**MACHINE LEARNING**
**Lecturer: Esteban Abelardo Hernandez Vargas**

The project test consists of 2 specific problems and a third one which is open.

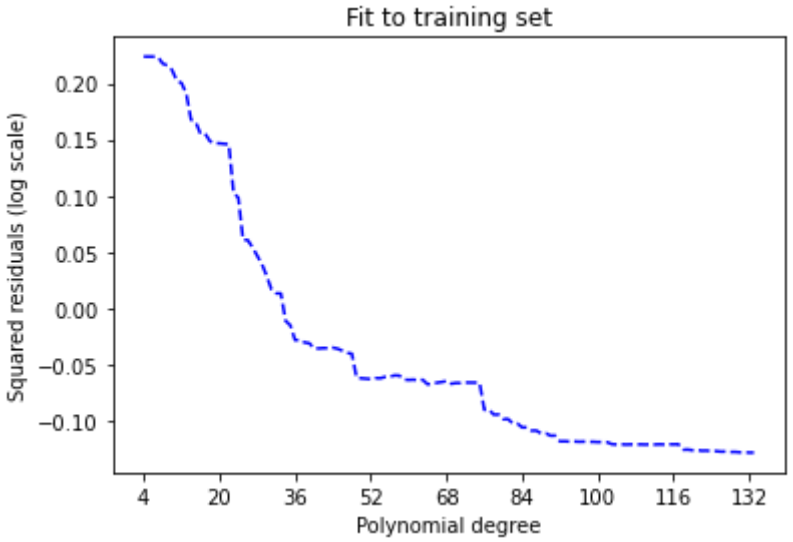
A brief written report with pictures and explanations of the results should be provided.

**Remark:** Provide a brief reasoning in the report why you choose certain algorithm.

**Student: Abraham Martínez López** **Matriculate number: 315671513**

**Problem 1**. Using the data set called „problem1.csv (x_training, y_training)“ :

a) Find the polynomial that fits the best the training data
**R:** We try for several degrees. From a constant (or zero polynomial) to a polynomial of the same degree as the amount of the data (133).



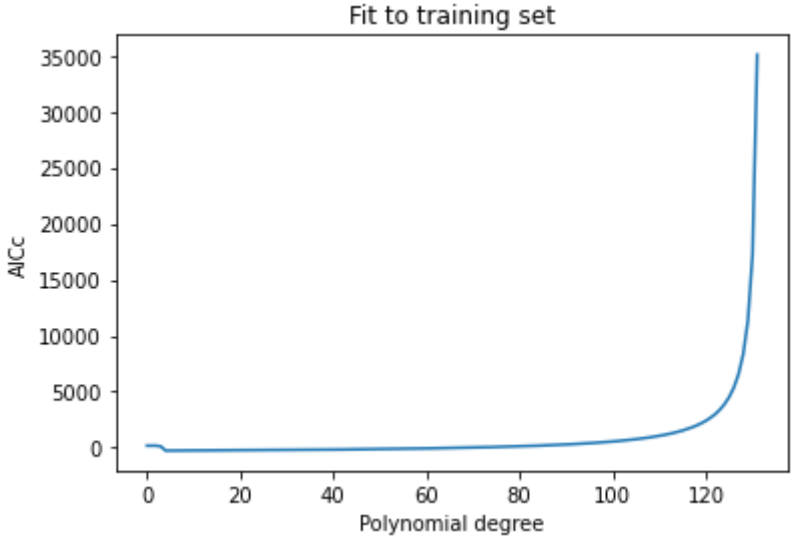

When we fit to the training data we can see the phenomenon of overfitting.

