MINERÍA DE DATOS – TÉCNICAS PREDICTIVAS DE MODELIZACIÓN

- TÉCNICAS DE MINERÍA DE DATOS.
- TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN.
- MODELO DE REGRESIÓN MÚLTIPLE.
- MODELOS DE ELECCIÓN DISCRETA.
- CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE.

TÉCNICAS DE MINERÍA DE DATOS

TÉCNICAS DE MINERÍA DE DATOS

- LA FASE DE **TÉCNICAS DE MINERÍA DE DATOS PROPIAMENTE DICHAS** ENGLOBA:
  - **TÉCNICAS PREDICTIVAS ENFOCADAS A LA MODELIZACIÓN Y CLASIFICACIÓN AD HOC**.
  - **TÉCNICAS DESCRIPTIVAS ENFOCADAS GENERALMENTE A LA CLASIFICACIÓN POST HOC Y OTRO TIPO DE TÉCNICAS VARIADAS**.

TÉCNICAS DE MINERÍA DE DATOS

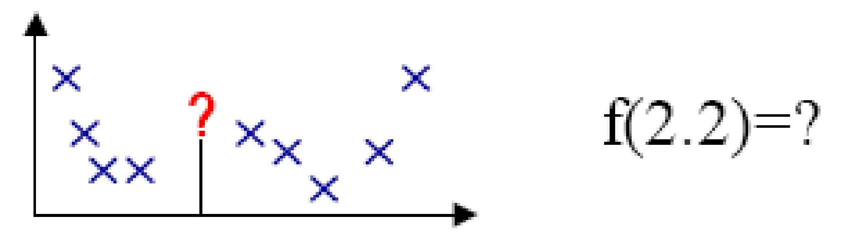
- **TÉCNICAS PREDICTIVAS**:
  - ESPECIFICAN EL MODELO PARA LOS DATOS EN BASE A UN **CONOCIMIENTO TEÓRICO PREVIO**.
  - EL MODELO SUPUESTO DEBE **CONTRASTARSE** DESPUÉS DEL PROCESO DE MINERÍA DE DATOS ANTES DE ACEPTARLO COMO **VÁLIDO**.
  - INCLUYEN TODOS LOS TIPOS DE:
    - REGRESIÓN.
    - SERIES TEMPORALES.
    - ANÁLISIS DE LA VARIANZA Y COVARIANZA.
    - ANÁLISIS DISCRIMINANTE.
    - ÁRBOLES DE DECISIÓN.
    - REDES NEURONALES.

TÉCNICAS DE MINERÍA DE DATOS

- **TÉCNICAS PREDICTIVAS**:
  - LOS **ÁRBOLES DE DECISIÓN**, LAS **REDES NEURONALES** Y EL **ANÁLISIS DISCRIMINANTE** SON A SU VEZ **TÉCNICAS DE CLASIFICACIÓN**:
    - PUEDEN **EXTRAER PERFILES** DE COMPORTAMIENTO O CLASES, SIENDO EL OBJETIVO CONSTRUIR UN MODELO QUE PERMITA **CLASIFICAR** CUALQUIER NUEVO DATO.
  - LOS **ÁRBOLES DE DECISIÓN** PERMITEN **CLASIFICAR** LOS DATOS EN GRUPOS BASADOS EN LOS VALORES DE LAS VARIABLES:
    - EL MECANISMO CONSISTE EN ELEGIR UN **ATRIBUTO COMO RAÍZ** Y DESARROLLAR EL ÁRBOL SEGÚN LAS VARIABLES MÁS SIGNIFICATIVAS.

TÉCNICAS DE MINERÍA DE DATOS

- **EJEMPLOS PREDICTIVOS**:
  - **INTERPOLACIÓN**:



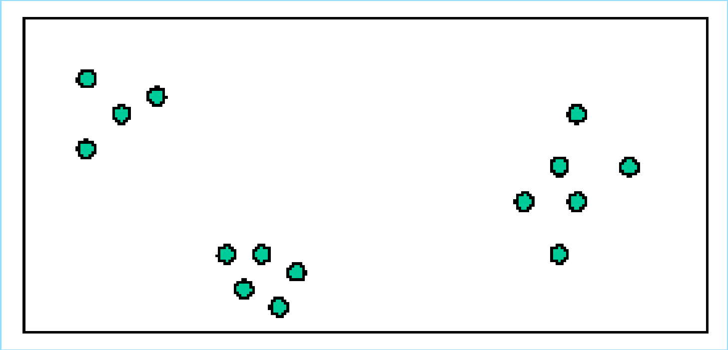
  - **PREDICCIÓN SECUENCIAL**:
    - 1, 2, 3, 5, 7, 11, 13, 17, 19, ... ?.
  - **APRENDIZAJE SUPERVISADO**:
    - 1 3 $\rightarrow$ 4.
    - 3 5 $\rightarrow$ 8.
    - 7 2 $\rightarrow$ 9.
    - 4 2 $\rightarrow$?.

TÉCNICAS DE MINERÍA DE DATOS

- **TÉCNICAS DESCRIPTIVAS**:
  - NO SE ASIGNA NINGÚN PAPEL PREDETERMINADO A LAS VARIABLES.
  - NO SE SUPONE LA EXISTENCIA DE VARIABLES DEPENDIENTES NI INDEPENDIENTES Y TAMPOCO SE SUPONE LA EXISTENCIA DE UN MODELO PREVIO PARA LOS DATOS.
  - **LOS MODELOS SE CREAN AUTOMÁTICAMENTE** PARTIENDO DEL RECONOCIMIENTO DE PATRONES.
  - INCLUYEN:
    - **CLUSTERING** Y **SEGMENTACIÓN** (QUE TAMBIÉN SON TÉCNICAS DE CLASIFICACIÓN EN CIERTO MODO).
    - **ASOCIACIÓN** Y **DEPENDENCIA**.
    - **ANÁLISIS EXPLORATORIO DE DATOS**.
    - **REDUCCIÓN DE LA DIMENSIÓN FACTORIAL, COMPONENTES PRINCIPALES, CORRESPONDENCIAS,** ETC.

