Clase 18 Modelos de clasificación usando Regresión logística

Diplomado en Análisis de Datos y Modelamiento Predictivo con Aprendizaje Automático para la Acuicultura

Dra. María Angélica Rueda Calderón

Pontificia Universidad Católica de Valparaíso

01 July 2023

PLAN DE LA CLASE

1.- Introducción

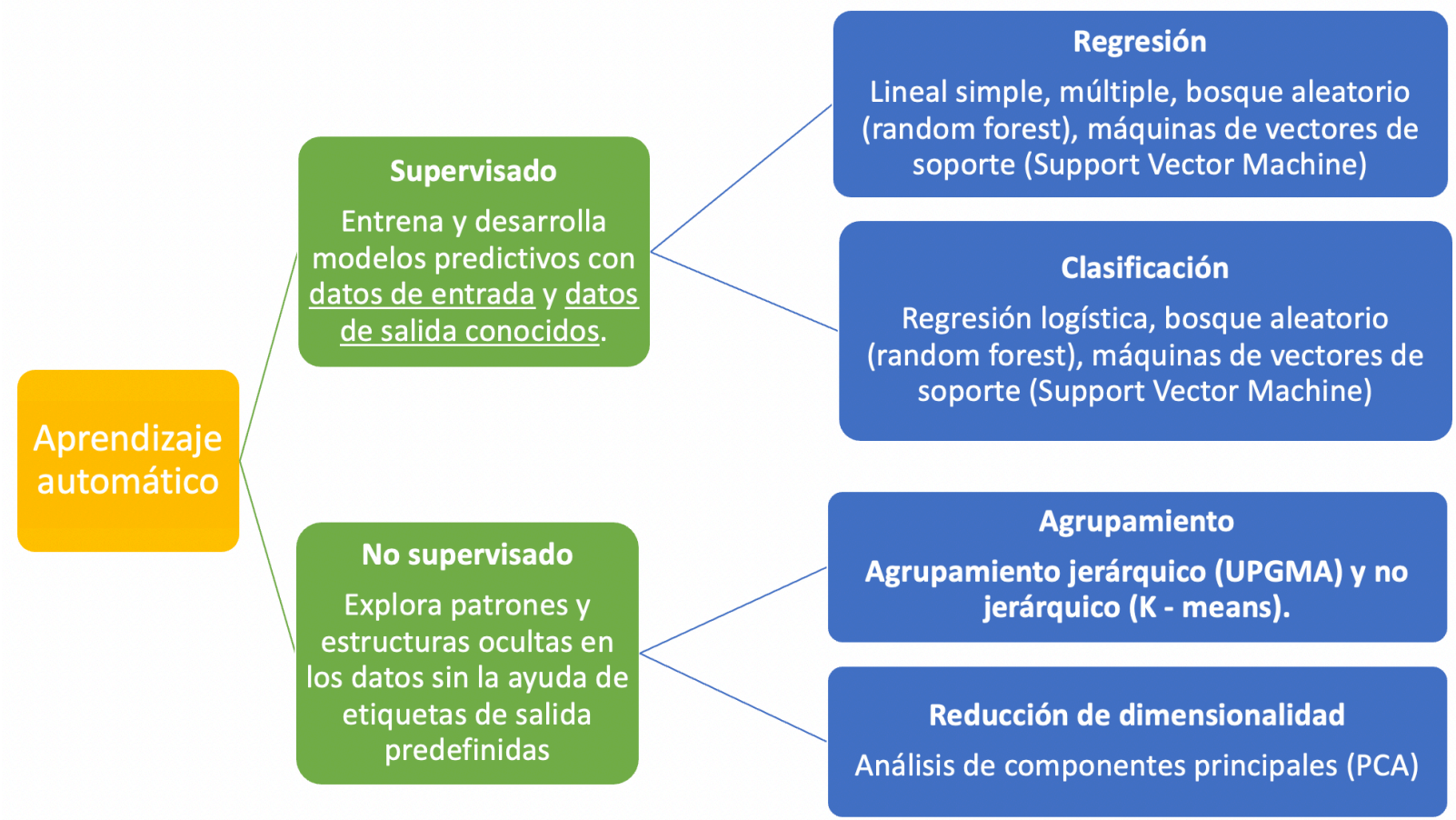
- Tipos de modelos predictivos.
- Regresión logística.
- Validación split-sample.
- ¿Cómo medir la precisión/desempeño del modelo?

