Dario Jauregui Lankenau
A00827837
9/8/21

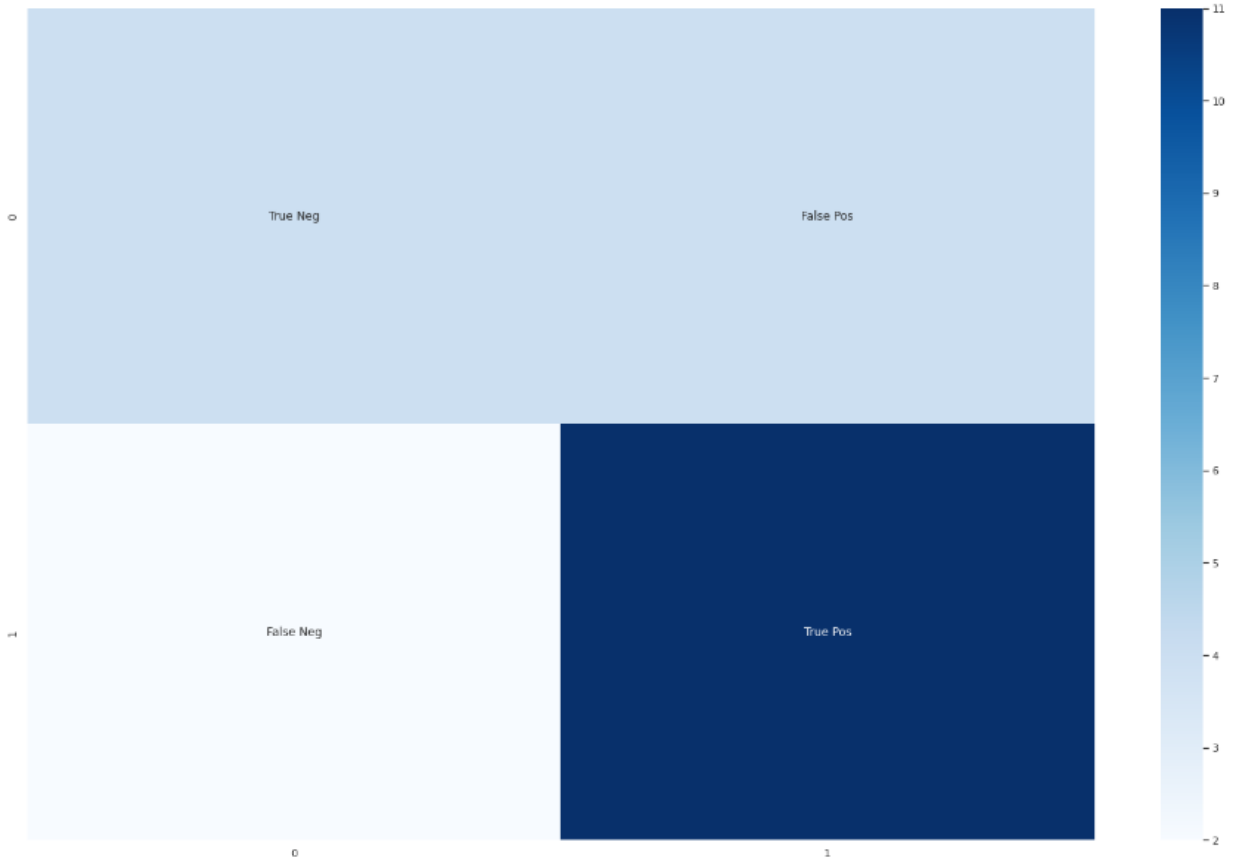
Implementación de un clasificador con **Regresión logística** para predecir el sexo de zarigüeya.

Del paquete DAAG R: "El datatset de la zarigüeya consta de nueve medidas morfométricas en cada una de las 104 zarigüeyas cola de cepillo de montaña, atrapadas en siete sitios desde el sur de Victoria hasta el centro de Queensland".

