Análisis Predictivo

Introducción Análisis Estadístico

Breve resumen: La estadística es el estudio de los fenómenos aleatorios. En este sentido la estadística tiene un alcance ilimitado de aplicaciones en disciplinas que van desde la ingeniería, la medicina, ciencias económicas, etc. La estadística ha alcanzado gran importancia debido al creciente aumento de generación de datos provenientes de fuentes digitales, las cuales se usan con fines investigativos y/o de negocios. La mayor parte de la información viene expresada en forma de tablas o gráficos estadísticos, por lo que un conocimiento básico es necesario para la correcta interpretación de la información.

Se divide en dos ramas:

- **Estadística Descriptiva**: se relaciona con la recolección de datos, resumen, presentación y descripción de los mismos, por lo que resulta útil para realizar *análisis exploratorios*.
- **Estadística Inferencial**: se relaciona con el proceso de utilizar los datos experimentales de una muestra proveniente de una población para tomar decisiones. Es fundamental para hacer *análisis predictivos*.

La estadística sirve para:

1. Descubrir patrones en los datos —> Análisis descriptivo y exploratorio.
2. Rechazar/confirmar una hipótesis —> Análisis predictivo.

3. Obtener la mejor solución a un problema —> Análisis prescriptivo.

Los método de la *Estadística Descriptiva* nos permiten:

- Determinar la **tendencia central** de una variable: para esto se usan parámetros de posición (media, moda, mediana y cuartiles para posición relativa).
- Determinar la **variabilidad** de una variable: para esto se utilizan parámetros de dispersión (varianza, desvío estándar.
- Determinar cómo es la **distribución** de una variable: esto lo podemos conocer mediante los parámetros de forma (distribución normal, etc.)

▼ *Características de la distribución normal:*

Los datos normales son simétricos con respecto al promedio.

La media, moda y la mediana son aproximadamente iguales.

Se cumple generalmente la regla empírica:

- el 68% de los datos se alejan 1 desvío estándar de la media.
- el 95% de los datos se alejan 2 desvío estándar de la media.
- el 99.7% de los datos se alejan 3 desvío estándar de la media.
- Los valores atípicos se consideran más allá de 3 desvío.

Análisis Predictivo

Imagen tomada de https://www.lifeder.com/poblacion-estadistica/.

Es el proceso de realizar **predicciones** sobre una **población** basándonos en una **muestra.**

El término **análisis predictivo** describe la aplicación de una técnica estadística o de **aprendizaje automático** (ML) para crear una predicción sobre el futuro.

En este proceso se crean **modelos predictivos** para predecir eventos futuros.

Un *modelo predictivo* utiliza métodos matemáticos y de cálculo para predecir un evento o un resultado. Estos modelos pronostican un resultado del tipo **estados (Si/No)** o **valor numérico**.

> 💡 Un ejemplo podría ser con el clasificador de mails en *spam/no spam*: la clasificación de los mails se realiza gracias al modelo predictivo que se construye para **reconocer patrones**. Estos patrones pueden bien ser la probabilidad o frecuencia de ocurrencia de cada palabra en los textos de los mails de spam y en los textos de no-spam.

