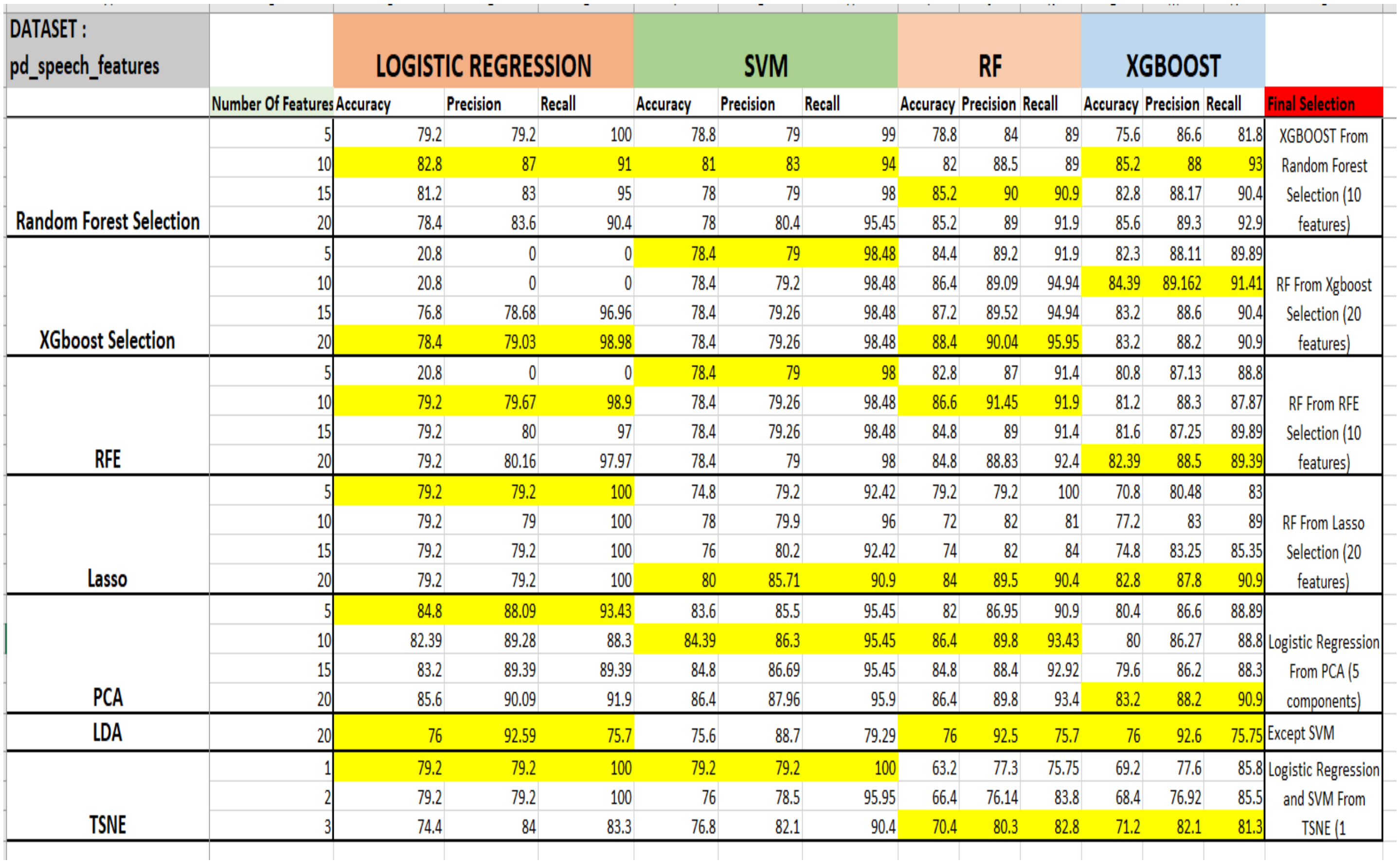
Used 7 methods of Feature Selection and Dimensionality Reduction on Classification and Regression dataset.

Used Logistic Regression, SVM, Random Forest and Xgboost for classification and applied all 7 methods to theses classifiers.

Used MLR and Voting Regression for regression and applied all 7 methods to these regressors.

