Crab Age Prediction

Juan Diego Espinosa Hernandez

Juan David Garcia Zapata

Pregado en estadistica
Universidad Nacional de Colombia

Noviembre 2023

UNIVERSIDAD NACIONAL DE COLOMBIA

Indíce

UNIVERSIDAD NACIONAL DE COLOMBIA

¿De dónde proviene la base?

La base de datos fue obtenida de la plataforma Kaggle y se centra en la predicción de la edad del cangrejo en función de los atributos fisicos[1].

Descripción de las variables

1. **Genero:** Macho, Hembra e Indeterminado
2. **Longitud:**Medida de la Longitud del cangrejo (en pies)
3. **Diámetro:** Medida del diametro del cangrejo (en pies)
4. **Altura:** Medida de la altura del cangrejo (en pies)
5. **Peso(1):** Medida del peso del cangrejo (en onzas)
6. **Peso(2):** Peso sin cáscara (en onzas)
7. **Peso(3):** El peso que envuelve los órganos abdominales en el profundo del cuerpo (En onzas)
8. **Peso (4):** Peso del caparazón(en libras)
9. **Edad:** Edad del Cangrejo(en meses)

UNIVERSIDAD NACIONAL DE COLOMBIA

La determinación del sexo de los cangrejos es esencial para la cría comercial de cangrejos. Los machos y las hembras tienen características físicas diferentes que pueden utilizarse para determinar su sexo. Los criadores comerciales deben poder determinar el sexo de los cangrejos para seleccionar los reproductores más sanos y fértiles, controlar la reproducción y clasificar las crías. Por lo tanto las variables que posee nuestra base de datos que se refieren a caracteristicas fisicas de los cangrejos son sumamente importante.

UNIVERSIDAD NACIONAL DE COLOMBIA

Variable de interés

Genero: Macho y Hembra
Nivel de interés:

1:Hembra
0:Macho

¿Se puede discriminar entre ambos generos?

Para responder a esta pregunta haremos una prueba de hipotesis de diferencia de vectores de medias.

$$\textbf{P.H:} \begin{cases} H_0 : \mu_H - \mu_M = 0 \\ H_1 : \mu_H - \mu_M \neq 0 \end{cases}$$

E.P:

$$X_c = [(\overline{X}_n - \overline{Y}_m - \delta_0)]' \left[\frac{1}{n}S_1 + \frac{1}{m}S_2\right]^{-1} [(\overline{X}_n - \overline{Y}_m - \delta_0)] \xrightarrow{d} \chi^2(p)$$

R.C:
Para un valor α dado , se rechaza H_0 si:

$$X_c > \chi^2_\alpha(p)$$

UNIVERSIDAD NACIONAL DE COLOMBIA

$$\overline{X}_H = \begin{bmatrix} 1.447 \\ 1.137 \\ 0.395 \\ 29.664 \\ 12.631 \\ 6.538 \\ 8.573 \\ 11.140 \end{bmatrix}; \overline{Y}_M = \begin{bmatrix} 1.404 \\ 1.099 \\ 0.379 \\ 28.095 \\ 12.243 \\ 6.110 \\ 7.994 \\ 10.723 \end{bmatrix}$$

$$X_c = 62.84716; \chi^2_{0.05}(8) = 15.50731$$

Como nuestro estadistico de prueba X_c, es mayor a el cuantil de una chi-cuadro con 8 grados de libertad, tenemos suficiente evidencia estadistica para rechazar H_0, por lo tanto se concluye que las variables tomadas son suficientes para discriminar entre ambos sexos de cangrejos.

UNIVERSIDAD NACIONAL DE COLOMBIA

Train y test

Total de datos: 2660

Total de train(70%): 1862

Total de test(30%): 798

UNIVERSIDAD NACIONAL DE COLOMBIA

Discriminante lineal de Fisher's

Fisher propone una estadística de clasificación lineal al transformar observaciones multivariadas X en observaciones univariadas $y = a'X$. El objetivo es maximizar la separación entre las observaciones univariadas derivadas de las poblaciones 1 y 2.

