UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

NASSP
NATIONAL ASTROPHYSICS AND SPACE SCIENCE PROGRAMME
www.star.ac.za • nassp.ukzn.ac.za

Machine Learning For Radio Source Host Galaxy Classification

Praisegod Thutho Ndlovu 216001741

Supervisor: Dr Khadija El Bouchefry (SARAO)

January 12, 2021

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

Department of Physics

Preface

The research detailed in this dissertation was completed by the candidate while based in the Discipline of NASSP, School of Chemistry and Physics of the College of Agriculture, Engineering and Science, University of KwaZulu Natal, Westville campus, South Africa. The research was financially supported by National Astrophysics and Space Science Programme (NASSP).

The contents of this work have not been submitted in any form to another university and, except where the work of others is acknowledged in the text, the results reported are due to investigations by the candidate.

signed: **Dr Khadija El Bouchefry** date: **12/01/2021**

Plagiarism

I, Praisegod Thutho Ndlovu, declare that:

1. The research detailed in this thesis, except where something else demonstrated or acknowledged, is my unique work.

2. This thesis has not been submitted for any degree or examination to any other college/university.

3. This thesis does not contain other people information, pictures, charts or other data, unless particularly acknowledged as being sourced from other people.

4. This thesis does not contain other people writing, unless particularly acknowledged as being sourced from other researcher.

5. Where I have used material for which publications followed, I have indicated in detail my role in the work.

6. This dissertation is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included.

7. This thesis does not contain content, illustrations or tables replicated and glued from the Web, unless particularly acknowledged, and the source being detailed within the thesis and within the References sections.

signed: **Praisegod Thutho ndlovu** date: **12/01/2021**

Abstract

We discuss the optical data from Sloan Digital Sky Survey(SDSS) DR9 and infrared data from Wide-field Infrared Survey Explorer(WISE). We then investigate the combination of colors with magnitude to classify between Star Forming Galaxy(SFG) and Active Galactic Nuclei(AGN) using supervised machine learning algorithms. We have applied the machine learning algorithms, K-Nearest Neighbour(KNN) and Random Forest(RF) to perform classification between SFG & AGN using Unified Radio Catalog(URC) data and the cross matched data separately. Both algorithms (KNN and Random forest) are trained with 90% of our data, Also we have used 10% for testing our algorithm. For URC the training data contain 17837 sources(1 1310 SFG & 6 527 AGN), also 1982 sources (1 241 SFG & 741 AGN) for testing. The cross matched data contain 8 829 training objects(5 111 SFG and 3 718 AGN), and we have used 982 objects(566 SFG and 416 AGN) for testing our algorithm. The predicted number SFG and AGN are presented in the tables based on the data used. The Random forest approach is more accurate than KNN algorithm. Both Random forest and KNN have distinguished SFG with a higher accuracy (or more SFG), independent of the decision of the input parameters.

Acknowledgments

I would like to express my very great appreciation to my supervisor: Dr Khadija El Bouchefry for her important and useful recommendations during the arranging and improvement of this research work. Her eagerness to grant her time so generously it is very much appreciated.

I would also like to thank Prof. Sivakumar Venkataraman, for his advice and help in keeping my progress on schedule.

My grateful thanks are also extended to my colleagues and friends who also helped me along the way.

I would also like to extend my thanks to NASSP & NRF for financially help it really helped me to do my research smoothly.

Finally, I wish to thank my parents for their support and encouragement all through my study.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Machine Learning (ML)

Machine Learning is the field of getting computers to learn and behave like people do, and upgrade their learning over time in autonomous fashion, by providing them with data and information in the form of observations and real-world interactions.
Machine learning algorithm come in different flavors but it can be grouped into three groups reinforced, supervised, and unsupervised techniques. In this project we will use the supervised machine learning to classify host galaxies.
In Supervised learning, we use a labelled data in order to train our algorithm. There are two types of Supervised Learning techniques: Regression(The goal is to predict continuous values, e.g. home prices) and Classification(The goal is to predict discrete values, e.g.True/False). There are many different types of supervised learning algorithms but in this project we will use K-Nearest neighbors(KNN)(which is used for both classification and regression) and Random forests(which is also used for both classification and regression).
The application of Machine learning it is most useful to astronomers. Machine learning can process data faster than other techniques, it can also analyse that data without being told what to do (Ivezić et al., 2019; Kremer et al., 2017; Heinis et al., 2016).

1.2 Galaxy

A galaxy it's a large collection of dust, gas and millions to trillions of stars held together by gravity. It is made up of dead and living stars. The well known galaxy which is closed to our planet (Earth) it's called a Milky Way galaxy. The Milky Way galaxy it contain a supermassive black hole in the middle (see figure1). In fact approximately all galaxies contain a black hole at the center. There are many galaxies in the universe grouped in their sizes, shapes and colour (Ebeling et al., 2014).

