

# ITESO

## DEPARTAMENTO DE MATEMÁTICAS Y FÍSICA

Asignatura: Ciencia de Datos e Inteligencia de Negocios

EXAMEN (Medición Estadística de Datos y Medidas de Similitud)

Nombre: Eduardo Castillo Martínez



(2 punto) Explique cuáles y sus características de los diferentes tipos de datos que podemos encontrar.

- Cuantitativos: datos numéricos que representan cierta jerarquía dependiendo su magnitud.
- Cualitativos:
  - Multiestado: Representan clasificaciones de una variable.
  - Enteros positivos: Normalmente son datos Multiestado transformados a numéricos.
    - Nominales: no representa jerarquía
    - Ordinarios: dependiendo de la magnitud se establece jerarquía
- Binarios: datos con solo dos clasificaciones de tipo booleano, cierto o falso normalmente representando por 1 y 0.
- Genéticos: clasificación de especies.



2. (2 punto) Considerando las características numeradas en la siguiente lista, complete la tabla 1 con la información faltante colocando el número de la característica en el lugar de la tabla que indique si es su **utilización**, **tipo de aprendizaje**, **ventaja** o **desventaja** en el algoritmo que le corresponda (si es que aplica).


Algoritmo	Utilización	Tipo de Aprendizaje	Ventaja	Desventaja
Hierarchical Clustering (HC)	P	K	M	B A Q
K-Means (KM)	F	K	M	N A Q
Regresión Logística (RLog)	E	C G		J L
Regresión Lineal (Rlin)	O D	C R		J
Regresión Polinomial (Rpol)	O D	C R		J
Componentes Principales (PCA)	H		S.	J

Tabla 1. Complete la información faltante


3. Si se tuvieran dos bases de datos pequeñas como las siguientes:

Num	Estado	Zona Geográfica	Índice de delincuencia
1	Jalisco	1	60
2	Edo. México	1	50
3	Sonora	2	76
4	Guerrero	3	20
5	Chiapas	3	30

Num	Municipio	Sector productivo	Población Actual (hab) <sup>1</sup>
1	Zapopan	6	100,000.00
2	Guadalajara	2	300,000.00
3	Tlaquepaque	1	259,236.05
4	Tequila	4	540,689.00
5	Zapotlanejo	3	200,000.00

a.  1 punto) Indique cuantos tipos de variables hay en estas dos bases de datos. Haga una lista del nombre de variable y el tipo al que pertenece.


- Estado: Cualitativa, Multiestado
- Zona geográfica: Cualitativa, Enteros positivos nominales.
- Índice de delincuencia: Cuantitativa
- Municipio: Cualitativa, Multiestado
- Sector productivo: Cualitativa, Enteros positivos nominales.
- Población Actual: Cuantitativa