TÉCNICAS DE MINERÍA DE DATOS

- **EJEMPLOS DESCRIPTIVOS**:
  - **SEGMENTACIÓN (APRENDIZAJE NO SUPERVISADO)**:
    - ¿CUÁNTOS GRUPOS HAY?.
    - ¿QUÉ GRUPOS FORMO?.



TÉCNICAS DE MINERÍA DE DATOS

- *Técnicas de minería*
  - *Predictivas*
    - *Modelización*
      - *Regresión*
      - *Modelos de elección discreta*
      - *Análisis de la varianza*
      - *Análisis canónico*
      - *Redes Neuronales*
    - *Clasificación ad hoc*
      - *Discriminante*
      - *Árboles de decisión*
  - *Descriptivas*
    - *Clasificación post hoc*
      - *Clustering*
      - *Segmentación*
    - *Asociación*
    - *Dependencia*
    - *Reducción de la dimensión*
    - *Análisis exploratorio*

TÉCNICAS DE MINERÍA DE DATOS

- LAS **TÉCNICAS DE CLASIFICACIÓN** PUEDEN PERTENECER:
  - AL GRUPO DE TÉCNICAS **PREDICTIVAS:** DISCRIMINANTE, ÁRBOLES DE DECISIÓN Y REDES NEURONALES.
  - AL GRUPO DE TÉCNICAS **DESCRIPTIVAS:** CLUSTERING Y SEGMENTACIÓN.
- LAS TÉCNICAS DE **CLASIFICACIÓN PREDICTIVAS** SUELEN DENOMINARSE **TÉCNICAS DE CLASIFICACIÓN AD HOC**:
  - CLASIFICAN INDIVIDUOS U OBSERVACIONES DENTRO DE **GRUPOS PREVIAMENTE DEFINIDOS**.
- LAS TÉCNICAS DE **CLASIFICACIÓN DESCRIPTIVAS** SE DENOMINAN **TÉCNICAS DE CLASIFICACIÓN POST HOC**:
  - REALIZAN CLASIFICACIÓN **SIN ESPECIFICACIÓN PREVIA DE LOS GRUPOS**.
- LAS **REDES NEURONALES** PUEDEN UTILIZARSE TANTO PARA LA **MODELIZACIÓN** COMO PARA LA **CLASIFICACIÓN**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **REVISIÓN DE CONCEPTOS PREVIOS**
- **VARIANZA**
- SI SE TIENE UN CONJUNTO DE DATOS DE UNA MISMA VARIABLE, LA **VARIANZA** SE CALCULA DE LA SIGUIENTE FORMA:
  - $$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2$$
  - X_i: CADA DATO.
  - *n*: N° DE ELEMENTOS.
  - $\overline{X}$: MEDIA ARITMÉTICA DE LOS DATOS.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **REVISIÓN DE CONCEPTOS PREVIOS**
- **COVARIANZA**
- PARA HACER EL **ESTUDIO CONJUNTO** DE LAS **VARIABLES CUANTITATIVAS** *X* E *Y*, SE SUPONE QUE SE DISPONE DE UNA **MUESTRA** DE *n* PARES DE **OBSERVACIONES** DE *X* E *Y*:
  - $$(x_1, y_1), \ldots, (x_n, y_n)$$
- LA **COVARIANZA** MUESTRAL ENTRE LAS OBSERVACIONES DE *X* E *Y* SE DEFINE COMO:
  - $$cov_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
  - $$\begin{aligned} cov_{x,y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right) \end{aligned}$$

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **REVISIÓN DE CONCEPTOS PREVIOS**
- **MODELO DE REGRESIÓN LINEAL**
- LA **RECTA DE REGRESIÓN DE *Y* SOBRE *X* ES LA RECTA *y* = *a* + *bx* QUE MINIMIZA** EL ERROR CUADRÁTICO MEDIO (E.C.M.):
  - $$E.C.M. = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$
  - $$a = \bar{y} - \frac{cov_{x,y}}{v_x}\bar{x} \qquad ; \qquad b = \frac{cov_{x,y}}{v_x}$$
- EL **COEFICIENTE DE CORRELACIÓN** LINEAL ENTRE *X* E *Y* SE DEFINE COMO:
  - $$r = \frac{cov_{x,y}}{\sqrt{v_x v_y}}$$