2.- Práctica con R y Rstudio cloud.

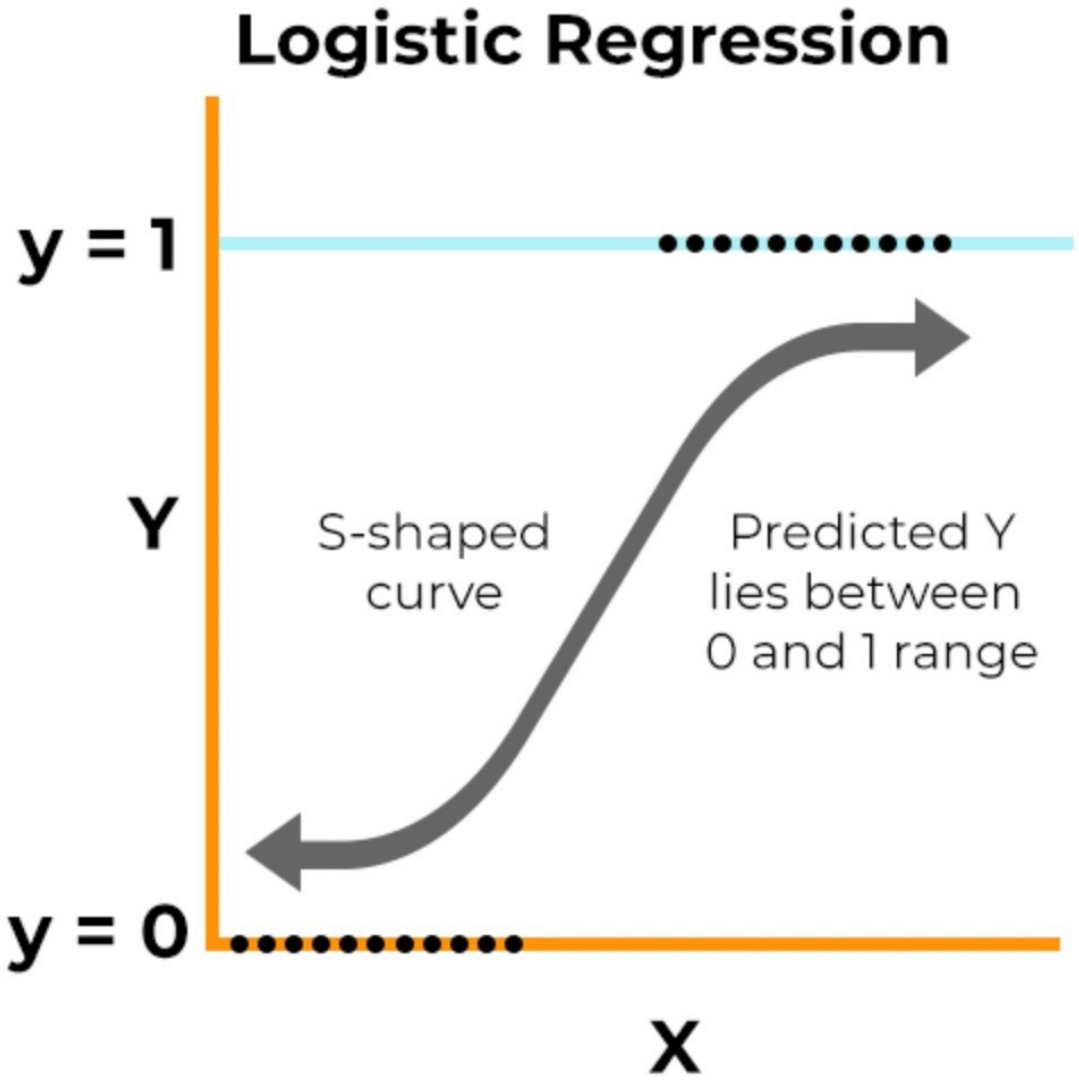
- Realizar modelamiento predictivo para regresión logística.
- Realizar gráficas avanzadas con ggplot2.