Comparación entre modelos

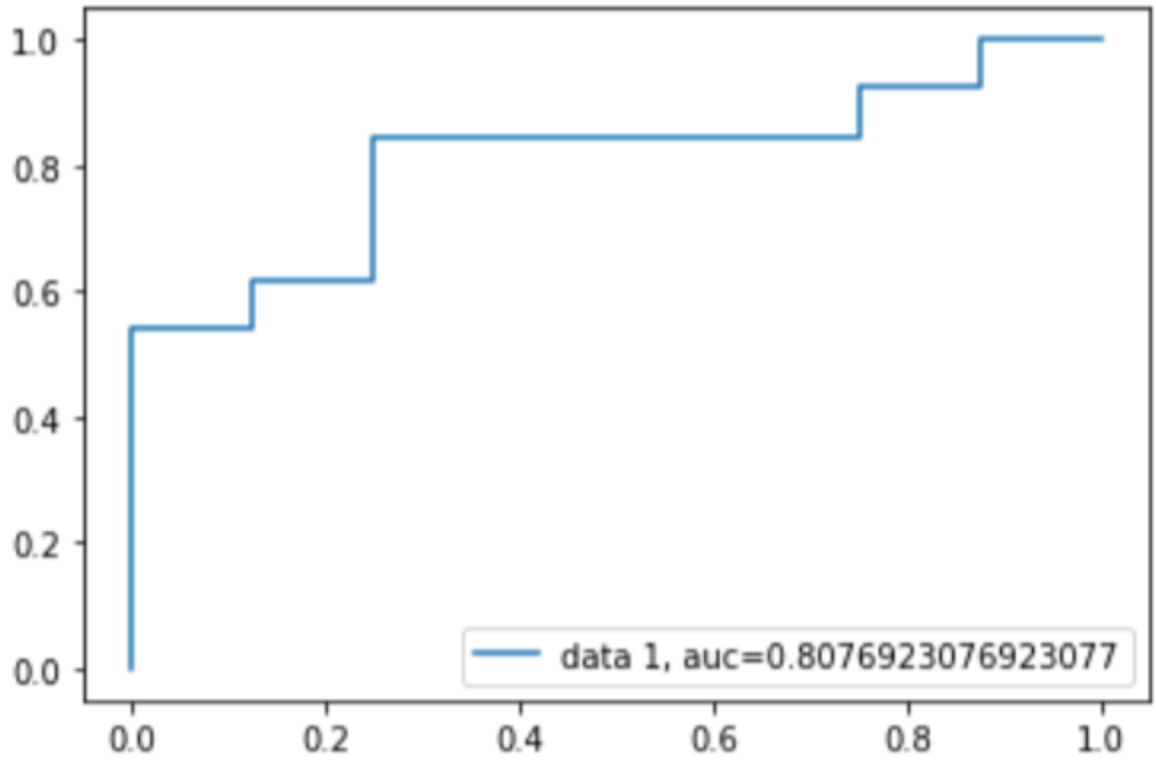
Modelo 1. (Precisión de 0.71).

Matriz de confusión



Vemos que este modelo nos da valores falsos positivos al igual que verdaderos negativos, lo cual indica una mala predicción. (Un buen modelo tendría mayores valores y por ende un color azul más oscuro en los cuadros de “True Neg” y “True Pos” mientras que los valores de “False Pos” y “False Neg” tendrían que ser cerca a 0 y de color transparente o blanco.

Gráfica ROC



La puntuación AUC (área bajo la curva) para el modelo es 0,80. La puntuación AUC 1 representa un clasificador perfecto y 0,5 representa un clasificador sin valor.