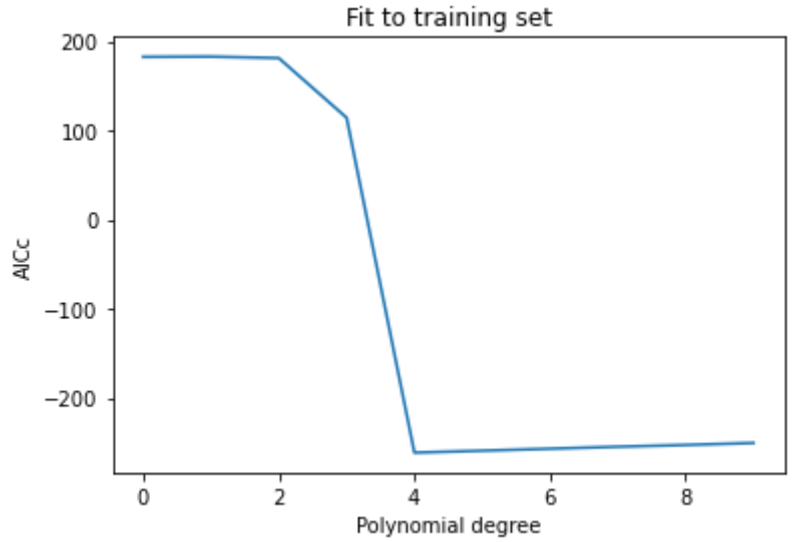
$$Squared\ residual\ = \sum_{i=0}^{133} (y_i - p_k(x_i))^2$$

where $p_k$ is the polynomial of order $k$ fitted to the data (Also note is the same seen in the course multiplied by 2). Due to numerical errors **the polynomial that fits the best the training data is** $k = 132$**,** still the difference between 132 and 133 is less than $0.0001$. And when we have a polynomial of degree greater than $36$we are warned about poorly conditioned fit. Even though the methods are ill conditioned we know of the existence of the best polynomial fitting that is the interpolation polynomial of degree 133.

b) Using the AIC criteria, find the best polynomial that can fit the data
**R:** We define Akaike information criterion (AIC) as

