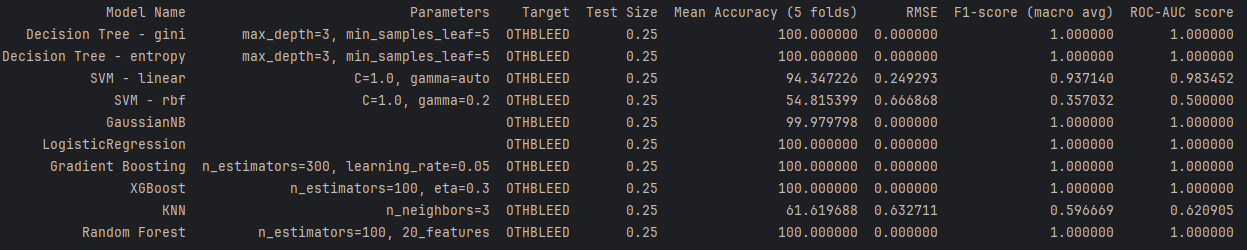
**Model Notes**

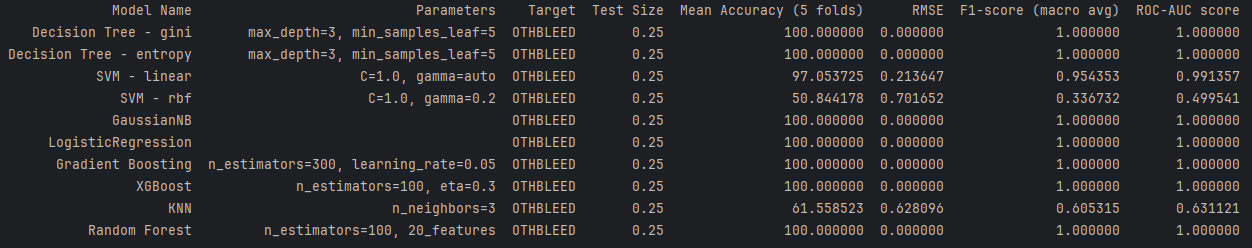
# Iteration 1: benchmark

* Year 2018-2020
* all features with less than 50% missing
* features: 128, observations: 4953 (shape: 4953x129)
* imputations applied, no standardization
* dataset: to be provided



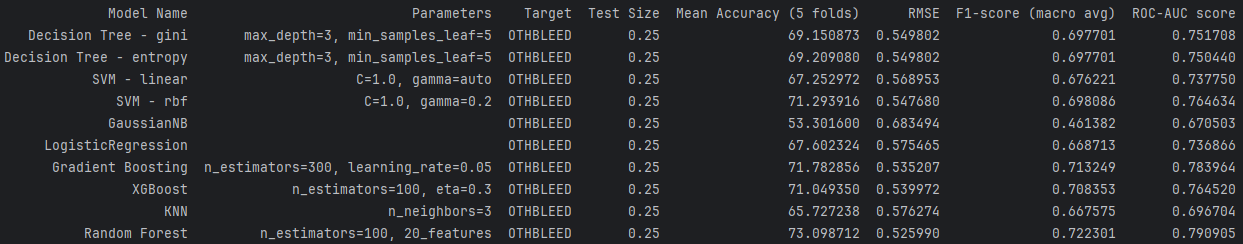
# Iteration 2: more recent data – extended to 2022

* Year 2018-2022
* all features with less than 50% missing
* features: 128, observations: 8587 (shape: 8587x129)
* imputations, no standardization
* dataset: [CABG\_5yr\_baseline.csv](https://github.com/jennytsai32/Capstone/blob/10937f26fe011d25038340276e3ba2534b3ab566/code/main_code/processed_data/2018_2022/CABG_5yr_baseline.csv)



