Minería de datos:

Se pueden tener muchos enfoques, pero desde lo general, es un conjunto de métodos para procesar y analizar una colección de datos ya sean grandes o pequeñas (pueden existir problemas técnicos con los conjuntos grandes “big data” ejemplo el almacenaje, procesamiento etc), para poder encontrar asociaciones, patrones o cualquier hallazgo interesante que explique los datos, ejemplo lo que hace Amazon o Google, o tiendas departamentales, que mediante un análisis explican los datos mediante asociaciones, y los organiza y los explica para poder hacer algo con ellos, ejemplo las personas adultas compran por la mañana y preparar algo para ese público, aunque la toma de decisiones ya no le compete mucho a la minería, esas decisiones las toma una persona o un sistema.

Un patrón lo podemos definir como una regularidad ósea que se repite, es discernible y predecible, los patrones no son fáciles de descubrir, para eso se ayuda de la minería, puede haber patrones naturales como la noche y el día, o distinguir el diseño en la una cortina o alfombra, aunque también existe la aleatoriedad, pero cuando pasa eso se puede encontrar una tendencia, ejemplo no todos los adultos mayores compran en el día, pero la tendencia dice que la mayoría lo harán

Lo principal es extraer información en forma de patrones o resúmenes o hallazgos de una colección de datos y transfórmalo en una estructura entendible, para que al final una persona o sistema automático pueda tomar decisiones. ejemplos los banco al asignar un crédito.

El concepto nace entre los 80s y 90s aunque los términos tienen un auge que cambia con el tiempo a otros nombres como pueden ser, analítica de datos, ciencia de datos, big data (aunque no es lo mismo), ingeniería de datos, extracción de conocimiento, inteligencia de negocios

Ejemplo dentro de los datos se minan los patrones o conocimiento a partir de una colección de datos

Ejemplos de minería de datos que podemos encontrar o cosas que podemos descubrir en minería de datos

Clusters (ramificaciones o grupos): son agrupamientos, es tratar de encontrar en que se parece un elemento con otro dentro de un grupo, edad, géneros, etc, se mide la similitud entre dos elementos de información y podemos hacer un subgrupo ejemplo patrones de compra, afiliaciones, y estos subgrupos se pueden asocias a otras(difusos), dependiendo de la variable, pero también existen que un elemento solo pertenezca a uno como en el de la edad solo puede tener 1

Anomalías: Detectar algo que sale de la normalidad que estamos observando, un ejemplo detección de fraudes, en bancos o en patrones de conducta o como escribe una persona, no todas las anomalías son malas, un ejemplo cuando en Gmail ingresamos en otro

dispositivo, o en medicina mediante análisis imágenes médicas como el cáncer, es la detección de algo que no debería estar ahí, pero lo esta

Reglas: reglas de asociación, aunque pueden ser de muchos tipos, como en las compras, por decir si una persona compra herramienta, le salen anuncios de ello en la web, esto se puede ver mucho en los clusters, o están muy relacionados, estas asociaciones nos ayudan a establecer reglas

La minería de datos es multidisciplinar para poder dar resolución a los problemas que se le presentan

Estadística: es un pilar fundamental en conjunto con el modelado matemático, modelos que ya existen, mediante ecuaciones, como medias, medianas, cuartiles, correlaciones etc. Nos ayuda a analizar y entender los datos.

Aprendizaje de maquina: inteligencia artificial, es construir un sistema inteligente, mediante reglas, imitando la inteligencia de un experto en el tema, el aprendizaje de maquina es un colección de métodos van a aprender a imitar el comportamiento del experto mediante datos, anteriormente se construían mediante réplicas de reglas que daba un experto y se codificaban teniendo así un árbol de decisiones muy extenso con el aprendizaje que maquina toma un conjunto de datos los procesa y crea el conocimiento para hacer una regla.

Visualización: visualización de la información, como presentamos la información, mediante gráficas, imágenes, tablas, para que la información mostrada a los clientes sea más agradable y no tan cruda, un ejemplo que quien te compra más, quien debe más, en la actualidad todo es visual, de una forma en que una imagen o graficas se puede dar toda la información necesaria para alguna toma de decisiones (comunicación visual) o hasta puede ser mediante infografía

base de datos: administración y gestión de las ciencias de la información, como los accedo, donde los guardo, como los recolecto, de forma estructurado, por ejemplo, en tablas y consultas sql de pendiendo de sistema de manejo y almacenaje que se tenga para la información

¿porque minería de datos?