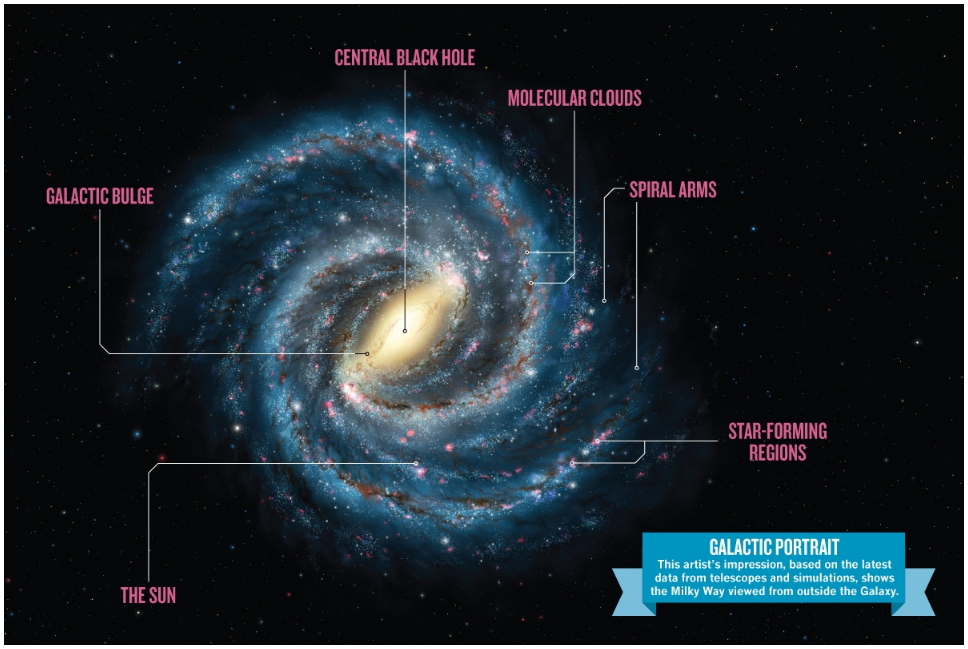
CENTRAL BLACK HOLE
MOLECULAR CLOUDS
GALACTIC BULGE
SPIRAL ARMS
STAR-FORMING REGIONS
THE SUN
GALACTIC PORTRAIT
This artist's impression, based on the latest data from telescopes and simulations, shows the Milky Way viewed from outside the Galaxy.

Figure 1: The Milky-Way Galaxy

1.3 Star Forming Galaxy

Star formation is the process by which dense regions within molecular clouds in star-forming regions, collapse and form stars.
The star formation is one of the most challenging problem in astronomical researches. The evolution of stars lead to evolution of stellar systems i.e. galaxies and galaxy cluster. The star formation is an ongoing process. The stars are continuing being formed in the galaxies. The stars are formed within the cloud of dust and scattered throughout the most galaxies. As the cloud collapses ,the material at the center start to heat up, it is this hot core at the heart of the collapsing cloud that will one day become the star. Not all of this materials end up being part of the star, the remaining dust may become asteroids, planets, comets or it may remain as a dust. The bigger the galaxy the more stars are being formed (Tacconi et al., 2010; Guo et al., 2012; Sparke & Gallagher, 2006).

1.4 Active Galactic Nuclei

Active galactic nuclei (AGN) is a region at the center of the galaxy that has higher than normal luminosity i.e. It is brighter than the rest of the galaxy. It is found only on active galaxies. AGNs are not in the black hole it must be around supermassive black hole since there is no light exist in black hole. Supermassive black hole is surrounded by accretion disc(the structure composed of gas, dust orbiting around), since the accretion disc is rotating, as it start to heat up becoming hot it shine more brightly than the rest of the galaxy, then it create jets that fire out particle approximately with speed of light (Osterbrock & Ferland, 2006; Kauffmann et al., 2003).

1.5 Redshift

Redshift is a situation where electromagnetic radiation from an object undergoes an increase in wavelength.
Astronomers use the term redshift to describe how far the objects are really. The object in the universe are redshifted. The further away the galaxy is the faster is moving, the faster the galaxy is moving the more it is reshifted. So by measuring the amount of redshift is a great way to measure the distance of a galaxy relative to earth (doppler redshift). By using the cosmic red shift the astronomers are able to measure the distance of the galaxies. We use the two method to find redshift which are **Spectroscopic redshift** and **Photometric redshift**.

The foremost exact way to measure redshift is by using spectroscopy. Astronomers can see at the spectra made by different elements and compare these with the spectra of stars. In case the absorption or emission lines they see within the star's spectra are moved, they know the object is moving either away from us or towards us (Rahman et al., 2015; Mobasher et al., 2007a).

For distant objects such as quasars, some of which are too faint to be watched by spectroscopy, astronomers estimate the photometric redshifts. In this case the astronomers watch the peak brightness of the object through different filters. For a redshifted object its peak brightness will show up through filters towards the red end of the spectrum (Mobasher et al., 2007b; Loh & Spillar, 1986).

2 Materials and Methods

2.1 Data

The data used in this project contains different wavelengths in order to better train the machine learning algorithms: Optical data from SDSS and infrared data from WISE & 2MASS.

2.1.1 Unified Radio Catalogue (URC)

The catalog contain millions of radio sources, created by combined large area of radio and optical surveys, which are Green bank 6 Centimetres (GB6(6cm)), Faint Images of the Radio Sky at 20 Centimetres (FIRST(20cm)), NRAO VLA Sky Survey 20 Centimetres (NVSS(20cm)), Westerbork Northern Sky Survey 92 Centimeters (WENS(92cm)), Very Large Array (VLA) Low-Frequency Sky Survey redux 4 Metres (VLSSr(4m)), and Sloan Digital Sky Survey(SDSS (optical)). Then finally the catalog has 2 866 856 objects within the sky of the North of 40^{deg} declination secured by the NVSS. The region enclosed by all five radio surveys and by the SDSS photometric study has the common region of 3269 deg^2. The one third of its data has been detected optically (by SDSS), while more than 160,000 sources detected at 20 cm. The catalogue matches FIRST with SDSS DR9 which overlap almost completely.

The class of SDSS spectroscopic source can be found from the spectrum by the spectroscopic survey. Where our classes are STAR, GALAXY and QSO. The subclass of SDSS is further classified, where QSO and Galaxy have subclass of BROADLINE, STARBURST, AGN or STARFOMING.