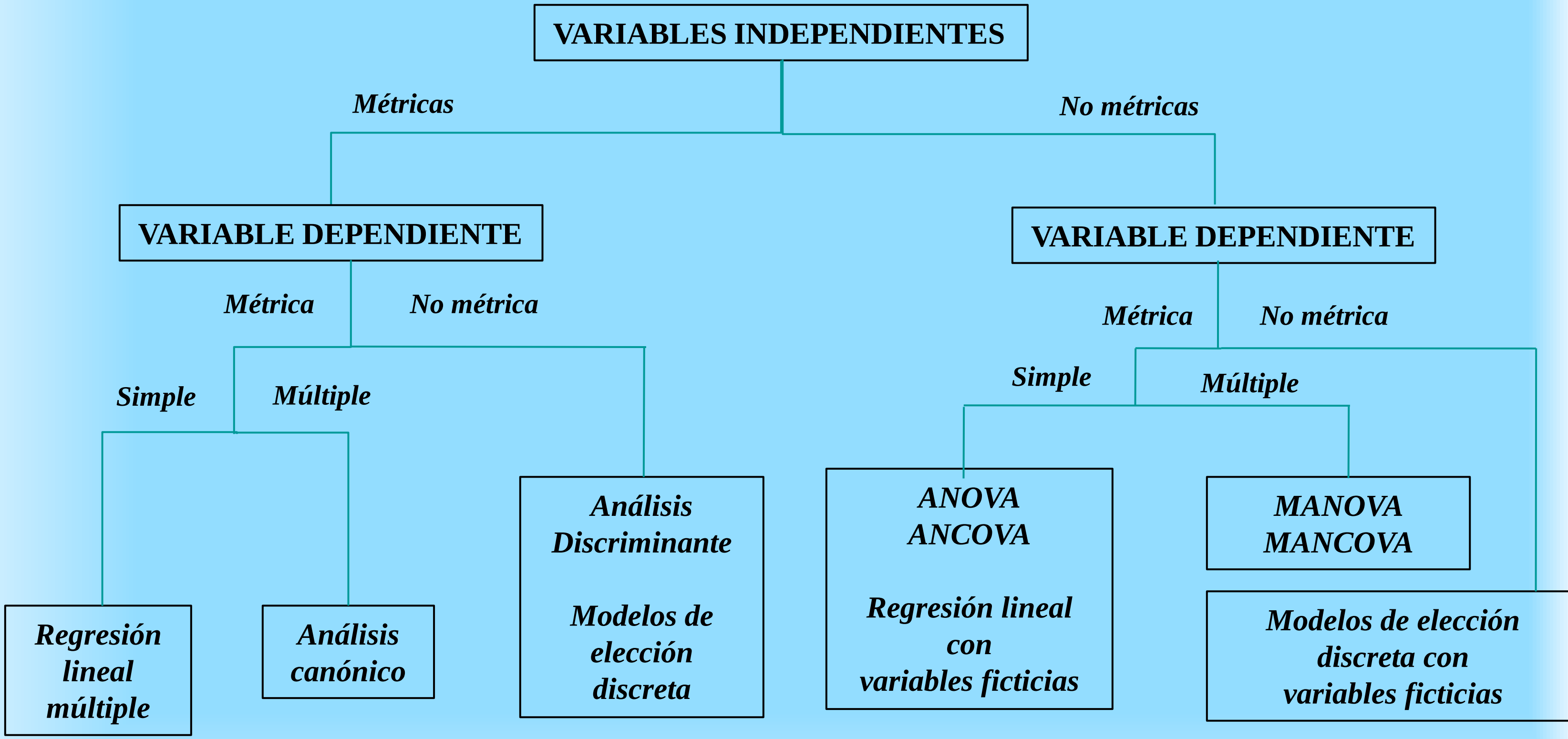
TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **REVISIÓN DE CONCEPTOS PREVIOS**
- **COEFICIENTE DE CORRELACIÓN PARCIAL**
- ES LA **RELACIÓN ENTRE DOS VARIABLES** CUANDO SE HA ELIMINADO DE CADA UNA DE ELLAS EL EFECTO QUE SOBRE ELLAS TIENE UNA TERCERA VARIABLE:
  - *X* , *Y* SON LAS VARIABLES OBJETO DEL ESTUDIO.
  - *Z* ES LA VARIABLE DE CONTROL.
  - CONSISTE EN ESTUDIAR LAS CORRELACIONES Y COMBINARLAS:
    - $r_{XY}\,,\ r_{XZ}\ ,\ r_{YZ}$
    - $$r_{XZ.Y} = \frac{r_{XZ} - \left(r_{XY}\right)\left(r_{YZ}\right)}{\sqrt{1 - r_{XY}^2}\,\sqrt{1 - r_{XZ}^2}}$$

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **TÉCNICAS PARA LA MODELIZACIÓN**
- LA CLASIFICACIÓN DE LAS TÉCNICAS DISCRIMINA ENTRE LA EXISTENCIA O NO DE **VARIABLES EXPLICATIVAS** Y **EXPLICADAS**.
- **TÉCNICAS PREDICTIVAS** O **MÉTODOS EXPLICATIVOS**:
  - EXISTE UNA **DEPENDENCIA** ENTRE LAS **VARIABLES** EXPLICADAS Y SUS VARIABLES EXPLICATIVAS, QUE PUEDA PLASMARSE EN UN **MODELO**.
- ESTAS **TÉCNICAS DE ANÁLISIS DE LA DEPENDENCIA**:
  - PUEDEN **CLASIFICARSE** EN FUNCIÓN DE LA **NATURALEZA** MÉTRICA O NO MÉTRICA DE LAS **VARIABLES** INDEPENDIENTES Y DEPENDIENTES.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN



TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS DE REGRESIÓN MÚLTIPLE**:
  - ES UTILIZADO PARA ANALIZAR LA RELACIÓN ENTRE:
    - UNA VARIABLE **DEPENDIENTE** (O **ENDÓGENA**) MÉTRICA.
    - VARIAS VARIABLES **INDEPENDIENTES** (O **EXÓGENAS**) TAMBIÉN MÉTRICAS.
  - EL **OBJETIVO** ESENCIAL ES UTILIZAR LAS VARIABLES **INDEPENDIENTES**, CUYOS VALORES SON CONOCIDOS, PARA **PREDECIR** LA ÚNICA VARIABLE CRITERIO (**DEPENDIENTE**) SELECCIONADA POR EL INVESTIGADOR.
- LA **EXPRESIÓN** ES LA SIGUIENTE:
  - $y = F(x_1, x_2, \ldots, x_n)$
  - DONDE INICIALMENTE, TANTO LA VARIABLE DEPENDIENTE y COMO LAS INDEPENDIENTES x_i SON MÉTRICAS.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- TAMBIÉN SE PUEDE TRABAJAR CON VARIABLES INDEPENDIENTES **NO MÉTRICAS** SI SE EMPLEAN VARIABLES FICTICIAS PARA SU TRANSFORMACIÓN EN MÉTRICAS:
  - **MODELOS DE REGRESIÓN CON VARIABLES FICTICIAS.**

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS CANÓNICO** O **ANÁLISIS DE LA CORRELACIÓN CANÓNICA:**
  - ES UNA TÉCNICA PARA ANALIZAR LA RELACIÓN ENTRE **MÚLTIPLES** VARIABLES **DEPENDIENTES** (O **ENDÓGENAS**) MÉTRICAS Y VARIAS VARIABLES **INDEPENDIENTES** (O **EXÓGENAS**) TAMBIÉN MÉTRICAS.
  - EL **OBJETIVO** ESENCIAL ES UTILIZAR LAS VARIABLES INDEPENDIENTES, CUYOS VALORES SON CONOCIDOS, PARA **PREDECIR** LAS VARIABLES CRITERIO (**DEPENDIENTES**).
- LA **EXPRESIÓN** ES LA SIGUIENTE:
  - $G(y_1, y_2,\ldots, y_n) = F(x_1, x_2,\ldots, x_n)$
  - DONDE INICIALMENTE, TANTO LAS VARIABLES DEPENDIENTES y_i COMO LAS INDEPENDIENTES x_i SON MÉTRICAS.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- ES UNA **AMPLIACIÓN** DEL MODELO DE REGRESIÓN MÚLTIPLE AL CASO DE **VARIAS VARIABLES DEPENDIENTES**.
- TAMBIÉN PUEDE EXTENDERSE AL CASO DE VARIABLES **DEPENDIENTES NO MÉTRICAS** Y AL CASO DE VARIABLES **INDEPENDIENTES NO MÉTRICAS**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS DISCRIMINANTE:**
  - SE USA PARA ANALIZAR LA **RELACIÓN** ENTRE UNA VARIABLE **DEPENDIENTE** (O **ENDÓGENA**) NO MÉTRICA (**CATEGÓRICA**) Y VARIAS VARIABLES **INDEPENDIENTES** (O **EXÓGENAS**) MÉTRICAS.
  - EL **OBJETIVO** ES UTILIZAR LOS VALORES CONOCIDOS DE LAS VARIABLES INDEPENDIENTES PARA **PREDECIR** CON QUÉ **CATEGORÍA** DE LA VARIABLE DEPENDIENTE SE CORRESPONDEN.
  - SE PUEDE **PREDECIR EN QUÉ CATEGORÍA** DE RIESGO CREDITICIO SE ENCUENTRA UNA PERSONA, EL ÉXITO DE UN PRODUCTO EN EL MERCADO, ETC.
- LA **EXPRESIÓN** ES:
  - $y = F(x_1, x_2, \ldots, x_n)$
  - DONDE y (**DEPENDIENTE**) ES **NO MÉTRICA** Y LAS VARIABLES **INDEPENDIENTES** SON **MÉTRICAS**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- ES UN **CASO PARTICULAR** DEL ANÁLISIS DE **REGRESIÓN MÚLTIPLE.**
- ES UNA **TÉCNICA DE CLASIFICACIÓN** QUE PERMITE:
  - **AGRUPAR** A LOS ELEMENTOS DE UNA MUESTRA EN DOS O MÁS **CATEGORÍAS** DIFERENTES, **PREDEFINIDAS** EN UNA VARIABLE **DEPENDIENTE NO MÉTRICA**, EN FUNCIÓN DE UNA SERIE DE VARIABLES **INDEPENDIENTES MÉTRICAS** COMBINADAS LINEALMENTE.
- PARA VALORES DADOS DE LAS VARIABLES INDEPENDIENTES SE DEBE **PREDECIR LA PROBABILIDAD DE PERTENENCIA A UNA CATEGORÍA O CLASE** DE LA VARIABLE DEPENDIENTE:
  - **EJEMPLO**: SEGÚN ALGUNAS VARIABLES MEDIDAS EN EL INDIVIDUO, PREDECIR LA PROBABILIDAD DE QUE:
    - UN INDIVIDUO COMPRE UN PRODUCTO.
    - UN INDIVIDUO DEVUELVA UN CRÉDITO.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **MODELOS DE ELECCIÓN DISCRETA:**
  - TIENEN LA **MISMA NATURALEZA** QUE EL **MODELO DISCRIMINANTE**.
  - SE **PREDICE LA PROBABILIDAD** DE PERTENENCIA A UNA CATEGORÍA (**CLASE**) PARA VALORES DADOS DE LAS VARIABLES DEPENDIENTES.
  - **PREDICEN** DIRECTAMENTE LA **PROBABILIDAD DE OCURRENCIA DE UN SUCESO** QUE VIENE DEFINIDO POR LOS VALORES DE LAS VARIABLES INDEPENDIENTES.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- UN CASO PARTICULAR DEL MODELO DE REGRESIÓN MÚLTIPLE ES EL **MODELO LINEAL DE PROBABILIDAD**:
  - $P_i = F(x_i, \beta) + u_i$
- SI ***F*** ES LA **FUNCIÓN DE DISTRIBUCIÓN DE UNA VARIABLE ALEATORIA**, ENTONCES ***P*** VARÍA ENTRE 0 Y 1.
- SI ***F*** ES LA **FUNCIÓN LOGÍSTICA** SE TIENE EL MODELO **LOGIT O REGRESIÓN LOGÍSTICA**:
  - $P_i = F(x_i, \beta) + u_i = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} u_i$
- SI ***F*** ES LA **FUNCIÓN** DE DISTRIBUCIÓN DE UNA **NORMAL** UNITARIA SE TIENE EL **MODELO PROBIT**:
  - $P_i = F(x_i, \beta) + u_i = (2\pi)^{-1/2} \int_{-\infty}^{x_i\beta} e^{\frac{-t^2}{2}} dt + u_i$