Se vive en la era del big data, estos se distribuyen y de comparte, diario de crean, estos son muchos, en todos partes los tenemos estos grandes volúmenes de datos, sabemos que tenemos cosas así y queremos descubrirlas.

¿Qué tan grande es el big data? Son 2000 tb de datos creados en internet en un minuto, 3 petabytes de datos están en la base de datos Google

¿Qué es la dataficacion? Es la tendencia tecnológica de transformar muchos aspectos de nuestra vida en datos, como por decir todas las aps que tenemos en nuestro celular, todo

lo que subimos a redes sociales, todo lo que compramos en e-commerce, esto está muy relacionado con el internet de las cosas como Alexa, todos estos aparatos siempre están generando datos que nos devuelven información útil, estos datos de digitalizan se comparten y nos ayudan a tomar decisiones mediante otros aparatos o nos los presentan para nosotros tomarlas

Proceso general de la minería de datos:

Definición del problema: que quiero saber o que quiero hacer, dependiendo del problema, se generaran los datos que se necesitan de acuerdo con el contexto y relevancia por lo regular este conjunto de datos los da el cliente (o elemento externo) dependiendo de sus necesidades (privado o público)

Recopilación de datos:

Depende del contexto, elementos a tener en cuenta, son el acceso que nos da el elemento externo a la información (permisos, cobros, regulaciones, contratos u otra consideración), otra cosa es el muestreo de comportamientos de la información, se tomara toda la data o solo tomar una parte de la información, otro elemento es el guardado de la información, ejemplo peso, formatos, permisos

Preprocesado de datos: limpieza, quietar cosas que no deberían estar ahí, limpieza de registros nulos, equivocados, errores o ruido en la información. Normalización, llevar todos los datos en un atributo a la misma escala por decir todos cm o todos metros, kilos o gramos, grados centígrados u otros. Transformación, pasar los datos a una representación que podemos acceder mediante un lenguaje de datos y este debe de tener una estructura de datos para trabajar con ellos como por ejemplo matrices arreglos, listas, diccionarios, tuplas, conjuntos

Extracción de patrones:

Proceso de análisis, métodos para tratar de encontrar dichos patrones o las relaciones, o las reglas o los conjuntos, anomalías, depende de lo que apliquemos nos dice que proceso de transformación debemos de hacer con los datos

Despliegue del conocimiento:

todas las relaciones o hallazgos encontrados, mediante reportes, graficas, tablas, o pasar estas a una aplicación externa para toma de decisiones, otros sistemas que se alimentan o alimentan a otros para la visualización de la información

Métodos en la minería de datos:

Descriptivo:

Clusters se describirán los datos que se tienen en términos de conjuntos de datos y los vamos a agrupar en un numero reducido de grupos de acuerdo a sus similitudes en sus características se puede hacer mediante tablas

Modelado estadístico se encargará de resumir los datos en unas cuantas características relevantes así ayudándonos para poder hacer los clusters. Por ejemplos en lugar de presentar una serie de datos mediante tablas, se puede calcular una serie de estadísticas que nos ayudara a resumir todos los datos en unas cuantas características que los describan de manera mas simple

Predictivos:

Funciones que predigan una variable basados en una función como una ecuación, se le pasan una serie de parámetros y tratara de darme el valor de la función, si compre cereal y leche me recomendara un tazón

Reglas jerárquicas o asociaciones, se comportan como un diagrama de flujo para toma de decisiones, dictamos una serie de reglas que deben de irse siguiendo de acuerdo con una asociación un ejemplo si se compra un bote de basura también se compran bolsas para basura para el contenedor o lo que hace apple

Con las funciones nos dan asociaciones conforme a los valores dados, y al revés con las asociaciones nos dan valores buscados

Modalidades de los datos o información

Se está haciendo referencia a como la información de representa, una es la forma estructurada como tablas con sus atributos que describen a una entidad, como en las bases de datos relacionales, con una estructura predefinida, donde sabemos qué tipo de dato se está extrayendo de la entidad. Otra forma son los datos no estructurados, cada entidad tiene su propia estructura (o en desorden) tales como texto o imágenes o algún otro.