CLASSIFICATION

DATASET: Parkinson's Disease Classification
The data used in this study were gathered from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87.



DATASET : pd_speech_features		LOGISTIC REGRESSION			SVM			RF			XGBOOST			
	Number Of Features	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Final Selection
Random Forest Selection	5	79.2	79.2	100	78.8	79	99	78.8	84	89	75.6	86.6	81.8	XGBOOST From Random Forest Selection (10 features)
	10	82.8	87	91	81	83	94	82	88.5	89	85.2	88	93	
	15	81.2	83	95	78	79	98	85.2	90	90.9	82.8	88.17	90.4	
	20	78.4	83.6	90.4	78	80.4	95.45	85.2	89	91.9	85.6	89.3	92.9	
XGboost Selection	5	20.8	0	0	78.4	79	98.48	84.4	89.2	91.9	82.3	88.11	89.89	RF From Xgboost Selection (20 features)
	10	20.8	0	0	78.4	79.2	98.48	86.4	89.09	94.94	84.39	89.162	91.41	
	15	76.8	78.68	96.96	78.4	79.26	98.48	87.2	89.52	94.94	83.2	88.6	90.4	
	20	78.4	79.03	98.98	78.4	79.26	98.48	88.4	90.04	95.95	83.2	88.2	90.9	
RFE	5	20.8	0	0	78.4	79	98	82.8	87	91.4	80.8	87.13	88.8	RF From RFE Selection (10 features)
	10	79.2	79.67	98.9	78.4	79.26	98.48	86.6	91.45	91.9	81.2	88.3	87.87	
	15	79.2	80	97	78.4	79.26	98.48	84.8	89	91.4	81.6	87.25	89.89	
	20	79.2	80.16	97.97	78.4	79	98	84.8	88.83	92.4	82.39	88.5	89.39	
Lasso	5	79.2	79.2	100	74.8	79.2	92.42	79.2	79.2	100	70.8	80.48	83	RF From Lasso Selection (20 features)
	10	79.2	79	100	78	79.9	96	72	82	81	77.2	83	89	
	15	79.2	79.2	100	76	80.2	92.42	74	82	84	74.8	83.25	85.35	
	20	79.2	79.2	100	80	85.71	90.9	84	89.5	90.4	82.8	87.8	90.9	
PCA	5	84.8	88.09	93.43	83.6	85.5	95.45	82	86.95	90.9	80.4	86.6	88.89	Logistic Regression From PCA (5 components)
	10	82.39	89.28	88.3	84.39	86.3	95.45	86.4	89.8	93.43	80	86.27	88.8	
	15	83.2	89.39	89.39	84.8	86.69	95.45	84.8	88.4	92.92	79.6	86.2	88.3	
	20	85.6	90.09	91.9	86.4	87.96	95.9	86.4	89.8	93.4	83.2	88.2	90.9	
LDA	20	76	92.59	75.7	75.6	88.7	79.29	76	92.5	75.7	76	92.6	75.75	Except SVM
TSNE	1	79.2	79.2	100	79.2	79.2	100	63.2	77.3	75.75	69.2	77.6	85.8	Logistic Regression and SVM From TSNE (1
	2	79.2	79.2	100	76	78.5	95.95	66.4	76.14	83.8	68.4	76.92	85.5	
	3	74.4	84	83.3	76.8	82.1	90.4	70.4	80.3	82.8	71.2	82.1	81.3	

1- **Random Forest Selection Interpretation:**

- Logistic Regression with Random Forest selection performs well on 10 features. the accuracy is 82.8, Precision 87 and recall 91.

- SVM with Random Forest selection performs well on 10 features as accuracy is 81, precision is 83, recall is 94.
- RF with random forest selection performs very well on 15 features as accuracy is 85.2, precision is 90 and recall is 90.9.
- Xgboost with random forest selection performs well on 10 features as accuracy is 85.2, precision is 88 and recall is 93.

We concluded that LR, SVM, Xgboost with 10 features works really better with Random Forest selection as it is useful in minimizing the features. So, if we want to select one model from random forest selection, we will choose Xgboost (with 10 features).