MACHINE LEARNING Y MODELOS PREDICTIVOS

Existen diferentes modelos predictivos, los cuales se clasifican bajo el marco del aprendizaje automático, machine learning en inglés, como algoritmos o metodos supervisados y no supervisados.



REGRESIÓN LOGÍSTICA



VALIDACIÓN SPLIT-SAMPLE

- División de datos: 70-80% de los datos para el conjunto de entrenamiento y el 30-20% restante para el conjunto de prueba.
- Evaluación del desempeño del modelo predictivo: Se usan métricas como la precisión, el área bajo la curva (AUC), la sensibilidad, la especificidad, entre otras.
- Variabilidad de los resultados: Es recomendable realizar múltiples particiones y promediar los resultados para obtener una evaluación más confiable del rendimiento del modelo.
- Sobreajustar y subajustar: Detectar problemas de sobreajuste (overfitting) y subajuste (underfitting) del modelo.

VALIDACIÓN CRUZADA SPLIT-SAMPLE

Dividir el set de datos en dos partes

Split-sample validation

70% de los datos para entrenar el modelo (Training dataset)

30% de los datos para probar/testear el modelo (Testing dataset)

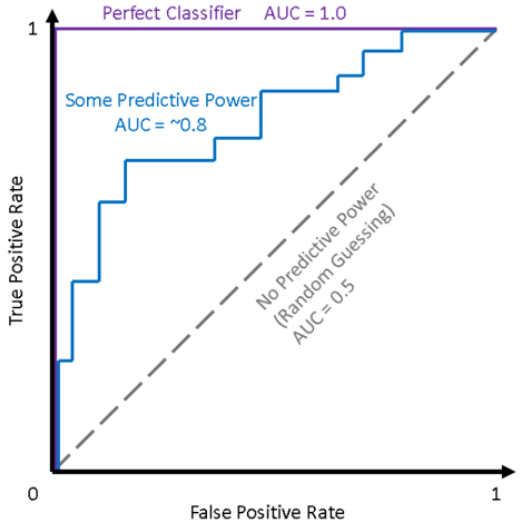
¿CÓMO MEDIR EL DESEMPEÑO DEL MODELO?

Métricas de evaluación

Rasgos binarios/categóricos con dos niveles

- Maduración sexual temprana
- Mortalidad

Área bajo la curva ROC (AUC)



$0 \leq AUC \leq 1$
AUC: Cercano a 1 indica que el modelo tiene capacidad para discriminar

ESTUDIO DE CASO: REGRESIÓN LOGÍSTICA EN CAMARONES

Y: Estado del Camaron: (0) fresco o en mal estado: (1).

X: Nitrógeno Volátil Total (Método destructivo).

Deep learning detection of shrimp freshness via smartphone pictures

Yuehan Zhang[1] · Chencheng Wei[1] · Yi Zhong[1,2] · Handong Wang[1] · Heng Luo[2] · Zuquan Weng[1,2,3]

Received: 31 March 2022 / Accepted: 25 May 2022 / Published online: 21 June 2022

Abstract

Shrimp is a type of aquatic product that is easy to deteriorate and the freshness has an essential influence on both its taste and nutritional value. Scientists have developed various approaches to measure shrimp freshness; however, the existing methods are usually destructive, complicated and costly. To develop a fast, non-destructive and low-cost alternative, we utilized deep learning models to identify the freshness of shrimp based on photos taken by smartphones. The models were trained on photographs of 306 shrimp along with their total volatile basic nitrogen values as freshness indicators. Our models achieved an area under receiver operating characteristic above 0.90 for freshness classification and root mean square error of prediction no more than 4.67 mg/100 g on fresh samples during the independent tests. Furthermore, the model performance was evaluated on datasets of shrimp photographed for 7 consecutive days and shrimp placed on different backgrounds and light settings. Our study suggested deep learning as an accurate, easy and low-cost method to detect shrimp freshness, which may have broader applications in food safety.

