Data Science Final Project

Context

Irrespective of whether or not data and images are stored/analyzed in a centralized manner, variability in scanner models, acquisition protocols and reconstruction settings are unavoidable in the current clinical practice. Yet radiomics are notoriously sensitive to such protocol variations. Hence, there is a clear need for the harmonization of features in order to allow consistent findings in radiomics multicenter studies.

Objective

The objective of this project is to develop different models to predict failure (endpoint) of the radiomics signature based from MRI, PET and CT scans.

Dataset

radiomics.csv contains **197** rows and **431** columns:

Failure.binary: binary property to predict

You can split the dataset as you want to create the training/validation/test datasets

Models

You have to deliver three different models:

Model1

- Create an **ensemble classification model** (atleast 3 models of your choice).
- Preprocess the data
  - Check for null and missing values
  - Check for normality, if not, normalized the data
  - Get the correlation of the whole data expect the categorical variables
- Split the data into training (80%) and testing (20%)
- Print the AUC values during Training
- Print the Top 20 important features during Training
- Print the AUC values during Testing

Model2

- Create a neural **network-based classification model**.
- Create five hidden layers with 256, 128, 128, 64 and 64 neurons, respectively with activation functions of Sigmoid

- Create an output layer with two neurons respectively with activation functions of Softmax.
- Every layer is followed by a dropout to avoid overfitting.
- Copy the slide 15 backpropagation compiler approach.
- Copy the slide 33 model compiler approach.
- Train the model with epoch = 10, batch size = 128 and validation split = 0.15 (reference slide 33).
- Evaluate the trained model using the testing dataset.
- Get the model prediction using the testing dataset.

Model3

- Without considering the binary output and categorical variables in the dataset, compare the following clustering technique results:
    - K-Means
    - Hierarchical
    - Model Based

Application

To deliver:

- 3 (Model1, Model2, and Model13) R Markdown
- 3 (Model1, Model2, and Model13) PDF Files from an R Markdown Outputs
- These should be pushed into your final github repository.
- Name the repo as INFS692

Readme file

This md file must have the documentation about your application, models, packaging, setup and dockerization

Git

Use git to version your code and push it to any public repository and send/upload the link in the MyCourses before December 16, 2022 at 11:59PM.

Evaluation

You will be evaluated on:

- Quality and structure of your codes
- Models architecture
- Git commits quality
- Data preparation and preprocessing
- Documentation
- The quality of your answers

The accuracy of the model is important but it will not be the main criteria to evaluate your project. Please be aware that I will be able to track if you share the same code with your classmates or copy the codes from an existing Github/Kaggle available online.

Good Luck!!!