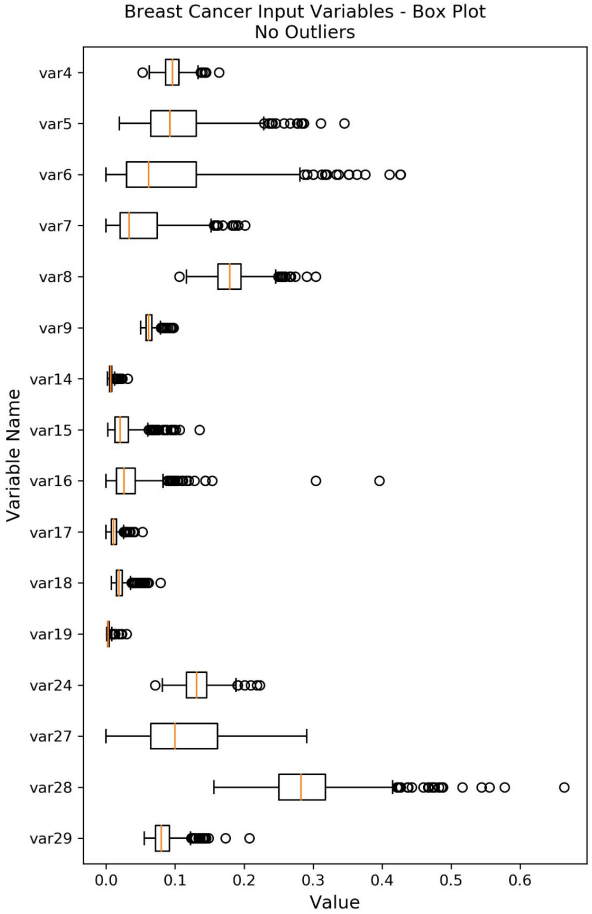
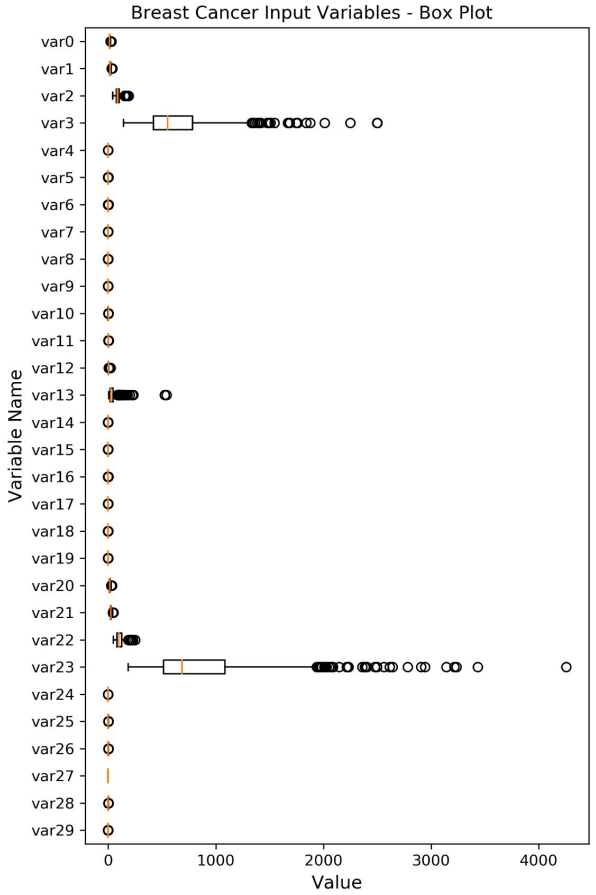
DSI Project

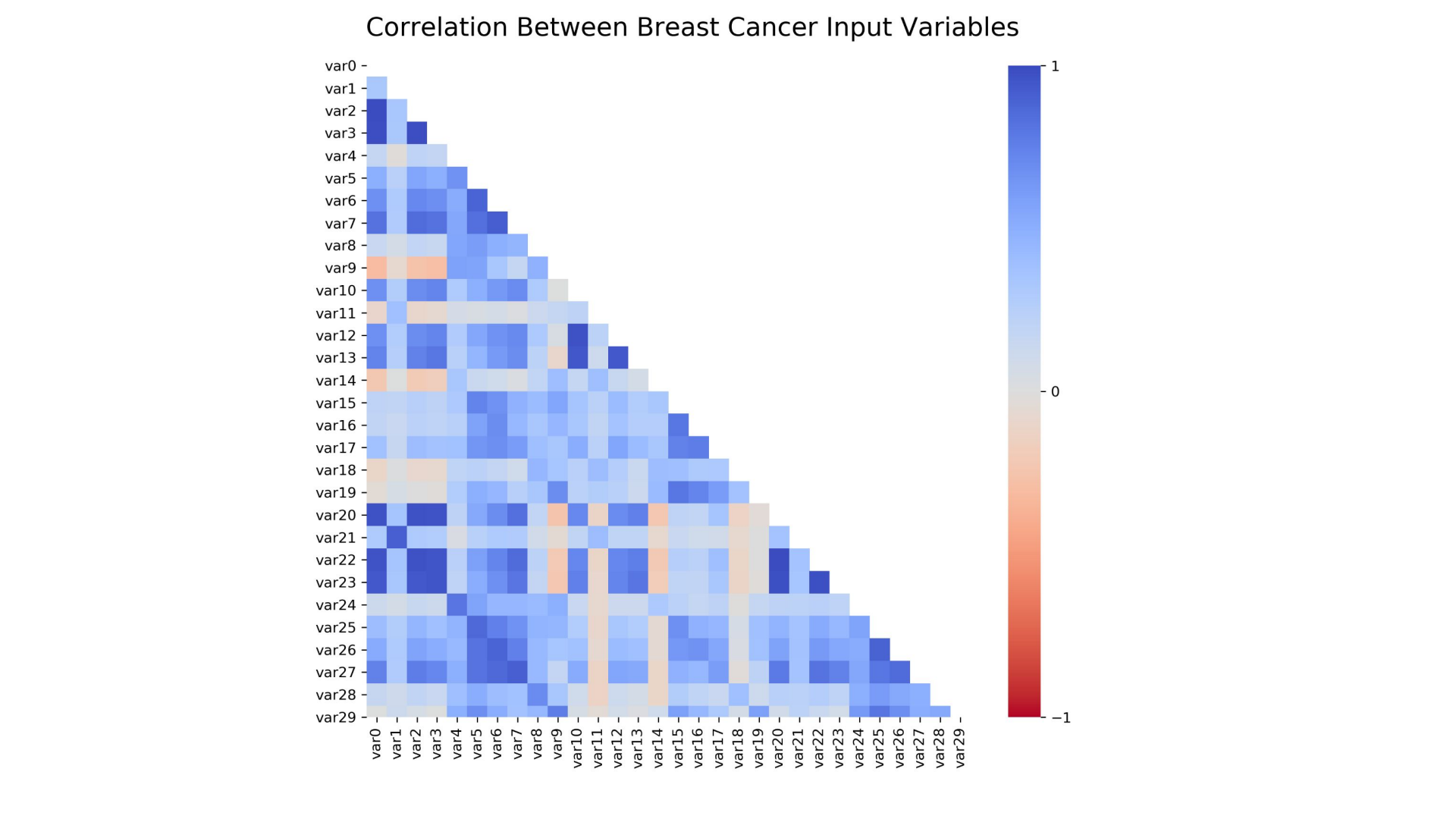
Combined Data Visualizations
John Koenig - 16JUN2021