CALIDAD DE CAMARÓN

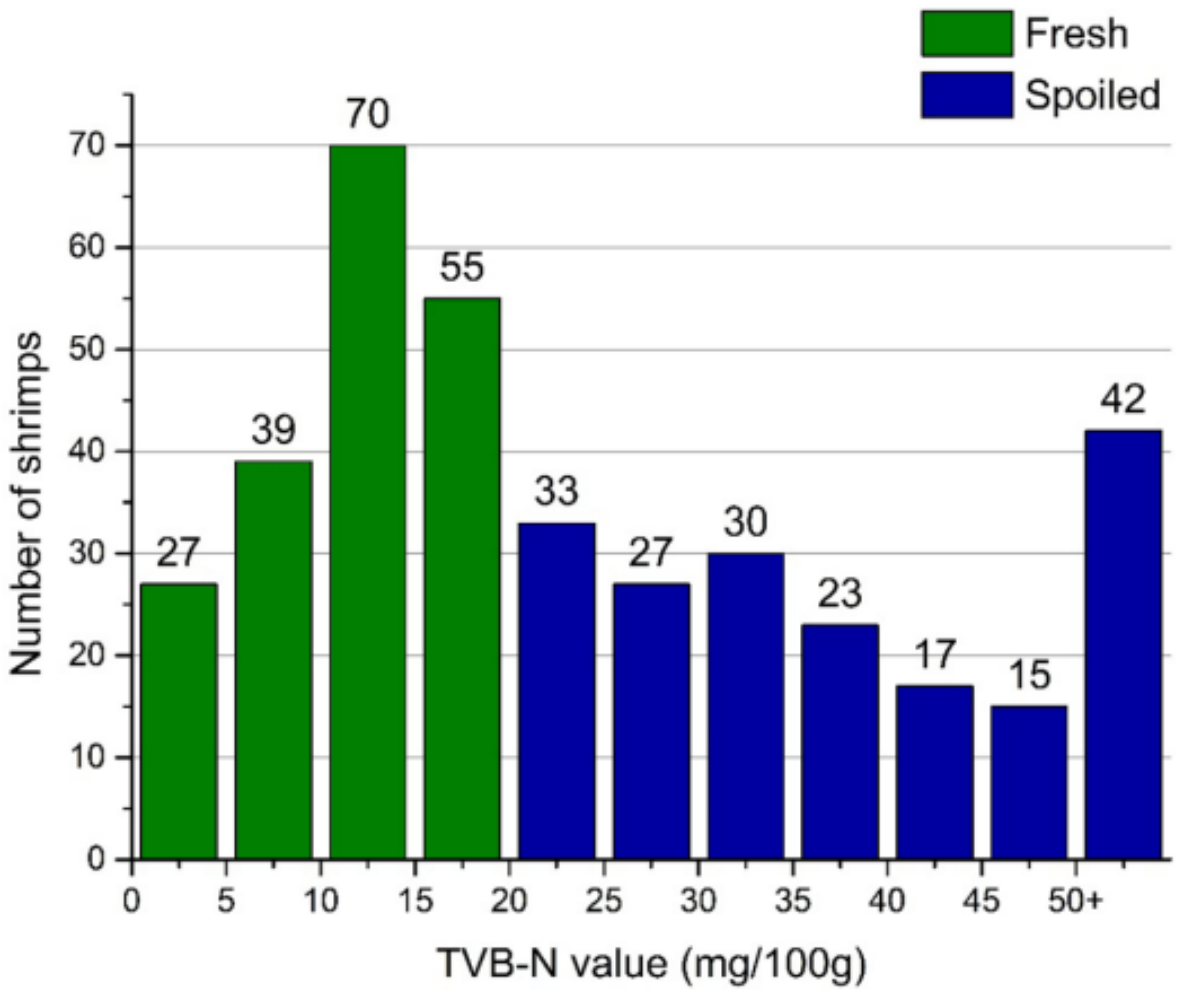
