STREAMLINE Training Summary Report: 2022-05-20 13:37:40.022164

General Pipeline Settings:

Data Path: /home/ryanurb/idata/datasets/HCC UCI

Output Path: /home/ryanurb/idata/output Experiment Name: HCC_PipeTest_FullRep

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

 $\label{thm:continuous} Training \ Subsample \ for \ KNN, ANN, SVM, and \ XGB: 0 \\ Uniform \ Feature \ Importance \ Estimation \ (Models): \ True$

 $Hyperparameter\ Sweep\ Number\ of\ Trials:\ 50$

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 Naive Bayes, AUC=0.814 1.0 Logistic Regression, AUC=0.877 0.95 Decision Tree, AUC=0.735 0.9 Random Forest, AUC=0.902 0.8 Gradient Boosting, AUC=0.958 0.90 Extreme Gradient Boosting, AUC=0.941 Light Gradient Boosting, AUC=0.937 True Positive Rate 0.7 Category Gradient Boosting, AUC=0.966 0.85 Support Vector Machine, AUC=0.944 Artificial Neural Network, AUC=0.920 K-Nearest Neightbors, AUC=0.837 Genetic Programming, AUC=0.814 0.80 ExSTraCS, AUC=0.905 0.4 --- No-Skill 0.3 0.75 0.2 Light Gradient Boosting Light Gradient Boosting Grategory Gradient Boosting Decision Tree Naive Bayes Gradient Boosting Extreme Gradient Boosting Support Vector Machine Logistic Regression Artificial Neural Network K-Nearest Neightbors Genetic Programming 0.3 0.4 0.5 0.6 0.7 0.8 0.9 False Positive Rate PRC Naive Bayes, AUC=0.722, APS=0.711 Logistic Regression, AUC=0.780, APS=0.786 0.9 Decision Tree, AUC=0.645, APS=0.572 Random Forest, AUC=0.832, APS=0.836 Gradient Boosting, AUC=0.937, APS=0.938 0.85 Extreme Gradient Boosting, AUC=0.919, APS=0.917 Light Gradient Boosting, AUC=0.919, APS=0.920 Precision (PPV) 0.80 Category Gradient Boosting, AUC=0.949, APS=0.950 Support Vector Machine, AUC=0.882, APS=0.886 O.75 Artificial Neural Network, AUC=0.854, APS=0.857 K-Nearest Neightbors, AUC=0.755, APS=0.748 Genetic Programming, AUC=0.709, APS=0.714 0.70 ExSTraCS, AUC=0.840, APS=0.844 No-Skill 0.65 0.3 0.2 ogistic Regression Decision Tree **Extreme Gradient Boosting** Light Gradient Boosting Gategory Gradient Boosting K-Nearest Neightbors Artificial Neural Network 0.1 0.0 0.4 0.5 0.6 0.0 0.1 0.2 0.3 0.7 0.8 0.9 Recall (Sensitivity)

Average Model Prediction Statistics (Rounded to 3 Decimal Points)

D1 = hcc-data_example_rep																
ML Algorithm		Accuracy	F1	Sensitivity	Specificity		TP	TN	FP	FN	NPV	LR+	LR-	ROC	PRC	PRC
	Accuracy		Score	(Recall)		(PPV)								AUC	AUC	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		0.836
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		0.917
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		0.92
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		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		3.778	0.394	0.814	0.709	0.714
EvSTroCS	0.707	0.838	0.741	0.667	0.028	0.855	42.0	94 667	7 333	21.0	0.824	0.701	0.355	0.005	0.84	0.844