

STREAMLINE Training Summary Report: 2022-05-20 13:23:22.764894

General Pipeline Settings:

ML Modeling Algorithms:

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

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 Neighbors: 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

Datasets:

D1 = hcc-data_example
D2 = hcc-data_example_no_covariates

Univariate Analysis of Each Dataset (Top 10 Features for Each)

D1 = hcc-data_example

Feature: P-Value

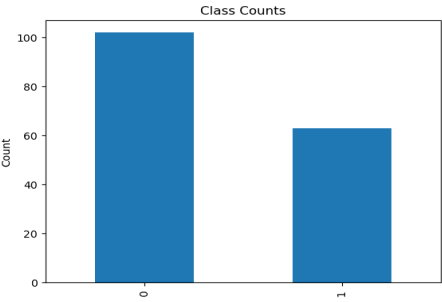
Performance Status*: 3.2548676278782114e-05
Symptoms : 0.0006092985105592
Liver Metastasis: 0.0029935882248699
Ascites degree*: 0.0038134308539161
Portal Vein Thrombosis: 0.0117430411554256
Age at diagnosis: 0.035683237512087
Encephalopathy degree*: 0.0367398682254197
Diabetes: 0.2071781828192029
Hepatitis C Virus Antibody: 0.2152844001545551
Endemic Countries: 0.3741454960813042

D2 = hcc-data_example_no_covariates

Feature: P-Value

Performance Status*: 3.2548676278782114e-05
Symptoms : 0.0006092985105592
Liver Metastasis: 0.0029935882248699
Ascites degree*: 0.0038134308539161
Portal Vein Thrombosis: 0.0117430411554256
Encephalopathy degree*: 0.0367398682254197
Diabetes: 0.2071781828192029
Hepatitis C Virus Antibody: 0.2152844001545551
Endemic Countries: 0.3741454960813042
Chronic Renal Insufficiency: 0.3855402814015594

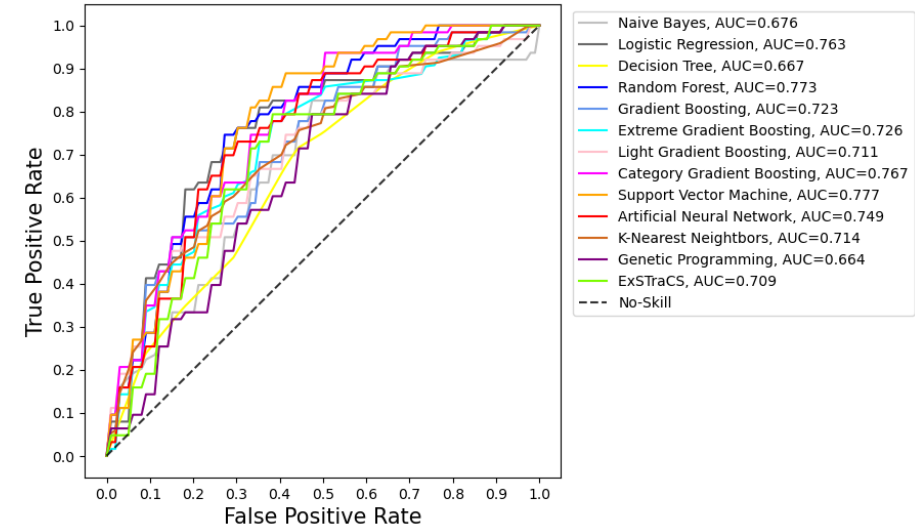
Dataset and Model Prediction Summary: D1 = hcc-data_example



