ANÁLISIS MULTIVARIANTE

Limpieza y transformación de datos

Mg. Víctor Guevara P

Universidad Nacional José Faustino Sánchez Carrión

Mayo 2023

1 Análisis previo de los datos

Normas de participación.

- Ingresar puntual
- Mantener silenciado el micrófono durante la sesión (salvo cuando se pida participación)
- Las preguntas se realizarán por el chat/en caso sea necesario se habilita el micrófono.
- Realizar las actividades encomendadas.

Análisis previo de los datos

Análisis previo de los datos

La preparación de datos incluye los procesos de obtención de datos, limpieza, ingeniería de características y análisis exploratorio. Los analistas de datos informan que estos procesos suelen consumir del 60 al 90 por ciento del tiempo de un proyecto.

Caso 1: Nivel de obesidad

Conjunto de datos para la estimación de niveles de obesidad basados en hábitos alimentarios y condición física en individuos de Colombia, Perú y México.

Objetivo:

- El informe presenta datos para la estimación de los niveles de obesidad en individuos de los países de México, Perú y Colombia, con base en sus hábitos alimentarios y condición física.
- Los datos contienen 17 atributos y 2111 registros, los registros están etiquetados con la variable de clase NObesidad (Nivel de Obesidad), que permite clasificar los datos utilizando los valores de Peso Insuficiente, Peso Normal, Sobrepeso Nivel I, Sobrepeso Nivel II, Obesidad Tipo I , Obesidad Tipo II y Obesidad Tipo III.

Revisar: https://www.sciencedirect.com/science/article/pii/S2352340919306985

Caso 1: Nivel de obesidad

```r
# Importar el conjunto de datos
obesidad_df<-read.csv("https://raw.githubusercontent.com/sombragris1/SMedicas/main/Obesidad.csv",
                      sep = ";", stringsAsFactors = TRUE, encoding = "latin1")
```

```r
# Mostramos las primeras 6 observaciones
head(obesidad_df)
```

```
##      Genero Edad Estatura Peso Fam_sobrepeso FAVC FCVC NCP          CAEC Fuma
## 1  Femenino   21     1.62 64.0            Si   no    2   3 Algunas veces   no
## 2  Femenino   21     1.52 56.0            Si   no    3   3 Algunas veces   Si
## 3 Masculino   23     1.80 77.0            Si   no    2   3 Algunas veces   no
## 4 Masculino   27     1.80 87.0            no   no    3   3 Algunas veces   no
## 5 Masculino   22     1.78 89.8            no   no    2   1 Algunas veces   no
## 6 Masculino   29     1.62 53.0            no   Si    2   3 Algunas veces   no
##   CH2O SCC FAF TUE          CALC         Transporte            N_Obesidad
## 1    2  no   0   1         Nunca Transporte público           Peso normal
## 2    3  Si   3   0  Algunas veces Transporte público           Peso normal
## 3    2  no   2   1 Frecuentemente Transporte público           Peso normal
## 4    2  no   2   0 Frecuentemente            Caminar  Nivel de sobrepeso I
## 5    2  no   0   0  Algunas veces Transporte público Nivel de sobrepeso II
## 6    2  no   0   0  Algunas veces          Automóvil           Peso normal
```

Caso 2: Predicción del cáncer de mama

El conjunto de datos de datos para esta práctica corresponde a una muestra de distintas características de pacientes que se sometieron a la detección de cáncer de mama.

