Task: Build a Colon Cancer Prediction Model from Multiple Datasets

Description:

You are given two separate datasets (Dataset 1 and Dataset 2) with different columns, and these datasets are unrelated. Your task is to build a machine learning model from scratch that predicts the type of colon cancer a patient has. The final model should be able to take inputs from both datasets and return a prediction for the type of colon cancer, with maximum accuracy.

Deliverables:

1. **Data Preprocessing**:

   Initial analysis of each dataset, with visualizations showing relationships, distributions, and any potential correlations between features.

   Figure out how Dataset 1 and Dataset 2 can be merged appropriately, so that it can be used to train the model.

   Treat outliers effectively.

2. **Model Building**:

   Build the model from scratch using any appropriate algorithm, experiment with at least two different machine learning models. Ensure that the model predicts the correct type of colon cancer based on the provided datasets.

3. **Model Optimization**:

   Derive new features based on domain knowledge (e.g., combining features that might indicate health risk factors) to improve the model's predictive power.

   Perform optimization technique, feature engineering ensemble techniques etc to optimize the model.

   Evaluate the model using appropriate metrics such as accuracy, precision, recall, F1-score, etc.

4. **Additional Enhancements**:

   Include feature importance analysis (using techniques such as SHAP values or model feature importances).

5. **Documentation**:

   Upload the entire code to a GitHub repository and on google collab to run the code, and share the links, make sure links are public.

   Provide detailed documentation explaining each step of the process, in the readme file.

   Take screenshots of the key steps (data preprocessing, model training, evaluation, etc.) and add them to the documentation.

Submission: -**Deadline**: 3 days from the date the assignment is received.