Hyper-Parameter Optimization of Machine Learning Techniques for Classification of Astronomical Sources

MOGAMMAD YAASEEN JONES

November 2020

A thesis presented to the University of Cape Town in full fulfilment of the degree: Honours in Astrophysics and Space Science

Supervisor: Prof. Mattia Vaccari
Co-Supervisor: Chaka Mofokeng

Abstract

This thesis aims to find a suitable machine learning classifier model to classify celestial bodies into three classes of astronomical sources (stars, star-forming galaxies and quasars) using data from the Sloan Digital Sky Survey (SDSS) Data Release 16 (DR16) and the Wide-field Infrared Survey Explore (WISE). The Random Forest (RF) and K-Nearest Neighbour (KNN) classifiers have been adopted, and the RF is found to have the best accuracy, so its hyper-parameters are optimized to further improve the accuracy of the classifier. The accuracy of the RF classifier is improved upon optimization, showing the importance that some wavebands have in classifying astronomical sources. The waveband feature importance is also found to change slightly after hyper-parameter optimization. Optimization of hyper-parameters is especially important for dealing with extremely large datasets like those from upcoming wide-area surveys in the optical and in the radio.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Mattia Vaccari, for guiding me throughout the project.

I would also like also to thank Chaka Mofokeng for providing a classification code that could be used for analysis.

Lastly I would like to thank my peer, Walter Silima, for his constant support throughout the project.

Contents

1 Introduction

As we move further into the future, technological advancements are improving considerably. Astronomy is no exception, and the advent of digital electronics means that the amount of data produced by modern telescopes is increasing very rapidly. Big Data is the term conventionally used for extremely large datasets. These large datasets may contain patterns, trends and relationships which may not be otherwise seen in smaller data samples. Since in astronomy we deal with an incredibly large amount of astronomical sources, classifying them all in any given sky survey may be a burden for anyone. It is also worth noting that when observing the night sky with the naked eye, the extremely large amount of astronomical sources within it appear ambiguous and unresolved. Stars, gas and dust particles are scattered across the sky, each having their own size and location while collectively creating a source that emits/absorbs electromagnetic frequencies (EMFs). Applying colour filters to detected EMFs allow the flux received for each source to be split up into different colours. Different astronomical sources have signature colours which when paired with other properties like location and size, make them classifiable. These properties can be determined for an increasingly large amount of sources because of advancements made in charged-couple devices (CCDs) and telescopes, allowing for the unambiguous classification of astronomical sources.

1.1 Astronomical Sources

1.1.1 Stars

Stars are hot balls of gas held together by their own gravity. Stars have different observed colours, sizes and temperatures owing to stars having different spectral classes (Carroll et al., 2014). Stars are classified based on their temperature where the most common method of classification of them is the Harvard Classification which labels stars as "O B A F G K M L T" (Seeds et al., 2011). Each letter is a classification of a different temperature where the hottest stars are O stars and the coolest ones are T stars. Hotter stars are generally bluer in colour and cooler stars are generally redder in colour. This classification is then further expanded to late and early type stars by adding a decimal subdivision for each spectral class where an A0 star is an early type A star and a A9 is a late type A star. The emission/absorption lines observed from stellar spectra indicate that spectral strength is dependant on spectral types and temperature. Taking a large dataset of observed stars with their observed properties indicate a trend as seen below in Figure (1.1) as the Hertzsprung-Russel (HR) diagram. As stars age, they burn their fuel, become cooler, appear fainter and shrink. The HR Diagram shows the plot of luminosity of absolute magnitude of the stars against their varying effective temperature. It is worth noting the *Main Sequence* stars which lie along a diagonal starting from the top left corner where O stars are, to the bottom right corner where M stars are. Another trendline is seen right above the main sequence known as the *Horizontal Branch* where a band of giants and super-giants are found. The HR diagram is an indicator that stars may have the same luminosity yet have different temperatures and/or visa versa.

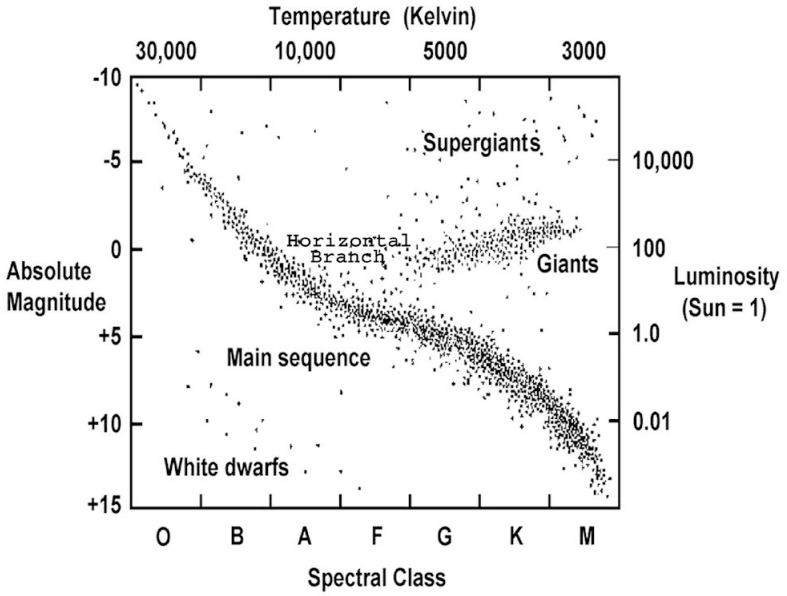


Figure 1.1: HR Diagram showing the spectral classes of stars and their associated temperatures (Image provided by Chandra X-Ray Observatory)

Since there is an overlap between luminosity and temperature , the need for other forms of classification, like the Morgan Keenan luminosity classes are needed. The Morgan Keenan luminosity classes are helpful in differentiating classes of same spectral type but with a variety of temperatures. All together, the dependence of strength of observed emission/absorption lines on temperature, the Stefan-Boltzmann law and the HR Diagram create a more simpler means in which to classify stars. By obtaining the spectra of stars, properties like effective temperature, radius and luminosity can be approximated.

