Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Semestre

Agosto 2020 ~ Enero 2021

Licenciatura en Actuaría

MINERÍA DE DATOS

Docente: Mayra Berrones

Grupo:003

# Resumen
# "Técnicas de Minería de Datos"


Alumnos:
Daniel de la Rosa Coss

Matricula:
1723784

# 1.-Detección de Outliers

Estudia el comportamiento de los valores extremos que difieren del patrón general de una muestra, a dichos valores se les conoce como “valor atípico”.

Dicho de otra forma, los valores atípicos son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos y estos datos atípicos son consecuencia de:

- Errores de entrada de datos y procedimiento.
- Acontecimientos extraordinarios.
- Valores extremos y/o faltantes.
- Causas no conocidas.

La razón por la cual se desea detectar dichos valores atípicos es debido a que estos datos distorsionan los resultados de los análisis.

Para detectar los valores atípicos existen distintos tipos de técnicas las cuales se pueden dividir en dos categorías principales: Métodos univariantes de detección de outliers y Métodos multivariantes de detección de outliers.

➔ **Técnicas para la detección de valores atípicos:**

**Prueba de Grubbs:** Dentro de dicha prueba hay dos formas para detectar un outlier, una de estas es mediante el p-valor, recurso estadístico en el cuál se calcula la probabilidad y dependiendo del valor Alpha se rechazará la hipótesis nula o alternativa, dependiendo de dicho valor. La segunda forma aplicar dicha prueba es mediante el valor crítico, para esto se calcula un estadístico en el cual se toma un valor que sea sospechoso de ser atípico y a este se le resta el valor de la media de los datos, se aplica el valor absoluto y se divide entre la desviación estándar de los datos, mediante un valor obtenido en las tablas de Grubbs (el cual se toma en base a la significancia y el tamaño de la muestra) se compara con el estadístico de prueba y el valor de tablas y dependiendo de cual es mayor se clasifica o no como outlier.

**Prueba de Dixon:** Esta prueba es similar a la prueba de Grubbs solo que esta prueba solo se puede aplicar para una muestra de datos que sea menor a 26 elementos.

**Prueba de Turkey:** Esta prueba es sencilla, ya que solo se ocupa hacer un diagrama de caja en el cual vemos la concentración de los datos (entre otras medidas de tendencia central) y nos permite observar de manera clara si hay un outlier en nuestros datos.

**Análisis de Valores (Mahalanobis):** Para esta técnica se requiere de programas de computador específicos para aplicar dicha técnica.

**Regresión Simple:** Para esta técnica existen distintas formas de detectar outliers, tomando el método de mínimos cuadrados podemos ver los residuos de cada dato al aplicar la

regresión, dicho residuo sale de la distancia que hay entre en dato y el valor ajustado a la regresión y existen aún otras formas de detectar de manera gráfica un outlier y para esto también existen distintos softwares de programación.

➔ **¿Qué hacer cuando se detecta un outlier?**

Existen distintas opiniones al respecto, una de estas que 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, pero, si no se debe a un error, eliminarlo o sustituirlo puede modificar las inferencias que se realicen a partir de esa información, debido a tres razones:

- Introduce un sesgo,
- disminuye el tamaño muestras y
- puede afectar a la distribución y las varianzas.

Lo recomendable es quitarles peso a esas observaciones atípicas mediante técnicas robustas.

➔ **Aplicaciones de la minería de datos en outliers:**

- Detección de fraudes financieros.
- Tecnología informática y telecomunicaciones.
- Nutrición y salud.
- Negocios.

## 2.-Regresión.

Una regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir, conocer si existe relación entre ellas. Existen 2 tipos de regresión: la Regresión Lineal Simple, cuando una variable independiente ejerce sobre otra variable dependiente y la Regresión Lineal Múltiple donde dos o más variables independientes influyen sobre una variante dependiente.

En la minería de datos, la regresión es usada como un predictor, es decir, 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.

➔ **Análisis de Regresión**

El análisis de regresión permite examinar la relación entre dos o más variables e identificar cuáles son las que tiene mayor impacto en un tema de interés. Dentro de este análisis observamos que tanto realmente se ajustan nuestros datos al modelo de regresión lineal (simple o múltiple). Una de las formas para saber que el modelo se ajusta bien a nuestros datos es mediante la significancia, si tiene un p-valor mayor a alpha entonces la regresión se ajusta a los datos, otra forma es mediante el grafico de probabilidad normal. Si los puntos

de los residuos se ajustan a la recta, entonces podemos decir que nuestros datos tienen una distribución normal estándar, otros gráficos que también ayudan a esto ultimo es el histograma que de igual manera debe tener forma de una distribución normal estándar. Existen otras formas para comprobar que la regresión se ajusta a nuestros datos como la R y la $R^2$ y de igual manera para la regresión múltiple.