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS DE LA VARIANZA SIMPLE** SE UTILIZA PARA ANALIZAR LA RELACIÓN ENTRE:
  - UNA VARIABLE DEPENDIENTE (O **ENDÓGENA**) **MÉTRICA** Y
  - VARIAS VARIABLES INDEPENDIENTES (O **EXÓGENAS**) **NO MÉTRICAS**.
- EL **OBJETIVO** ES DETERMINAR SI DIVERSAS **MUESTRAS** PROCEDEN DE **POBLACIONES CON IGUAL MEDIA**.
- LOS **VALORES NO MÉTRICOS** DE LAS VARIABLES **INDEPENDIENTES** DETERMINARÁN UNA SERIE DE **GRUPOS** EN LA **VARIABLE DEPENDIENTE**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL MODELO **ANOVA** MIDE LA **SIGNIFICACIÓN ESTADÍSTICA** DE LAS **DIFERENCIAS ENTRE LAS MEDIAS** DE LOS **GRUPOS** DETERMINADOS EN LA VARIABLE DEPENDIENTE POR LOS VALORES DE LAS VARIABLES INDEPENDIENTES:
  - $y = F(x_1, x_2, \ldots, x_n)$
  - DONDE LA VARIABLE **DEPENDIENTE** *y* ES **MÉTRICA** Y LAS VARIABLES **INDEPENDIENTES** SON **NO MÉTRICAS**.
- SE TRATA POR TANTO DE OTRO **CASO PARTICULAR** DEL MODELO DE **REGRESIÓN MÚLTIPLE.**

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS DE LA COVARIANZA SIMPLE** ES UNA TÉCNICA UTILIZADA PARA ANALIZAR LA **RELACIÓN** ENTRE UNA VARIABLE **DEPENDIENTE** (O **ENDÓGENA**) **MÉTRICA** Y VARIAS VARIABLES **INDEPENDIENTES** (O **EXÓGENAS**), PARTE DE LAS CUALES SON **NO MÉTRICAS**, SIENDO LA OTRA PARTE **MÉTRICAS** (COVARIABLES):
  - $y = F(x_1, x_2, \ldots, x_n)$
  - DONDE LA VARIABLE DEPENDIENTE ***y*** ES **MÉTRICA** Y LAS VARIABLES **INDEPENDIENTES** SON ALGUNAS **MÉTRICAS** Y OTRAS **NO MÉTRICAS**.
- ES OTRO **CASO PARTICULAR** DEL MODELO DE **REGRESIÓN MÚLTIPLE.**

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS DE LA VARIANZA MÚLTIPLE** ES UNA TÉCNICA UTILIZADA PARA ANALIZAR LA **RELACIÓN** ENTRE:
  - **VARIAS** VARIABLES **DEPENDIENTES** (O **ENDÓGENAS**) **MÉTRICAS** Y
  - **VARIAS** VARIABLES **INDEPENDIENTES** (O **EXÓGENAS**) **NO MÉTRICAS.**
- EL **OBJETIVO** ES **CONTRASTAR** SI LOS VALORES NO MÉTRICOS DE LAS VARIABLES **INDEPENDIENTES** DETERMINARÁN LA **IGUALDAD** DE VECTORES DE MEDIAS DE UNA SERIE DE **GRUPOS** DETERMINADOS POR ELLOS EN LAS VARIABLES DEPENDIENTES.
- EL MODELO **MANOVA** MIDE LA **SIGNIFICACIÓN ESTADÍSTICA** DE LAS **DIFERENCIAS** ENTRE LOS VECTORES DE **MEDIAS** DE LOS **GRUPOS** DETERMINADOS EN LAS VARIABLES DEPENDIENTES POR LOS VALORES DE LAS VARIABLES INDEPENDIENTES.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- LA EXPRESIÓN ES:
  - $G(y_1, y_2, \ldots, y_m) = F(x_1, x_2, \ldots, x_n)$
  - DONDE LAS VARIABLES **DEPENDIENTES** SON **MÉTRICAS** Y LAS VARIABLES **INDEPENDIENTES** SON **NO MÉTRICAS**.
- ES OTRO **CASO PARTICULAR** DE LA **REGRESIÓN MÚLTIPLE**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **ANÁLISIS DE LA COVARIANZA MÚLTIPLE** SE USA PARA ANALIZAR LA RELACIÓN ENTRE:
  - **VARIAS** VARIABLES DEPENDIENTES (O **ENDÓGENAS**) **MÉTRICAS** Y
  - **VARIAS** VARIABLES INDEPENDIENTES (O **EXÓGENAS**) MEZCLA DE VARIABLES **MÉTRICAS** Y **NO MÉTRICAS**.
- LA EXPRESIÓN ES:
  - $G(y_1, y_2,\ldots, y_m) = F(x_1, x_2,\ldots, x_n)$
  - DONDE LAS VARIABLES **DEPENDIENTES** SON **MÉTRICAS** Y LAS VARIABLES **INDEPENDIENTES** SON UNA PARTE **MÉTRICAS** Y OTRA PARTE **NO MÉTRICAS**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EN EL **ANÁLISIS DE LA COVARIANZA** (SIMPLE Y MÚLTIPLE):
  - LAS VARIABLES MÉTRICAS INDEPENDIENTES (**COVARIABLES**) TIENEN COMO **OBJETIVO ELIMINAR** DETERMINADOS EFECTOS QUE PUEDAN **SESGAR** LOS RESULTADOS **INCREMENTANDO LA VARIANZA** DENTRO DE LOS GRUPOS:
    - **ELIMINAR**, MEDIANTE UNA **REGRESIÓN LINEAL**, LA VARIACIÓN EXPERIMENTADA POR LAS VARIABLES DEPENDIENTES PRODUCIDA POR LA **COVARIABLE** O **COVARIABLES** DE EFECTOS INDESEADOS.
    - HACER UN ANÁLISIS **ANOVA** O **MANOVA** SOBRE LAS **VARIABLES DEPENDIENTES AJUSTADAS** (RESIDUOS DE LA REGRESIÓN ANTERIOR).