La información estructurada, la información se separa del contenido, cada observación es una instancia de la estructura predefinida, la estructura es la colección de atributos o variables de cada entidad de interés, por ejemplo cada estudiante es una instancia u observación, cada columna representa un tipo diferente de variable o atributo (nominal, de rango, ordinales), generalmente es fácil ingresar manipular, almacenar y de leer, es fácil hacer consultas y procesar la información, también se le puede hacer procesos estadísticos, como encontrar valores máximos o mínimos, sumatorias, moda, mediana, entropía, probabilidad, variancia, media, y aplicar procesos para su graficas ion(graficas de pay, de puntos, grafos, de calor, de barras) para la visualización de la información,

Esto tiene varios propósitos, una es encontrar asociaciones entre variables, en ejemplo ventas a lo largo día de dos productos, otra cosa que podemos hacer con la información estructurada se pueden hacer predicciones de acuerdo con como estén actuando, un

ejemplo los precios de la gasolina a lo largo del tiempo por medio de modelos estadísticos y sus variantes

Procesos de los datos estructurados: estadística descriptiva.

Máximos, mínimos, sumatorias, modas, medianas, variancias, desviación estándar entropía, valores únicos, probabilidad, graficas para la visualización

Las tareas que podemos tener en los datos estructurados

Asociación de variables por ejemplo como se comporta un par de variables mediante un cierto tiempo teniendo así una asociación entre ellas, todo lo podemos visualizar mediante una gráfica de barras dobles.

Predicción se pueden hacer estudio para predicciones mediante un modelo estadístico para encontrar tendencias o patrones por igual se pueden hacer para una o un par de variables mediante el paso del tiempo, teniendo en cuenta que la primera parte de los datos son recopilados y la segunda parte de los datos a representar son los datos de predicción

Detalles Issues

que se pueden presentar con la información estructurada, cuando se tienen muchas varíales, estos puede contraer muchas dificultades al procesarlas, se debe de tener solo la información importante para la resolución del problema, se puede tener información incompleta un ejemplo como los valores nulos en sql, puede que los datos sean incoherentes o que causen ruido en la data puesto que no estamos seguros de que el dato sea real, ejemplo estamos leyendo datos de estaturas y nos aparece el nombre de la entidad, los outliers son valores alejados de lo que esperamos ver con estos datos se tiene que decidir que hacer o si son importantes para el análisis.

información textual

Es un conjunto coherente escrito con signos que transmite un mensaje informativo, los signos con el alfabeto que se use, no es lo mismo las que varían del origen, en el latín que los de medio oriente, este alfabeto debe de ser coherente para su traducción y transmisión del mensaje, esto puede tener un mensaje crudo o en su parte cruda solo es un conjunto de letras separas o agrupadas por signos por ejemplo lo que hace Google para las búsquedas por palabras clave, en el significado semántico es lo que el conjunto de símbolos nos quiere decir , por ejemplo mama es progenitora, un ejemplo es Alexa que de un audio lo transforma a texto para saber su significado, la información textual la podemos encontrar, en emails, redes sociales, documentos generales, páginas web, reportes médicos, libros, etc,. El texto es más fácil y menos costoso para su envío y almacenamiento y producir

Procesos que se pueden hacer son filtrados de palabras para la limpieza del texto y siga siendo inteligible, se pueden hacen frecuencia de palabras para saber la importancia del texto, la transformación de texto es para pasar del significado crudo para una matriz, por

ejemplo, cuando se hace una frecuencia de palabras, pasando de un data sin estructura a una con estructura para no trabajar con strings a pasar a trabajar con matrices para su análisis

Tareas con información textual, podemos clasificación, por ejemplo, como en Gmail al clasificar nuestra correspondencia, o en los nuevos mensajes de texto que nos dicen que pueden llegar a ser anuncios o maliciosos, un ejemplo se puede ver en la clasificación de argumento u opinión de algún artículo en venta si la opinión es buena mala o neutral por medio de palabras clave, también se puede detectar un plagio en textos, dadas las palabras utilizadas en el mismo,

Detalles que podemos encontrar como dificultades, podemos tener un gran número de palabras en el texto, otra cosa que podemos encontrar la polisemia (la misma palabra tiene varios significados, jaguar carro, jaguar animal) y la sinonimia (varias palabras tienen el mismo significado), y el idioma nos puede traer problemas por sus reglas y su alfabeto (nos dan restricciones) debemos de especificar el idioma a analizar, un ejemplo es el Spanglish