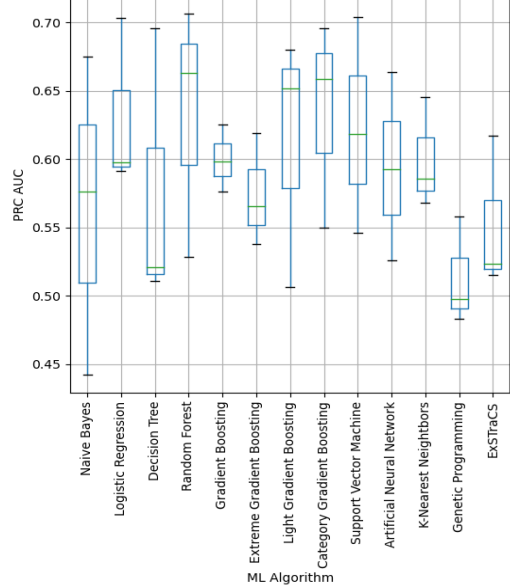
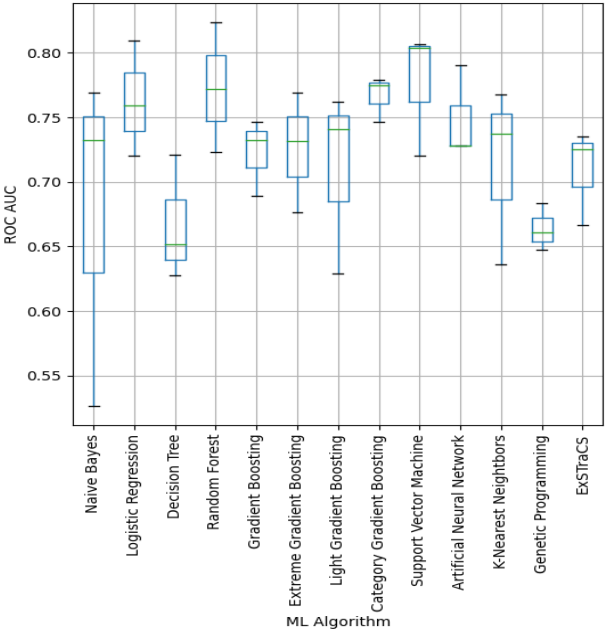
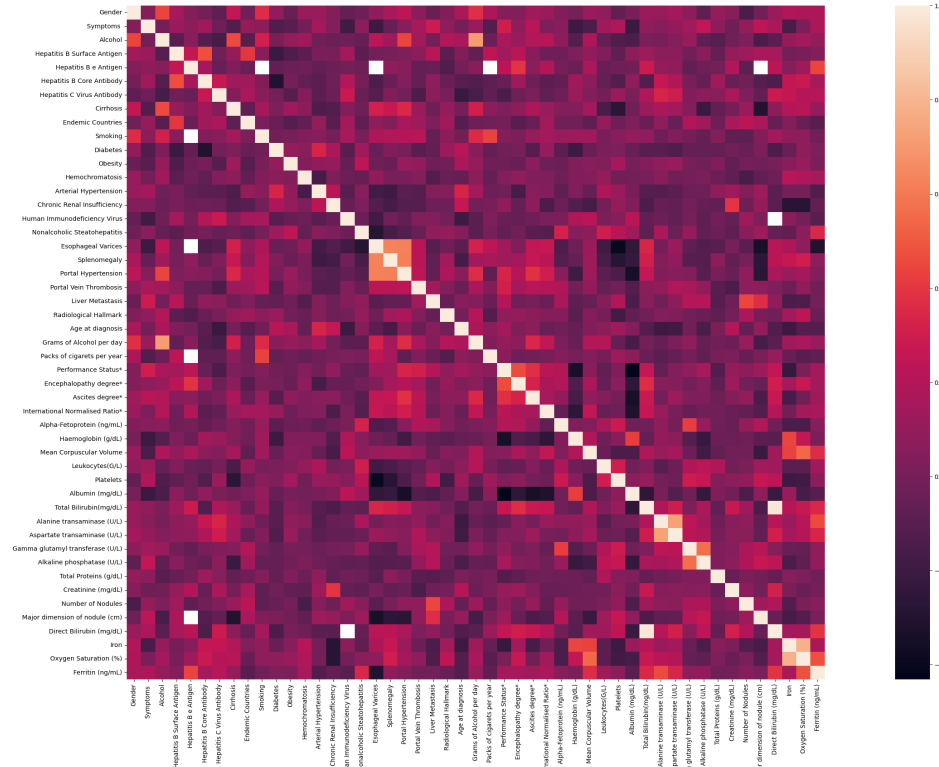
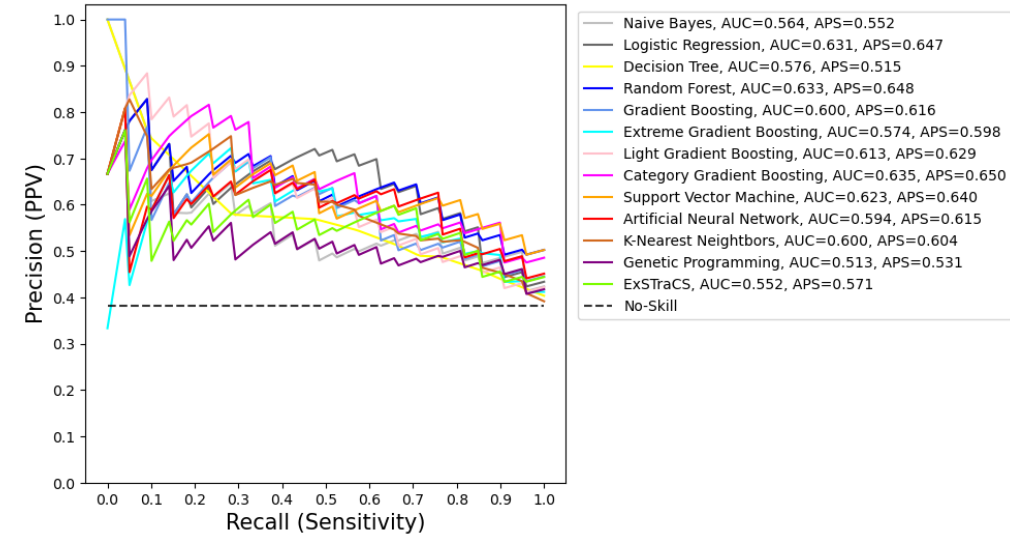
Dataset Counts Summary:
instances: 165.0
features: 49.0
categorical_features: 27.0
quantitative_features: 22.0
missing_values: 826.0
missing_percent: 0.10216

Top ML Algorithm Results (Averaged Over CV Runs):
Best (ROC_AUC): Support Vector Machine = 0.777
Best (Balanced Acc.): Random Forest = 0.724
Best (F1 Score): Random Forest = 0.662
Best (PRC AUC): Category Gradient Boosting = 0.635
Best (PRC APS): Category Gradient Boosting = 0.650

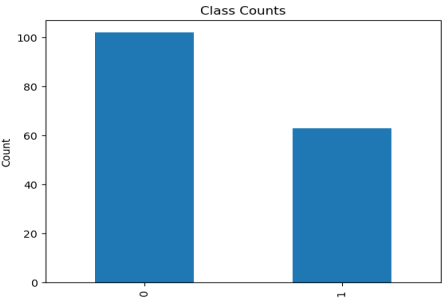
ROC



PRC



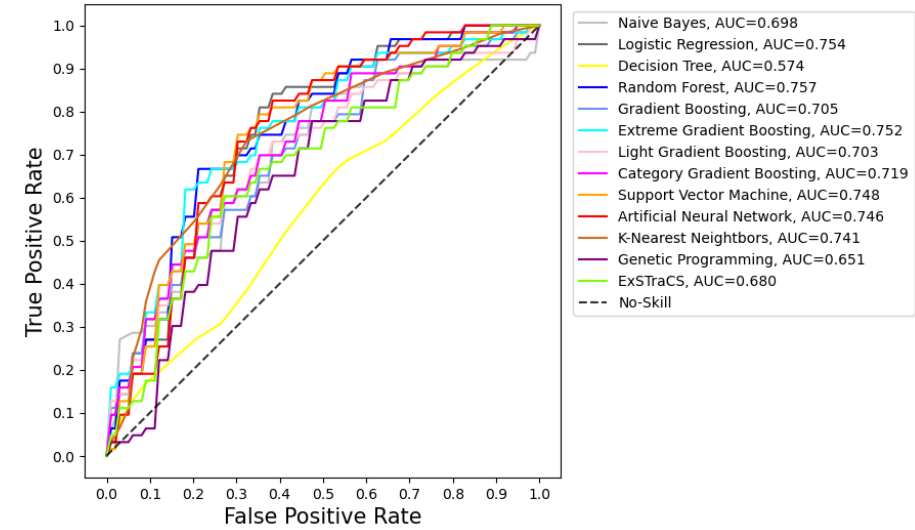
Dataset and Model Prediction Summary: D2 = hcc-data_example_no_covariates