$$AICc\ = N\ log(\frac{Squared\ residual}{N}) + \frac{2MN}{N-M-1}$$

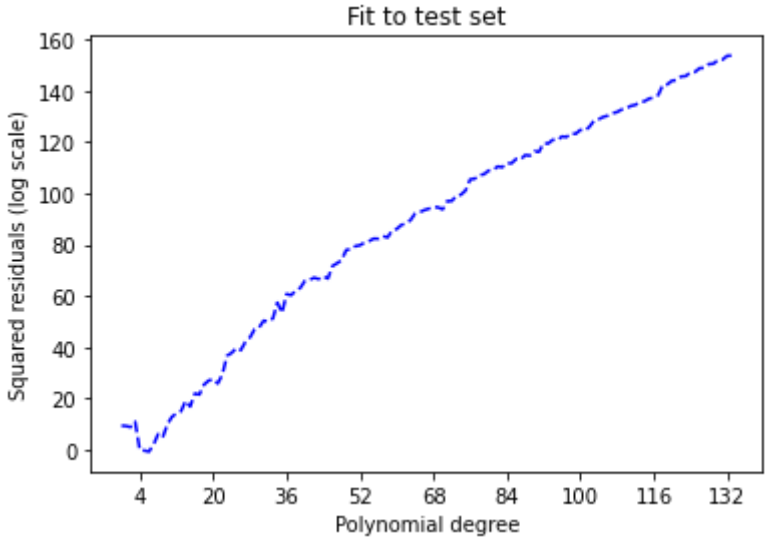


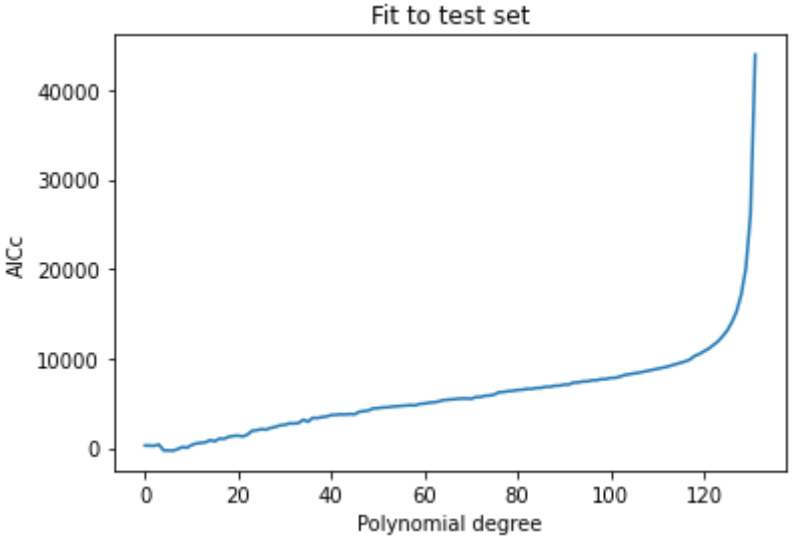




Taking the value with the less AICc value we got $k = 4$ with $AICc = -261.54$.

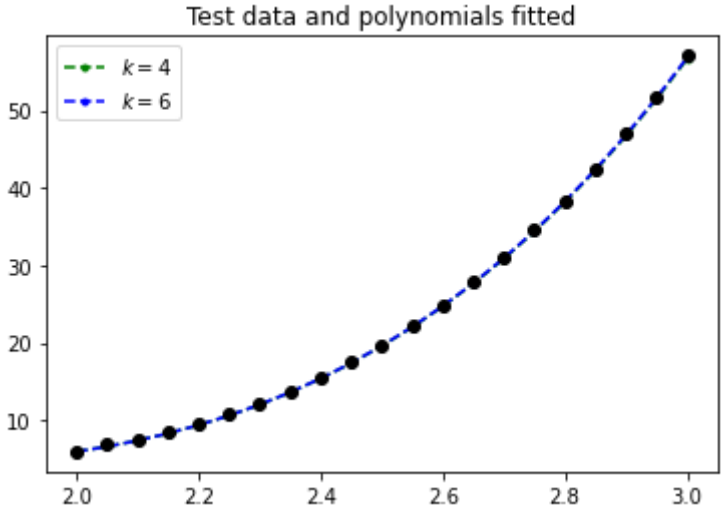
c) Cross validate the polynomial with the data set called "problem1.csv (x_test, y_test)"

**R:** In general, for every polynomial we got that the new scores are:

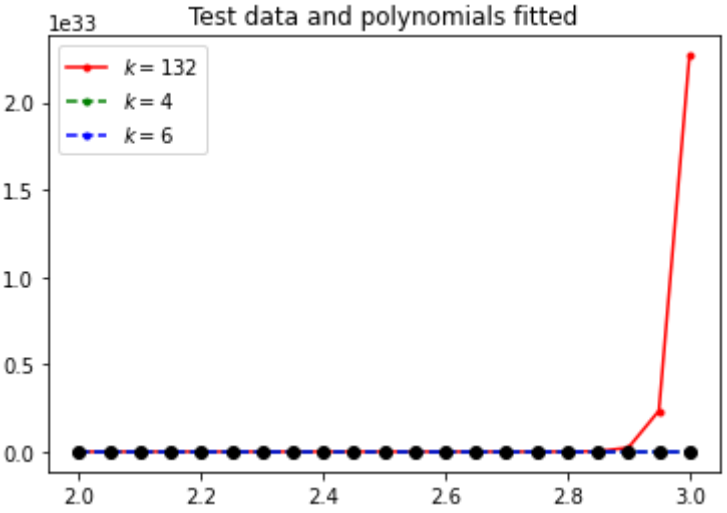





In both cases, the best was the polynomial of degree 6.