2- Xgboost Selection Interpretation:

- Logistic Regression with xgboost feature selection does not perform well on the smaller number of features it perform well on 20 features with the accuracy 78.4, precision 79 and recall 98.
- SVM with xgboost feature selection perform almost same with more or less features so it's better to select with less features. (Accuracy 78.4, precision 79, recall 98.48)
- RF with xgboost selection performs well on 20 features with accuracy 86.6, precision 90.04 and recall 95.95.
- Xgboost with xgboost selection performs well on 10 features with accuracy 84.39, precision 89.16, recall 91.41.

We concluded that from xgboost selection RF model (with 10 features) performs well from all of them.

3- RFE Selection Interpretation:

- Logistic Regression with RFE selection performs well with 10 features with an accuracy of 79, precision 79.67, recall 98.9.
- SVM with RFE selection perform well with 5 features with an accuracy 78, precision 79, recall 98.
- RF with RFE selection perform well with 10 features with an accuracy of 86.6, precision 91.45, recall 91.9.
- Xgboost with RFE selection performs well with 20 features with an accuracy of 82.39, precision 88.5, recall 89.39

We concluded that from RFE selection RF performs (with 10 features) model performs well from all of them.

4- Lasso Selection:

- Logistic Regression with lasso selection performs well with 5 features with an accuracy of 79, precision 79, recall 100.
- SVM with lasso selection performs well with 20 features with an accuracy of 80, precision 85, recall 90.
- RF with lasso selection performs well with 20 features with an accuracy of 84, precision 89, recall 90.
- Xgboost with lasso selection performs well with 20 features with an accuracy of 80, precision 85, recall 90.

We concluded that lasso selection of less features does not perform with classifiers. RF from lasso selection (20 features) performs well from all of them.

5- **PCA Dimensionality Reduction:**

- Logistic Regression with PCA performs well with 5 components with an accuracy of 84, precision 88, recall 93.
- SVM with PCA performs well with 10 components with an accuracy of 84, precision 86, recall 95.
- RF with PCA performs well with 10 components with an accuracy of 86, precision 89, recall 93.
- Xgboost with PCA performs well with 20 components with an accuracy of 83, precision 88, recall 90.

We concluded that PCA is useful in dimensionality reduction of the dataset as Logistic Regression with 5 components perform well from all of them.

6- **LDA Dimensionality Reduction:**

TSNE does not perform on a smaller number of components. It does not reduce the dimensionality of dataset. Logistic Regression, Xgboost, RF all perform well with 20 components except SVM.