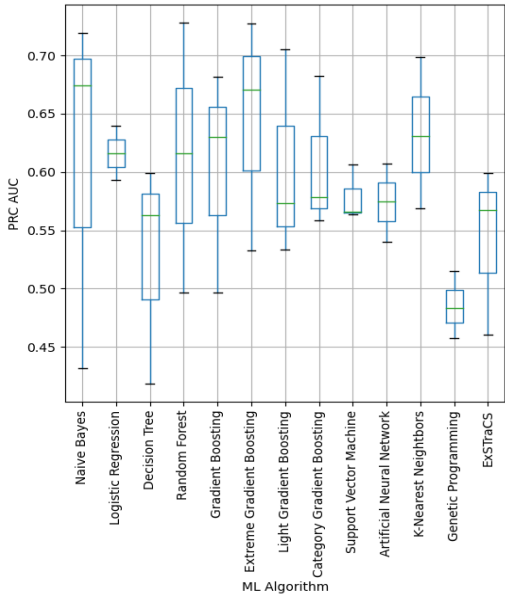
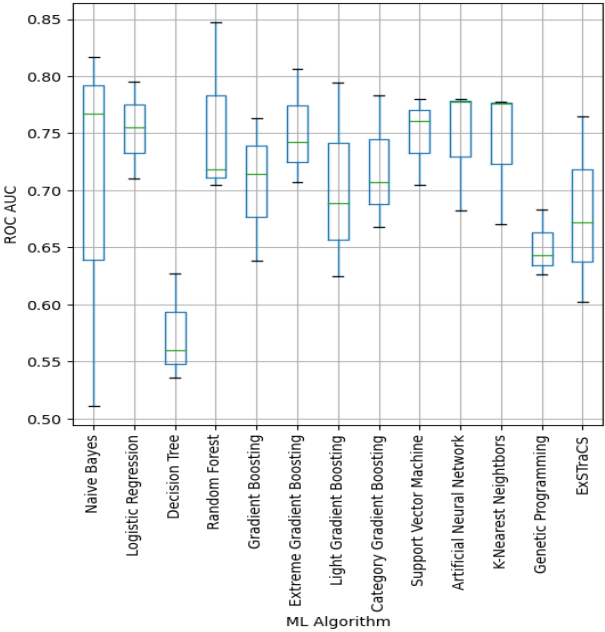
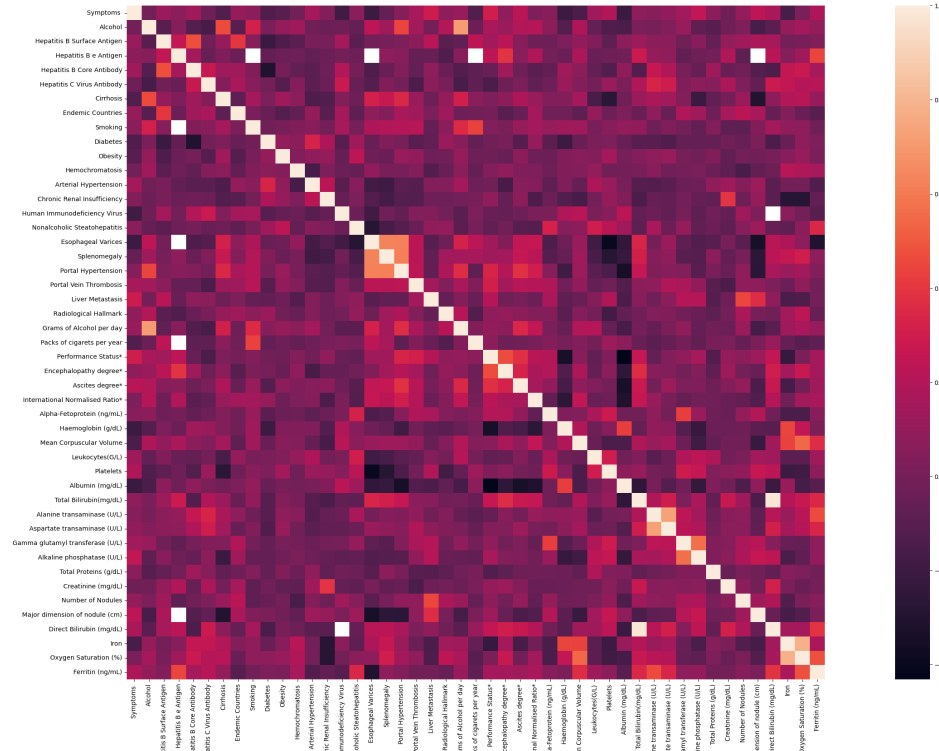
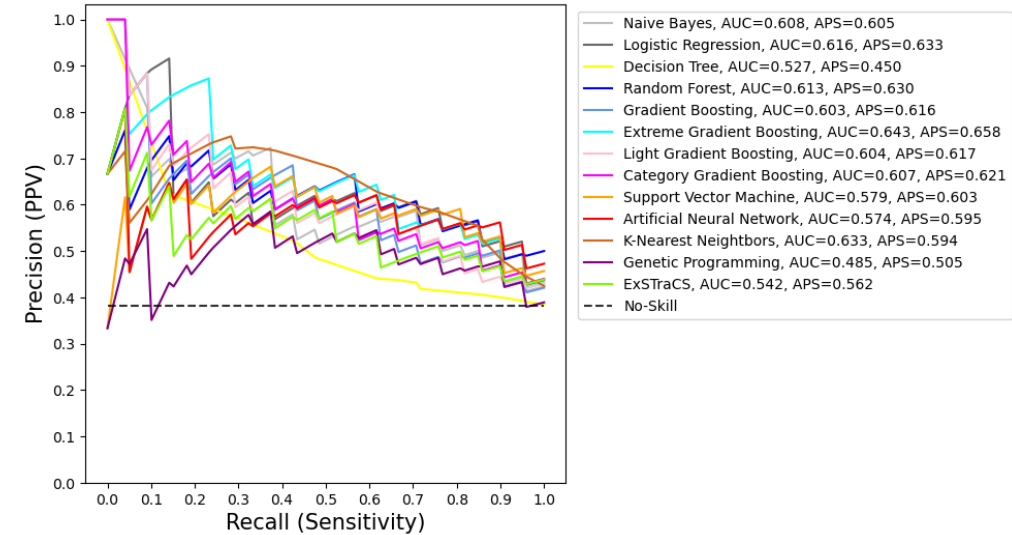
Dataset Counts Summary:
instances: 165.0
features: 47.0
categorical_features: 26.0
quantitative_features: 21.0
missing_values: 826.0
missing_percent: 0.10651

