UNIVERSIDAD
NACIONAL
DE COLOMBIA

Universidad Nacional de Colombia
Sede Medellín

Análisis de Relaciones entre variables

Profesora: Patricia Jaramillo A. Ph.D

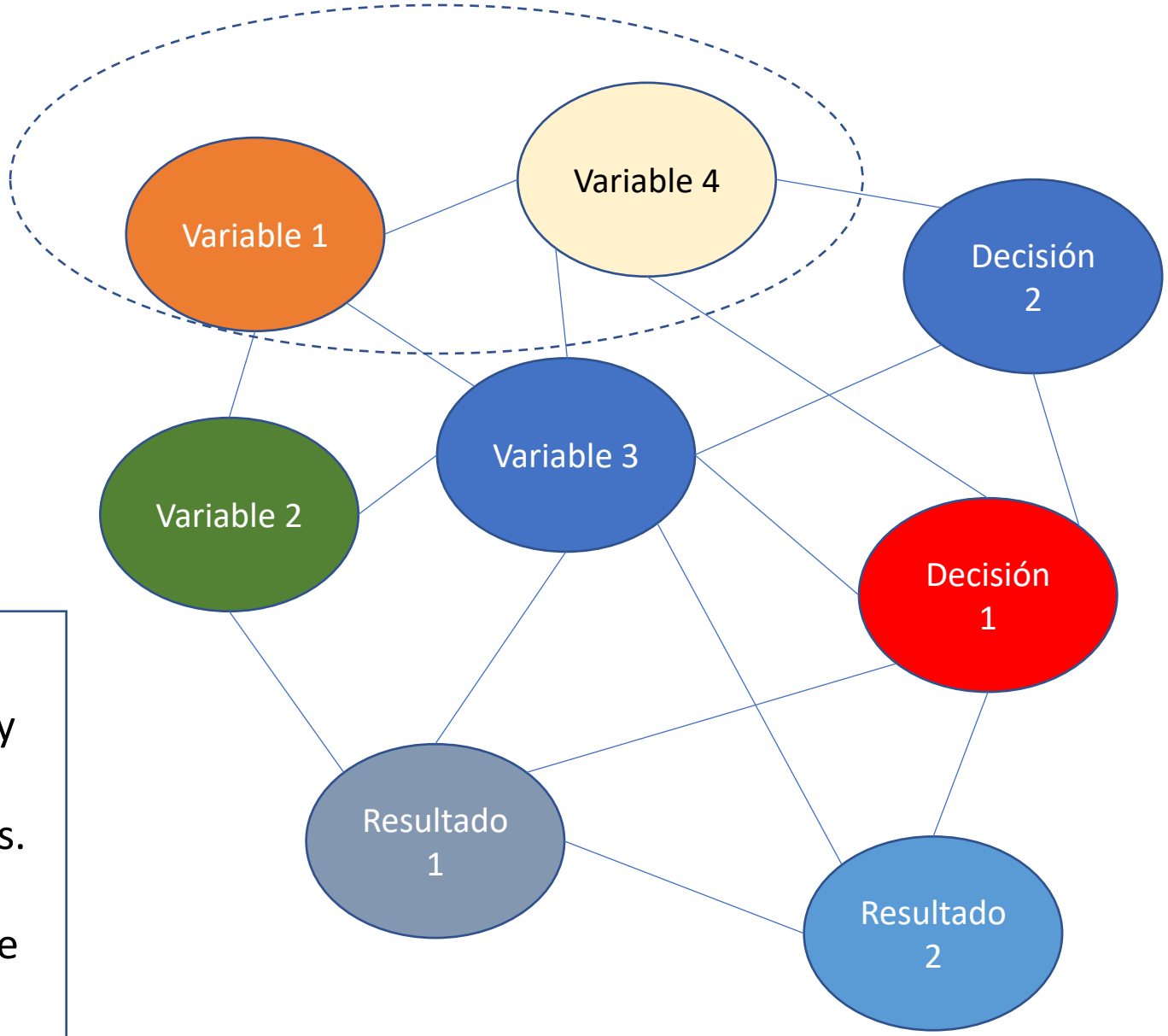
Sistema

Es un conjunto de componentes interrelacionados entre sí que poseen una estructura y propiedades específicas

Una tarea compleja es descubrir las interrelaciones entre componentes.