Bias y Varianza

```
Average Bias :  0.17587261904761906
Average Variance :  0.13817500000000002
```

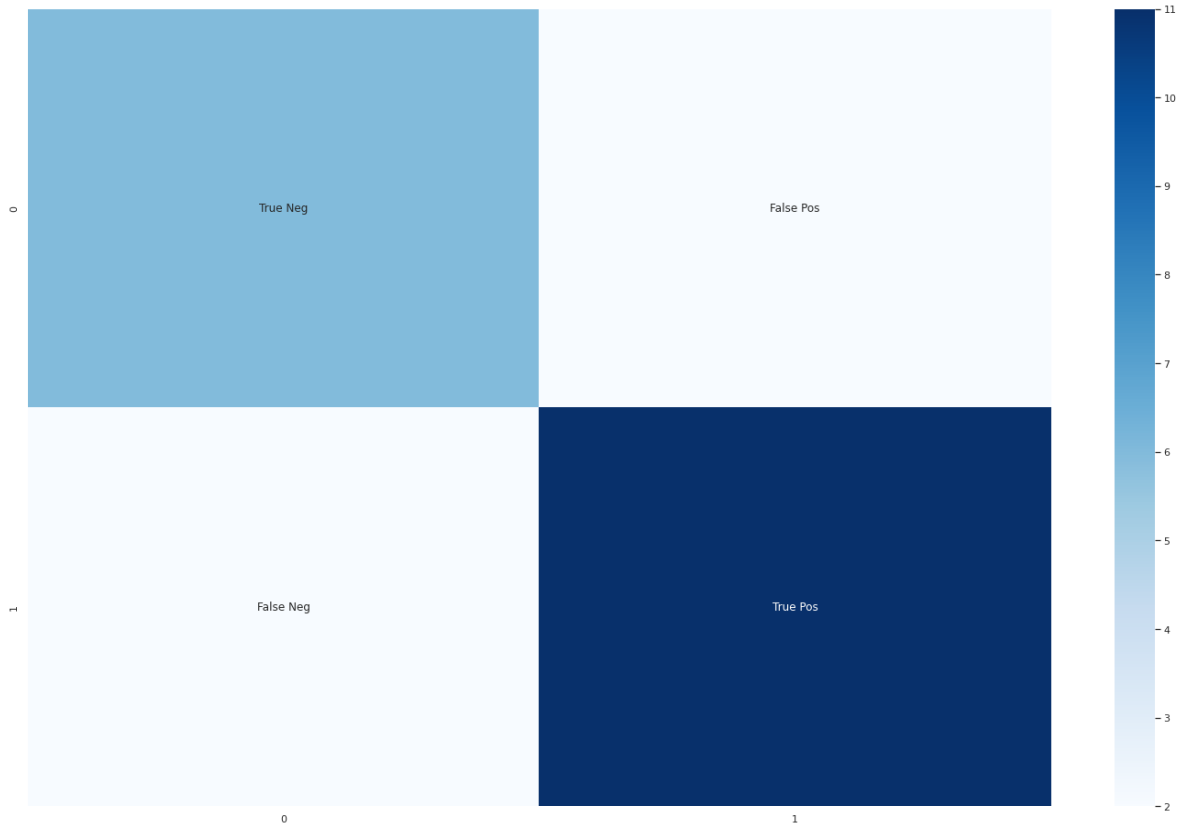
Vemos que existe un **bias alto relativo** en comparación con la varianza, más que nada debido a que estamos utilizando una regresión logística. El valor de la **varianza es relativamente poco**, ya que la suma de cuadrados es similar a diferentes datasets. Esto refleja que nuestro modelo dará buenas predicciones, más no excelentes. Además, refleja un **underfitting**, ya que el bias es alto

en comparación con la varianza, lo que significa que el modelo no es lo suficientemente complejo como para capturar bien el patrón en los datos de entrenamiento y, por lo tanto, también sufre de bajo rendimiento en datos no vistos. Esto es principalmente por que el modelo de regresión logística es lineal y no puede seguir la curva de las y's reales para las predicciones que realiza.

Modelo 2. (Precisión de 0.81).