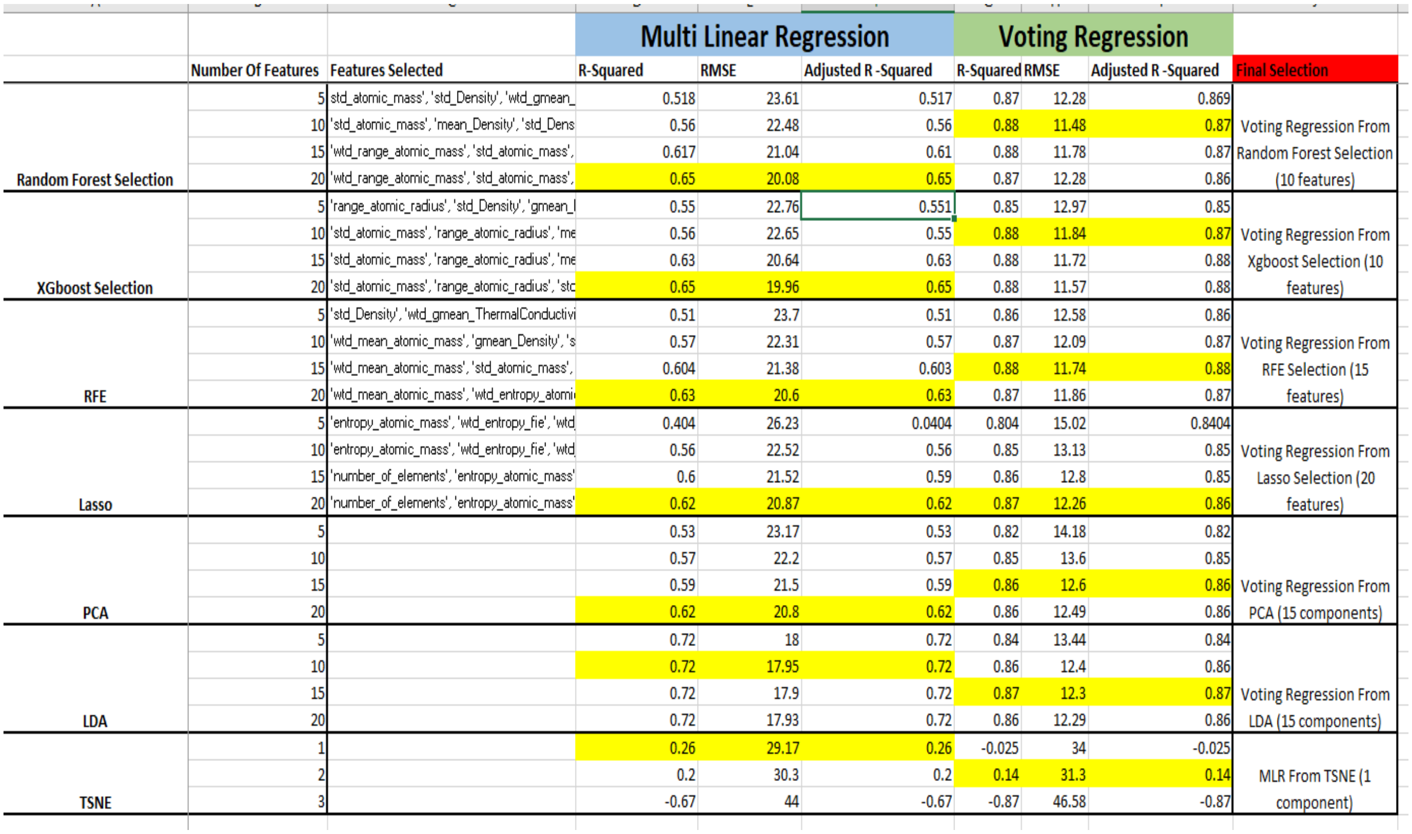
7- **TSNE:**

- Logistic Regression with TSNE performs well with 1 component with an accuracy of 79, precision 79, recall 100.
- SVM with TSNE performs well with 1 component with an accuracy of 79, precision 79, recall 100.
- RF with TSNE performs well with 3 components with an accuracy of 70, precision 80.3, recall 82.8.
- Xgboost with TSNE performs well with 3 components with an accuracy of 71, precision 82, recall 81.

We concluded that TSNE is useful in dimensionality reduction of the dataset as Logistic Regression with 1 component perform well from all of them.

REGRESSION

DATASET: Superconductivity Data
The data contains 21263 superconductors and their relevant features.
Superconductors along with the critical temperature in the 82nd column.