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- LA **REGRESIÓN MÚLTIPLE** ADMITE LA POSIBILIDAD DE TRABAJAR CON **VARIABLES INDEPENDIENTES NO MÉTRICAS** SI SE EMPLEAN **VARIABLES FICTICIAS** PARA SU TRANSFORMACIÓN EN **MÉTRICAS**:
  - A CADA CLASE DE LA VARIABLE NO MÉTRICA SE LE **ASIGNA UN VALOR NUMÉRICO**.

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- EL **MODELO DE REGRESIÓN MÚLTIPLE CON VARIABLES FICTICIAS:**
  - ES **SIMILAR** AL ANÁLISIS DE LA **REGRESIÓN MÚLTIPLE**.
  - LA **DIFERENCIA** ES QUE LAS VARIABLES **INDEPENDIENTES** PUEDEN SER TAMBIÉN **NO MÉTRICAS**.
- SE USA PARA ANALIZAR LA **RELACIÓN** ENTRE UNA VARIABLE DEPENDIENTE (O **ENDÓGENA**) **MÉTRICA** Y VARIAS VARIABLES INDEPENDIENTES (O **EXÓGENAS**) **MÉTRICAS**, **NO MÉTRICAS** O **MEZCLA** DE AMBAS.
- EL **OBJETIVO** ES UTILIZAR LAS VARIABLES INDEPENDIENTES, CUYOS VALORES SON CONOCIDOS, PARA **PREDECIR** LA ÚNICA VARIABLE CRITERIO (**DEPENDIENTE**).
- LA EXPRESIÓN ES:
  - $y = F(x_1, x_2, \ldots, x_n)$

TÉCNICAS PREDICTIVAS PARA LA MODELIZACIÓN

- **MÉTODOS DEL ANÁLISIS MULTIVARIANTE DE LA DEPENDENCIA, SEGÚN LA NATURALEZA DE SUS VARIABLES DEPENDIENTES E INDEPENDIENTES:**

TÉCNICA	VARIABLES DEPENDIENTES	VARIABLES INDEPENDIENTES
ANOVA Y MANOVA	Métrica (métricas)	No métricas
ANCOVA Y MANCOVA	Métrica (métricas)	Métricas y no métricas
REGRESIÓN MÚLTIPLE	Métrica	Métricas
REGRESIÓN MÚLTIPLE (VARIABLES FICTICIAS)	Métrica	Métricas y no métricas
CORRELACIÓN CANÓNICA	Métricas y no métricas	Métricas y no métricas
ELECCIÓN DISCRETA	No métrica	Métricas
ELECCIÓN DISCRETA (VARIABLES FICTICIAS)	No métrica	Métricas y no métricas

MODELO DE REGRESIÓN MÚLTIPLE

MODELO DE REGRESIÓN MÚLTIPLE

- LA **REGRESIÓN MÚLTIPLE** TIENE COMO **OBJETIVO** ANALIZAR UN MODELO QUE PRETENDE **EXPLICAR EL COMPORTAMIENTO** DE UNA VARIABLE (**ENDÓGENA**, **EXPLICADA** O **DEPENDIENTE**), Y, UTILIZANDO UN CONJUNTO DE VARIABLES EXPLICATIVAS (**EXÓGENAS** O **INDEPENDIENTES**), $X_1, X_2, \ldots, X_K$.
- EL MODELO **LINEAL** (MODELO **ECONOMÉTRICO**) VIENE DADO POR:
  - $Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + u$
- LOS COEFICIENTES (**PARÁMETROS**) $b_1, b_2, \ldots, b_k$ DENOTAN LA **MAGNITUD DEL EFECTO** QUE LAS VARIABLES EXPLICATIVAS (**EXÓGENAS** O **INDEPENDIENTES**) $X_1, X_2, \ldots, X_K$ TIENEN SOBRE LA VARIABLE EXPLICADA (**ENDÓGENA** O **DEPENDIENTE**) Y.
- EL COEFICIENTE b_0 SE DENOMINA **TÉRMINO CONSTANTE** (O **INDEPENDIENTE**) DEL MODELO.
- EL TÉRMINO u SE DENOMINA **TÉRMINO DE ERROR** DEL MODELO.