1.1.2 Star-Forming Galaxies

In any region of space with dense interstellar dust clouds, gravity eventually takes over, creating stars or star clusters that are gravitationally bound. These gravitationally bound stars and star clusters are known as a galaxy. Galaxies can also be created by the collision of other galaxies. Properties that define galaxies are their colour and rotational velocities. Galaxies appear as either elliptical, spiral or irregular in shape, each having different structures and star formation rates. Elliptical galaxies do not have spiral arms and contain very little interstellar medium. They consist mostly of old super-giant and dwarf stars. Spiral galaxies however, have spiral arms, a central bulge and disks. If the central bulge is elongated, a spiral galaxy is also called a barred spiral galaxy. Spiral galaxies contain interstellar medium and very little super-giant and dwarf stars (Seeds et al., 2011). Irregular galaxies are shapeless and appear any way in which they please. They do not have a recognizable structure and are mostly gas and dust. When galaxies interact, their gas and dust are compressed and trigger rapid star formation. Star formation may be detected in the infrared region of the electromagnetic spectrum and are indicators of star formation galaxies (SFGs) (Seeds et al., 2011). The Hubble Tuning Fork is a visual representation of the different types of galaxies. In Figure (1.2) the morphological differences between types of galaxies can be seen.

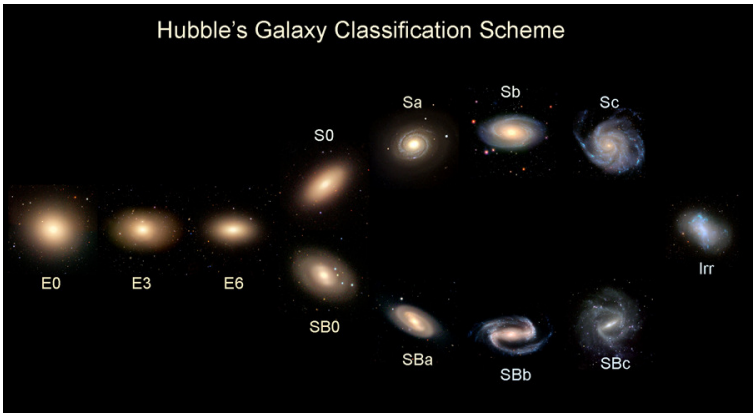


Figure 1.2: Hubble Tuning Fork Diagram (Carroll et al., 2014)

1.1.3 Active Galactic Nuclei (AGNs) / Quasars

Many galaxies display a very bright compact nucleus, or an Active Galactic Nucleus (AGN). AGNs are between some of the most luminous objects in the Universe, and as such can be detected up to very large distances. Since they often outshine their host galaxies altogether, they may appear as point-like sources. The first AGNs to be observed were thus referred to as Quasi-Stellar (Radio) Sources, or Quasars. The observed properties of AGNs are rather diverse as they can be observed in all wavebands of the EM spectrum. AGNs are believed to have a super-massive black hole at their centre and an accretion disk surrounding it that radiates in the X-ray. AGNs are highly luminous in their centres where relativistic jets blast outward from. Spectral lines of AGNs are often very broad, indicating extremely hot temperatures. This implies that thermal radiation plays a part in the AGNs radiative processes. Processes likes synchrotron radiation and non-thermal radiation are also present. A popular AGN classification scheme divides them into radio-quiet and radio-loud (Carroll et al., 2014; Seeds et al., 2011). Radio-quiet AGNs emit weak nuclear emission-line regions. Radio-loud AGNs show similar properties as radio-weak AGNs but their emissions are stronger as they include a radio jet as seen in sources like blazars and radio galaxies.

2 Multi-Wavelength Photometry

Photometry is the study/science of the measurement of light or electromagnetic radiation (Bass, 1995). Light may be collected using 2D array detectors such as charged-coupled devices (CCDs), which are currently the workhorse detectors installed on telescopes for wide-area optical sky surveys. Electromagnetic spectrum can be detected over a wide wavelength range, and multi-wavelength photometry is thus the study of light from different sources over several wavebands. Combining measurements in different wavebands, properties like temperature and chemical composition of the source can be determined.

2.1 SDSS

The Sloan Digital Sky Survey (SDSS) telescope is a 2.5 meter telescope at Apache Point Observatory in New Mexico, USA. The telescope has a wide-area, multi-band CCD with a pair of fibre-fed double spectrographs (Gunn et al., 2006). The SDSS carried out an imaging survey in 5 broad-band filters (known as u, g, r, i and z) down to limiting magnitudes of 20.2, 22.2, 22.2, 21.3 and 20.5 in the AB system respectively (Djorgovski et al., 2013). This can be seen below in Figure (2.1).