Top ML Algorithm Results (Averaged Over CV Runs):
Best (ROC_AUC): Random Forest = 0.757
Best (Balanced Acc.): Random Forest = 0.691
Best (F1 Score): Random Forest = 0.630
Best (PRC AUC): Extreme Gradient Boosting = 0.643
Best (PRC APS): Extreme Gradient Boosting = 0.658

ROC



PRC



Average Model Prediction Statistics (Rounded to 3 Decimal Points)

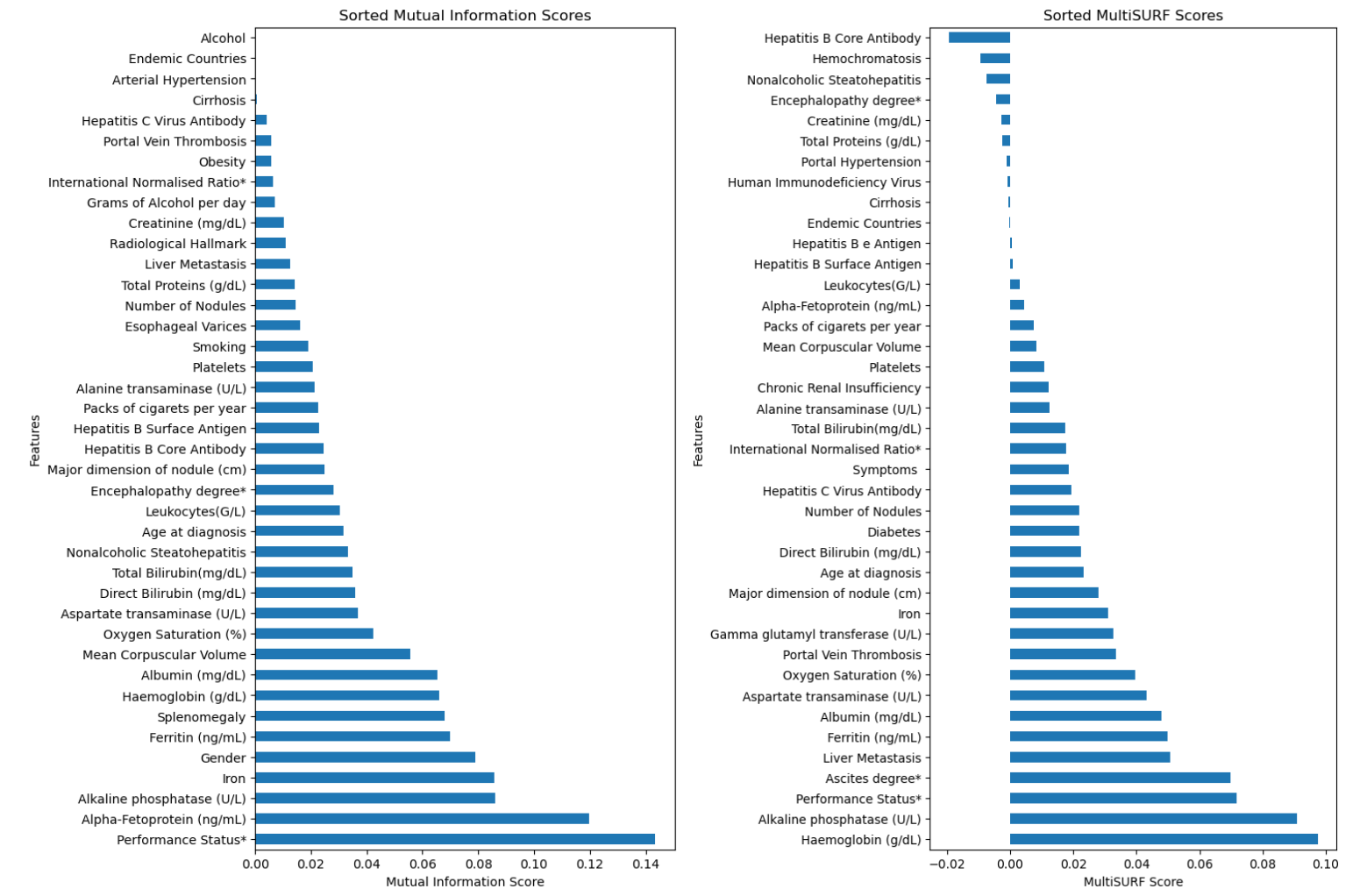
D1 = hcc-data_example

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	0.545	0.558	0.445	0.492	0.598	0.579	10.333	20.333	13.667	10.667	0.522	3.67	2.66	0.676	0.564	0.552
Logistic Regression	0.719	0.727	0.656	0.683	0.755	0.632	14.333	25.667	8.333	6.667	0.795	2.788	0.42	0.763	0.631	0.647
Decision Tree	0.651	0.655	0.586	0.635	0.667	0.556	13.333	22.667	11.333	7.667	0.746	2.159	0.551	0.667	0.576	0.515
Random Forest	0.724	0.733	0.662	0.683	0.765	0.649	14.333	26.0	8.0	6.667	0.797	3.112	0.413	0.773	0.633	0.648
Gradient Boosting	0.66	0.685	0.575	0.556	0.765	0.598	11.667	26.0	8.0	9.333	0.735	2.557	0.589	0.723	0.6	0.616
Extreme Gradient Boosting	0.658	0.697	0.553	0.492	0.824	0.636	10.333	28.0	6.0	10.667	0.725	2.968	0.618	0.726	0.574	0.598
Light Gradient Boosting	0.633	0.667	0.528	0.492	0.775	0.578	10.333	26.333	7.667	10.667	0.713	2.375	0.658	0.711	0.613	0.629
Category Gradient Boosting	0.655	0.697	0.546	0.476	0.833	0.639	10.0	28.333	5.667	11.0	0.72	2.961	0.631	0.767	0.635	0.65
Support Vector Machine	0.718	0.715	0.659	0.73	0.706	0.606	15.333	24.0	10.0	5.667	0.814	2.496	0.377	0.777	0.623	0.64
Artificial Neural Network	0.714	0.721	0.651	0.683	0.745	0.622	14.333	25.333	8.667	6.667	0.793	2.698	0.428	0.749	0.594	0.615
K-Nearest Neighbors	0.527	0.618	0.158	0.143	0.912	0.176	3.0	31.0	3.0	18.0	0.638	0.607	0.926	0.714	0.6	0.604
Genetic Programming	0.607	0.63	0.512	0.508	0.706	0.517	10.667	24.0	10.0	10.333	0.699	1.744	0.699	0.664	0.513	0.531
ExSTraCS	0.624	0.667	0.484	0.444	0.804	0.57	9.333	27.333	6.667	11.667	0.711	2.177	0.679	0.709	0.552	0.571

