Juan Pablo Botero Suaza

Juan Carlos Ghiringhelli

41024416Y

Tipología de vida y ciclo de los datos

Practica 2 – Limpieza y validación de los datos

Contenido

[Practica 1 – Web scraping 1](#_Toc6085298)

[1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información. 3](#_Toc6085299)

[2. Definir un título para el dataset. Elegir un título que sea descriptivo. 4](#_Toc6085300)

[3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido). 4](#_Toc6085301)

[4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente. 5](#_Toc6085302)

[5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido. 6](#_Toc6085303)

[6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay). 6](#_Toc6085304)

[7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. 7](#_Toc6085305)

[8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: 8](#_Toc6085306)

[9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. 9](#_Toc6085307)

[10. Dataset. Presentar el dataset en formato CSV. 9](#_Toc6085308)

## 1. Descripción del dataset

El conjunto de datos elegido registra, en el contexto de consultas para diagnosticar diabetes, datos predictores médicos provenientes de una batería de pruebas realizadas a mujeres de la India.

Como suele ocurrir con datos predictivos médicos, es posible anticipar el diagnóstico de la condición médica para pacientes sin necesidad de realizar la prueba específica. Para casos de alto riesgo se podrían tomar las medidas adecuadas como notificación al paciente o tomar precauciones ante el ingreso del paciente a una operación o una emergencia. También sirve para comprender mejor las causas de la condición y la correlación de diferentes valores con la posibilidad de padecerla.

El contenido se descargó del siguiente enlace de [Kaggle](https://www.kaggle.com/uciml/pima-indians-diabetes-database) y contiene nueve atributos:

1. Pregnancies: Cantidad de veces que la paciente estuvo embarazada.
2. Glucose: Concentración de glucosa en sangre dentro de dos horas de una prueba de resistencia oral a la glucosa, medido en miligramo por decilitro.
3. BloodPressure: Presión de sangre, en milímetros por mercurio, una medida médica equivalente a la presión de una columna de mercurio de un mm de alto a 0℃ a una atmosfera.
4. SkinThickness: grosor de la piel en mm en la zona del pliegue del tríceps.
5. Insulin: Insulina administrada por suero en la ultima hora, en unidades por mililitro.
6. BMI: de *Body Mass Index*, índice de masa corporal, medido en (kilos/altura)².
7. DiabetesPedigreeFunction: función de la condición presente en parientes para asignar una probabilidad genética de heredarla.
8. Age: edad en años.
9. Outcome: Presencia de la condición diabetes, siendo 1 positivo y 0 negativo. El conjunto presenta 268 casos positivos sobre un total de 768.

## 2. Integración y selección de los datos de interés a realizar

El conjunto de datos es independiente y por lo tanto no es necesario integrarlo con otros. Dada la complejidad y falta de detalle de la función de pedigre de diabetes no utilizaremos este atributo. Los demás datos son simples de comprender y potencialmente valiosos por lo que los mantendremos.

## 3. Limpieza de datos

### 3.1 Tratamiento de ceros o elementos vacíos

El conjunto de datos no presenta campos vacíos ni nulos, por lo que no será necesario tratarlos en este aspecto.

El atributo de edad, el rango de edades va de 21 años a 81, por lo que parece correcto y no haremos nada con estos datos.

Los atributos de cantidad de embarazos y insulina en sangre presentan casos con 0, lo cual es coherente, tanto por no haber estado embarazada nunca la persona como por no consumir insulina. Si bien puede resultar de una falta de datos, se asumirá que todos estos casos son válidos.

Los valores de glucosa, presión de sangre, grosor de la piel en el tríceps, e índice de masa corporal presentan valores en cero. Consideraremos que esto denota una falta de datos, ya que en cualquiera de estos casos el paciente estaría muerto o herido de gravedad. Para cada caso:

* Glucosa: el nivel de glucosa en sangre, a diferencia de la presión, no tiene un valor constante que refleja un buen estado de salud. Si bien los extremos siempre son peligrosos, esta presenta normalmente una gran varianza dependiendo de la última vez que se consumió glucosa, que cantidad, en que forma, con que alimentos y dependiendo del metabolismo y actividad inmediata. Dentro de los valores analizados, es el que presenta más variabilidad y valores extremos, por lo que tomaremos el valor mediano para imputar.
* Presión de sangre: Muchos valores presentan una variación muy alta o baja, y ni el promedio ni la mediana representan realmente un valor neutro, lo que haría que puedan tener un peso sobre la predicción siendo un valor desconocido. Utilizando registros médicos, seleccionaremos el valor 105, valor considerado normal para mujeres de cualquier edad, dando por supuesto que si no se sabe el valor es porque no se consideró importante anotarlo o medirlo.
* Grosor de piel: Los valores presentan un rango mediano y valores bien distribuidos, por lo que consideraremos el atributo como simétrico y usaremos el promedio.
* Índice de masa corporal (IBM): similar caso al grosor de piel, usaremos el promedio.

3.2 Identificación y tratamiento de valores extremos

## 4. Análisis de los datos

### 4.1 Selección de grupos de datos a analizar

### 4.2 Comprobación de la normalidad y homogeneidad de la varianza

### 4.3 Pruebas estadísticas

## 5. Representación de los resultados

## 6. Conclusiones

## 7. Código

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación Previa | JPBS, JCG |
| Redacción de las respuestas | JPBS, JCG |
| Desarrollo de código | JPBS, JCG |

## Bibliografía

* Calvo M., Subirats L., Perez D. (2019). Introduccion a la limpieza y analisis de los datos.
* Editorial UOC.
* Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.
* Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*.
* Morgan Kaufmann.
* Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme*
* *Scores.* Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
* Peter Dalgaard (2008). *Introductory statistics with R*. Springer Science & Business Media.
* Wes McKinney (2012). *Python for Data Analysis*. O’Reilley Media, Inc.
* Tutorial de Github https://guides.github.com/activities/hello-world.