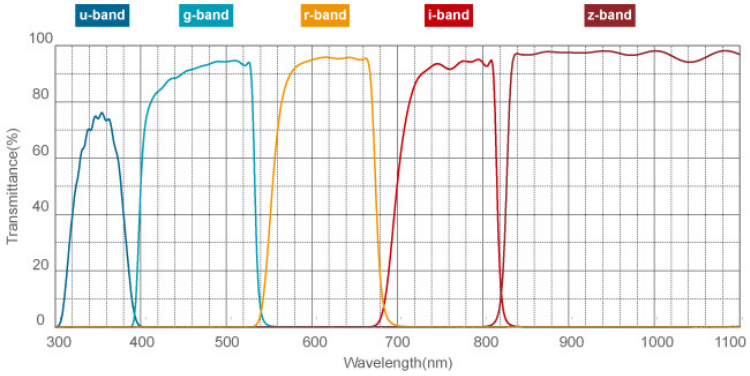


Figure 2.1: SDSS photometric bands and their absolute transmittance. (Fukugita et al. (2006))

It then went on to perform spectroscopic observations of stars, galaxies and quasars selected on the basis of the multi-band photometric observations, thus assembling an unrivaled dataset for the classification of different classes of astronomical sources. The SDSS survey is periodically extended by improving the resolving power of the high-performing spectroscopic telescopes. The most recent extension of the SDSS survey is the SDSS-IV that covers $\sim 14,555\ deg^2$ of the sky (Albareti et al., 2016). This SDSS-IV uses three main programs: the Extended Baryonic Spectroscopic Survey (eBOSS), Mapping Nearby Galaxies at APO (MaNGA) and Apache Point Observatory Galactic Evolution Experiment-2 (APOGEE-2). eBOSS has one core program, the Sloan Extended Quasar Emission-line and Luminous red galaxy (SEQUELS) which handles optical spectroscopy of astronomical sources. The Time Domain Spectroscopic Survey (TDSS) and Spectroscopic Identification of Erosita Survey (SPIDERS) are sub-programs of the eBOSS. The TDSS is used to select variable sources while SPIDERS is used for selecting X-ray sources (Albareti et al., 2016). APOGEE-2 is used to detect light in the infrared waveband. This reveals astronomical sources behind dusty regions and to provide calibrated element abundance measurements. MaNGA is used to map and observe kinematics of the nearby galaxies. Data Release 16 (DR16) is the final data release for the main cosmological program of the eBOSS and all raw and reduced spectra from that project are in DR16 (Ahumada et al., 2020).

2.2 WISE

The Wide-field Infrared Survey Explorer (WISE) is a mid-infrared survey that scans the entire sky using a 40 cm telescope, with a field-of-view of 47 arc-minutes, using four infrared bands (W1, W2, W3 and W4) centered at 3.4, 4.6, 12 and 22 μm, respectively, as shown below in Figure. (2.2). WISE is in a Sun-synchronous low Earth polar orbit and surveys the entire sky along great circles resulting in the entire sky being scanned in a period of a half a year (Wright et al., 2010)). WISE presents the results as images of the astronomical sources in different regions in the sky. The processing of these images is done as described by (Wright et al., 2010) and data is extracted from these images.

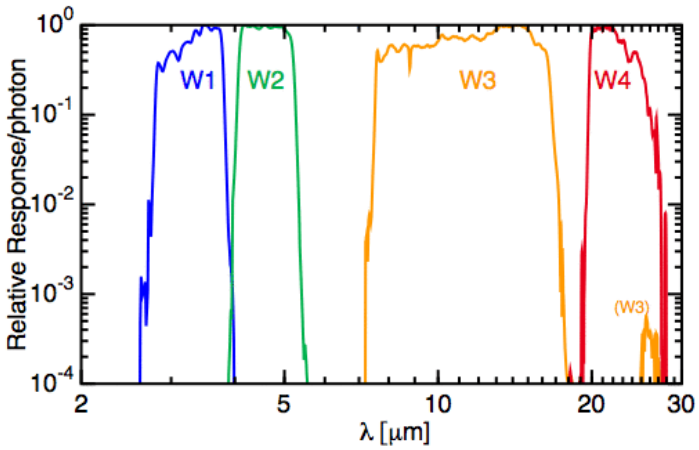


Figure 2.2: WISE photometric bands and their relative transmittance (Wright et al., 2010)

2.3 The SDSS-WISE DR16 Dataset

Using data acquired from the Wide-field Infrared Survey Explorer (WISE) and the Sloan Digital Sky Survey (SDSS), a large dataset of sources and their properties was determined and stored. Since WISE was a survey that included the entire sky and the SDSS was a survey of just a portion of the sky, some astronomical sources are catalogued in the dataset of both surveys. A SDSS-WISE dataset was produced by cross-matching sources by either by using pointers or matching properties that were similar. This SDSS-WISE dataset would be an incredibly large dataset with multi-wavelength measurements with magnitudes ranging from the visible (As is with the SDSS in Figure (2.1)) to mid-infrared (as seen in WISE in Figure (2.2)).