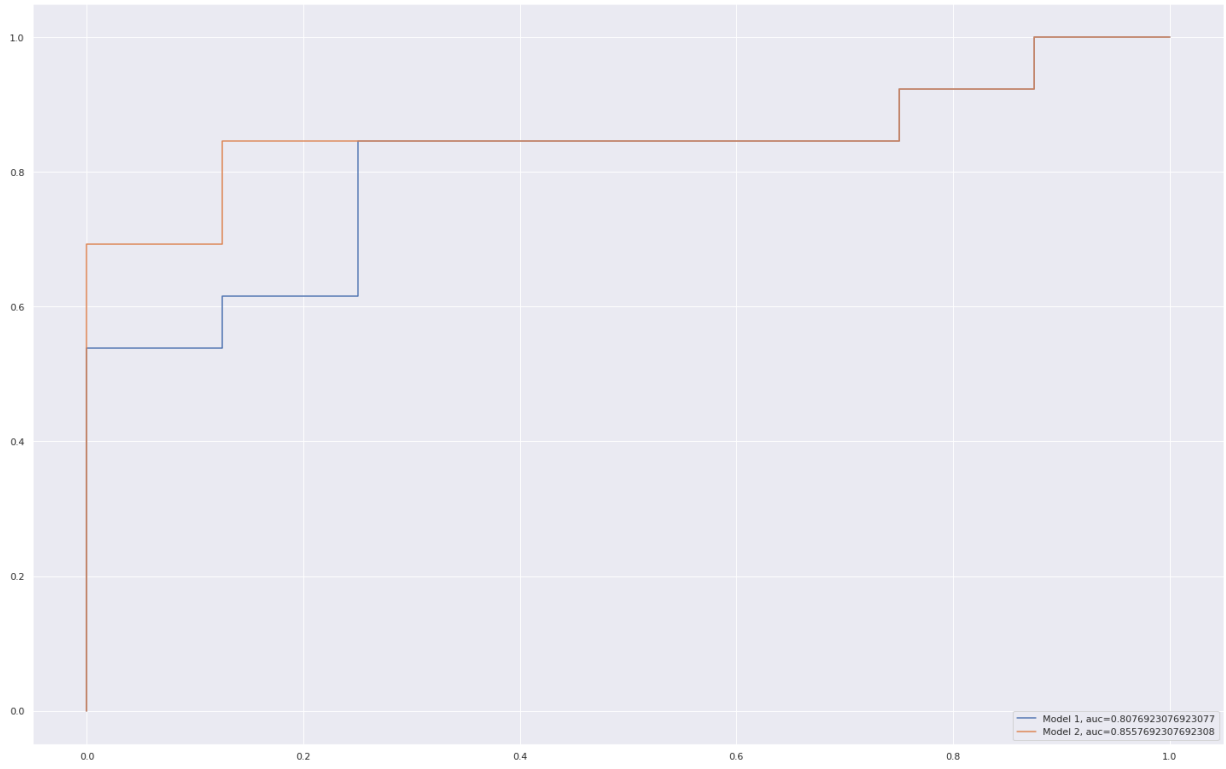
Debido a que es regresión logística, realmente no se tienen hiperparámetros críticos para ajustar. A veces, se pueden ver diferencias útiles en el rendimiento o la convergencia con diferentes solucionadores (solver). En nuestro caso, el solver es el más adecuado para el modelo y el dataset. Sin embargo, se modificó el dataset para mejorar nuestras predicciones.

Matriz de confusión



Se mejora el modelo y sus predicciones para cada cuadro de la matriz. Vemos que el valor de Falso Positivo disminuye en gran cantidad y que el valor de Verdadero Negativo incrementa en valor y por ende en su tono fuerte de color.

Gráfica ROC



La puntuación AUC para el modelo nuevo es 0,85, una mejora de 5% en comparación del modelo anterior. La puntuación AUC 1 representa un clasificador perfecto y 0,5 representa un clasificador sin valor.

Bias y Varianza

```
Model 2: Average Bias :  0.16223928571428572
Model 2: Average Variance :  0.13942738095238097
```

Vemos que **el bias en general disminuyo** y que **no existió mucho cambio en la varianza,** debido a que es un modelo de regresión logística y es un método lineal, por lo cual se dificulta acoplarse a la curva real de los datos.