Existen diferentes modalidades de observación, por ejemplo, un artículo podemos mostrar su información en la tabla donde diga el modelo, nombre, garantía, etc, pero también se puede hacer por medio de un reporte complementando la información como extensión de esta

Análisis de datos estructurados

Observaciones, es la información recolectada acerca del objeto de interés dentro de un contexto, como una persona con contexto puede ser médico, administrativo, social y de ello de extraerán las variables, por ejemplo si es médico se tomaran edad, enfermedades, peso, vacunas, de administrativo sus ventas o sus compras, por periodos de tiempo, en proceso de observación intervine el observador puede ser un ser humano o un sistema que este monitoreando por medio de sensores, el observador, es aquello que está recopilando la información del objeto observado y este no debe de intervenir, no se debe de tener parcialidad que es la inclinación de presentar y mantener una perspectiva parcial de la observación que implica la falta de un punto de vista neutral

Después de encontrar los atributos los llamamos variables, estas son el registro de la medida de las observaciones con un valor específico, uno decide que eventos tiene que observar dependiendo del contexto, las variables nos dicen que tipo de valor debo de ir guardando para cada observación

En estadística tendremos 3 tipos de variables.

Nominales describen una categoría o cualitativo, describe un atributo por ejemplo sexo, color, tipo de chocolate se dice que nombra algo que describe una categoría predefinida que describe la propiedad, describe una cualidad, generalmente se nombran con etiquetas o nombres aunque también se puede hacer por índices o códigos, como por decir un código

para un color, 1 = rojo, 2=amarillo, donde el número no ordena nada o no importa su relevancia, por decir las ladas de teléfono que son una etiqueta numérica para una zona, o se puede mostrar como un catalogo

Ordinales, son aquellas donde el orden importa, por ejemplo, un rango militar o un nivel de satisfacción al ver una película, aunque existe un orden tal vez no se conoce la escala de este, o sus proporciones entre sí, también se puede poner nombres o etiquetas con códigos o índices, pero aquí el orden si importa, 1 bueno, 2 regular, 3 malo

Con las variables nominales y ordinales se pueden estimar frecuencias y porcentaje, diciéndonos la escala de valor de las etiquetas estudiadas o saber tendencias, por ejemplo en saber si a alguien le gusta un color o están satisfechos con una compra, para graficar datos nominales de puede ser una gráfica de barras o de pay para conocer la tendencia del atributo, para las variables ordinales se puede hacer por grafica de barras porque se tiene un orden de variable y en la barra de puede ver la frecuencia del valor del atributo

Para variables ordinales algunas veces se puede calcular la media, pero no es recomendable, un ejemplo es cual es la media de un nivel de satisfacción

Intervalo o razón, un ejemplo es la edad porque si se puede saber la escala del mismo, son variables numéricas y tienen que ver con la cantidad numérica que están midiendo, representan un atributo físico o cantidad, como años, altura o peso, área, salarios,

Para poder trabajar con estos datos primero se grafican como en una gráfica de dispersión, en el eje x se pone la muestra y en él Y se pone la edad y donde se atraviesan se pone un punto por cada observación, esto ayuda la representación, como segundo paso los datos se ordenan y se vuelven a graficar en una gráfica de dispersión ordenada, así podemos obtener valores máximos y mínimos, lo siguiente es encontrar otras métricas de la data, una de ellas es la mediana, esta métrica divide nuestros datos en dos, si en la gráfica ponemos una línea en la mitad podemos encontrar la mediana, también si trazamos a un cuarto entre máximo y mínimo, y como complemento podemos trazar otra línea a los ¾ de la mínimo y máximo, también se le conoce como percentiles, el tercer cuartil se le conoce como percentil 75 y el otro a percentil 25 y la mediana es conocida como percentil 50, esto indica que tendremos un porcentaje de los datos hacia debajo de la gráfica, con esto se puede conseguir el resumen de los 5 números que está compuesto por mínimo, máximo, primer cuartil, mediana y tercer cuartil, y con esos 5 números se obtiene un diagrama de caja o un boxplot, máximo y mínimo se conocen como whiskers o bigotes de la caja.

