McMaster University

STATS 780

Data Science

Final Report

Author
Sean Somogyvari
001226262

Supervisor
Dr. Sharon McNicholas

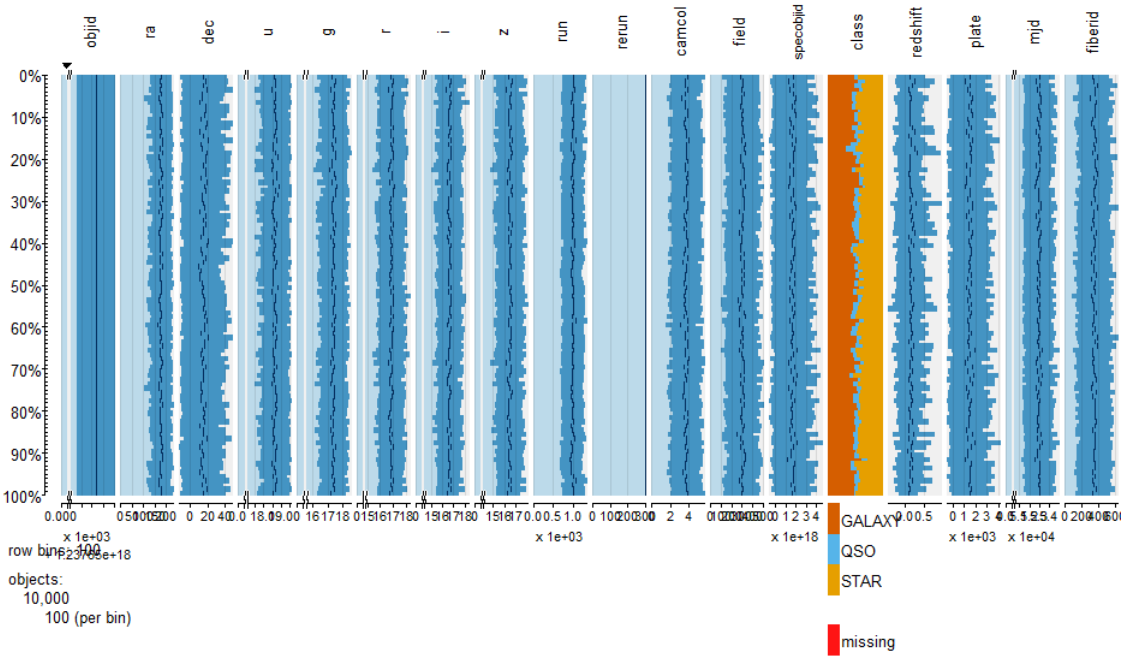
November 27, 2018

1 Purpose

The purpose of the following report was to examine the classification of the *Sloan Digital Sky Survey RD14* [1] by exploring popular classification methods including decision trees, ensemble methods, clustering, model-based clustering, mixture discriminant analysis, and neural networks.

2 Description of Data

The data examined is the *Sloan Digital Sky Survey DR14* explaining Sloan Digital Sky Survey (SDSS) telescope data from July 2016. The data consists of 10,000 sample observations described by 17 feature columns and one response column which identifies the sample as either a star, galaxy, or quasar. The variables included within the data were the Object ID, center point right ascension, center point declination, u, g, r, i, z, run, rerun, camcol, field, spec object ID, redshift, plate, mjd, and fiber ID. A full description of all variables involved may be found on the SDSS website [2].

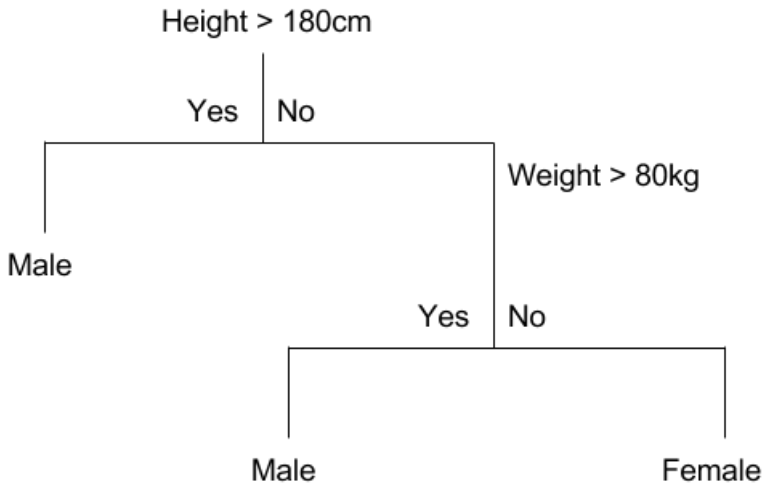


3 Techniques

3.1 Decision Tree Learning

Decision tree learning is the process of generating "trees" based on observations from a dataset in order to form a conclusion regarding a datasets target values. Decision trees are used typically for classification (classification trees) and predictive regression problems (regression trees). The decision tree takes the shape of an inverted tree, where the

initial node at the top of the tree is regarded as the root. The root splits off into branches based on conditions formed from the data, and when a branch comes to an end the final node is considered a leaf [3].



