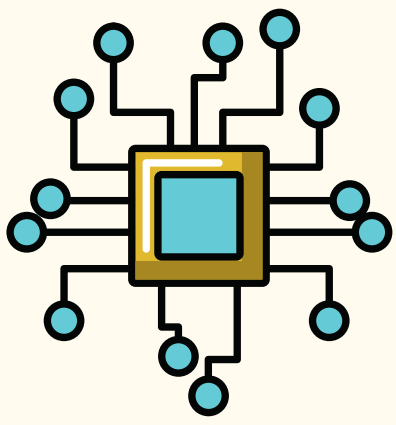
CHALLENGE

Survival Analysis vs Components' Production

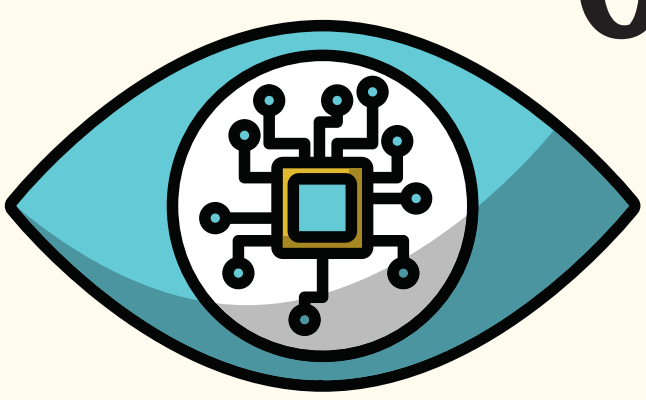
In 1988, an experiment was designed and implemented at one AT&T factory.
The goal was to investigate alternatives in the "wave soldering" procedure for mounting electronic components to printed circuit boards.
The response (or dependent variable) is the number of visible solder skips.

01 VARIABLE TRANSFORMATION

The dependent variable needs to be configured to be 0 if the event has not happened, or 1, if it has happened. That is a general rule for Data Mining: the inputs must be numeric.



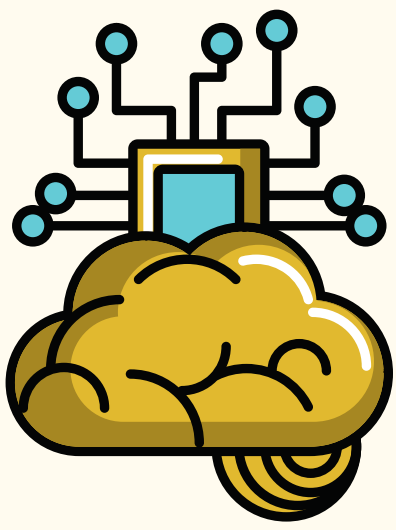
02 KAPLAN-MEYER ESTIMATOR

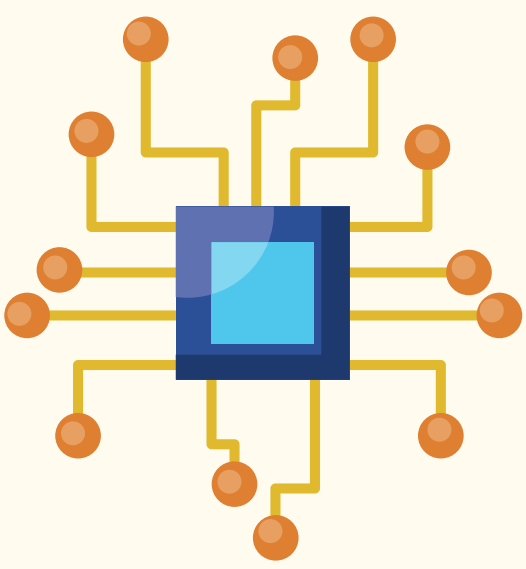


The Kaplan-Meyer Estimator is the algorithm to compute the Survival Curves. The goal is to measure the time until the event occurs.

03 VISUALIZATION

Plotting is of the key outputs of a typical survival analysis model. The usual shape is an inverted S. The goal is to see how long until a significant portion of the events start happening and until when is it expected for the event to happen.





04 LOG-RANK TEST

We can often segment the data into gender, age, clusters, etc. The log-rank test aims to test whether there are significant differences among the different groups. We need to use the function multivariate_logrank_test. Have a look and explore to see how it works.

Diogo Resende