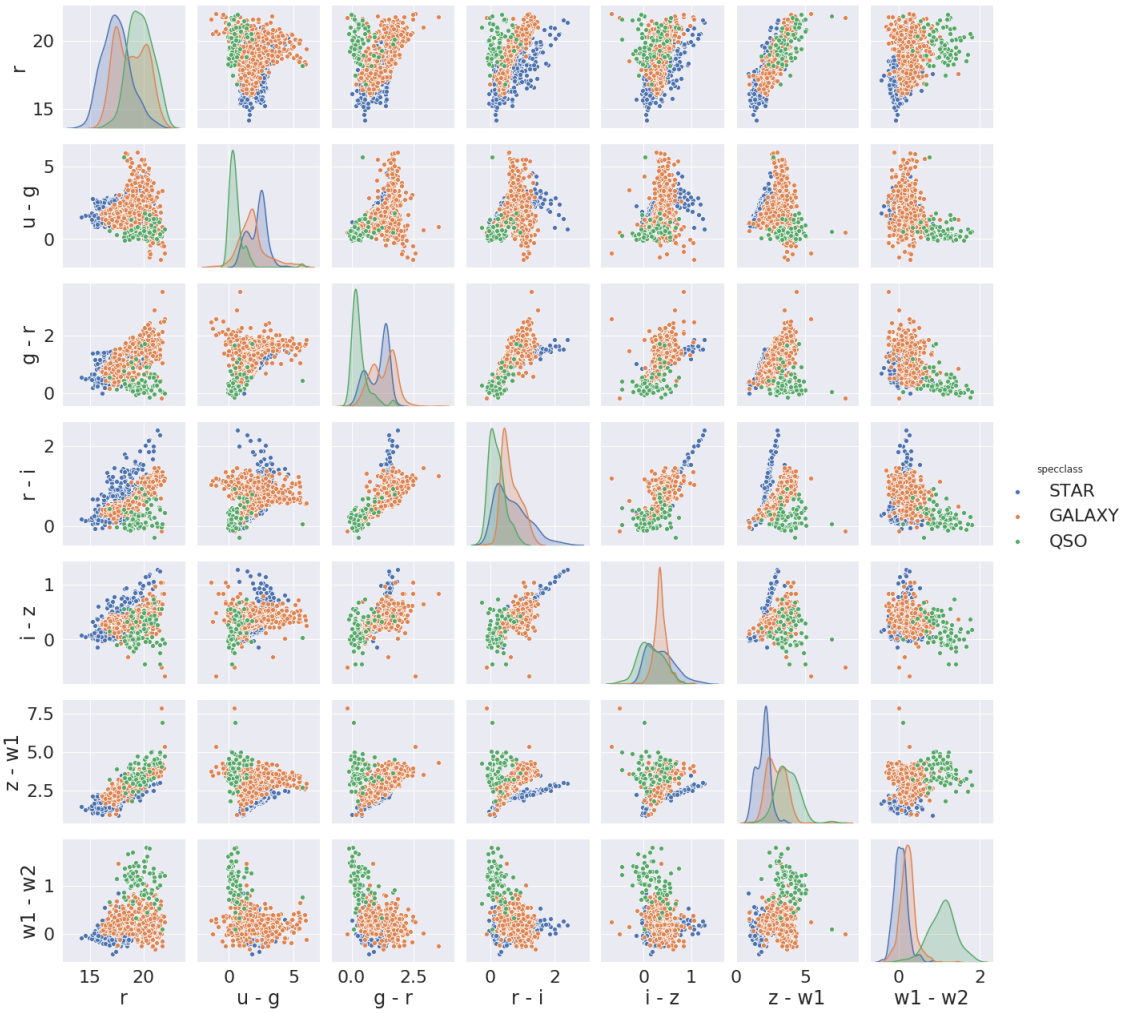


Figure 2.3: Colour plot of the spectral classes of sources under different colour filters

From Figure (2.3) we see that properties of stars, galaxies and quasars appear to merge and be similar under the same filters. Some sources are seen to have the same colour under spectroscopic measurements. Since the SDSS-WISE dataset is so large, machine learning techniques can be used to identify trends and better parameters to classify the astronomical sources.

3 Machine Learning

Machine learning is a branch of Artificial Intelligence (AI) that provides the ability to learn trends, patterns and experience without being explicitly programmed to do so. Supervised machine learning is used in this project as the system trains the machine using training dataset and then makes predicated output values for missing information. This type of machine learning may also test its predictions against its training dataset to find its errors and modify itself to reduce such errors (Ivezic et al., 2014). Classification in this paper was performed using the AstroML Python module, [1]. This module was used for analysing and visualising of astronomical data from the SDSS-WISE DR16. AstroML contains tools which are built on numerical modules like astropy, numpy, scipy, scikit-learn and matplotlib. These tools are necessary to perform data mining and machine learning algorithms. In this paper, data visualisation and loading were done using AstroML while machine learning algorithms were performed using the scikit-learn module. To see the performance of the classifiers which follow, classier methods are tested on SDSS-WISE data to investigate how well each classifier can differentiate the three classes (galaxies, stars and quasars). Features that were fed to the classifiers were $(r, u-g, g-r, r-i, i-z, z-w1, w1-w2)$ which are the colour computed from photometric magnitudes contained in SDSS-WISE DR16. A standard approach to test the performance of machine learning algorithms on the same dataset is used to draw conclusions. Performance of the algorithms are determined using precision and recall as explained by Ivezic et al. (2014).

$$precision = \frac{TP}{TP + FP} \tag{1}$$

Where TP (True Positive) is the number of correctly classified objects and FP (False Positive) are misclassified.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

Where FN (false negative) is the number of misclassified objects. It can be interpreted that precision is a measure of correctly classified objects whereas recall measures the fraction of classified to misclassified objects. Another more detailed measure is f1 score, defined by,

$$f1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

This a weighted average of precision and recall. Ideally, high precision and high recall will correctly classify objects.

