Supplementary Materials

Evaluation of Bias Across Multiple Machine Learners in Epidemiologic Case-Control Studies of Pancreatic Cancer in the Prostate, Lung, Colorectal, and Ovarian Cancer (PLCO) Trial

Ryan Urbanowicz[1]†, Pranshu Suri[1], Yuhan Cui[1], Jason Moore[1]
[1]*Institute for Biomedical Informatics, University of Pennsylvania*
Philadelphia, PA, USA
†*Email: ryanurb@pennmedicine.upenn.edu*
Email: psuri@seas.upenn.edu; YuhanCui@seas.upenn.edu ; jhmoore@upenn.edu

Karen Ruth[1], Shannon M. Lynch‡[2]
[1]*Biostatistics and Bioinformatics Facility, Fox Chase Cancer Center*
[2]*Cancer Prevention and Control, Fox Chase Cancer Center*
Philadelphia, PA, USA
Email: Karen.Ruth@fccc.edu
‡*Email: Shannon.Lynch@fccc.edu*

Rachael Stolzenberg-Solomon[1]
[1]*Division of Cancer Epidemiology and Genetics, National Cancer Institute*
Shady Grove, MD, USA
Email: Rachael.solomon@nih.gov

S1. Cross Validation

We wrote our own python method to perform cross validation partitioning such that each CV partition balanced the relative numbers of cases and controls for case/control datasets 1 and 2. In order to ensure that the age/sex/race matching was preserved in the partitions of case/control dataset 3, we implemented a secondary CV partitioner that kept matched sets (each including 2 controls and 1 case) together within a given partition.

S2. Machine Learning Hyperparameter settings.

Below we review the specific hyperparameters explored in the respective grid searches for each ML algorithm. Any hyperparameters not mentioned were left at their default settings[1–3]. For logistic regression, these included 'penalty', 'max_iter', and 'C', with values (l1, l2), (100, 1000), and (0.001, 0.01, 0.1, 1, 10, 100, 1000), respectively[1]. For decision tree these included 'max_depth', 'min_samples_split', 'min_samples_leaf', and 'criterion', with values (3, 6, None), (2, 5, 10), (1, 5, 10), and (gini, entropy), respectively[1]. For random forest, these included the same parameters as decision tree but added 'n_estimators' and 'max_features', with values (10, 100, 1000) and (auto, log2), respectively[1]. For XGBoost, these included 'learning_rate', 'gamma', subsample', colsample_bytree', 'max_depth', and 'min_child_weight', with values (0.01, 0.3, 0.5), (0, 1, 5), (0.6, 0.8, 1.0), (0.6, 0.8, 1.0), (3, 6), and (1, 5, 10), respectively[3]. For support vector machine (with a radial based kernel function), these included 'C' and 'degree', with values (0.1, 1, 10) and (2, 3, 4), respectively[1]. For artificial neural network, these include 'learning_rate', 'momentum', 'activation', 'solver', and 'hidden_layer_sizes', with values (constant,invscaling, adaptive), (0.1, 0.5, 0.9), (logistic, tanh, relu), (adam, sgd), and ((50,50,50), or (100,100,100)), respectively[1]. Guassian Naïve Bayes had no hyperparameters available in the scikilearn implementation to optimize[1]. Lastly the LCS algorithm, ExSTraCS was applied with default hyperparameters utilized in previously published studies (i.e. rule population size = 1000, training iterations = 200,000)[2,4,5], since it had been previously designed to perform well without significant hyperparameter tuning, and to reduce the overall computational time required to run this somewhat more computationally expensive methodology.

S3. Feature importance score estimation

Below we specify the strategy utilized for feature importance estimation for each ML algorithm. For the Logistic Regression model, each feature's importance score was obtained by taking the exponential of the regression coefficient 'coef_'[1]. For the Decision Tree, Random Forest, and XGBoost models, feature importance scores were obtained via 'feature_importances_', a scikit-learn attribute for each of the 3 models[1]. Next, for Artificial Neural Networks, Support-Vector Machine, and the Bayesian classifier, feature importance scores were obtained by running the models on the dataset a set number of times, corresponding to the number of features present in the dataset. For each time the model is run, one feature is left out, and a balanced accuracy for that model is obtained. Comparing this balanced accuracy to the original balanced accuracy of the model with all of the features included, feature importance scores were obtained, where the lower the balanced accuracy of the newly-run model (one feature omitted) compared to the original model (all features included), the more "important" the feature is. Lastly, feature importance scores for LCS were obtained based on feature specificity count used in previous ExSTracS studies to quantify feature importance[2,6]. Furthermore, for each dataset, MultiSURF scores offer an additional, model independent estimate of feature importance that we report independent of the respective model feature importance sores[7].

S4. Feature importance normalization, weighting, and visualization

To allow for comparisons, feature importance scores were scaled and visualized using compound bar plots. Specifically, since all ML methods have different feature importance value scales, and a variety of feature importance strategies were applied (custom to each method), respective scores were first normalized (0-1) and then the magnitude of feature importance in each plot was scaled by the relative performance (i.e. balanced accuracy) of each ML algorithm. If an algorithm yielded a balanced accuracy of 0.5 or less (i.e. no predictive signal found), then this scaling would eliminate its contribution to the feature importance visualization plot. Algorithms that performed better had a higher overall maximum feature importance contribution on a 'per-feature' basis.

Since all ML methods have different feature importance value scales, and a variety of feature importance strategies were applied (custom to each method), respective scores were first normalized (0-1) and then the magnitude of feature importance in each plot was scaled by the relative performance (i.e. balanced accuracy) of each ML algorithm. If an algorithm yielded a balanced accuracy of 0.5 or less (i.e. no predictive signal found), then this scaling would eliminate its contribution to the feature importance plot. Algorithms that performed better have a higher overall maximum feature importance contribution on a 'per-feature' basis.

S5. Supplementary Results

Notably, LCS yields the best recall for each dataset, and the best F1-score for datasets 1 and 2, as well as the second best F1-score for dataset 3 (in line with the pattern observed for balanced accuracy results.

S6. Supplementary Discussion

In this study we assessed the predictive ability and compared feature importance scores across 8 different ML modeling algorithms and 3 different cohort-based case/control datasets. In addition to establishing a rigorous machine learning pipeline for investigating cohort data, we sought to compare machine learning performance, as well as examine the impact of bias due to (1) missing data (case-control Set #2), and (2) confounding introduced by control selection (case-control set #1, 2).

Some limitations of our study related to missing data and potential for imbalanced data. Datasets with a large percentage of missing values can be problematic for machine learners. While our missing data was minimized in the first (ALL cases and control) and third dataset (matched set), it was greater than 30% for some features and this missingness was differential for cases (up to 30%) and controls (up to 15%).

LCS performed best in this study despite the fact that its default hyperparameters were used in contrast to all other 'out-of-the-box' methods examined in this study (except Naïve Bayes which has no hyperparameters) which were given the opportunity to optimize their hyperparameters via a grid search. While the argument could certainly be made that a much more extensive hyperparameter sweep could be conducted for these other algorithms, doing so can become extremely computationally expensive particularly when the dataset sizes scale up. The two algorithms that performed, particularly poorly in this study (i.e. SVM and neural networks), are not completely unexpected. First, since logistic regression was utilized in this study, we only implemented a radial based function kernel for SVM, as opposed to also exploring a linear kernel. Second, neural networks, while potentially quite powerful, can be difficult to optimize in terms of both their hyperparameters and overall model architecture, which also needs to be specified. Neural network models are also notoriously difficult to interpret, an algorithm like LCS offers a strategy that is not only powerful with little to no hyperparameter optimization, and able to detect complex patters, but also offers a fundamentally interpretable model comprised of a set of human readable IF:THEN rule expression[5].

S7. Future Work

Future work may focus on conducting sensitivity analyses across different sample definitions to assess reproducibility and as well as conduct leave-one-out feature analyses to further improve the interpretability of machine learning models. Further given the success of the LCS algorithm we will utilize it along with other established ML algorithms to scale up this study of pancreatic cancer to incorporate a multitude of candidate genetic risk factors.

Given it's predictive success, we expect to leverage the ability of LCS to automatically detect heterogeneous patient subgroups in future work and further characterize any complex patterns of association through manual inspection of the IF:THEN rules that comprise an LCS model as well as through global rule patterns[2,8,9].

S8. Supplementary Tables

Table S1: Average accuracy (with standard deviation) for each dataset and ML modeler.

	Case/Control Set 1	Case/Control Set 2	Case/Control Set 3
Logistic Regression	0.8464 (0.0054)	0.8479 (0.0068)	0.6666 (0.0262)
Decision Tree	0.7930 (0.0300)	0.7989 (0.0255)	0.6487 (0.0434)
Random Forest	0.8428 (0.0109)	0.8420 (0.0074)	0.6640 (0.0374)
XGBoost	0.8340 (0.0076)	0.8262 (0.0106)	0.6341 (0.0531)
SVM	0.8216 (0.0180)	0.8430 (0.0001)	0.6651 (0.0141)
Naïve Bayes	0.8179 (0.0091)	0.8109 (0.0095)	0.6490 (0.0405)
Neural Networks	0.7754 (0.1831)	0.8418 (0.0041)	0.6651 (0.0322)
LCS	0.7820 (0.0161)	0.7826 (0.0245)	0.6138 (0.0506)

Table S2: Average specificity (with standard deviation) for each dataset and ML modeler.

	Case/Control Set 1	Case/Control Set 2	Case/Control Set 3
Logistic Regression	0.9860 (0.0053)	0.9813 (0.0051)	0.9314 (0.0306)
Decision Tree	0.8906 (0.0376)	0.8983 (0.0368)	0.8751 (0.0655)
Random Forest	0.9746 (0.0083)	0.9709 (0.0087)	0.91151 (0.0358)
XGBoost	0.9560 (0.0118)	0.9439 (0.0166)	0.8451 (0.1026)
SVM	0.9683 (0.0261)	1.0 (0.0)	1.0 (0.0)
Naïve Bayes	0.9174 (0.0086)	0.90182 (0.0127)	0.7924 (0.0413)
Neural Networks	0.8897 (0.2702)	0.9951 (0.0078)	0.9326 (0.0535)

LCS	0.8373 (0.0199)	0.8254 (0.0383)	0.7195 (0.0889)

Table S3: Average recall (with standard deviation) for each dataset and ML modeler.

	Case/Control Set 1	**Case/Control Set 2**	**Case/Control Set 3**
Logistic Regression	0.0962 (0.0316)	0.1312 (0.0371)	0.1401 (0.0725)
Decision Tree	0.2687 (0.0545)	0.265 (0.0572)	0.1965 (0.0812)
Random Forest	0.1349 (0.0310)	0.1500 (0.0311)	0.1724 (0.0562)
XGBoost	0.1787 (0.0414)	0.1937(0.0407)	0.2164 (0.0855)
SVM	0.0337 (0.0379)	0.0 (0.0)	0.0 (0.0)
Naïve Bayes	0.2837 (0.0500)	0.3225 (0.0585)	0.3623 (0.0719)
Neural Networks	0.1612 (0.2968)	0.0187 (0.0275)	0.1336 (0.0503)
LCS	0.485 (0.0721)	0.5525 (0.0696)	0.3996 (0.1192)

Table S4: Average precision (with standard deviation) for each dataset and ML modeler.

	Case/Control Set 1	**Case/Control Set 2**	**Case/Control Set 3**
Logistic Regression	0.5634 (0.1221)	0.5669 (0.0948)	0.5053 (0.1547)
Decision Tree	0.3285 (0.0808)	0.3378 (0.0521)	0.4550 (0.1270)
Random Forest	0.5039 (0.1267)	0.4975 (0.0839)	0.5001 (0.1440)
XGBoost	0.4337 (0.0511)	0.4030 (0.0746)	0.4388 (0.1180)
SVM	0.0936 (0.0998)	0.0 (0.0)	0.0 (0.0)
Naïve Bayes	0.3882 (0.0453)	0.3783 (0.0366)	0.4682 (0.0718)
Neural Networks	0.1464 (0.1992)	0.2208 (0.2875)	0.5574 (0.1829)
LCS	0.3568 (0.0354)	0.3745 (0.0289)	0.4194 (0.0831)

Table S5: Average F1-score (with standard deviation) for each dataset and ML modeler.

	Case/Control Set 1	**Case/Control Set 2**	**Case/Control Set 3**
Logistic Regression	0.1627 (0.0492)	0.2116 (0.0520)	0.2136 (0.0928)
Decision Tree	0.2900 (0.0497)	0.2899 (0.0463)	0.2666 (0.0943)
Random Forest	0.2124 (0.0491)	0.2286 (0.0394)	0.2549 (0.0793)
XGBoost	0.2504 (0.0441)	0.2572 (0.0360)	0.2754 (0.0827)
SVM	0.0494 (0.0548)	0.0 (0.0)	0.0 (0.0)
Naïve Bayes	0.3270 (0.0479)	0.3466 (0.0458)	0.4072 (0.0708)
Neural Networks	0.1033 (0.1379)	0.0331 (0.0475)	0.2061 (0.0671)
LCS	0.4100 (0.0450)	0.4434 (0.0246)	0.4044 (0.0892)

Balanced Accuracy Statistical Comparisons

Kruskal Wallis Test – Comparing ML results over 10 fold CV

```
EpiOnly_20180710_Clean
KruskalResult(statistic=57.34439352276202, pvalue=5.105699883503447e-10)

Epi_DietAdj_20180710_Clean
KruskalResult(statistic=66.96047135196164, pvalue=6.063992352649637e-12)

Epi_DietAdj_Matched_20180710_Clean
KruskalResult(statistic=19.399236766902266, pvalue=0.007024523838197663)
```

Mann Whitney U Test – Comparing ML pairwise results over 10 fold CV

EpiOnly_20180710_Clean

Algorithm 1	Algorithm 2	statistic	p-value
logistic_regression	decision_tree	8	0.000853
logistic_regression	random_forest	26	0.037831
logistic_regression	xgboost	11	0.001805
logistic_regression	naive_bayes	2	0.000165
logistic_regression	svm	5	0.000367
logistic_regression	ann	20	0.011349
logistic_regression	lcs	0	9.13E-05
decision_tree	random_forest	21	0.015605
decision_tree	xgboost	34	0.120661
decision_tree	naive_bayes	28.5	0.056138
decision_tree	svm	1	0.000117
decision_tree	ann	18	0.007488
decision_tree	lcs	2	0.000165
random_forest	xgboost	30.5	0.075387
random_forest	naive_bayes	9	0.001101
random_forest	svm	1	0.000117
random_forest	ann	20	0.011349
random_forest	lcs	0	9.13E-05
xgboost	naive_bayes	15	0.004554
xgboost	svm	0	8.63E-05
xgboost	ann	20	0.011349
xgboost	lcs	0	9.13E-05
naive_bayes	svm	0	8.63E-05
naive_bayes	ann	12	0.001888
naive_bayes	lcs	8	0.000853
svm	ann	50	0.483499
svm	lcs	0	8.63E-05
ann	lcs	2	0.000122

Epi_DietAdj_20180710_Clean

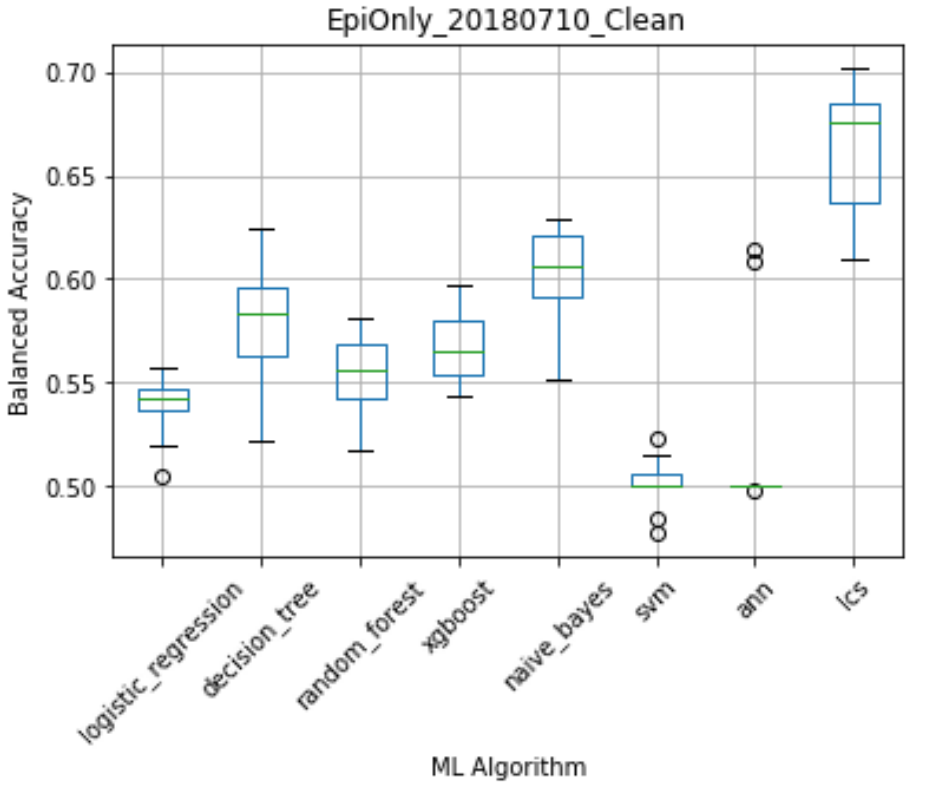
Algorithm 1	Algorithm 2	statistic	p-value
logistic_regression	decision_tree	18	0.008608
logistic_regression	random_forest	41	0.260261
logistic_regression	xgboost	32	0.092938
logistic_regression	naive_bayes	3	0.00022
logistic_regression	svm	0	3.19E-05
logistic_regression	ann	4	0.000289
logistic_regression	lcs	0	9.13E-05
decision_tree	random_forest	22	0.018782
decision_tree	xgboost	26	0.037776
decision_tree	naive_bayes	18	0.008608
decision_tree	svm	0	3.17E-05
decision_tree	ann	1	0.000121
decision_tree	lcs	0	9.08E-05
random_forest	xgboost	36	0.153745
random_forest	naive_bayes	4	0.000291
random_forest	svm	0	3.19E-05
random_forest	ann	3	0.000218
random_forest	lcs	0	9.13E-05
xgboost	naive_bayes	7	0.000657
xgboost	svm	0	3.19E-05
xgboost	ann	1	0.000122
xgboost	lcs	0	9.13E-05
naive_bayes	svm	0	3.19E-05
naive_bayes	ann	0	9.03E-05
naive_bayes	lcs	2	0.000165
svm	ann	30	0.048
svm	lcs	0	3.19E-05
ann	lcs	0	9.03E-05