For SDSS photometry, parameters labelled near correlate with the closest SDSS photometric match within 30".

The URC contains 2 866 856 sources, which include 12551(0.44%) STARFOMING and 7268(0.25%) AGN which are known.

For each sort of object (SFG, AGN) it is possible to recognize the object dependent on the run of estimations of the picked parameters (Clarke et al., 2020a).

2.1.2 Sloan Digital Sky Survey (SDSS)

By measuring the redshifts of a million galaxies, the Sloan Digital Sky Survey will provide a three-dimensional picture of our local neighborhood of the universe.
The latest data release for SDSS is DR16 which include all previous data release i.e.DR15,DR14,DR13,DR12...
With wave bands u, g, r, i & z of wavelengths 3551, 4686, 6165, 7481 & 8931. SDSS DR16 covers an area of 14555 square degrees. The SDSS DR16 contains data of approximately 430000 in APOGEE. The eBOSS covers 7500 square degrees, with approximately 4500 square degrees in the North galactic cap and 3000 square degrees in the South galactic cap.
DR16 contain six types of data which are images, optical spectra, infrared spectra, IFU spectra, stellar library spectra, and catalog data. But we will use optical data(SDSS) (Ahumada et al., 2019; Sharma et al., 2019; Ball et al., 2006).

2.1.3 AllWISE Source Catalog

The AllWISE Source Catalog[1] it has astrometry and photometry for 747,634,026 sources identified in AllWISE Atlas Intensity Images. The Informative Supplement to the WISE All-Sky Data Release Products is a common guide to the users of WISE data (Cutri et al., 2012). The magnitudes, positions, astrometric and photometric uncertainties, flags showing the reliability and nature of the source characterizations, and the relationship with 2MASS Point and Expanded

[1] https://wise2.ipac.caltech.edu/docs/release/allwise/

Source Catalog sources are introduced for each source.
The photometric accuracy in each of the four bands has upgraded, since the faint source flux bias it is improved and background evaluation has been made more powerful.
Astrometry precision is better as a result of the amendment for the appropriate motion of 2MASS astrometric reference stars in 11 years between the two overviews, and the fuse of the different free source estimations in the image cover regions into the astrometric arrangements.
The estimations of the apparent motion of sources are accommodated the first run through, and improved source flux variability measurements have been processed (Shajib & Wright, 2016; Kurcz et al., 2016)

2.1.4 Wide-field Infrared Survey Explorer (WISE)

The Wide-field Infrared Survey Explorer (WISE) is mapping the whole sky following its launch on 2009 December 14. WISE map the whole sky in four infrared wavelength 3.4, 4.6, 12 and 22 μm which is represented as W1, W2, W3, W4 respectively. WISE produces an image atlas in its four colors that will be a stack of all the multiple frames covering each part of the sky. The WISE All Sky Data Release include all data collected from 7 January 2010 to 6 August 2010. This data Release include an Atlas of 18240 match-filtered, calibrated and codded image sets, a source catalog containing positional and photometric information for over 563 million objects detected on the WISE images (Koenig & Leisawitz, 2014; Mainzer et al., 2011; Wright et al., 2010).

2.1.5 Topcat & Python

TOPCAT is an intelligently graphical viewer and editor for tabular information. Its point is to supply most of the facilities that astronomers require for analysis and control of source catalogues and other tables, in spite of the fact that it can be utilized for non-astronomical information as well. It gets it a number of diverse astronomically vital formats (counting FITS, CDF and VOTable) and much formats can be included. It is particularly great at intuitively investigation of huge tables (Taylor, 2005).

And Python is an translated, object oriented, high level programming language with energetic semantics. Its high level built in information structures, combined with energetic writing and dynamic authoritative, make it exceptionally appealing for Quick Application Improvement, as well as for utilize as a scripting to associate existing components together. Python is simple, easy to memorize language structure emphasizes coherence and so decreases the cost of program upkeep. Python supports modules and packages, which empowers program measured quality and code reuse. The Python interpreter and extensive standard library are accessible in source for complimentary for all major stages, and can be openly distributed (Rossum, 1995).

This will be used together to analyse astronomical data i.e. accessing tables, plotting some graphs and doing analysis.

2.2 The Machine Learning Toolkit

2.2.1 AstroML

AstroML[2] is a Python module for data mining and machine learning built on numpy, matplotlib, scipy, scikit-learn, and astropy. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in Python (VanderPlas et al., 2014). The purpose of AstroML is to supply an open store for quick python usage of statistical procedure commonly used in astronomy, and to supply available examples of astrophysical data investigation utilizing methods created within the fields of statistics and machine learning. A quality of astroML lies within the truth that the python code utilized to download, handle, analyze, and plot the information is all open-source and openly accessible (VanderPlas et al., 2012).

2.2.2 Scikit-learn

Scikit-learn (Sklearn)[3] is a library in Python that provides many unsupervised and supervised learning algorithms. It gives a choice of productive tools for statistical modeling and machine learning including classification, clustering and regression through a consistence interface in Python. Scikit-learn is built upon NumPy, SciPy and Matplotlib in python. The library is focused on modeling data. It is not focused on loading, manipulating and summarizing data. (Pedregosa et al., 2011).

2.3 Classification

Classification is a type of supervised machine learning. It indicates the class to which data have a place to and is best used when the yield has finite and discrete values. There are a number of classification models. Classification models include K-Nearest Neighbour(KNN), decision tree, random forest, gradient-boosted tree, logistic regression, multilayer perceptron, one vs rest, and Naive Bayes. The model that we are going to use is K-Nearest Neighbour Classifier (KNNC) and Random Forest Classifier (RFC) (Ivezić et al., 2019).