3.1 Classifiers

The classifiers used in this paper were used based on the scikit-learn module. [2]

[1] `http://www.astroml.org/`

[2] `https://scikit-learn.org/stable/supervised_learning.html`

3.1.1 K-Nearest Neighbour

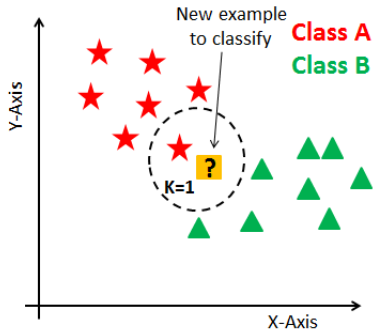
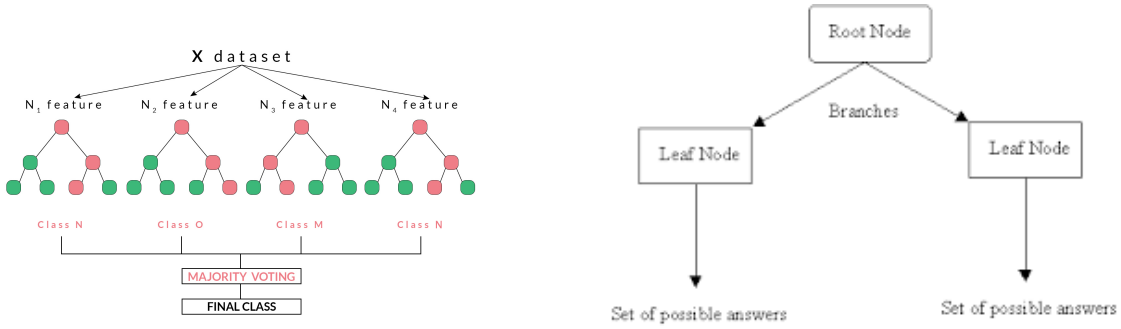


Figure 3.1: Visual Representation of the kNN Technique (Leuschke, 2020)

Instance learning algorithms like that of the kNN classifier, are based on the principle that when features contained in a dataset are nearby each other, they have similar properties (Kotsiantis., 2007). The kNN algorithm uses features to compute the distance metric in such a way that the distance between similar features of the same class is minimized, while distances to features of other classes maximized as shown above in Figure 3.1. In most cases, different distance metrics (Euclidean, Minkowsky, Chebyshev) and k (number of nearest neighbours), are the parameters that affect the performance of the kNN algorithm.

3.1.2 Random Forest



(a) (b)

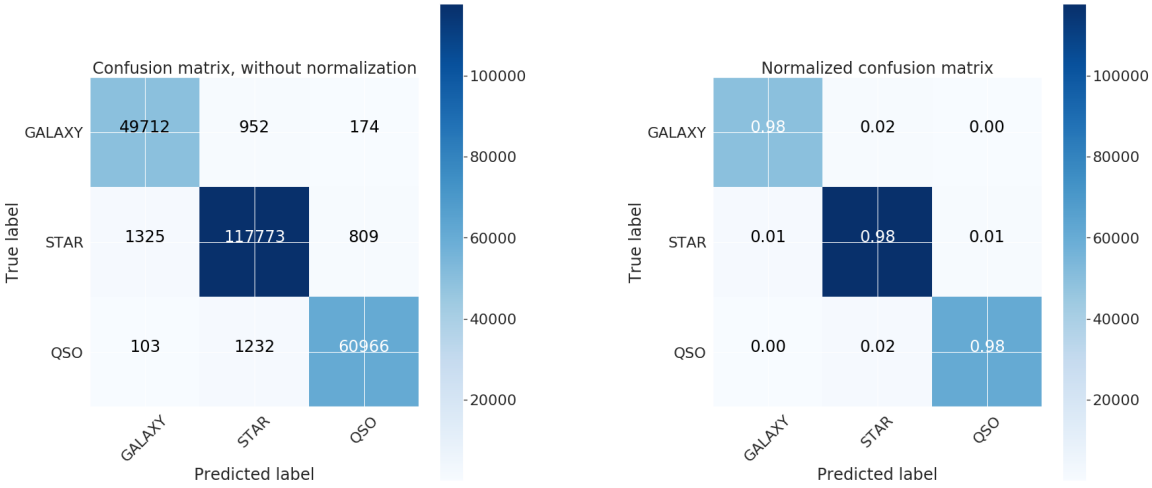
Figure 3.2: Visual Representation of the Random Forest Technique a) (Tahsildar, 2020), b) `https://kgpdag.wordpress.com/`

Logic based algorithms like that of the RF classifier is an example of ensemble classifier (Breiman., 2001). This uses decision trees to perform classification. The unknown sources are classified by sorting them based on their feature values and sources are places in decision trees. As seen in Figure (3.2b), each node represents one of the features associated with the unknown source to be classified. Unknown sources are classified from the root node until they reach the leaf node where final classification is determined. RF generates multiple decision trees based on the features of the source data. Randomly, each decision tree is generated from a randomly selected subset of all the features of the source data. The total number of trees and the maximum amount

of features after splitting a tree at each level are the most important parameters that affect the performance of the RF algorithm (Ivezic et al., 2014).