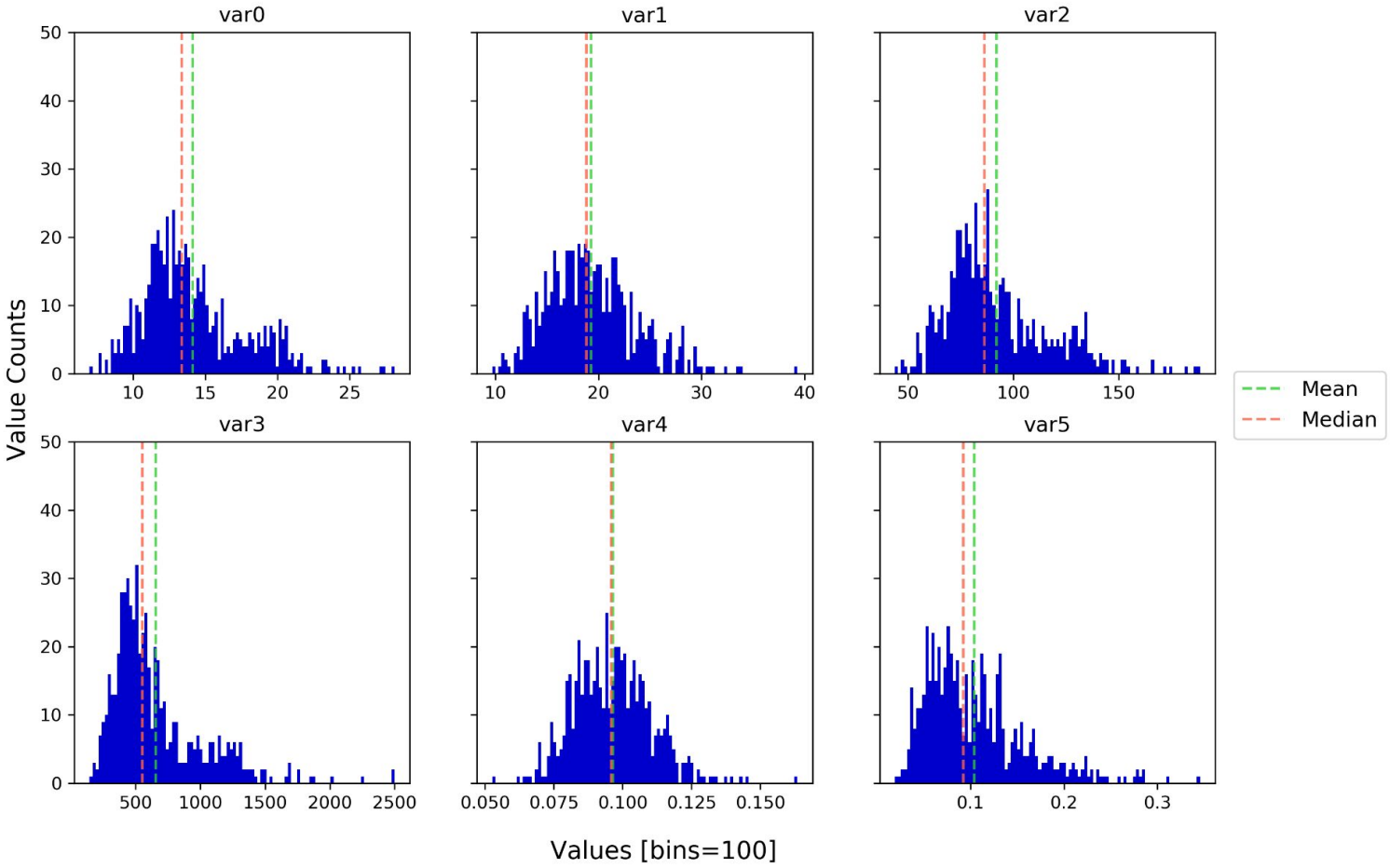
Combined Data Visualizations

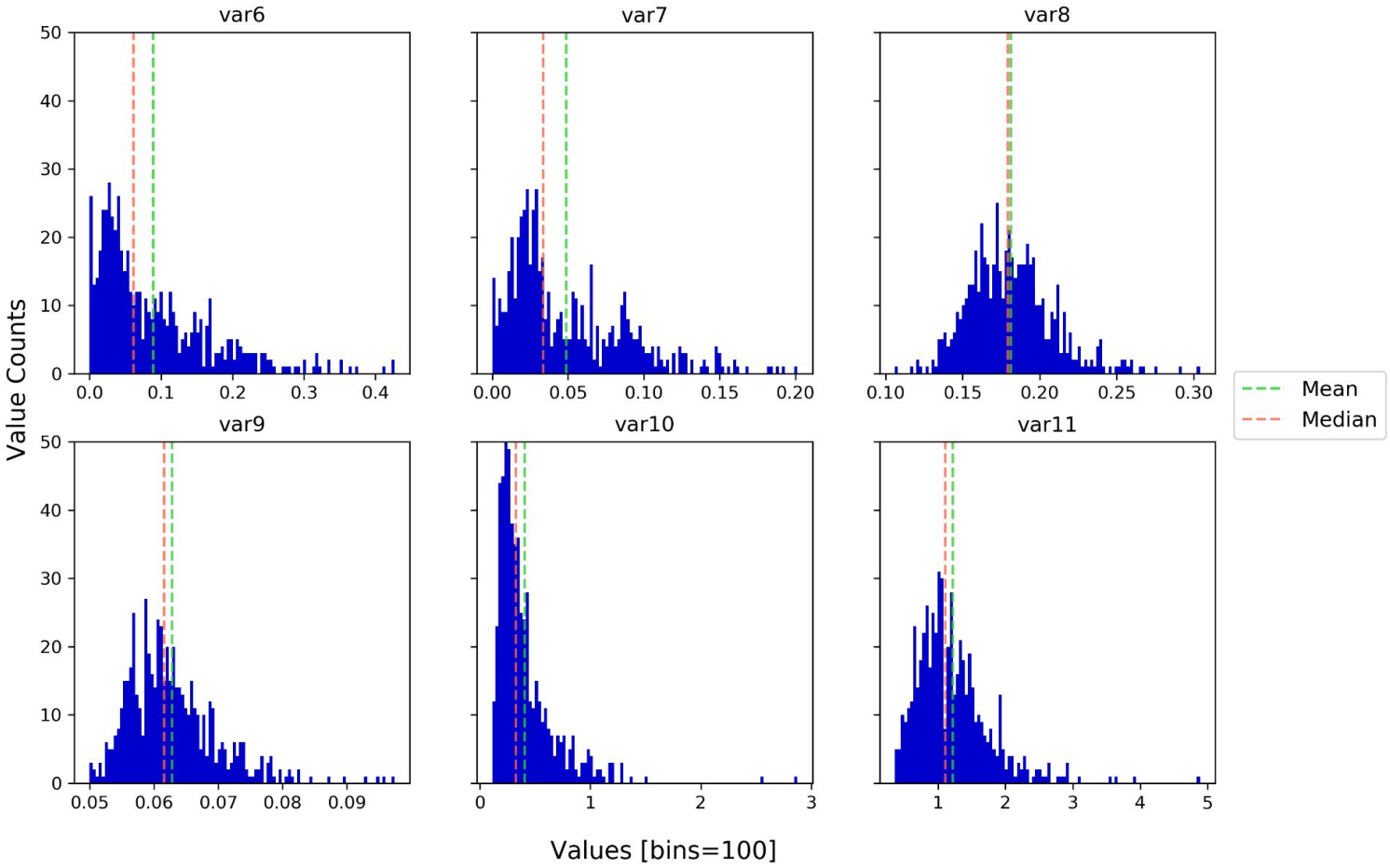
- Box Plots
- Correlation Heatmap
- 6 x Histogram Multi-plots
- Confusion Matrix Heatmaps
- Top 10 Models Table
- Visualize Decision Tree
- Multi-ROC Curve Plot

