UREKA
de educatione flagranti

What is Correlation?

Correlation is a measure of how strongly one variable depends on another.

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases
- So, correlation varies between -1 and 1

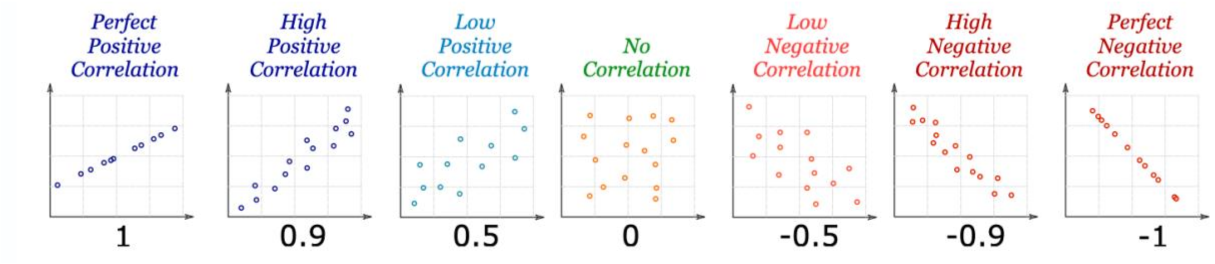


Image Credit: https://www.mathsisfun.com/data/correlation.html

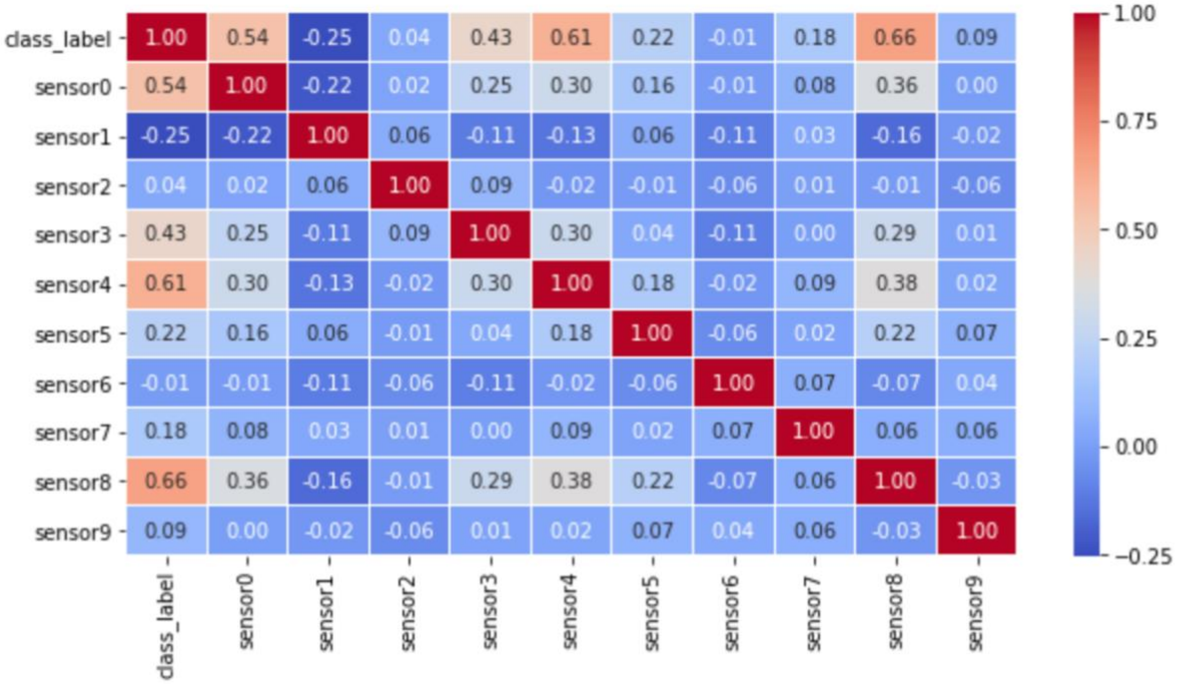
Remember that Correlation is not causation

"**Correlation is not causation**" means that just because two things **correlate does not** necessarily **mean** that one causes the other

The classic example of correlation not equaling causation can be found with ice cream and murder. That is, the rates of violent crime and murder have been known to jump when ice cream sales do. So, indeed buying ice cream and murders are highly correlated but it doesn’t mean that one causes the other.

Here is an example of Correlation Matrix:

We can infer that class_label is strongly positively correlated with sensor 8 and has low correlation with Sensor 2 and 6



	class_label	sensor0	sensor1	sensor2	sensor3	sensor4	sensor5	sensor6	sensor7	sensor8	sensor9
class_label	1.00	0.54	-0.25	0.04	0.43	0.61	0.22	-0.01	0.18	0.66	0.09
sensor0	0.54	1.00	-0.22	0.02	0.25	0.30	0.16	-0.01	0.08	0.36	0.00
sensor1	-0.25	-0.22	1.00	0.06	-0.11	-0.13	0.06	-0.11	0.03	-0.16	-0.02
sensor2	0.04	0.02	0.06	1.00	0.09	-0.02	-0.01	-0.06	0.01	-0.01	-0.06
sensor3	0.43	0.25	-0.11	0.09	1.00	0.30	0.04	-0.11	0.00	0.29	0.01
sensor4	0.61	0.30	-0.13	-0.02	0.30	1.00	0.18	-0.02	0.09	0.38	0.02
sensor5	0.22	0.16	0.06	-0.01	0.04	0.18	1.00	-0.06	0.02	0.22	0.07
sensor6	-0.01	-0.01	-0.11	-0.06	-0.11	-0.02	-0.06	1.00	0.07	-0.07	0.04
sensor7	0.18	0.08	0.03	0.01	0.00	0.09	0.02	0.07	1.00	0.06	0.06
sensor8	0.66	0.36	-0.16	-0.01	0.29	0.38	0.22	-0.07	0.06	1.00	-0.03
sensor9	0.09	0.00	-0.02	-0.06	0.01	0.02	0.07	0.04	0.06	-0.03	1.00

Side-notes:

What is Multicollinearity?

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related. We have perfect **multicollinearity** if, for example, the correlation between two independent variables is equal to 1 or −1.

One of the assumptions of Classic Vanilla Multiple Regression is no multicollinearity. **Multiple regression** assumes that the independent variables are not highly correlated with each other.

Why is Multicollinearity a problem with Multiple Regression?

A simple way to look at it is to consider the following linear equation:

$$Y = mX + nZ + c$$

here m and n are coefficients of X and Z respectively; here coefficient m is increase in Y for every unit increase in X while holding Z constant. However, in practice it is impossible to hold Z constant. So, the positive correlation between X and Z mean that a unit increase in X is accompanied with some increase in Z at the same time.

So, it might be a good idea to remove highly correlated variables when you perform Linear Regression.

However, algorithms like Regression with Regularization or Decision Trees have feature selection embedded in them. So, it is not a bad idea to keep all the features and then let these kinds of non-parametric models or regularization techniques handle multicollinearity.

What is the difference between parametric and non-parametric models?

Nonparametric statistics is the branch of statistics that is not based solely on parametrized families of probability distributions (common examples of parameters are the mean and variance). Nonparametric statistics is based on either being distribution-free or having a specified distribution but with the distribution's parameters unspecified. The term parametric refers to parameters that define the distribution of the data.

Examples of parametric models are Logistic Regression, Naïve Bayes while Decision Trees doesn’t make any assumptions regarding the distribution of the data hence, they are non-parametric.