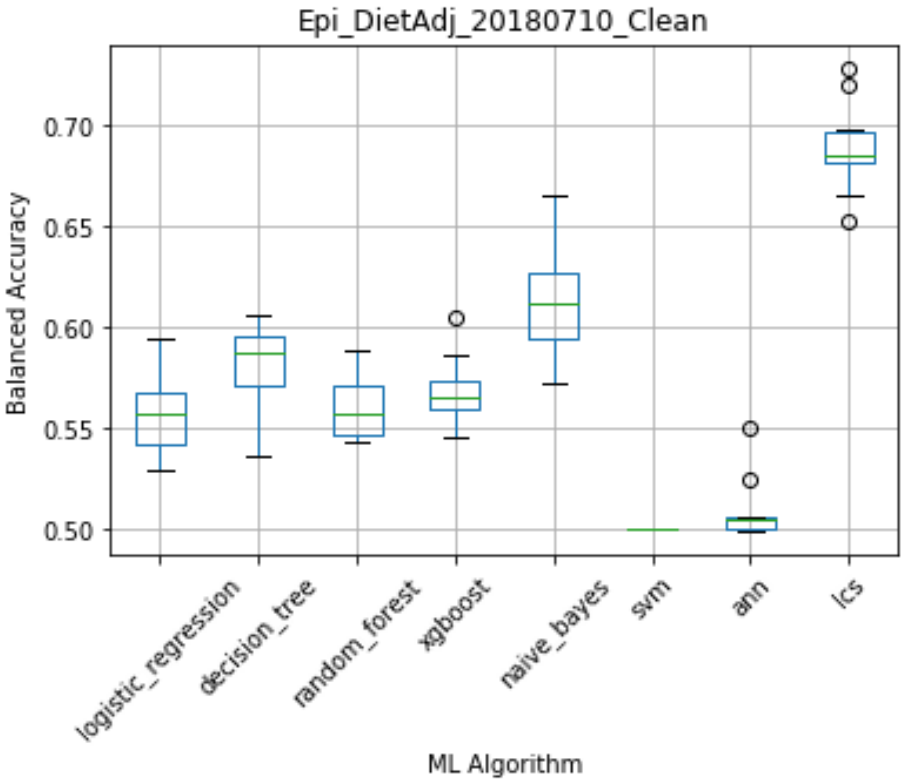
Epi_DietAdj_Matched_20180710_Clean

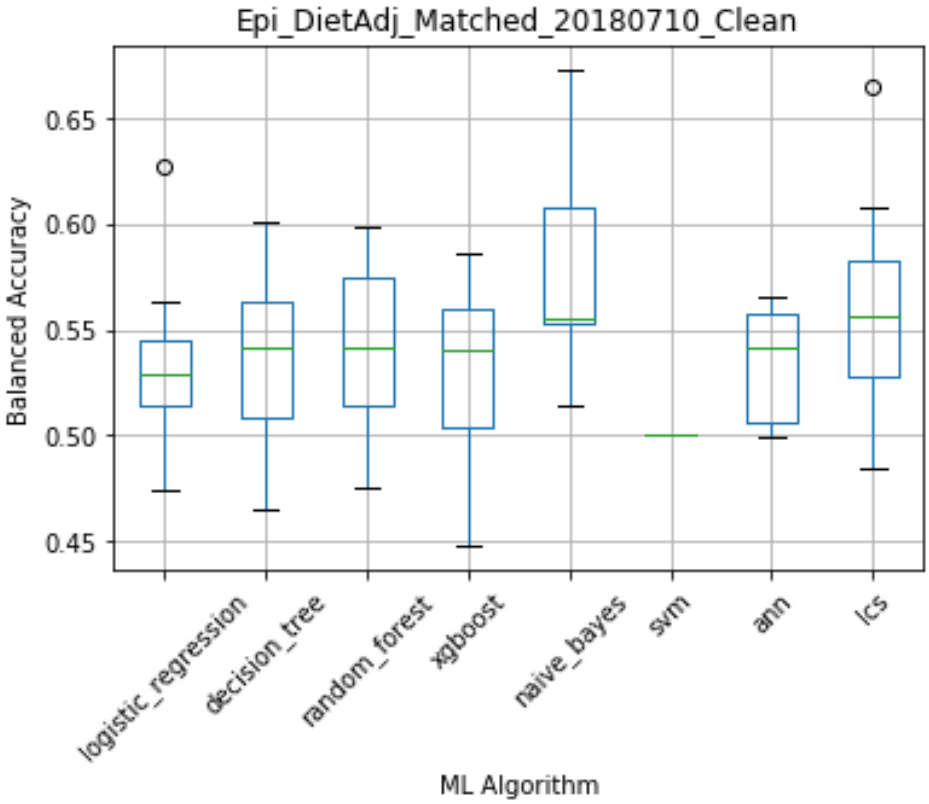
Algorithm 1	Algorithm 2	statistic	p-value
logistic_regression	decision_tree	45	0.366865
logistic_regression	random_forest	43	0.311588
logistic_regression	xgboost	47	0.425053
logistic_regression	naive_bayes	22	0.018818
logistic_regression	svm	10	0.00071
logistic_regression	ann	46	0.39563
logistic_regression	lcs	34	0.120661
decision_tree	random_forest	45.5	0.381141
decision_tree	xgboost	48	0.454861
decision_tree	naive_bayes	24	0.026951
decision_tree	svm	20	0.008594
decision_tree	ann	46	0.39563
decision_tree	lcs	36	0.153745

random_forest	xgboost	41	0.260261
random_forest	naive_bayes	27	0.044487
random_forest	svm	20	0.008594
random_forest	ann	43	0.311523
random_forest	lcs	42	0.285375
xgboost	naive_bayes	27	0.044487
xgboost	svm	30	0.057629
xgboost	ann	49	0.484919
xgboost	lcs	36	0.153745
naive_bayes	svm	0	3.19E-05
naive_bayes	ann	24	0.026906
naive_bayes	lcs	38	0.192337
svm	ann	20	0.005918
svm	lcs	10	0.00071
ann	lcs	34	0.120572

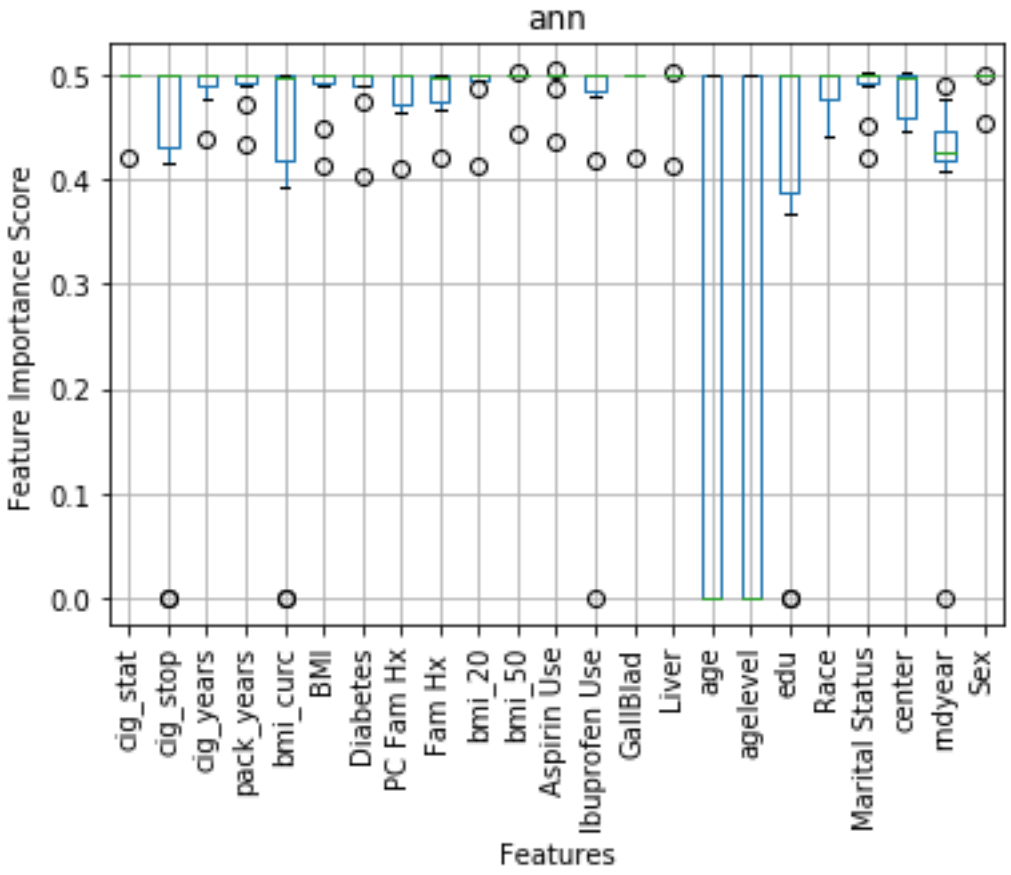
Boxplots comparing ML Balanced accuracy performance