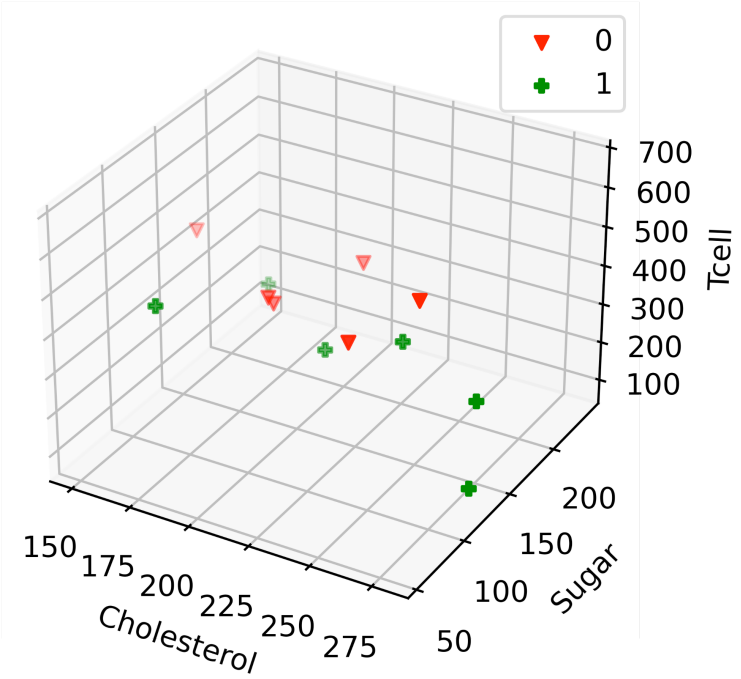

After plotting the results of fitting a polynomial of degree 4 and 6 we couldn't notice any differences. But the 132 degree polynomial performs terribly wrong in comparison.

