Clase 25 Técnicas avanzadas de modelamiento predictivo: Máquinas de Soporte Vectorial (SVM)

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

Dra. María Angélica Rueda Calderón y Dr. Jose Gallardo

Pontificia Universidad Católica de Valparaíso

22 July 2023

PLAN DE LA CLASE

1.- Introducción

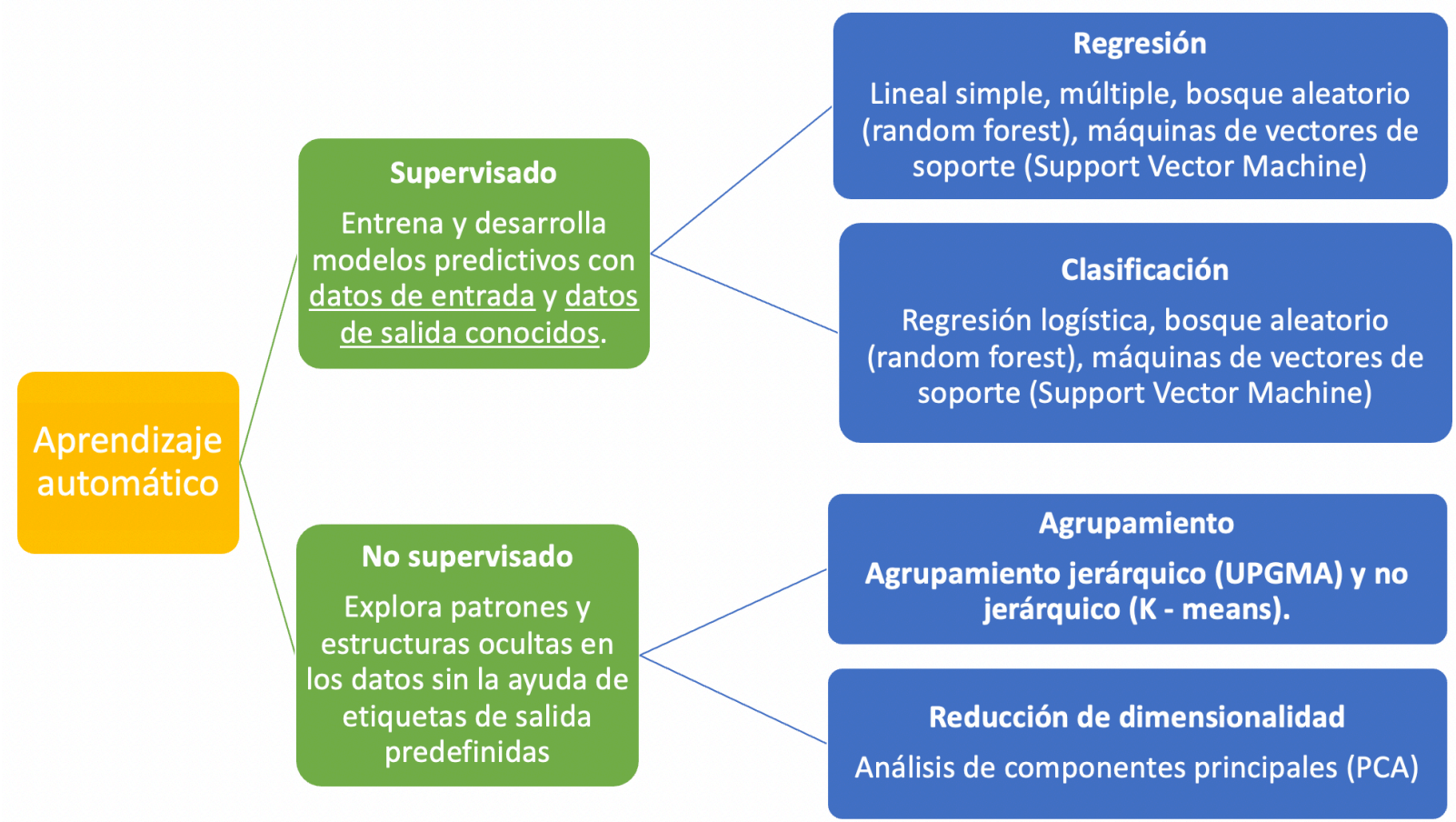
- Tipos de modelos predictivos.
- Estudio de caso: SVM en camarones irradiados.
- Support vector machine (SVM): ¿Qué son y para que sirven?.
- Estudios de caso: Predicción de peso y predicción de abundancia de microalgas tóxicas.
- Métricas para medir el desempeño de un modelo de regresión.
- Práctica con ejemplo de maduración temprana.

2.- Práctica con R y Rstudio cloud.

- Realizar modelamiento predictivo con Máquinas de Soporte Vectorial.
- Realizar gráficas avanzadas con plotly.

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.



ESTUDIO DE CASO: CAMARON IRRADIADO

- ¿Cómo distinguir camarón irradiado v/s no irradiado?