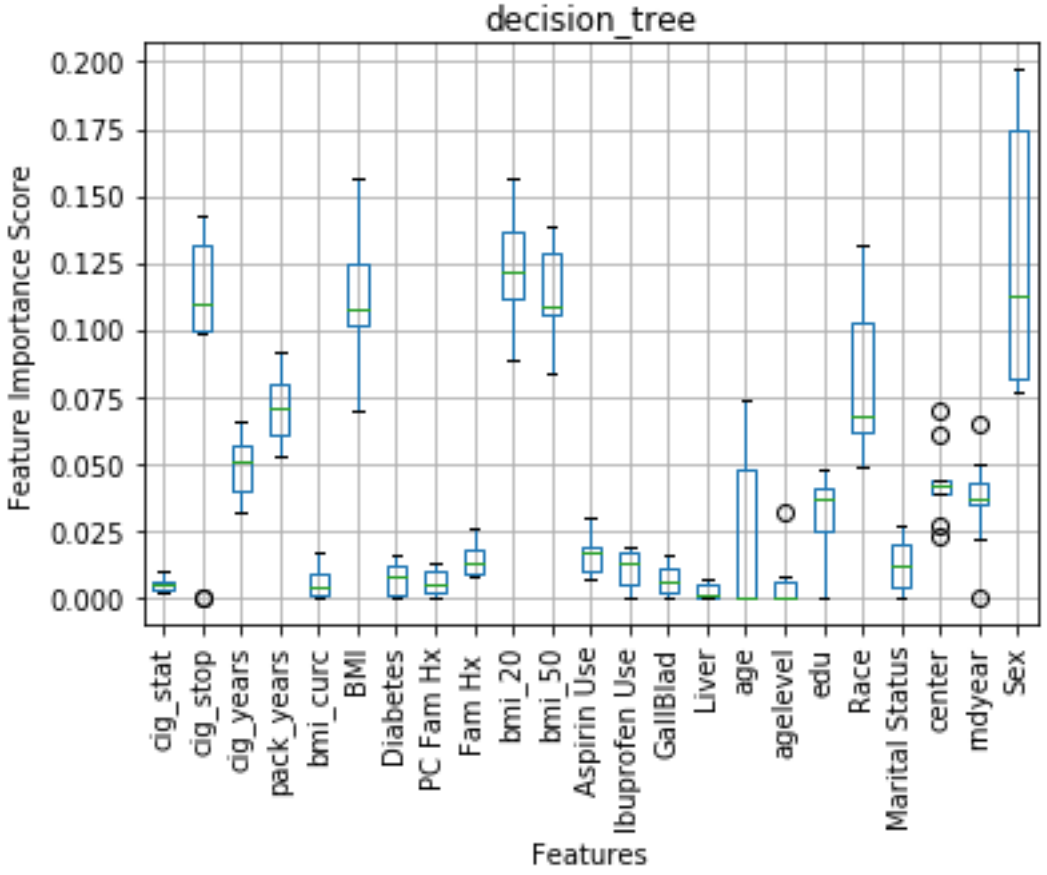


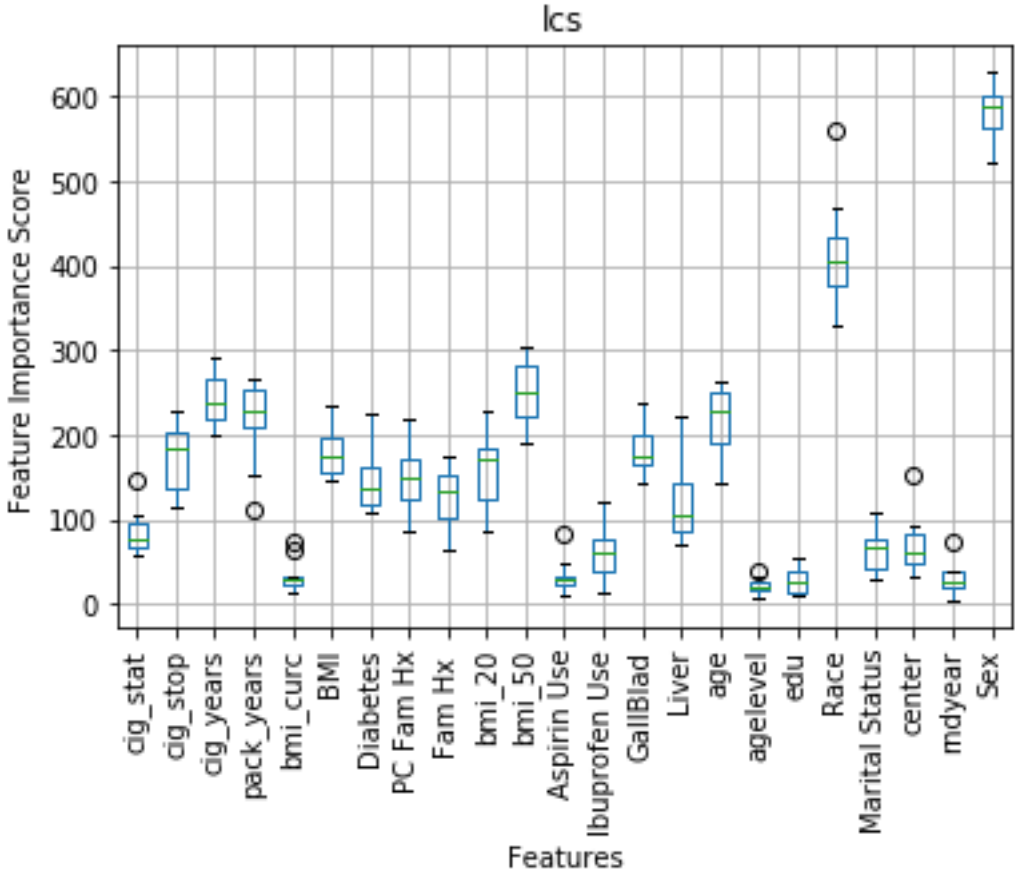


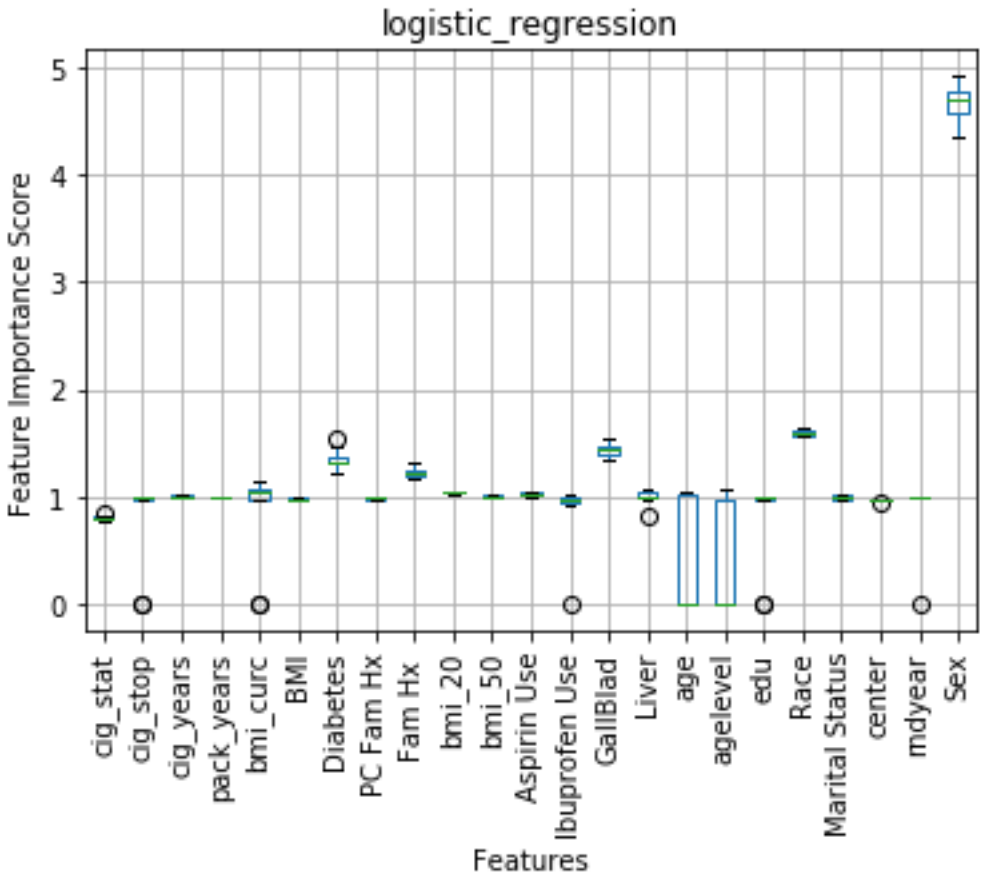


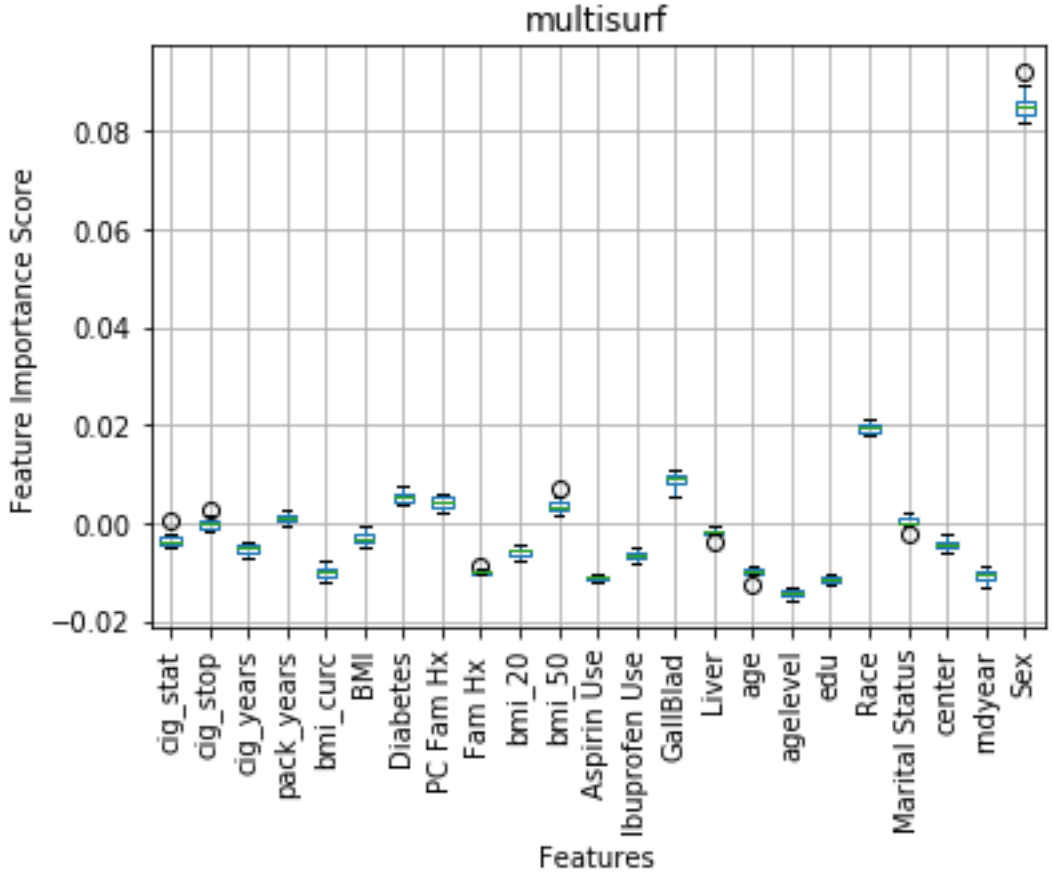
Dataset 1 Feature Importance Boxplots

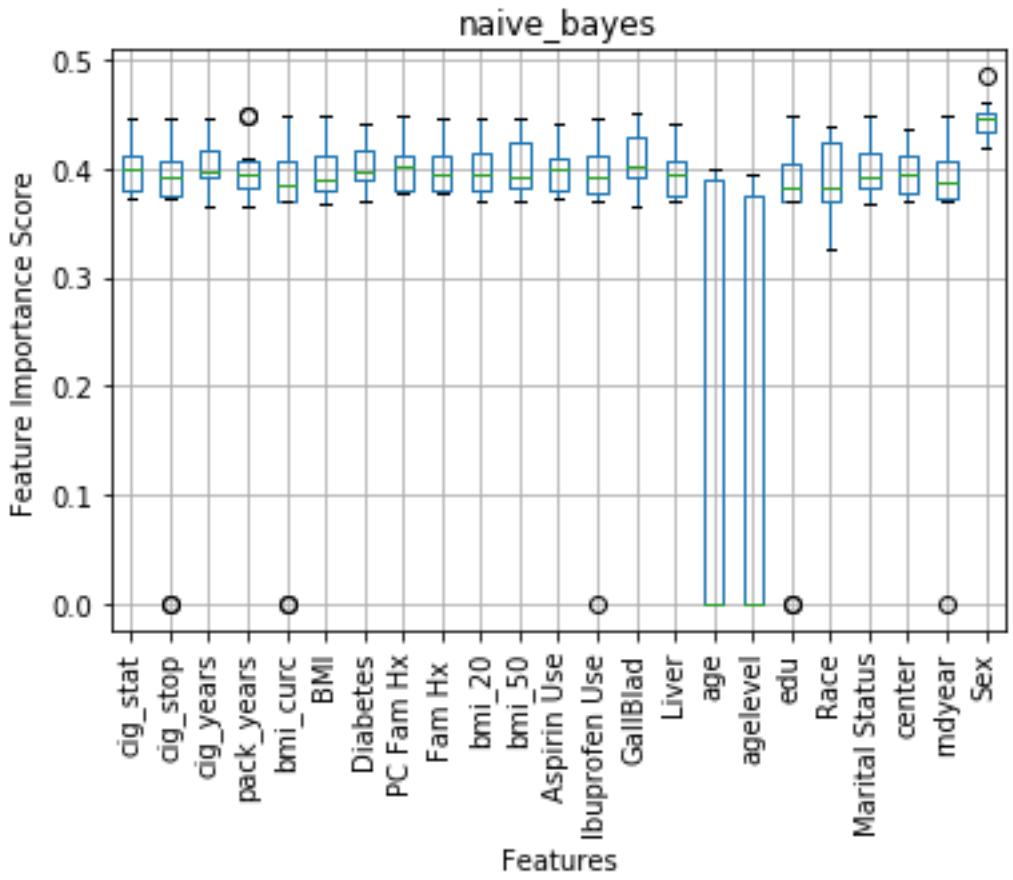

ann
Feature Importance Score
0.0
0.1
0.2
0.3
0.4
0.5
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

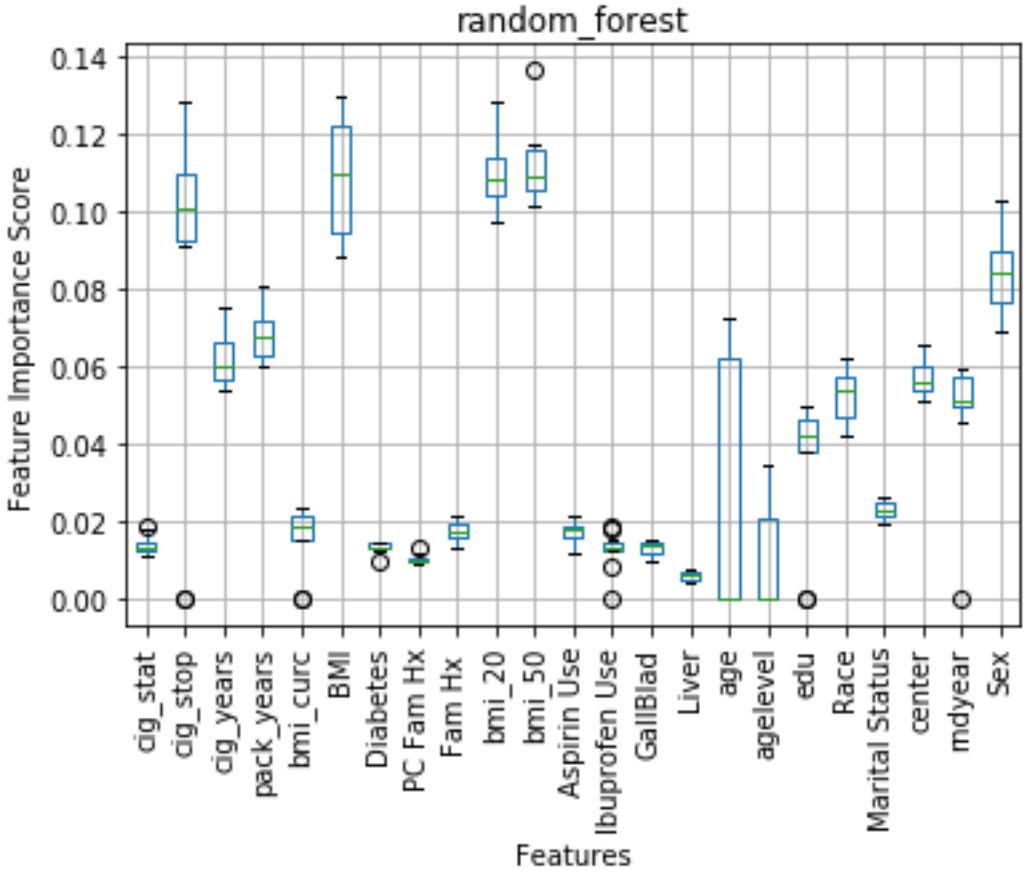

decision_tree
Feature Importance Score
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

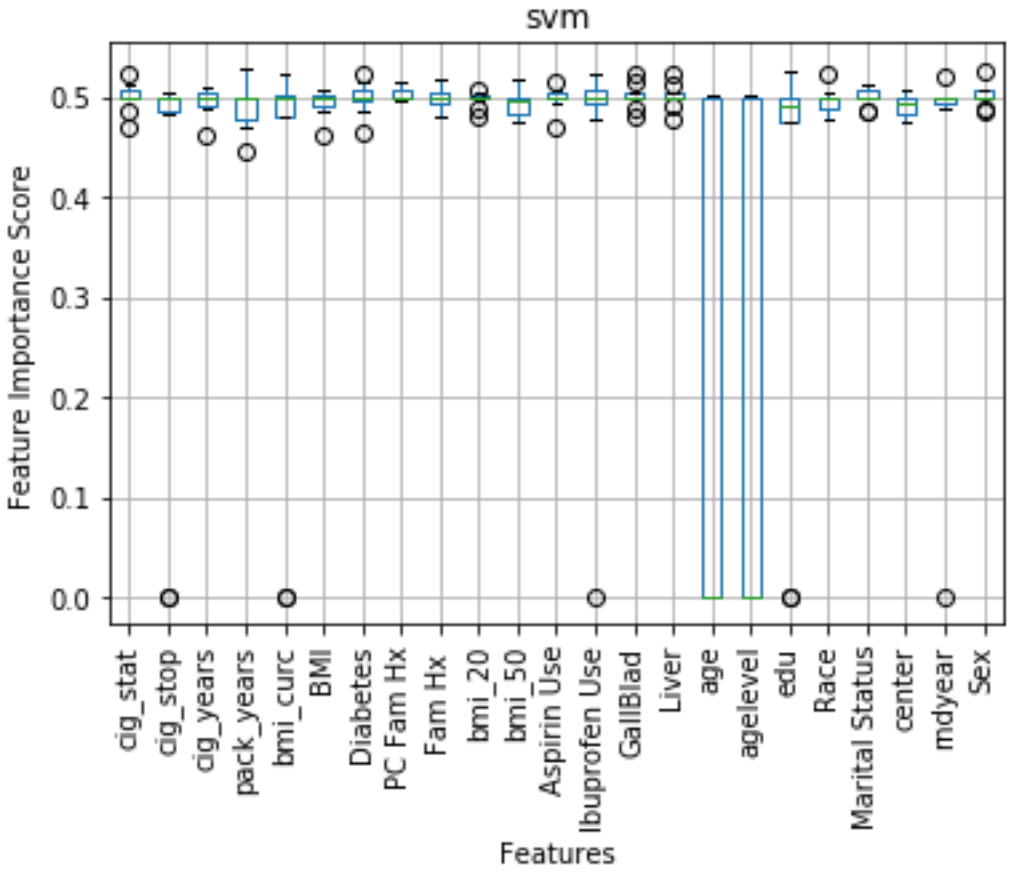

lcs
Feature Importance Score
0
100
200
300
400
500
600
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

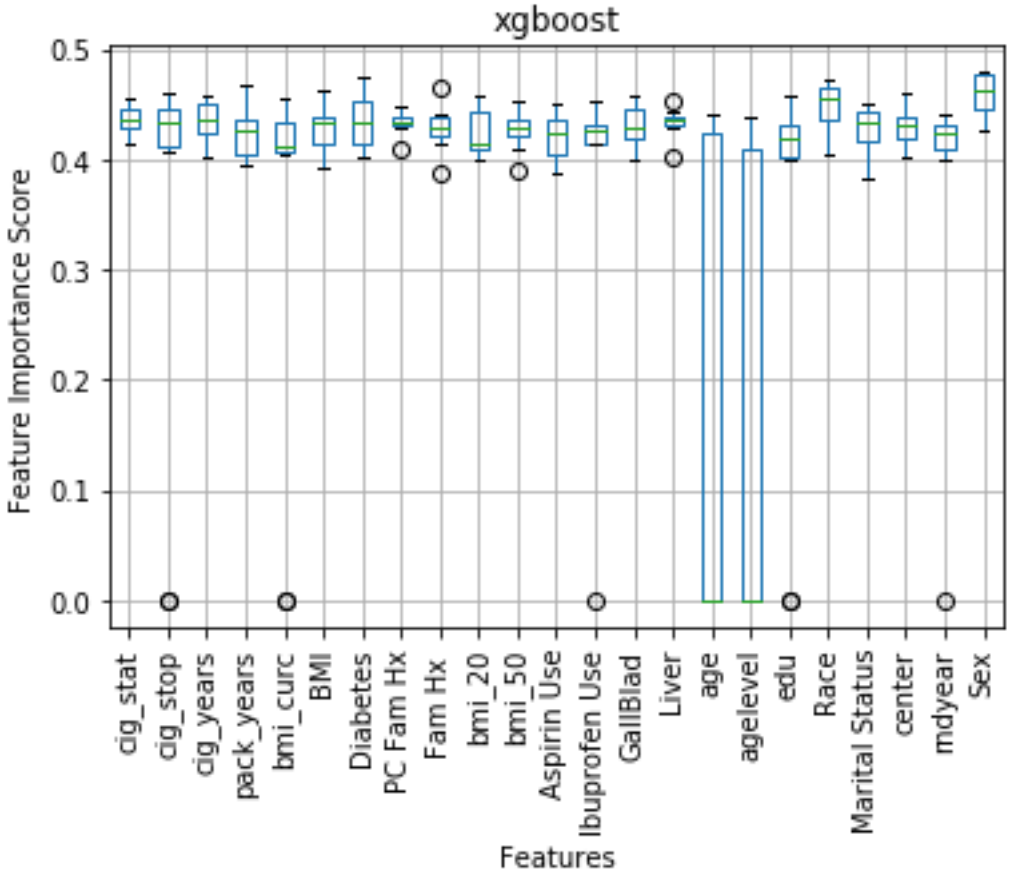

logistic_regression
Feature Importance Score
0
1
2
3
4
5
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features


multisurf
Feature Importance Score
0.08
0.06
0.04
0.02
0.00
−0.02
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