ejemplo se tiene un conjunto de datos y se tiene que sacar el resumen de los 5 números, primero se ordenan los datos de menor a mayor, obteniendo así el máximo y el mínimo, para obtener la mediana es el valor que se encuentra la mitad, para buscarlo de manera de programación se hace con la formula p = (NumElem – 1) * 0.5 = x, el número que representa x es la posición en el ordenamiento empezando a contar desde 0, después de obtienen la posición superior y la posición inferior (floor y ceil), después de esto se aplica para el primer

cuartil pero en lugar de usar el percentil 50 o 0.5 de hará para el 0.25 y 0.75, obteniendo los 5 números podemos trazar la caja

para cuando existen medianas donde el cálculo nos arroja decimales, se toman las posiciones de floor y ceil (pl y pu) y la media se calcularía, mediana= pl+(pu-pl)*0.5, el 0.5 representa el valor del percentil 50, para los demás es igual solo cambia el valor del percentil 0.25 y 0.75

boxplot modificada, después de ordenar los datos se observan valores en el máximo o mínimo un tanto alejados de los de más números se dice que estos valores son outliers (valores atípicos) y para saberlo la estadística nos ayuda, primero se mide el rango Inter cuartil (IQR) es la diferencia entre tercer cuartil y el primer cuartil, después de buscan los valores de las inner fences o (vallas) superior e inferior, para la upper inner fence = 3er cuartil +1.5(IQR) y para lower inner fence = 1er cuartil -1.5(IQR), tomando en cuentas las vallas podemos encontrar el upper whisker= dato más cercano igual o menor a la valla interior superior, y para el lower whisker = dato más cercano igual o mayor a la valla interior inferior, los puntos que quedan fuera de los whiskers son los outlier, los outliers pueden ser objetos de atención y vale la pena estudiarlos y saber por qué, puesto que pueden ser errores o si las observaciones tienen algo en particular o si son ruido de captura en la información y después de decide si se dejan o se eliminan, si de dejan se puede distorsionar algunas métricas

la mediana es una métrica de centralidad para perfilar los demás valores.

Centro de datos o media o promedio

El promedio también se le conoce en la estadística como valor esperado o centro aritmético, es la sumatoria de los valores u observaciones entre el número de observaciones

La media no es una estadista robusta porque no es resistente a los valores extremos de las observaciones, se ve modificada por casos de outliers, por el contrario, la mediana si es una estadística robusta, La estadística robusta es una aproximación alternativa a los métodos estadísticos clásicos. El objeto es producir estimadores que no sean afectados por variaciones pequeñas respecto a las hipótesis de los modelos.

Cuando queramos representar un valor típico de los datos vamos a preferir la mediana porque no se ve afectada

Para que la media sea más afectiva se utiliza la media cortada o versión modificada que consiste en eliminar los k valores más largos de los datos y los k valores menores de los datos (k=(a/100)*n) donde n es el número de datos y a el porcentaje que queremos cortar la media y nos dice cuántos datos debemos borrar, si el número de k es igual a decimales entonces se toma el entero menor por lo tanto la media cortada es más robusta que la media, se debe de estar probando que tanto se debe recortar la media, hasta que no varie, nos mide la centralidad de los datos,

Dispersión de los datos es que tan alejados están nuestros datos del centro, para medir la dispersión, una forma es el rango = máximo – mínimo este rango que contienen todos los datos, otra forma es con el rango Inter cuartil IQR 3er quartil – 1er quartil, la mitad de los datos deben de estar en este rango, si la dispersión es pequeña va hacer más probable que yo observe un dato alrededor del centro y al contrario, si el rango es alto y tienen menos puntos, mayor es la dispersión, Estas medidas no consideran todos los valores de los datos, solo algunos valores resumidos. Para solucionar estos problemas se utiliza la desviación estándar o la raíz cuadrada de la varianza, si la varianza es pequeña, la desviación estándar será pequeña, ósea que los datos no están tan lejos del centro.

Mediana y el rango Inter cuartil son robustas

La media, el rango y la desviación estándar no son robustas

Cuando se presenta un informe estadístico se eligen las métricas robustas, si se utiliza la media se debe utilizar junto a la desviación estándar para explicar las variaciones que se pudieran dar respecto a la media.