2.3.1 K-Nearest Neighbour Classifier (KNNC)

K-Nearest Neighbor algorithm is utilized to relegate a data point to clusters based on closeness estimation. It employments a supervised strategy for classification.The steps that we are following when writing a k-means algorithm are:

1. We choose k-Nearest Neighbour

[2] https://www.astroml.org/

[3] https://scikit-learn.org/stable/index.html

2. We calculate the distance of the nearest neighbour
3. Sort the distance and find nearest neighbour by looking the k-th minimum distance
4. In these k-Nearest Neighbour, check the number of the data points in each class
5. Allot the new data points to a class for which the number of the neighbour is the greatest
6. Now our model is ready (Wang, 2011).

There is no defined statistical methods that is used to find the value of k.
While utilizing k-means method we will have to be discover the ideal k value for which the framework gives more precise results. This may effectively be done by allotting k-values from a run (i.e.1-20) and comparing outcomes. We are plotting a graph of accuracy vs k-value, then we are ready to study of how the accuracy shifts over distinctive k-values, this will offer assistance to recognize a k-value that yield the most excellent results (Subhashini, 2009a).
We can also derive a plot of error rate vs k-value indicating values in a characterized range. At that point select the k-value having a least error rate.
The low value of k is sensitive to outliers and a higher value of k is more reliable to outliers as it considers more voters to decide prediction.
So it is wise to select a larger k if different classes in training set are widely separated, since large values of k reduce the effect of noise on the classification, but make boundaries between classes less distinct.

There are many distance functions to calculate the distance of the nearest neighbour but Euclidean is the most commonly used measure. It is used when data is continuous. So we will use Euclidean distance to find the distance between test data and trained data. We measure the distance on a straight line from point (x_1, y_1) to (x_2, y_2).

Euclidean distance,

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

Manhattan distance,

$$d(x,y) = \sum_{i=1}^{m}|x_i - y_i|$$

Minkowski distance,

$$d(x,y) = \left(\sum_{i=1}^{m}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

The idea to use distance measure is to find the distance between new sample and training cases and then finds the k-closest SFG/AGN to unknown SFG/AGN (Altman, 1992; Ye, 2013; Polsterer et al., 2013; Borne, 2013; Hastie et al., 2009).

2.3.2 Random Forest Classifier (RFC)

Random Forest is a learning technique that works by building numerous decision trees. The steps indicating how random forest works are:

1. We begin with the determination of random tests on a dataset given.
2. Then the algorithm can build a decision tree for each test. At that point it will get the prediction result from each decision tree.
3. Then voting will be done for each anticipated outcome.
4. We finally select the foremost voted prediction result as the ultimate prediction result (Gao et al., 2009).

3 Discussion and Results

The SFG and AGN can be essentially distinguished from their spectrum. So we have decided to use photometric parameters in classifying between SFG and AGN. The model magnitude values of an object at a different wavelengths will be used for classification of SFG and AGN because this values are at distinctive points within the spectrum. The color values will moreover offer assistance in the classification since these will be better parameters for contrasting the spectrum of the objects.

There is various sky surveys that are being controlled by different groups which helps us to gather information about the celestial objects. This surveys are SDSS which assembles multi-wavelength data on five distinctive bands, FIRST survey have radio data,and WISE/2MASS has infrared data. All these surveys use different coordinate systems to map the sky. The SDSS employ five filters which are, Ultraviolet(u) 3543Å, Green(g) 4770Å, Red(r) 6231Å, Near Infrared(i) 7625Å, and Infrared(z) 9134Å. The u,g,i,r & z values and the names of objects from SDSS database will be the five essential parameters for classification of SFG and AGN. Also based on these five filters we derive the color values (i.e. g-r, u-g, r-i, g-r, i-z, r-i) to further improve the quality of our classification (Subhashini, 2009b).

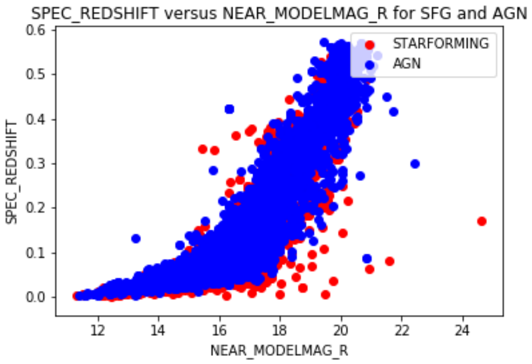

Figure 2: SPEC REDSHIFT vs NEAR MODELMAG R for both SFG and AGN

In figure 2, the plot of the spectroscopic redshift of our SFG and AGN classification in the training set as a function of modelmagnitude r. The SFG are shown as red dots, while AGN are in blue. The graph shows that as the magnitude increases also the spectroscopic redshift increases as expected.

In astronomy, the color is defined by the difference between the two magnitudes. The sky surveys select the filters to see an extensive extend of colors, while concentrating on the colors of fascinating celestial objects (Subhashini, 2009b).
The classification sources utilizing only photometry in numerous wavebands, & naming them by looking their colors is very quick, compared to spectroscopy and multi-wavelength observations (due to complexity of getting point by point surveying millions of individual sources is time expending). The foremost widely used colors for source classification are u-g and g-r, which is presented below (Clarke et al., 2020b). We use the u-g, g-r, r-i, and i-z colors to demonstrate photometric classification of the objects. The magnitudes used are the model magnitudes , while the true redshift measurements come from the spectroscopic pipeline.

Astronomical objects of equivalent course results in having comparable color values.