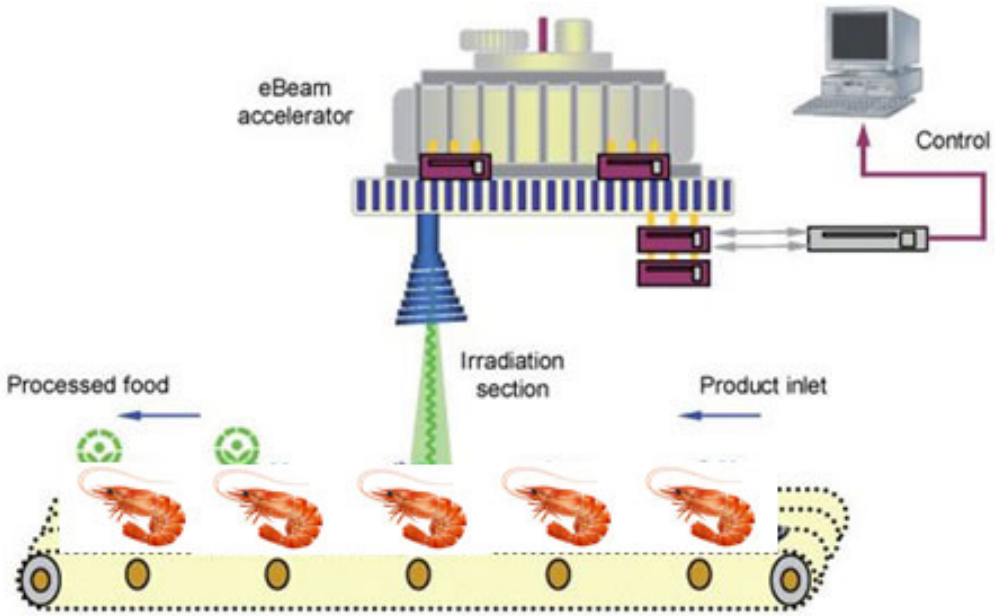


Fig. 1: *Electron Beam Irradiation (Morata et al., 1989)*

Fuente: Xiong et al. 2016

ESTUDIO DE CASO: ANÁLISIS IMÁGENES MULTIESPECTRAL

- Análisis de 19 variables en diferentes longitudes de onda.

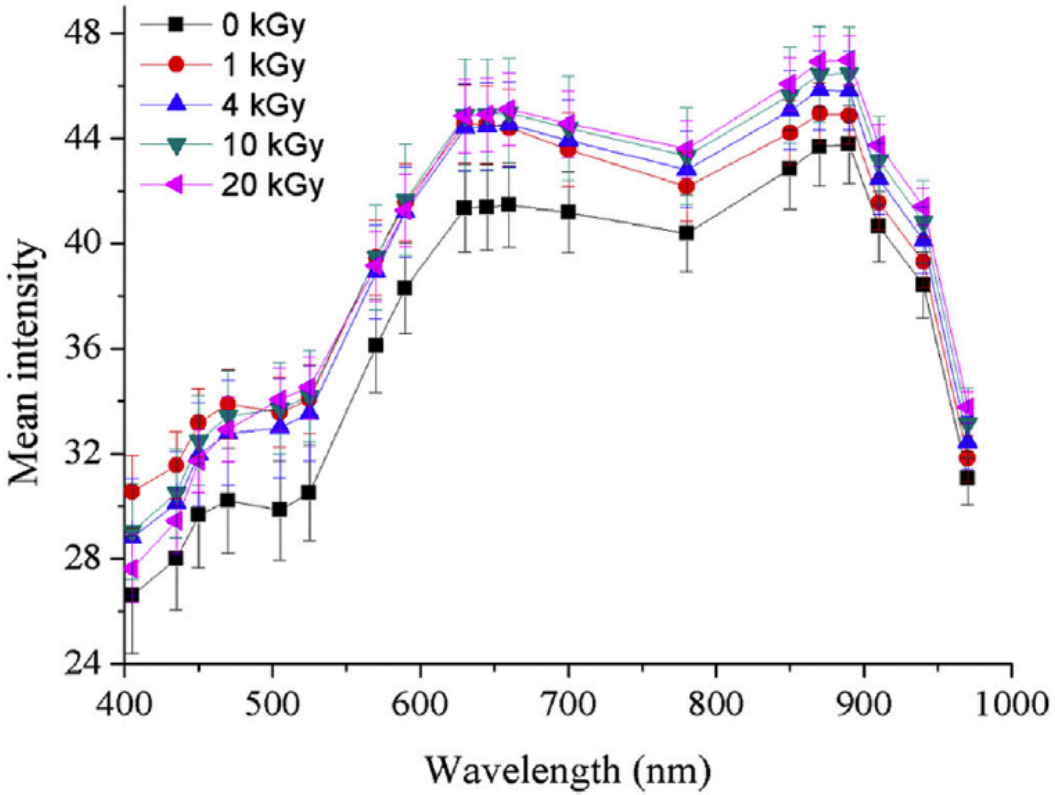
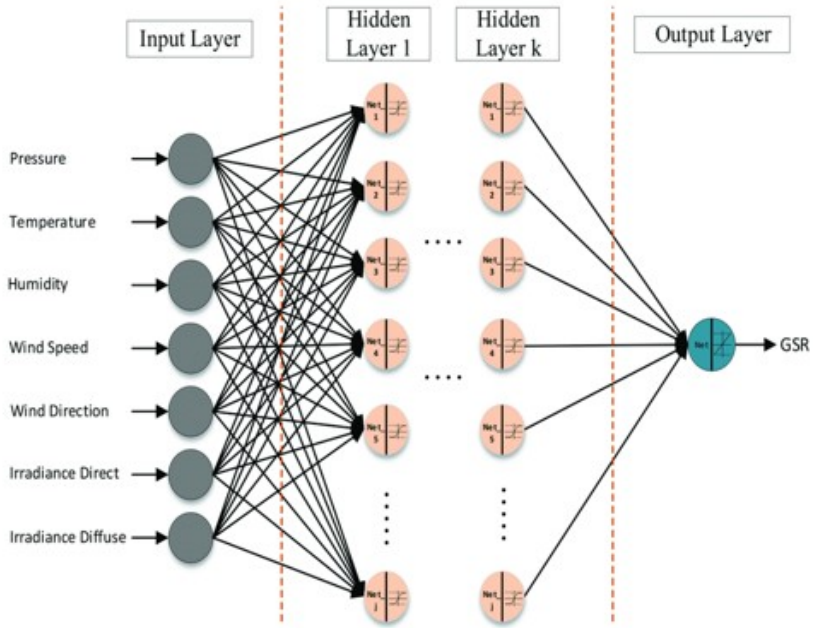
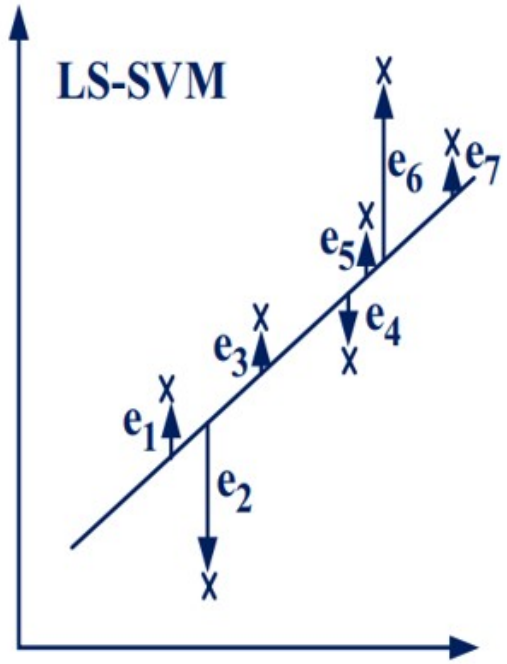


