- **Dataset**

  The project is based on breast cancer Wisconsin (diagnostic) dataset. The dataset was obtained from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

  However, some changes have been made in the original dataset. Therefore, I will be providing you the dataset in csv format.

- **Attribute Information:**

  Features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image.

  1. ID number
  2. Diagnosis (M = malignant, B = benign)

  Attributes 3 to attribute 32

  The mean, standard error and "worst" (largest) of these features were computed for each image, resulting in 30 features.

  1. radius (mean of distances from center to points on the perimeter)
  2. texture (standard deviation of gray-scale values)
  3. perimeter
  4. area
  5. smoothness (local variation in radius lengths)
  6. compactness (perimeter^2 / area - 1.0)
  7. concavity (severity of concave portions of the contour)
  8. concave points (number of concave portions of the contour)
  9. symmetry
  10. fractal dimension ("coastline approximation" - 1)

- **Instructions**

1. The project must be completed individually.

2. Conduct your exploratory analytics and try to understand your features
    a. Examine the type of variables (response and predictors)
    b. Apply descriptive statistics on your variables and provide a summary
        i. Mean, median, standard deviation, …
        ii. frequencies
    c. Inspect missing values
    d. Examine distribution of variables

3. Prepare data
    a. Transform variables where you see fit
    b. Did you engineer new features?
    c. Explain how you handled the outliers and missing data.

4. Examine relationship between diagnosis and other variables
    a. For numeric variables, create correlations matrix

5. Conduct classification data mining techniques
    a. Train classification algorithms (at least 3) on the dataset (tuning)
    b. Test the trained models on the test dataset
    c. For train-test split you can use single train-test split, 5-fold cross validation (cross-val-score) or bootstrapping. Note that you are free to select one.

6. Present findings
    a. Interpret results in a manner that is understandable to your manager.
    b. Present data exploration and analysis results.
    c. Present accuracy, precision, recall, ROC curve, AUC.
    d. Present in an organized and appealing style.

- **Deliverables**
    a. A report of at most 8 pages (Single Column, Font 11.5 Times New Roman and Single Space) including all tables and figures.
    b. Your python notebook (jupyter) that was used to prepare, train and test your dataset for your analysis.
    c. The submission deadline is May 7 by midnight. Feel free to submit it before the deadline.