STREAMLINE Training Summary Report: 2022-06-15 02:58:59.781548

General Pipeline Settings:

Data Path: /home/ryanurb/idata/datasets/HCC_UCI
Output Path: /home/ryanurb/idata/output
Experiment Name: HCC_PipeTest_Full
Class Label: Class
Instance Label: InstanceID
Ignored Features: None
Specified Categorical Features: None
CV Partitions: 3
Partition Method: S
Match Label: None
Categorical Cutoff: 10
Statistical Significance Cutoff: 0.05
Export Feature Correlations: True
Export Univariate Plots: True
Random Seed: 42
Run From Jupyter Notebook: False
Use Data Scaling: True
Use Data Imputation: True
Use Multivariate Imputation: True
Use Mutual Information: True
Use MultiSURF: True
Use TURF: False
TURF Cutoff: 0.5
MultiSURF Instance Subset: 2000
Max Features to Keep: 2000
Filter Poor Features: True
Top Features to Display: 40
Export Feature Importance Plot: True
Overwrite CV Datasets: False
Primary Metric: balanced_accuracy
Training Subsample for KNN,ANN,SVM,and XGB: 0
Uniform Feature Importance Estimation (Models): True
Hyperparameter Sweep Number of Trials: 50
Hyperparameter Timeout: None
Export Hyperparameter Sweep Plots: True
Export ROC Plot: True
Export PRC Plot: True
Export Metric Boxplots: True
Export Feature Importance Boxplots: True
Metric Weighting Composite FI Plots: balanced_accuracy
Top Model Features To Display: 40

ML Modeling Algorithms:

Naive Bayes: True
Logistic Regression: True
Decision Tree: True
Random Forest: True
Gradient Boosting: True
Extreme Gradient Boosting: True
Light Gradient Boosting: True
Category Gradient Boosting: True
Support Vector Machine: True
Artificial Neural Network: True
K-Nearest Neightbors: True
Genetic Programming: True
eLCS: False
XCS: False
ExSTraCS: True

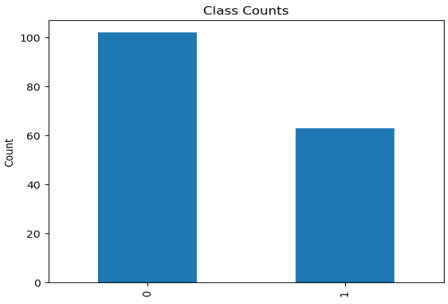
LCS Settings (eLCS,XCS,ExSTraCS):

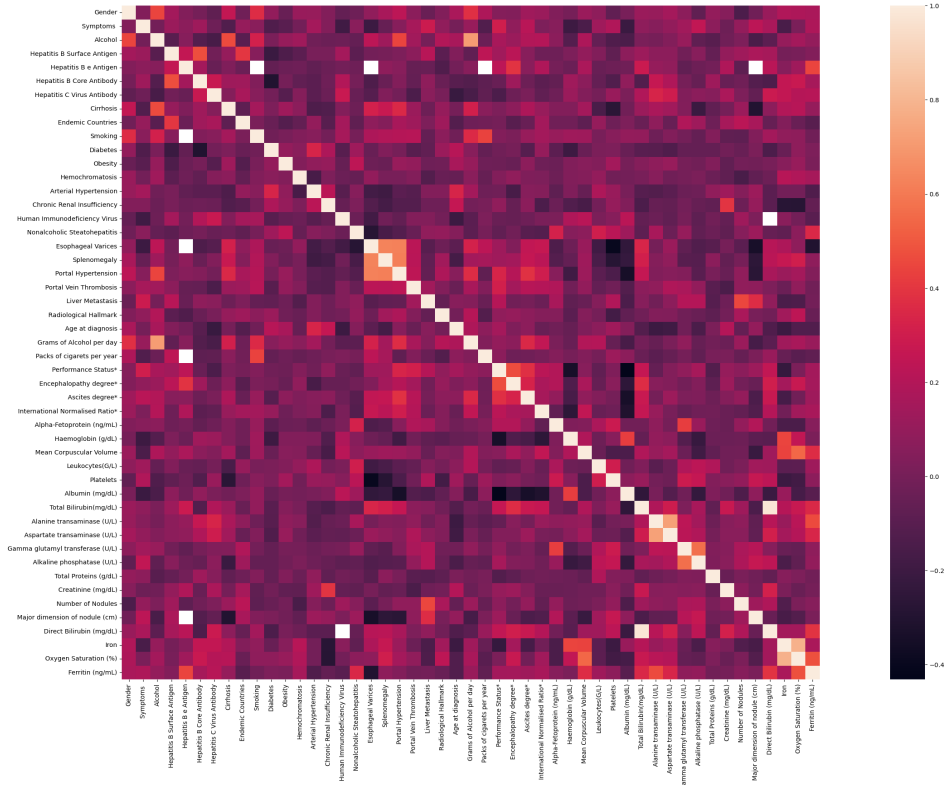
Do LCS Hyperparameter Sweep: False
nu: 1
Training Iterations: 200000
N (Rule Population Size): 2000
LCS Hyperparameter Sweep Timeout: 1200