Splitting is performed by recursively partitioning data such that groups are formed into branches with similar responses. There are multiple algorithms used to decide a splitter, the Gini impurity is used by the classification and regression tree algorithm (CART). The Gini impurity measures the probability, $\hat{p}_{mg}$, of obtaining two distinct outputs, given node, m, and class, g [4].

$$\hat{p}_{mg} = \frac{1}{N_m} \sum_{\mathbf{x_i} \in R_m} I(y_i = g) \tag{3.1}$$

$$\sum_{g=1}^{G} \hat{p}_{mg}(1 - \hat{p}_{mg}) \tag{3.2}$$

3.2 Ensemble Methods

Decision trees are pleasant for visualization, however a combination of decision trees make more robust decisions. There are three main methods, Bagging, Boosting, and Random Forests. Bagging is the process of creating several subsets of data given a training set. Each subset trains their own distinct decision tree, where the output of bagging is then the average across all trees.

$$\hat{f}_{bag}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \hat{f}^m(\mathbf{x}) \tag{3.3}$$

Random forests is an extension of bagging. When growing the tree, a random sample of predictors is chosen instead of all being available. Each random split is different across all

trees grown in the random forest. This method takes all predictors into account regardless of weight. In boosting, trees are grown sequentially where the response is the previous trees residuals.

3.3 Clustering

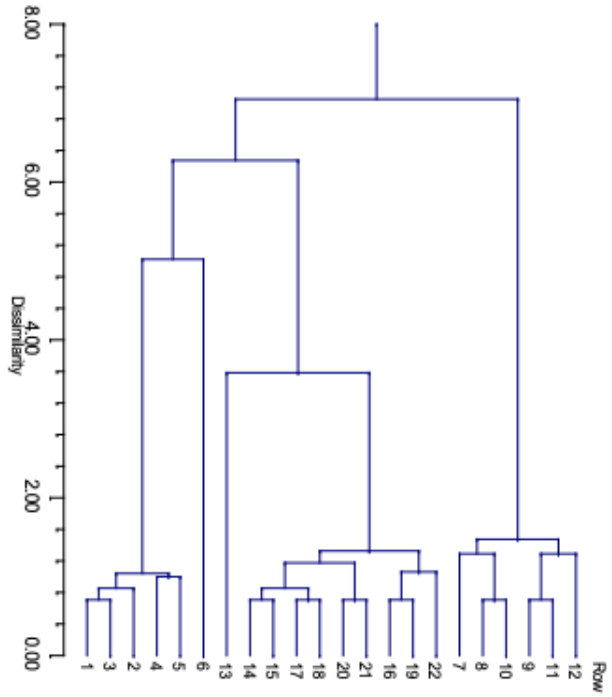
Clustering, unlike decision tree classification, is a form of unsupervised learning. The basic process of clustering is to look for relationships between objects of a group which are more similar to each other than other objects.

Agglomerative Hierarchical Clustering is a bottom-up approach. Each sample observation is assigned to its own cluster, then the two clusters which reside nearest to each other are combined into one larger cluster. The distance between two observations is decided using dissimilarity, while the distance between clusters is decided using linkage. A popular dissimilarity and linkage are the Euclidean distance and complete linkage, respectively.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \tag{3.4}$$

$$d(A, B) = \max_{\mathbf{x} \in A, \mathbf{y} \in B} d(\mathbf{x}, \mathbf{y}) \tag{3.5}$$

A dendrogram is used to visualize hierarchical clustering, where the vertical axis represents the dissimilarity between clusters, and the horizontal axis differentiates the clusters [5].



Another method of clustering is known as k-Means or $k-$Medoids clustering. Where k clusters are determined based on the minimization of the Euclidean distance between points. The cluster centre is chosen as the mean or medoid average of al the points within the cluster. The downfall of k-Means or k-Medoids clustering arises when the clusters are not of spherical nature. Since clustering is an unsupervised method, the number of clusters may be unknown. Thus a methods such as "the elbow" and silhouette approach are used to determine the number of clusters.

3.4 Model-Based Clustering

Because of the ambiguity of how certain clusters may be shaped, it's important to explore other methods of clustering. Model-based clustering is a method based on the use of statistical mixture models. Mixture models help identify sub-populations of clusters within data that may overlap each other. The density of a finite mixture model given the $g_t h$ mixin proportion, ϑ of parameters, and a vector π with $\sum \pi = 1$.

$$f(\mathbf{x}|\theta_g) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\theta_g) \tag{3.6}$$

The Gaussian mixture model is widely used in data science, withe likelihood function for $\mathbf{x}$.

$$\mathcal{L}(\vartheta, \mathbf{x}) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(\mathbf{x_i}|\mu_\mathbf{g}, \mathbf{\Sigma_g}) \tag{3.7}$$

Parameter estimation is performed using the expectation-maximization algorithm [6]. Specifically the Gaussian mixture model has $Gp(p+1)/2$ free parameters. The Gaussian parsimonious clustering models (GPCMs) contain the best known model-based clustering family. The eigenvalue decomposition of the component covariance models is as follows [7] with the family of models shown in the table [8].