# Iteration 3: selected features; standardization

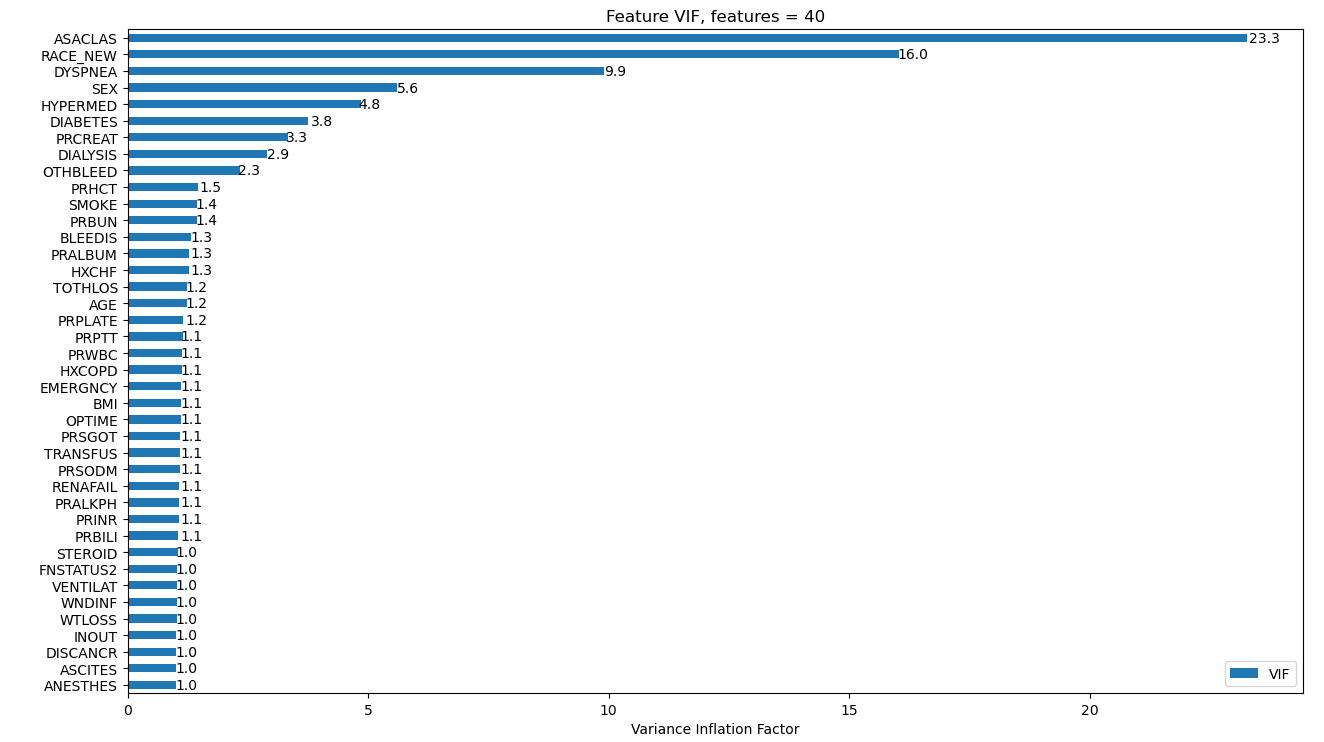
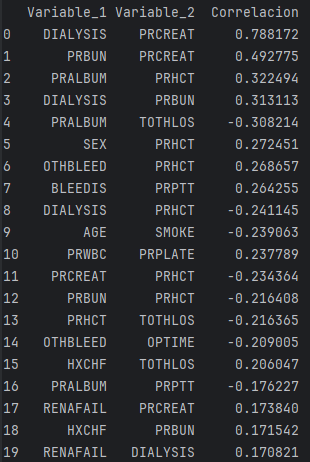
* Year 2018-2022
* features: 40, observations: 8587 (shape: 8587x41)
  + Dr. Gupta selected 43 features
  + 'HEIGHT','WEIGHT','ETHNICITY\_HISPANIC' was dropped for obvious multicollinearity issue (see graph below)
* imputations, standardization
* dataset: [GABG\_5yr\_preselect40.csv](https://raw.githubusercontent.com/jennytsai32/Capstone/10937f26fe011d25038340276e3ba2534b3ab566/code/main_code/processed_data/2018_2022/CABG_5yr_preselect40.csv)



# Iteration 4: feature selection

* Year 2018-2022
* features: 40, observations: 8587 (shape: 8587x41)

a. Features VIF Ranking b. Features Correlation Pairs Ranking

# Iteration 2:

* addressing multicollinearity – VIF and correlation check (then drop highly correlated features)
* 43 features – selected by Dr. Gupta

# Iteration 3:

* Extending to Year 2018-2022
* addressing missing values – imputations
* addressing multicollinearity – VIF and correlation check (then drop highly correlated features)
* 40 features included

# Iteration 4: Feature Selection

# Iteration 5: Feature Engineering (AutoML)

BACK-UP Notes

43 features