Por simplicidad, a veces, es posible modelar las interrelaciones entre subconjuntos de variables y este modelo es input de otros modelos de interrelaciones entre otros conjuntos de variables.

Ejemplo: modelar la interrelación entre un par de variables.



Modelación de las interrelaciones entre variables del sistema

Permiten:

- Predecir el valor de las variables de salida, ante un nuevo caso (X1,X2,X3)
- Ayudar al entendimiento de las relaciones (LO MAS IMPORTANTE)
- Incorporar estas relaciones en modelos mas complejos de decisión

Modelación de las interrelaciones entre variables del sistema

Pueden ser establecidas mediante (partir de análisis de datos):

- En lo posible, **relaciones físicas y comprensibles**
- Correlaciones
- Análisis de regresión
- Regresión logística
- Series de tiempo
- Redes neuronales
- Arboles de clasificación)
- FIS (Fuzzy Inference System)
- Etc.

Esta relaciones pueden ser estáticas o dinámicas

En lo posible, **relaciones físicas y comprensibles**

Ejemplo:

Lo cantidad de usuarios conectados en un periodo dado es:

$$I_t=I_{t-1}+F_t -V_t-P_t$$

Donde:

I_t = personas que quedan conectadas al final del periodo t (variable de salida)

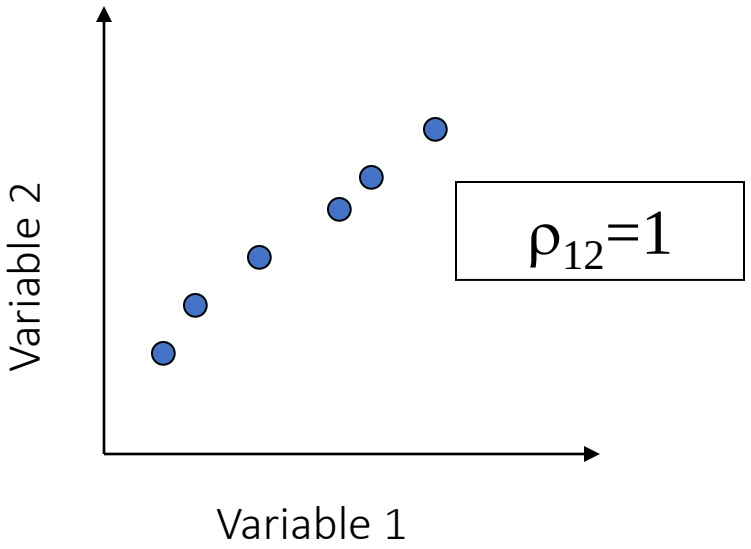
F_t= personas que se conectan en el periodo t (variable de entrada)

V_t= personas que se desconectan en el periodo t (variable de entrada)

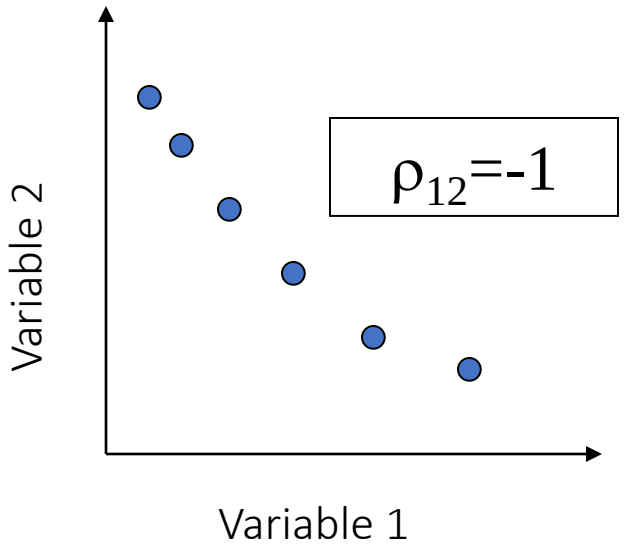
P_{t-1}= personas que se desconectan por algún problema de conexión (variable de entrada)

Pero…, lamentablemente, en muchos problemas no es posible establecer estas relaciones explicativas, acudimos a aprender de la información recolectada en el pasado.

