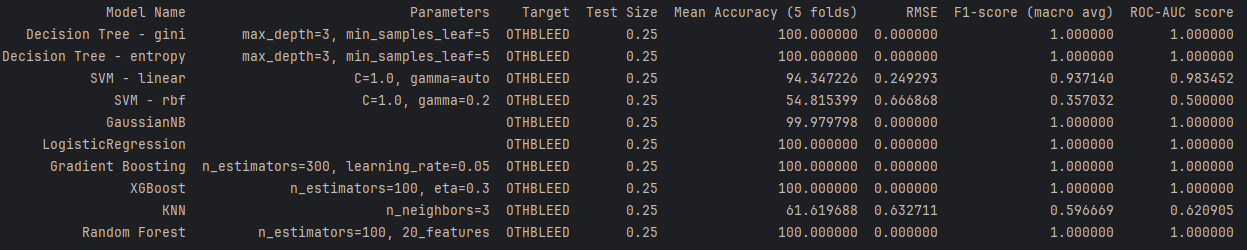
**Model Notes**

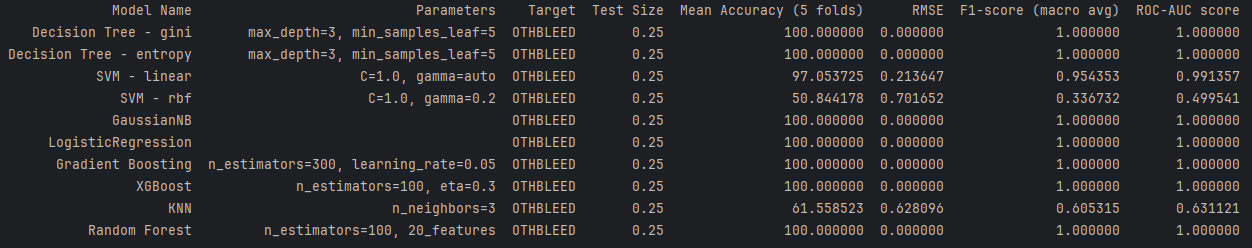
# Iteration 1: benchmark

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



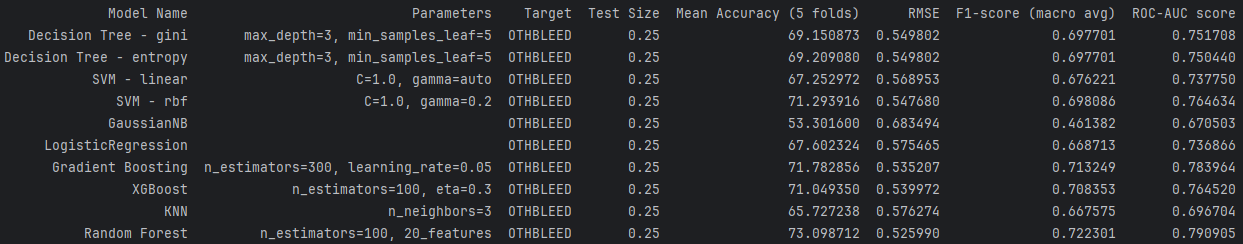
# Iteration 2: more recent data – extended to 2022

