englobada por completo el conjunto de datos.

Reglas de asociación

Búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elementos u objetos en bases de datos de transacciones, bases de datos de relaciones y otros repositorios de información disponibles.

Conceptos básicos:

- **Soporte:** fracción de transacciones que contiene un itemset.
- **Conjunto de elementos frecuente:** un conjunto de elementos cuyo soporte es mayor o igual que un umbral mínimo.

- **Conjunto de elementos:** una colección de uno o mas artículos, por ejemplos, {leche, pan, mermelada}. K-itemset, un conjunto de elementos que contiene k elementos**.**
- **Recuento de soporte: frecuencia de ocurrencia de un itemset.**
- **Confianza (c):** mide que tan frecuentes ítems en Y aparecen en transacciones que contiene X.

Aplicaciones:

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

Objetivo: dado un conjunto de transacciones T, el objetivo de la minería de reglas de asociación es encontrar todas las reglas teniendo:

- Umbral mínimo de soporte.
- Umbral mínimo de confianza.

Enfoque de fuerza bruta:

- Lista todas las reglas de asociación posibles.
- Compruebe el soporte y la confianza para cada regla.
- Elimine las reglas que fallan en los umbrales mínimos.

Métodos de reglas de asociación

- **Enfoque de dos pasos:**
  1. **Generación de elementos frecuentes:** generar todos los conjuntos de elementos cuyo soporte $\geq$ min soporte.
  2. **Generación de reglas:** generar reglas de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una partición binaria de un conjunto de elementos frecuente.
- **Principio a priori:** si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes, el soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos (propiedad anti-monótona de soporte).

Detección de outliers

Estudia el comportamiento de valores extremos que difieren del patron general de una muestra.

Valor atipico: observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos, distorsionan los resultados de los análisis.

Técnicas para la de detección de valores atípicos:

- Prueba de Grubbs.
- Prueba de Dixion.
- Prueba de Tukey (diegrama de caja).
- Análisis de valores atípicos de Mahalanobis.
- Regresión simple (regresión por mínimos cuadrados).

Se puede eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variable.
Si no se debe a un error, eliminarlo o sustituirlo puede modificar las inferencias que se realicen a partir de esta información, debido a que:

- Introduce un sesgo.
- Disminuye el tamaño muestral.
- Puede afectar a la distribución y a las varianzas.

Visualización

Es la presentación de información en formato ilustrado o gráfico.
Al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos.

Tipos de visualización de datos:

- **Gráficos:** es el tipo más común y conocido, lo utilizamos en nuestro día a día con las hojas de calculo para representar datos de manera sencilla, como gráficos circulares, líneas, columnas y barras.
- **Mapas:** se empezó a utilizar mas seguido con la popularización de Google Maps, nos sirve para la localización de nuestra flota de vehículos en tiempo real o bien la de las tiendas de un supermercado.
- **Infografías:** es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente.

- **Cuadros de mando:** es una herramienta que permite saber en todo momento el estado de los indicadores del negocio: de ventas, económicos, de producción, de recursos humanos y más.

A continuación, empezaremos a hablar sobre las tareas predictivas:

Regresión

Es un modelo matemático para determinar el grado de dependencia entre una o mas variables, es decir conocer si existe relación entre ellas.

Existen dos tipos de regresión:

- **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.

Tiene como objetivo analizar los datos de un conjunto y en base a eso, predecir lo que puede ocurrir con ese conjunto de datos en un futuro.
Permite examinar la relación entre dos o mas variables e identificar cuales son las que tienen mayor impacto en un tema de interés.

- **Variable dependiente:** es el factor mas importante, el cual se está tratando de entender o predecir.
- **Variable independiente:** es el factor que tú crees que pueda impactar en tu variable dependiente.

Podemos explicar un fenómeno y predecir cosas acerca del futuro, por lo que nos será de ayuda para tomar decisiones y obtener los mejores resultados.

Clasificación

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

Métodos de la clasificación:

- **Analisis discriminante:** se utiliza para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos.
- **Reglas de clasificación:** buscan términos no clasificados de forma periodica, si se encuentra una coincidencia se agrega a los datos de clasificación.

- **Arboles de decisión:** es un método analítico que a través de una representación esquemática facilita la toma de decisiones.
- **Redes neuronales artificiales:** también se le conoce como sistema conexionista, es un modelo de unidades conectadas para transmitir señales.

Características de los métodos de clasificación:

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

Patrones secuenciales

Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el paso del tiempo.
EL orden de los acontecimientos es considerado.
Se buscan asociaciones de la forma "si sucede de la forma X en el instante de tiempo t entonces sucederá en el evento Y en el instante t+n". El objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.
Expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos en el tiempo.

Características:

- El orden importa.
- Objetivo: encontrar patrones secuenciales.
- El tamaño de una secuencia es su cantidad de elementos.
- La longitud de la secuencia es la cantidad de ítems.
- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

Aplicaciones:

- **Medicina:** predecir si un compuesto químico causa cáncer.
- **Análisis de mercado:** comportamiento de compras.
- **Web:** reconocimiento de spam de un correo electrónico.

Predicción

Se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir de un evento.
En muchos casos, el simple hecho de reconocer y comprender las tendencias históricas es suficiente para trazar una predicción algo precisa de lo que sucederá en el futuro.

Características de la predicción:

- Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo.
- Los valores son generalmente continuos.
- Las predicciones son a menudo (no siempre) sobre el futuro.

Aplicaciones:

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.
- Predecir si va a llover en función de la humedad actual.
- Predecir el precio de venta de una propiedad.
- Predecir la puntuación de cualquier equipo durante un partido de futbol.

Técnicas de predicción:

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