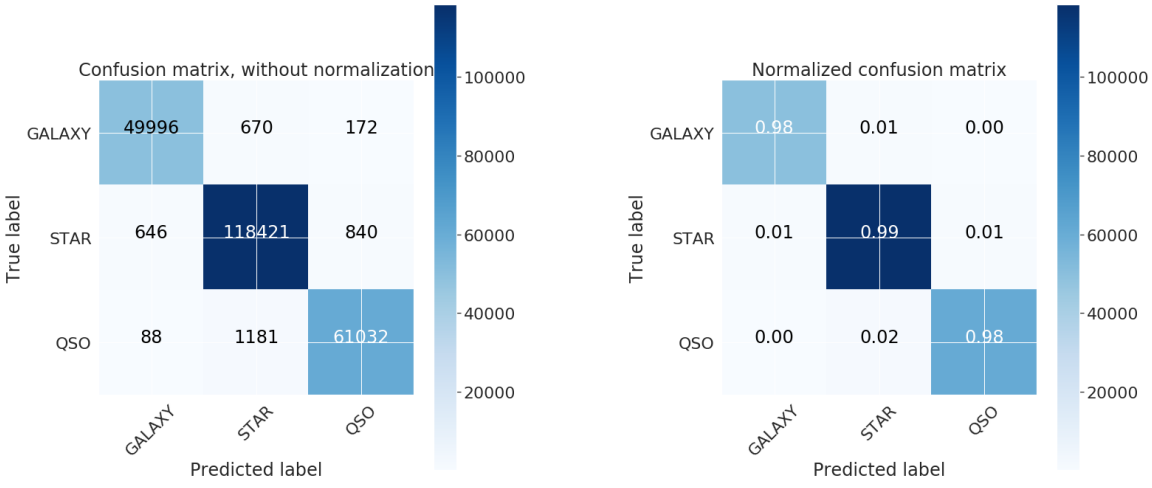
3.1.3 kNN vs. RF

A good indicator of the how well classification occurred is the confusion matrix as seen below in Figure (3.3) and Figure (3.4). The confusion matrices shows how well astronomical sources are predicted to their true value. Normalization confusion matrices are confusion matrices where each row element is divided by the total number of true labelled sources.



(a) (b)

Figure 3.3: Confusion matrices of kNN classifier: (a) without normalization; (b) with normalization.



(a) (b)

Figure 3.4: Confusion matrices of RF classifier: (a) without normalization; (b) with normalization.

Observing entries in the normalised confusion matrices indicate that for the kNN

method, more galaxies were classified as stars whereas in the RF method, more stars were classified correctly. When looking at the two techniques, the RF classifier stands out, being more precise in classification of astronomical sources. The kNN classifier produces an accuracy (f-score) of 98.03% while the RF technique had an accuracy (f1-score) of 98.46%. However, the kNN classifier does outperform the RF classifier in speed. The kNN classifier took 46 seconds to run while the RF classifier took 187 seconds.

4 Hyper-Parameter Optimization

Due to the RF classifier being more accurate, it was decided to optimize the hyper-parameters of this technique. The following hyper-parameters of the random forest classifier method are analysed to find the best possible values which will increase the accuracy of the classification. The method in which the following hyper-parameters are analysed is by observing validation plots which find the best accuracy for any given parameter range (Meinert, 2019).

4.0.1 n_estimators

This hyper-parameter is seen as the number of trees in the forest or the total number of decisions that will be used to classify each astronomical source. By default, this is set to 100.

4.0.2 max_depth

This hyper-parameter is the maximum depth of each tree. By default this value is set to 'None' nodes of a decision tree are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

4.0.3 min_samples_split

This hyper-parameter is the minimum number of samples required to split an internal node. By default it is set to 2.

4.0.4 min_samples_leaf

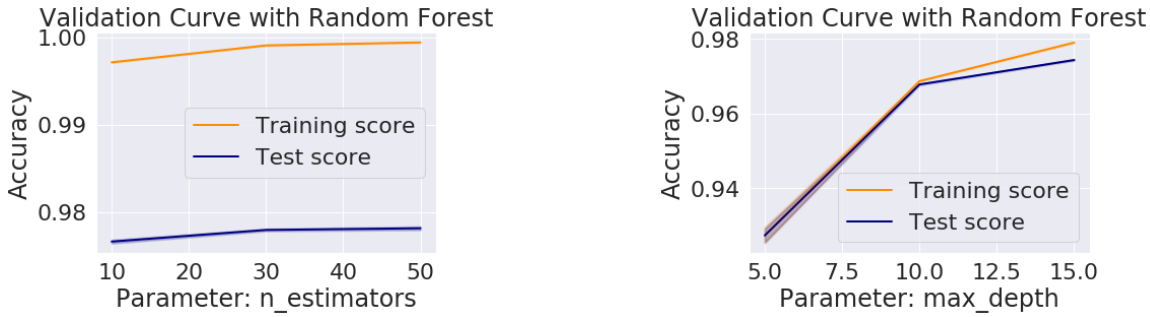
This hyper-parameter is the the minimum number of samples required to be at a leaf node. "A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches." (ScikitLearn, 2020).

5 Results

Initially, the SDSS-WISE dataset was analysed using the Random Forest Classifier with the following hyper-parameters:

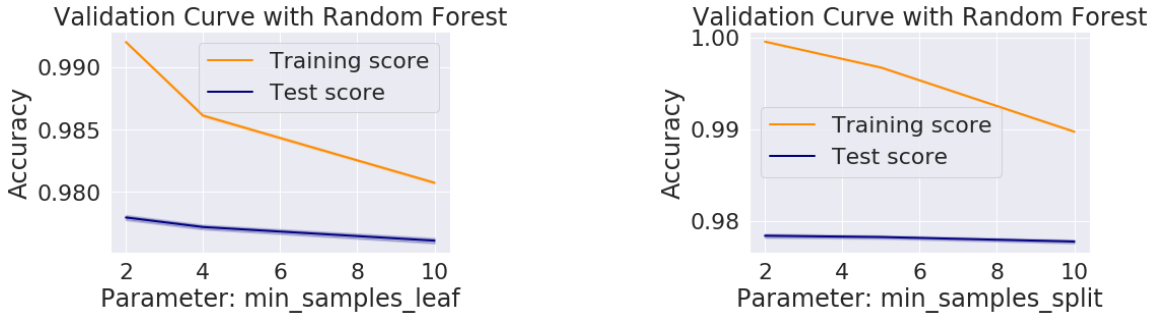
Hyper-Parameter	Initial Values
n_estimators	30
max_depth	none
min_samples_split	2
min_samples_leaf	1