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g^T \tag{3.8}$$

Model	Volume	Shape	Orientation	$\boldsymbol{\Sigma}_g$	No. Covariance Parameters
EII	Equal	Spherical	–	$\lambda \mathbf{I}$	1
VII	Variable	Spherical	–	$\lambda_g \mathbf{I}$	G
EEI	Equal	Equal	Axis-Aligned	$\lambda \boldsymbol{\Delta}$	p
VEI	Variable	Equal	Axis-Aligned	$\lambda_g \boldsymbol{\Delta}$	$p + G - 1$
EVI	Equal	Variable	Axis-Aligned	$\lambda \boldsymbol{\Delta}_g$	$pG - G + 1$
VVI	Variable	Variable	Axis-Aligned	$\lambda_g \boldsymbol{\Delta}_g$	pG
EEE	Equal	Equal	Equal	$\lambda \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}'$	$p(p+1)/2$
EEV	Equal	Equal	Variable	$\lambda \boldsymbol{\Gamma}_g \boldsymbol{\Delta} \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2 - (G-1)p$
VEV	Variable	Equal	Variable	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta} \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2 - (G-1)(p-1)$
VVV	Variable	Variable	Variable	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2$
EVE	Equal	Variable	Equal	$\lambda \boldsymbol{\Gamma} \boldsymbol{\Delta}_g \boldsymbol{\Gamma}'$	$p(p+1)/2 + (G-1)(p-1)$
VVE	Variable	Variable	Equal	$\lambda_g \boldsymbol{\Gamma} \boldsymbol{\Delta}_g \boldsymbol{\Gamma}'$	$p(p+1)/2 + (G-1)p$
VEE	Variable	Equal	Equal	$\lambda_g \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Gamma}'$	$p(p+1)/2 + (G-1)$
EVV	Equal	Variable	Variable	$\lambda \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2 - (G-1)$
VVV	Variable	Variable	Variable	$\lambda_g \boldsymbol{\Gamma}_g \boldsymbol{\Delta}_g \boldsymbol{\Gamma}_g'$	$Gp(p+1)/2$

The best model selection from the GPCMs is done using the Bayesian Information Criterion [9].

$$BIC = 2l(\vartheta) - \rho log n \tag{3.9}$$

For higher dimensional data a mixture of factor analyzers is introduced as a data reduction technique.

$$\mathbf{X}_i = \mu + \boldsymbol{\Lambda} \mathbf{U}_i + \epsilon_i \tag{3.10}$$

$$f(\mathbf{x}_i|\theta_g) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i|\mu_g, \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^T + \boldsymbol{\Psi}_g) \tag{3.11}$$

McNicholas and Murphy [10] developed a family of parsimonious Gaussian mixture models using a component covariance structure as follows.

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g^T + \omega_g \boldsymbol{\Delta}_g \tag{3.12}$$

Expanded PGMM Nomenclature					
$\mathbf{\Lambda}_g = \mathbf{\Lambda}$	$\mathbf{\Delta}_g = \mathbf{\Delta}$	$\omega_g = \omega$	$\mathbf{\Delta}_g = \mathbf{I}_p$	PGMM Equiv.	$\mathbf{\Sigma}_g$
C	C	C	C	CCC	$\mathbf{\Lambda\Lambda}' + \omega\mathbf{I}_p$
C	C	U	C	CUC	$\mathbf{\Lambda\Lambda}' + \omega_g\mathbf{I}_p$
U	C	C	C	UCC	$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega\mathbf{I}_p$
U	C	U	C	UUC	$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g\mathbf{I}_p$
C	C	C	U	CCU	$\mathbf{\Lambda\Lambda}' + \omega\mathbf{\Delta}$
C	C	U	U	–	$\mathbf{\Lambda\Lambda}' + \omega_g\mathbf{\Delta}$
U	C	C	U	UCU	$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega\mathbf{\Delta}$
U	C	U	U	–	$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g\mathbf{\Delta}$
C	U	C	U	–	$\mathbf{\Lambda\Lambda}' + \omega\mathbf{\Delta}_g$
C	U	U	U	CUU	$\mathbf{\Lambda\Lambda}' + \omega_g\mathbf{\Delta}_g$
U	U	C	U	–	$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega\mathbf{\Delta}_g$
U	U	U	U	UUU	$\mathbf{\Lambda}_g\mathbf{\Lambda}_g' + \omega_g\mathbf{\Delta}_g$

3.5 Mixture Discriminant Analysis

Mixture discriminant analysis is the process of using model-based clustering on labeled observations in order to classify unlabeled observations. The maximum likelihood estimation is calculated for all known observations. The BIC is then used for best model selection.

$$\mathcal{L}(\vartheta|x_{1:n}) = \prod_{i=1}^{n}\sum_{g=1}^{G}\pi_g\phi(x_i|\mu_g,\Sigma_g) \tag{3.13}$$

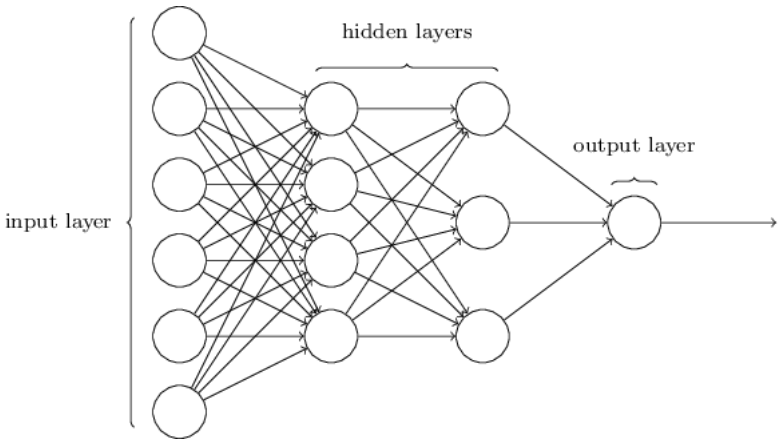
The resulting mixture model fit to the known observations ,x1:n, and resultant parameters, is used to calculate the class, g , which the current unknown observation belongs to corresponding components h as seen below.