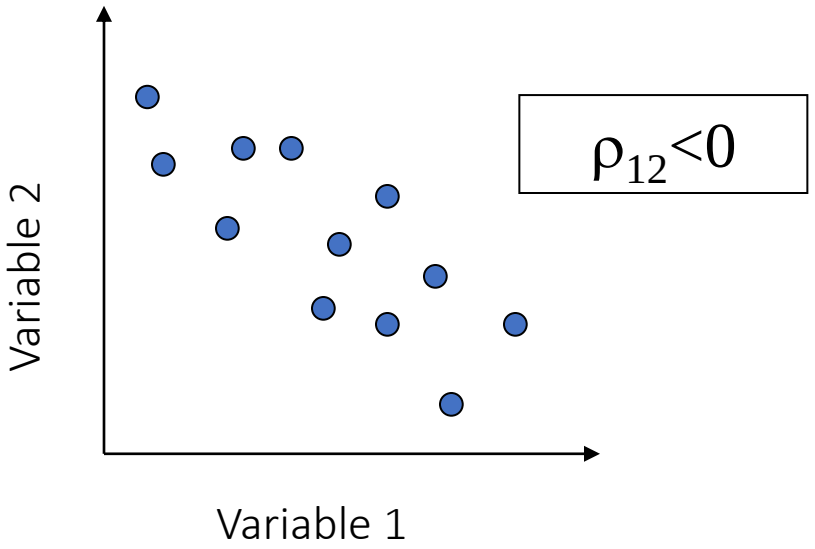
Correlación entre 2 variables



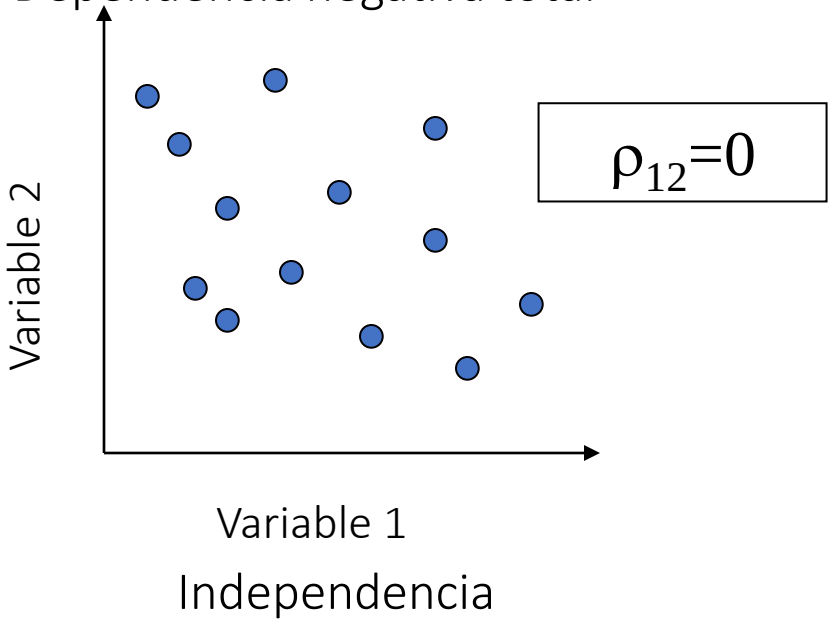
Dependencia positiva total



Dependencia negativa total



Dependencia negativa difusa



Independencia

OJO
No establecen relaciones causa-efecto

$$\rho_{12} = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{(x_{1i} - \bar{x}_1)}\sqrt{(x_{2i} - \bar{x}_2)}}$$

Concurso estime la correlación
http://guessthecorrelation.com/

Ejemplo: Redes logística

Una vez entrenado a partir de casos diagnosticados por médicos, permite:

- Predecir, para un nuevo paciente, la probabilidad de tener diabetes y de que tipo, si tiene valores de variables de entrada (X1,X2,X3).
- Ayudar al **entendimiento** de las relaciones para para toma de decisiones

Aprendizaje

Cuando las relaciones entre variables no son fácilmente identificables solo a partir de la capacidad cognitiva humana, pero se cuenta con suficiente datos de ellas, a partir de aprendizaje computacional, sus relaciones se pueden describir como funciones, reglas, etc

O puede que esas relaciones sean fáciles de identificar por los humanos pero son difíciles de especificar formalmente, ejemplo: reconocimiento del habla o del contenido en imágenes.