Fig. 2. Average spectral from multispectral imaging of shrimp by five different irradiation doses. Vertical bars represent standard deviations from measurements.

MODELOS PREDICTIVOS





BPNN (Back propagation neural network): Red neuronal.

LS-SVM (Least squares-support vector machines): Nuevo método de máquinas de soporte vectorial (SVM).

Fuente: Aljanad et al. 2021

Fuente: Dameshghi & Refan, 2021

¿CÓMO MEDIR EL DESEMPEÑO DEL MODELO DE CLASIFICACIÓN?

Para problemas con dos clases:

$$SEN = \frac{a}{a+b}, \quad SPE = \frac{a}{a+c}, \quad G = \sqrt{SEN \cdot SPE},$$

$$JAC = \frac{a}{a+b+c}, \quad MCC = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b)\cdot(c+d)\cdot(a+c)\cdot(b+d)}},$$

$$ACC = \frac{a+d}{a+b+c+d}$$

where a are the true positive, i.e. the objects correctly included in class A; b the false negative, i.e. the class A objects not included in the class; c are the false positive, i.e. the not-A objects included in class A; d are the true negative, i.e. the not-A objects not included in class A.

SEN: Sensitivity
SPE: Specificity
G: Geometric mean
JAC: Jaccard's coefficient
MCC: Matthews correlation coefficient or (Pearson's coefficient)
ACC: Overall prediction accuracy

Table 2
Classification Results of BPNN and SVM on external testing set for four considered data sets.

Data set	Predicted class by BPNN						Predicted class by LS-SVM					
	SEN	SPE	G	JAC	MCC	ACC	SEN	SPE	G	JAC	MCC	ACC
0, 1	80.0	76.7	55.7	78.3	59.3	78.0	70.0	80.0	50.0	74.8	53.9	76.0
0, 4	80.0	90.0	70.7	84.9	69.6	86.0	85.0	93.3	79.1	89.1	77.3	90.0
0, 10	100.0	96.7	98.3	95.2	96.0	98.0	100.0	100.0	100.0	100.0	100.0	100.0
0, 20	95.0	100.0	97.5	95.0	95.9	98.0	100.0	100.0	100.0	100.0	100.0	100.0

All results are the means in percent calculated with 40 replications together.
SEN: sensitivity; SPE: specificity; G: geometric mean; JAC: Jaccard's coefficient; MCC: Matthews correlation coefficient; ACC: overall prediction accuracy.

¿CÓMO MEDIR EL DESEMPEÑO DEL MODELO DE CLASIFICACIÓN?

Para problemas multi-clase:

$$CR = \frac{N_c}{N_c + N_{nc}} \times 100\%$$

$$ER = \frac{N_{nc}}{N_c + N_{nc}} \times 100\% = 1 - CR$$

N_c: Number of correct classification of samples.
N_{nc}: Number of wrong classification of samples.

CR: Correct **R**ate or called percentage of correct classification

ER: Error **R**ate or called percentage of wrong classification.

Table 1
Classification correct rates of LS-SVM and BPNN on the calibration and prediction data sets. Numbers in parentheses indicate correct classified samples of total samples.

Data set	Calibration set		Prediction set	
	LS-SVM	BPNN	LS-SVM	BPNN
0, 1, 4, 10, 20	84.2% (345/410)	90.7% (372/410)	73.6% (81/110)	70.0% (77/110)
0, 1	86.5% (147/170)	99.4% (169/170)	76.0% (38/50)	78.0% (39/50)
0, 4	96.5% (164/170)	100.0% (50/170)	90.0% (45/50)	86.0% (44/50)
0, 10	99.4% (169/170)	100.0% (50/170)	100.0% (50/50)	98.0% (49/50)
0, 20	100.0% (170/170)	100.0% (50/170)	100.0% (50/50)	98.0% (49/50)

$N = 520 \quad N_{CS\_all} = 410\ (\sim 80\%) \quad N_{PS_{all}} = 110\ (\sim 20\%)$

$$CR_{CS_{LS-SVM}} = \frac{345}{345 + 65} \times 100 = \frac{345}{410} \times 100 = 84.2\%$$

$$CR_{PS_{LS-SVM}} = \frac{81}{81 + 29} \times 100 = \frac{81}{110} \times 100 = 73.6\%$$