$$z_{jg} = \frac{\pi_g\phi(x_j|\mu_g,\Sigma_g)}{\sum_{h=1}^{G}\pi_h\phi(x_j|\mu_h,\Sigma_h)} \tag{3.14}$$

$$h = argmax_g(z_{jg}) \tag{3.15}$$

3.6 Neural Networks

A neural network is a system modeled after neurons within the brain used for regression or classification. A neural network consists of a series of weighted sums forming hidden layers with m nodes. The output is similarly a weighted sum of hidden nodes activated using a non-linear function.



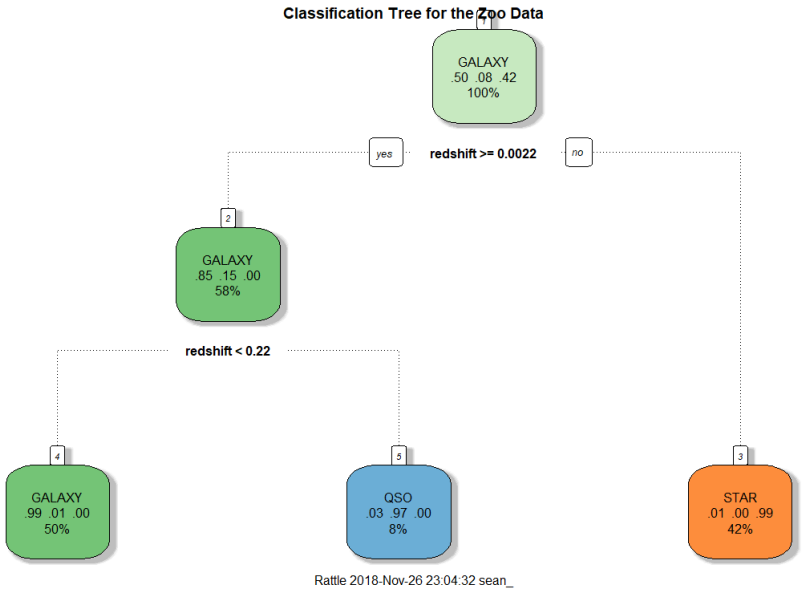
$$Z_m = \alpha_{0m} + \alpha_m^T \mathbf{X} \tag{3.16}$$

$$T_k = \beta_{0k} + \beta_k^T \sigma \mathbf{X} \tag{3.17}$$

Considering one hidden layer, where Z_m is each hidden node belonging to the single hidden layer. The activation function σ, is the sigmoid function, however there are multiple possibilities. There are k output nodes T_k [4]. The output is produced using an output function $g_k(\mathbf{T})$, dependent on the goal. For regression $g_k(\mathbf{T}) = T_k$ and, classification the softmax function is used, $g_k(\mathbf{T}) = \frac{e^{T_k}}{\sum_{h=1}^{K} e^{T_k}}$.

4 Results

As discussed in section 3.1, a classification tree can is grown by recursively partitioning data based on the best splitter. This can be performed with thanks to the rpart function in R. The following tree was grown for the Sloan data.

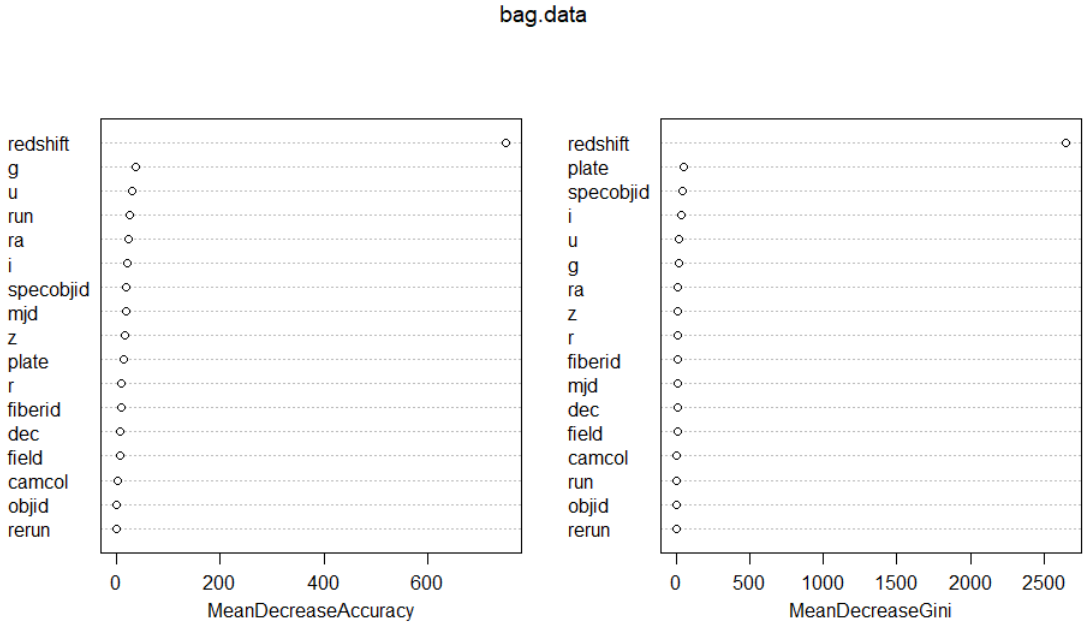


The classification tree shows that the redshift predictor dominated all the other predictors. According to the tree, Stars were classified for redshift < 0.0022, Galaxies for

redshift $>$ 0.22, and Quasars in between.