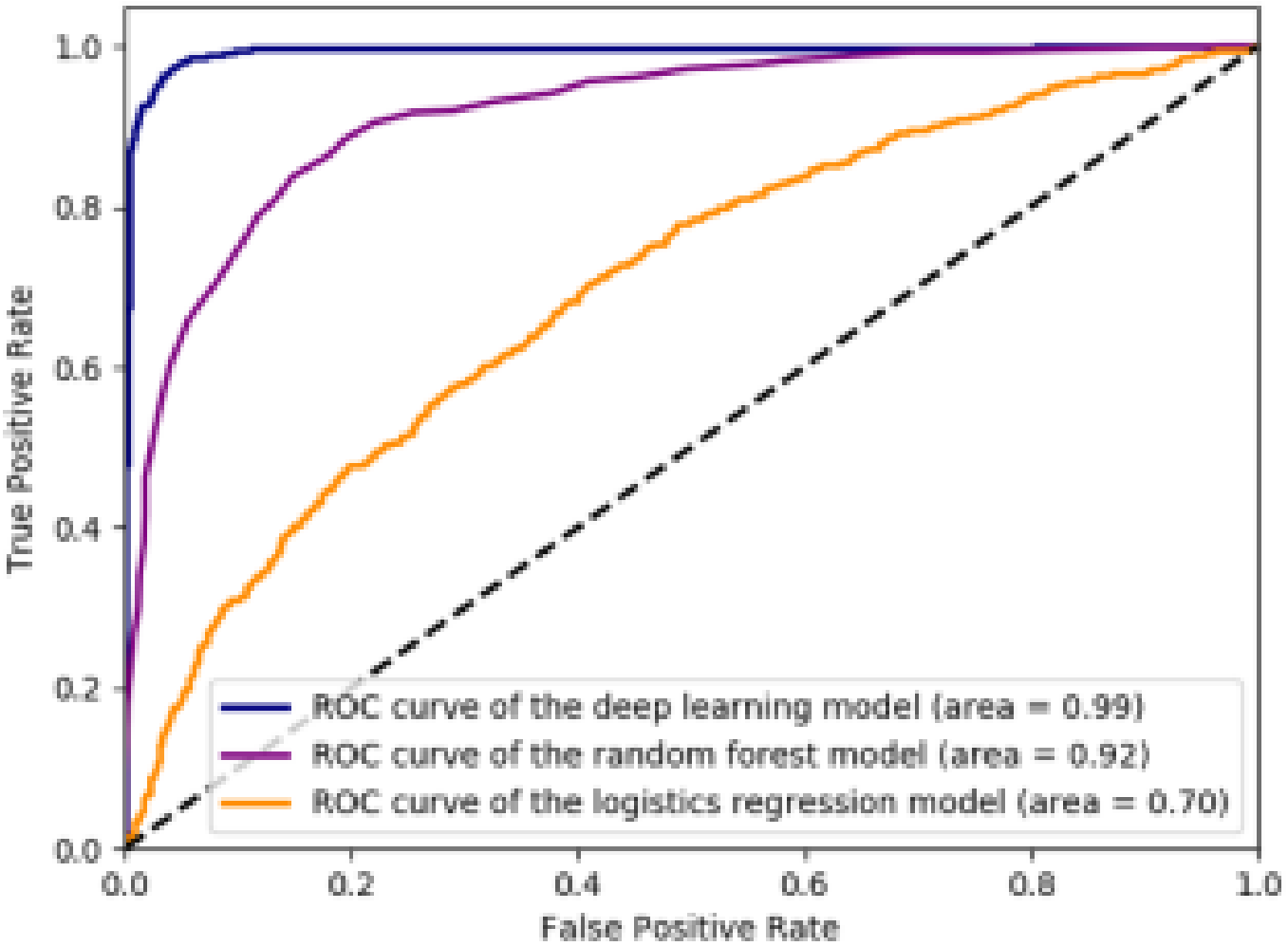


