UNIVERSIDAD AUTONOMA DE NUEVO LEON

“Facultad de Ciencias Físico Matemáticas”

MINERIA DE DATOS

RESUMENES

Kevin Franco González González 1805425

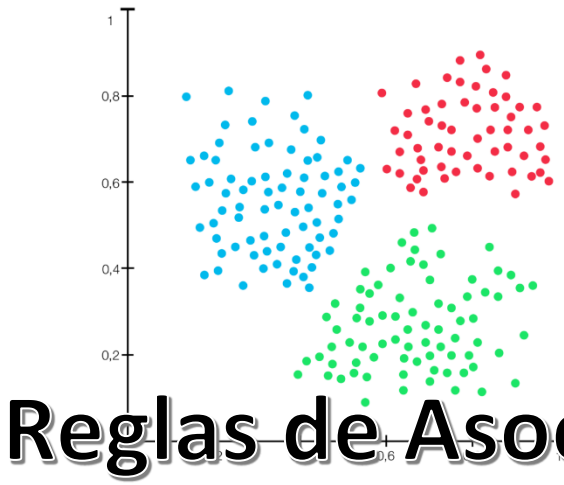
Descripción de Técnicas

DESCRIPTIVAS

Clustering

Este primeramente consiste en formar grupos que estén dentro de la base de datos, con cada uno de ellos se busca contener una cercanía relativa dentro del área de trabajo para ello se utilizan distintos métodos con el fin de poder identificar la cantidad adecuada de centroides y/o clústeres, parecido al método elbow, en donde dentro del mismo procedimiento que se realiza se utiliza el proceso de K medias.

Se puede emplear este tipo de técnica haciendo una investigación de mercado para poder realizar la segmentación de clientes que por su definición lo dice es la delimitación de unidades, es el resultado de la segmentación de la muestra. Si nos ponemos a analizar las métricas que hay en los clientes históricos podemos a los diferentes grupos que tenemos definirlos y a su vez con ello poder mejorar las áreas que se encuentren en el proceso que se está realizando.



Reglas de Asociación

La regla de asociación como su nombre lo dice asociar, esta se encarga de buscar para poder analizar relaciones o patrones relevantes que surgen investigando a fondo entre los datos, obteniendo la información recaudada podemos predecir los patrones estudiados que sucederán a continuación después de que un elemento aleatorio es añadido a la base de datos pertenecida.

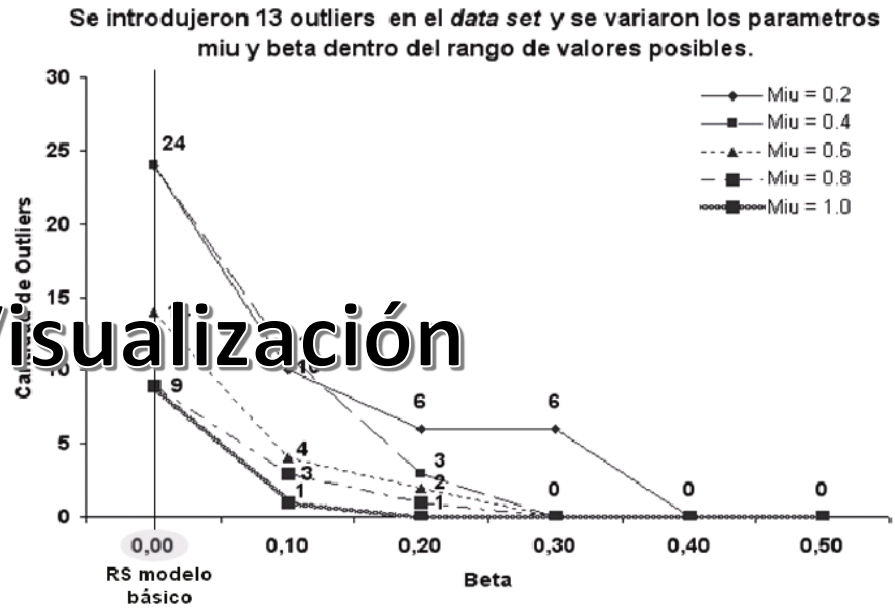
Un claro ejemplo es Un sistema de detección de intrusiones:
Se ha utilizado para la detección de intrusiones, estudiando los patrones de mal uso en la seguridad de la información, encontrando patrones de acceso a los recursos, procesando los registros de ataques a la red, para así descubrir comportamientos secuenciales de intrusión y diseñar estrategias para la detección de varias etapas de ataque.
Un principio muy importante utilizado es el principio de a priori donde este menciona que si un conjunto de elementos es frecuente entonces todos sus subconjuntos también deben de ser frecuentes se mantiene debido a la propiedad de medida de soporte donde esta nos dice que el soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos, a esto se le conoce como la propiedad anti. Monótona de soporte.

Detección de Outliers

“Un outlier es una observación que se desvía tanto de otras observaciones que despierta la sospecha de haber sido generado por un mecanismo diferente” [Hawkings, 1980].