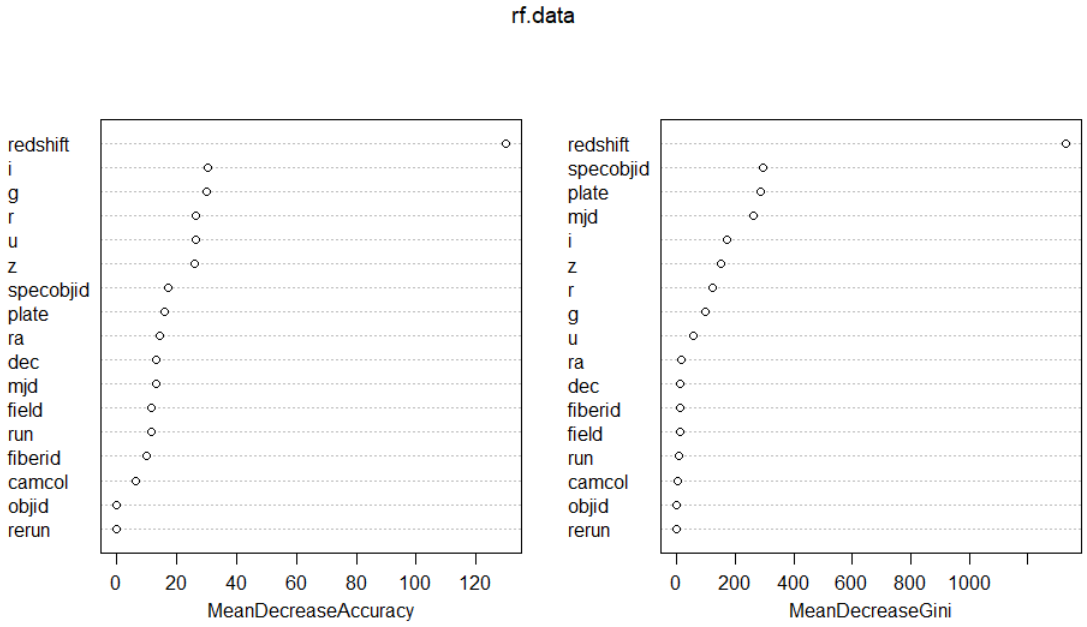
In order to form a more robust classification, bagging is performed. The randomForest function in R allows the following plots to be generated for N training data (half of the data set). The classification error and confusion matrix for the training data was produced. The variable importance plot shows that the redshift dominates the variables, agreeing with the classification tree.

```
Confusion matrix:
       GALAXY QSO STAR class.error
GALAXY   2514  14    9 0.009065826
QSO        23 386    1 0.058536585
STAR        5   0 2048 0.002435460
```



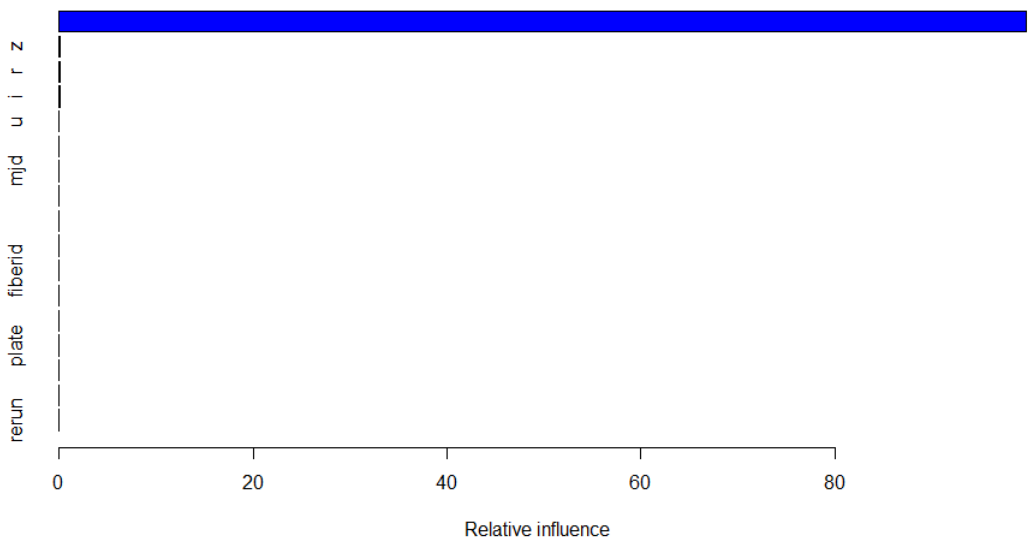
The bagging was used to classify the remaining testing data, resulting in a misclassifcation mean of 0.013. The randForest function was also used to generate a random forest on the same training data.

```
       GALAXY QSO STAR  class.error
GALAXY   2516  10   11 0.0082774931
QSO        26 383    1 0.0658536585
STAR        2   0 2051 0.0009741841
```