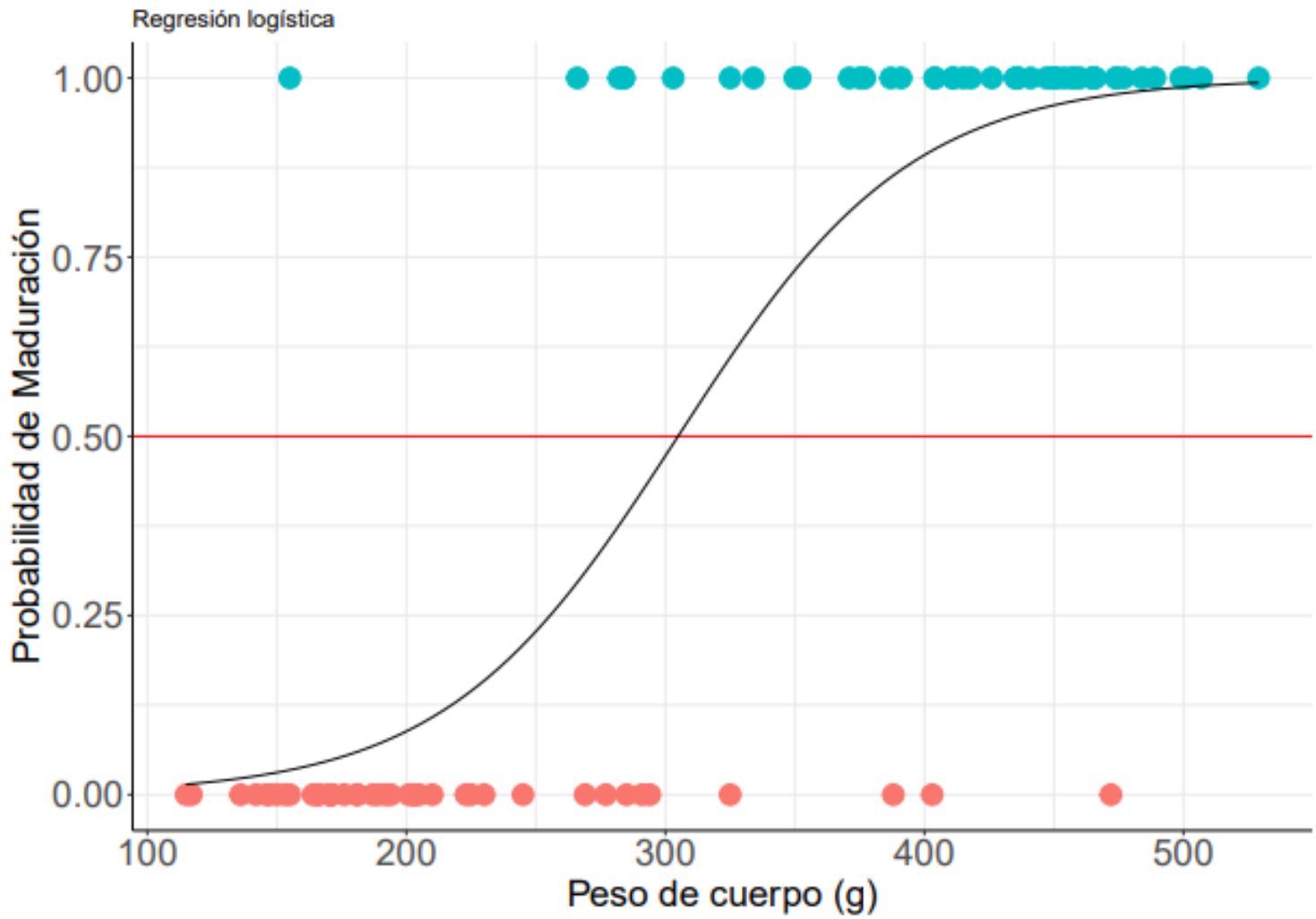
Fig. 3 Distribution of TVB-N values across all shrimp