$$ER_{CS_{LS-SVM}} = 1 - CR_{CS_{LS-SVM}} = 15.8\%$$

$$ER_{PS_{LS-SVM}} = 1 - CR_{PS_{LS-SVM}} = 26.4\%$$

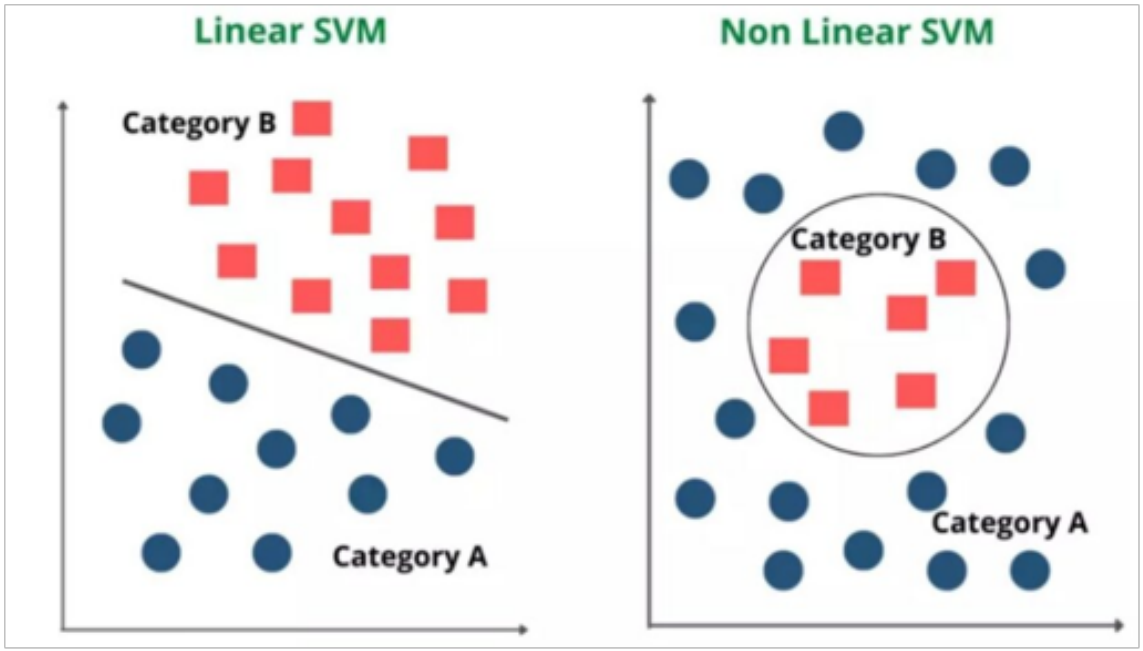
ALGORITMO SUPERVISADO MÁQUINAS DE SOPORTE VECTORIAL (SVM): CLASIFICACIÓN

SVM fue creado por Vladimir Vapnik y su equipo en los años 90 en los laboratorios de AT&T Bell en los Estados Unidos.

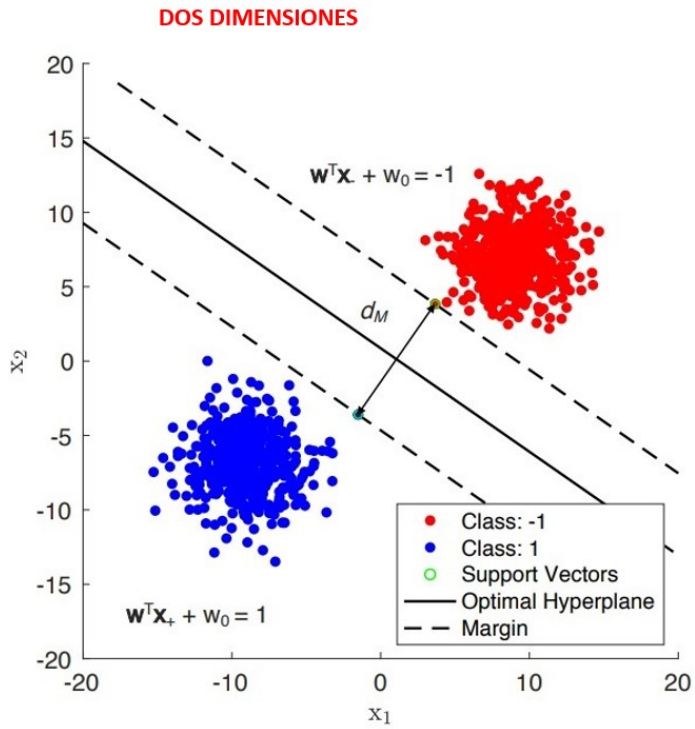
- SVM es un algoritmo de aprendizaje supervisado utilizado para clasificación y regresión.
- SVM encuentra un hiperplano en un espacio de alta dimensión que mejor separa las clases de datos.
- Se enfoca en encontrar los vectores de soporte, que son las observaciones más cercanas a la separación entre clases.
- Es efectivo incluso en espacios de alta dimensión ($>$ Variables que observaciones).

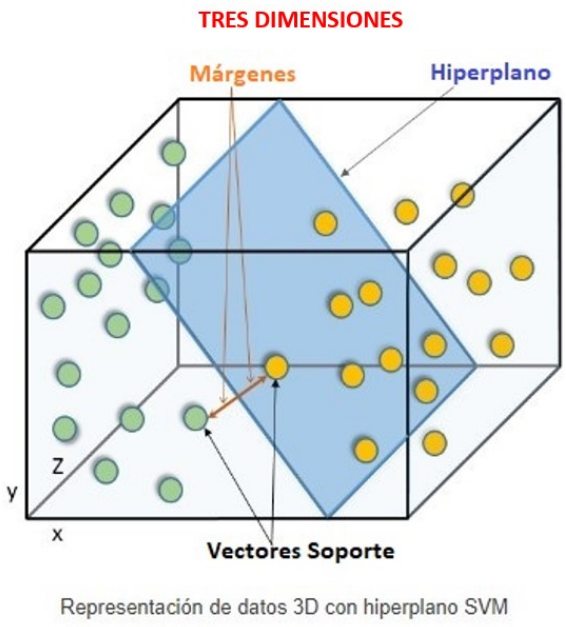
MÁQUINAS DE SOPORTE VECTORIAL (SVM)

- Hay problemas de clasificación y regresión linealmente separables (***Kernel lineal***).
- SVM no lineales permiten modelar relaciones complejas y no lineales entre las variables predictoras y la variable respuesta (***Kernel de base radial***, ***Kernel polinomial*** y ***Kernel sigmoidal***).