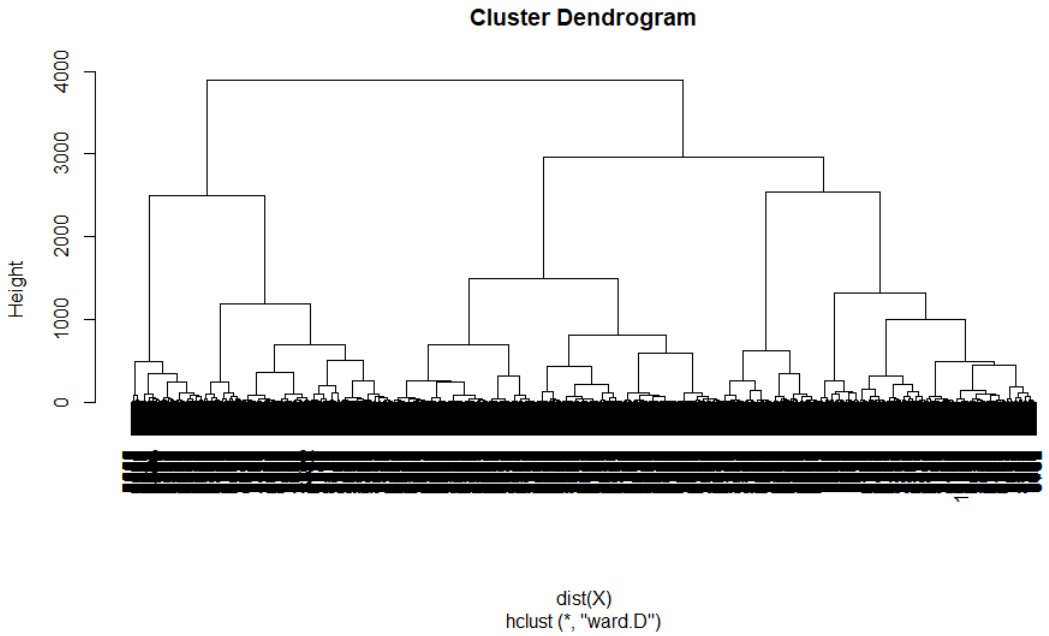
As expected, the variable importance plot shows that redshift is dominant. However, due to the random predictor nature of the random forest, other variables are taken into consideration. The random forest was used to classify the same testing data resulting in a mean misclassifcation of 0.128, performing better than the bagging.

Boosting is performed in R using the gbm function. The relative influence plot once again shows that the redshift dominates the rest of the variables. Classifying the same testing data resulted in a mean misclassifcaiton of 0.0124. Boosting performed the best out of the three methods.



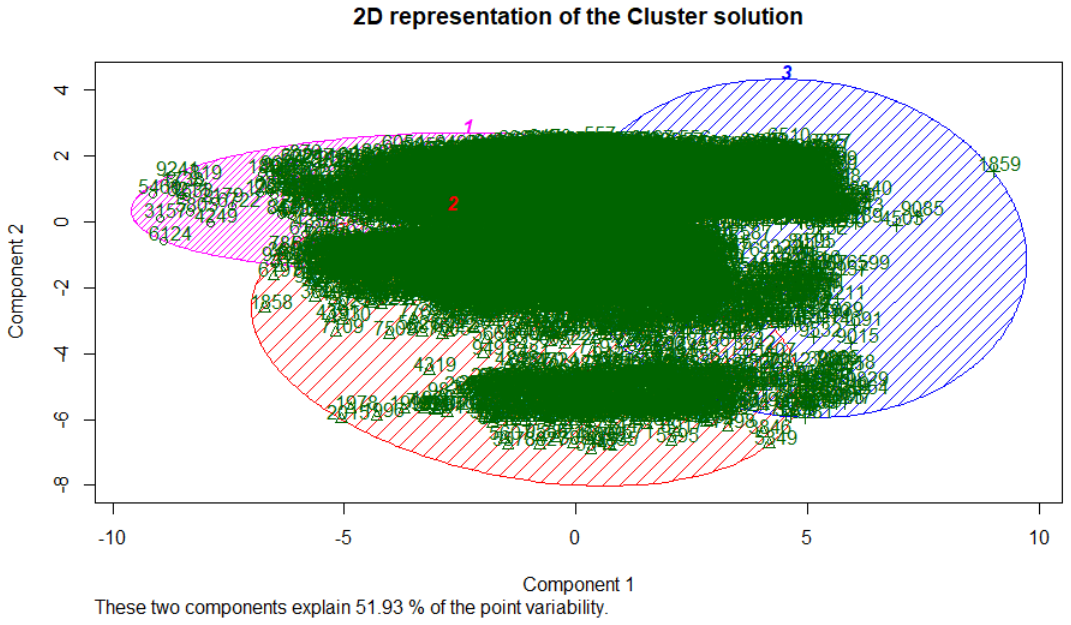
Hierarchical clustering performed on the data set in R using the hclust function using wards linkage, results in the following dendrogram and confusion matrix. However, the classification error is quite high.