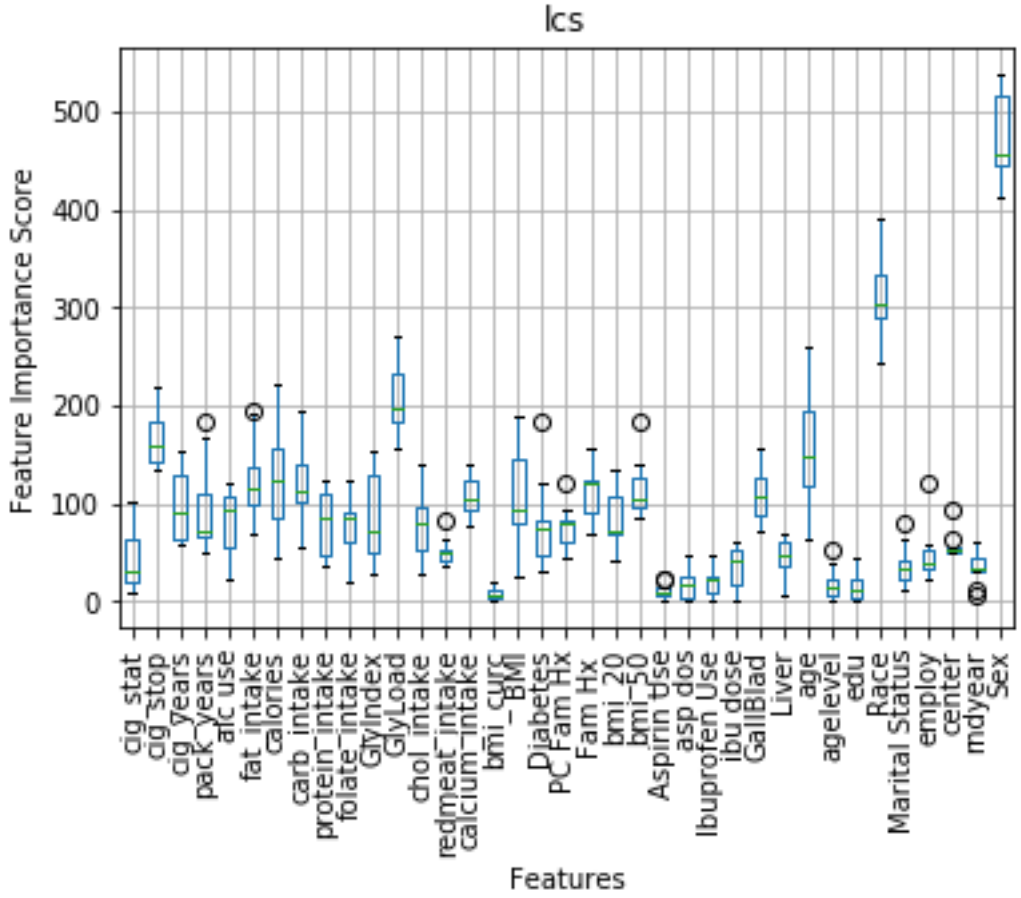

naive_bayes
Feature Importance Score
0.5
0.4
0.3
0.2
0.1
0.0
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

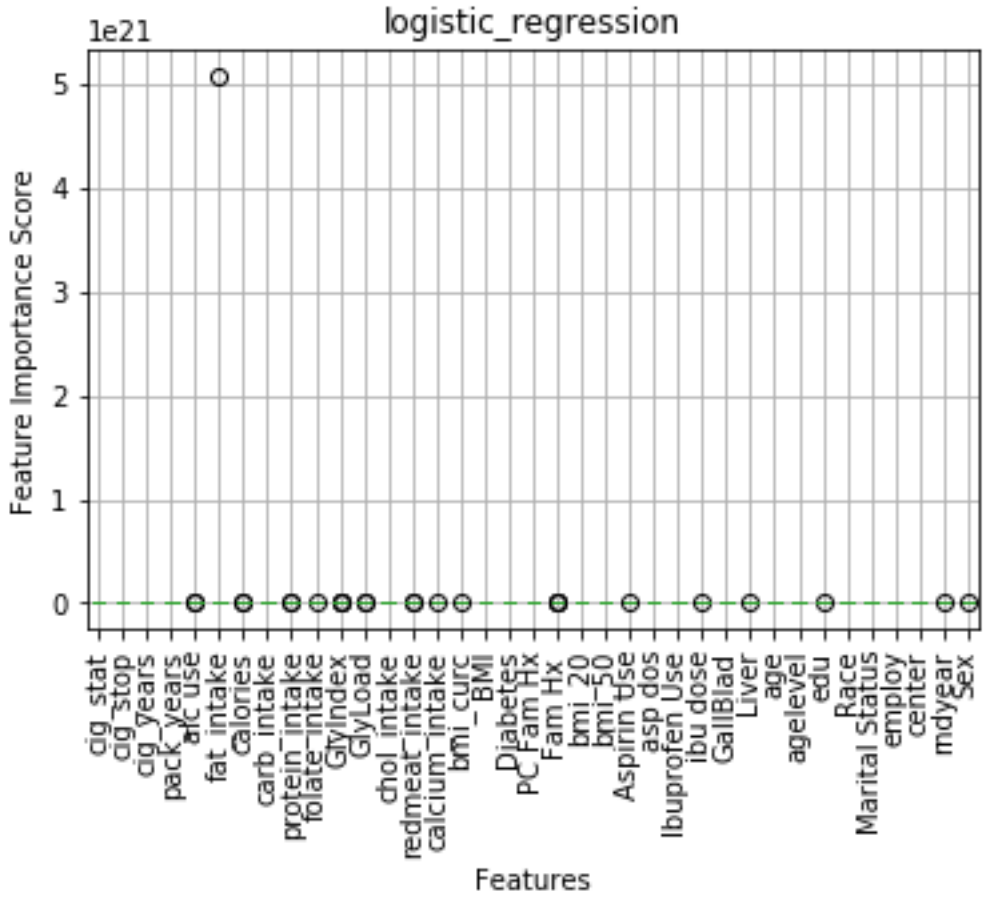

random_forest
Feature Importance Score
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

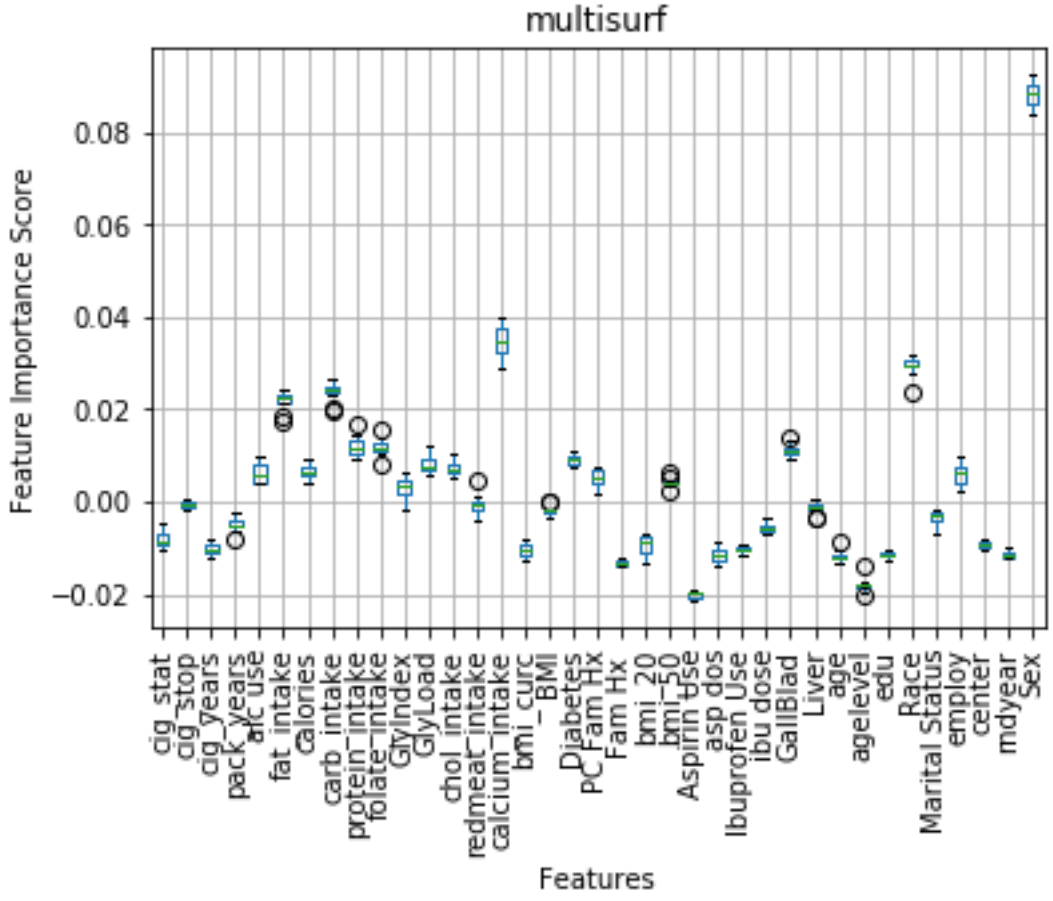

svm
Feature Importance Score
0.5
0.4
0.3
0.2
0.1
0.0
cig_stat
cig_stop
cig_years
pack_years
bmi_curc
BMI
Diabetes
PC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin Use
Ibuprofen Use
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
center
mdyear
Sex
Features

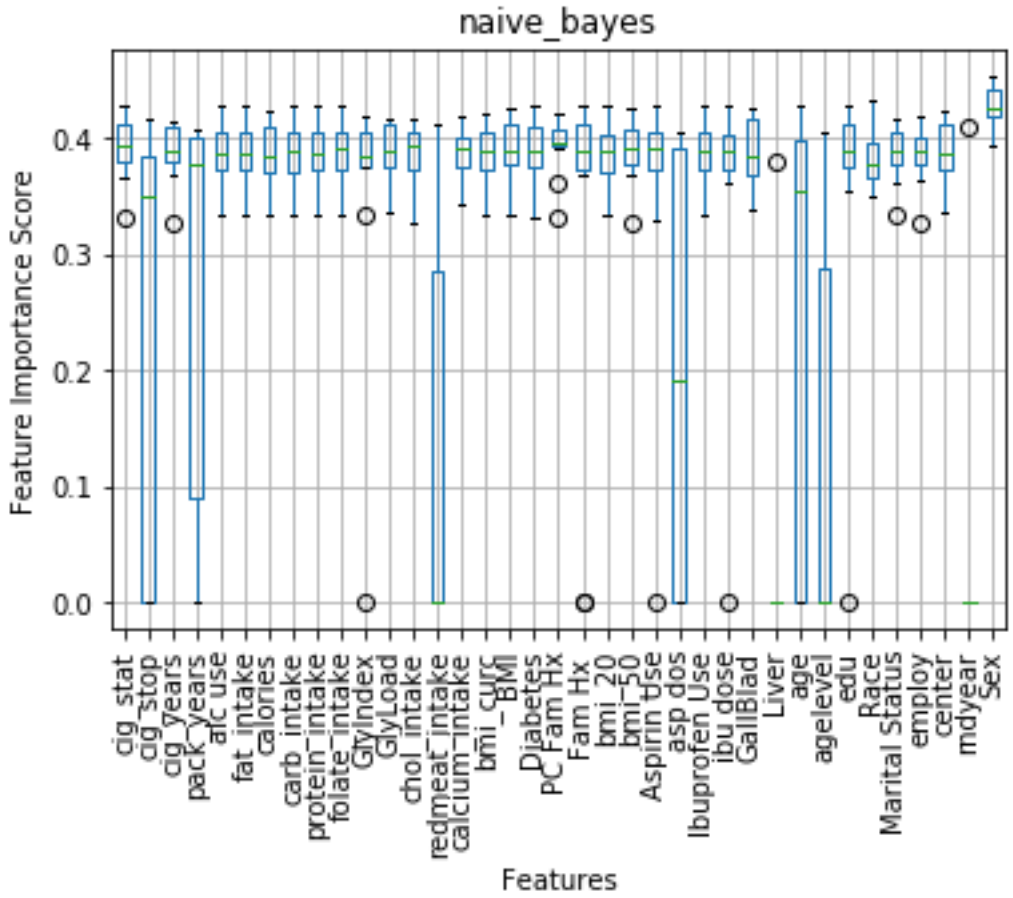


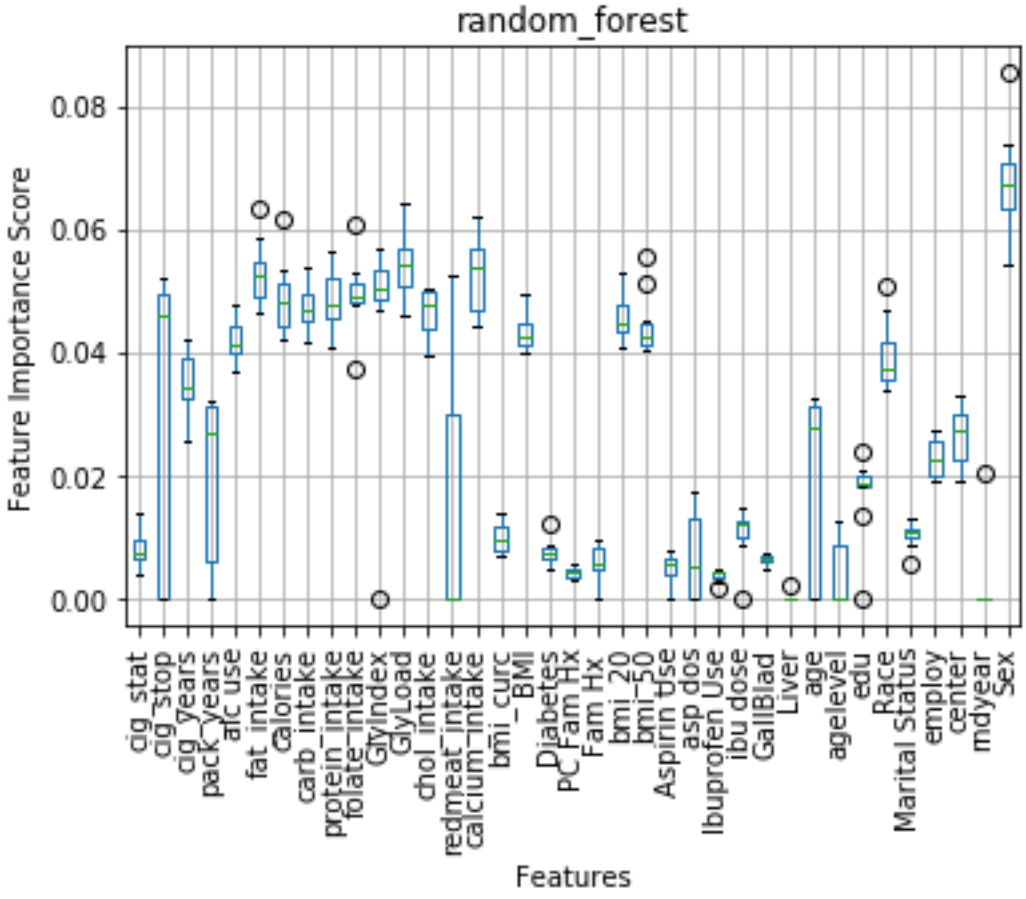
Dataset 2 Feature Importance Boxplots

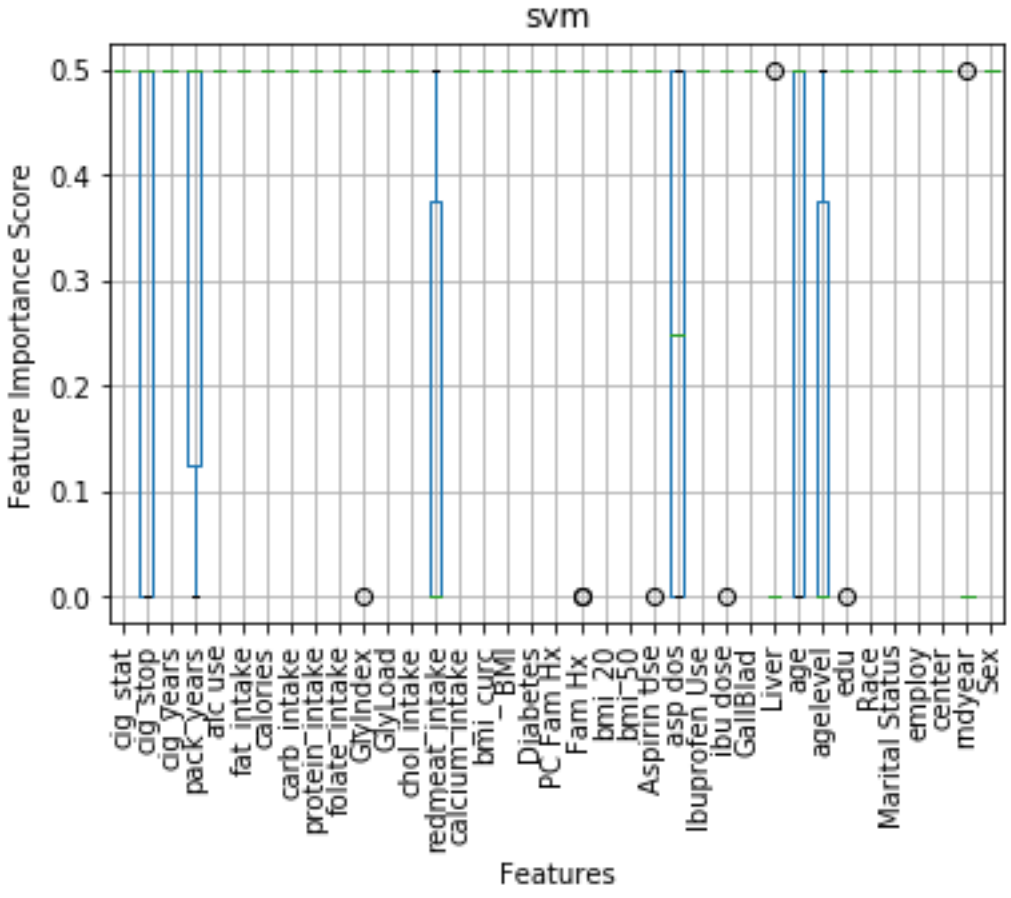

lcs
Feature Importance Score
0
100
200
300
400
500
cig_stat
cig_stop
cig_years
pack_years
afc use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
BMI
Diabetes
PC_Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