* 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

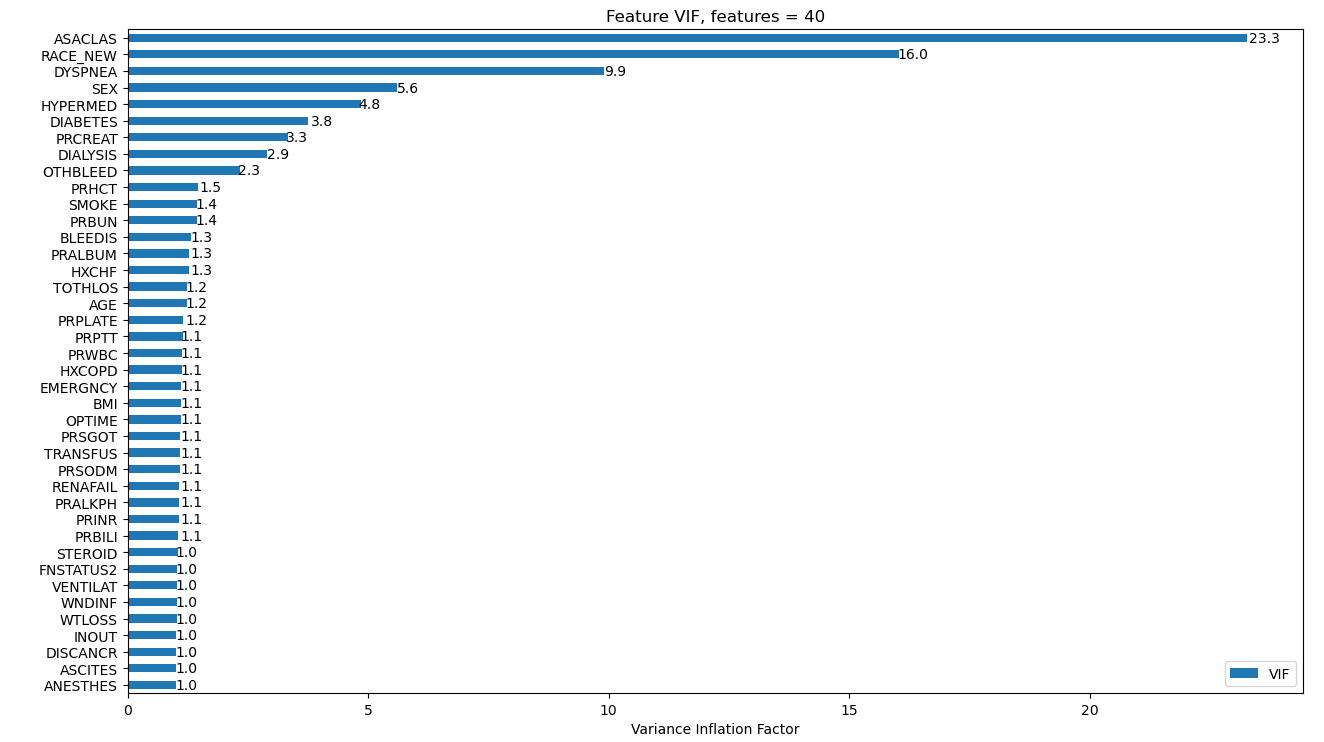
* Year 2018-2022
* features: 40, observations: 8587 (shape: 8587x41)
  + Dr. Gupta selected 43 features
  + 'HEIGHT','WEIGHT','ETHNICITY\_HISPANIC' was dropped for obvious multicollinearity issue (see graph below)
* imputations, standardization
* dataset: [GABG\_5yr\_preselect40.csv](https://raw.githubusercontent.com/jennytsai32/Capstone/10937f26fe011d2503834027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