	1	2	3
GALAXY	123	2948	1927
QSO	707	52	91
STAR	2062	642	1448



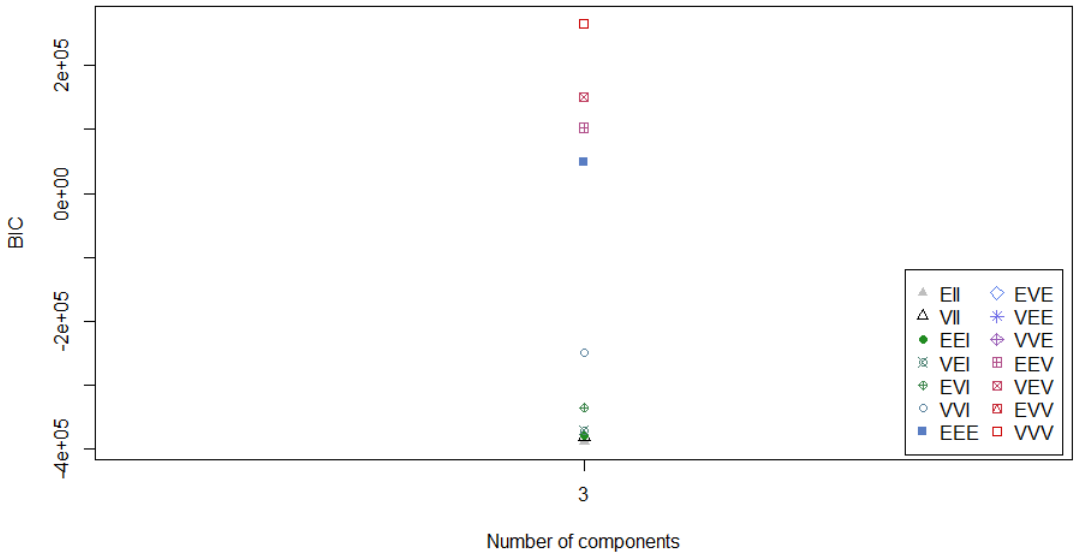
The k-means can be performed for 3 clusters in R using the kmeans function. However, much like from the hierarchical clustering, this clusters are not distinct and spherical. This overlap causes alot of misclassifcation, the attempted clustering is seen in the below figure.

	1	2	3
GALAXY	3328	118	1552
QSO	34	40	776
STAR	600	2369	1183



Model based clustering was attempted in hopes for an appropriate fit, using the Mclust function in R. Unfortunately, classification error was too high as the GPCMs did not fit that data well. The BIC was plotted, showing that VVV model fit the best.

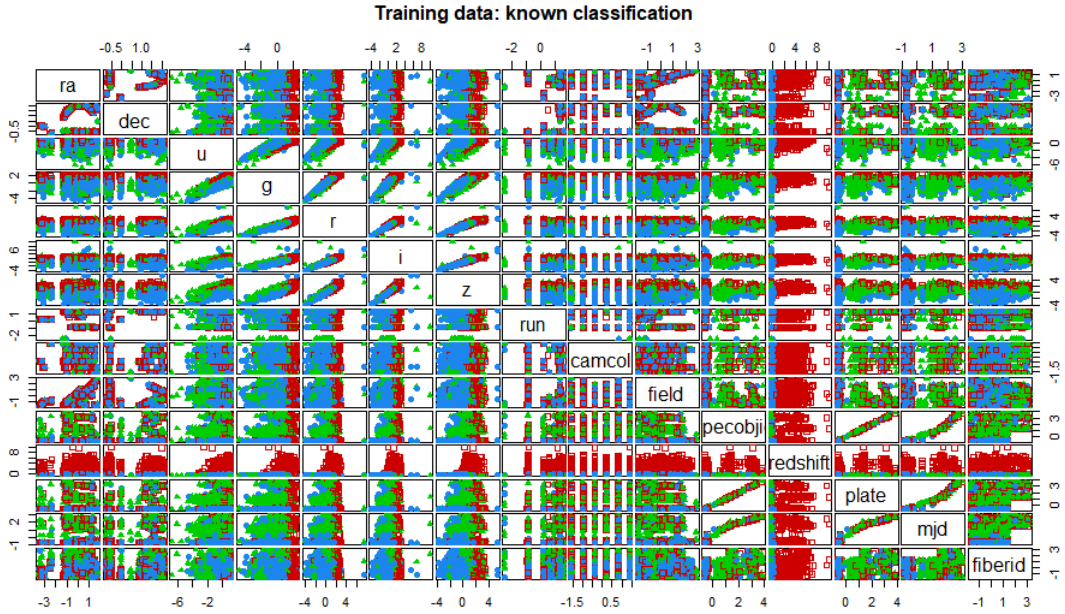
	1	2	3
GALAXY	699	1878	2421
QSO	10	803	37
STAR	254	1675	2223



Since the GPCM did not fit, mixed discriminant analysis was performed using Gaussian finite mixture modeling in R thanks to the MclustDA function. Training data was formed excluding every 4th sample from the data set. The training data used resulted in a