Histogramas nos ayuda a entender la forma de los datos, está ligada a la distribución de los datos, es una gráfica de barras de los posibles rangos donde yo quiero agrupar mis datos y en el eje y estarán las frecuencias de mis datos, los histogramas divide los datos en intervalos, estos son excluyentes un dato solo puede estar en uno, y es exhaustivos que incluye todos los datos la sumatoria de todos los datos agrupados en los bins debe de ser el total de los datos, para cada bin o intervalos para saber en qué bin entra cada datos se sabe:

Dato igual o mayor al límite bajo

Dato menor que el límite alto

Para construir el histograma, primero debemos saber el rango y en cuentos bins dividiremos el total de los datos luego se calculas los limites inferior y límite superior para cada bin, no se sabe un mejor número de bins en que agrupar, se recomienda ir viendo cómo se comportan los datos, puesto 5 bins producen un histograma y 6 bins producirán uno diferente, si se hacen varios histogramas nos ayuda a entender más los datos, se puede observar un pico de datos, es el rango donde más datos se encuentran, si un valor se observa que se repite en ese pico significa que es la moda de los datos, otra cosa observable es la cola izquierda y la cola derecha, con estas tres observaciones se puede clasificar las distribuciones en formas como:

Unimodal, simétrico solo existe un pico en medio del histograma

Sesgado a la derecha

Sesgado a la izquierda

Bimodal (2 picos)

Multimodal (varios picos)

Uniforme y simétrico la mayoría de los picos tienen la misma altura

Si en la gráfica en uno de los bins es cero, talver exista un outlier, aunque de un histograma puede tener varias formas como unimodal sesgado a la derecha con un posible outlier

Si compara un diagrama de caja junto con un histograma podemos ver las distribuciones de puntos y verificar si existen los outliers, aunque de manera numérica también se puede representar y saber hacia qué lado los datos están corridos hacia alguna dirección y el histograma tiene un sesgado

Dependiendo de donde este el botplox podemos saber o darnos una idea de cómo será nuestro histograma, un ejemplo si la caja es larga por lo regular el histograma es uniforme, si están hacia algún lado nos dice que se contiene un sesgo y si es centrada posiblemente sea unimodal y simétrica

En estadística se estudia más la distribución normal o de campana que es unimodal, es simétrica y no tiene outliers, a esta distribución se le puede aplicar lo que se conoce como regla empírica, esta nos dice que si yo encuentro la media y la desviación estándar:

68% de los datos se encuentran entre media- (1 sigma) y media + (1 sigma)

95% de los datos están entre media- (2 sigma) y media + (2 sigma)

99,7% de los datos se encuentran entre media- (3 sigma) y media + (3 sigma)

Sigma= desviación estándar

La distribución de los datos está ligada a la probabilidad, nos quiere decir la posibilidad de encontrar cierto número de datos de acuerdo con la media en una distribución normal, en el centro es mayor la posibilidad de encontrar un valor que en las colas

Métrica para encontrar la relación entre dos variables se llama la correlación, ahí varios tipos, la más común en estadística es coeficiente de correlación de Pearson, este mide la dependencia lineal que existe entre dos variables X y Y, el rango tomado de los valores están dentro de 1 y -1, donde 1 quiere decir que X y Y están totalmente linealmente correlacionados positivamente, 0 nos dice que no tiene correlación lineal y -1 dice que tiene una correlación lineal negativa, en otras palabras si es uno positivo el valor de una aumente la otra también lo hace, si es 0 cada una cambia de formas diferentes y si es -1 si una baja la otra sube. En una gráfica se traza una línea y se ve su comportamiento. Es una correlación lineal perfecta cuando todos los pintos en la línea se mueven y cuando no es 0 y si se acercan los puntos, pero no la tocan los valores dependiendo del comportamiento se dice que el valor oscila entre 1 y 0 dado el caso.

La correlación no implica una causalidad, esto quiere decir que X no es la causa de que, Y cambie o viceversa, solo dice que tiene una dependencia, a esto se le dice correlación es

spuria, que una no causa la otra, por ejemplo, las ventas videojuegos y los doctores graduados. Aunque las dos graficas se comporten aparentemente igual, no se existe una causa que ligue una con la otra, solo se comportan parecido o igual, cuando esto pasa puede que esto se explique con una tercera variable un ejemplo seria la variable de la población, si la población aumenta más gente juega y más gente tiene doctorado

Nota: trabaja parecido a la pendiente de la gráfica o se apoya de ella

Tabla de contingencia, (se mide frecuencia de intersección de valores) Muestra la distribución de frecuencia (multivariante) de variables, es una tabla que toma varias variables y se muestras múltiples variantes, por decir cuántos hombres o mujeres tienen una consola y en esta podemos leer toda la información, se cuentan para cada posible combinación de variables sus valores