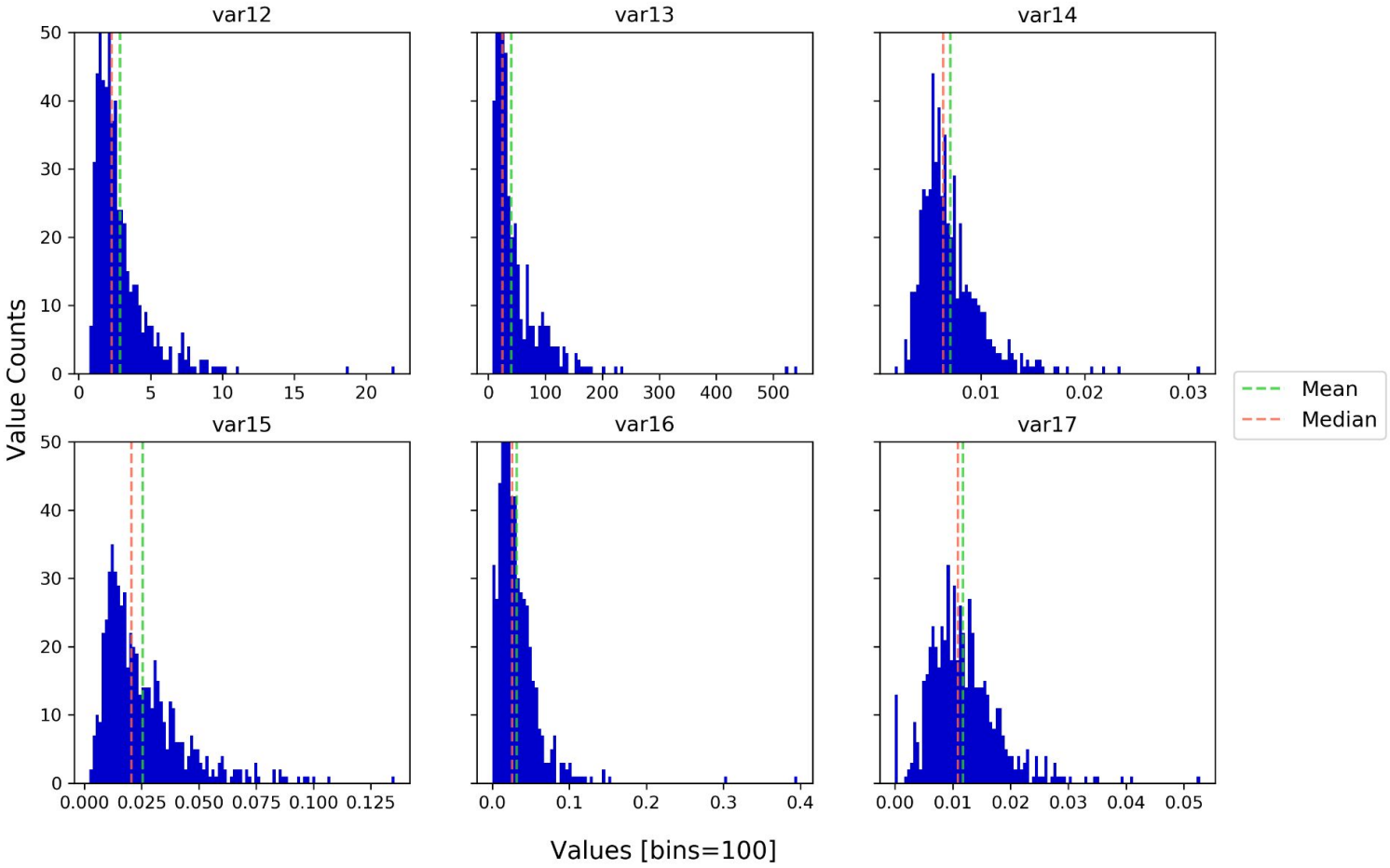

Breast Cancer Input Variables - Box Plot
No Outliers
Variable Name
var4
var5
var6
var7
var8
var9
var14
var15
var16
var17
var18
var19
var24
var27
var28
var29
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Value

