Imagine that we have below medical data. We got some clinical measurements **Chest Pain, Good Blood Circulation, Blocked Arteries and weight**. and we want to apply a machine learning method to them to predict whether or not someone will develop **Heart Disease.**





To do this, we could use Logistic Regression or K-Nearest Neighbours or a Random Forests or some other methods. There are tons to choose from.

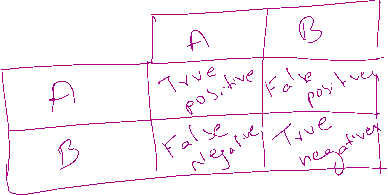


**How do we decide which one works best with our data?**

1. We start by dividing data into **Training** and **Testing** sets.
2. Then we train all of the methods **(Logistic Regression or K-Nearest Neighbours or a Random Forests or some other methods)** we are interested in with the T**raining** data.
3. And Test each method on the testing set.
4. Now we need to summarize how each method performed on the **Testing** data. One way to do this is by creating a Confusion Matrix for each method **(Logistic Regression or K-Nearest Neighbours or a Random Forests or some other methods)**







**A: Has Heart Disease**

**B: Does not have Heart Disease**

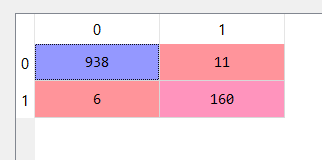
**Rows** in the **Confusion Matrix** correspond to what machine learning algorithm predicted and the **Columns** corresponds to the known **truth**. Since there are only two categories to choose from: **A or B.**

The top left corner contains **True Positives:** These are the patients that had heart disease that were correctly identified by the algorithm

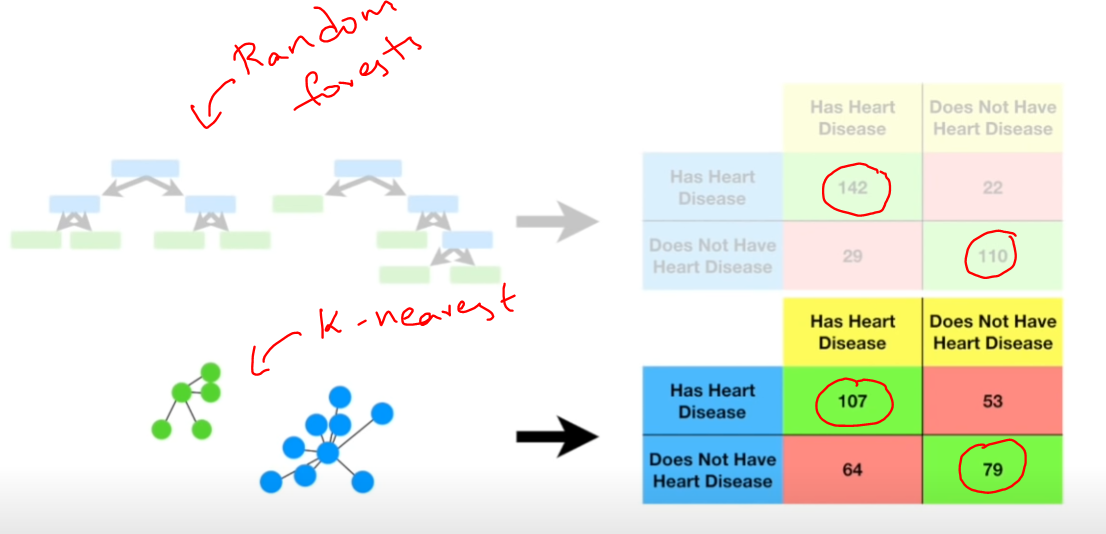
**True Negatives**: These are the patients that does not have heart disease and are correctly identified by the algorithm

**False Negatives:** These are the patients that had heart disease but the algorithm said they didn’t.

**False Positives:** These are the patients that doesn’t not have heart disease but the algorithm says they do.



We have to compare the confusion matrix of one method (K- nearest) to another method (Random Forests)



In the figure above, we can see **that Random forests method performed better**