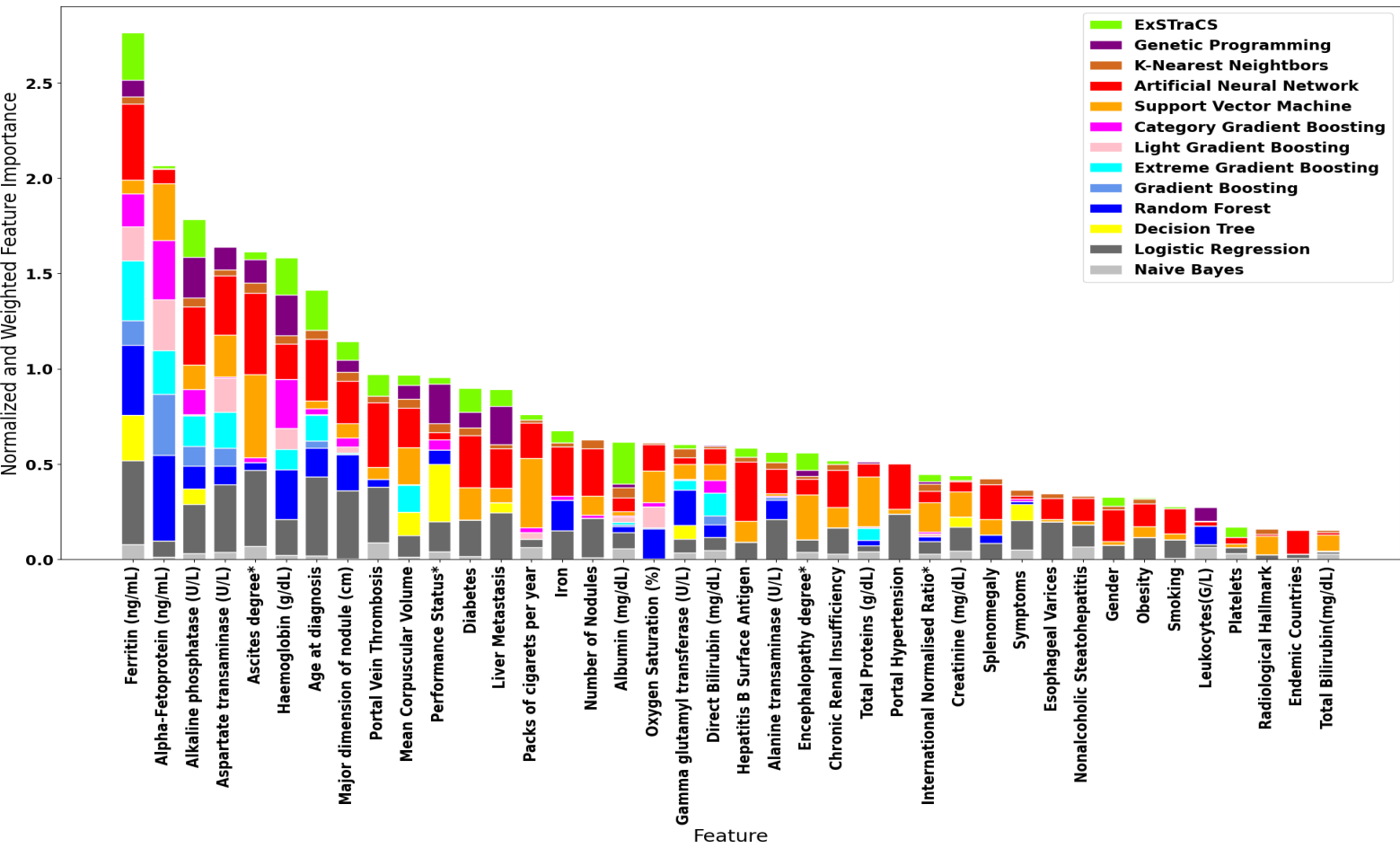
D2 = hcc-data_example_no_covariates

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	0.558	0.57	0.463	0.508	0.608	0.599	10.667	20.667	13.333	10.333	0.53	3.876	2.632	0.698	0.608	0.605
Logistic Regression	0.68	0.691	0.611	0.635	0.725	0.59	13.333	24.667	9.333	7.667	0.762	2.362	0.506	0.754	0.616	0.633
Decision Tree	0.556	0.564	0.477	0.524	0.588	0.442	11.0	20.0	14.0	10.0	0.667	1.289	0.809	0.574	0.527	0.45
Random Forest	0.691	0.697	0.63	0.667	0.716	0.599	14.0	24.333	9.667	7.0	0.774	2.644	0.479	0.757	0.613	0.63
Gradient Boosting	0.637	0.661	0.547	0.54	0.735	0.556	11.333	25.0	9.0	9.667	0.722	2.039	0.626	0.705	0.603	0.616
Extreme Gradient Boosting	0.669	0.703	0.573	0.524	0.814	0.635	11.0	27.667	6.333	10.0	0.735	2.814	0.585	0.752	0.643	0.658
Light Gradient Boosting	0.654	0.685	0.558	0.524	0.784	0.599	11.0	26.667	7.333	10.0	0.728	2.523	0.61	0.703	0.604	0.617
Category Gradient Boosting	0.633	0.655	0.544	0.54	0.725	0.55	11.333	24.667	9.333	9.667	0.718	1.99	0.635	0.719	0.607	0.621
Support Vector Machine	0.635	0.673	0.52	0.476	0.794	0.607	10.0	27.0	7.0	11.0	0.713	2.638	0.654	0.748	0.579	0.603
Artificial Neural Network	0.678	0.685	0.614	0.651	0.706	0.586	13.667	24.0	10.0	7.333	0.764	2.381	0.5	0.746	0.574	0.595
K-Nearest Neighbors	0.552	0.648	0.182	0.143	0.961	0.25	3.0	32.667	1.333	18.0	0.65	1.619	0.886	0.741	0.633	0.594
Genetic Programming	0.623	0.642	0.536	0.54	0.706	0.533	11.333	24.0	10.0	9.667	0.712	1.889	0.656	0.651	0.485	0.505
ExSTraCS	0.597	0.648	0.449	0.381	0.814	0.556	8.0	27.667	6.333	13.0	0.682	2.035	0.759	0.68	0.542	0.562