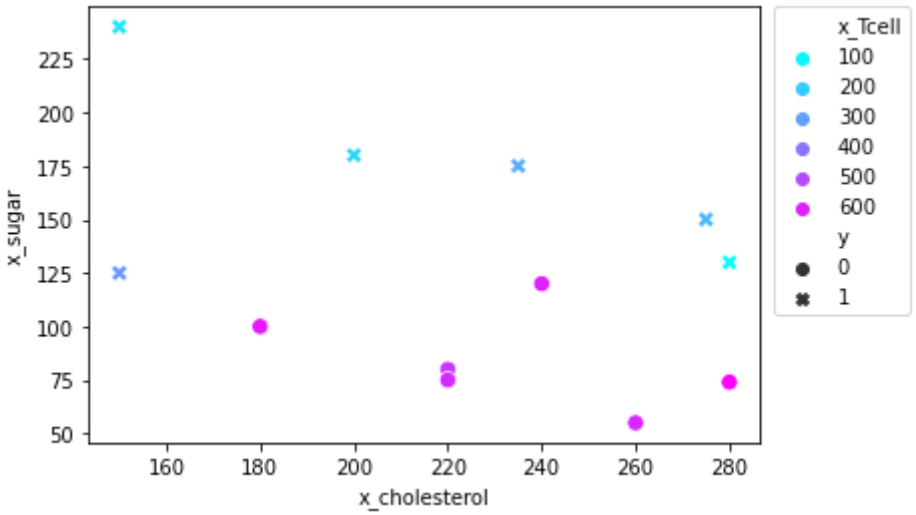


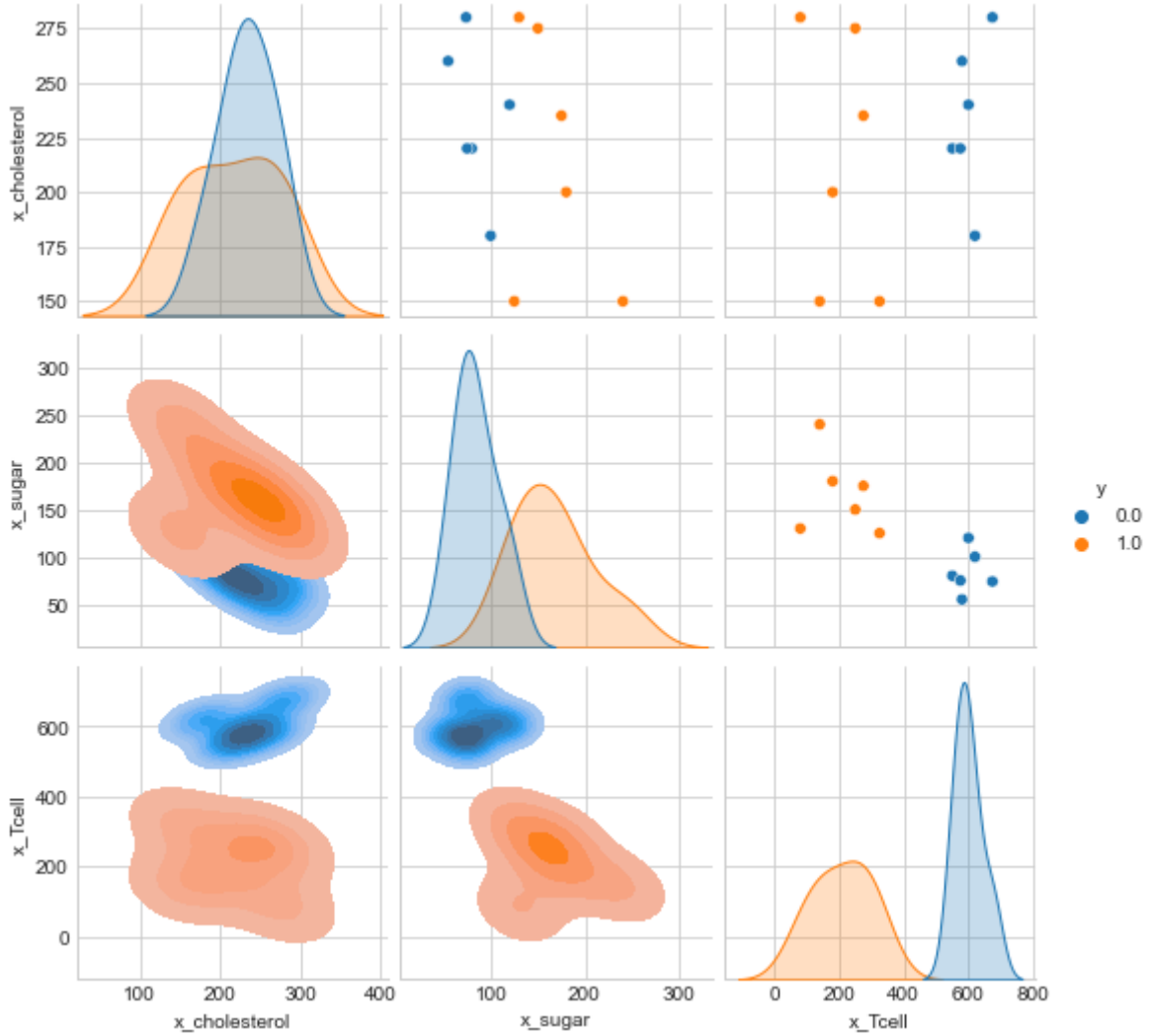
**Problem 2**. From a clinical trial, we have 12 patients with HIV infection. After treatment, the disease progressed in 6 patients (1) and in 6 patients the infection did not progress (0). Four measurements are taken in the 12 patients (Age, sugar levels, T cell levels and Cholesterol). Which measurement can be used as a marker to describe progression of the disease? Which will be the criteria to predict the progression? The data can be found in „problem2.csv (x_age, x_sugar, x_Tcell, x_cholesterol, outcome). Arrange the data and briefly explain your results. The variable "y" (target) is a vector of 0 and 1 to represent the progression.

**R:** We have a binary classification problem where HIV progressed. So we will look for the best profile to determine progress or not.

