**Introduction:**

I am planning to work on Breast Cancer prediction. As Cancer is troubling the world and especially Breast Cancer in women around the world, I plan to work on the available public datasets, which will help us identify cancer at its early stages depending on the various attributes provided by the patients who have been diagnosed with Cancer.

 I will explore the Breast Cancer dataset and develop a Logistic Regression model to try classifying suspected cells as Benign or Malignant. Also, with the other datasets which are using patient details like age, health conditions, previous cancer, etc., I want to develop a Decision Tree Classification to predict whether new patient data received has Cancer or not (depending on the symptoms stated).

For this project, I will be using datasets from the below websites, which are public:

1. <https://www.kaggle.com/code/architkuiya/breast-cancer-detection/data>
2. <https://www.kaggle.com/datasets/adhyanmaji31/breast-cancer-prediction>
3. <https://www.kaggle.com/datasets/haithemhermessi/breast-cancer-screening-data-set>
4. <https://www.kaggle.com/datasets/puspitasaha/breast-cancer-prediction>
5. <https://www.kaggle.com/code/kiransheshma/breast-cancer-exploratory-data-analysis/data>

**What types of models do you plan to use and why:**

I am planning to use the below models:

1. **Logistic Regression:** This model is used to predict if a given patient has a Malignant tumor or a Benign tumor using the various attributes provided in the dataset.
2. **Decision Tree Implementation:** This model tree-based classification model. With the available attributes, it will predict the target variable, which is cancer tumor or not, using various independent variables (points that can help predict cancer like age, gender, etc.)

I am using two different datasets; one will be used for Logistic Regression and the other for Decision Tree Classification.

If time permits and any idea strikes my mind, I will try other models that best suit my analysis.

**How do you plan to evaluate your results:**

I will use an accuracy score and confusion matrix to evaluate the results using a train test data split (80-20) which I have been doing so far.

**What do you hope to learn:**

Using this analysis, I am hoping to learn what are the factors which would cause Breast Cancer. Also, what are the symptoms of Breast Cancer prediction? When we know these details, it can help people get tested ahead and avoid Cancer deaths.

**Assess any risks and ethical implications with your proposal:**

1. As the data is sourced from a public website and not a government health departmental source, we are unsure if the information is accurate.
2. This is not recent data, and as research evolves and might have added a few more attributes, we might be missing those extra parameters. However, I will be researching more datasets to get the latest data.

**Identify a contingency plan if your original project plan does not work out:**

This dataset should suffice with the initial research done for finalizing the dataset. However, I will look for the most recent data, primarily from the sources like Health Departments, to see if that will add more value to the analysis.

In case of any data imbalance issues, I will use techniques like SMOTE to make the prediction accurate.

If the Breast Cancer data is not helping to proceed further, I have another contingency plan to use Online fraud detection datasets.

**Include anything else you believe is important:**

Apart from all the details stated above, I will also look for any other ways we have to evaluate the results.