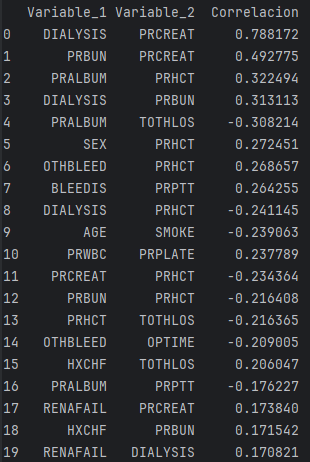
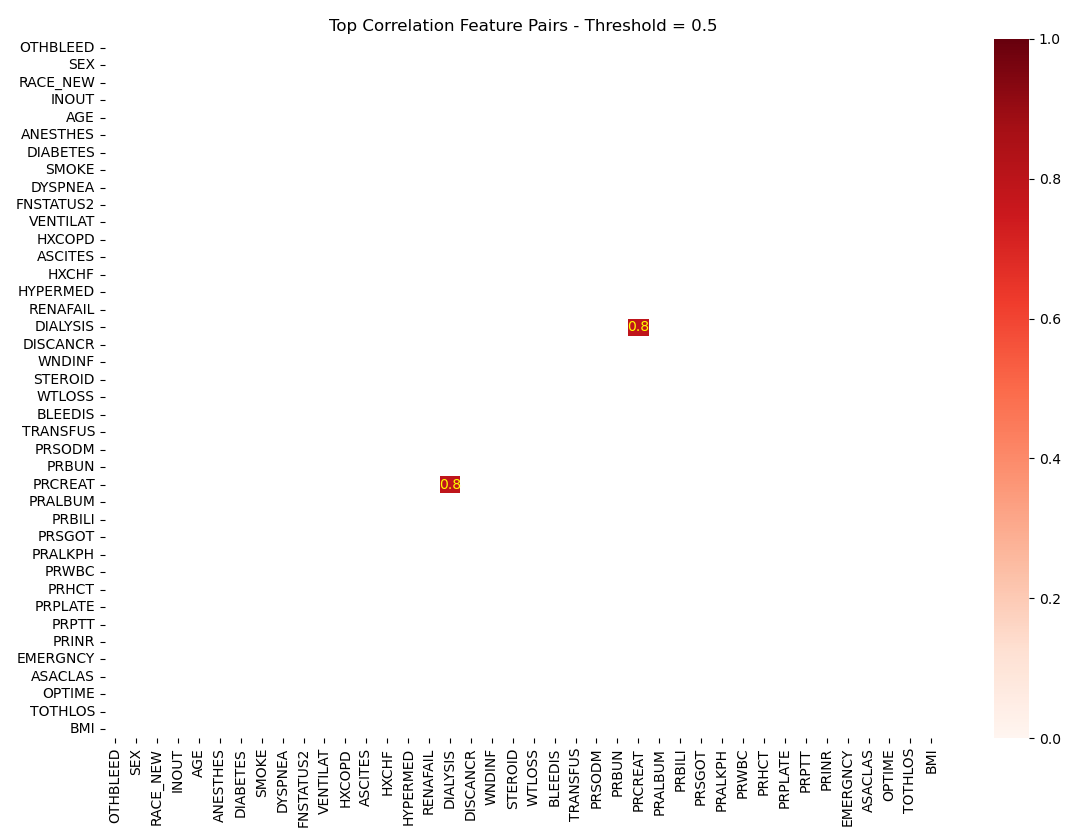
# Iteration 4: feature selection

