## THE CHALLENGE: Zeta Disease Prediction

### INTRODUCTION

Mars Mission Control needs a good data-driven system for predicting Zeta Disease infection on the International Mars Colony.

Use the `\_zeta-disease\_training-data\_` dataset to build a model that can predict who will be infected by Zeta Disease.

Apply your model to the `\_zeta-disease\_prediction-data\_` dataset to predict who will be infected by Zeta Disease.

### DATASET

The dataset includes 9 columns with information on 800 people.

1. age : in years

2. weight : body weight in pounds (lbs)

3. bmi : Body Mass Index (weight in kg/(height in m)2)

4. blood\_pressure : resting blood pressure (mm Hg)

5. insulin\_test : inuslin test value

6. liver\_stress\_test : liver\_stress\_test value

7. cardio\_stress\_test : cardio\_stress\_test value

8. years\_smoking : number of years of smoking

9. zeta\_disease :

1 = yes;

0 = no

Solution approach:

Two scenarios have been created. Details are mentioned in respective Jupyter notebooks:

Scenario 1:

Using all the columns.

Classification matrix:

precision recall f1-score support

0 0.82 0.88 0.85 156

1 0.74 0.64 0.69 84

accuracy 0.80 240

macro avg 0.78 0.76 0.77 240

weighted avg 0.79 0.80 0.79 240

Scenario 2:

Dropping couple of columns

precision recall f1-score support

0 0.79 0.90 0.84 154

1 0.76 0.56 0.64 86

accuracy 0.78 240

macro avg 0.77 0.73 0.74 240

weighted avg 0.78 0.78 0.77 240

First scenario is providing better results. So model trained based upon scenario 1 should be used. But input data requires revisit, since in many columns a large number of data points are 0, which looks unusual based upon column definitions. Details are mentioned in respective jupyter notebooks regarding the data patterns.