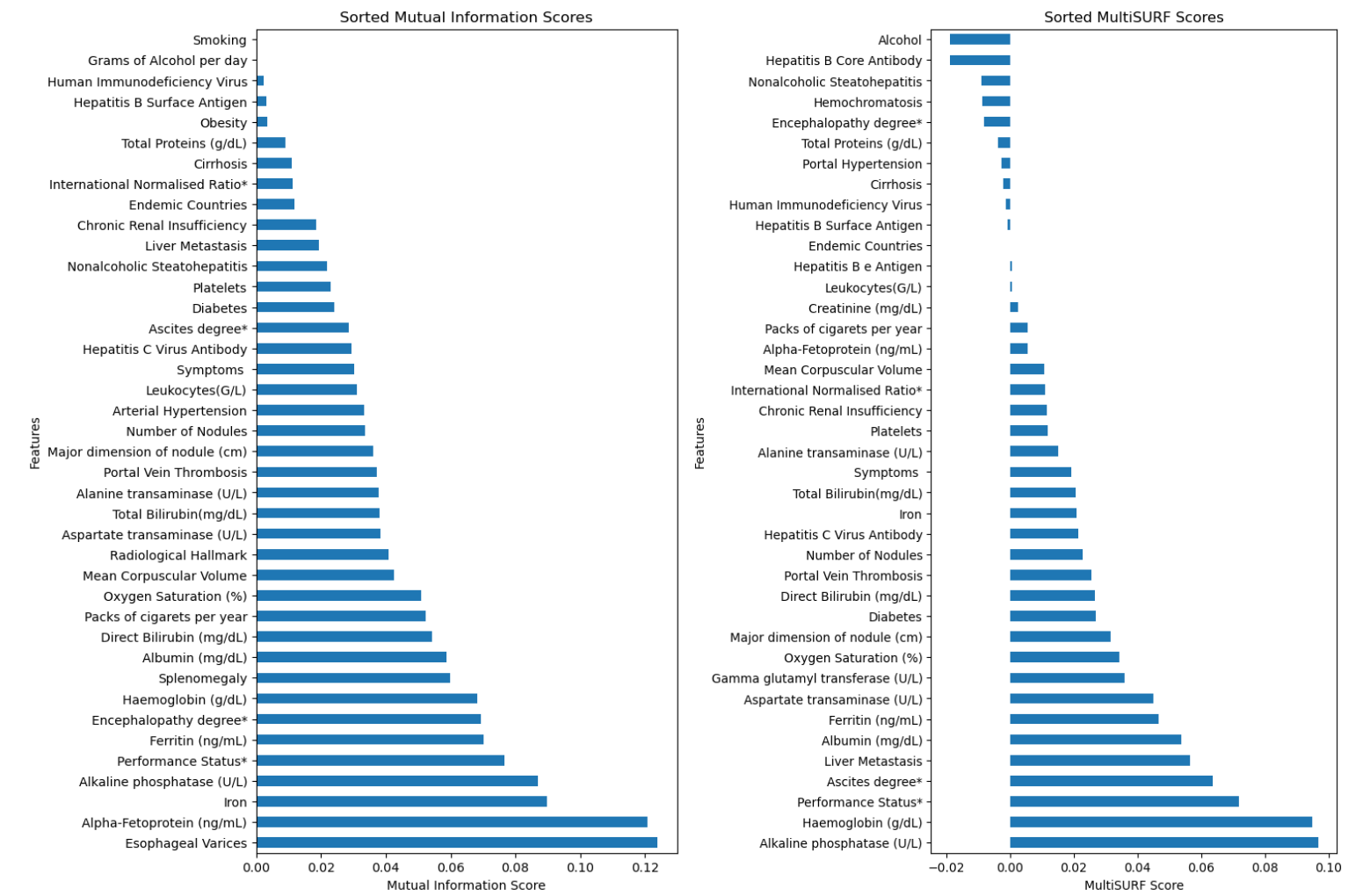
Feature Importance Summary: D1 = hcc-data_example



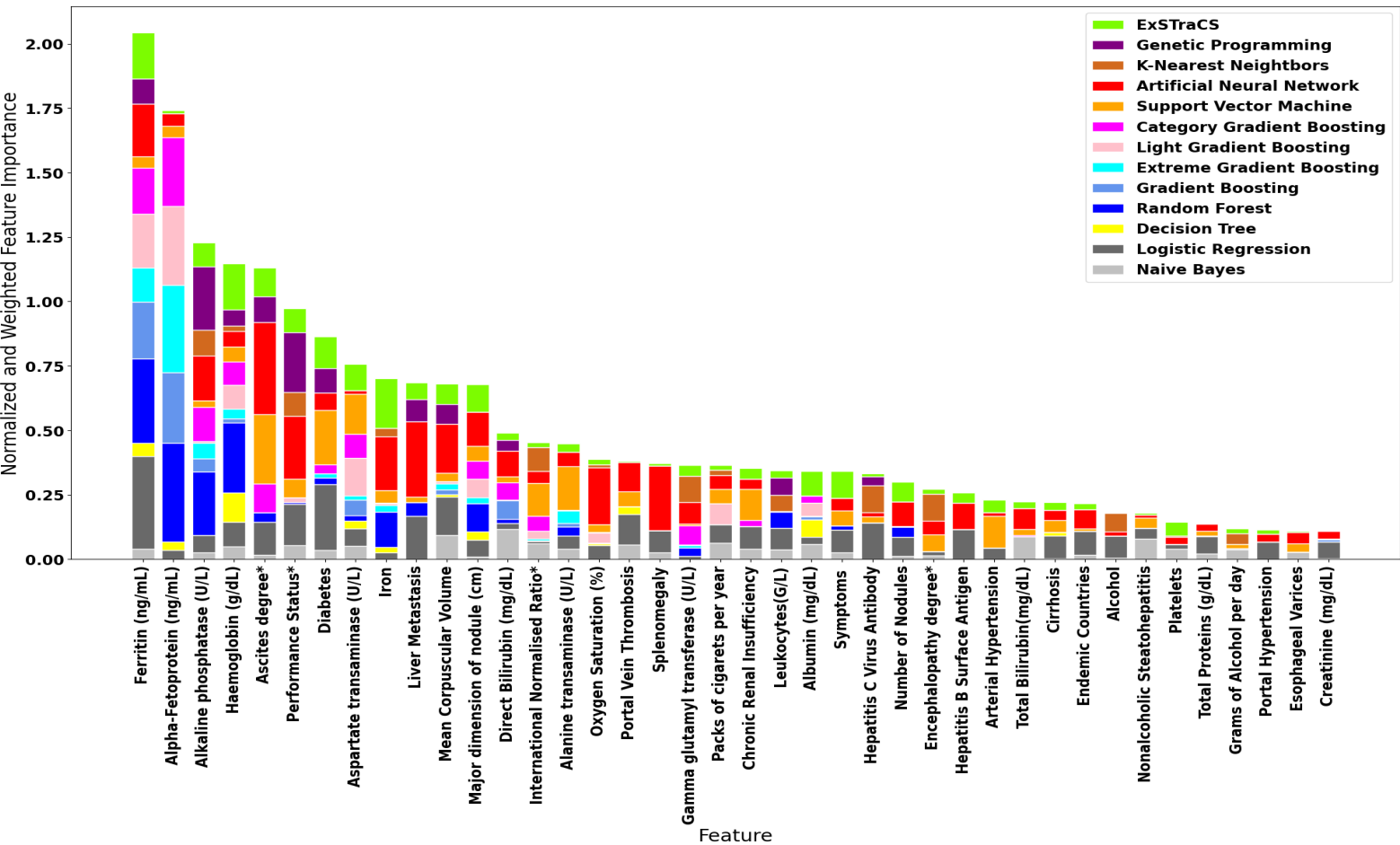
Composite Feature Importance Plot (Normalized and Performance Weighted)