Target Training Dataset: hcc-data_example

Applied Datasets:
D1 = hcc-data_example_rep

Dataset and Model Prediction Summary: D1 = hcc-data_example_rep

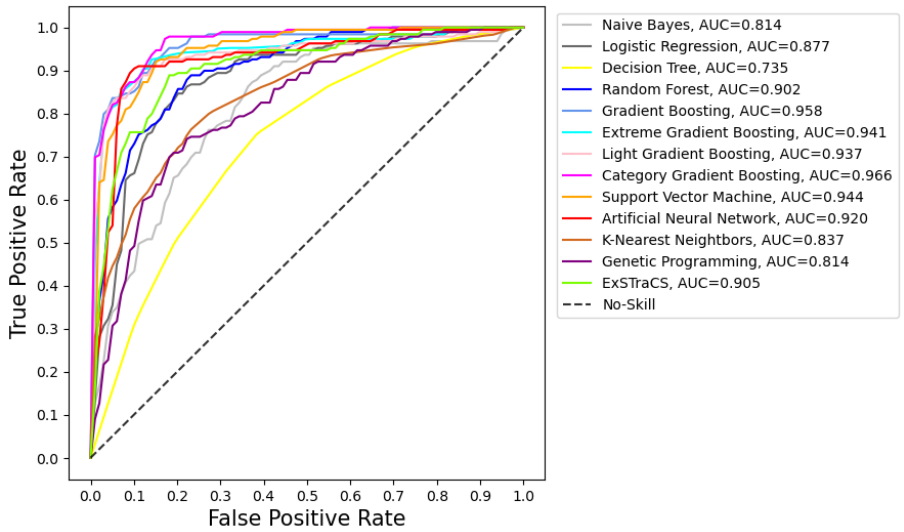


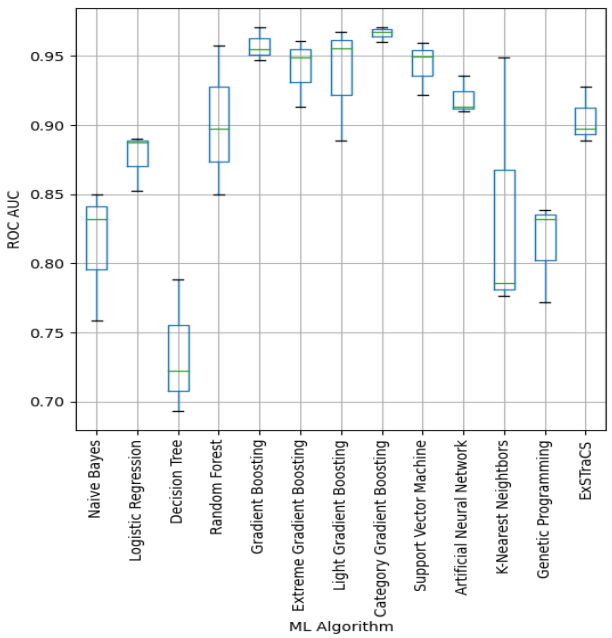


Dataset Counts Summary:
instances: 165.0
features: 49.0
categorical_features: 27.0
quantitative_features: 22.0
missing_values: 826.0
missing_percent: 0.10216