Con el objetivo de proponer un algoritmo que permita identificar eficaz y eficientemente outliers en grandes bases de datos se seleccionó la aproximación por celdas propuesta por Edwin Knorr y Raymond NG en 1998 en el trabajo “Algorithms for Mining Distance-Based Outliers in Large Datasets” [Knorr y otros, 1998]. Este método puede procesar de forma muy eficiente hasta 4 dimensiones (5 en algunos casos) pero luego decrece su rendimiento e incluso puede imposibilitarse su ejecución. Consiste en analizar factores que sobresalen dentro de la información de los datos, todos estos datos que son lejanos a una media se convierten en un outlier, con lo que se generan un sesgo con estos dentro de la información.

Poder detectarlos es conveniente para poder identificar algún error dentro de la información que merezca ser analizada con mayor detalle dentro de la información.

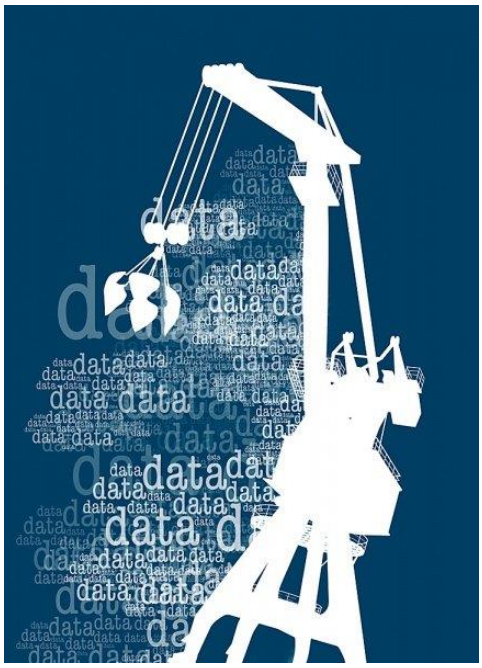


Visualización

Regresión

Esta consiste en involucrar todas las técnicas que se mencionen en la misma visualización ya que es un apoyo casi indispensable para analizar completamente la información y así mismo esto nos permita crear mayor conciencia del comportamiento de datos.

Es importante conocer diferentes tipos de visualización de datos, ya que uno de los grandes retos que enfrentan los usuarios de empresas es cuál de los tipos de visuales se deben utilizar para representar la información de la mejor forma, los cuales se dividen en gráficos infografías, cuadros de mando y mapas. También hay aplicaciones que podemos comprender de la visualización de datos que son comprender la información con rapidez identificar relaciones y patrones e identificar tendencias emergentes. Cuando hacemos énfasis en cualquier empleo la importancia de la visualización es considerable ya que los conjuntos de habilidades están cambiando para adaptarse a que un mundo este basado en los datos como lo menciona la presentación. Cada vez es mas valioso usar los datos a la hora de tomar decisiones, realizar un análisis ya que con una base solida datos y una buena visualización obtienes un mejor resultado.