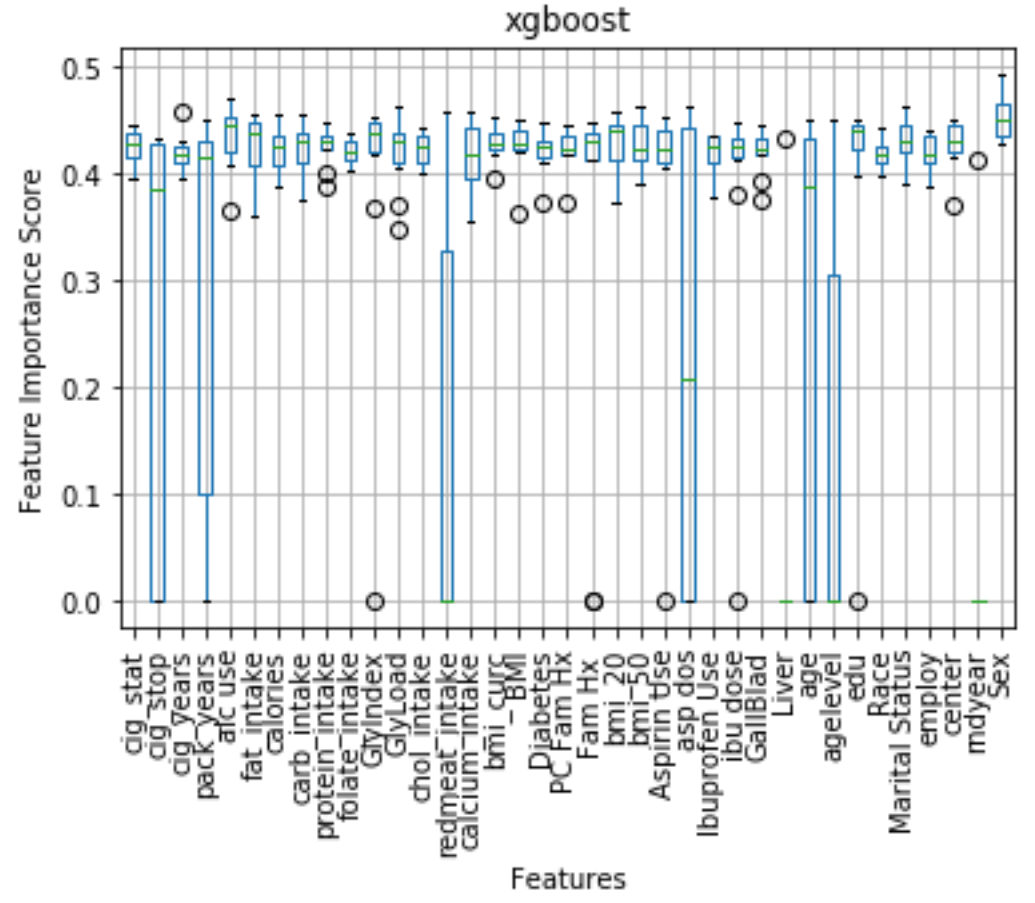

logistic_regression
1e21
Feature Importance Score
0
1
2
3
4
5
cig_stat
cig_stop
cig_years
pack_years
afc use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
BMI
Diabetes
PC_Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

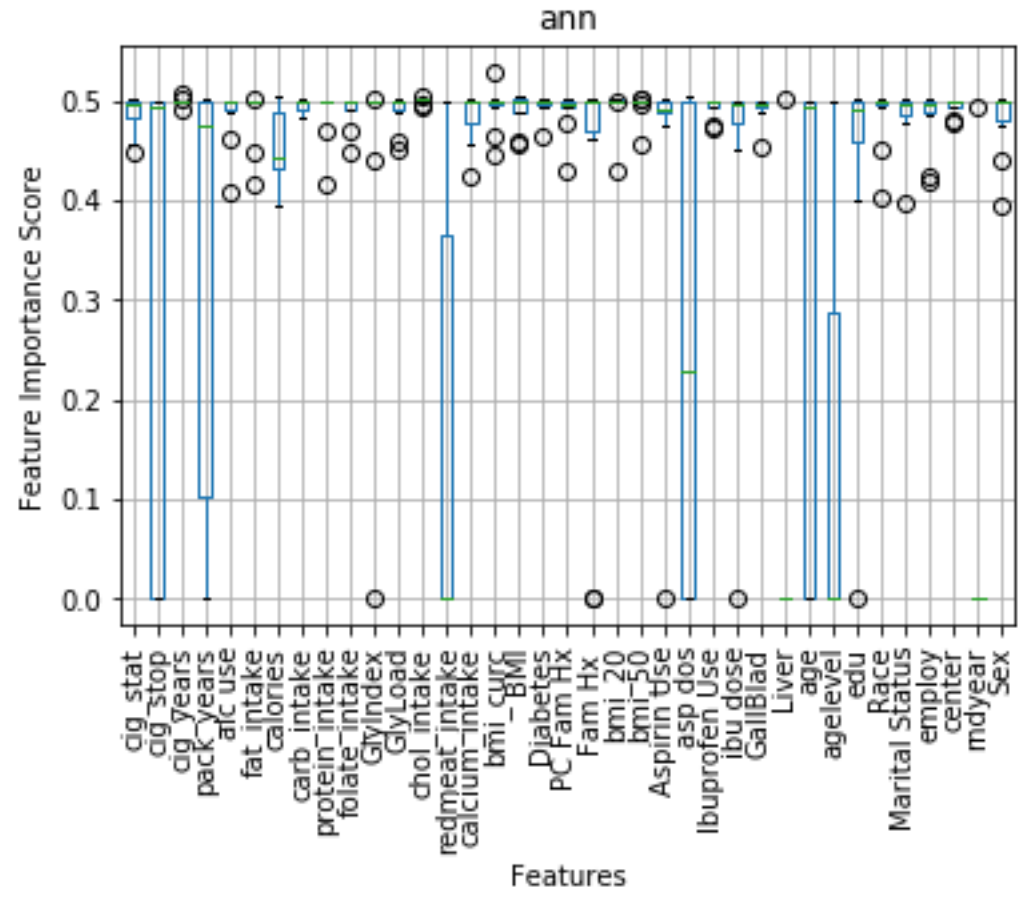

multisurf
Feature Importance Score
0.08
0.06
0.04
0.02
0.00
−0.02
cig_stat
cig_stop
cig_years
pack_years
afc use
fat intake
calories
carb intake
protein intake
folate intake
GlyIndex
GlyLoad
chol intake
redmeat intake
calcium intake
bmi_curc
BMI
Diabetes
pC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

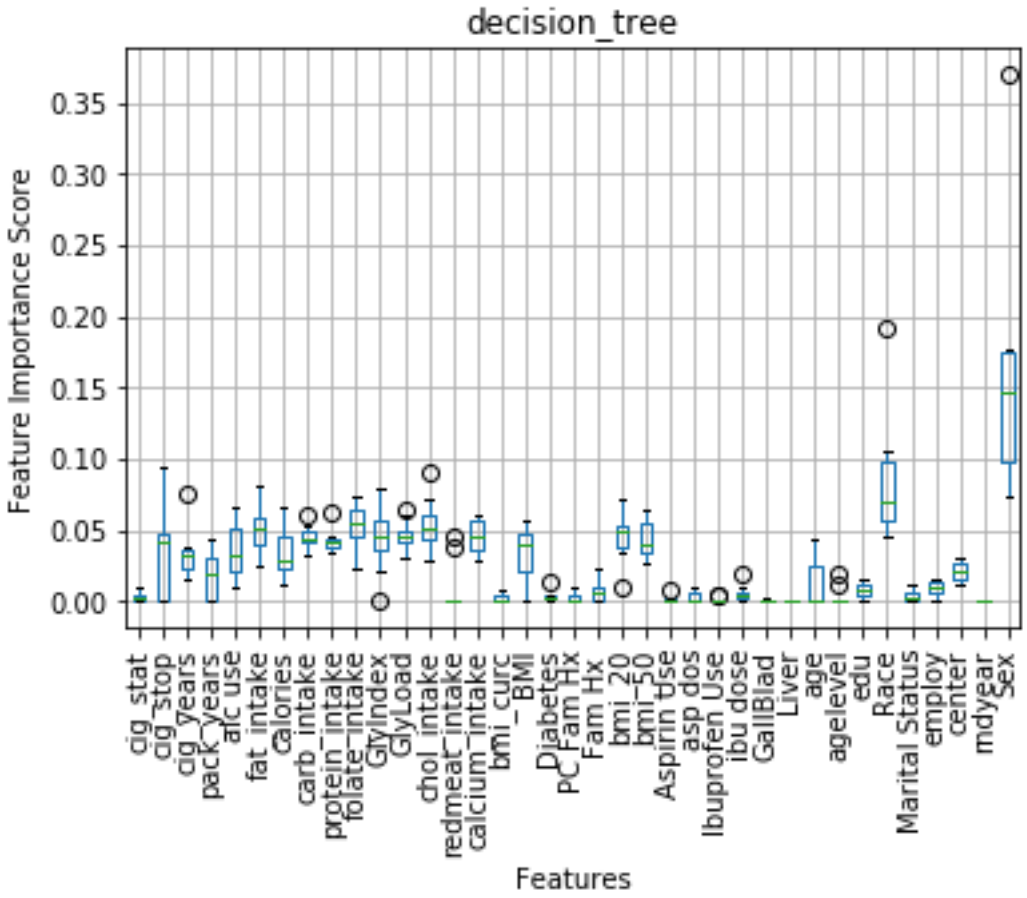

naive_bayes
Feature Importance Score
0.4
0.3
0.2
0.1
0.0
cig_stat
cig_stop
cig_years
pack_years
afc use
fat intake
calories
carb intake
protein intake
folate intake
GlyIndex
GlyLoad
chol intake
redmeat intake
calcium intake
bmi_curc
BMI
Diabetes
pC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features


random_forest
Feature Importance Score
0.00
0.02
0.04
0.06
0.08
cig_stat
cig_stop
cig_years
pack_years
arc use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
BMI
Diabetes
pC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

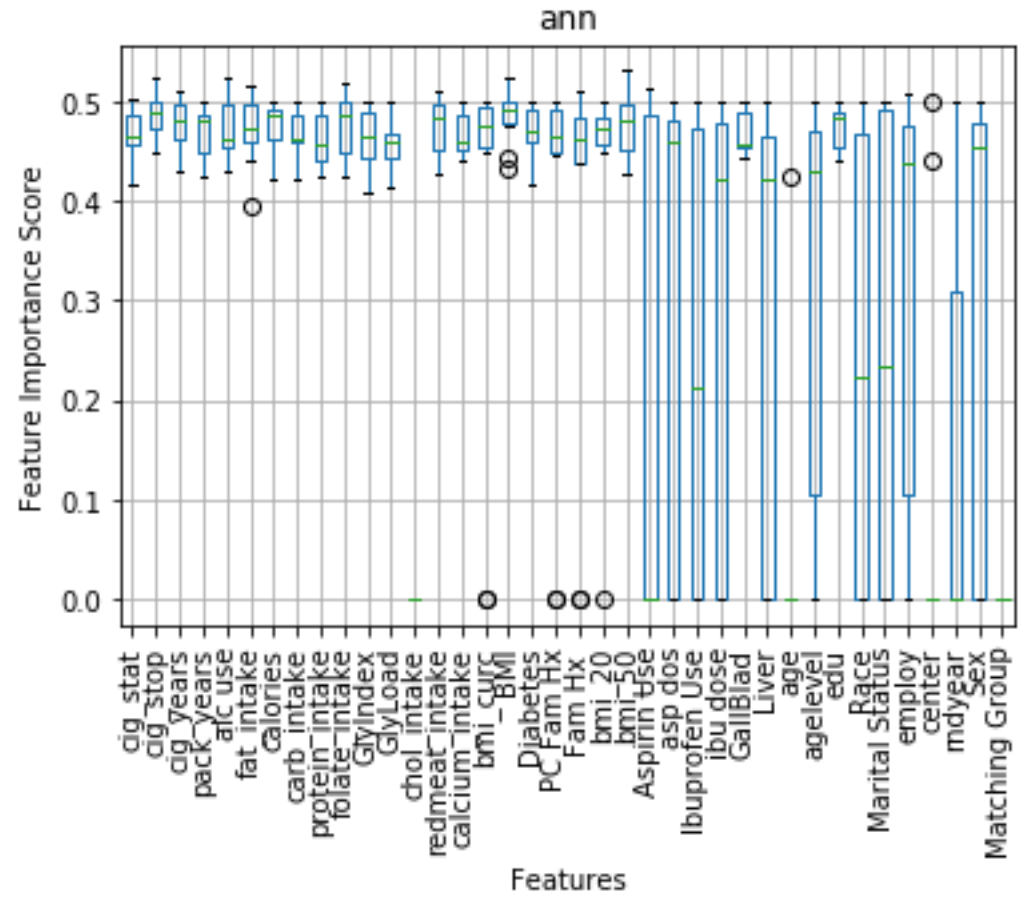

svm
Feature Importance Score
0.0
0.1
0.2
0.3
0.4
0.5
cig_stat
cig_stop
cig_years
pack_years
arc use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
BMI
Diabetes
pC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

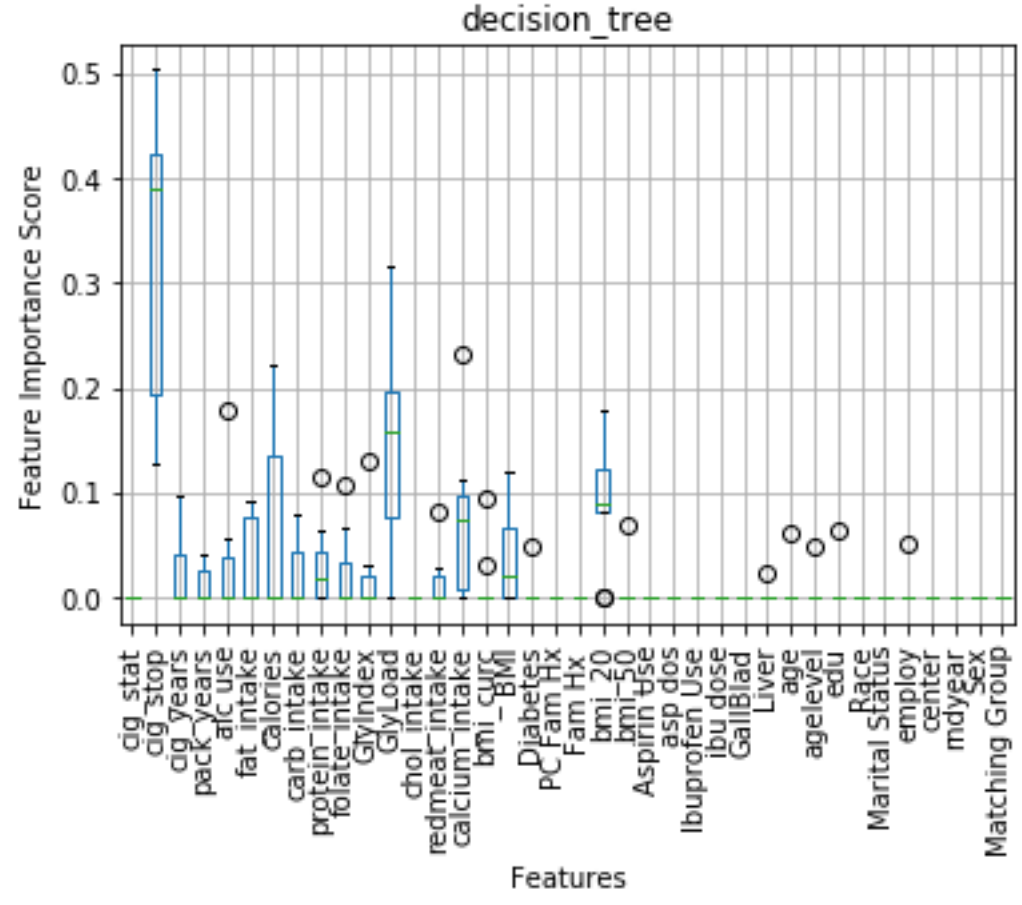

xgboost
Feature Importance Score
0.0
0.1
0.2
0.3
0.4
0.5
cig_stat
cig_stop
cig_years
pack_years
afc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curr
bmi_BMI
Diabetes
pC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