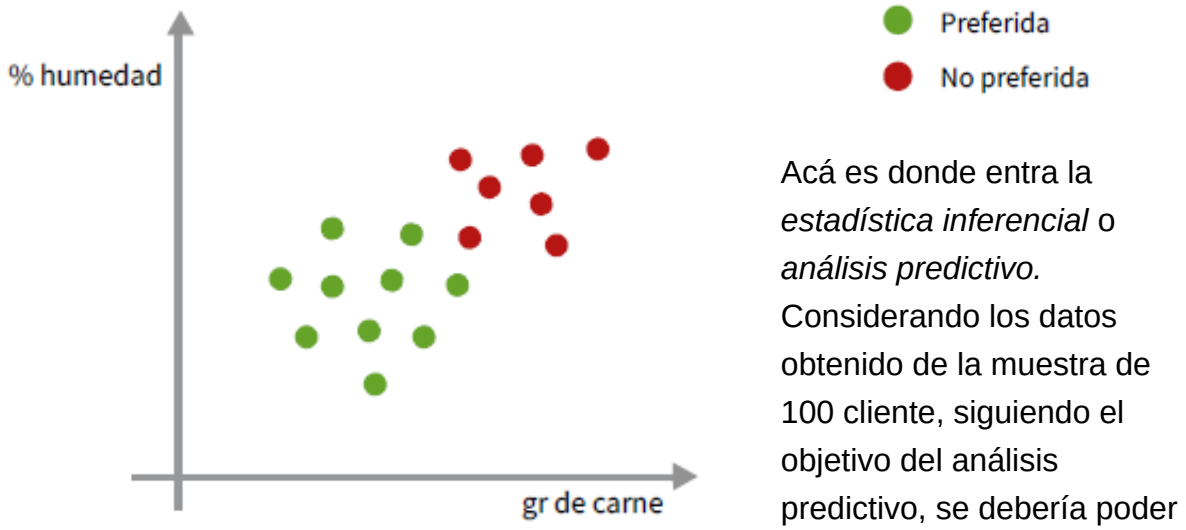
Objetivo del Análisis Predictivo

Tiene por objetivo: generar un **modelo** analítico de datos históricos para predecir (o inferir) el comportamiento de la población de la cual proceden los datos muestrales.

Supongamos un ejemplo: una empresa de hamburguesas quiere **mejorar sus ventas**. Mediante una encuesta realizada en 100 clientes, descubre que el **tamaño de la hamburguesa y la humedad importan**. Buscando la receta perfecta, cocinó nuevas hamburguesas variando sus **características**, y se las dio a otros **nuevos** 100 clientes. Esta vez encuestaron sobre el nivel de satisfacción de cada cliente. Para esto había dos posibles respuesta:

- Preferida: la volverían a elegir.
- No preferida: no la volverían a elegir.

Luego de este análisis estadístico elaboraron una gráfica con los datos de los clientes que probaron la nueva hamburguesa:



Puntos verdes: preferida. Puntos rojos: no preferida.

Acá es donde entra la *estadística inferencial* o *análisis predictivo.* Considerando los datos obtenido de la muestra de 100 cliente, siguiendo el objetivo del análisis predictivo, se debería poder generar un modelo que infiera el comportamiento de los clientes en base a las características de la hamburguesa: [tamaño,humedad].

Modelo

Un modelo es una **relación matemática** que relaciona una variable que queremos predecir con una o varias variables asociadas.

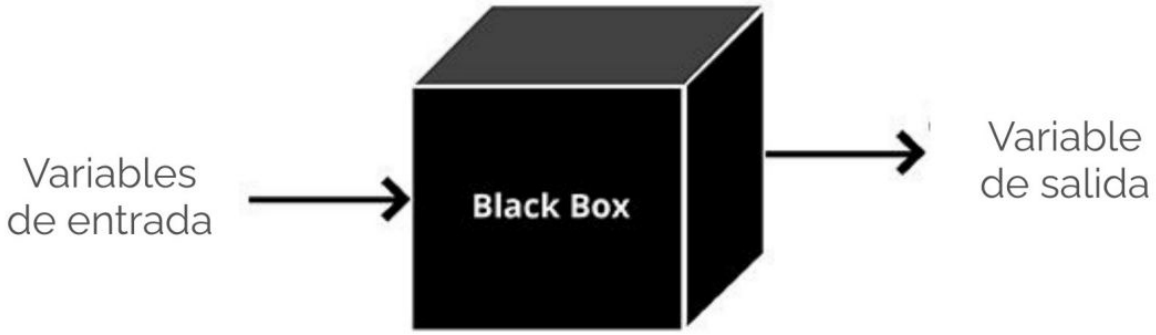
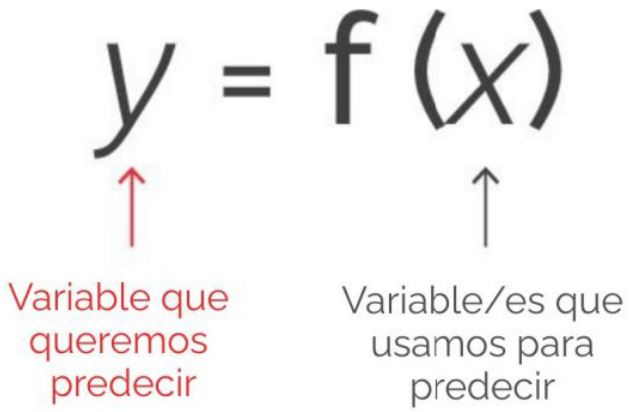


Diagrama de un modelo de Machine Learning.

> Un modelo es una función matemática.