Cuando existen variables continuas, una alternativa es discretizarlo en grupos, utilizando las etiquetas de un grupo podemos ir haciendo subgrupos el grupo principal teniendo como resultado una lista de grupos con sus datos de frecuencia y así construir la tabla de contingencia

Una de las métricas que se pueden hacer a una tabla de contingencia es el odds ratio (razón de probabilidad) que es Para dos variables binarias (X e Y), mide la razón de las probabilidades de X en presencia de Y y las probabilidades de X en ausencia de Y, solo para variables que su valores sean 1 o 0 como true o false, mujer o hombre esto se logra por medio de la formula OR =a los productos cruzados de la tabla divididos el primero por el segundo, siendo la primera la frecuencia (n1,1)*(n0,0) n = a la posición de la tabla, En caso de que uno (o más) celda (s) contiene un cero, agregar 0.5 a todas las celdas (Haldane Anscombe corrección)

Ya cuando se calcula Las variables son independientes si y solo si la relación es 1. Para una relación>1, las variables son positivamente asociado. Para una razón <1, las variables están asociados negativamente. Es parecido a la correlación lineal solo que el resulta se define si es mayor o menor a uno

Tabla de contingencia: coeficiente phi de Pearson a diferencia de odds ratio, los valores obtenidos si tienen un límite de intervalo

Medida de asociación para dos variables binarias, interpretado de manera similar al coeficiente de correlación de Pearson (- 1 a 1).

Tabla de contingencia: prueba de chi-squared de Pearson, esta métrica no está limitada a variables que binarias, pueden ser variables que tomen múltiples valores.

Determina si existe un valor estadísticamente significativo. diferencia entre las frecuencias esperadas y las frecuencias observadas en una o categorías de una tabla de contingencia.

k: Número de categorías

xi : Frecuencia observada para la categoría i

mi : Frecuencia esperada para la categoría

Hipótesis nula: el tipo del trabajo es independiente al barrio de residencia

Para cada celda de calcula la prueba

Chi square es un valor de probabilidad que sigue una distribución que es la distribución del mismo nombre

Grados de libertad = (número de filas-1) (número de columnas -1) = (3-1) (4-1) = 6

Tabla de contingencia: correlación biseral puntual

Igual que Pearson coeficiente de correlación, pero para una variable binaria y una variable continua, p.ej. consola y altura

Tabla de contingencia: probabilidad condicional

Mide la probabilidad de que ocurra un evento, dado que ya ha ocurrido otro evento. Para discreto valores:

$P\ A\ B$: Probability of A occurring given that B has occurred

$P(A \cap B)$: Probability of A and B occurring together

$P(B)$: Probability of B occurring

Si el resultado es 1 es positiva sino es negativa la correlación

Probabilidad: probabilidad condicional Calcule la probabilidad condicional de morir dado que unpersona es un hombre o una mujer, según los siguientes datos.

Análisis de datos no estructurados.

Dentro de las modalidades de datos se tiene el texto para aplicarle un análisis se entiende que dentro de la minería de datos se tiene un sobrenombre como minera de texto que tiene varias aplicaciones.

Palabras que sean poco frecuentes o muy frecuentes no nos aportan mucho de información del texto, al graficar en barras la frecuencia de palabras en un texto podemos ver un fenómeno donde pocas palabras tendrán frecuencias altas y muy rápidamente la pendiente de la gráfica cae muy rápida dejando a las demás pablaras con una frecuencia muy baja este efecto se le conoce como ley de Zipf.

Al momento de analizar texto nos centraremos en sustantivos y verbos que son las que nos darán la información que necesitamos a las palabras restantes se les conoce como palabras vacías o stopwords y estas deben de ser eliminadas de la lista de frecuencia de palabras

Al modelo de representar a un usuario o un género como colección de frecuencias de palabras, se le llama como Bolsa de Palabras (Bag-of_Words)

El índice de Jaccard.

Métrica de similitud

J(A, B) = |A ^ B| / |A V B|

J --> [0, 1] ==> 0 = Los conjuntos no se parecen

1 = Los conjuntos son idénticos

Vectorizacion, cada palabra se representa por medio de vectores por medio de documentos formando una base vectorial A[0,0,0]