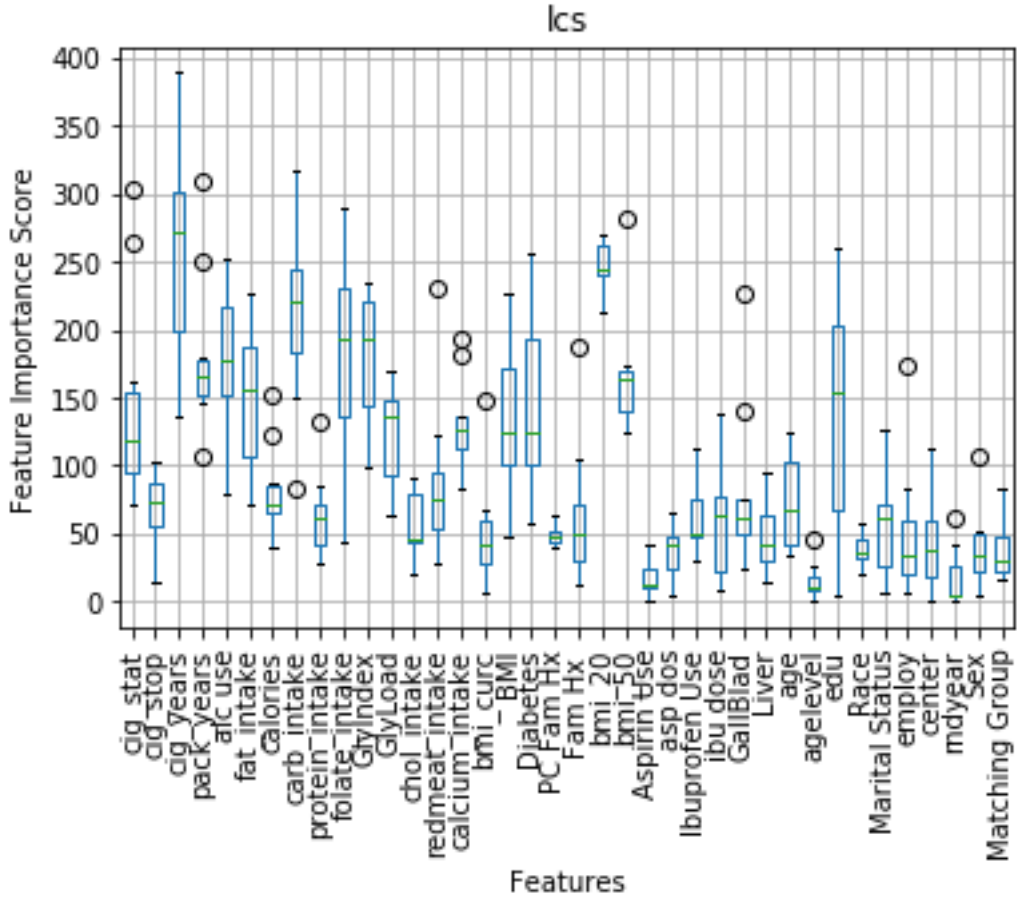

ann
Feature Importance Score
0.0
0.1
0.2
0.3
0.4
0.5
cig_stat
cig_stop
cig_years
pack_years
afc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curr
bmi_BMI
Diabetes
pC Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Features

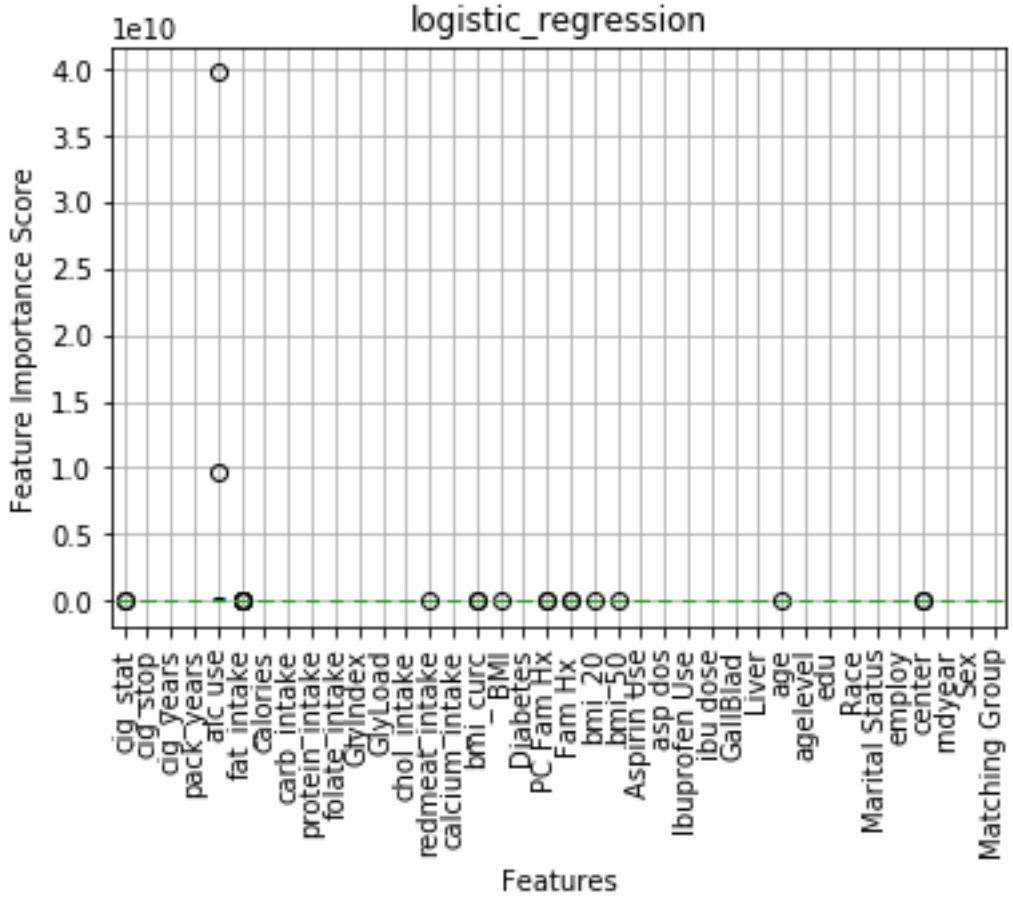


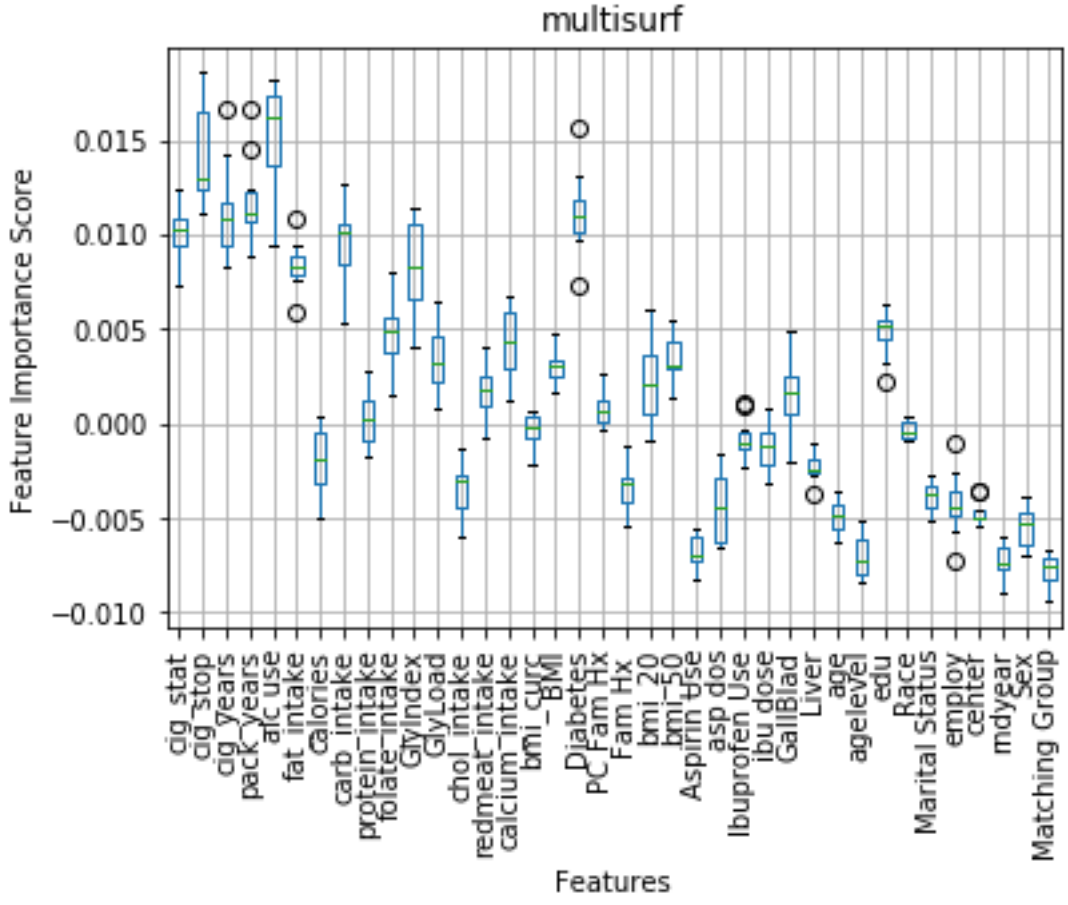
Dataset 3 Feature Importance Boxplots

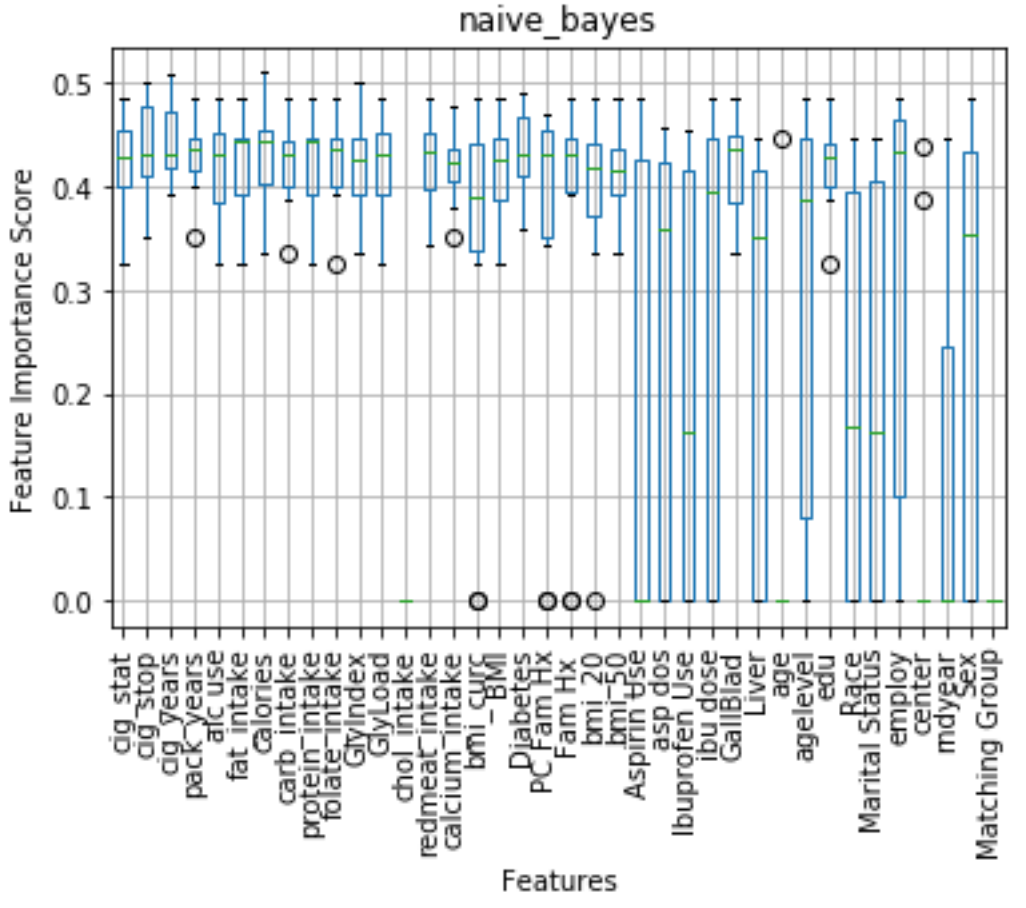

ann
Feature Importance Score
0.0
0.1
0.2
0.3
0.4
0.5
cig_stat
cig_stop
cig_years
pack_years
arc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curr
BMI
Diabetes
PC_Fam_Hx
Fam_Hx
bmi_20
bmi_50
Aspirin_use
asp_dos
Ibuprofen_Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital_Status
employ
center
mdyear
Sex
Matching_Group
Features

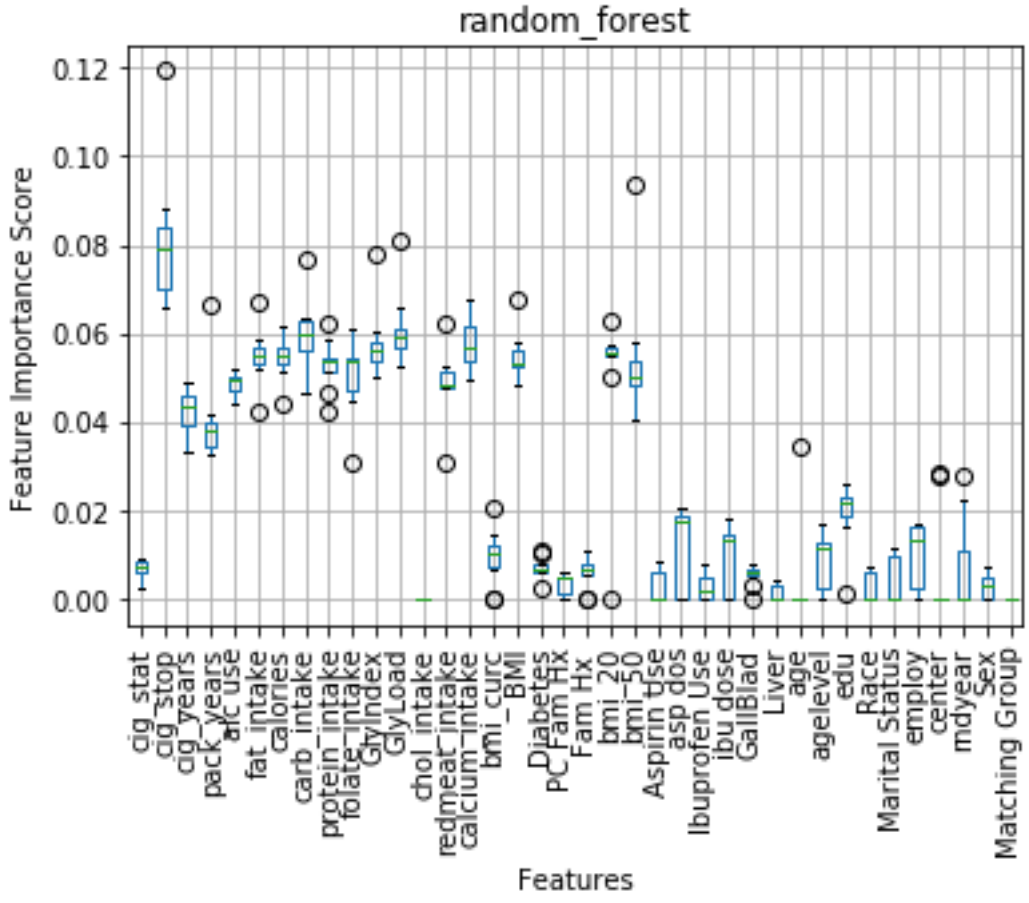

decision_tree
Feature Importance Score
0.0
0.1
0.2
0.3
0.4
0.5
cig_stat
cig_stop
cig_years
pack_years
arc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curr
BMI
Diabetes
PC_Fam_Hx
Fam_Hx
bmi_20
bmi_50
Aspirin_use
asp_dos
Ibuprofen_Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital_Status
employ
center
mdyear
Sex
Matching_Group
Features

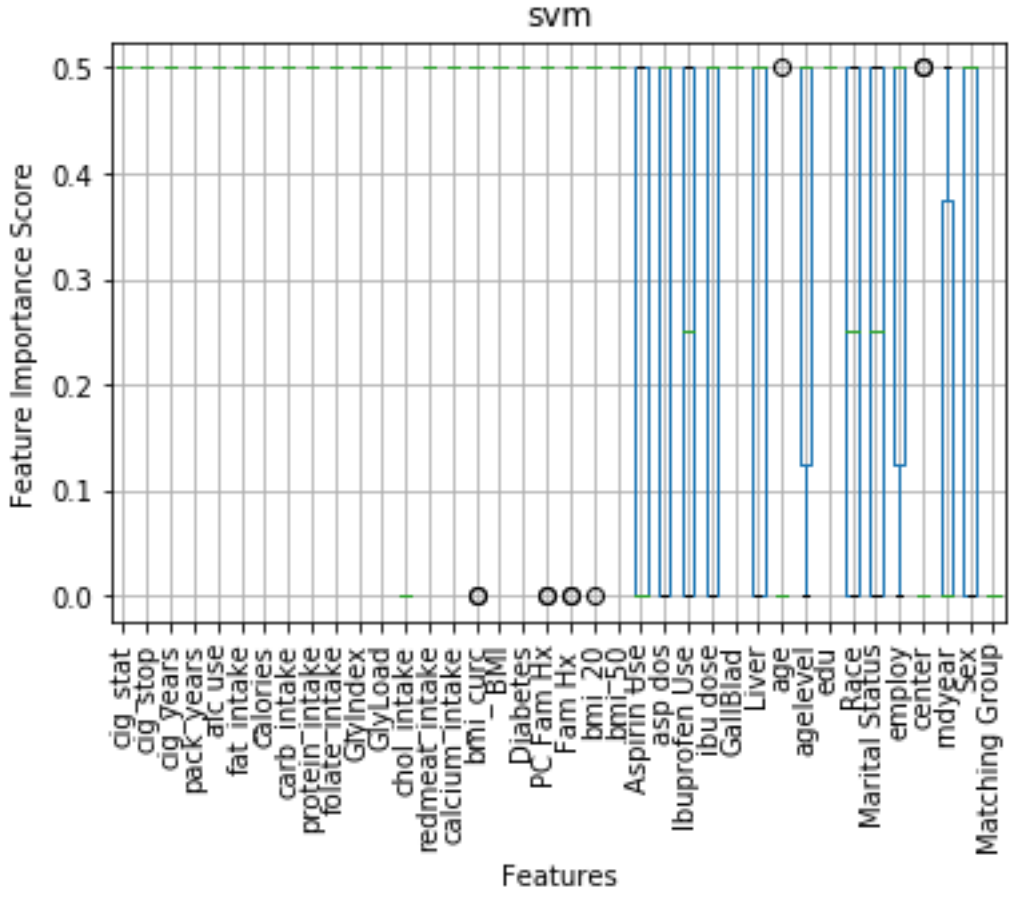

lcs
Feature Importance Score
Features
0
50
100
150
200
250
300
350
400
cig_stat
cig_stop
cig_years
pack_years
alc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
bmi_BMI
Diabetes
PC_Fam_Hx
Fam_Hx
bmi_20
bmi_50
Aspirin_use
asp_dos
Ibuprofen_Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital_Status
employ
center
mdyear
Sex
Matching_Group

