innovaccer

# Problem Statement for "Data Science - Intern" role

**Question:**

This breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg's office. Each record represents follow-up data for one breast cancer case. These are consecutive patients seen by Dr. Wolberg since 1984, and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis.

This dataset contains 3 measures (mean, standard deviation, and worst) for 10 different cell features - radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. All these features have been computed from a digitized image of a fine needle aspirate of a breast mass, and they describe characteristics of the cell nuclei present in the image.

Then, the tumor size represents the diameter of excised tumor in centimeters, and lymph node status indicates the number of positive axillary lymph nodes. Outcome column represents whether the tumor in a given patient was recurrent (R) or non-recurrent (N). Time represents the recurrence time in outcome is recurrent (R), and disease-free time if the outcome is non-recurrent (N).

Analyse this dataset to answer the following questions.

a. Build a classifier to predict the outcome of a new patient with high accuracy. Also, remember that as a data-scientist working on healthcare problems, your intent should also be to minimize the number of false-negatives.
b. Build a regression model to predict the recurrence time for patients whose outcome is R.

Please find the Dataset for your reference: Click Here

---

***Note:** Please submit any codes that you write, a ppt (max 5 slides) explaining the algorithm (word limit 100) and results obtained, and an excel file showing prediction for the test dataset.

***Timeline:** Students need to upload their solutions latest by May 06, 2020 (Wednesday) before 9 AM

**Upload your Solutions:** Upload Here