Dataset Proposal

Link to the Dataset

https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence/

Dataset description and predictive problem

This data set contains 13 clinicopathologic features aiming to predict recurrence of well differentiated thyroid cancer. The data set was collected in duration of 15 years and each patient was followed for at least 10 years.
The predictive task is to predict the recurrence of thyroid cancer among the participants in the following years.

Dataset Profile

- Number of Rows: 383
- Number of Attributes: 16
- Attributes:
    - ***Age*** *- The participant's age. integer*
    - ***Gender*** *- The gender class of participants. categorical*
    - ***Smoking*** *- Whether the participant is currently smoking or not. categorical*
    - ***Hx Smoking*** *- Whether the participant used to be a smoker. categorical*
    - ***Hx Radiotherapy*** *- Whether the participant used to be threated with radiations. categorical*
    - ***Thyroid Function*** *- The participant’s outcome of the thyroid function test. categorical*
    - ***Physical Examination*** *- The participant’s outcome of his physical condition. categorical*
    - ***Adenopathy*** *- Describes in which adenopathy the participant has cancer. categorical*
    - ***Pathology****- The type of thyroid cancer. categorical*
    - ***Focality*** *- The focal of the participant's cancer. categorical*
    - ***Risk*** *- Whether the thyroid cancer is life threatening. categorical*
    - *T - Size of tumour. categorical*
    - ***N*** *- Lymph node involvement. categorical*
    - ***M*** *- The distant metastasis of the participant tumour. categorical*

- ***Stage -*** *Whether the cancer has spread to other organs or parts of the body. categorical*
- ***Response -*** *categorical*

Challenge/difficulty exist in the dataset

1. **Dataset size** - The dataset is medical data in the field of thyroid function. There are 383 participants that assemble this data set.
2. **Sampling time -** The samples in the datasets were sampled during 15 years and not in the same time, that might affect the means of determining the value of each feature for each participant. This difference might harden creating an efficient model.
3. **Dependencies between Attributes** - some of the attributes have dependencies which need to be considered, for example:*Smoking and Hx Smoking*.
4. ***Partially* participant's *medical background*** - *not all the background of each* participant is contained in the data *because this is a medical dataset which has limitations of privacy and ethical reasons. This is a challenge we need to deal with while we build the mode..*