**Step 1: VIF**

a. Features VIF Ranking



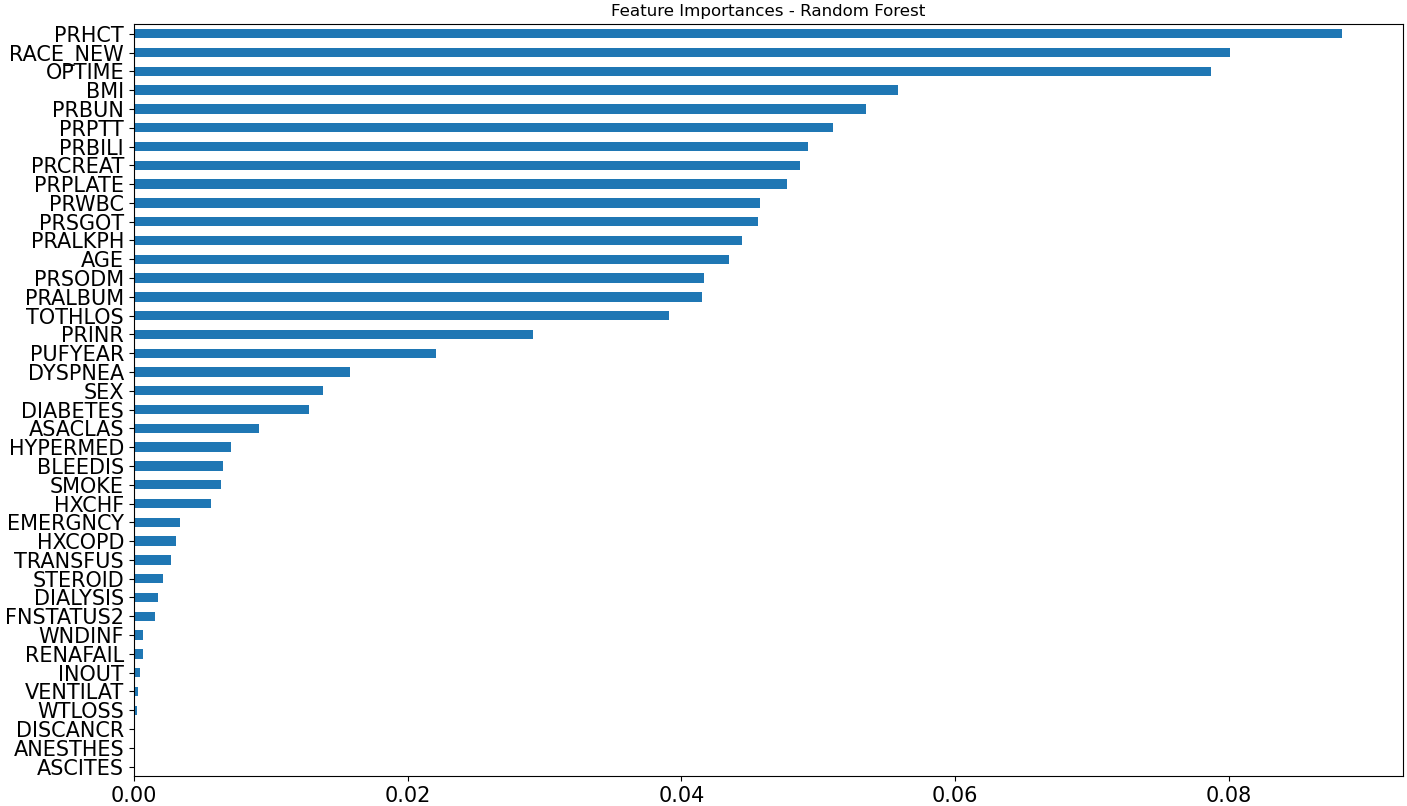
b. Features Correlation Pairs Ranking

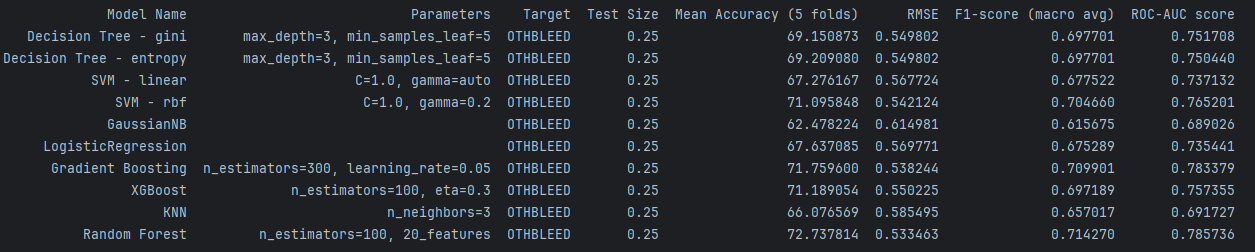
Pre-operative serum creatinine

Currently on dialysis

**Step 2: Features Importances (from Random Forest)**

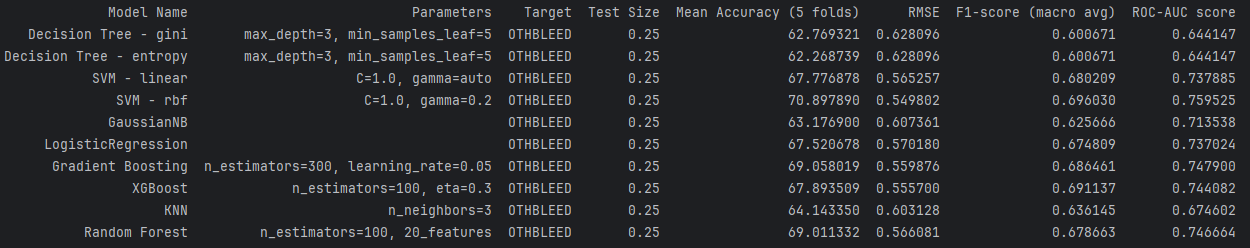


* Year 2018-2022
* features: 20, observations: 8587 (shape: 8587x21)
  + Dr. Gupta selected 43 features
  + Picked top 20 from feature importances (random forest)