The first approach is to see how the accuracy of the model changes when each parameter is changed independently. For this, a validation plot is implemented, which runs the model against a predefined parameter range and plots the accuracy of the classifier for each predefined parameter point. Upon doing so, the following Figures (5.1) and (5.2) were produced:



(a) (b)

Figure 5.1: Validation Curve plots for hyper-parameters that show considerable improvement in accuracy of the RF classifier.



(a) (b)

Figure 5.2: Validation Curve plots for hyper-parameters that show a decline in accuracy of the RF classifier.

From the validation plots, it can be seen that only 2 hyper-parameters are really of importance, namely being n_estimators and max_depth. From Figure (5.1a) we see that increasing n_estimators increases the training score and test score accuracies. It appears that this converges to 100% as the hyper-parameter increases. However, this

means that more processing time is required. It can also be seen that max_depth also increases abruptly for high values of max_depth. Because the training and test score accuracies decrease for higher values for the min_samples_leaf and min_samples_split, it can be concluded that only the first 2 hyper-parameters are of importance as they increase accuracy. The random forest was then run with the best found hyper-parameters. The following outcome was seen:

Hyper-Parameter	**Initial Values**
n_estimators	50
max_depth	15
min_samples_split	2
min_samples_leaf	1

Table 5.1: The best hyper-parameters after optimization

The accuracy (f1-score) for the RF classifier after optimization was determined to be 98.36%. This is slightly less than the 98.46% before optimization. A look into the feature importance accuracies show:

	Precision	**Recall**	**f1-score**	**support**
SFGs	0.8470	0.8429	0.8449	50838
STARs	0.9189	0.9204	0.9196	119907
QSOs	0.9483	0.9492	0.9487	62301

Table 5.2: Accuracy for RF feature selection before optimization

	Precision	**Recall**	**f1-score**	**support**
SFGs	0.8446	0.8439	0.8443	50838
STARs	0.9190	0.9189	0.9189	119907
QSOs	0.9476	0.986	0.9481	62301

Table 5.3: Accuracy for RF feature selection after optimization

The RF classifier has a feature selection accuracy of 91.11% before optimization and 91.05% after optimization. However, when looking at the feature importance plots before and after optimization as seen below in Figure (5.5), it can be seen that the importance of some features shift. 'w1-w2' remains a 1st priority feature whose importance becomes higher. 'u-g' shifts importance from 2nd priority to 4th priority with a very noticeable drop in importance. The 'i-z' and 'r' features switch place places too. This is an indication that when accuracy of the RF classifier increases, the importance of certain features decrease while the importance of other features increase.

Figure 5.3: Before Optimization

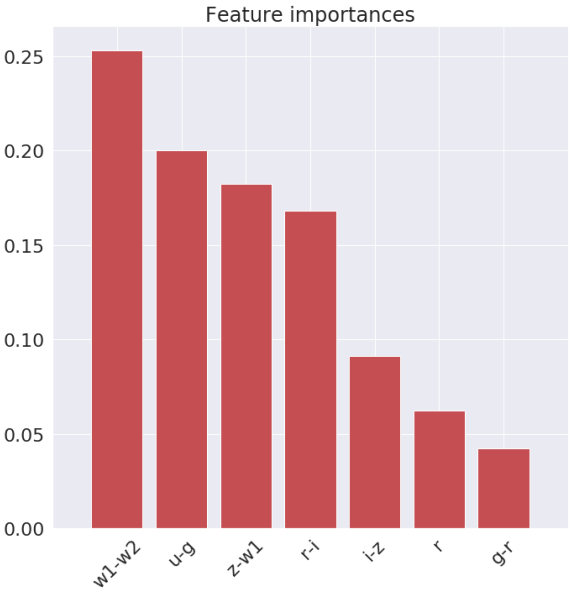
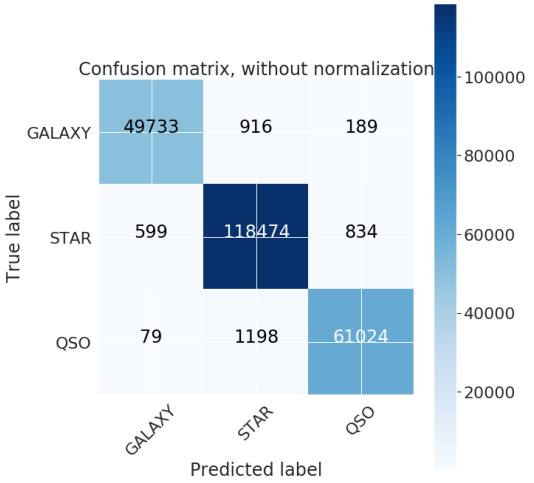


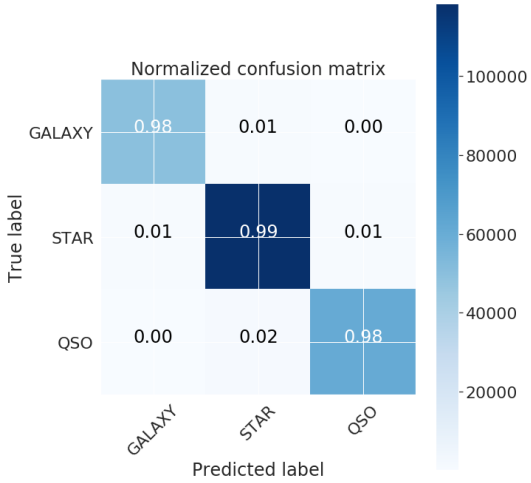
Figure 5.4: After Optimization

Figure 5.5: Plots of feature importance showing more resolved colour magnitudes

Once again, the confusion matrices are analysed to show really how well classification has occurred after optimization of the hyper-parameters.