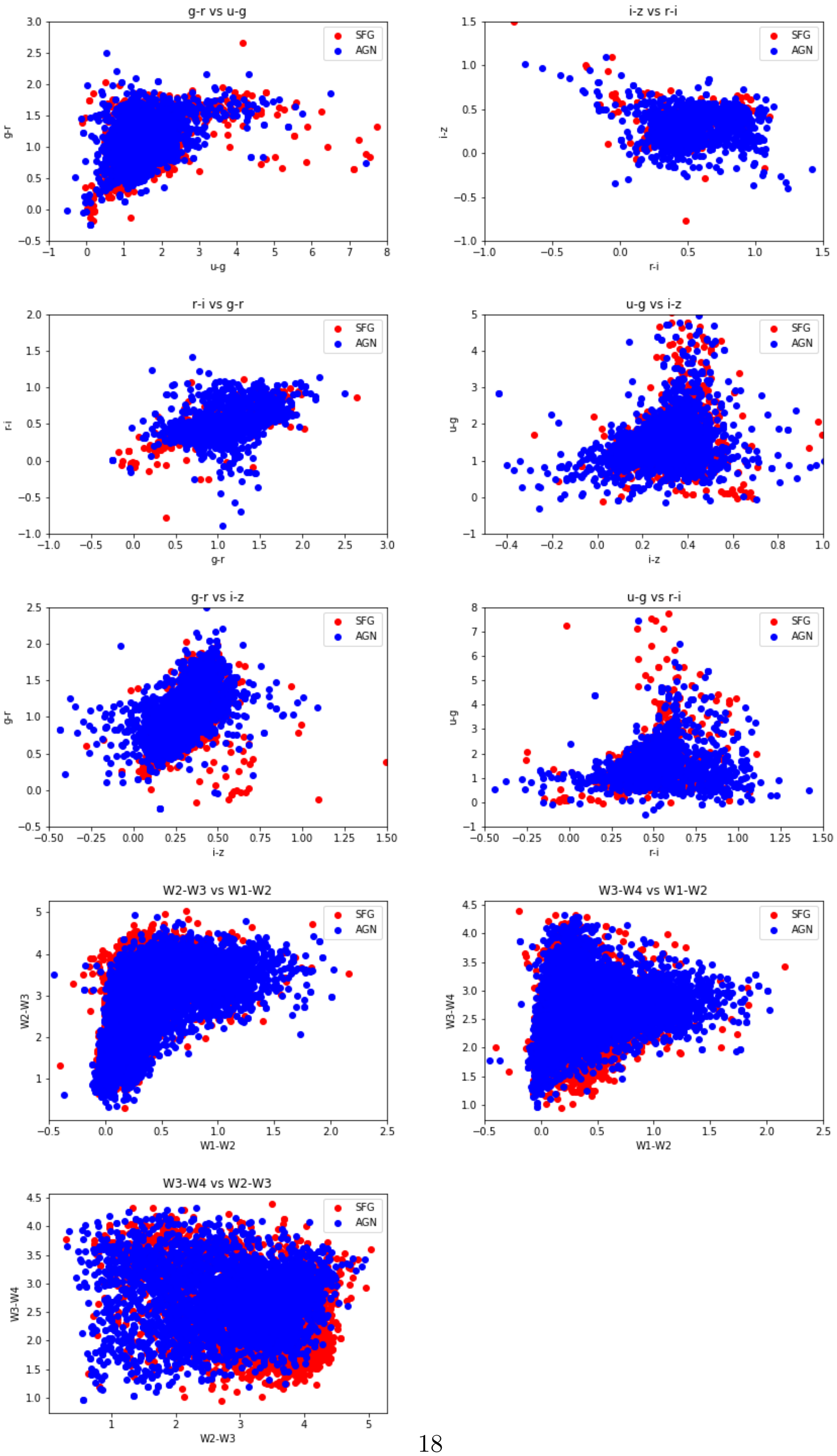

Figure 3: The color-color diagrams for photometric classification for both SFG and AGN

The figure3 shows the possible color-color diagrams for photometric classification for SFG and AGN, by looking at this diagrams the points are spread all over, which will be hard for classification, except the first diagram which is for g-r vs u-g. So we choose g-r vs u-g for our classification problem.

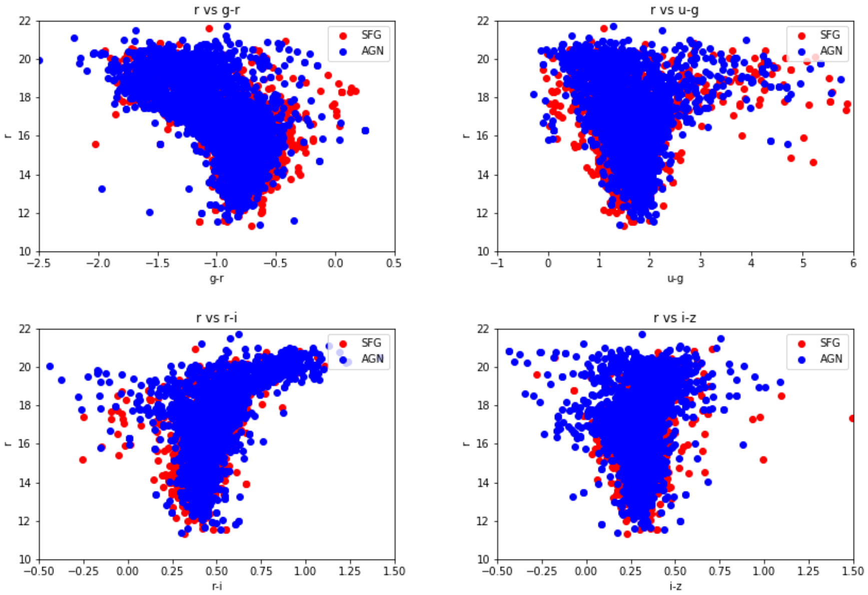

Figure 4: The color-magnitude diagrams for photometric classification for both SFG and AGN

Also figure4 shows the possible color-magnitude diagrams for photometric classification for SFG and AGN. We chose the r-magnitude for the plot since it has been working very well compared to other magnitudes. By looking at our diagrams we can say that this will improve our classification accuracy since since the point is not spread all over as in some color-color diagrams.

