A Comparison of Classification Algorithms:

Classification of celestial objects: stars, galaxies and quasars

Classifiers	Accuracy Score	Strengths	Weaknesses
1. Logistic Regression **cross validation logistic regression** **with regularization** important featrures: psfMag_u, psfMag_g , petromag_g , gr, ri, ug	0.97 0.968 0.9693333	1.Most interpretable machine learning algorithms 2.Regularized to avoid overfitting	Underperform when there are multiple or non-linear decision boundaries
2. SVM classifier **Using "OneVsRestClassifier"**	0.954	1.Non-linear decision boundaries 2.Robust against overfitting, especially in high-dimensional space 3. Best classification performance (accuracy) on the training data.	1.Don't scale well to larger datasets 2.Random forests are usually preferred over SVM's.
3. KMeans	-	Fast, simple, and surprisingly flexible	If the true underlying clusters in the data are not globular, then K-Means will produce poor clusters
4. KNN classifier	0.904	1.Robust to noisy training data 2.Effective for large training data	1.It is costly and lazy, 2.Requires full training data plus depends on the value of k 3. Has the issue of dimensionality because of the distance
5. Ensemble Classifications • **Random Forest Classifier** • **XGB classifier**	0.97466667 0.9727	1.Perform well in practice 2.Robust to outliers, 3. scalable, 4. Naturally model non-linear decision boundaries 5.Overfitting is less	1.Analysing theoretically is difficult 2.Large number of decision trees can slow down the algorithm in making real-time predictions. 3.If the data consists of categorical variables with different number of levels, then

Important Features: Ug, iz, ri, psfFlux_u		6.Fast but not in all cases 7.Most effective and versatile 8.More robust to noise. 9.Can be grown in parallel. 10.Runs efficiently on large databases. 11.Has higher accuracy	the algorithm gets biased in favour of those attributes
7. Decision Tree classifier	0.9384	1.can handle missing values nicely 2.best suited when the target function has discrete output values	1.The more the number of decisions in a tree, less is the accuracy 2.do not fit well for continuous variables and result in instability and classification plateaus. 3.creating large decision trees that contain several branches is a complex and time-consuming task.