Análisis de Datos I

Segunda fase de proyecto

Moisés Monge Cordonero - Joshua Cervantes Artavia

Datos

Descripción general de los datos

Los datos fueron obtenidos del repositorio de Machine Learning de la Universidad de California Irving, la cual se puede accesar en el siguiente link: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease y este dataset tiene por nombre Chronic_Kidney_Disease el cual contiene una serie de variables relacionadas a 400 pacientes de los cuales 250 presentan enfermedad del riñón y 150 no. Esta data fue accesada el día 28 de Agosto del 2023. Los individuos en cuestión responden a pacientes de Apollo Hospitals en Karaikudi, Taminaldu, India, durante un proceso de recolección de datos de al alrededor de 2 meses. Los datos fueron recogidos a partir de Julio del año 2015.

Variables

El dataset se compone por total de 25 variables de las cuales 11 son numéricas y 14 nominales. Las variables en cuestión son:

- age: Edad en años (numérica)
- bp: presión sanguínea en mm/Hg (numérica)
- sg: gravedad específica con valores entre (1.005, 1.01,1.015,1.020.1.025) (nominal)
- al: albúmina con valores en (0,1,2,3,4,5) (nominal)
- su: azúcar con valores en (0,1,2,3,4,5) (nominal)
- rbc: células de sangre roja con valores en (normal, anormal) (nominal)
- pc: células pus con valores en (normal, anormal) (nominal)
- pcc: grupos de cédulas de pus con valores en (presente, no presente) (nominal)
- ba: bacteria con valores en (presente, no presente) (nominal)
- bgr: glucosa en sangre aleatoria en mgs/dl (numérica)
- bu: urea en sangre en mgs/dl (numérica)
- sc: creatinina sérica en mgs/dl (numérica)
- sod: sodio en mEq/L (numérica)
- pot: potasio en mEq/L (numérica)

- hemo: hemoglobina en gms (numérica)
- pcv: hematocrito en % (numérica)
- wc: conteo de células blancas en cells/cumm (numérica)
- rc: conteo de células rojas en millones/cmm (numérica)
- htn: hipertensión con valores en (sí, no) (nominal)
- dm: diabetes mellitus con valores en (sí, no) (nominal)
- cad: enfermedad de arteria coronaria con valores en (sí, no) (nominal)
- appet: apetito con valores en (bueno, pobre) (nominal)
- pe: edema del pie con valores en (sí, no) (nominal)
- ane: anemia con valores en (sí, no) (nominal)
- class: si presenta o no enfermedad crónica del riñón -ckd-con valores en (ckd, notckd)

A las variables mencionadas en cuestión, se le han añadido cuatro variables relativas adicionales que se han construido, las cuales corresponden a las relaciones: hemo/wbcc, hemo/rbcc, bu/bgr, sod/bp, denotadas con un punto en lugar /. Como variable complementaria se considerará la variable class.

Para efectos de la reducción de la dimensión a efectuar, se consideran únicamente las variables numéricas.

Data faltante

La data presenta algunos datos faltantes, para esto primeramente se han contabilizado aquellas observaciones que tuvieran 4 datos faltantes o más y se han excluido, dejando un total de 357 observaciones y posteriormente, sobre dichas 357 observaciones se ha realizado una imputación de datos por medio de un KNN.

Resultados

Se ha decido implementar como método de reducción el análisis de componentes principales dada la cantidad de variables con las que se cuente el mismo puede ser útil al tener un total de 16. Al implemente el método se obtienen los resultados mostrados mostrados en el cuadro 1, y el cuadro 2.

En el cuadro 1 se puede observar que se alcanza el 75.88 % de inercia acumulada para los primeros 5 autovalores, por lo que podría decirse tomarse estos 5 para así reducir la dimensión y tener una buena aproximación de las variables originales, sin embargo, puede ser necesario observar algún otro diagnóstico como puede ser el gráfico del codo que es comentado más adelante.