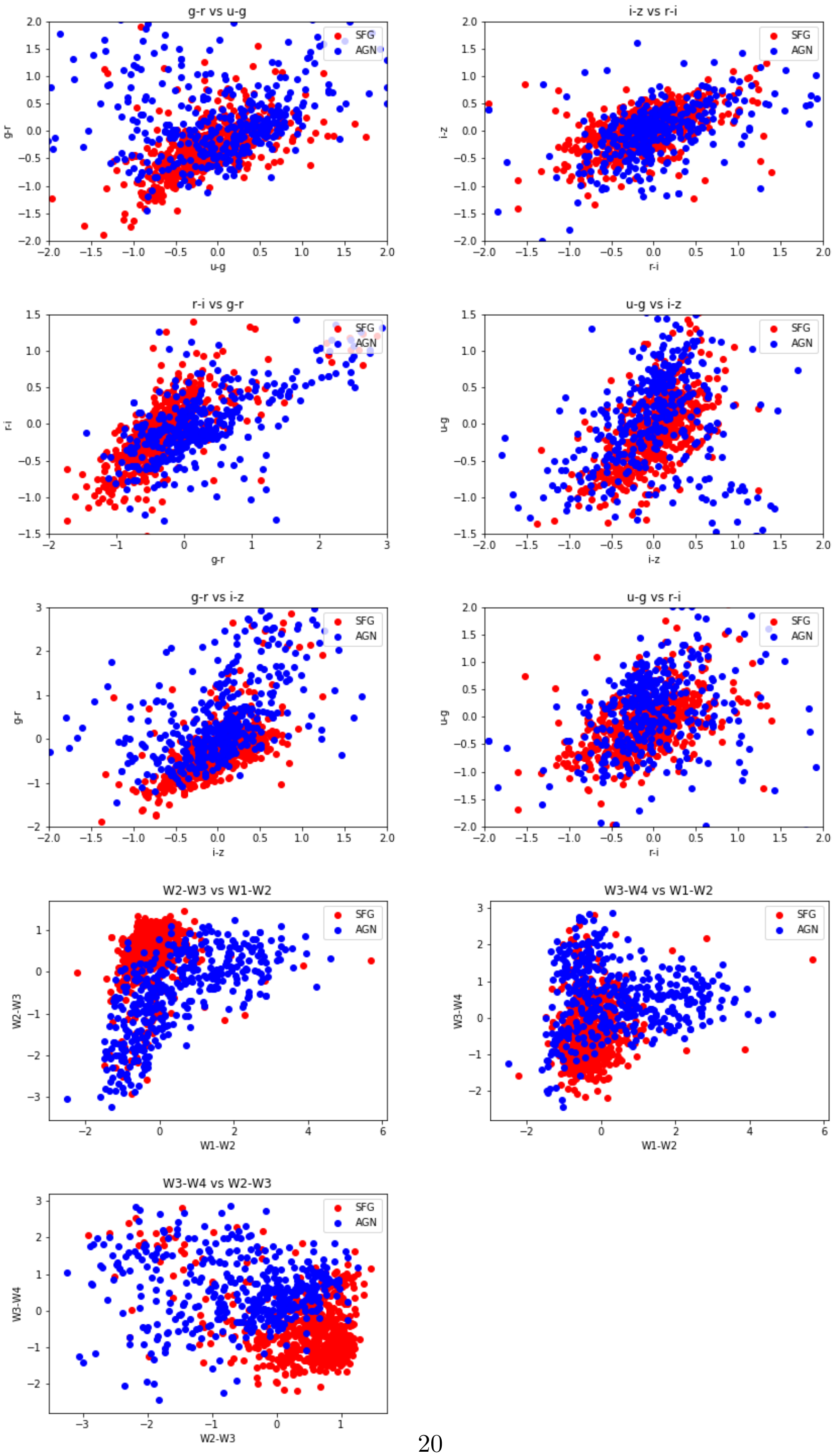

Figure 5: The color-color diagrams for photometric classification for both SFG and AGN after classification using Random forest

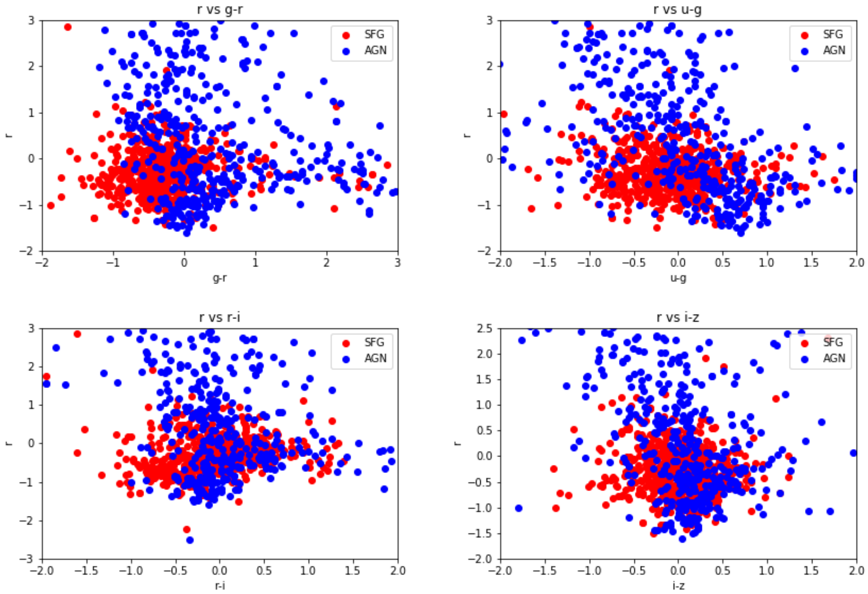

Figure 6: The color-magnitude diagrams for photometric classification for both SFG and AGN after classification using Random forest

Figure5 & 6 shows the color-color and color-magnitude diagrams for photometric classification for both SFG and AGN after classification. And we can see from the plots that the SFG & AGN are now very well classified, hence our process of classification went well.

