README

Yanfei Chen 261018456

Throughout this project, I implement three different models to predict the failure of the radiomics signature using all the continuous features. The dataset I used is the radiomics.csv file provided on MyCourses and the label used to train model is the Failure.binary. The dataset contains 197 rows and 498 columns.

There are three PDF file and three RMD file in this repository. Each pdf together with the corresponding rmd file denotes to one model.

First task for all three models is data pre-processing work to check null/missing value and normalize all continuous values. Then the dataset was splitted into training and test dataset with the probability of 80%.

Model 1 (Model 1.PDF + Model 1.RMD): The first model is an ensemble classification model using KNN, Decision Tree, and Random Forest. By simply using the KNN, Decision Tree, and Random Forest model, we gain the AUC value of 83.7%, 92%, and 95.9%, correspondingly. By ensemble the three models together using the average of the predict probability, we gain a test accuracy of 92.5%.
Package used: rsample, readr, rpart, caret, rpart.plot, ROCR, pROC, dplyr, vip

Model 2 (Model 2.PDF + Model 2.RMD): The second model is a neural network-based model using 5 hidden layers with sigmoid activation function and 1 output layer with softmax activation function. The test accuracy for the second model is 65.8%.
Package used: readr, caret, dplyr, rsample, keras, tensorflow

(**Note:** The instruction requires 10 neurons in the output layer, but our label contain only two class which is 1 and 0. So, I set the number of neurons at the output layer as 2)

Model 3 (Model 3.PDF + Model 3.RMD): The third model does not consider the output value. I used K-Means, Hierarchical, Model Based to cluster all the data. For K-means, I choose the model with 2 clusters. For Hierarchical, the silhouette method suggests 2 clusters while gap statistic method suggest 9 clusters. I also cut the dendogram at height 70.5 and created the sub plots. For Mode Bases, I choose the model with 10 clusters.
Package used: rsample, dplyr, readr, factoextra, cluster, stringr, gridExtra, mclust, tidyverse

(**Note:** There is an Rmarkdown output error with the Hierarchical plot when I tried to output as PDF file. The error is "! LaTeX Error: Unicode character ^^[(U+001B)] not set up for use with LaTeX." But I can knit the same RMD file into Word file successfully. I spend 6 hours trying to fix it, but still failed to knit the PDF. So, I knit the RMD into Word and transfer the Word file into a PDF one. Hope it won't influence my feedback for the project)