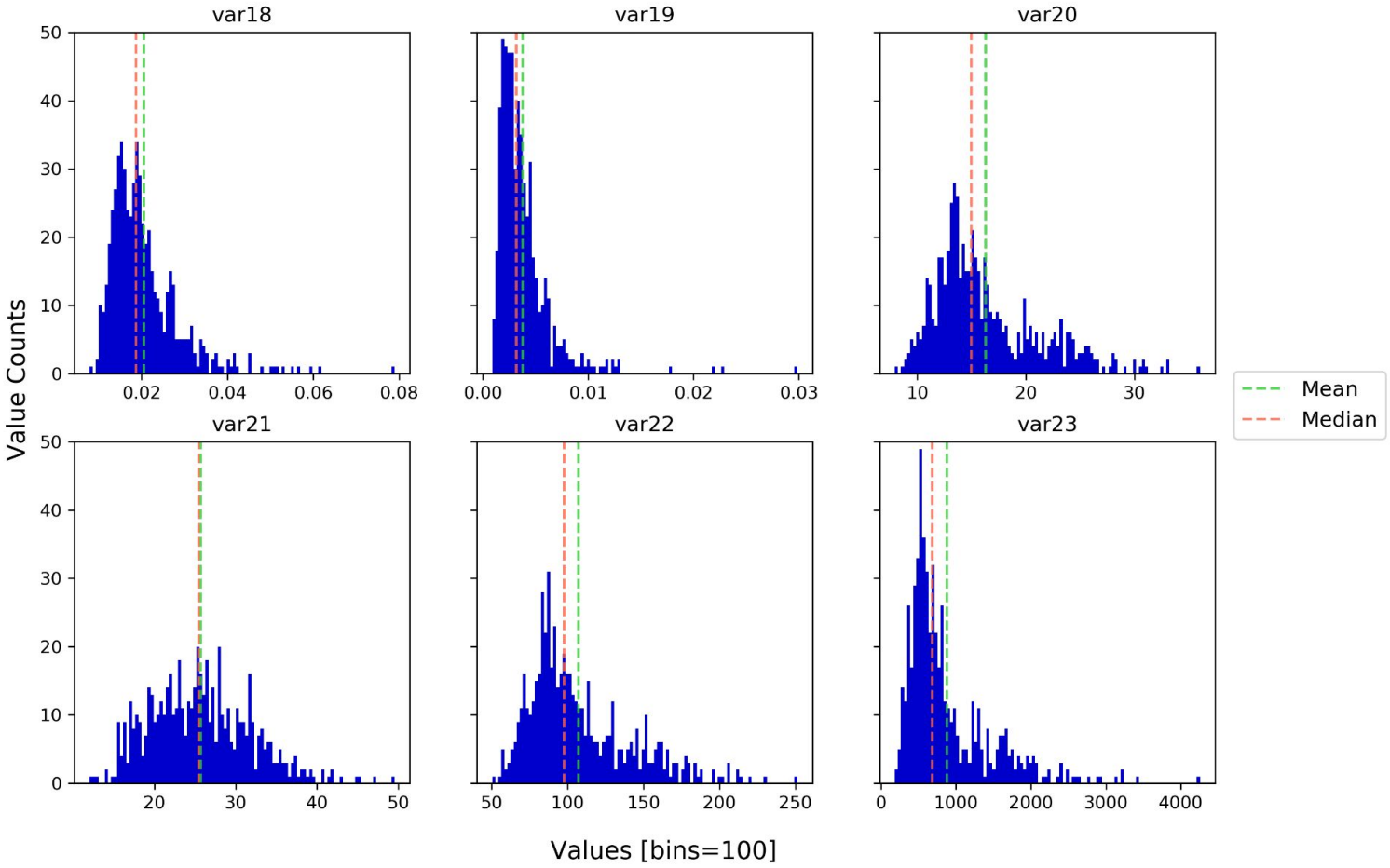

Breast Cancer Input Variables - Box Plot
Variable Name
var0
var1
var2
var3
var4
var5
var6
var7
var8
var9
var10
var11
var12
var13
var14
var15
var16
var17
var18
var19
var20
var21
var22
var23
var24
var25
var26
var27
var28
var29
0
1000
2000
3000
4000
Value

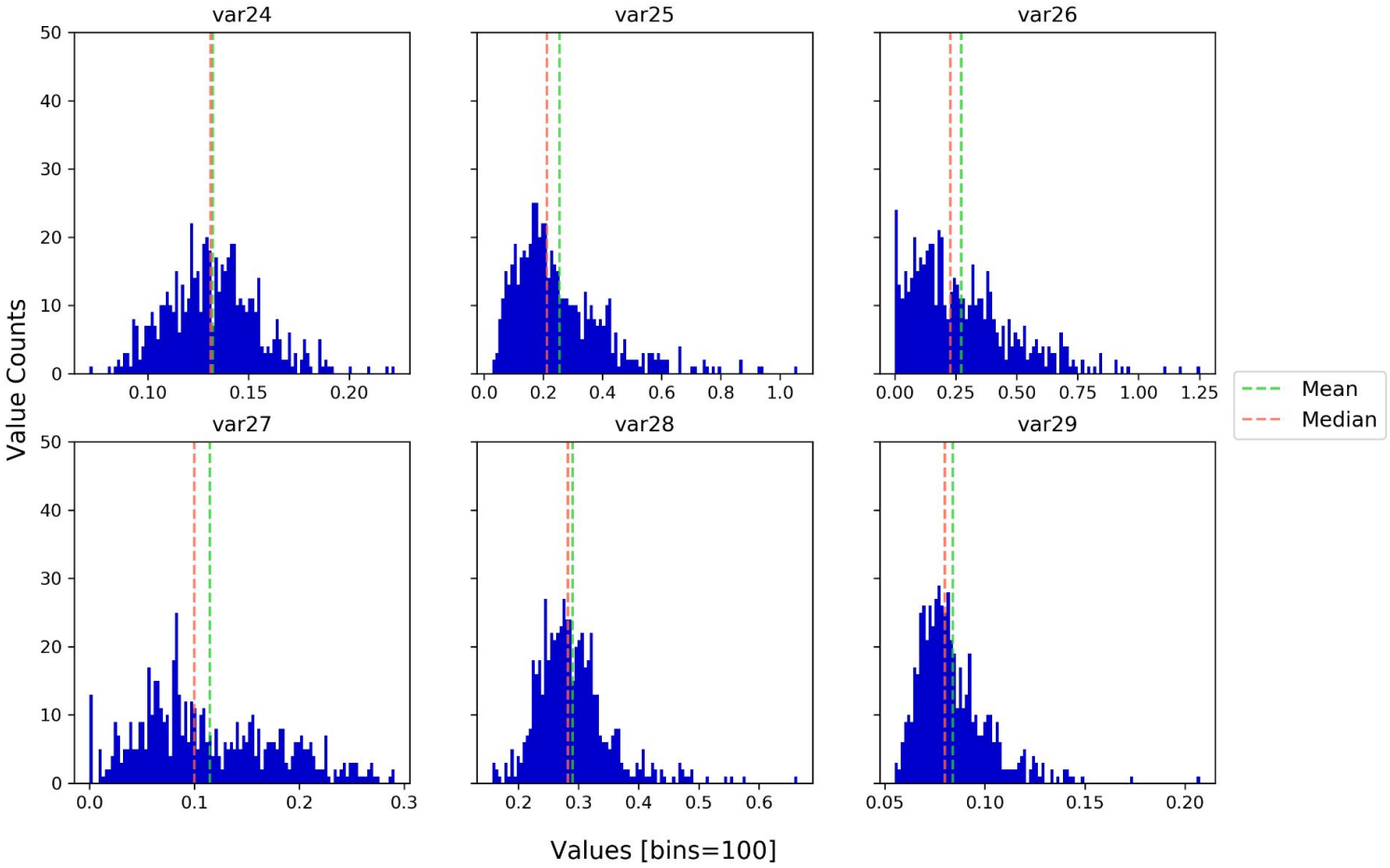

Correlation Between Breast Cancer Input Variables
var0
var1
var2
var3
var4
var5
var6
var7
var8
var9
var10
var11
var12
var13
var14
var15
var16
var17
var18
var19
var20
var21
var22
var23
var24
var25
var26
var27
var28
var29
1
0
−1


Histograms of Breast Cancer Input Variables var0 - var5
Note: X-axis is not constant between plots
var0
var1
var2
var3
var4
var5
Value Counts
Values [bins=100]
Mean
Median