Las características se calculan a partir de una imagen digitalizada de una aspiración con aguja fina (FNA) de una masa mamaria. Describen las características de los núcleos celulares presentes en la imagen.

url: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.707&rep=rep1&type=pdf

Caso 2: Predicción del cáncer de mama

Información de atributo:

- radio (media de las distancias desde el centro a los puntos en el perímetro)
- textura (desviación estándar de los valores de escala de grises)
- perímetro
- área
- suavidad (variación local en las longitudes de los radios)
- compacidad (perímetro**2/área-1.0)
- concavidad (severidad de las porciones cóncavas del contorno)
- puntos cóncavos (número de porciones cóncavas del contorno)
- simetría
- dimensión fractal (“aproximación de la línea de costa” - 1)
- Diagnóstico (M= maligno, B= benigno)

Caso 2: Predicción del cáncer de mama

```
# Importar el conjunto de datos
cancer<-read.csv("https://raw.githubusercontent.com/VictorGuevaraP/Estadistica-R/master/cancer.csv",
                 sep = ";", stringsAsFactors = T)
```

```
#Mostrar los primeros registros
head(cancer)
```

```
##   radio textura perimetro   area Suavidad compacidad concavidad puntos_conc
## 1 17.99   10.38    122.80 1001.0  0.11840    0.27760     0.3001     0.14710
## 2 20.57   17.77    132.90 1326.0  0.08474    0.07864     0.0869     0.07017
## 3 19.69   21.25    130.00 1203.0  0.10960    0.15990     0.1974     0.12790
## 4 11.42   20.38     77.58  386.1  0.14250    0.28390     0.2414     0.10520
## 5 20.29   14.34    135.10 1297.0  0.10030    0.13280     0.1980     0.10430
## 6 12.45   15.70     82.57  477.1  0.12780    0.17000     0.1578     0.08089
##   simetria dim_fractal peor_radio peor_perimetro peor_area peor_concav
## 1   0.2419     0.07871      25.38          184.60    2019.0      0.7119
## 2   0.1812     0.05667      24.99          158.80    1956.0      0.2416
## 3   0.2069     0.05999      23.57          152.50    1709.0      0.4504
## 4   0.2597     0.09744      14.91           98.87     567.7      0.6869
## 5   0.1809     0.05883      22.54          152.20    1575.0      0.4000
## 6   0.2087     0.07613      15.47          103.40     741.6      0.5355
##   peor_pun_conc diagnostico
## 1        0.2654           M
## 2        0.1860           M
## 3        0.2430           M
```

Limpieza de Datos

Limpieza de Datos

La limpieza de los datos es uno de los tres mayores problemas en el data warehousing Ralph Kimball La limpieza de datos es el problema número uno en el data warehousing - DCI survey.

Tareas en la limpieza de datos

Estimar valores perdidos. Identificar outliers y suavizar datos. Corregir datos incosistentes. Resolver redundancias causadas por la integración

Valores missing

Impacto de los valores faltantes

- 1% Datos faltantes
- 1-5% Manejable
- 5-15% Utilizar metodos sofisticados
- MÁS DEL 15% Interpretación perjudicial

Valores missing

Solución missing

- Ignorar la variable.
- Rellene el valor perdido manualmente.
- Usa una constante global para completar el valor faltante.
- Use una medida de la tendencia central para que el atributo
- complete el valor faltante
- Utilice el valor más probable

Mecanismos y aleatoriedad de datos faltantes

Missing completamente al azar (MCAR)

Los datos faltantes son MCAR cuando la probabilidad de que falten datos en una variable no está relacionada con ninguna otra variable medida y no está relacionada con la variable con valores perdidos. En otras palabras, la falta de la variable es completamente no sistemática.

Missing completamente al azar (MCAR)

Ejemplo 1:

Cuando faltan datos para los encuestados de los cuales se perdió su cuestionario en el correo. Esta suposición se puede probar separando los casos faltantes y los casos completos y examine las características del grupo. Si las características no son iguales para ambos grupos, el supuesto de MCAR no se cumple.

Ejemplo 2:

En un estudio donde se tiene las variables ingreso y edad. Estaremos bajo un modelo MCAR cuando al analizar conjuntamente edad e ingresos, suponemos que la falta de respuesta en el campo ingresos es independiente del verdadero valor de los ingresos y la edad.