In a dataset a training set is executed to construct a model, whereas a test set is to approve the built up model. The data points in the preparing set are avoided from the test set. If the test set does contain examples from the training set, it will be difficult to assess whether the algorithm has learned to generalize from the training set or has simply memorized it. Ordinarily, a dataset is separated into a preparing set, a test set in each cycle. We utilize the training information to fit the model and testing information to test it.

Topcat was used to cross match the unified radio catalog (2 866 856 sources) with AllWISE (747 634 026 sources). Then the number of sources was reduced to 402 871, which has 5 677 SFG and 4 134 AGN. From the cross matched data we have taken the corresponding data objects with parameters of magnitudes then derived the colors(difference of the magnitudes) from the magnitudes. The input data object with parameters: g, r, i, z, W1, W2, W3, W4, J, H, K, u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K for both KNN and RF.

We have applied the machine learning algorithms (KNN & RF) to perform classification of between SFG & AGN using URC data and the cross matched data separately. Both algorithms (KNN and Random forest) are trained with 90% of our data, Also we have used 10% for testing our algorithm. For URC the training data contain 17837 sources(1 1310 SFG & 6 527 AGN), also 1982 sources (1 241 SFG & 741 AGN) for testing. The cross matched data contain 8 829 training objects(5 111 SFG and 3 718 AGN), and we have used 982 objects(566 SFG and 416 AGN) for testing our algorithm. The predicted number SFG and AGN are presented in the tables below (i.e. table1,2,3,4,5 &6).

KNN Classification

	SFG	AGN	Total
Training	11310	6527	17837
Testing	1241	741	1982
Predicted	1322	660	1982

Random Forest Classification

	SFG	AGN	Total
Training	11310	6527	17837
Testing	1241	741	1982
Predicted	1282	700	1982

Table 1: The input parameters u-g, g-r, r-i, i-z for both KNN & Random Forest Classification using URC

KNN Classification

	SFG	AGN	Total
Training	11310	6527	17837
Testing	1241	741	1982
Predicted	1300	682	1982

Random Forest Classification

	SFG	AGN	Total
Training	11310	6527	17837
Testing	1241	741	1982
Predicted	1265	717	1982

Table 2: The input parameters u-g, g-r, r-i, i-z, r, g, u, i, z for both KNN & Random Forest Classification using URC

KNN Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	553	429	982

Random Forest Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	594	388	982

Table 3: The input parameters u-g, g-r, r-i, i-z for both KNN & Random Forest Classification using cross-matched data

KNN Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	631	351	982

Random Forest Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	588	394	982

Table 4: The input parameters u-g, g-r, r-i, i-z, r, g, u, i, z for both KNN & Random Forest Classification using cross-matched data

KNN Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	537	445	982

Random Forest Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	555	427	982

Table 5: The input parameters u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K for both KNN & Random Forest Classification using cross-matched data

KNN Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	556	426	982

Random Forest Classification

	SFG	AGN	Total
Training	5111	3718	8829
Testing	566	416	982
Predicted	563	419	982

Table 6: The input parameters u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K, g, r, i, z, W1, W2, W3, W4, J, H, K for both KNN & Random Forest Classification using cross-matched data

We used F1 score, recall and precision as measurements to evaluate the performance of the model:
This metrics contain details, which we get by looking at the anticipated labels from our model with the spectroscopic label. Lets clarify this by using an example of a AGN classifier:
(1) TP- true positive: an object with true label AGN is classified as AGN.
(2) TN- true negative: an object with true label non-AGN is classified as non-AGN.
(3) FP- false positive: an object with true label non-AGN is classified as AGN.
(4) FN-false negative: an object with true label AGN is classified as non-AGN (Rahman et al., 2015).

Precision

Precision indicate how great the classifier is at recognizing TP-true positives, which are the accurately recognized sources. For a low precision for a certain class will show small proportion of positive recognition.

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

Recall
Recall shows how great the classifier is at limiting FN-false negatives. For a low recall of an individual class will demonstrate it is frequently misclassified as another class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score
F1-score is the harmonic mean of recall & precision, also is utilized altogether execution measurements. Here these measurements are determined per class to show the relative execution of each class.