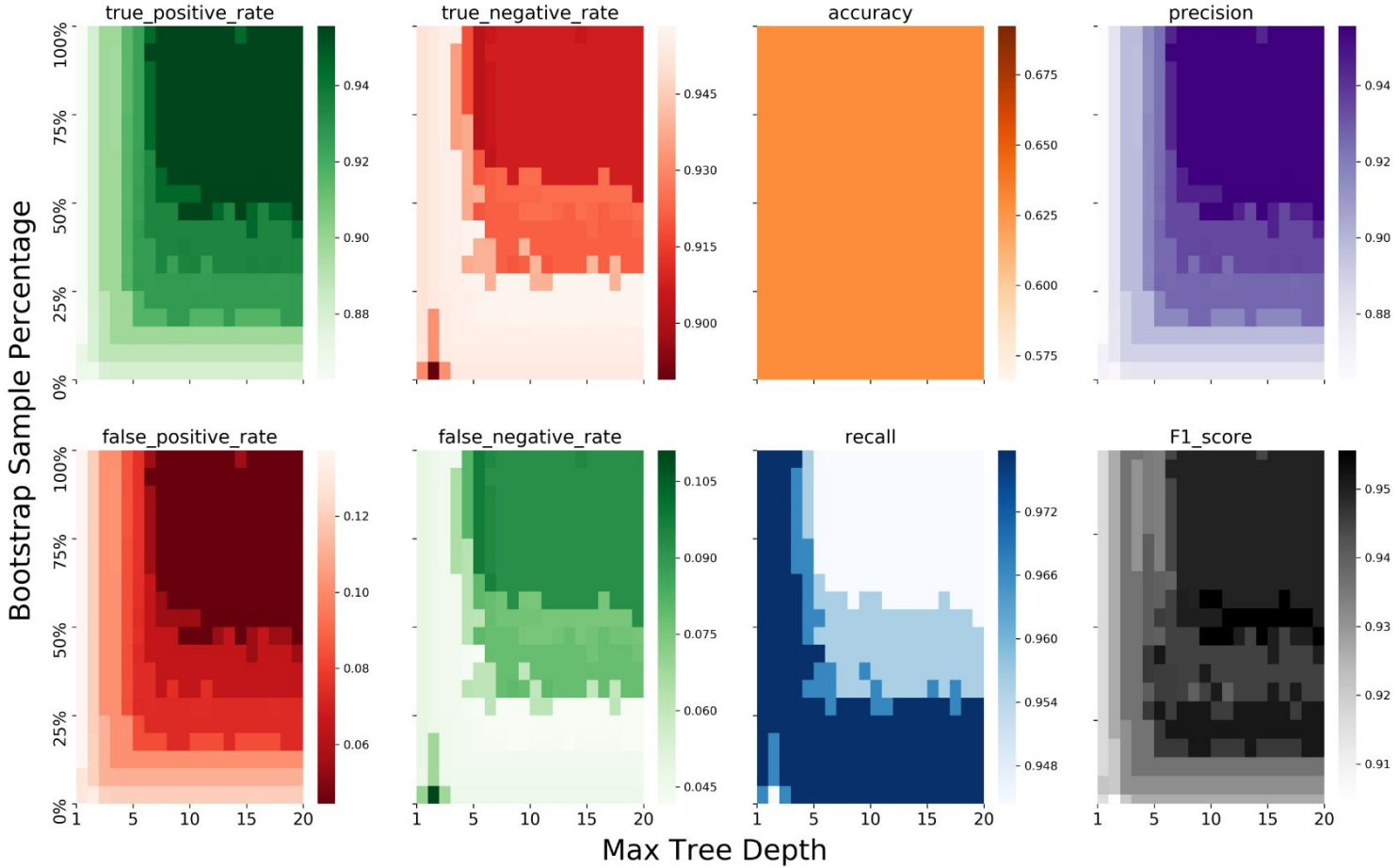

Histograms of Breast Cancer Input Variables var6 - var11
Note: X-axis is not constant between plots
var6
var7
var8
var9
var10
var11
Value Counts
Values [bins=100]
Mean
Median


Histograms of Breast Cancer Input Variables var12 - var17
Note: X-axis is not constant between plots
var12
var13
var14
var15
var16
var17
Value Counts
Values [bins=100]
Mean
Median


Histograms of Breast Cancer Input Variables var18 - var23
Note: X-axis is not constant between plots
var18
var19
var20
var21
var22
var23
Value Counts
Values [bins=100]
Mean
Median


Histograms of Breast Cancer Input Variables var24 - var29
Note: X-axis is not constant between plots
var24
var25
var26
var27
var28
var29
Value Counts
Values [bins=100]
Mean
Median