Del cuadro 2 se puede destacar que las variables que tienen una mejor representación sobre el plano principal son las de hemoglobina (hemo), hematrocrito (pcv) y la relativa urea en sangre respecto a glucosa en sangre. La de potasio es la que muestra la peor representación, repecto a las demás estas presentan una comunalidad sobre el eje principal superior al 0.2.

De la figura 1 se puede destacar el gran quiebre que se presenta posterior a la primera componentes principal, de tal forma que este es el primer codo, sin embargo dados los niveles de inercia explicada comentados anteriormente se podría no decidir cortar aquí el modelo. Como se puede observar seguido de la quinta componente se obtiene que la inercia explicada se mantiene casi constante por lo que podría decidirse tomar únicamente las primeras 5 componentes principales. De tal forma que

Cuadro 1
Autovalores e inercia explicada

Valor propio	Valor	Porcentaje de inercia	Porcentaje de inercia acumulado
1	5.44	36.20	36.20
2	1.93	12.84	49.04
3	1.60	10.64	59.68
4	1.28	8.53	68.21
5	1.15	7.67	75.88
6	1.01	6.69	82.57
7	0.83	5.49	88.06
8	0.71	4.74	92.81
9	0.51	3.39	96.20
10	0.21	1.43	97.62
11	0.15	1.02	98.64
12	0.09	0.63	99.27
13	0.06	0.40	99.67
14	0.04	0.25	99.92
15	0.01	0.08	100.00

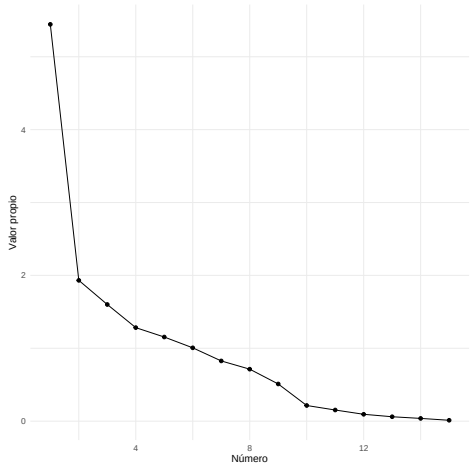
Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

Cuadro 2
Comunalidades

Variable	Comunalidad
age	0.20
bp	0.24
bgr	0.34
bu	0.77
sc	0.63
sod	0.31
pot	0.17
hemo	0.79
pcv	0.78
wbcc	0.53
rbcc	0.66
hemo.wbcc	0.73
hemo.rbcc	0.09
bu.bgr	0.70
sod.bp	0.37

Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

Figura 1
Gráfico del codo para análisis de componentes principales



Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

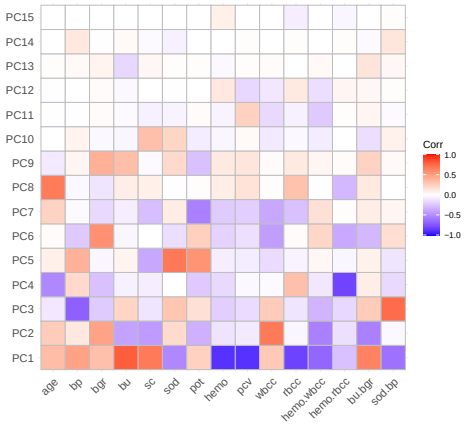
esto también permita reducir de manera significativa la cantidad de variables, ya que como se observa posterior a la quinta componente principal existen más codos.

En la figura 2 se puede observar la correlación entre las componentes principales y las variables originales. Se puede observar que en este caso la hemoglobina y el hematocrito que son los con el valor más de calidad de representación en el plano principal tienen una correlación negativa cercana -1 respecto a la componente principal 1. Un aspecto a destacar es que las componentes principales con el índice más alto como es de esperarse tienen una menor correlación respecto a las variables originales y en su gran mayoría las correlaciones más altas se obtienen las primeras componentes principales.