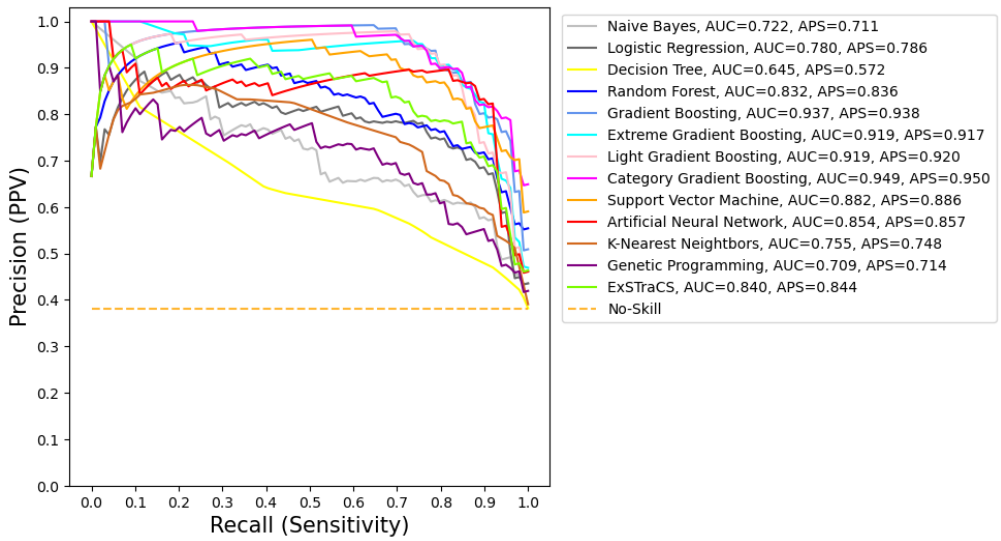
Top ML Algorithm Results (Averaged Over CV Runs):
Best (ROC_AUC): Category Gradient Boosting = 0.966
Best (Balanced Acc.): Artificial Neural Network = 0.901
Best (F1 Score): Artificial Neural Network = 0.876
Best (PRC AUC): Category Gradient Boosting = 0.949
Best (PRC APS): Category Gradient Boosting = 0.950

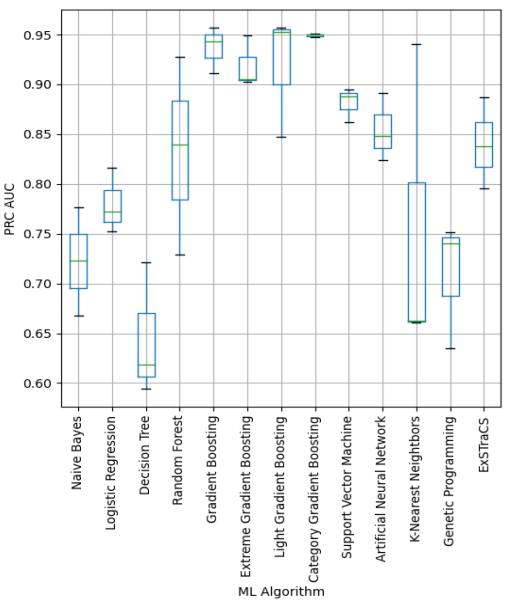
ROC





PRC





Generated with the URBS-Lab STREAMLINE: (https://github.com/UrbsLab/STREAMLINE)

Average Model Prediction Statistics (Rounded to 3 Decimal Points)