ÁREA BAJO LA CURVA



ESTUDIO DE CASO: REGRESIÓN LOGÍSTICA SIMPLE

Queremos predecir la probabilidad de madurar en función del peso.

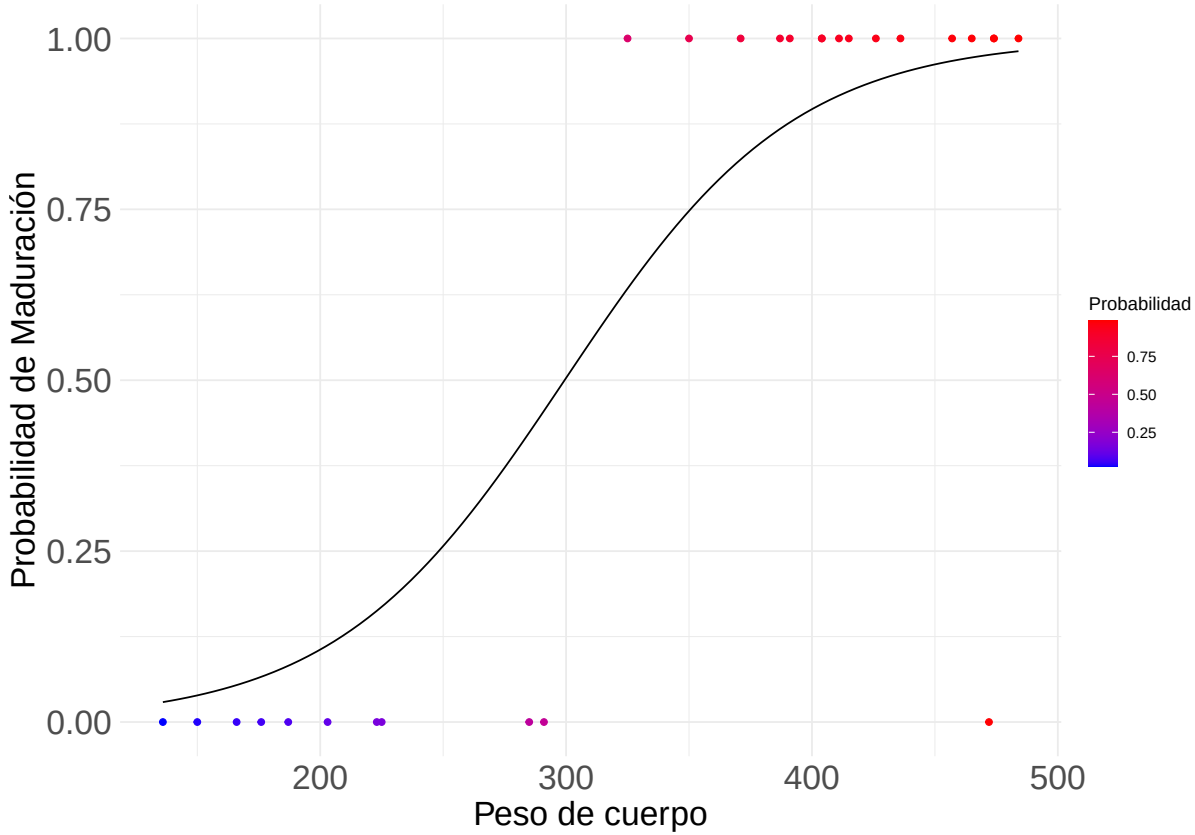


VALIDACIÓN SPLIT-SAMPLE

Fish	Genotype	Gonad	Mass	Maturation	Probabilidad
1	EE	8.88	391	1	0.88
2	EL	0.10	203	0	0.11
3	EE	10.74	474	1	0.98
4	EE	9.83	436	1	0.95
11	LL	0.07	136	0	0.03
19	EE	10.64	415	1	0.92
20	EE	12.04	484	1	0.98
24	EL	2.90	285	0	0.42
28	EE	9.43	426	1	0.94
35	EE	11.38	474	1	0.98

PREDICCIONES DE LA REGRESIÓN LOGÍSTICA SIMPLE

n: 88 datos train_data: 61 y test_data: 27



MATRIZ DE CONFUSIÓN Y MÉTRICAS DE EVALUACIÓN

		y (Maduración actual)	
		Immature *Salmo salar* 	Mature *Salmo salar* 
$\hat{y}$ (Maduración predicha)	Immature *Salmo salar* 	10	0
	Mature *Salmo salar* 	1	16