			Multi Linear Regression			Voting Regression			
	Number Of Features	Features Selected	R-Squared	RMSE	Adjusted R -Squared	R-Squared	RMSE	Adjusted R -Squared	Final Selection
Random Forest Selection	5	std_atomic_mass', 'std_Density', 'wtd_gmean_	0.518	23.61	0.517	0.87	12.28	0.869	Voting Regression From Random Forest Selection (10 features)
	10	'std_atomic_mass', 'mean_Density', 'std_Dens	0.56	22.48	0.56	0.88	11.48	0.87	
	15	'wtd_range_atomic_mass', 'std_atomic_mass',	0.617	21.04	0.61	0.88	11.78	0.87	
	20	'wtd_range_atomic_mass', 'std_atomic_mass',	0.65	20.08	0.65	0.87	12.28	0.86	
XGboost Selection	5	'range_atomic_radius', 'std_Density', 'gmean_I	0.55	22.76	0.551	0.85	12.97	0.85	Voting Regression From Xgboost Selection (10 features)
	10	'std_atomic_mass', 'range_atomic_radius', 'me	0.56	22.65	0.55	0.88	11.84	0.87	
	15	'std_atomic_mass', 'range_atomic_radius', 'me	0.63	20.64	0.63	0.88	11.72	0.88	
	20	'std_atomic_mass', 'range_atomic_radius', 'stc	0.65	19.96	0.65	0.88	11.57	0.88	
RFE	5	'std_Density', 'wtd_gmean_ThermalConductivi	0.51	23.7	0.51	0.86	12.58	0.86	Voting Regression From RFE Selection (15 features)
	10	'wtd_mean_atomic_mass', 'gmean_Density', 's	0.57	22.31	0.57	0.87	12.09	0.87	
	15	'wtd_mean_atomic_mass', 'std_atomic_mass',	0.604	21.38	0.603	0.88	11.74	0.88	
	20	'wtd_mean_atomic_mass', 'wtd_entropy_atomi	0.63	20.6	0.63	0.87	11.86	0.87	
Lasso	5	'entropy_atomic_mass', 'wtd_entropy_fie', 'wtd	0.404	26.23	0.0404	0.804	15.02	0.8404	Voting Regression From Lasso Selection (20 features)
	10	'entropy_atomic_mass', 'wtd_entropy_fie', 'wtd	0.56	22.52	0.56	0.85	13.13	0.85	
	15	'number_of_elements', 'entropy_atomic_mass'	0.6	21.52	0.59	0.86	12.8	0.85	
	20	'number_of_elements', 'entropy_atomic_mass'	0.62	20.87	0.62	0.87	12.26	0.86	
PCA	5		0.53	23.17	0.53	0.82	14.18	0.82	Voting Regression From PCA (15 components)
	10		0.57	22.2	0.57	0.85	13.6	0.85	
	15		0.59	21.5	0.59	0.86	12.6	0.86	
	20		0.62	20.8	0.62	0.86	12.49	0.86	
LDA	5		0.72	18	0.72	0.84	13.44	0.84	Voting Regression From LDA (15 components)
	10		0.72	17.95	0.72	0.86	12.4	0.86	
	15		0.72	17.9	0.72	0.87	12.3	0.87	
	20		0.72	17.93	0.72	0.86	12.29	0.86	
TSNE	1		0.26	29.17	0.26	-0.025	34	-0.025	MLR From TSNE (1 component)
	2		0.2	30.3	0.2	0.14	31.3	0.14	
	3		-0.67	44	-0.67	-0.87	46.58	-0.87	

1- **Random Forest Selection Interpretation:**
R Squared values increases in Multi Linear Regression when number of features increases. RMSE decreases in Multi Linear Regression when number of features increases.
R Squared values does not affect much in Voting Regression when number of features increases. RSME also does not affect much in Voting Regression when number of features increases. Also Voting Regression has low RMSE values and high R squared values which indicates Voting Regression better predicts the data.

Random Forest Selection works better/useful with Voting Regression

2- **Xgboost Selection Interpretation:**
R squared values of Voting Regression increases and becomes constant after 10 features RSME of voting regression decreases and becomes constant after 10 features.

R squared values of MLR increases with number of features. RSME decreases with number of features.

The values of R squared of Voting Regression is much larger means it overfits the model. MLR from random forest selection is more suitable. But MLR from xgboost selection is not useful as RSME keeps decreasing and not becoming constant.

3- RFE Selection Interpretation:

R squared values of MLR increases with the number of features. RSME decreases with number of features.

Values of R squared and RSME becomes constant 15 features. Voting regression with 15 features is better as the RSME is minimum and R square is maximum.

4- Lasso Selection:

R squared values of MLR increases with the number of features. RSME of MLR decreases with number of features.

R squared values of Voting Regression increases with the number of features. RSME of MLR decreases with number of features.

Voting regression is better as the RSME is minimum and R square is maximum. L1 feature based selection is not much useful.

5- PCA Dimensionality Reduction:

After increasing components of PCA each algorithm performs well therefore we can’t reduce the original data to extremely low dimension space using PCA.

We choose Voting Regression with 15 components.

6- LDA Dimensionality Reduction:

LDA performs well on MLR as we can reduced the original data to extremely low dimension space using LDA because R squared values and RSME values are constant.

We choose Voting Regression with 15 components because it gives less RSME.

7- TSNE:

TSNE performs badly on both the models (MLR and Voting Regression) we can reduced the original data to low dimension space because after 2 components the RSME increases and R squared values decreases but the values of R squared is negative or very low. MLR and Voting Regression with TSNE gives poor values of RSME and R squared values.
We choose MLR with 1 component because the RSME is low.