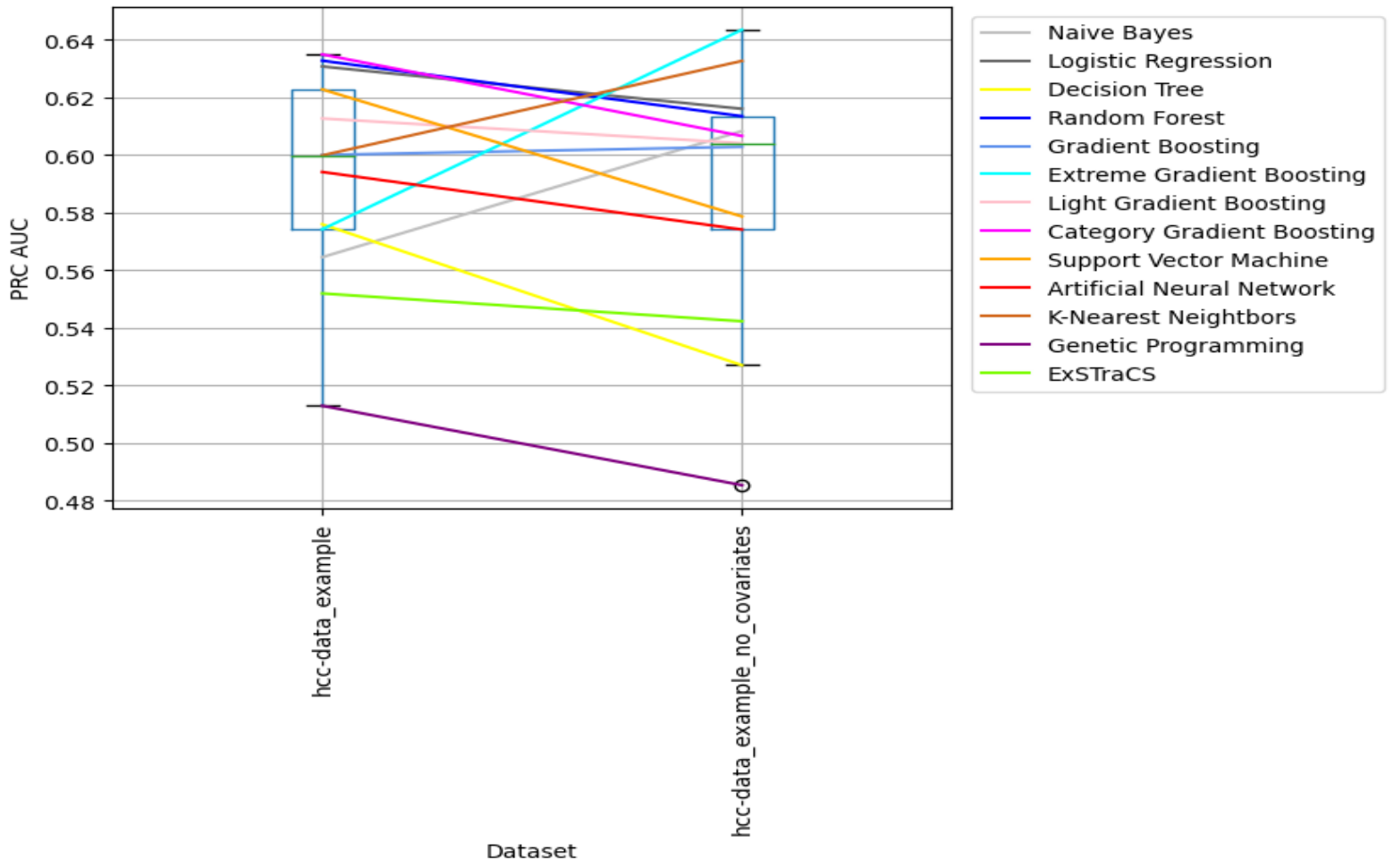
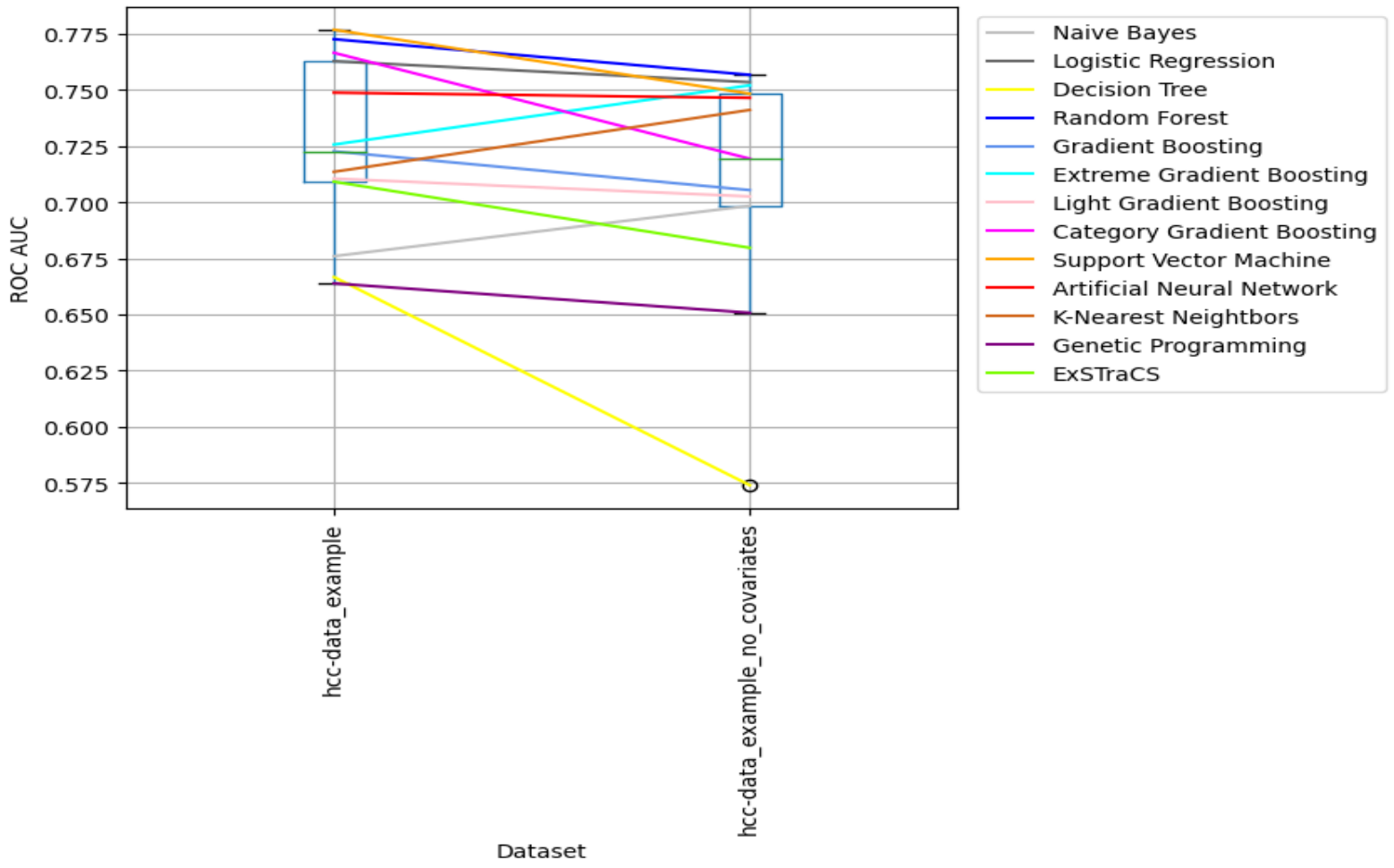
Feature Importance Summary: D2 = hcc-data_example_no_covariates



