

The background is a dark blue gradient. On the left, there is a large, semi-transparent circular image of a circuit board. Overlaid on the top left of this circle are two overlapping triangles: a blue one in front of a green one. In the top right corner, there is a 3D perspective view of a circuit board's traces.

Fundamentos de Confusion Matrix

Prof. Jhoan Steven Delgado V.



Intro

Edad	Ingresos	Tiene carro?
24	1'200.000	NO
23	4'500.000	SI
45	1'250.000	SI
32	1'100.000	NO



Logistic
Regression?

?

KNN?

Confusion Matrix (Matriz de confusión)

*Una matriz de dimensión $n \times n$ donde n es la cantidad de clases (o categorías) de la variable objetivo

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	142 (VP)	22 (FN)
	No tienen Carro	29 (FP)	110 (VN)

Número de personas que sí tienen carro y que el modelo correctamente predijo que SI. (VP)

Número de personas que sí tienen carro y que el modelo incorrectamente predijo que NO. (FN - TIPO II)

Número de personas que no tienen carro y que el modelo incorrectamente predijo que SI. (FP - TIPO I)

Número de personas que no tienen carro y que el modelo correctamente predijo NO. (VN)



Confusion Matrix (Matriz de confusión)

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	142	29
	No tiene Carro	22	110

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	139	32
	No tiene Carro	20	112

Accuracy

Accuracy= $(VP+VN)/(VP+VN+FP+FN)$ (Fracción de clasificaciones correctas)

Accuracy KNN = $(142+110)/(142+110+22+29) = 0.83 \rightarrow 83\%$ de todas las predicciones fueron correctas

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	142 (VP)	29 (FN)
	No tiene Carro	22 (FP)	110 (VN)

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	139 (VP)	32 (FN)
	No tiene Carro	20 (FP)	112 (VN)

Recall (Sensitivity), Specificity

- Recall (Sensitivity): Porcentaje de verdaderos **positivos** que fueron correctamente predichos por nuestro modelo

$$\text{Recall (Sensitivity)} = \text{VP} / (\text{VP} + \text{FN}) \rightarrow \text{TRP (True Positive Rate)}$$

- Specificity: Porcentaje de verdaderos **negativos** que fueron correctamente predichos por nuestro modelo

$$\text{Specificity} = \text{VN} / (\text{VN} + \text{FP}) \rightarrow \text{TNR (True Negative Rate)}$$

Recall (Sensitivity) KNN = $142 / (142 + 29) = 0.83 \rightarrow 83\%$ de las personas con carro, fueron correctamente predichas por nuestro modelo KNN.

Specificity LR = $112 / (112 + 20) = 0.85 \rightarrow 85\%$ de las personas sin carro, fueron correctamente predichas por nuestro modelo LR.

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	142 (VP)	29 (FN)
	No tiene Carro	22 (FP)	110 (VN)

KNN

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	139 (VP)	32 (FN)
	No tiene Carro	20 (FP)	112 (VN)

Logistic Regression

Precision

- Precision: Porcentaje de los positivos **predichos** de nuestro modelo que fueron correctamente predichos (correctos)

$$\text{Precision} = \text{VP} / (\text{VP} + \text{FP})$$

Precision KNN = $142 / (142 + 22) = 0.87$ -> De las 164 personas que predcimos que tenían carro, solo el 87% realmente tienen carro.

Precision nos da una noción de la calidad de los positivos predichos

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	142 (VP)	29 (FN)
	No tiene Carro	22 (FP)	110 (VN)

KNN

		Predicción	
		Tiene Carro	No tiene carro
Realidad	Tiene Carro	139 (VP)	32 (FN)
	No tiene Carro	20 (FP)	112 (VN)

Logistic Regression

F1 Score

- Favorece a los clasificadores que tienen alto recall y precision.

$$F1 \text{ Score} = 2 * \{ \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \}$$

		Predicción	
		Tiene Carro	No tiene carro
KNN			
Realidad	Tiene Carro	142 (VP)	29 (FN)
	No tiene Carro	22 (FP)	110 (VN)

		Predicción	
		Tiene Carro	No tiene carro
Logistic Regression			
Realidad	Tiene Carro	139 (VP)	32 (FN)
	No tiene Carro	20 (FP)	112 (VN)



Resumen

- $\text{Accuracy} = (\text{VP} + \text{VN}) / (\text{VP} + \text{VN} + \text{FP} + \text{FN})$ (Fracción de clasificaciones correctas)
- $\text{Mala clasificación} = (\text{FP} + \text{FN}) / (\text{VP} + \text{VN} + \text{FP} + \text{FN})$: (Fracción de clasificaciones incorrectas)
- $\text{Precisión} = \text{VP} / (\text{VP} + \text{FP})$: (Fracción de verdaderos positivos a positivos predichos)
- $\text{Recall (sensibilidad o TPR)} = \text{VP} / (\text{VP} + \text{FN})$ (Fracción de verdaderos positivos sobre todos los positivos)
- $\text{Especificidad (o TNR)} = \text{VN} / (\text{VN} + \text{FP})$: (Fracción de verdaderos negativos sobre todos los negativos)
- $\text{Tasa de falsos positivos (o FPR)} = \text{FP} / (\text{VN} + \text{FP}) = 1 - \text{TNR}$



Referencias

- Notas de clase, Prof. Javier Díaz, Fundamentos de analítica.
- The Statquest Illustrated guide to Machine Learning, Josh Starmer.