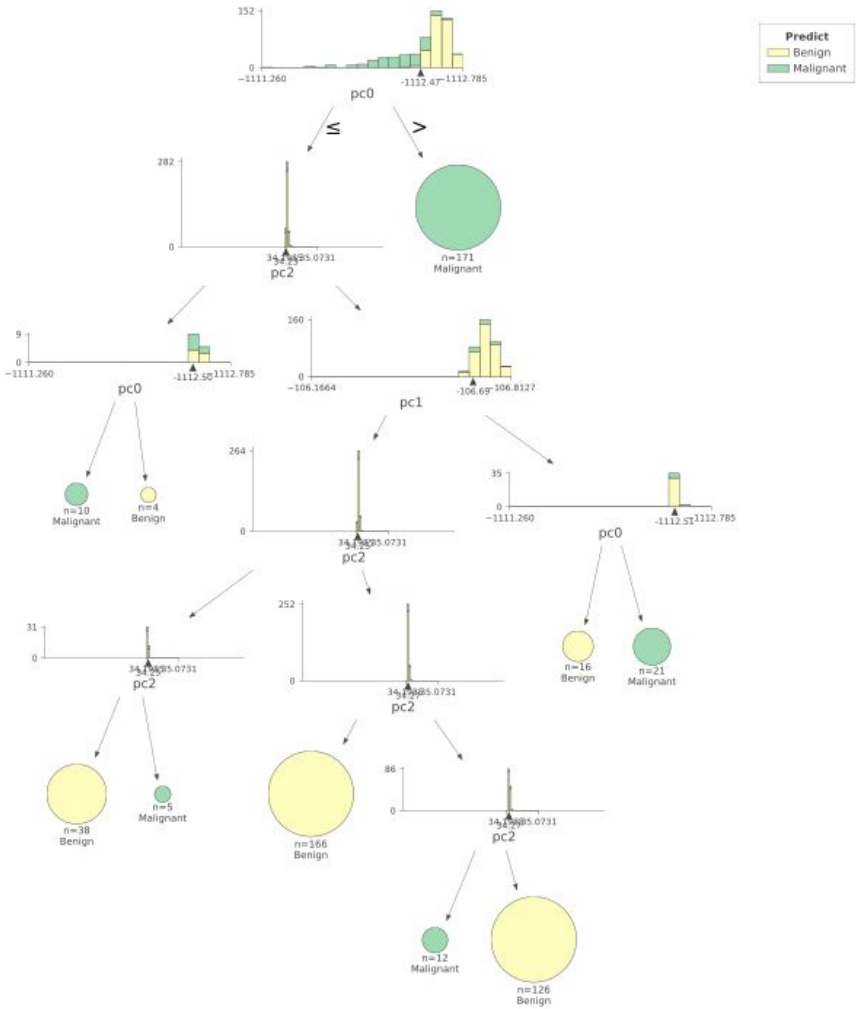
Test Performance Heatmaps

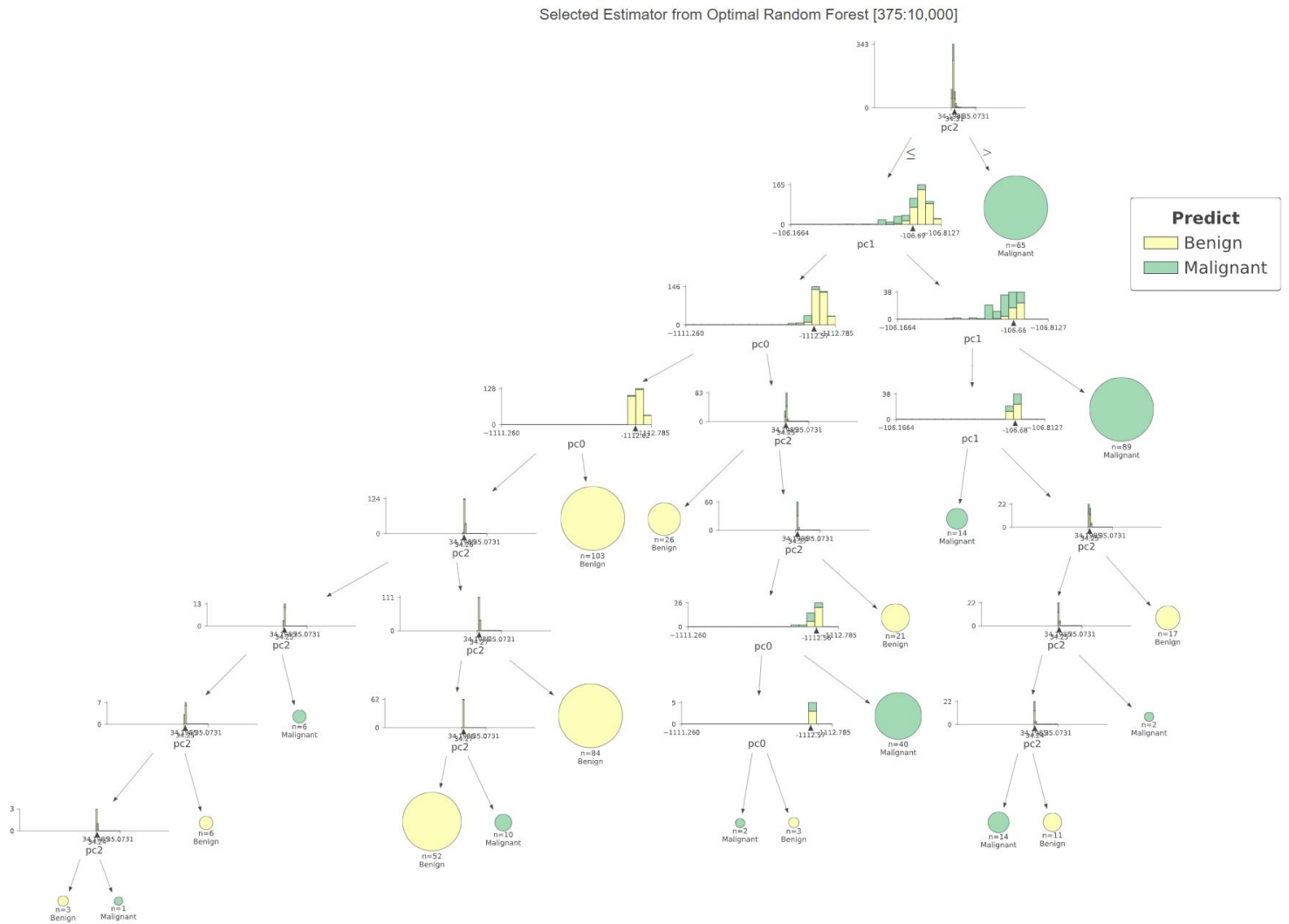


Top 10 Potential Random Forest Models

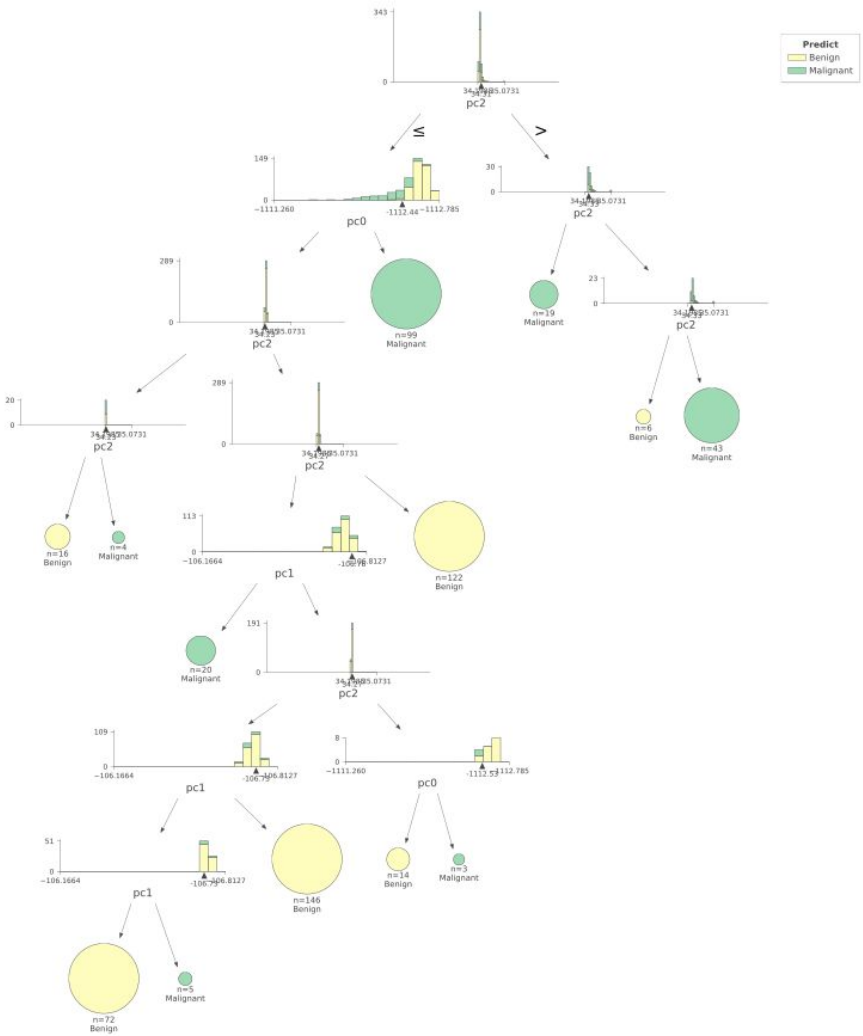
	test_split	total_records	n_trees	sample_percent	depth	true_positive	true_negative	false_positive	false_negative	true_positive_rate	true_negative_rate	false_positive_rate	false_negative_rate	accuracy	precision	recall	F1_score	ROC_area
RF_trees10000_sample0.2_depth9	0.25	143	10000	0.2	9	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.2_depth10	0.25	143	10000	0.2	10	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.2_depth14	0.25	143	10000	0.2	14	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.2_depth19	0.25	143	10000	0.2	19	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.2_depth20	0.25	143	10000	0.2	20	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.25_depth7	0.25	143	10000	0.25	7	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.25_depth8	0.25	143	10000	0.25	8	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.25_depth9	0.25	143	10000	0.25	9	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.25_depth10	0.25	143	10000	0.25	10	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229
RF_trees10000_sample0.25_depth11	0.25	143	10000	0.25	11	88	46	7	2	0.9263	0.9583	0.0737	0.0417	0.6294	0.9263	0.9778	0.9514	0.9229

Selected Estimator from Optimal Random Forest [9119:10,000]