PARTES DE SVM





Representación de datos 3D con hiperplano SVM

Fuente: Bianco et al., 2019

Página web

TIPOS DE KERNELS EN SVM

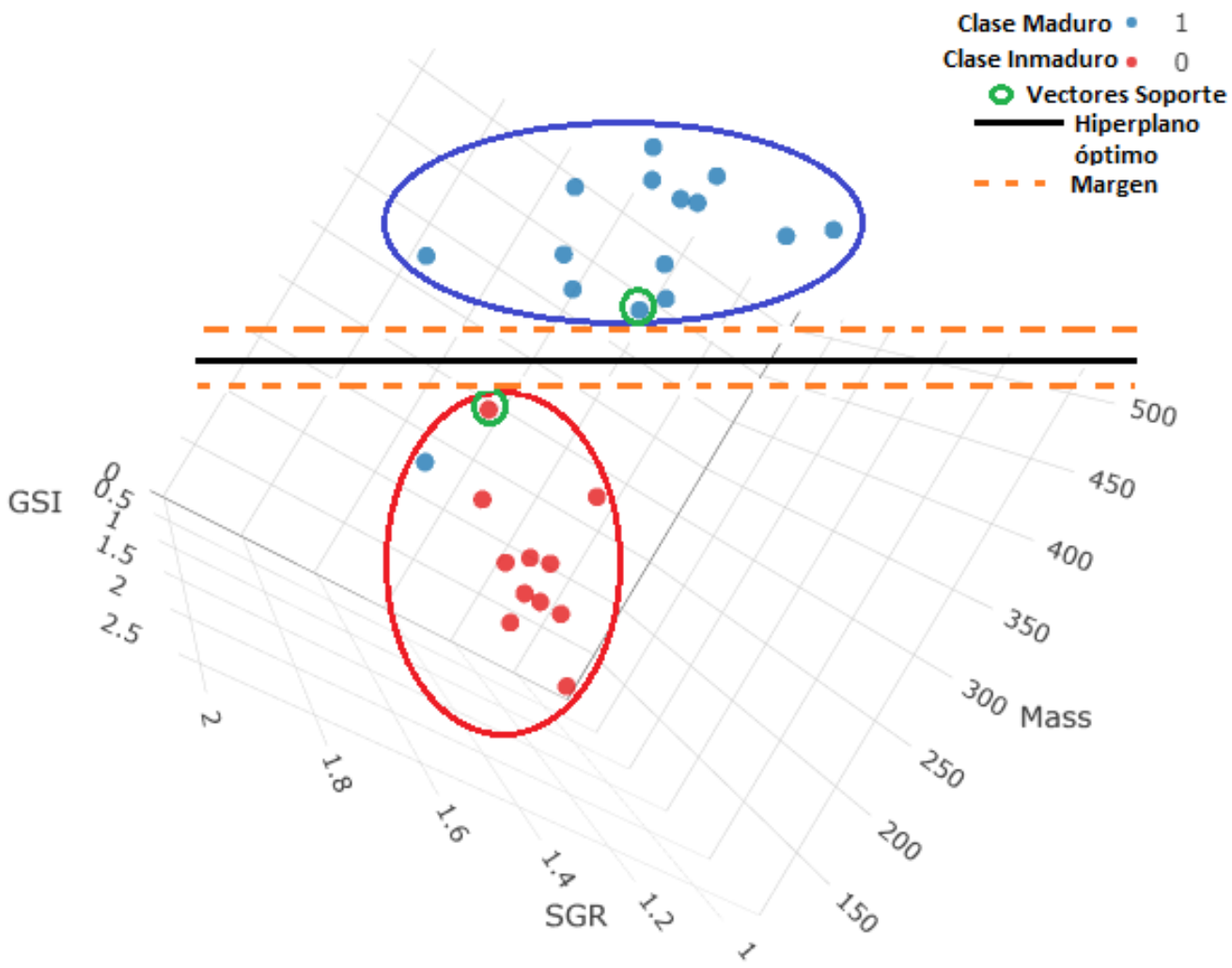
kernel es una función matemática que mide la similitud entre pares de datos en un espacio de características. El kernel en SVM es una herramienta especial que nos permite convertir datos difíciles de separar en un espacio donde sí pueden ser separados.

Tipo de Kernel	Hiperparámetros	Uso
Lineal	C (parámetro de costo)	Clasificación y regresión lineal
Base Radial (RBF)	C (parámetro de costo) gamma (parámetro del kernel RBF)	Clasificación y regresión no lineal

EJERCICIO: MADURACIÓN TEMPRANA CON SVM

Mass	SGR	Length	GSI	Maturation
391	1.64	30.8	2.271	1
203	1.23	26.3	0.049	0
474	1.77	31.8	2.266	1
436	1.72	31.8	2.255	1
282	1.93	27.8	2.241	1
277	1.08	29	0.065	0
171	1.11	25.2	0.041	0
171	1.06	25.5	0.041	0
284	1.76	28.2	1.57	1
466	1.72	32.6	2.693	1