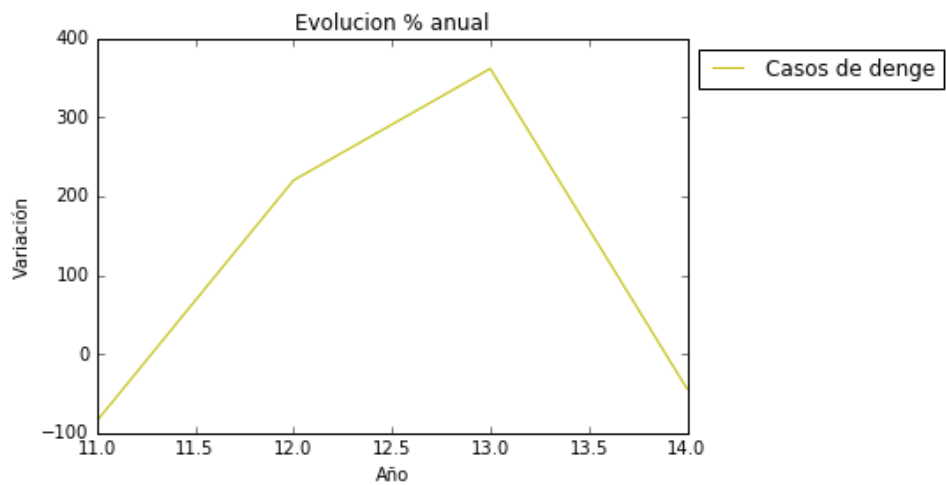
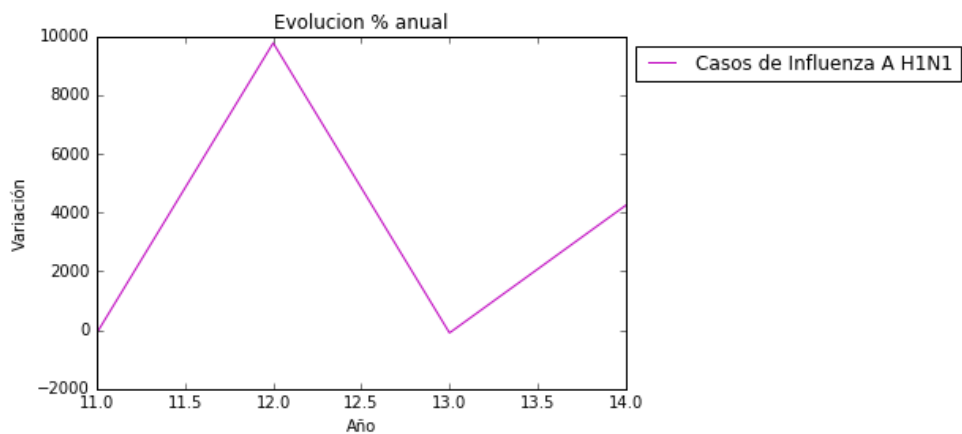
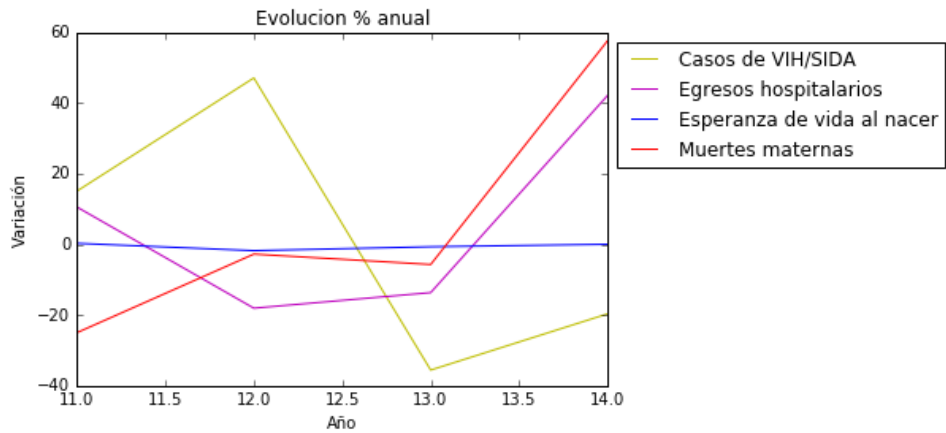
b.  punto) ¿Es posible aplicar directamente un análisis de componentes principales a estas dos bases de datos? Explique o justifique su respuesta.


No es posible, el ACP requiere que las variables categóricas(Multiestado) sean convertidas en variables tipo “dummy”.

4. Considerando la siguiente base de datos (en el Moodle se encuentra el archivo csv si les es más fácil), responda las siguientes preguntas y justifique su respuesta:


nombre	2014	2013	2012	2011	2010
Casos de Dengue	1446	2584	560	175	1171
Casos de Influenza A H1N1	608	14	592	6	108
Casos de VIH/SIDA	506	630	978	665	578
Egresos hospitalarios	221364	155789	180462	220280	199288
Esperanza de vida al nacer	75.36	75.36	75.89	77.28	77.07
Muertes maternas	52	33	35	36	48

a.  1 puntos) Realicé un gráfico donde muestre la evolución porcentual de cada una de las categorías al pasar de los años.



b)  punto) Porcentualmente ¿qué categoría es la que ha tenido más variaciones en el transcurso de los años?

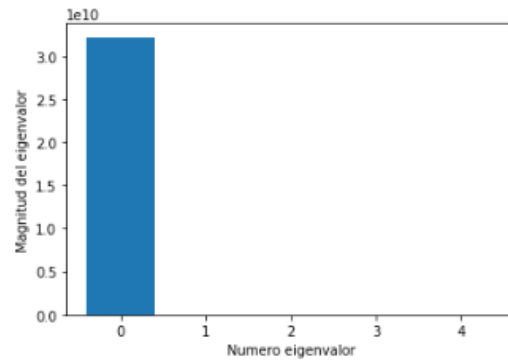
Casos de Influenza A H1N1

c)  punto) ¿Qué categorías han tenido un desempeño porcentual similar en el transcurso de los años?