Regla descriminate de Fisher's
La regla de Fisher's es equivalente a la discriminante lineal, asumiendo varianzas iguales. costos y probabilidades iguales.

$$\bar{y}_0 = (\bar{x}_1 - \bar{x}_2)' * S_p^{-1} * x_0; \qquad \hat{m} = \frac{1}{2} * (\bar{y}_1 + \bar{y}_2)$$

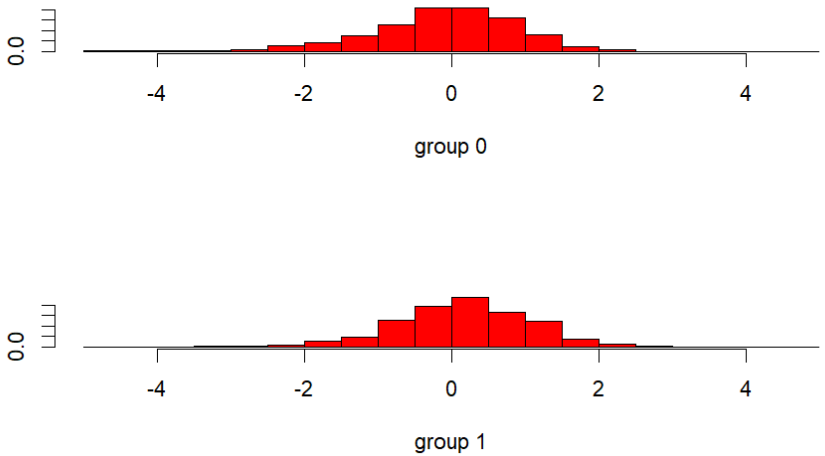
$$\begin{cases} Si: \ \bar{y}_0 \geq \hat{m} = 1(Hembra) \\ Si: \ \bar{y}_0 < \hat{m} = 0(Macho) \end{cases}$$

UNIVERSIDAD NACIONAL DE COLOMBIA

Matriz de confusión

	1	0
1	316	115
0	232	135

Tasa de mala clasificación: $100 * \frac{347}{798} = 43.4\%$



UNIVERSIDAD NACIONAL DE COLOMBIA

Rregresión logística

La regresión logística resulta útil para los casos en los que se desea predecir la presencia o ausencia de una característica o resultado según los valores de un conjunto de predictores. Es similar a un modelo de regresión lineal pero está adaptado para modelos en los que la variable dependiente es dicotómica.

$$log(\frac{p(X}{1-p(X}) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+...+\beta_p X_p}}$$

UNIVERSIDAD NACIONAL DE COLOMBIA

Regla discriminante

Una observación $\mathbf{X}_0 = [x_{01}, x_{02}, ..., x_{0p}]'$ es asignida a la poblacion π_1 si la razón de odds estimada es mayor que 1, o si $\hat{p}(X_0) > 0.5$, es decir:

$$\frac{\hat{p}(X_0)}{1 - \hat{p}(X_0)} > 1$$

$$\hat{p}(X_0) = \frac{e^{\hat{\beta}_0 + x_0' \hat{\beta}}}{1 + e^{\hat{\beta}_0 + x_0' \hat{\beta}}} > 0.5$$

UNIVERSIDAD NACIONAL DE COLOMBIA

En nuestro caso

$$\hat{p}(X_0) = \frac{e^{-2.43+0.58X_1+1.66X_2+0.77X_3-0.03X_4-0.06X_5+0.09X_6+0.04X_7+0.01X_8}}{1+e^{-2.43+0.58X_1+1.66X_2+0.77X_3-0.03X_4-0.06X_5+0.09X_6+0.04X_7+0.01X_8}}$$