PREDICCIÓN EN SVM DE CLASIFICACIÓN



MATRIZ DE CONFUSIÓN

Table 3: Matriz de Confusión SVM

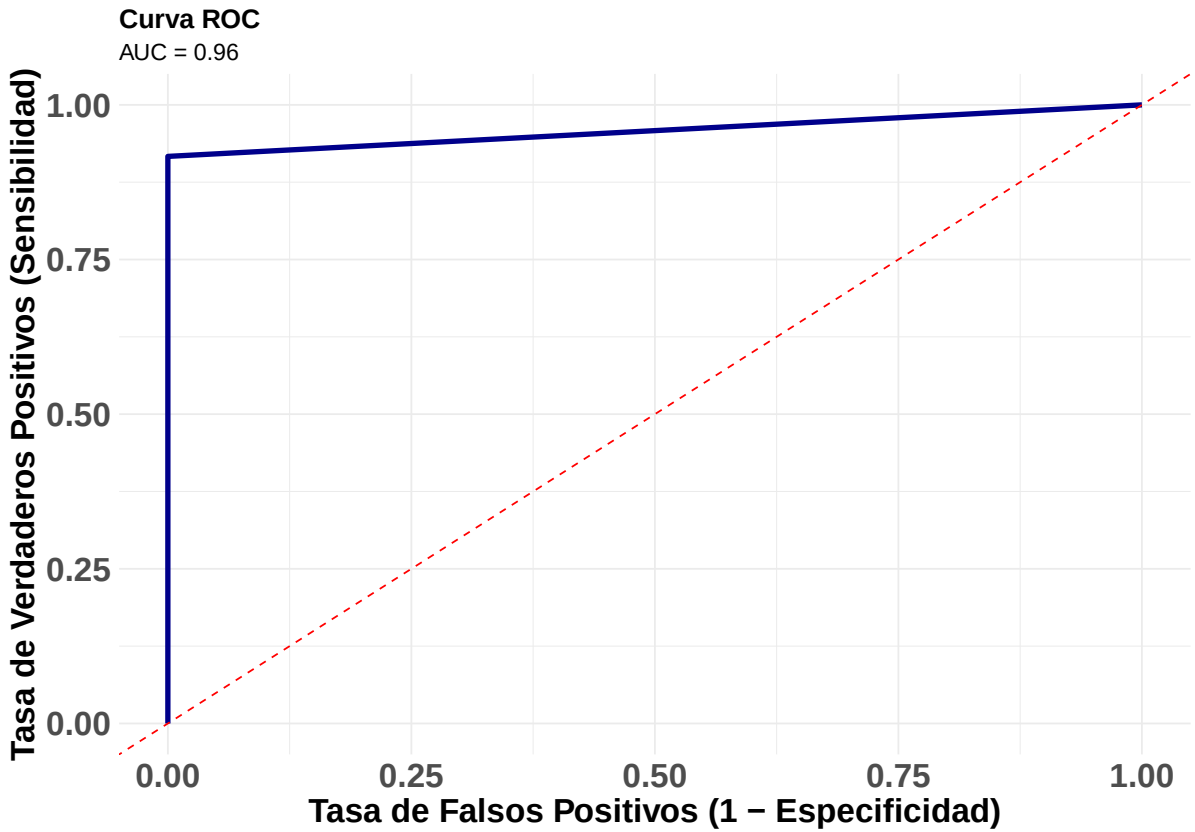
	Maduro	Inmaduro
Maduro	14	1
Inmaduro	0	11

$$ACC = \frac{(14+11)}{(14+0+1+11)} = 0.96$$

$$SEN = \frac{(14)}{(14+0)} = 1$$

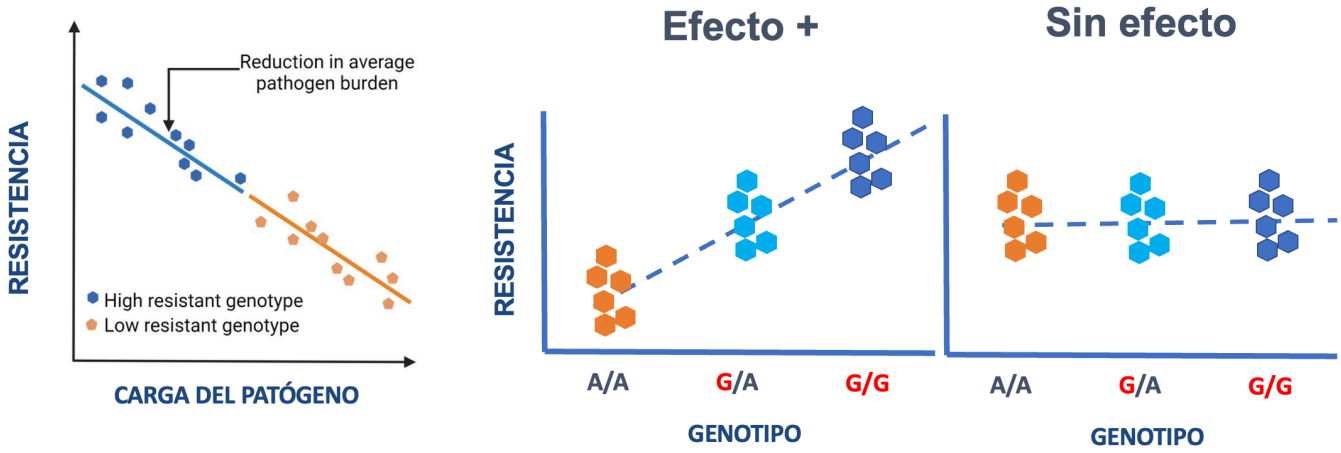
$$SPE = \frac{(11)}{(11+1)} = 0.92$$

AREA BAJO LA CURVA ROC (AUC) PARA SVM



ESTUDIO DE CASO 2: RESISTENCIA A PATÓGENOS

- Variable predicha: Resistencia a koi herpes virus en carpa (Resistente = 0; Susceptible =1).
- Variables predictoras: 15.615 genotipos SNP.