(a)



(b)

Figure 5.6: Confusion matrices of RF classifier after optimization: (a) without normalization; (b) with normalization.

The confusion matrices in Figures (3.4) are compared to the confusion matrices in Figures (5.6). The normalized matrices in both figures look exactly the same. This is expected since the accuracy RF classifier before and after optimization does not have any large accuracy differences. However, upon close inspection of the matrices without normalization, it can be seen that after optimization, more sources were actually predicted incorrectly but because such a large dataset is used, the inaccuracies aren't seen in the normalized confusion matrices.

Since results show that the best hyper-parameters do not necessarily create the highest accuracy for RF classification, a new approach is attempted. This technique is called the Grid Search Cross Validation (CV) and uses a more rigorous approach to determining the best hyper-parameters. This CV method uses all possible combinations of predefined parameter ranges, to find the best possible hyper-parameters. Even when using this approach, the same results were seen.

The max_depth parameter may be the cause for the collective decrease in f1-score for the RF classifier. This is because this parameter was initially set to 'None'. When max_depth is set to the default parameter, the RF decision trees are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. By setting a limit for max_depth creates time

6 Conclusion

In conclusion, increasing the n_estimators and max_depth hyper-parameters collectively does not increase the f1-score accuracy of RF classifier. However, it is seen that the importance of each feature used to classify any astronomical source changes based on the accuracy of the model. It is therefore recommended that the only hyper-parameter that needs to be changed to increase accuracy is the n_estimator (number of decision trees). When this parameter is maximised, processing time is sacrificed and the increase in accuracy only changes by a small bit as seen in the validation curve in Figure (5.1a). The RF classifier appears to be the best way to classify any large set of data where the only hyper-parameters that make a difference are n_estimators and max_depth. Upon using larger values of these hyper-parameters, the total time taken to compute classifications increase considerably. Even so, using the RF classifier with minimal amount of decision trees like that at it's default can produce the best possible accuracies. The RF classifier shows it's dominance in classifying and proves to be a necessary tool for extremely large datasets like those processed by the Square Kilometer Array (SKA), Large Synoptic Survey Telescope (LSST) and more. With upcoming surveys from the SKA and LSST, petabytes of observational data will be produced and will need to be analysed at a much faster rate than astronomers. For this, the RF classifier seems to be a good solution.

References

Ahumada R., *The 16th Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra*, *Astrophysical Journal*

Albareti F. D., SDSS Calloboration., Allende P. C., Almeida A., Anders F., Anderson S., Andrews B. H., Aragon-Salamanca A. and et al., 2016., *ArXiv e-prints*, 1608.02013

Bass M., *Handbook of Optics Volume II - Devices, Measurements and Properties*, McGraw-Hill, 1995, pp 847

Breiman L., *Machine Learning*, Kluwer Academic Publishers, 2001., 45:5

Carroll B. W., Ostlie A. D., *An Introduction to Modern Astrophysics*, Pearson Education Limited, 2014, pp 230

Djorgovski S. G., Mahabal A. A., Drake A. J., Graham M. J., Donalek C., 2013., *ArXiv e-prints*, 1203.5111

Fukugita M., Ichikawa T., Gunn J. E., Doi M., Shimasaku K., Schneider D. P., 2006., *AJ*, 111, 1748

Gunn J. E., *The 2.5 m Telescope of the Sloan Digital Sky Survey*, 2006, *Astronomy Journal*, 131, pp 2332-2359.

Ivezić Ž., Conolly A. J., VanderPlas J. T., Gray A., *Statistics, Data Mining, and Machine learning in Astronomy*, Princeton University Press, 2014, pp 230

Kotsiantis S. B., *Supervised Machine Learning: A Review of Classification Techniques*, Informatica, 2007, 31:249-268

Leuschke A., *KNN Classification using Scikit-learn*, `https://morioh.com/p/620a3a2faf6c`, Accessed Online: 1 Nov 2020

Meinert R., *Optimizing Hyperparameters in Random Forest Classification*, `https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6`, Accessed Online: 15 Oct 2020

Scikit Learn, *sklearn.ensemble.RandomForestClassifier*, `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`, Accessed Online: 1 Nov 2020

Seeds M. A., Beckman D., *Foundations of Astronomy*, Brooks/Cole - Cengage Learning, 2011, pp 230

Tahsildar S., *Random Forest Algorithm In Trading Using Python*, `https://blog.quantinsti.com/random-forest-algorithm-in-python/`, Accessed Online: 10 Nov 2020

Wright E. L., Eisenhart P. R. M., Mainzer A. K., Ressler M. E., Cutri R. M., and et al., 2010., *Astronomy Journal*, 1868-1881

Appendix

The following link attached shows the python code and processes used for analysis submitted in this paper:
`https://github.com/yaaseenjones/honours_project.git`