*UNIVERSIDAD DE SANTIAGO DE CHILE*
*FACULTAD DE INGENIERÍA*
*DEPTO. DE INGENIERÍA INFORMÁTICA*

# PEP 2
## Análisis de Datos

*Prof: Max Chacón*
*Diciembre 2015*

1. Un analista de marketing tiene la hipótesis que existe una fuerte relación (más del 50%) entre las personas que están suscritas al diario el mercurio y estrategia, con las personas que toman sus vacaciones en el extranjero y en el sur del país. Para comprobar esta hipótesis posee las bases de datos de las suscripciones a los periódicos y de lan.com. Muestre una base de datos de no mas de 10 ejemplos que compruebe la hipótesis. Considerando un soporte de mas del 10%, cuando el universo esta constituido por las personas comunes a las tres bases de datos. (1)

2. El profesor de electro-fisiología cognitiva de la U de Mancheter Wael El-deredy, estudia los trastornos bipolares mediante experimentos de juegos de azar. Para modelar estos trastornos plantea un modelo de toma de decisiones Bayesiano. Aquí las probabilidades a priori corresponden a las expectativas que tiene los sujetos de ganar y las probabilidades a posteriori es lo que determina finalmente si se apostará o no. Para probar su hipótesis genera un experimento con dos grupos de sujetos (bipolares y sanos), los que realizan apuestas en un juego al azar, según el nivel de dinero apostado. Del experimento se obtienen las siguientes probabilidades a priori en función del dinero apostado para sujetos sanos como: $Ps = \left(1 - \frac{d}{1020}\right)$, las probabilidades a priori para los bipolares son: $Pb = \left(1 - \frac{d}{1990}\right)$. Además se tiene la siguiente tabla de verosimilitudes.

| Nivel de la apuesta (*d*) | Verosimilitud Bipolares | Verosimilitud normales |
|---|---|---|
| £ 1000 | 0.9 | 0,94 |
| £ 900 | 0.93 | 0,85 |
| £ 800 | 0,87 | 0,89 |
| £ 700 | 0,86 | 0,86 |
| £ 500 | 0,95 | 0,95 |
| £ 400 | 0,91 | 0,88 |
| £ 300 | 0,89 | 0,87 |
| £ 200 | 0,87 | 0,89 |
| £ 100 | 0,92 | 0,91 |

Determine la decisión de apostar para cada nivel de apuesta, tanto para bipolares y como para sanos. Con estos resultados identifique si hay diferencias entre bipolares y sanos. Si estas diferencias existen, indique a su juicio cuales son las causas de estas diferencias, en términos de las expectativas de los sujetos. (1.5)

3. Para evaluar la diabetes Mellitus existen dos variables de interés, el nivel sobrepeso y el nivel de glucosa en la sangre del paciente. De un hospital se tiene la siguiente tabla.

| Nivel de Sobrepeso | Glucosa en mg/100 ml sangre | Diabético |
|---|---|---|
| Normal | 70 | No |
| Normal | 80 | No |
| Sobrepeso | 90 | No |
| Normal | 100 | No |
| Obeso | 110 | No |
| Sobrepeso | 120 | Si |
| Sobrepeso | 130 | Si |
| Obeso | 140 | Si |
| Obeso | 150 | Si |

Usando un clasificador Bayesiano ingenuo, determine si un paciente obeso con 90 mg/ml de glucosa en sangre, será diabético o no. (1,5)

4. . La SIAT de Carabineros de Chile posee una reducida base de datos, en la cual dos variables son importantes para determinar la accidentabilidad en los caminos rurales que están en manteción en Chile. Una es la edad del conductor (mayor o menor que 25 años) y el estado del camino durante la manteción (bueno o malo). Se sabe que la probabilidad de accidente en estos caminos es de 40%. Además se sabe que la probabilidad que los caminos en reparación estén en buen estado es de 30% y los probabilidad que los conductores sean menores de 25 años es de 40%. Usando el método de árboles de decisión propuesto por Quinlan, determine que variable es más relevante para determinar la accidentabilidad, cuando se sabe que las probabilidades conjuntas están dadas por la siguiente tabla.

| | *Camino mal estado* | *Camino buen estado* | *Mayor de 25* | Menos de 25 |
|---|---|---|---|---|
| Accidente | 2/5 | 1/5 | 6/10 | 0 |

(2)

***Formulas***

- Ingenuo: $p(c_i / a_i) = p(c_i)\prod_j p(a_j / c_i)$ ; para $a_i$=$x$ continua $p(x / c_i) = e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \Big/ \sigma\sqrt{2\pi}$

- Árboles de decisión: $Ganancia(V) = -\sum_i p(c_i) ldp(c_i) + \sum_j p(v_j) \sum_i p(c_i / v_j) ldp(c_i / v_j)$;

$RazonGanancia(V) = -Ganancia(V) \Big/ \sum_j p(v_j) ldp(v_j)$.