Fuente: Palaiokostas 2021 Imagen referencial modificada: Seal. et al. 2020

MÉTODOS DE PREDICCIÓN Y PERFORMANCE

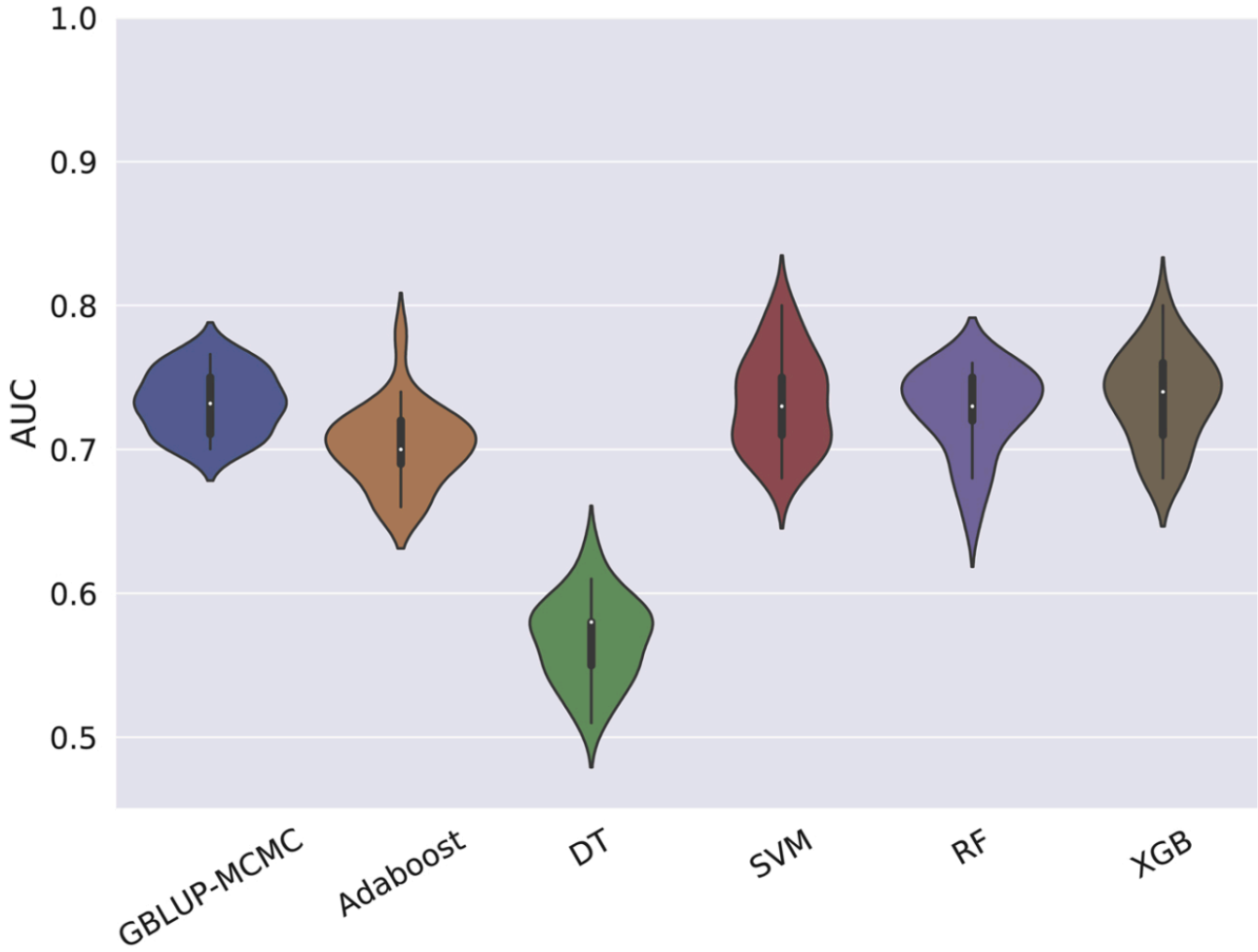
1.- MÉTODOS

- GBLUP: Predicción lineal insesgada genómica (Una para cada genotipo, Línea base).
- ML: SVM, RF, Decision Trees (DT), otros.