El aprendizaje permite que mediante modelos computacionales seamos capaces de enfrentar problemas complejos de decisión que podrían parecer subjetivas o no son triviales.

El aprendizaje permite:

- Identificar patrones, tendencias y relaciones entre datos recolectados, especialmente patrones no obvios. Por ejemplo, descubrir que las personas que compran leche descremada con alta frecuencia también compran pañales para adultos.
- Adaptarse a nuevas circunstancias.
- Crear nuevos comportamientos con base en patrones identificados.
- Hacer decisiones con base en los éxitos o fracasos de esos comportamientos

Aprender, matemáticamente, corresponde a ajustar una función desconocida Y= f(x) a partir de datos muestrales (Xi, Yi).

Puede ser a partir de:

- Modelos estadísticos: requieren de un modelo predefinido (lineal, polinomial, etc
- Redes Neuronales y Sistemas difusos

  Aproximadores universales de funciones

  Aprendizaje selectivo

Aprendizaje

Se divide en:

- **Supervisado**: se conoce, a priori, casos con datos de salida Y (variable dependiente) y entradas X (variables independientes) del problema.

- **No supervisado**: no se conoce a priori las salidas Y. No hay variable dependiente de otras variables. Se basa en la adaptación ante un entorno.

No supervisado

Encontrar patrones de los datos sin saber que existen.

- Ejemplo: **Clustering**: identifica características interesantes para separar grupos estadísticamente diferenciados.
    - Por ejemplo, de cada persona tener datos sobre altura, azúcar en sangre, edad, frecuencia cardíaca en reposo, si hace ejercicio o no, nivel de stress etc.
    - Se podría **encontrar grupo de los sanos y vida saludable** que tienen unos patrones y rasgos comunes y grupo de los que **no tienen vida saludable**, no se cuidan demasiado, no hacen deporte y tienen unos patrones y rasgos también comunes.

Supervisado

Existen dos tipos de problemas supervisados:

- **Problemas de regresión** (como por ejemplo la regresión lineal) salida de valor continuo

- **Problemas de clasificación** (como el modelo logístico) salida de variable categórica

Métodos de clasificación supervisada

Clasifica elementos en categorías conocidas, relacionando variables de entrada con una variable categórica de salida (por ejemplo binaria).

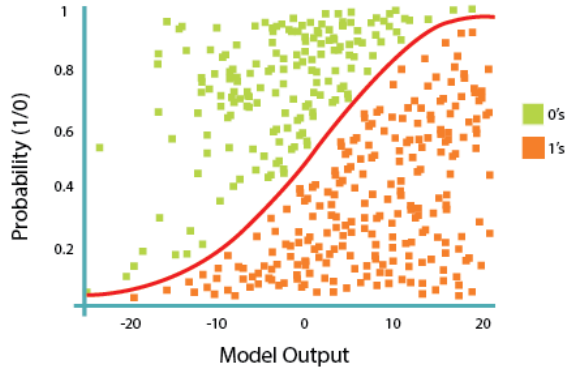
Ejemplo:los clientes de tarjetas de crédito pueden clasificarse en los que pagan sus créditos en un tiempo acordado y los que no lo hacen.

Intenta deducir variables explicativas (edad, ingresos, estado civil) que ayuden a predecir en cuál de estas dos categorías se encuentra un nuevo cliente.

Ejemplo supervisado: Regresión logística

Se cuenta con datos de las características del análisis de sangre (nivel de glucosa etc…) de varias personas y se desea clasificar a los pacientes en dos grupos, a partir de diagnóstico médico ya dado:

- Grupo 1 – Diabetes tipo I (menos grave)
- Grupo 2 – Diabetes tipo II (más grave)



Entrenar un algoritmo **clasificador** que permita calcular la probabilidad de que un nuevo paciente sea tipo I y la probabilidad de que sea tipo II.