Our 12 data look like this:

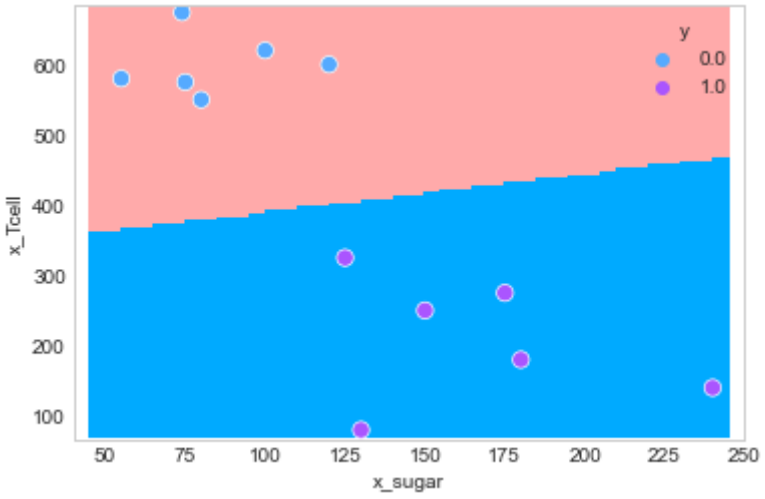


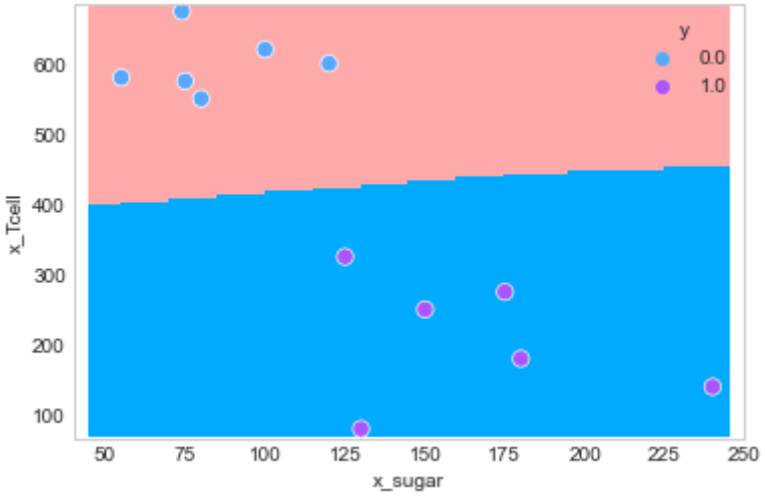



Notice how there is a trend. Where the amount of sugar and Tcells seem to play a major role in if a person will have a disease progress.

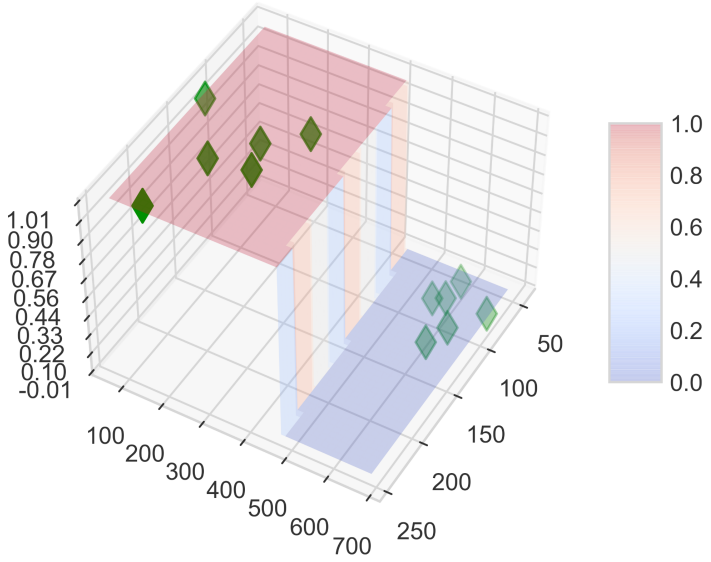
Now, doing a **Linear Discriminant Analysis** using Tcells and sugar we got a nice classification with a score on the training set of 1.




The difference between the linear and the quadratic are minimum. So we remain with the linear analysis. And with the K-N classifier is the same result.




