**Predicting cancer using classification**

A health facility has been experiencing a lot of death due to cancer with a few survivors. They’ve recently been told on how AI can be used to predict whether a cancer case is benign or malignant on the fly without having the patient go through rigorous/expensive procedures. They now seek your expertise as a Data Scientist and have provided you with a dataset with the link below:

<https://drive.google.com/file/d/1GgLd-Pmtl6HPoxpdgn9MgvGQJb88gJNp/view?usp=sharing>

For this task, the stakeholders are interested in ensuring that the model predicts malignant cases correctly i.e if a case is malignant, the model should be able to identify accurately that it is malignant.

1. Define an approach you would take to solve the problem and document it
2. Get the data and determine what type of machine learning problem it is
3. Outline possible algorithms you would use to create the model.

Hint :

So far you know about Logistic regression. There are a host of other regression algorithms : Decision Tree, Random Forest, K-Nearest Neighbours (KNN) etc. Do your research.

1. Conduct exploratory analysis to understand the distribution of variables, identify any correlations, and gain insights into the dataset.
2. Handle missing values, encode categorical variables, and scale numerical features if necessary.

Hint: Encoding is converting from categorical to numerical using libraries like Label Encoder and OHE (One Hot Encoder). Scaling features is ensuring there’s no disparity in the variations for the features so that no feature is given preference. There are libraries like Standard Scaling and Min-Max scaling for that

1. Extract additional features if needed, such as interaction terms or polynomial features.
2. Train the selected model on the training dataset.
3. Evaluate the trained model's performance on the test dataset using appropriate evaluation metrics such as accuracy, precision and recall.