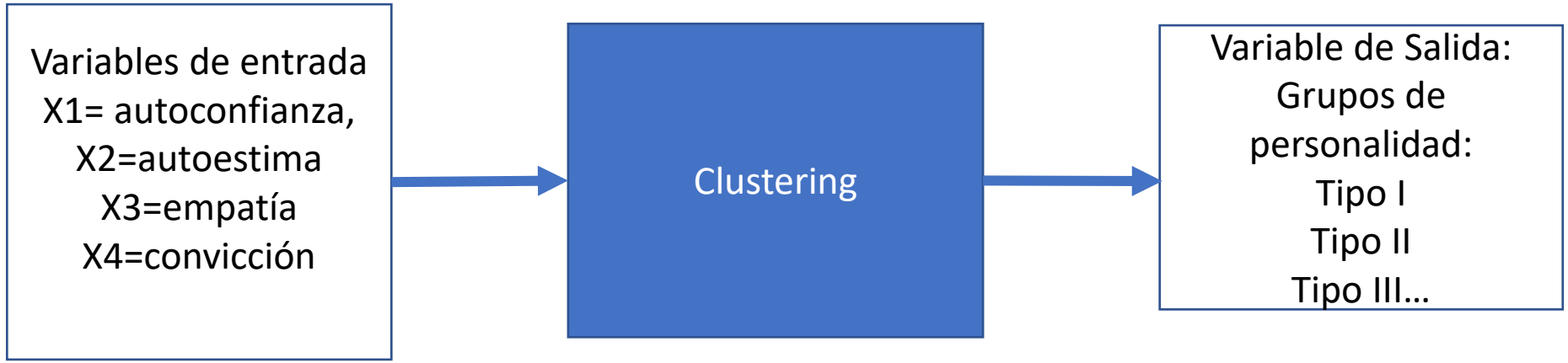
Ejemplo: Clasificación de diabetes

Una vez entrenado a partir de casos diagnosticados por médicos, permite:

- Predecir, para un nuevo paciente, la probabilidad de tener diabetes y de que tipo, si tiene valores de variables de entrada (X1,X2,X3).
- Ayudar al **entendimiento** de las relaciones para para toma de decisiones

Ejemplo no supervisado: clustering

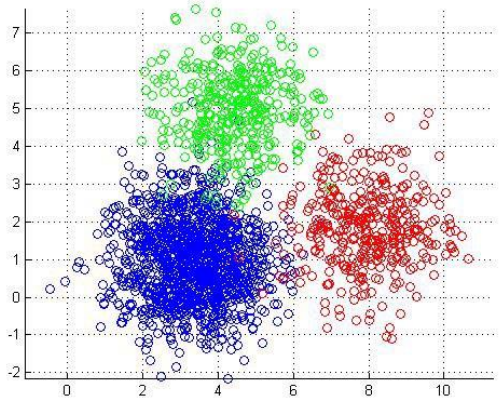
Encontrar grupos de personas con diferentes rasgos de personalidad, por ejemplo: el líder, el emprendedor, el artista etc)



Una vez ajustado permite:

- Predecir para un nuevo persona a que tipo de personalidad pertenece
- Ayudar al entendimiento de las relaciones para para toma de decisiones

Clusterización



- Es una técnica no supervisada porque no tiene variable dependiente.
- busca patrones de similitud entre las variables.
- Agrupa las observaciones (clientes, compañías, ciudades, etc.) de modo que al interior de cada cluster se puedan considerar similares, y no similares a las observaciones de otros cluster.
- Por ejemplo, para los clientes de un concesionario de carros un grupo de interés podrían ser hombres de mediana edad, casados, que ganan más de 15 millones y que suelen comprar carros deportivos de alta gama.

- Para saber si dos observaciones pertenecen a diferentes clusters debe considerarse una medida de disimilitud.

- Por ejemplo, si 2 clientes tienen el mismo género su disimilitud en esa variable es 0.
- Respecto a la variable ingresos, 2 clientes son disimiles en la diferencia entre sus ingresos.