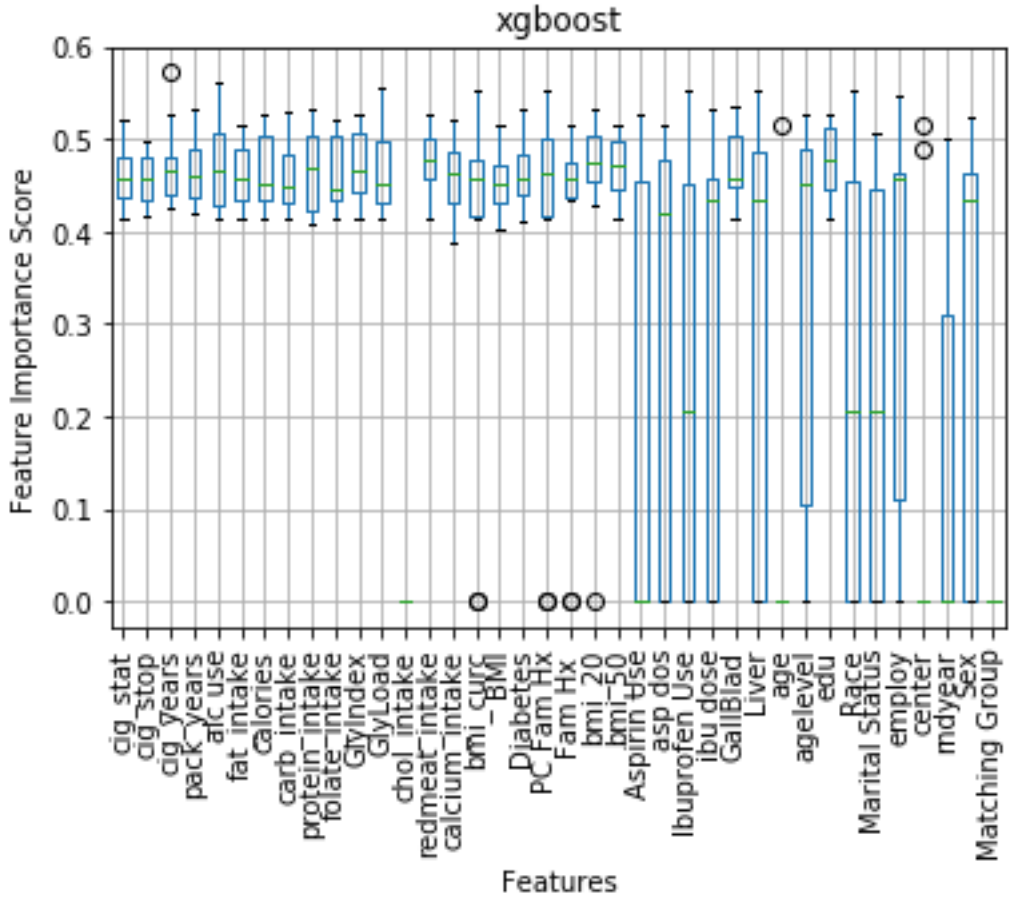

logistic_regression
1e10
Feature Importance Score
Features
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
cig_stat
cig_stop
cig_years
pack_years
alc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
bmi_BMI
Diabetes
PC_Fam_Hx
Fam_Hx
bmi_20
bmi_50
Aspirin_use
asp_dos
Ibuprofen_Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital_Status
employ
center
mdyear
Sex
Matching_Group


multisurf
Feature Importance Score
0.015
0.010
0.005
0.000
−0.005
−0.010
cig_stat
cig_stop
cig_years
pack_years
arc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curr
BMI
Diabetes
PC_Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Matching Group
Features


naive_bayes
Feature Importance Score
0.5
0.4
0.3
0.2
0.1
0.0
cig_stat
cig_stop
cig_years
pack_years
arc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curr
BMI
Diabetes
PC_Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu_dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Matching Group
Features

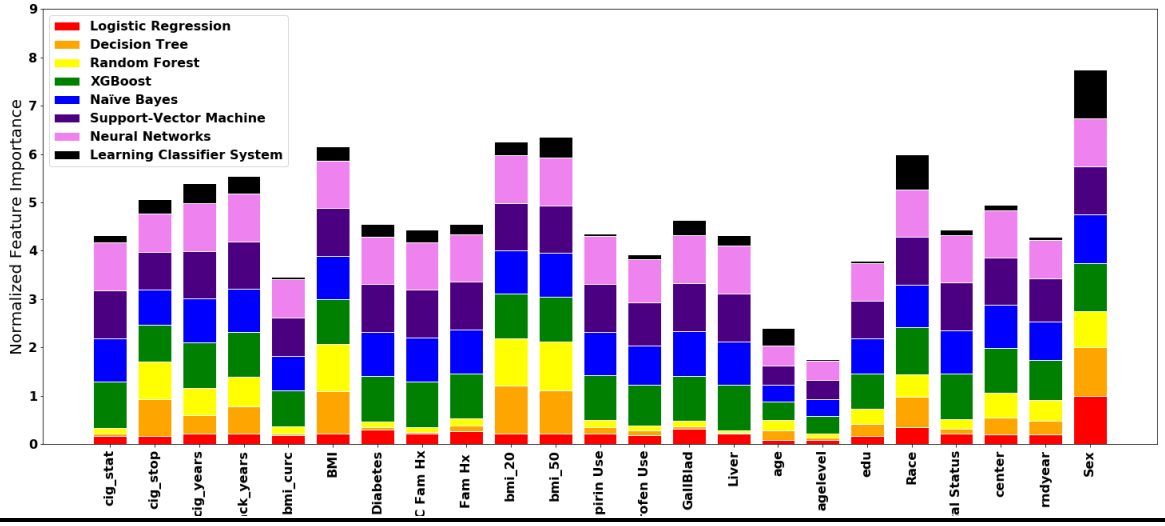

random_forest
Feature Importance Score
0.12
0.10
0.08
0.06
0.04
0.02
0.00
cig_stat
cig_stop
cig_years
pack_years
alc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
BMI
Diabetes
PC_Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Matching Group
Features

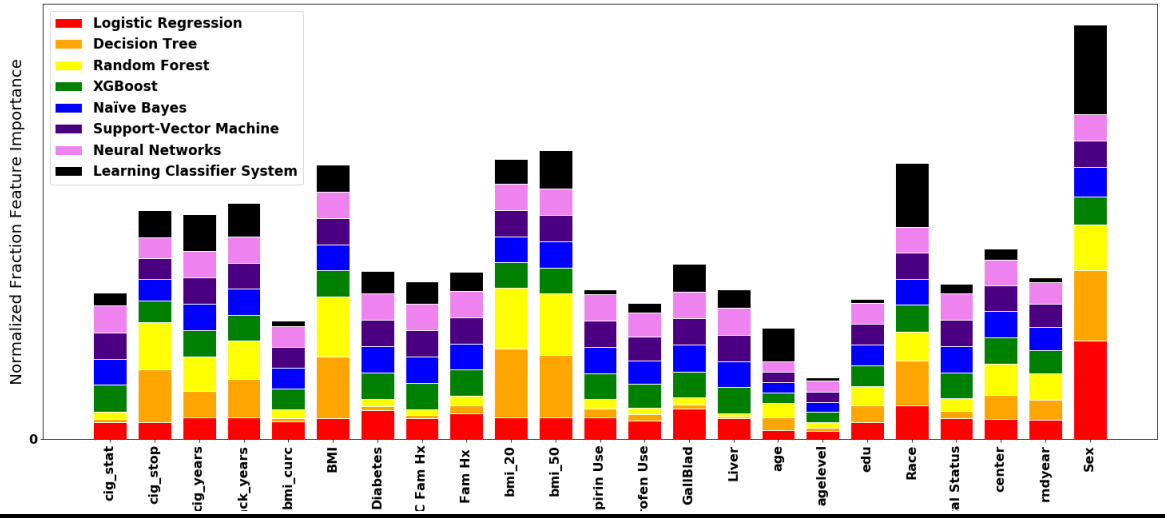

svm
Feature Importance Score
0.5
0.4
0.3
0.2
0.1
0.0
cig_stat
cig_stop
cig_years
pack_years
alc_use
fat_intake
calories
carb_intake
protein_intake
folate_intake
GlyIndex
GlyLoad
chol_intake
redmeat_intake
calcium_intake
bmi_curc
BMI
Diabetes
PC_Fam Hx
Fam Hx
bmi_20
bmi_50
Aspirin use
asp dos
Ibuprofen Use
ibu dose
GallBlad
Liver
age
agelevel
edu
Race
Marital Status
employ
center
mdyear
Sex
Matching Group
Features

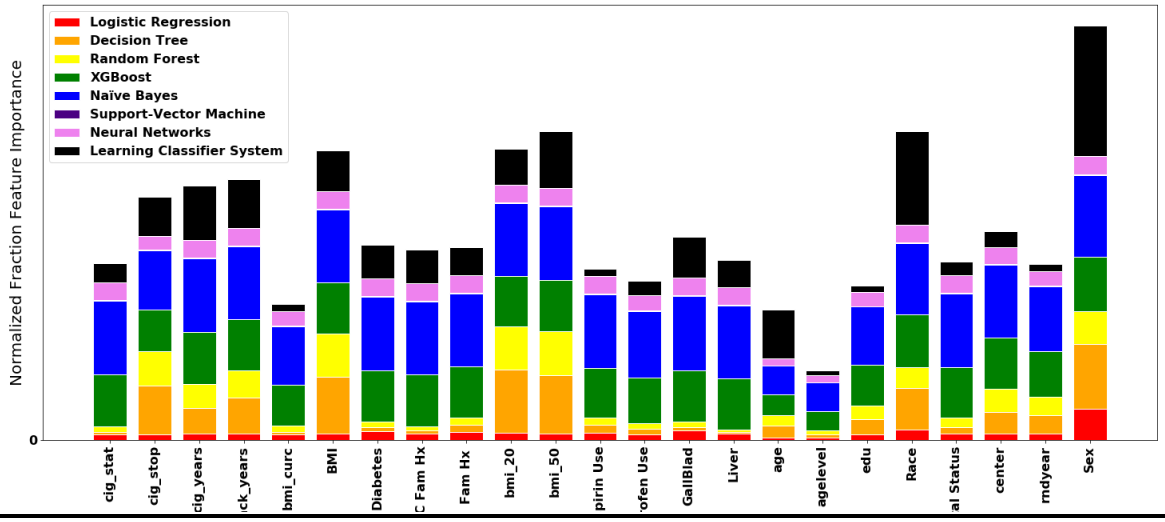


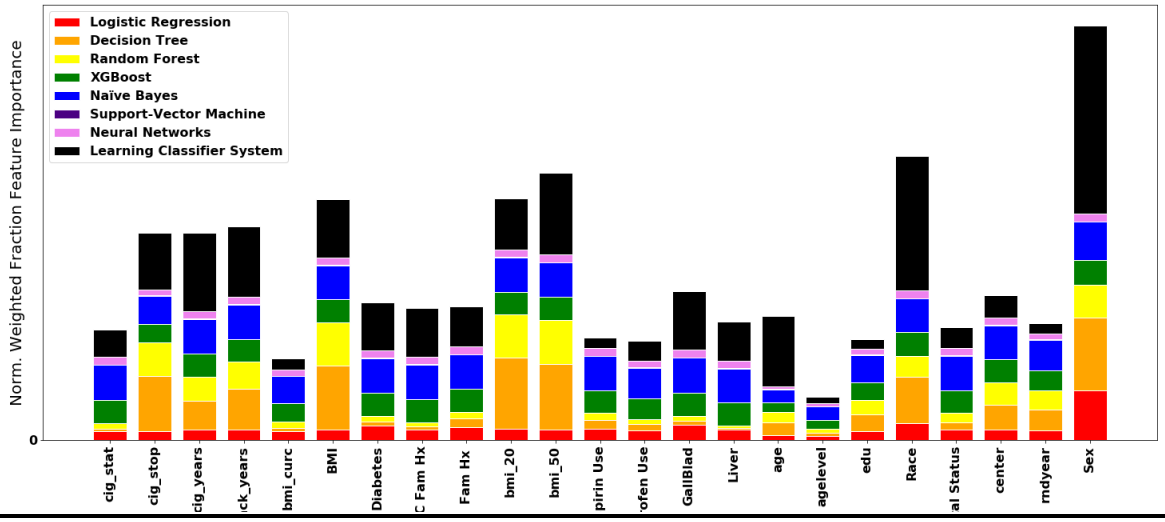
Compound Feature Importance Plots

Dataset 1

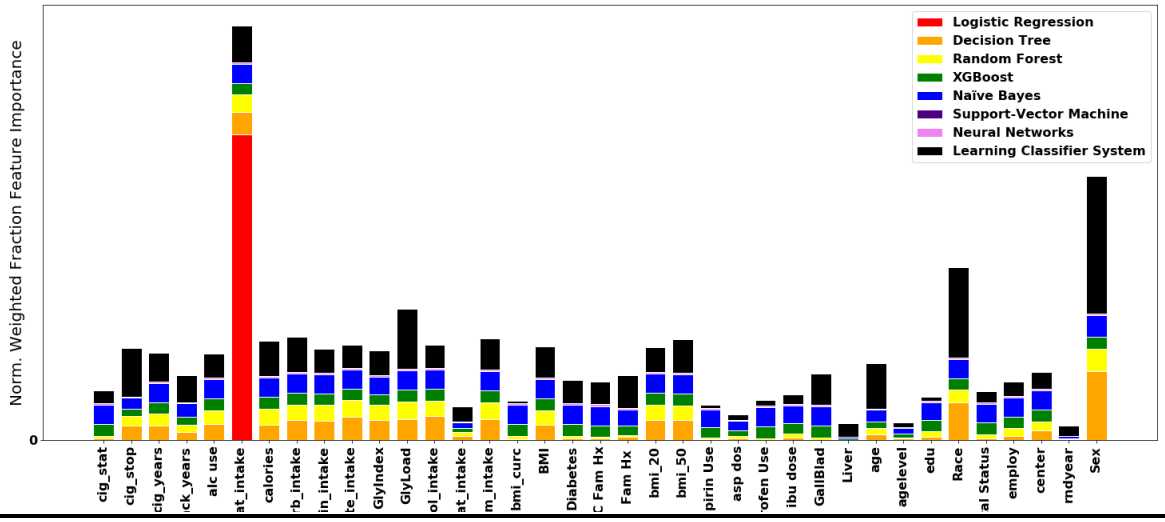

Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
Normalized Feature Importance
0
1
2
3
4
5
6
7
8
9
cig_stat
cig_stop
cig_years
ck_years
bmi_curc
BMI
Diabetes
C Fam Hx
Fam Hx
bmi_20
bmi_50
pirin Use
ofen Use
GallBlad
Liver
age
agelevel
edu
Race
al Status
center
mdyear
Sex