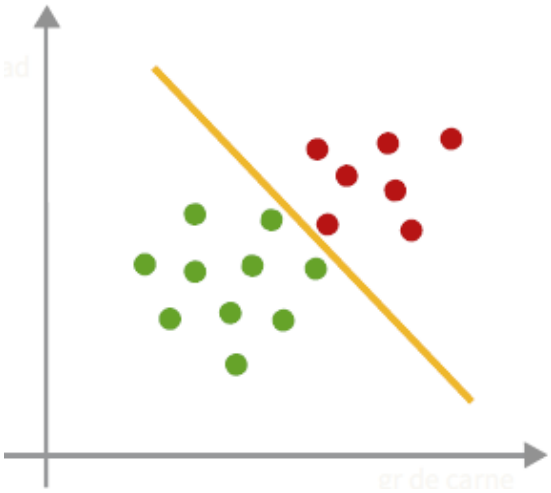
- **y:** variable dependiente, respuesta, explicada o target.
- **x:** variable independiente, regresora, explicativa o features.

Relación matemática entre variables.

Los modelos son *representaciones simplificadas* de la realidad, que utilizan una *porción* de la realidad empírica para **predecir** o estudiar el problema.

Por ejemplo: volviendo al caso de la ventas de hamburguesas. Un modelo sería capaz de establecer algún tipo de relación entre las variables **tamaño [gr]** y **humedad [%]** de manera de al tener nuevos datos ingresados de hamburguesas poder saber si los clientes las preferirán o no (en función de dichas variables). En la figura se puede observar una posible función

matemática que para valores que estén por encima de dicha línea los clientes prefieran las hamburguesas y viceversa.



¿Qué garantiza la calidad de un modelo?

1. Buena selección de los **factores relevantes**: ¿Cuáles variables debo tener en cuenta para representar mi modelo?.
2. Adecuada descripción de sus **relaciones funcionales**: ¿Cómo se relacionan estas variables con la variable que quiero predecir?

https://www.youtube.com/watch?v=JjEI5Qp0T_4

Supuestos o presunciones

Los modelos esconden supuestos, de otra forma sería más complicado encontrar una relación matemática para explicar la variable respuesta (target) en función de la variable predictora (feature).

La producción de leche de una granja era tan baja que el granjero pidió a la universidad local ayuda académica. La universidad reunió un equipo multidisciplinar de profesores, encabezado por un físico teórico, y estuvieron dos semanas haciendo investigación de campo intensiva. Los científicos volvieron a la universidad, con sus portátiles repletos de datos, y el encargo de escribir el informe se dejó para el líder del equipo. Poco después, el granjero recibió el informe, que empezaba así: *Tengo la solución, pero funciona solo en el caso de vacas esféricas en el vacío*. Harte, John (1988), Consider a Spherical Cow: A Course in Environmental Problem Solving, University Science Books, ISBN 978-0935702583.

Para calcular el volumen un objeto: ¿Cuáles son mis supuestos?

Construir un modelo es un proceso iterativo: se realiza un modelo sencillo, mediante un conjunto de datos acotado, se prueba y se valida para determinar su precisión, y si no cumple con lo esperado, se realiza otro modelo un poco más complejo hasta obtener el modelo más efectivo. ¿Nos serviría un modelo perfecto? ¿Existe?.

Modelizar es representar la realidad con una cantidad menor de información. Esto nos lleva a errores que son inherentes a todos los modelos, que pueden reducirse si, pero no ser eliminados por completo. Para esto tengo que conocer: ¿Cuánto es mi error?

Error del modelo

El error un modelo es la *diferencia* entre la *predicción* y el *valor real.*

El error es una variable aleatoria.

```
E = | Valor Real - Valor Predicho |
```

> La eliminación del error implicaría una **perfecta** identificación del **modelo** con el objeto **real**. Pero dejaríamos de tener un modelo y tendríamos la realidad con toda su complejidad. Debe buscarse un compromiso entre la **complejidad** del modelo y el **error** aceptable en los resultados.

Para saber que tan grande es el error de mi modelo existen métricas. Estas difieren un poco entre la aplicación final del modelo si es para **clasificación** o **regresión.**

Conocer el error permite mejorar:

- La selección de los componentes y mayor precisión.
- Mayor cantidad de componentes (aunque esto aumente la complejidad del modelo).