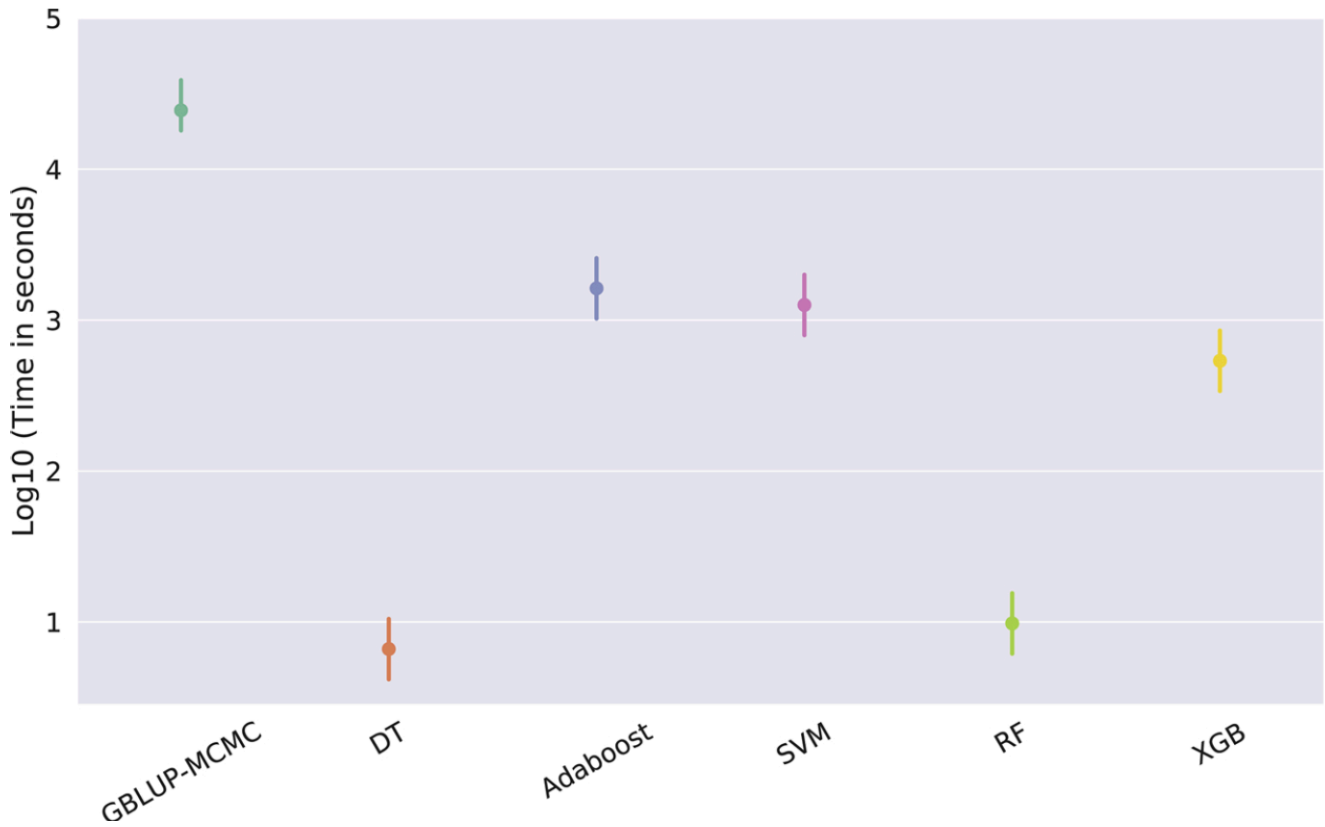
2.- MEDIDAS DE PERFORMANCE

- Curva ROC: Relación entre la tasa de verdaderos positivos (Sensibilidad) y la tasa de falsos positivos (1-Especificidad).
- AUC (Area Under the Curve): Valor numérico que se obtiene a partir de la curva ROC y es un indicador de la capacidad de discriminación del modelo de clasificación.
- Un AUC de 1.0 significa que el modelo tiene una capacidad perfecta para distinguir entre las dos clases, mientras que un AUC de 0.5 indica que el modelo no es mejor que una clasificación aleatoria.

COMPARACIÓN MODELOS: AUC



COMPARACION MODELOS: TIEMPO.



EJERCICIO CON R: RESISTENCIA A PATÓGENOS

n: 1259 observaciones y 9 genotipos SNP.

Table 4: Matriz de Confusión SVM Sobrevivencia

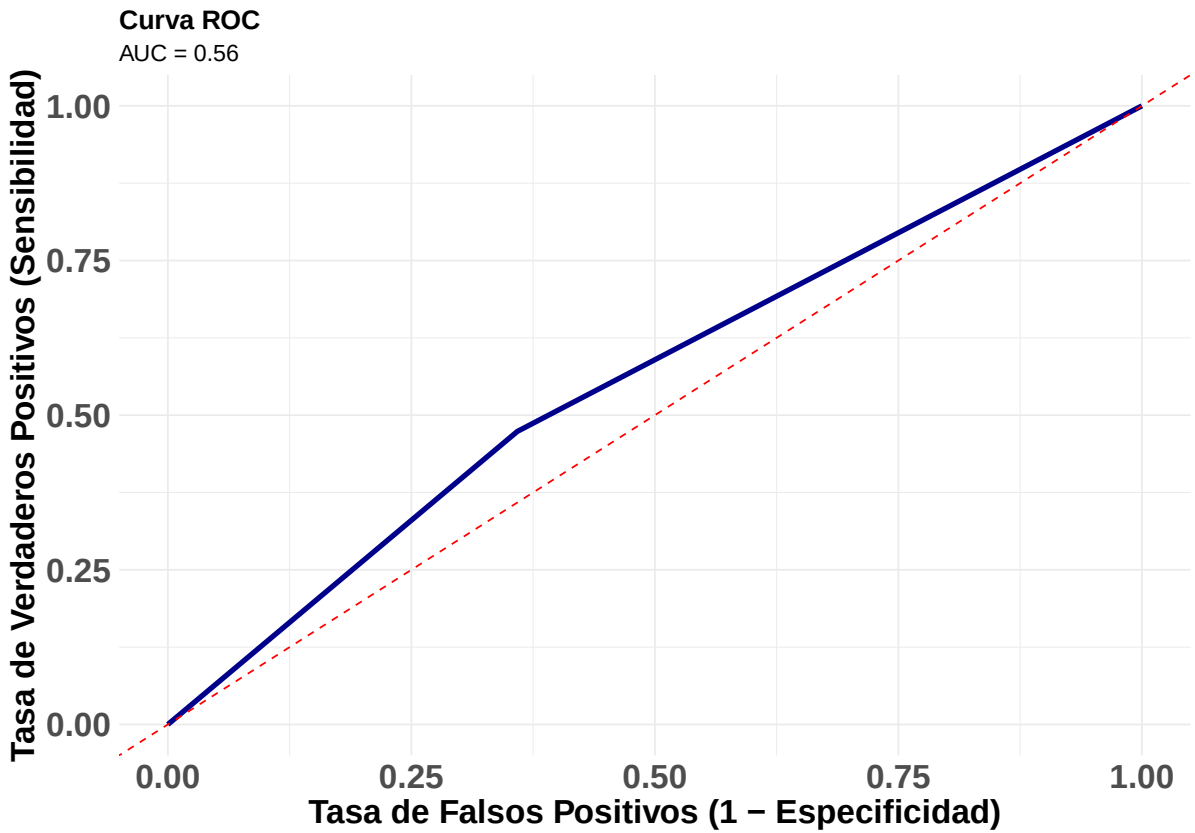
	Muerto	Vivo
Muerto	84	130
Vivo	47	117

$$ACC = \frac{(84+117)}{(84+130+47+117)} = 0.53$$

$$SEN = \frac{(84)}{(84+47)} = 0.64$$

$$SPE = \frac{(117)}{(117+130)} = 0.47$$

AREA BAJO LA CURVA ROC (AUC) PARA SVM



PRÁCTICA ANÁLISIS DE DATOS

- El trabajo práctico se realiza en Rstudio.cloud.

RESUMEN DE LA CLASE

- Obtener los mejores parámetros para los modelos de SVM.
- Realizar predicciones con SVM.
- Métricas de evaluación (AUC, Accuracy, MSE).