Ahora realizando el gráfico de circulo de correlaciones el cual se muestra en la figura 3 donde se puede observar que en su gran mayoría todas las variables se encuetran representadas por estas dos componentes principales. Además cabe destacar el hecho de que existen variables que se ven altamente relacionadas y otras que se ven claramente representadas en este plano principal al estar cerca de la frontera del círculo de correlaciones.

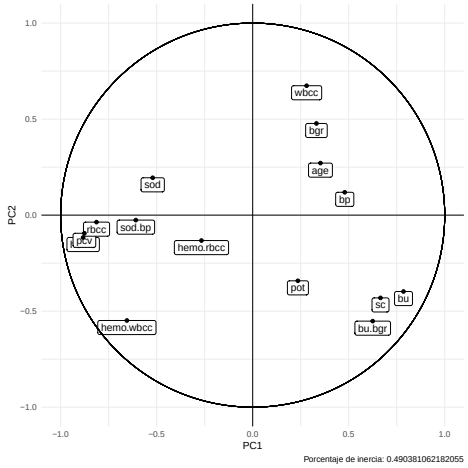
En esta primera bitácora no se ha realizado ningún método de clasificación automática que permitiera otorgar una etiqueta a cada observación y en la base original no se cuenta con una etiqueta previamente asignada, lo que no permite destacar algo de cada una de variables y que permita visualizar de mejor forma lo que se muestra en la figura 4. Sin embargo, se cuenta con la variable nominal de su, en este caso se agregan colores de esto de tal forma que se logre observar si existen clusters. En la figura 5 se puede observar que en su gran mayoría las observaciones corresponden a azúcar 0, además se encuentra que estos datos se encuentran dispersos en el tercer y cuarto cuadrante en su mayoría. Adicionado a lo anterior la mayoría de las restantes clases se ubican en el cuadrante 1.

Figura 2
Matriz de correlaciones componentes principales y variables originales



Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

Figura 3
Círculo de correlaciones (componentes principales 1 y 2)

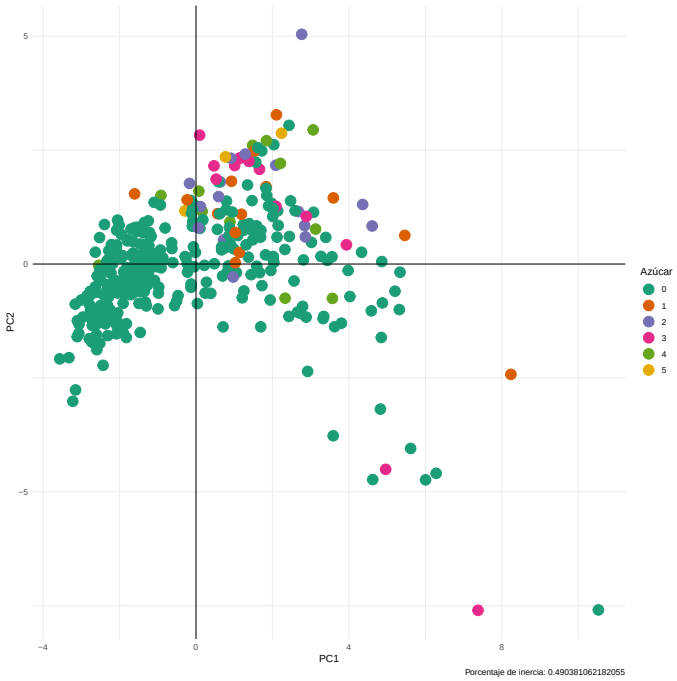


Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

Figura 4
Plano principal ACP

Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

Figura 5
Plano principal ACP con variable de azúcar



Fuente: Elaboración propia, con datos de Koklu, M. and Ozkan, I.A., (2020)

Bibliografía

Koklu, M. and Ozkan, I.A., (2020), Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques. *Computers and Electronics in Agriculture*, *174*, 105507.

Rubini,L., Soundarapandian,P., and Eswaran,P.. (2015). Chronic_Kidney_Disease. UCI Machine Learning Repository.