DuocUC | ESCUELA DE INFORMÁTICA Y TELECOMUNICACIONES

Informe Actividad 1.3.6

FMY-001V - Fundamentos de Machine Learning

Maximiliano Briones Troncoso - 07/04/2021

1. ¿Qué muestran en cada columna?

- **Geography**: Lugar de la persona a quien se le tomaron los datos
- **Year**: Año de toma de datos
- **Strata**: Estrato, tipos de categorización de personas, corresponde a: Total de población, Edad, Nivel Educacional, Ingresos, Raza-Etnia, Sexo.
- **Strata Name**: Nombre de estrato, subgrupos de estratos
- **Percent**: Porcentaje de prevalencia de diabetes
- **Lower 95% CL**: Cota inferior de intervalo de confianza de 95%. Corresponde a promedio - desviación estándar
- **Upper 95% CL**: Cota superior del intervalo de confianza de 95%. Corresponde a promedio + desviación estándar.
- **Standar Error**: Error estándar

2. ¿Cuántos años de información muestran?

Desde 2012 a 2018 ambos inclusives. 7 años.

3. Por cada año ¿qué datos se muestran?

Los valores correspondientes a cada columna y su grupos.

4. ¿Podrías indicar cuál sería un posible objetivo de minería de datos en este caso?

Identificar la preponderancia de cierto grupo de personas a presentar diabetes, lo que indicaría el riesgo del mismo.

6. Si tuvieras que eliminar algún dato a nivel año, ¿cuál elegirías? ¿por qué?

Ambas cotas superior e inferior no son a priori necesarias. Ambas pueden ser calculadas del resto de datos, por lo que eliminarlas no perjudicaría en la visualización o obtención de datos.

7. ¿Se puede decir que existen datos irrelevantes?

Sí. Ambas cotas por lo expuesto anteriormente.

8. Haz un resumen de los datos por año e indica si existen “outliers”

Sí. Gracias al gráfico anterior se puede notar que en el año 2016 se presentó un valor de ***Percent*** de 24.2

9. Busca si hay “missing values”. ¿A qué año corresponden?

Usando .describe() para cada columna y analizando los outputs, no se encontraron valores vacíos