Abdulaziz Almuhaidib

T5 Tuwaiq DS Bootcamp

21October 2021

Final Project

Abstract

The project determines whether the patient has survived respiratory cancer or not. Moreover, would early checkup make huge difference? The dataset used in this project is generated from the surveillance, epidemiology, and end results (SEER) center dataset which contains over 9.5 million records. This dataset has 182099 records of patient who have respiratory cancer in the years 2008 to 2012. The dataset has 13 features including one target.

Design

Two data sets were conducted from the main SEER dataset. The First dataset was all numerical, so it is ready to be used in classifier. On the other hand, the second dataset has 9 categorical feathers that had been converted to have a better understanding of the data. Both datasets had 12 feathers that had been selected based on common selection in many published papers and the 13th features (target) was determined as shown below:

If survival month is greater or equal to 60 months and vital status recode is alive:

Then the patient is a survival

Else:

If survival month is less than 60 months and cause of death is respiratory *cancer:*

Then the patient is not a survival

Else:

Then the patient is died of other cause

Data

The dataset contains 182099 patient records with 13 features for each, 9 of which are categorical. Most feature highlights include measurements of disease, summery stage, grade, race, vital and signs.

Algorithms

Models

Random Forest and k-nearest neighbors classifiers were used before settling on random forest as the model with strongest cross-validation performance. Random forest feature importance ranking was used directly to guide the choice and order of variables to be included as the model underwent refinement.

Model Evaluation and Selection

The entire training dataset of 182099 records was split into 80/20 train. and all scores reported below were calculated with 10-fold cross validation on the training portion only.

The metric for Survivability was classification rate (accuracy).

Final random forest 10-fold CV average scores:

- Accuracy 0.939

K-Nearest Neighbors 10-fold CV average scores:

- Accuracy: 0.926

Tools

- Glob, Re, and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib for plotting