Making a **logistic regression** we found the same:




So the **best criteria or measurement** is to check if the Tcells are above or below 400.

**Problem 3**. The third problem is flexible.

The student can either find any data set online and try to apply any of the tools we learn in the course

**<u>or</u>**

Make an assay of 500 words of one of the following papers

- Li *et al.* 2020. Accurate data-driven prediction does not mean high reproducibility
- LeCun_2015_Deep learning

Please explain in your own words what is the paper about, provide key aspects you consider relevant about.

**R:** Assay of Li *et al.* 2020. Accurate data-driven prediction does not mean high reproducibility. (In Spanish)

*No todo lo que brilla es oro*

Provervio Mexicano

Si tenemos un nuevo descubrimiento, entonces bajo las mismas condiciones y procedimientos otras personas también deberían poder descubrir lo mismo. Es lo que llamamos reproducibilidad de los resultados. El problema que expone la columna de Li *et al.* 2020 son los problemas que implica el uso de machine learning en los descubrimientos científicos.

La labor de un científico es responder a la pregunta del "¿Por qué?". Poder explicar las causas de los fenómenos que nos rodean. Para ello aplicamos una serie de cuestionamientos, propuestas y experimentos inspirados en lo que observamos día con día. Es por ello, que históricamente la tecnología y los nuevos instrumentos permiten a la humanidad observar más (los telescopios, microscopios, ondas de radio, sensores…). Ahora nos encontramos con una nueva herramienta, la de machine learning. Un conjunto de prácticas donde se combinan las matemáticas y la computación para poder obtener resultados de grandes volúmenes de información. El rasgo más sobresaliente de esta herramienta es su capacidad de predecir. ¿Pero basta para entender la naturaleza?

La respuesta es no. Machine learning tiene una gran capacidad para asociar datos y descubrir tendencias entre variables. Su valor es ese, que permite comprimir y resumir mucha información. Pero sigue siendo información no un descubrimiento. Y es que aunque se sabe que asociación no implica causalidad, pareciera que si se sabe más información de la asociación entonces ya descubrimos algo válido. Nada más alejado de la realidad. Hay múltiples formas de que esten conectados nuestros datos. Por ejemplo que haya una variable oculta que gobierne ambas (factor de confusión).

Una formulación en terminos probabilistas es: si tenemos X1,X2 como rasgos y Y como variable de respuesta, con machine learning encontraremos $P(Y|X1,X2)$ es decir, una distribución de probabilidad dada la evidencia. Pero por el teorema de bayes, este depende de $P(Y|X1,X2)=P(Y,X1,X2)/P(X1,X2)$ es decir de la probabilidad conjunta. Y la info de cómo se distribuye la conjunta no explica nada de cómo interactúan entre las variables. Así que aunque entendamos las tendencias de las variables X1,X2 y Y no sabemos nada de qué causa a qué.

Incluso hay discrepancias en los hallazgos de machine learning en el terreno práctico. Pues los datos que provienen de la misma fuente pueden crear un bias y al momento de hacer validación cruzada el bias se preserva. Sin contexto del experimento o fenómeno que estamos estudiando es probable que nuestro mismo modelo no explique las cosas. Incluso si agregamos más variables que de antemano sabemos que no determinan el fenómeno es fácil encontrar relaciones falsas. Un ejemplo es que los nutrientes y el calor hace que crezca una población de bacterias. Digamos que queremos saber dependiendo del número de bacterias la cantidad de nutrientes. Entonces ponemos toda la información (incluida la del calor) y resulta que el calor en nuestra regresión lineal es el rasgo más influyente, todo es correcto, pero si no sabemos nada de biología podemos cometer el error de afirmar que el calor es importante para la presencia de nutrientes (que es totalmente falso).

Los descubrimientos que se apoyan en machine learning tienen que ser estrictos y rigurosos con respecto a sus afirmaciones. También, no debemos concluir directamente sobre los buenos resultados en términos de precisión o validación cruzada. Pues a pesar de que mediante la máquina aprendemos, no significa que descubramos.