Métricas de error

Regresión

Las métricas de regresión se basan en el error entre cada observación real y el valor predicho, así tenemos las siguientes:

- Mean Absolute Error (mae):

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Mean Squared Error (mse):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- Root Mean Squared Error (rmse):

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

- Relative Squared Error: Si este valor está más cerca de 1, el modelo no tiene nada que mejorar, pero si el valor de R2 está más cerca de 0, el modelo debe mejorar más. R square = 1 - (MSEmodel/MSEbaseline) donde el *MSEbaseline* se obtiene utilizando el promedio de los valores para calcular el MSE.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

- Relative Squared Error Adjusted: si se agregan nuevas características el R2 no mejorará bien, aumentará o permanecerá igual pero nunca disminuirá, por lo que no podemos saber si **todas** las variables explicativas son significativas para el modelo. En estos casos podemos usar el **R2 Ajustado**. Esta nueva fórmula también tiene en cuenta el número de características y aumentará cuando las características sean significativas o puede disminuir si un efecto específico no mejora el modelo.

$$R^2_{ajust} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Clasificación

- Confusion Matrix: Esta es una métrica para problemas de clasificación. Es una matriz en la que nxn viene dado por el número de clases. Por ejemplo, en clasificación binaria tendremos matriz de confusión 2x2. Donde las filas representan los valores reales y las columnas representan el resultado de la predicción.

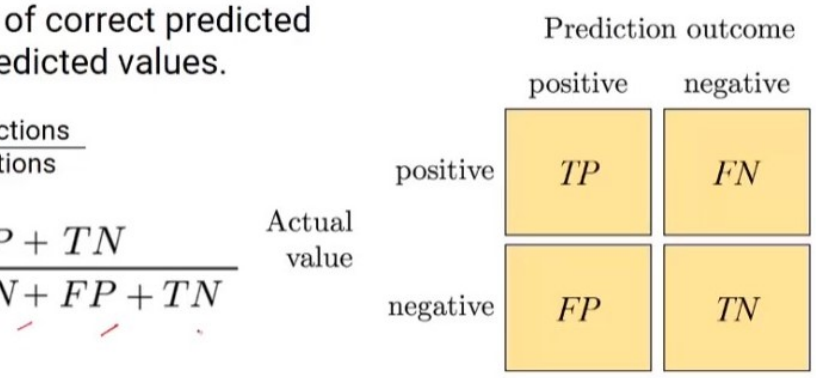
		Prediction outcome: positive	Prediction outcome: negative		
Actual value	positive	TP	FN	$TP + FN$	Total Actual positive
	negative	FP	TN	$FP + TN$	Total Actual negative

- Accuracy: Esta es una métrica **popular** para la tarea de clasificación. La definición formal dice: *es la proporción de valores de predicción correctos sobre el valor total de predicción*.

Accuracy: It is the ratio of correct predicted values over the total predicted values.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

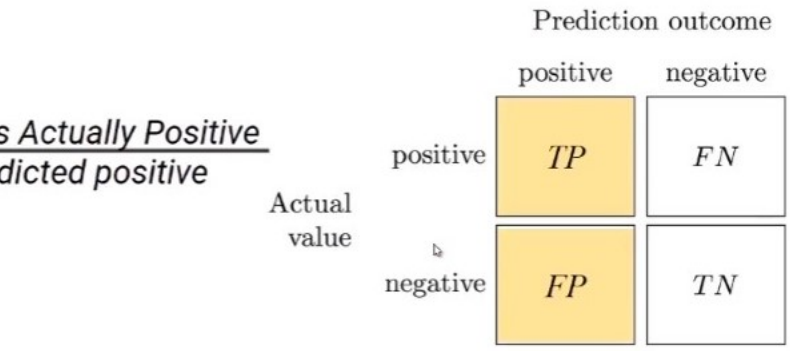


- Alternativas al accuracy (en problemas con dataset desbalanceado):
  - True Positive Rate: TPR = TP/(Tp + Fn) este valor alto es deseable.
  - False Negative Rate: FNR = FN/(Tp + Fn) este valor bajo representa un buen modelo.
  - True Negative Rate: TNR = TN/(Fp + Tn) este valor alto es deseable.
  - False Positive Rate: FPR = FP/(Fp + Tn) este valor bajo es deseable.
- Precision: Se define como la proporción de los valores predichos como positivos que en realidad fueron positivos (palabra clave: Predicción).

Out of all the positive predictions, how many are actually positive.

$$\textit{precision} = \frac{\textit{Predictions Actually Positive}}{\textit{Total Predicted positive}}$$

$$\text{precision} = \frac{TP}{TP + FP}$$



- Recall: De todos los positivos reales, ¿cuántos se han pronosticado como positivos?.