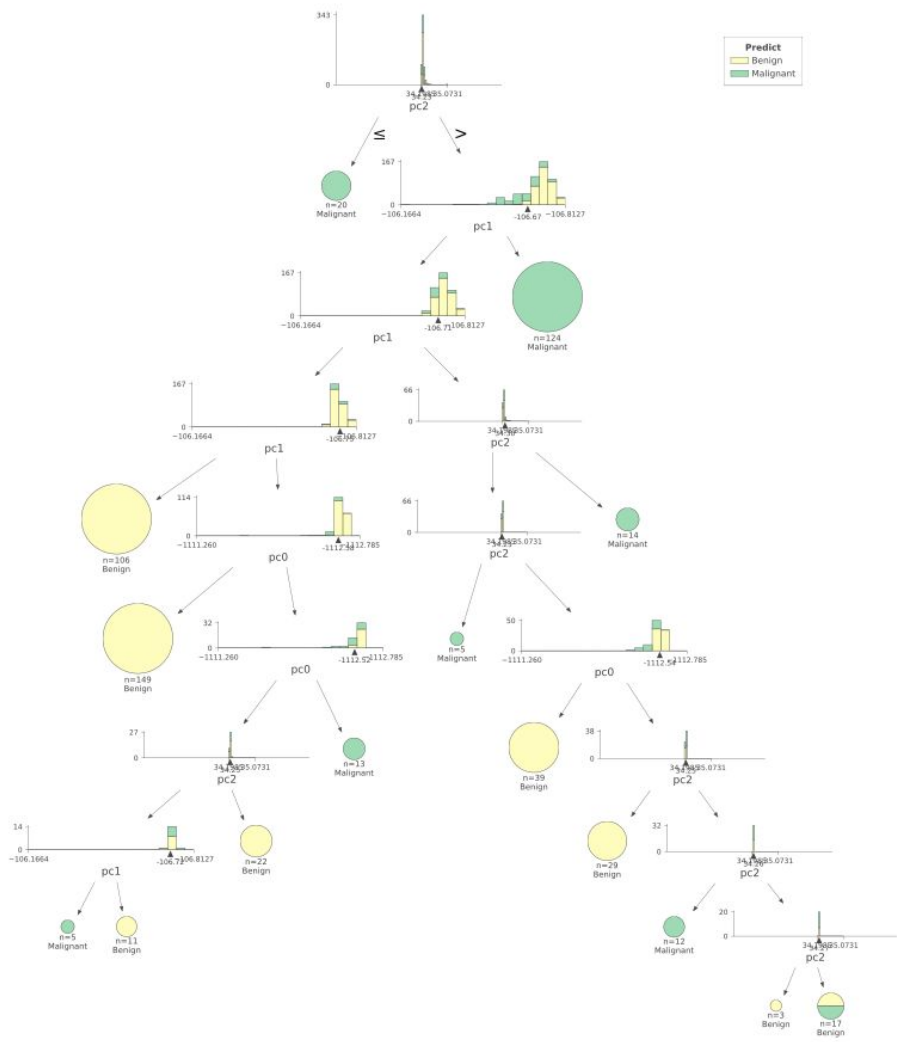


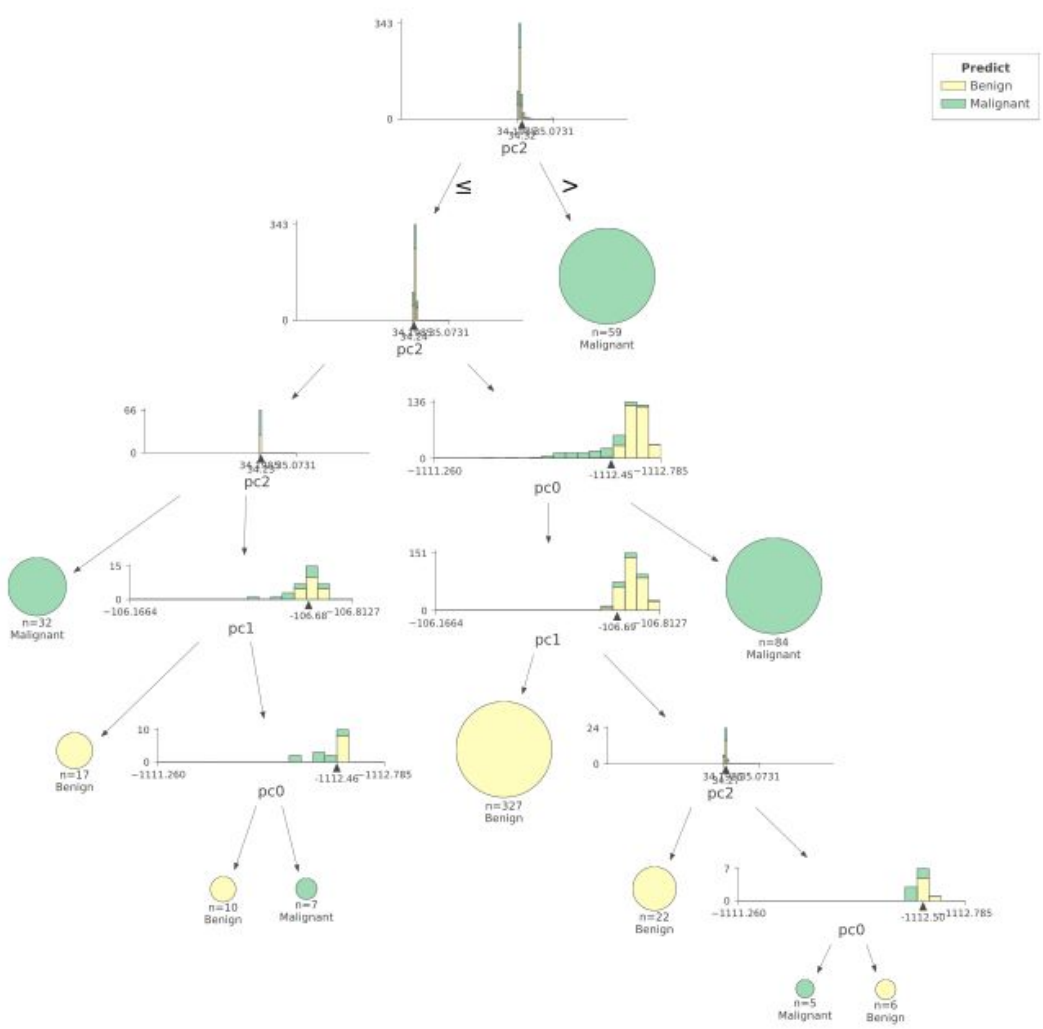

Selected Estimator from Optimal Random Forest [375:10,000]
Predict
Benign
Malignant
pc2
pc1
pc0
<=
>
n=65
Malignant
n=89
Malignant
n=103
Benign
n=26
Benign
n=14
Malignant
n=21
Benign
n=17
Benign
n=6
Malignant
n=84
Benign
n=40
Malignant
n=2
Malignant
n=6
Benign
n=52
Benign
n=10
Malignant
n=2
Malignant
n=3
Benign
n=14
Malignant
n=11
Benign
n=3
Benign
n=1
Malignant