Matriz de confusión

	1	0
1	316	115
0	230	137

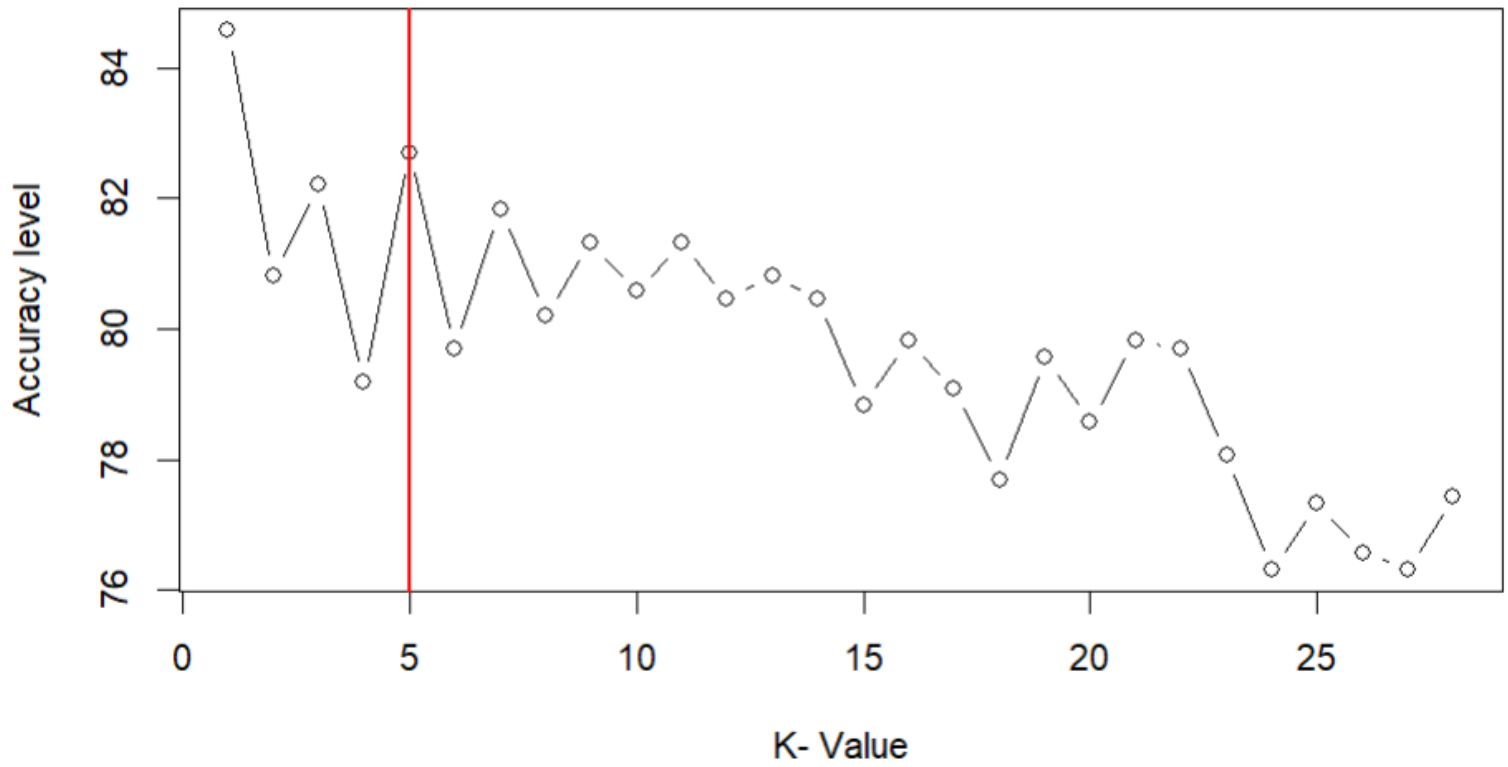
Tasa de mala clasificación: $100 * \frac{345}{798} = 43.2\%$

UNIVERSIDAD NACIONAL DE COLOMBIA

KNN (K-Nearest Neighbors)

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, partiendo de la suposición de que se pueden encontrar puntos similares cerca uno del otro.

K optimo



UNIVERSIDAD NACIONAL DE COLOMBIA

Matriz de confusión

	1	0
1	358	73
0	65	302

Tasa de mala clasificación: $100 * \frac{136}{798} = 17.3\%$

Conclusión

Al aplicar 3 metodos diferentes de clasificación Fische'r, Logistico y K-NN, donde en los dos primeros tuvimos resultados parecidos con una tasa de mala clasificación aproximadamente del 43% en cambio con K-NN nuestra tasa de mala calsificacion es de 17.3%, por lo cual el metodo que tiene un mejor rendimiento a la hora de clasificar el sexo de los cangrejos es **KNN**

UNIVERSIDAD NACIONAL DE COLOMBIA

Referencias I

[1] Kaggle. https://www.kaggle.com/datasets/sidhus/crab-age-prediction. Online;21 de noviembre. 2021.

UNIVERSIDAD NACIONAL DE COLOMBIA

Gracias por la atención

UNIVERSIDAD
NACIONAL
DE COLOMBIA