## 3.-Clustering.

El Clustering también se conoce como agrupamiento, el proceso consiste en la división de datos en grupos de objetos similares.

Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos, para esto se usa la información que brindan las variables que pertenecen a cada objeto, midiendo la similitud entre los mismos, y a la vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de diferentes clases.

➔ **Análisis de Cluster.**

Dado un conjunto de puntos de datos de tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

➔ **Aplicaciones:**

- *Aseguradoras:* Identificación de grupos de asegurados de automóviles con un alto costo promedio de reclamo.
- *Estudios de terremotos:* Los epicentros del terremoto observados deben agruparse a lo largo de fallas continentales.
- *Uso de suelo:* Identificación de áreas de uso similar de la tierra en una base de datos de observación de la tierra.
- *Marketing:* Ayudar a los profesionales de marketing a descubrir distintos grupos en sus bases de clientes.
- *Planificación de la ciudad:* Identificación de grupos de casas según su tipo de casa, valor y ubicación geográfica.

➔ **Métodos de Agrupación:**

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

## 4.-Visualización de Datos

La visualización de datos 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:**
  Este es el tipo más común y conocido que utilizamos en nuestro día a día con las hojas de cálculo para representar datos de manera sencilla como gráficos circulares (o de pastel), líneas, columnas y barras aisladas o agrupadas, burbujas, áreas, diagramas de dispersión y mapas tipo árbol.
- **Mapas:**
  Con la popularización de Google Maps y su conocida API todos conocemos la visualización de datos en mapas para conocer, por ejemplo, la localización de nuestra flota de vehículos en tiempo real.
- **Infografías:**
  Una infografía es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente. Son excelentes para ayudarnos a procesar más fácil la información compleja.
- **Cuadros de Mando (Dashboards):**
  En el entorno empresarial, un cuadro de mando es una herramienta que permite saber en todo momento el estado de los indicadores de negocios como los de ventas, económicos, de producción, de recursos humanos, etc.

## 5.-Reglas de Asociación.

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

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

➔ El 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

➔ Reglas de asociación: principio a priori:

Principio a priori: Si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes. El principio de A priori se mantiene debido a la siguiente propiedad de la medida de soporte:

$$\forall X, Y: (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

El soporte es el conjunto de elementos nunca excede de sus subconjuntos. Esto se conoce como la propiedad anti-monótona de soporte.

## 6.-Clasificación.

Tareas predictivas: Predecir un valor de un atributo en particular basándose en los datos recolectados de otros atributos. Dentro de las tareas predictivas tenemos:

- la predicción,
- los patrones secuenciales,
- la regresión y
- la clasificación.

La clasificación es una técnica de la minería de datos, es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

➔ **Métodos**
  - **Análisis discriminante:** método utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos.
  - **Arboles de decisión**: método analítico que a través de una representación esquemática facilita la toma de decisiones.
  - **Reglas de clasificación**: buscan términos no clasificados de forma periódica, si se encuentra una coincidencia se agrega a los datos de clasificación.
  - **Redes neuronales artificiales**: es un modelo de unidades conectadas para transmitir señales.

➔ **Características de los métodos:**
  - Precisión en la predicción.
  - Eficiencia.
  - Robustez.
  - Estabilidad.
  - Interpretabilidad.

## 7.-Patrones Secuenciales.

- Conceptos:

- Minería de Datos 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 acontecimiento es considerado.
- Reglas de asociación secuencial: 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.

➔ Ventajas y Desventajas

- Ventajas:
  - Flexibilidad.
  - Eficiencia.
- Desventajas:
  - Utilización.
  - Sesgado por los primeros patrones.

➔ Aplicaciones:

| Tipo de Dato | Aplicación | Ejemplo | |
|---|---|---|---|
| ADN y Proteínas | Medicina | Predecir si un compuesto químico causa cáncer | Agrupamiento de patrones secuenciales |
| Recorrido de clientes en supermercado | Análisis de Mercado | Comportamiento de compras | Agrupamiento de patrones secuenciales |
| Registros de accesos a una página web | Web | Reconocimiento de spam de un correo electrónico | Clasificación con datos secuenciales |

## 8.-Predicción.

La predicción es una técnica que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir el resultado 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.

Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo.

➔ Relación con otras técnicas

Cualquiera de las técnicas usadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido, junto con los datos históricos de esos ejemplos.

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

➔ Técnicas

La mayoría de las técnicas de predicción se basan en modelos matemáticos:

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

➔ Tipos de métodos de regresión
- Regresión lineal
- Regresión lineal multivariante
- Regresión no lineal
- Regresión no lineal multivariante

➔ Redes Neuronales.

Utiliza para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión.

Este proceso se conoce como entrenamiento de la red neuronal. Las redes neuronales de tres capas: de entrada, oculta y de salida.