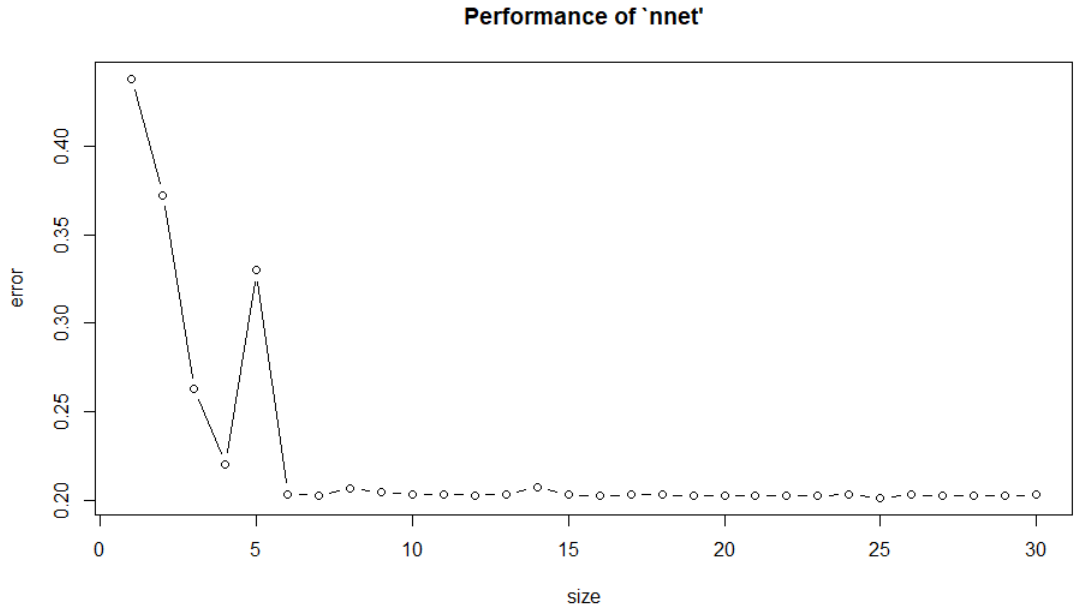
training error of 0.038. Predicting classes based on testing data (every 4th sample) resulted in a testing error of 0.0336.

	Predicted		
Class	GALAXY	QSO	STAR
GALAXY	1185	36	5
QSO	7	194	0
STAR	24	12	1037



Finally, a neural network was implemented to classify the data. Training data was formed using 8000 random samples, the rest used as testing data. Cross validation was used to see the optimal number of hidden units, using the tune.nnet function in R for the number of hidden units between 1 and 30. The plot between the size of the hidden layer and error shows that for 6 hidden units and above the error in minimized. The absolute best error of 0.20175 occurred at 25 hidden units. The function nnet in R was then used to create a neural network with 25 hidden units on the training data. The neural network then predicted the classes of the remaining testing data, shown in the confusion matrix. The neural network had a classification error of 0.015.

	GALAXY	QSO	STAR
GALAXY	993	11	5
QSO	12	160	0
STAR	1	1	817



5 Conclusion

Classification methods were examined using the *Sloan Digital Sky Survey RD14* [1], including decision trees, ensemble methods, clustering, model-based clustering, mixture discriminant analysis, and neural networks. The classification tree developed gave a strong insight as to what variables mattered the most, mainly being the redshift. Where stars were classified for a redshit < 0.0022, Galaxies for a redshift > 0.22 , and Quasars in between. Bagging, random forests, and boosting all agreed that the redshift was the most important variable. Each had a misclassification mean of 0.013, 0.128, and 0.0124 respectively. Proving that boosting performed the best. Clustering was explored using hierarchical clustering with wards linkage and k-means clustering, however both methods resulted in a large misclassification rate. This happened due to the clusters overlapping and not belonging to a spherical nature. To attempt to mediate the irregular shape, model based clustering was used. The VVV model from the GPCM family was chosen as the best fit from the BIC. However, misclassification was still high. Mixed discriminant analysis was implemented and resulted in a testing error of 0.0336. Finally, a neural network was formed based on 5-fold CV, using 25 hidden nodes. The classification error was only 0.015.

The neural network provided the best result for classification. In the future when the SDSS observes more data, the network may be used to help classify unknown space entities.

References

[1] Sloan Digital Sky Survey RD14
`https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey`

[2] Glossary of SDSS Terminology
`https://www.sdss.org/dr12/help/glossary/`

[3] Classification And Regression Trees for Machine Learning
`https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/`

[4] T. Hastie, R. Tibshirani, J. Friedman *The Elements of Statistical Learning* Spring Series in Statistics Second Edition, 2009.

[5] Dendrogram Service
https://www.creative-proteomics.com/services/dendrogram-service.htm

[6] G. McLachlan, T. Krishnan *The EM Algorithm and Extensions* Wiley Series in Probability and Statistics, 2008.

[7] J. Banfield, A. Raftery *Model-Based Gaussian and Non-Gaussian Clustering* International Biometric Society, pg 803-821, 1993.

[8] G. Celeux, G. Govaert *Choosing models in model-based clustering anddiscriminant analysis* URA CNRS 817, 1995.

[9] G. Schwarz *Estimating the Dimension of a Model* The Annals of Statistics vol. 6, pg 461-464, 1978.

[10] S. McNicholas, T. Murphy *Model Based Clustering of Microarray Expression Data via Latent Gaussian Mixture Modelsl* Bioinformatics vol. 26, pg 2705 - 2712, 2010.

[11] How Neural Networks Work
`https://chatbotslife.com/how-neural-networks-work-ff4c7ad371f7`