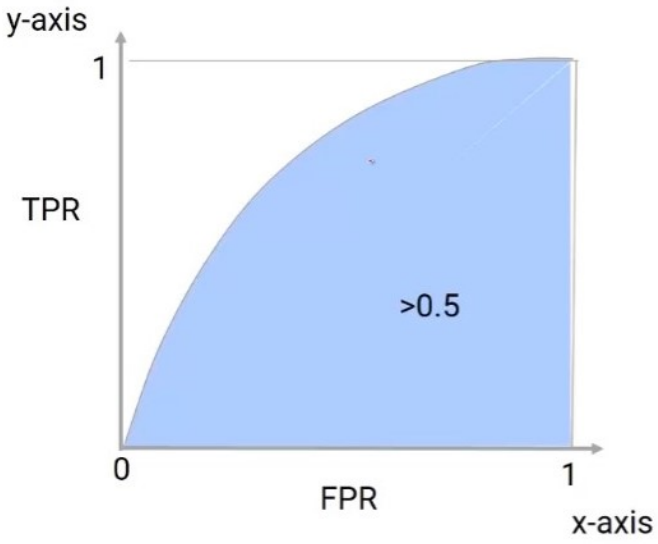
Out of all actual positive, how many are predicted positive.

$$recall = \frac{Predictions\ Actually\ Positive}{Total\ Actual\ Positive}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

		Prediction outcome: positive	Prediction outcome: negative
Actual value	positive	TP	FN
	negative	FP	TN

- AUC-ROC: respectivamente significan ***a**rea **u**nder **c**urve* y ***r**eceiver **o**perating **c**haracteristic.* Esta métrica es útil para evaluar clasificación **Binaria**. Nos muestra el equilibrio entre los verdaderos positivos y los falsos positivos.

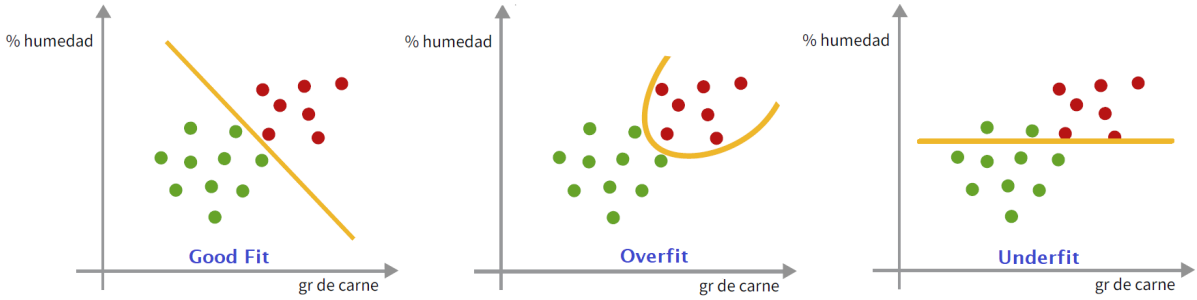


Optimización

Usando el ejemplo de las hamburguesas, podemos continuarlo estableciendo un modelo de **clasificación**, el cual lo que hará, es descubrir aquella relación matemática que mediante la entrada de datos para las dos variables **[tamaño,humedad]** se pueda predecir/inferir si el cliente quedará satisfecho o no.

Pero para asegurarnos de que esa predicción es valida habrá que estudiar las métricas del modelo de clasificación que vimos más arriba: *confusion matrix, accuracy, precision, recall, auc-roc (existen muchas más pero estas son las más usuales).*

En base a dichas métricas pueden aparecer tres posibles casos:



Formas de prevenir el overfitting:

1. Aumentar el numero de datos totales (censar más clientes, encuestas).
2. Ajustar los híperparámetros de nuestros modelos (depende del modelo).
3. Utilizar modelos más simples (reducir la cantidad de neuronas/capas en una red neuronal).
4. Controlar el número de iteraciones (Early Stopping).

Las formas de prevenir el underfitting son:

1. Tratamiento adecuado de los datos (limpieza y normalización).
2. Utilizar modelos más complejos (mayor cantidad de neuronas/capas en una red neuronal).
3. Ajuste de híperparámetros de nuestros modelos (depende del modelo).
4. Mayora cantidad de iteraciones del algoritmos (Early stopping).

Referencias

1. VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc."
2. Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.
3. Certificación Universitaria en Data Science. Mundos E - Universidad Nacional de Cordoba.
4. Certificación en Machine Learning - Machine Learning for Beginners. Analytics Vidhya.

Made with ❤️ Ignacio Bosch