

# TYPE OF DOCUMENT DATA SCIENCE & A.I.

## Project Name Assessment: Breast Cancer Classification

High Potential Program

### 1 > Assignment

You are asked to analyze the dataset included with this assignment using either Python 3.x or R. The use of Jupyter Notebook is appreciated.

The dataset contains data from 157 patients diagnosed with three different types of breast cancer, including their DNA measurements. Specifically, we are interested in the type of breast cancer ('class' variable), and whether it can be predicted from other variables in the dataset (i.e. the DNA measurements).

Your analysis should include the following parts:

- 1 Create a breast cancer type classification model
- 2 Explain which rules and patterns in the data your model captures
- 3 Estimate the performance of your model on unseen data
- 4 List the top 5 most important predictors
- 5 Discuss how confident you are about these results (and why, or why not)

### Deliverables

- > A presentation in .pptx format, aimed at fellow Data Scientists who would like to know more about your prediction model and results (without having to look at your code). Feel free to show more about your methods and results, thought process, etc.
- > A zip file containing your code for assessing code quality.
- > A small summary of your main field of expertise. A lot of algorithms are available in the field of Data Science (Machine Learning, Artificial Intelligence and the like). A subset of these algorithms and techniques is suitable for this assessment. The assessment tests your knowledge about these algorithm and needed techniques. It could be that you are an expert in for example time series analysis. Therefore we ask you to include a ten line summary of your field of expertise. The summary could help you get a stage further in the selection process.

### Grading

You will be graded based on used methods, code, your insights into the results and clarity of presenting them.

For Data Scientists, we expect to see a good level in many of the following areas: data handling, programming, machine learning, statistics, mathematics, eagerness to learn, analytical thinking, communicating results and time management. Extra effort is always appreciated, but focus first on the task at hand

at hand.

## Guidelines

This task should take no longer than 6 hours to complete. Good luck!

© Anchormen > 08 October 2018 #02