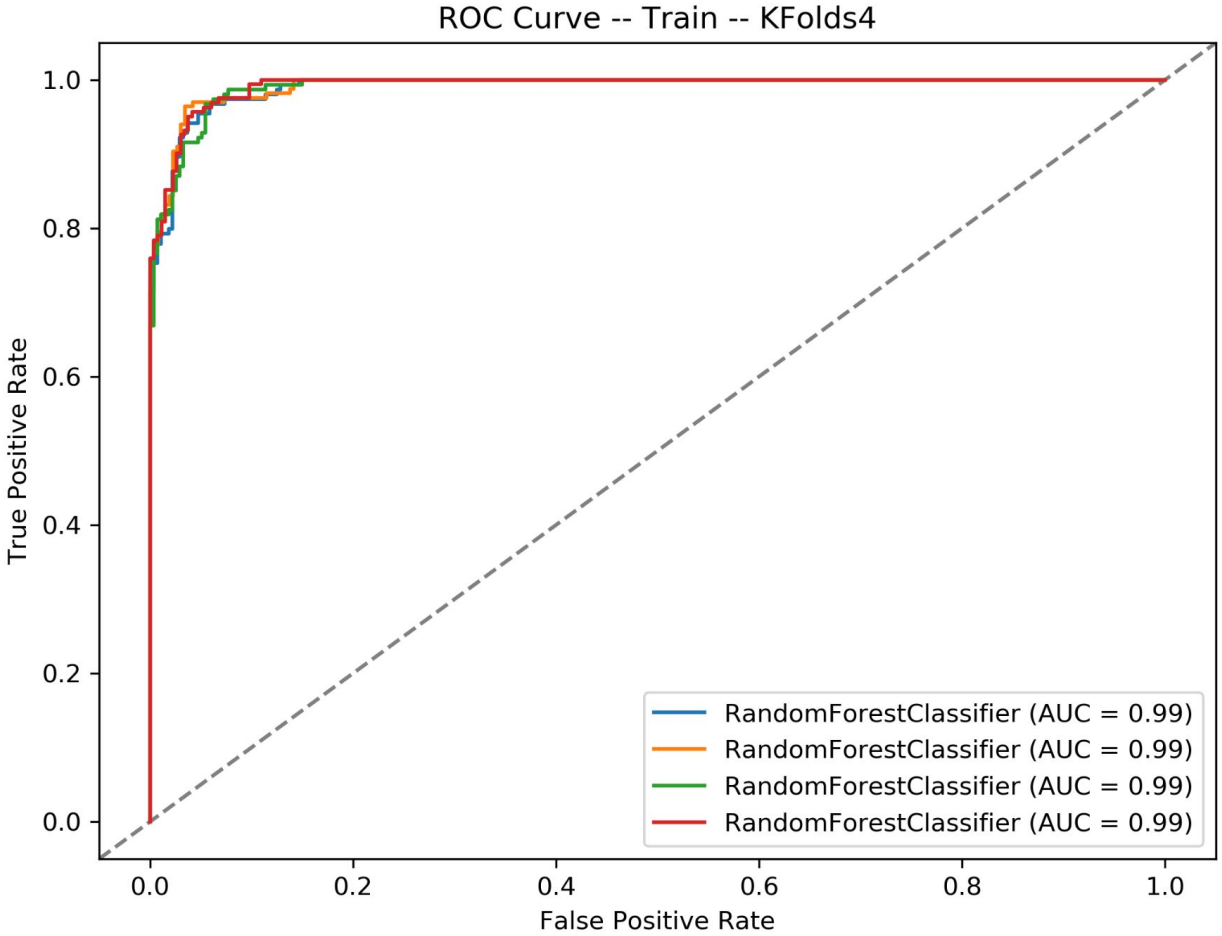
Selected Estimator from Optimal Random Forest [3333:10,000]

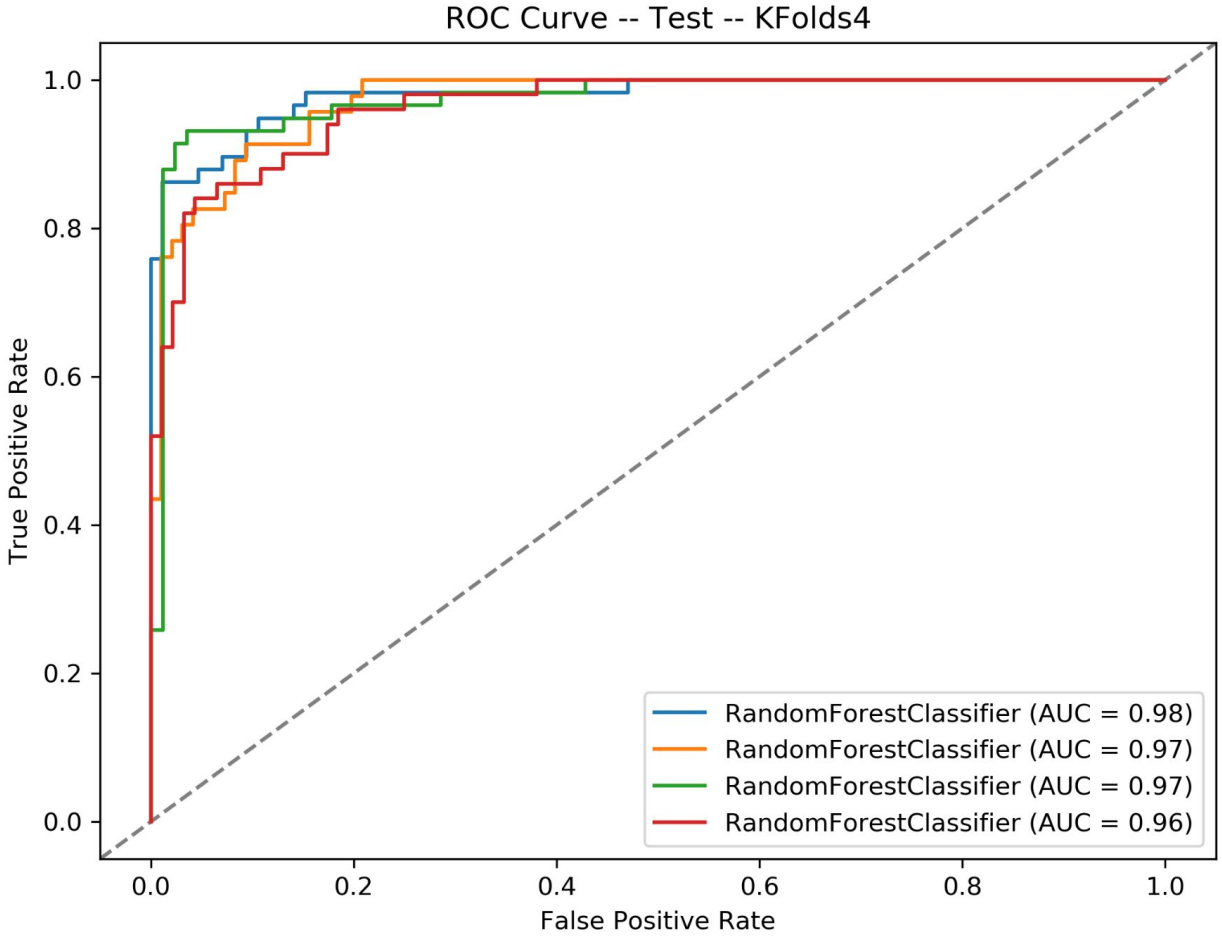


Selected Estimator from Optimal Random Forest [4456:10,000]




Selected Estimator from Optimal Random Forest [7474:10,000]
Predict
Benign
Malignant
pc2
pc0
pc1
n=59
Malignant
n=32
Malignant
n=84
Malignant
n=17
Benign
n=327
Benign
n=22
Benign
n=10
Benign
n=7
Malignant
n=5
Malignant
n=6
Benign


ROC Curve -- Train -- KFolds4
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
RandomForestClassifier (AUC = 0.99)
RandomForestClassifier (AUC = 0.99)
RandomForestClassifier (AUC = 0.99)
RandomForestClassifier (AUC = 0.99)


ROC Curve -- Test -- KFolds4
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
RandomForestClassifier (AUC = 0.98)
RandomForestClassifier (AUC = 0.97)
RandomForestClassifier (AUC = 0.97)
RandomForestClassifier (AUC = 0.96)

Contact

Thank you for reviewing my DSI project for General Assembly

If you have any comments or feedback about this course content, please contact:

John Koenig

Faculty Affiliate, Regis University

Master of Data Science Program

jkoenig0002@regis.edu