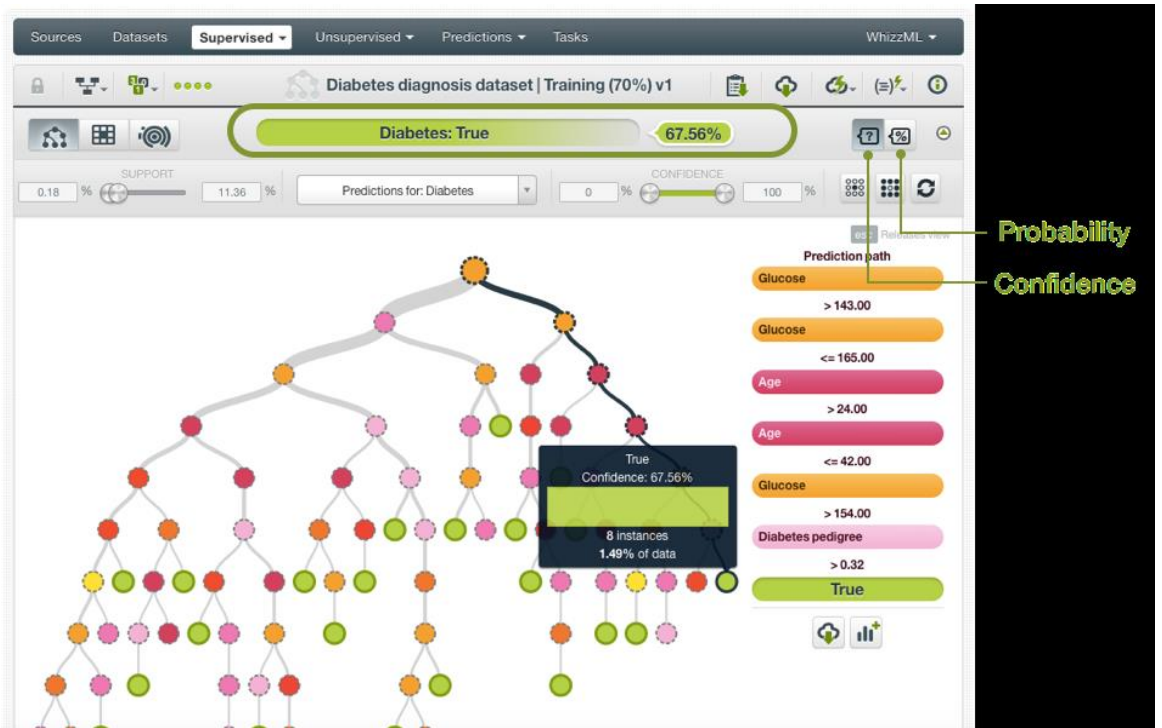
- Luego se suman todas las disimilitudes: como pueden estar en diferentes unidades, deben normalizarse (llevar a unidades adimensionales y a la misma escala) antes de sumarse.

Árbol de clasificación

- A partir de un conjunto de datos de diferentes variables, es capaz de descubrir relaciones no lineales entre ellas, de forma intuitiva.
- Salida: reglas simples de clasificación de nuevas observaciones.
- **Ejemplo:** A partir de datos de anteriores clientes de un banco, que han accedido a prestamos, se tiene registro de variables particulares: estado civil, edad, propiedades, ingresos, educación y si a pagado a tiempo o no.
- El árbol de clasificación selecciona las variables de mayor interés y construye una red jerárquica en forma de árbol partiendo de la raíz, derivando ramas y nodos para derivar nuevas ramas
- Cada nodo representa una variable que se divide en 2 ramas (o más) categorías, por ejemplo. ejemplo estado civil: casado o soltero. Cada una de ellas se divide en dos: casa propia o no, etc.

- Cada nodo hace una selección más fina que el nodo anterior.
- En cada nodo puede hacerse una medición de “pureza”, esto es el porcentaje de esa rama de los usuarios que pagan a tiempo y los que no

- El aprendizaje radica en seleccionar la mejor manera de dividir (orden jerárquico y umbrales de división) y el criterio de parada



Fuente: Classification and Regression with the BigML Dashboard
The BigML Team