**Especialización en Inteligencia Artificial**

**Taller Parte 2 semana 6 de Machine Learning**

**Profesor: Juan Carlos Valencia**

**Junio 2025**

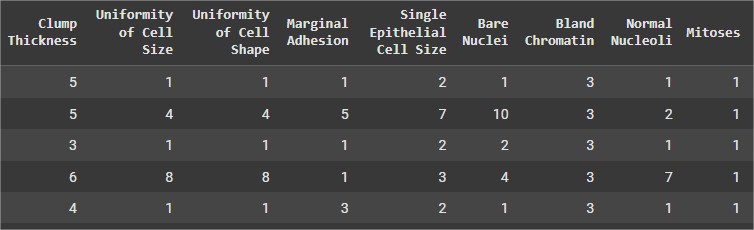
**¿Cuándo se aplican modelos de machine learning en problemas reales?**

Tenemos un dataset de cáncer de mama de Wisconsin. Estos datos fueron recolectados para estudio, poder tener un diagnóstico temprano y preciso, ya que este es fundamental para mejorar el tratamiento de los pacientes.

El objeto de estudio de este dataset es la clasificación de tumores que causan el cáncer de mama y esto se harán en dos posibles categorías:

* No maligno
* Maligno

A partir de características que se nos dan en el conjunto de datos que fueron recolectadas de las imágenes, se busca identificar formas y patrones que permitan predecir el diagnóstico final. Osea, el objetivo es desarrollar un modelo capaz de aprender, a partir de estas variables, si una o varias personas tiene alta probabilidad de tener células no malignas o lo contrario, maligna. Todo esto con el fin de asistir al personal médico en la toma de decisiones rápidamente.



Es la tabla anterior tenemos los primeros 5 registros del dataset para saber la info que se nos presenta. Esta info está clasificada de la siguiente forma:

**Clump Thickness**: Grosor del grupo de células

**Uniformity of Cell Size**: Tamaño celular

**Uniformity of Cell Shape**: Forma celular

**Marginal Adhesion**: Adhesión marginal

**Single Epithelial Cell Size**: Tamaño de las células epiteliales

**Bare Nuclei**: Núcleos sin citoplasma

**Bland Chromatin**: Cromatina normal

**Normal Nucleoli**: Nucléolos normales

**Mitoses**: Frecuencia de mitosis

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

En la imagen anterior utilizamos la columna Class solo para verificar el balanceo de clases, pero esta se omitirá para el diagnóstico.

Interfaz de usuario gráfica, Aplicación, Teams

El contenido generado por IA puede ser incorrecto.

Vemos que hay una alta correlación entre las siguientes variables:

* Cromatina y el tamaño celular -> 0.76
* Células epiteliales y el tamaño celular -> 0.75
* Cromatina y el forma celular -> 0.74

Texto

El contenido generado por IA puede ser incorrecto.

* El 96.4 de las predicciones fueron acertadas, es un buen indicio.
* El 93.9% de las predicciones que fueron realmente acertadas para los casos positivos, ese 2.5% de imprecisión que se pueden tener como falsos positivos, toca ajustar el modelo para no dar un mal diagnóstico. Sería decirle a una persona que tiene cancer aún cuando este está “sano”.
* Aunque hay pocos casos de errores, podemos ver que con el F1-Score se mantiene un buen equilibrio entre la precisión y el recall.

Gráfico

El contenido generado por IA puede ser incorrecto.

Y llevándolo a datos reales, vemos 3 casos de falsos positivos y solo 2 de falsos negativos lo que implica pacientes enfermos que no se les realizó tratamiento por mala predicción.

Podemos decir que el modelo, aunque preciso comete errores, los cuales se traducen a vidas perdidas o demandas por malos veredicto.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

La gráfica anterior indica que el malignancy\_index tiene un buen poder discriminativo entre ambas clases. Podemos ver los extremos en ambos boxes que deja por fuera valores que sobrepasan a estos, tienen valores atípicos, pero los malignos muestran una mayor cantidad de outliers.

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

En resumen, se puede concluir que el modelo de regresión es efectivo para estimar el índice de tumores malignos, lo que lo convierte en una herramienta muy buena para apoyar decisiones de diagnósticos y realizar seguimiento de pacientes con sospecha de enfermedad.

**Conclusión**

Se puede concluir que el índice de malignidad es de mucha utilidad para clasificar y predecir la presencia de tumores malignos en pacientes. Los modelos utilizados se pueden considerar “confiables” realizando ajustes para optimizar los resultados, reduciendo las fallas en las predicciones y que el porcentaje de casos acertados se acerque lo mas que se pueda al 100%.