Métrica	Estimación
Precisión (Precision)	16/(16+0) = 1
Sensibilidad	16/(16+1) = 0.94
Especificidad	10/(10+0) = 1
Exactitud (Accuracy)	(10+16)/(10+0+1+16)=0.96

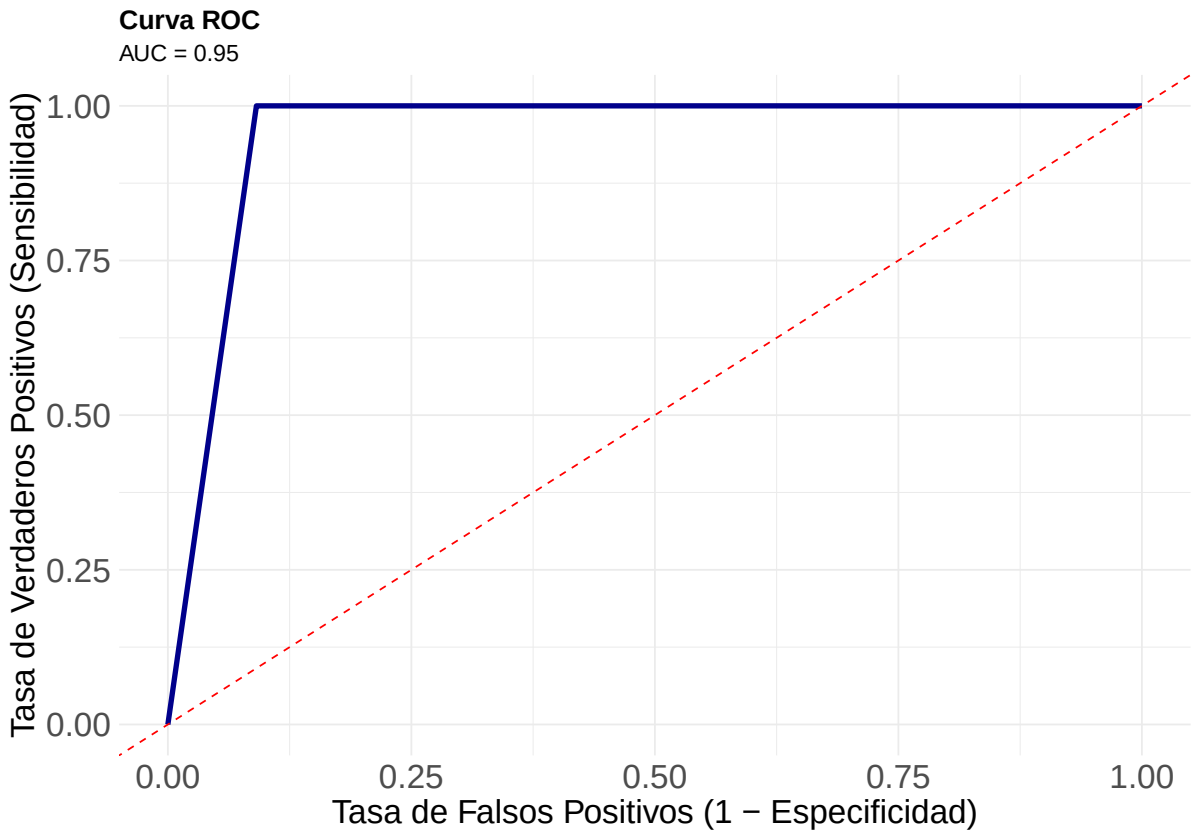
Precisión: **VP/(VP+FP)**

Sensibilidad: **VP/(VP+FN)**

Especificidad: **VN/(VN+FP)**

Exactitud: **(VP+VN)/(VP+ FP + FN+VN)**

AREA BAJO LA CURVA ROC (AUC)



PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.

RESUMEN DE LA CLASE

- Validación (Split-Sample).
- Métricas de evaluación (AUC, Especificidad, Precisión, Sensibilidad, Exactitud).