D1 = hcc-data_example_rep

ML Algorithm	Balanced Accuracy	Accuracy	F1 Score	Sensitivity (Recall)	Specificity	Precision (PPV)	TP	TN	FP	FN	NPV	LR+	LR-	ROC AUC	PRC AUC	PRC APS
Naive Bayes	0.61	0.622	0.517	0.561	0.66	0.674	35.333	67.333	34.667	27.667	0.704	6.452	0.683	0.814	0.722	0.711
Logistic Regression	0.823	0.824	0.78	0.82	0.827	0.745	51.667	84.333	17.667	11.333	0.882	4.83	0.219	0.877	0.78	0.786
Decision Tree	0.693	0.695	0.632	0.683	0.703	0.594	43.0	71.667	30.333	20.0	0.782	2.451	0.452	0.735	0.645	0.572
Random Forest	0.841	0.84	0.803	0.841	0.84	0.771	53.0	85.667	16.333	10.0	0.895	6.899	0.192	0.902	0.832	0.836
Gradient Boosting	0.887	0.895	0.861	0.852	0.922	0.871	53.667	94.0	8.0	9.333	0.91	11.634	0.161	0.958	0.937	0.938
Extreme Gradient Boosting	0.886	0.899	0.863	0.831	0.941	0.898	52.333	96.0	6.0	10.667	0.9	15.111	0.18	0.941	0.919	0.917
Light Gradient Boosting	0.878	0.889	0.851	0.831	0.925	0.873	52.333	94.333	7.667	10.667	0.899	12.008	0.183	0.937	0.919	0.92
Category Gradient Boosting	0.877	0.893	0.852	0.81	0.944	0.9	51.0	96.333	5.667	12.0	0.889	15.003	0.202	0.966	0.949	0.95
Support Vector Machine	0.87	0.867	0.837	0.884	0.856	0.797	55.667	87.333	14.667	7.333	0.922	7.055	0.137	0.944	0.882	0.886
Artificial Neural Network	0.901	0.903	0.876	0.894	0.908	0.858	56.333	92.667	9.333	6.667	0.933	9.958	0.117	0.92	0.854	0.857
K-Nearest Neightbors	0.623	0.705	0.289	0.275	0.971	0.455	17.333	99.0	3.0	45.667	0.708	3.98	0.734	0.837	0.755	0.748
Genetic Programming	0.749	0.766	0.687	0.677	0.82	0.698	42.667	83.667	18.333	20.333	0.805	3.778	0.394	0.814	0.709	0.714
ExSTraCS	0.797	0.828	0.741	0.667	0.928	0.855	42.0	94.667	7.333	21.0	0.824	9.791	0.355	0.905	0.84	0.844

Median Model Prediction Statistics (Rounded to 3 Decimal Points)

D1 = hcc-data_example_rep

ML Algorithm	Balanced Accuracy	Accuracy	F1 Score	Sensitivity (Recall)	Specificity	Precision (PPV)	TP	TN	FP	FN	NPV	LR+	LR-	ROC AUC	PRC AUC	PRC APS
Naive Bayes	0.625	0.709	0.547	0.492	0.892	0.738	31.0	91.0	11.0	32.0	0.688	4.563	0.736	0.832	0.723	0.7
Logistic Regression	0.84	0.836	0.8	0.857	0.824	0.75	54.0	84.0	18.0	9.0	0.903	4.857	0.173	0.888	0.772	0.782
Decision Tree	0.676	0.679	0.622	0.667	0.686	0.568	42.0	70.0	32.0	21.0	0.788	2.125	0.437	0.722	0.618	0.555
Random Forest	0.843	0.836	0.803	0.857	0.814	0.743	54.0	83.0	19.0	9.0	0.912	4.687	0.156	0.898	0.84	0.841
Gradient Boosting	0.886	0.897	0.862	0.841	0.931	0.883	53.0	95.0	7.0	10.0	0.905	12.259	0.17	0.955	0.943	0.943
Extreme Gradient Boosting	0.893	0.903	0.871	0.825	0.931	0.885	52.0	95.0	7.0	11.0	0.899	12.49	0.182	0.949	0.905	0.903
Light Gradient Boosting	0.88	0.885	0.85	0.841	0.922	0.862	53.0	94.0	8.0	10.0	0.907	10.119	0.167	0.956	0.953	0.953
Category Gradient Boosting	0.888	0.903	0.867	0.825	0.951	0.912	52.0	97.0	5.0	11.0	0.898	16.838	0.184	0.967	0.95	0.95
Support Vector Machine	0.887	0.891	0.859	0.873	0.892	0.838	55.0	91.0	11.0	8.0	0.92	8.39	0.141	0.949	0.888	0.89
Artificial Neural Network	0.892	0.897	0.866	0.889	0.912	0.859	56.0	93.0	9.0	7.0	0.929	9.894	0.125	0.913	0.848	0.851
K-Nearest Neightbors	0.503	0.618	0.031	0.016	0.99	0.5	1.0	101.0	1.0	62.0	0.62	1.619	0.994	0.786	0.663	0.657
Genetic Programming	0.761	0.776	0.704	0.698	0.824	0.71	44.0	84.0	18.0	19.0	0.816	3.958	0.366	0.832	0.74	0.744
ExSTraCS	0.769	0.812	0.705	0.587	0.941	0.854	37.0	96.0	6.0	26.0	0.789	9.444	0.434	0.898	0.838	0.84