MODELO DE REGRESIÓN MÚLTIPLE

- SI SE DISPONE DE UN **CONJUNTO** DE *T* **OBSERVACIONES** PARA C/U DE LAS VARIABLES ENDÓGENA Y EXÓGENAS, EL **MODELO** SE ESCRIBE DE LA FORMA:
  - $Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt} + u_t \qquad t=1,2,3,\ldots,T$
- LA APARICIÓN (NO NECESARIA) DE UN **TÉRMINO INDEPENDIENTE** EN EL MODELO PUEDE INTERPRETARSE COMO LA **PRESENCIA DE UNA PRIMERA VARIABLE** X_0 CUYO VALOR SEA **SIEMPRE** 1.
- **PROBLEMA FUNDAMENTAL**: SUPONIENDO QUE LA RELACIÓN ENTRE LA VARIABLE *Y* Y EL CONJUNTO DE VARIABLES X_1, X_2,…, X_K ES COMO SE HA DESCRITO EN EL MODELO, Y QUE SE DISPONE DE UN CONJUNTO DE *T* OBSERVACIONES PARA C/U DE LAS VARIABLES, LA ENDÓGENA Y LAS EXÓGENAS, ¿CÓMO PUEDEN ASIGNARSE VALORES NUMÉRICOS A LOS PARÁMETROS b_0, b_1, b_2,…, b_k, BASÁNDONOS EN LA INFORMACIÓN MUESTRAL?:
  - ESTOS VALORES SE LLAMARÁN **ESTIMACIONES DE LOS PARÁMETROS.**

MODELO DE REGRESIÓN MÚLTIPLE

- UNA VEZ ENCONTRADAS LAS **ESTIMACIONES** DE LOS PARÁMETROS DEL MODELO:
  - SE PODRÁ HACER **PREDICCIONES** ACERCA DEL COMPORTAMIENTO FUTURO DE LA VARIABLE *Y*.

MODELO DE REGRESIÓN MÚLTIPLE

- EL **MODELO LINEAL** SE FORMULA BAJO LAS SIGUIENTES **HIPÓTESIS**:
  - LAS VARIABLES $X_1, X_2, \ldots, X_K$, SON **DETERMINISTAS** (NO SON VARIABLES ALEATORIAS), YA QUE SU VALOR ES UN VALOR CONSTANTE PROVENIENTE DE UNA **MUESTRA** TOMADA.
  - LA VARIABLE u (**TÉRMINO DE ERROR**) ES UNA VARIABLE **ALEATORIA** CON ESPERANZA NULA Y MATRIZ DE COVARIANZAS CONSTANTE Y DIAGONAL (MATRIZ ESCALAR):
    - PARA TODO t, LA VARIABLE u_t, TIENE **MEDIA** CERO Y **VARIANZA** σ^2 NO DEPENDIENTE DE t, Y ADEMÁS $Cov(u_i, u_j)=0$ PARA TODO i Y PARA TODO j DISTINTOS ENTRE SÍ:
      - EL HECHO DE QUE LA **VARIANZA** DE u_t SEA **CONSTANTE** PARA TODO t (QUE NO DEPENDA DE t), SE DENOMINA **HIPÓTESIS DE HOMOSCEDASTICIDAD**.
      - EL HECHO DE QUE $Cov(u_i, u_j)=0$ PARA TODO i DISTINTO DE j SE DENOMINA **HIPÓTESIS DE NO AUTOCORRELACIÓN**.

MODELOS DE ELECCIÓN DISCRETA

MODELOS DE ELECCIÓN DISCRETA

- LA EXPRESIÓN DEL **MODELO DE ANÁLISIS DE LA REGRESIÓN MÚLTIPLE** ES:
  - $y = F(x_1, x_2, \dots, x_n)$.
- LA REGRESIÓN MÚLTIPLE ADMITE LA POSIBILIDAD DE TRABAJAR CON **VARIABLES DEPENDIENTES DISCRETAS** EN VEZ DE CONTINUAS PARA PERMITIR LA **MODELIZACIÓN DE FENÓMENOS DISCRETOS**.

MODELOS DE ELECCIÓN DISCRETA

- **MODELOS DE ELECCIÓN DISCRETA:**
  - LA VARIABLE **DEPENDIENTE** ES UNA VARIABLE **DISCRETA** QUE REFLEJA DECISIONES INDIVIDUALES EN LAS QUE EL CONJUNTO DE ELECCIÓN ESTÁ FORMADO POR **ALTERNATIVAS SEPARADAS Y MUTUAMENTE EXCLUYENTES**.
- LOS **MODELOS DE ELECCIÓN DISCRETA** EN LOS QUE EL **CONJUNTO DE ELECCIÓN** TIENE SÓLO **DOS ALTERNATIVAS** POSIBLES SE LLAMAN **MODELOS DE ELECCIÓN BINARIA**.
- CUANDO EL **CONJUNTO DE ELECCIÓN** TIENE **VARIOS VALORES DISCRETOS** SE TIENEN LOS MODELOS DE ELECCIÓN MÚLTIPLE O **MODELOS MULTINOMIALES**.