Composite Feature Importance Plot (Normalized and Performance Weighted)



Compare ML Performance Across Datasets



Using Best Performing Algorithms (Kruskall Wallis Compare Datasets)

Datasets:
D1 = hcc-data_example
D2 = hcc-data_example_no_covariates

index	P-Value	Best Alg D1	Mean D1	Best Alg D2	Mean D2
Balanced Accuracy	0.5127	Random Forest	0.7236	Random Forest	0.6912
Accuracy	0.1046	Random Forest	0.7333	Extreme Gradient Boosting	0.703
F1 Score	0.5066	Random Forest	0.6618	Random Forest	0.6299
Sensitivity (Recall)	0.5002	Support Vector Machine	0.7302	Random Forest	0.6667
Specificity	0.8222	K-Nearest Neighbors	0.9118	K-Nearest Neighbors	0.9608
Precision (PPV)	0.6579	Random Forest	0.6491	Extreme Gradient Boosting	0.6345
TP	0.5002	Support Vector Machine	15.3333	Random Forest	14.0
TN	0.8222	K-Nearest Neighbors	31.0	K-Nearest Neighbors	32.6667
FP	0.5127	Naive Bayes	13.6667	Decision Tree	14.0
FN	1.0	K-Nearest Neighbors	18.0	K-Nearest Neighbors	18.0
NPV	0.5127	Support Vector Machine	0.8141	Random Forest	0.774
LR+	0.8222	Naive Bayes	3.6703	Naive Bayes	3.8759
LR-	0.8222	Naive Bayes	2.6603	Naive Bayes	2.6324
ROC AUC	0.5127	Support Vector Machine	0.7768	Random Forest	0.7568
PRC AUC	0.8273	Category Gradient Boosting	0.6349	Extreme Gradient Boosting	0.6435
PRC APS	0.8273	Category Gradient Boosting	0.65	Extreme Gradient Boosting	0.6581

Pipeline Runtime Summary

hcc-data_example		hcc-data_example_no_covariates	
Pipeline Component	Time (sec)	Pipeline Component	Time (sec)
Exploratory Analysis	2.61	Exploratory Analysis	2.49
Preprocessing	0.38	Preprocessing	0.39
Mutual Information	0.21	Mutual Information	0.2
MultiSURF	0.9	MultiSURF	0.86
Feature Selection	1.03	Feature Selection	1.01
Naive Bayes	0.77	Naive Bayes	0.69
Logistic Regression	12.61	Logistic Regression	7.98
Decision Tree	11.82	Decision Tree	7.39
Random Forest	375.34	Random Forest	382.56
Gradient Boosting	97.98	Gradient Boosting	83.67
Extreme Gradient Boosting	370.12	Extreme Gradient Boosting	415.45
Light Gradient Boosting	26.81	Light Gradient Boosting	28.79
Category Gradient Boosting	7784.51	Category Gradient Boosting	4108.43
Support Vector Machine	11.36	Support Vector Machine	8.18
Artificial Neural Network	62.5	Artificial Neural Network	49.46
K-Nearest Neighbors	31.07	K-Nearest Neighbors	22.25
Genetic Programming	42681.75	Genetic Programming	48841.55
ExSTraCS	2413.38	ExSTraCS	2384.28
Stats Summary	23.58	Stats Summary	23.46