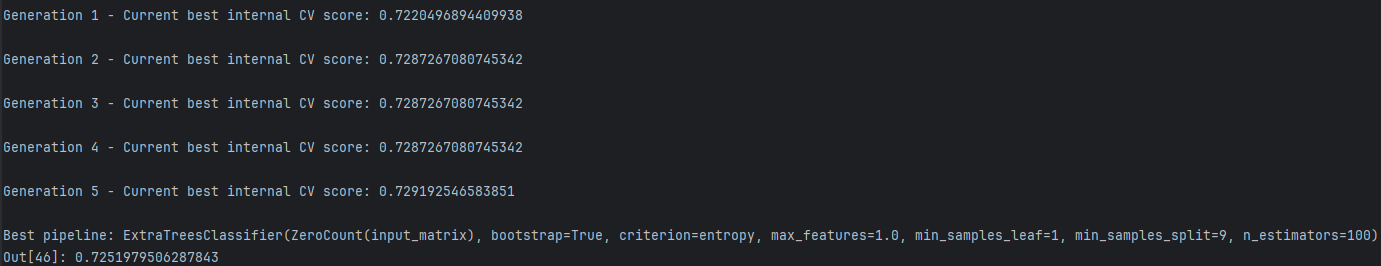
# Iteration 5: Feature Selection - PCA

* Run PCA on 20 features from iteration #4
* 19 features included (feature dimension reduced 1)



# Iteration 6: Feature Engineering - TPOT

* Run TPOT on 40 features dataset – the best model so far
* 0.25 on test size, generations=5, population\_size=20, verbosity=2



# Iteration 7: Feature Engineering - AutoFeat

=================

# Feature Selection

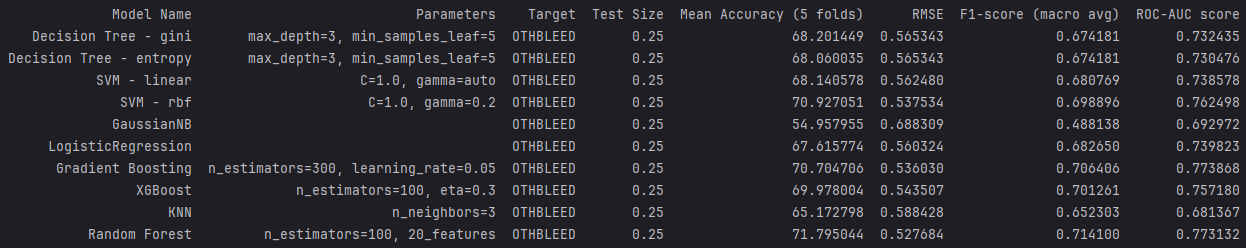
* sklearn.feature\_selection
* featuretools
* featurewiz

# Post-Training Analysis

* performance by class; check 0 and 1 balance
* Feature importance analysis - FI estimates the contribution of each feature (independent parameter) to the model output
* Tree-based models like Random **Forest, Gradient Boosting, Log** provide feature importances as a built-in feature.
* SHAP (SHapley Additive exPlanations)
  + shap library
* LIME (Local Interpretable Model-agnostic Explanations)
  + lime library
* Partial Dependence Plots (PDP)
  + Partial dependence plots show how the predicted outcome changes with variations in a particular feature while accounting for the average effects of all other features.
  + **pdpbox** and **sklearn.inspection** library in scikit-learn
* Feature Correlation Analysis
  + Corr matrix
  + Heatmap

BACK-UP Notes

43 features



Jenny’s 20 features after random forest