Mecanismos y aleatoriedad de datos faltantes

Missing al azar (MAR)

Los datos faltantes faltan al azar (MAR) cuando la probabilidad de que faltan datos en una variable está relacionada con alguna otra variable medida en el modelo, pero no con el valor de la variable con valores faltantes en sí.

Ejemplo 1:

Si suponemos que los ingresos de una persona son independientes de los ingresos del miembro del hogar pero puede depender de la edad estaremos bajo un modelo MAR.

Mecanismos y aleatoriedad de datos faltantes

Missing no al azar (MNAR)

La probabilidad de que una instancia tenga un valor faltante en un atributo depende de los valores faltantes en el conjunto de datos. Ocurre cuando las personas entrevistadas no quieren revelar algo muy personal acerca de ellas. El patron de valores faltantes no es aleatorio.

Ejemplo:

En el ejemplo anterior, se obtiene que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además de poder depender de otros factores

Missing - consideraciones

- Para conjuntos de datos con un bajo porcentaje de valores faltantes el mecanismo se puede considerar MCAR.
- Para conjuntos de datos con un alto porcentaje de valores faltantes el mecanismo se puede considerar MNAR.
- En muchas aplicaciones lo prudente sería considerar distintos modelos plausibles para el mecanismo de no respuesta y realizar un análisis de sensibilidad de las estimaciones.

Tratamiento de missing

Eliminación de casos

Consiste en eliminar las observaciones o variables que tengan los datos perdidos. Solamente debe realizarse si es poco el porcentaje de observaciones a eliminar y si es posible asumir que los valores faltantes provienen de un proceso MCAR.

Tratamiento de missing

Imputar missing

Reemplazar el valor perdido con un valores conocidos.Hay variedad de métodos, desde opciones sencillas (reemplazar por la media o mediana) hastas otras más complejas (modelos de regresión).

Tratamiento de missing

Mantener missing

A veces es factible analizar la información por separado. Por ejemplo, en algunas situaciones que se pueden manejar la estimación de parámetros en presencia de valores faltantes

Imputación

Imputación simple

Los valores faltantes son reemplazados con valores estimados basados en la información disponible. Se reemplaza cada valor faltante por un un solo número. * Imputación por la media: Reemplazar con la media de la columna. * Imputación por la mediana: Reemplazar con la mediana de la columna. * Imputación por la moda: Reemplazar con la moda de la columna.

Imputación

Imputación de KNN

El método consiste en que para cada valor faltante se encuentran las k-observaciones o instancias que están más cercanas considerando las otras variables. se reemplaza el valor faltante de la siguiente manera:

- Si la variable es categórica se reemplaza por la moda de las k-observaciones más cercanas.
- Si la variable es numérica se reemplaza por la media de las k-observaciones más cercanas.

Imputación

Metodos de regresión

El método consiste en estimar un modelo de regresión en función a las otras variables. Luego se reemplaza el valor faltante utilizando el modelo de regresión.

Imputación

Imputación missForest

Los método no paramétrico no hace suposiciones explícitas sobre la forma funcional de f(x). En su lugar, buscan una estimación de f (x) que se acerque lo más posible a los puntos de datos sin ser demasiado brusca. Dichos enfoques pueden tener una gran ventaja sobre los enfoques paramétricos: al evitar la suposición de una forma funcional particular para f(x), tienen el potencial de ajustarse con precisión a un rango más amplio de formas posibles para f(x).

Transformación de Datos

Algunas técnicas son:

- Suavizamiento: Remover datos ruidosos
- Normalización:
  - Normalización
  - Normalización min-max
  - Normamalización por escalamiento decimal
- Construcción de Atributos: Nuevos atributos construidos basados en los anteriormente especificados.

Discretización

- Es un método que transforma datos cuantitativos en cualitativos Algunas metodologias solo aceptan atributo categóricos. El proceso de aprendizaje es frecuentemente menos eficiente cuando las datos son solo cuantitativos