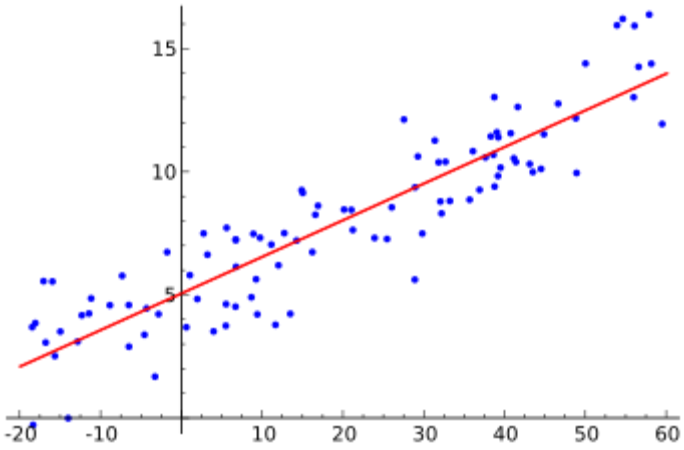


PREDICTIVAS

|

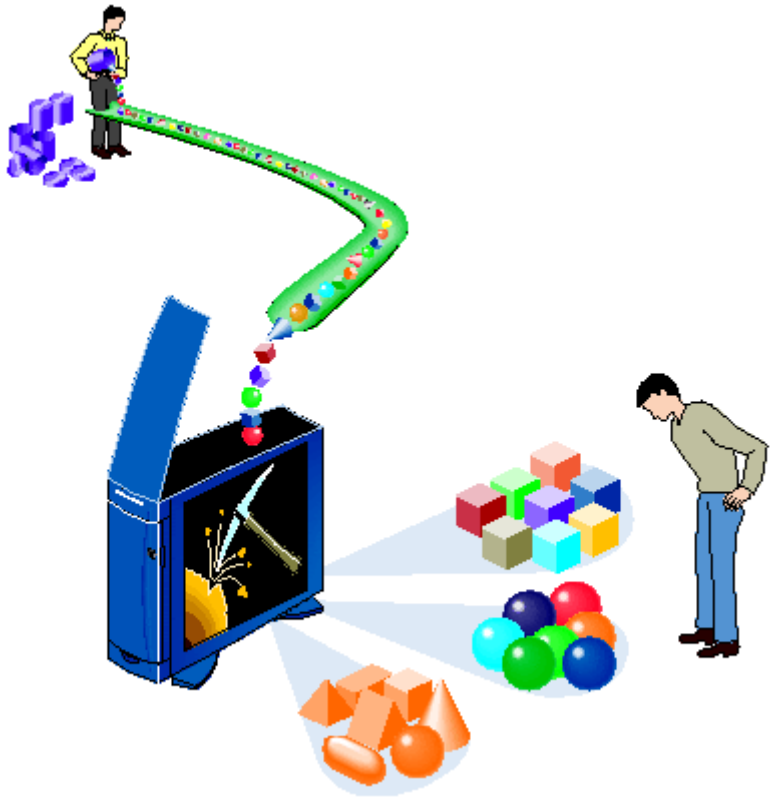
La regresión es un modelo matemático que nos ayuda a determinar la correlación entre una o más variables, esta se divide en dos que la primera viene siendo la regresión lineal donde en ella se comparan entre si 2 valores, una variable aleatoria independiente que ejerce influencia sobre la otra variable dependiente. Otra de ellas es la regresión múltiple que por su nombre lo dice nos da entender rápidamente que es a diferencia de la lineal, en esta se tiene dos o más variables independientes y dependiente.

Al realizar el análisis de regresión para visualizar el comportamiento de la grafica de probabilidad normal contra la de ajustes , el histograma y el de orden , nos hablo sobre las diversas funciones que debemos de utilizar para poder concluir que tan bueno fue el ajuste que nosotros realizamos, o si el modelo de regresión lineal es bueno.



Clasificación Patrón secuencia

La clasificación es una técnica que en minería de datos es empleada frecuentemente se aplica esta mapea u organiza un conjunto de atributos por clase dependiendo de sus características. Nos puede servir para hacer predicciones futuras en alguna clasificación de datos ya clasificados. existen varios tipos de técnicas de clasificación como la clasificación por inducción de árbol de decisión la clasificación bayesiana, de redes neuronales, basada en asociaciones, entre otras.

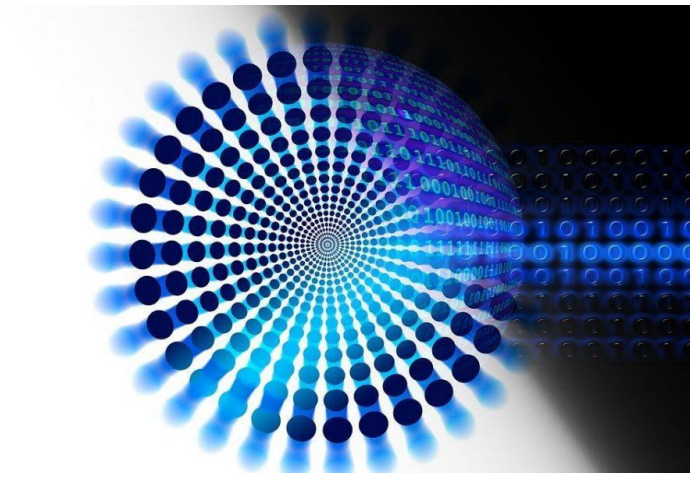


Predicción

En los patrones de secuenciales sucede que se extraen los patrones frecuentes relacionados con el tiempo u otro tipo de secuencia.

Las características que los patrones de secuencia tienen es que no importa el orden por que se enfoca principalmente en encontrar esa secuencia que está sucediendo en los patrones que se van analizando.

El tamaño de una secuencia es su cantidad de elementos que tiene, su longitud es la cantidad de los ítems que están, entre otros. Las desventajas es que los primeros valores pueden sesgarse por la forma en la que se acomodan. a su vez, tienen de ventaja su eficiencia y flexibilidad ya que no necesitan estar corriendo los datos constantemente, con una vez realizarlo es suficiente comúnmente.



En la predicción primeramente tenemos que definir el problema, objetivo o salida deseada que se desea obtener, una vez realizando ello tenemos que recopilar los datos obtenidos, después procederemos a elegir la medida o indicador de éxito.

En la predicción se utiliza un modelo llamado modelo predictivo que consiste en dividir el espacio de los predictores para agrupar las observaciones con valores similares para la variable dependiente o de respuesta. El espacio muestral se divide en subregiones aplicando una serie de reglas o decisiones para que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones. En su estructura tenemos el primer nodo o nodo raíz, nodos internos o intermedios y los nodos terminales u hojas. Existe el árbol de regresión que se utiliza en la predicción también para que las observaciones queden dentro de un hiper rectángulo y tengan el mismo valor estimado.