MODELOS DE ELECCIÓN DISCRETA

- LOS MODELOS DE ELECCIÓN DISCRETA SE DENOMINAN **MODELOS DE DATOS DE RECUENTO** CUANDO LOS VALORES DE LA **VARIABLE DEPENDIENTE** DISCRETA SON NÚMEROS QUE **NO REFLEJAN CATEGORÍAS.**
- EN CASO DE QUE LOS VALORES NUMÉRICOS DE LA **VARIABLE DEPENDIENTE** DISCRETA REFLEJAN **CATEGORÍAS** LOS MODELOS SE DENOMINAN **MODELO DE ELECCIÓN DISCRETA CATEGÓRICOS**:
  - SE CLASIFICAN EN:
    - **MODELOS DE ELECCIÓN DISCRETA CATEGÓRICOS ORDENADOS:** LOS **VALORES** NUMÉRICOS NO TIENEN SIGNIFICADO CUANTITATIVO **Y REFLEJAN UN ORDEN DE CATEGORÍAS.**
    - **MODELOS DE ELECCIÓN DISCRETA CATEGÓRICOS NO ORDENADOS:** LOS **VALORES** NUMÉRICOS **REFLEJAN ÚNICAMENTE CATEGORÍAS.**

CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE

CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE

- ES ÚTIL CUANDO SE DESEA CONSTRUIR UN MODELO **PREDICTIVO** PARA PRONOSTICAR EL **GRUPO AL QUE PERTENECE** UNA OBSERVACIÓN A PARTIR DE DETERMINADAS **CARACTERÍSTICAS** OBSERVADAS QUE DELIMITAN SU **PERFIL**.
- PERMITE **ASIGNAR O CLASIFICAR** NUEVOS INDIVIDUOS U OBSERVACIONES DENTRO DE **GRUPOS PREVIAMENTE DEFINIDOS**:
  - POR ELLO ES UNA **TÉCNICA DE CLASIFICACIÓN AD HOC**.
- SE LO CONOCE COMO **ANÁLISIS DE LA CLASIFICACIÓN**:
  - SU **OBJETIVO** FUNDAMENTAL ES:
    - PRODUCIR UNA **REGLA** O UN **ESQUEMA DE CLASIFICACIÓN.**
    - DEBE **PREDECIR LA POBLACIÓN** A LA QUE ES **MÁS PROBABLE** QUE TENGA QUE **PERTENECER** UNA NUEVA OBSERVACIÓN O **INDIVIDUO**.

CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE

- EL MODELO PREDICTIVO **DEFINE LA RELACIÓN** ENTRE:
  - UNA VARIABLE DEPENDIENTE (O **ENDÓGENA**) NO MÉTRICA (**CATEGÓRICA**), Y.
  - **VARIAS** VARIABLES INDEPENDIENTES (O **EXÓGENAS**) **MÉTRICAS**.
- LA EXPRESIÓN ES:
  - $\boldsymbol{y = F(x_1, x_2, \ldots, x_n)}$.
- LAS **CATEGORÍAS** DE LA VARIABLE **DEPENDIENTE** DEFINEN LOS POSIBLES **GRUPOS DE PERTENENCIA** DE LAS OBSERVACIONES O INDIVIDUOS.
- LAS VARIABLES **INDEPENDIENTES** DEFINEN EL **PERFIL** CONOCIDO DE CADA **OBSERVACIÓN**.

CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE

- EL **OBJETIVO** ESENCIAL:
  - ES **UTILIZAR** LOS VALORES CONOCIDOS DE LAS VARIABLES **INDEPENDIENTES** MEDIDAS SOBRE UN INDIVIDUO U OBSERVACIÓN (**PERFIL**).
  - PARA **PREDECIR** CON QUÉ **CATEGORÍA** DE LA VARIABLE **DEPENDIENTE** SE CORRESPONDEN PARA **CLASIFICAR** AL INDIVIDUO EN LA **CATEGORÍA** ADECUADA.

CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE

- LAS DOS GRANDES **FINALIDADES** SON:
  - LA **DESCRIPCIÓN DE DIFERENCIAS** ENTRE GRUPOS, Y.
  - LA **PREDICCIÓN DE PERTENENCIA** A GRUPOS.
- LA INTERPRETACIÓN DE LAS **DIFERENCIAS ENTRE LOS GRUPOS** RESPONDE AL OBJETIVO DE **DETERMINAR**:
  - EN QUÉ MEDIDA UN **CONJUNTO DE CARACTERÍSTICAS** OBSERVADAS EN LOS **INDIVIDUOS** PERMITE EXTRAER **DIMENSIONES** QUE **DIFERENCIAN A LOS GRUPOS**.
  - **CUÁLES** DE ESTAS **CARACTERÍSTICAS** SON LAS QUE EN MAYOR MEDIDA **CONTRIBUYEN** A TALES DIMENSIONES, ES DECIR, CUÁLES PRESENTAN EL MAYOR PODER DE **DISCRIMINACIÓN**.
- LAS **CARACTERÍSTICAS** USADAS PARA DIFERENCIAR ENTRE LOS GRUPOS RECIBEN EL NOMBRE DE **VARIABLES DISCRIMINANTES**.

CLASIFICACIÓN AD HOC: ANÁLISIS DISCRIMINANTE

- **HIPÓTESIS EN EL MODELO DISCRIMINANTE**
- LA APLICACIÓN DEL ANÁLISIS DISCRIMINANTE SE APOYA EN UNA SERIE DE **SUPUESTOS BÁSICOS:**
  - **NORMALIDAD** MULTIVARIANTE.
  - HOMOGENEIDAD DE MATRICES DE VARIANZA-COVARIANZA (**HOMOSCEDASTICIDAD**).
  - **LINEALIDAD** Y AUSENCIA DE MULTICOLINEALIDAD.