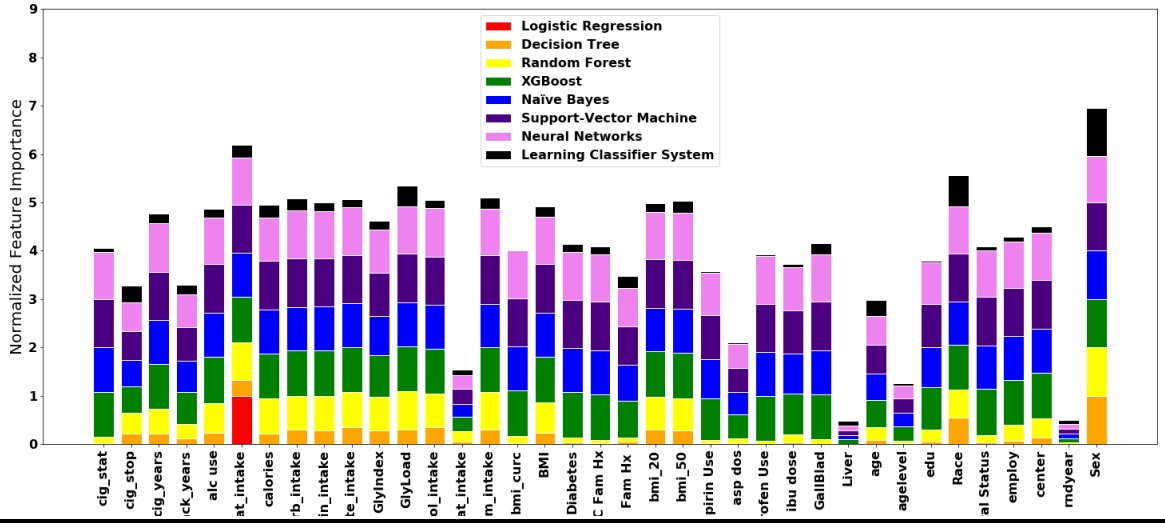

Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
Normalized Fraction Feature Importance
0
cig_stat
cig_stop
cig_years
ck_years
bmi_curc
BMI
Diabetes
C Fam Hx
Fam Hx
bmi_20
bmi_50
pirin Use
ofen Use
GallBlad
Liver
age
agelevel
edu
Race
al Status
center
mdyear
Sex

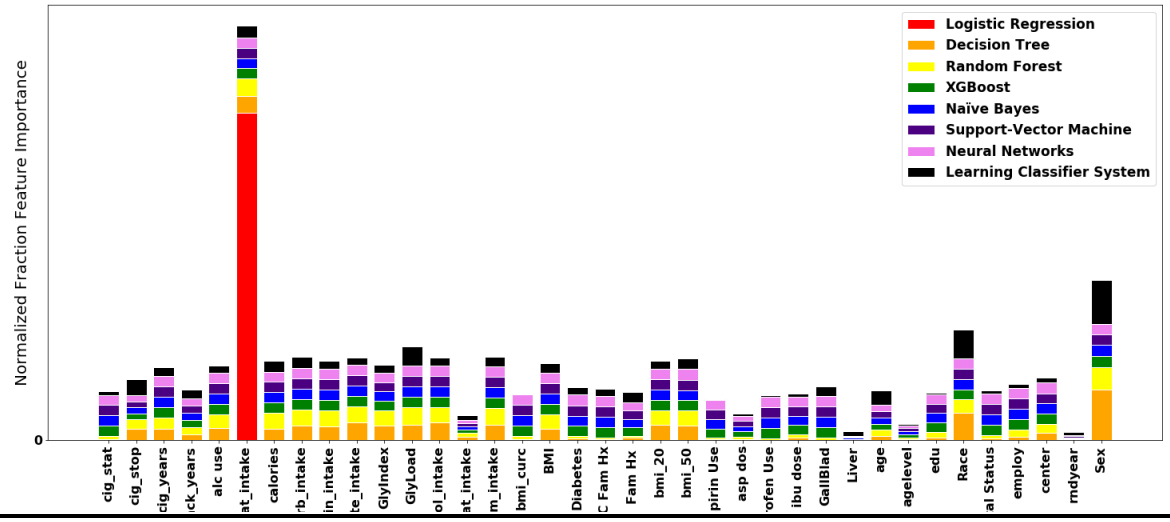

Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
Normalized Fraction Feature Importance
0
cig_stat
cig_stop
cig_years
ck_years
bmi_curc
BMI
Diabetes
C Fam Hx
Fam Hx
bmi_20
bmi_50
pirin Use
ofen Use
GallBlad
Liver
age
agelevel
edu
Race
al Status
center
rmdyear
Sex

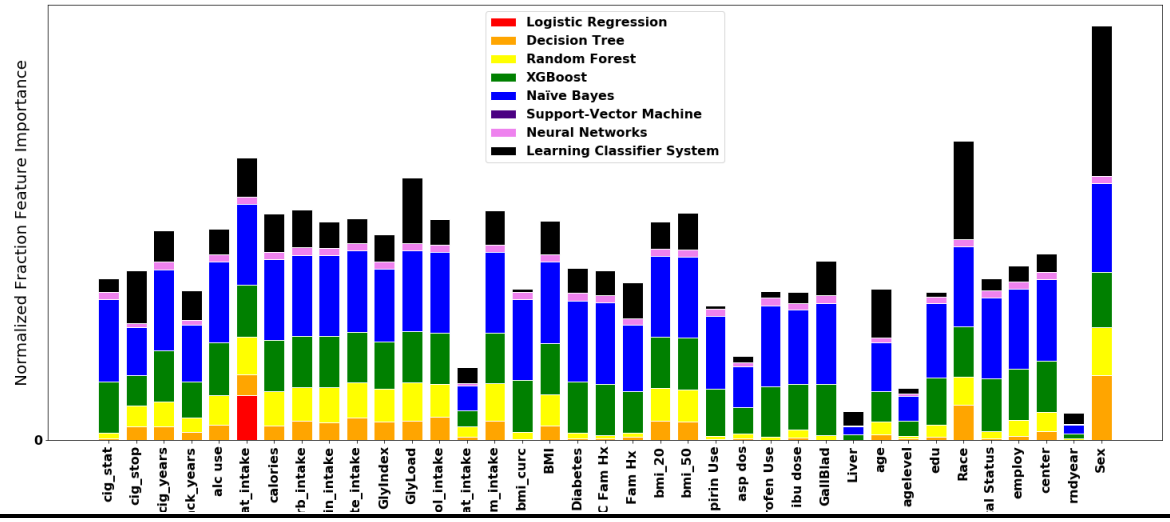

Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
Norm. Weighted Fraction Feature Importance
0
cig_stat
cig_stop
cig_years
ck_years
bmi_curc
BMI
Diabetes
C Fam Hx
Fam Hx
bmi_20
bmi_50
pirin Use
ofen Use
GallBlad
Liver
age
agelevel
edu
Race
al Status
center
rmdyear
Sex

Dataset 2

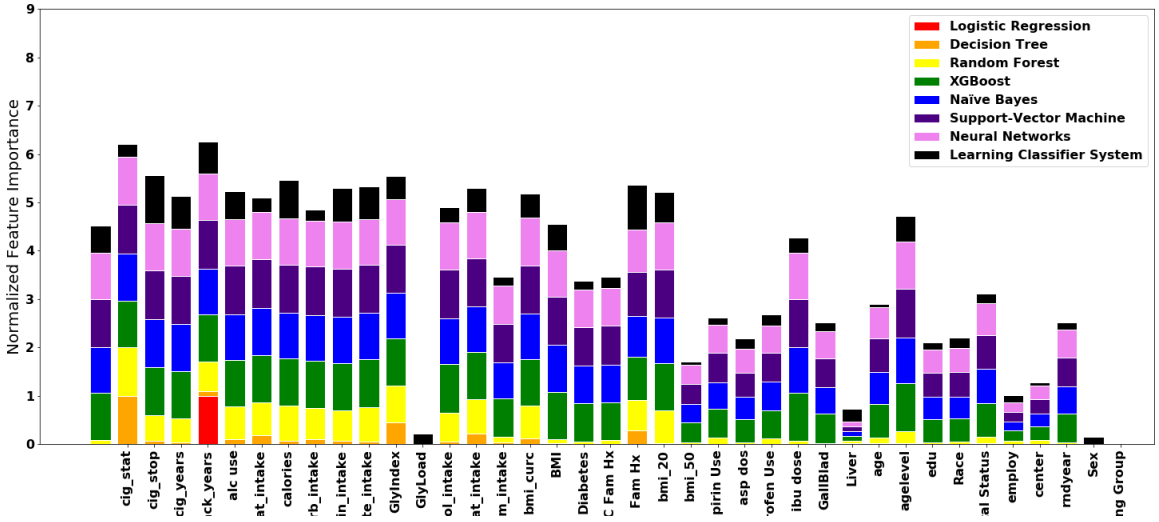

Norm. Weighted Fraction Feature Importance
Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
0
cig_stat
cig_stop
cig_years
alc use
at_intake
calories
in_intake
GlyIndex
GlyLoad
at_intake
m_intake
bmi_curc
BMI
Diabetes
Fam Hx
bmi_20
bmi_50
pirin Use
asp dos
ofen Use
ibu dose
GallBlad
Liver
age
edu
Race
employ
center
rmdyear
Sex

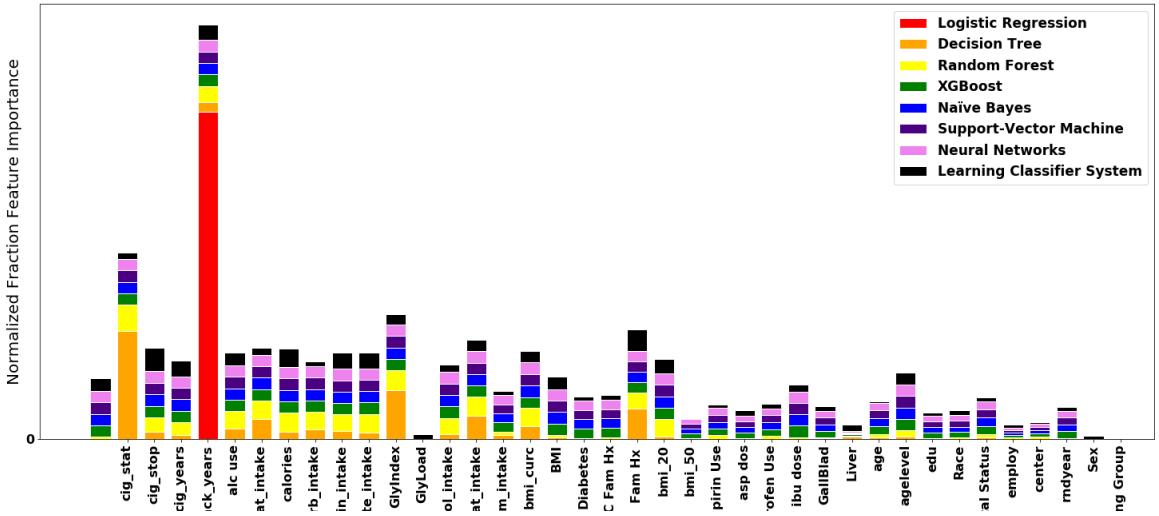

Normalized Feature Importance
Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
0
1
2
3
4
5
6
7
8
9
cig_stat
cig_stop
cig_years
alc use
at_intake
calories
in_intake
GlyIndex
GlyLoad
at_intake
m_intake
bmi_curc
BMI
Diabetes
Fam Hx
bmi_20
bmi_50
pirin Use
asp dos
ofen Use
ibu dose
GallBlad
Liver
age
edu
Race
employ
center
rmdyear
Sex

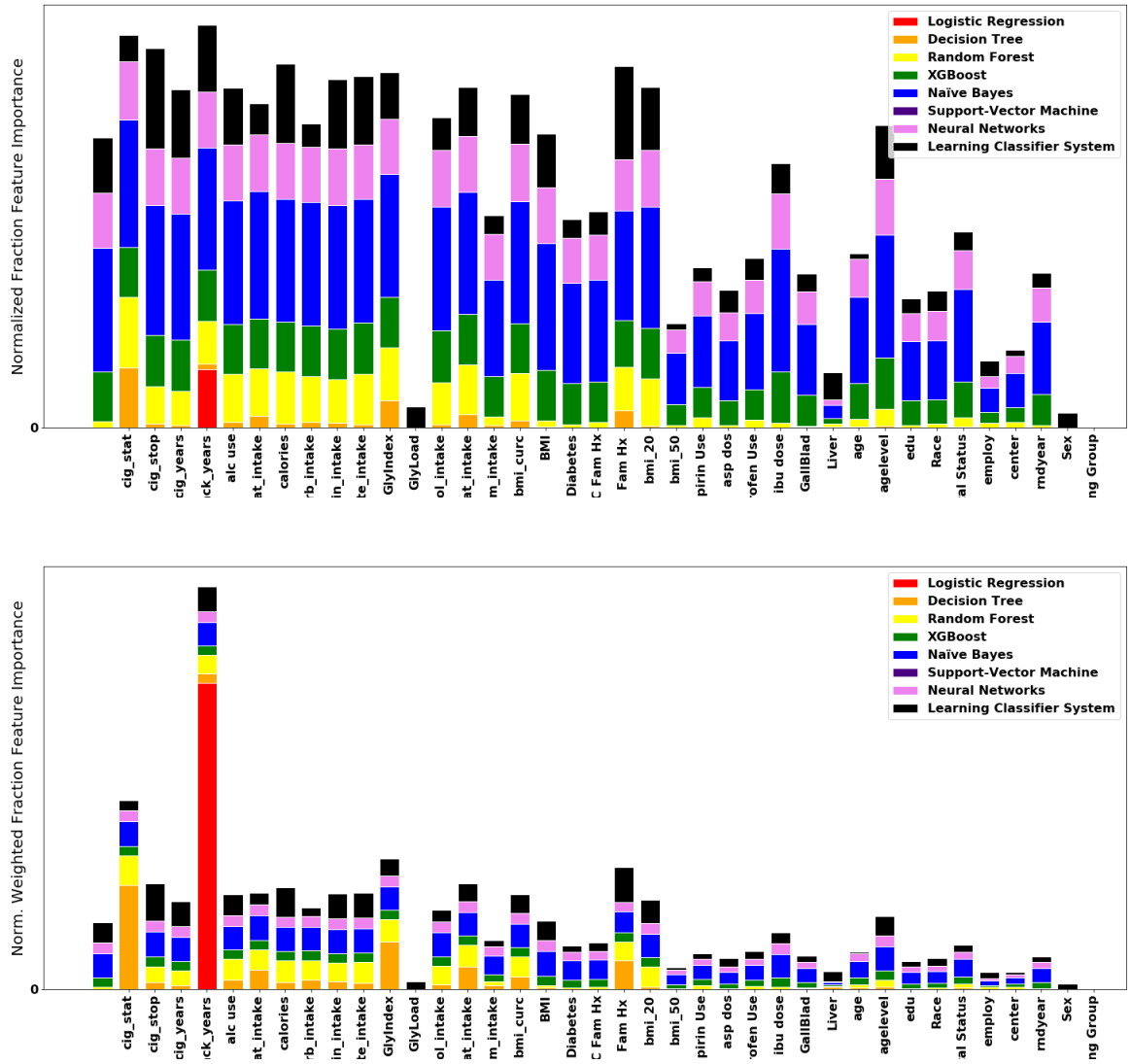

Normalized Fraction Feature Importance
0
Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
cig_stat
cig_stop
cig_years
alc use
at_intake
calories
GlyIndex
GlyLoad
at_intake
bmi_curc
BMI
Diabetes
Fam Hx
bmi_20
bmi_50
asp dos
GallBlad
Liver
age
edu
Race
employ
center
Sex


Normalized Fraction Feature Importance
0
Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
cig_stat
cig_stop
cig_years
alc use
at_intake
calories
GlyIndex
GlyLoad
at_intake
bmi_curc
BMI
Diabetes
Fam Hx
bmi_20
bmi_50
asp dos
GallBlad
Liver
age
edu
Race
employ
center
Sex

Dataset 3


Normalized Feature Importance
Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
cig_stat
cig_stop
cig_years
alc use
calories
GlyIndex
GlyLoad
BMI
Diabetes
Fam Hx
bmi_20
bmi_50
asp dos
GallBlad
Liver
age
edu
Race
employ
center
rmdyear
Sex


Normalized Fraction Feature Importance
Logistic Regression
Decision Tree
Random Forest
XGBoost
Naïve Bayes
Support-Vector Machine
Neural Networks
Learning Classifier System
cig_stat
cig_stop
cig_years
alc use
calories
GlyIndex
GlyLoad
BMI
Diabetes
Fam Hx
bmi_20
bmi_50
asp dos
GallBlad
Liver
age
edu
Race
employ
center
rmdyear
Sex



References

1. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825−2830 (2011).
2. Urbanowicz, R. J. & Moore, J. H. ExSTraCS 2.0: description and evaluation of a scalable learning classifier system. *Evol. Intel.* **8**, 89–116 (2015).
3. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016). doi:10.1145/2939672.2939785
4. Urbanowicz, R. & Browne, W. *Introduction to Learning Classifier Systems | Ryan J. Urbanowicz | Springer*. (Springer, Berlin, Heidelberg, 2017).
5. Urbanowicz, R. J. & Moore, J. H. Learning Classifier Systems: A Complete Introduction, Review, and Roadmap. *J. Artif. Evol. App.* **2009**, 1:1–1:25 (2009).

6. Urbanowicz, R. J., Granizo-Mackenzie, A. & Moore, J. H. An analysis pipeline with statistical and visualization-guided knowledge discovery for Michigan-style learning classifier systems. *IEEE Computational Intelligence Magazine* **7**, 35–45 (2012).
7. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M. & Moore, J. H. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of Biomedical Informatics* **85**, 168–188 (2018).
8. Urbanowicz, R. J., Andrew, A. S., Karagas, M. R. & Moore, J. H. Role of genetic heterogeneity and epistasis in bladder cancer susceptibility and outcome: a learning classifier system approach. *J Am Med Inform Assoc* **20**, 603–612 (2013).
9. Urbanowicz, R., Granizo-Mackenzie, A. & Moore, J. Instance-linked Attribute Tracking and Feedback for Michigan-style Supervised Learning Classifier Systems. in *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation* 927–934 (ACM, 2012). doi:10.1145/2330163.2330291