Basado en análisis visual de los gráficos, media y correlación, se determinó que las categorías con desempeño similar son "Muertes maternas" y "Egresos hospitalarios".

5. Considere nuevamente la pequeña base de datos del ejercicio 4. Si aplicamos el análisis de componentes principales (PCA) y se obtienen los siguientes eigenvalores.

$W = [3.22056079e+10, 7.02176974e+05, 4.93786877e+04, 2.86913156e+04, 3.63003228e+00]$



1 punto) ¿Es correcto pensar que con solo los datos del año 2014 de todas categorías es suficiente para poder distinguirlas? Justifique su respuesta.

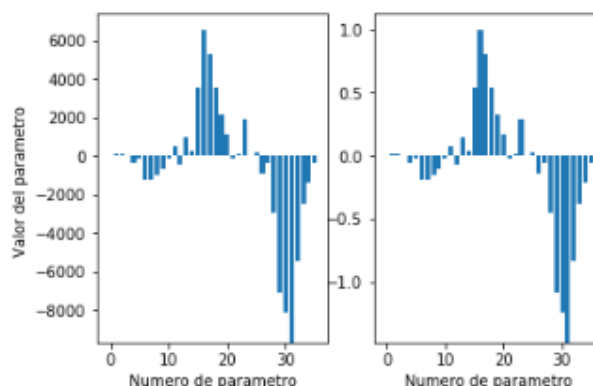
Si puede ser suficiente para distinguirla, dado que ese año es el eje donde se tiene una mayor dispersión de los datos



(1 punto) Normalmente el PCA se aplica a las columnas de una base de datos. ¿Qué interpretación se le puede dar a los resultados de aplicar el PCA a las filas de la base de datos?

Al aplicar el ACP a las filas de la base de datos, se estaría buscando cuales son las categorías en donde se presenta la mayor dispersión de los datos, a diferencia del análisis de columnas dónde se busca cuales son los años con mayor dispersión, para el análisis de columnas se proyectan los años con menos dispersión en los años con mayor, en el caso de análisis de filas se proyectan las categorías de menor sobre las de mayor dispersión con el fin de reducir la dimensionalidad de los datos.

6. 2 punto) Suponga que se requiere diseñar un modelo basado en regresión logística, y al momento de analizar los parámetros de dos modelos diseñados con los mismos datos se obtiene las siguientes figuras del modelo 1 y modelo 2 respectivamente.



¿Qué modelo conviene utilizar y por qué? o ¿Ambos tendrán el mismo desempeño?

Los dos modelos tendrían el mismo desempeño, parece ser que el modelo 2 está estandarizado, pero aún así los coeficientes están asignados igual en los dos modelos proporcionalmente, los parámetros de mayor orden siguen teniendo una ponderación mayor en el modelo. Podría ser conveniente aplicar una regresión regularizada para evitar ese problema.

7. 2 puntos) Se pretende diseñar un sistema de diagnóstico automático, donde el objetivo es identificar si una persona contrajo VIH. Un diagnóstico positivo del sistema permite aplicar una segunda fase exámenes médicos para determinar si es curable o no. Se propusieron dos modelos logísticos de los cuales las medidas de desempeño son las siguientes:

Tabla 2. Resumen de medidas estadísticas

Medida	Modelo 1	Modelo 2
Exactitud	0.99	0.98
Precision	1.0	0.5
Recall	0.5	1.0

Considerando que la base de datos que fue utilizada para el entrenamiento consta de 10000 pacientes de los cuales 200 están realmente enfermos. ¿Qué modelo es el que conviene utilizar, si se sabe que para una persona sana que se le haga la segunda ronda de estudios no sufre ninguna afectación en su salud? Justifique su respuesta

La exactitud no es un buen criterio en este caso, dado que el número de enfermos es pequeño en relación al número total de pacientes. Así el modelo clasifique que todos están sanos, la exactitud seguiría siendo alta.

En este caso es mejor enfocarse en el recall, si la precisión no es tan buena, la consecuencia sería que el modelo clasificaría como enfermo a pacientes que no lo están.

Al tener un mal desempeño en el recall no se estarían detectando pacientes que si están enfermos, causando posiblemente su muerte, por lo tanto es mejor utilizar para este caso el modelo 2.