By looking table 8,9,10,11,&12, Random forest has higher accuracy than K-Nearest Neighbour in all of the different input parameters (i.e. by looking at table KNN has accuracy of 78% while Random forest has accuracy of 88%).

$$\text{Recall} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

K Nearest Neighbour				
	Precision	Recall	f1-score	Support
AGN	0.73	0.65	0.69	741
SFG	0.81	0.86	0.83	1241
Accuracy			0.78	1982

Random Forest				
	Precision	Recall	f1-score	Support
AGN	0.87	0.82	0.84	741
SFG	0.89	0.92	0.91	1241
Accuracy			0.88	1982

Table 7: KNN & Random forest Classification Report for colors(u-g, g-r, r-i, i-z) only as an input using URC

K Nearest Neighbour				
	Precision	Recall	f1-score	Support
AGN	0.72	0.66	0.69	741
SFG	0.81	0.85	0.83	1241
Accuracy			0.78	1982

Random Forest				
	Precision	Recall	f1-score	Support
AGN	0.85	0.82	0.84	741
SFG	0.90	0.91	0.90	1241
Accuracy			0.88	1982

Table 8: KNN & Random forest Classification Report for colors(u-g, g-r, r-i, i-z) and magnitude (r, g, u, i, z) as an input using URC

K Nearest Neighbour				
	Precision	Recall	f1-score	Support
AGN	0.68	0.70	0.69	416
SFG	0.78	0.76	0.77	566
Accuracy			0.74	982

Random Forest				
	Precision	Recall	f1-score	Support
AGN	0.73	0.69	0.71	416
SFG	0.78	0.82	0.80	566
Accuracy			0.76	982

Table 9: KNN & Random forest Classification Report for colors(u-g, g-r, r-i, i-z) only as an input

K Nearest Neighbour				
	Precision	Recall	f1-score	Support
AGN	0.69	0.58	0.63	416
SFG	0.72	0.81	0.76	566
Accuracy			0.71	982

Random Forest				
	Precision	Recall	f1-score	Support
AGN	0.71	0.68	0.69	416
SFG	0.77	0.80	0.79	566
Accuracy			0.75	982

Table 10: KNN & Random forest Classification Report for colors(u-g, g-r, r-i, i-z) and magnitudes(r, g, u, i, z) as an input

K Nearest Neighbour				
	Precision	Recall	f1-score	Support
AGN	0.74	0.79	0.76	416
SFG	0.84	0.79	0.81	566
Accuracy			0.79	982

Random Forest				
	Precision	Recall	f1-score	Support
AGN	0.78	0.81	0.79	416
SFG	0.85	0.84	0.85	566
Accuracy			0.82	982

Table 11: KNN & Random forest Classification Report for colors(u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K) only as an input

K Nearest Neighbour				
	Precision	Recall	f1-score	Support
AGN	0.75	0.77	0.76	416
SFG	0.83	0.81	0.82	566
Accuracy			0.79	982

Random Forest				
	Precision	Recall	f1-score	Support
AGN	0.80	0.81	0.80	416
SFG	0.86	0.85	0.86	566
Accuracy			0.83	982

Table 12: KNN & Random forest Classification Report for colors(u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K)and magnitudes(u, g, r, i, z, W1, W2, W3, W4, J, H, K) as an input

We choose k-value in KNN by plotting error vs k and accuracy vs k. For error graph we choose k where the error is at the lowest while for accuracy graph we choose k where the accuracy is at the highest (see figure 7,8,9,10,11 &12). And the value of k that we got from error graph is the same as the k-value that we got from accuracy graph as expected.

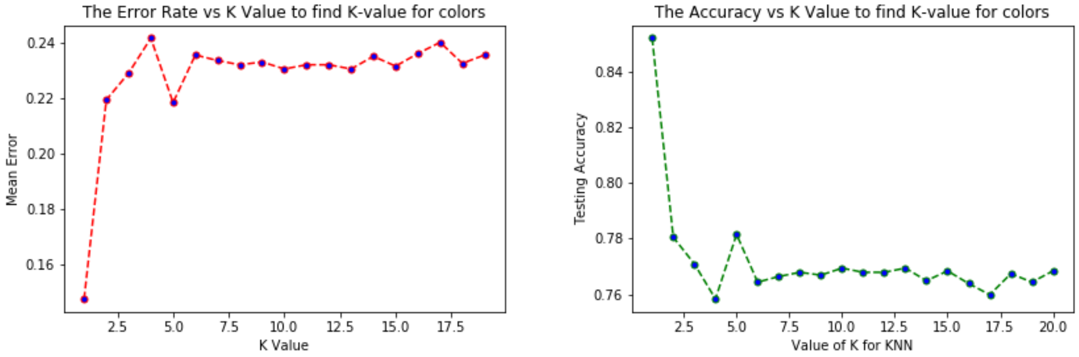

Figure 7: The Error Rate vs K Value and Accuracy vs K Value for u-g, g-r, r-i, i-z using URC

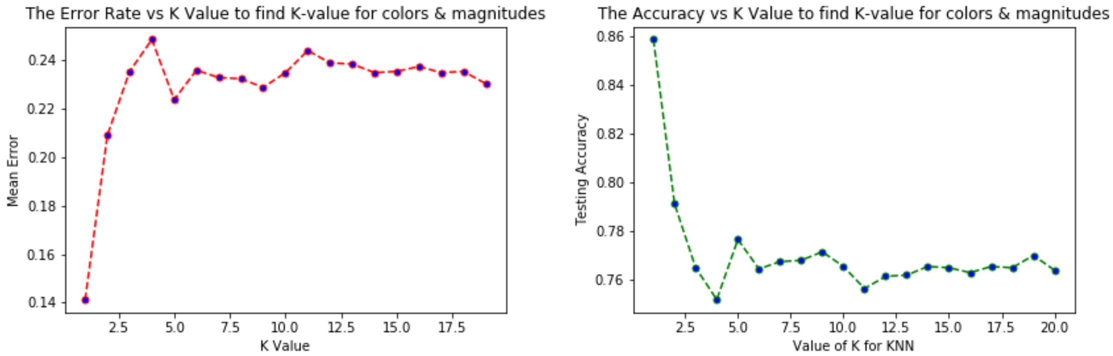

Figure 8: The Error Rate vs K Value and Accuracy vs K Value for u-g, g-r, r-i, i-z, u, g, r, i, z using URC

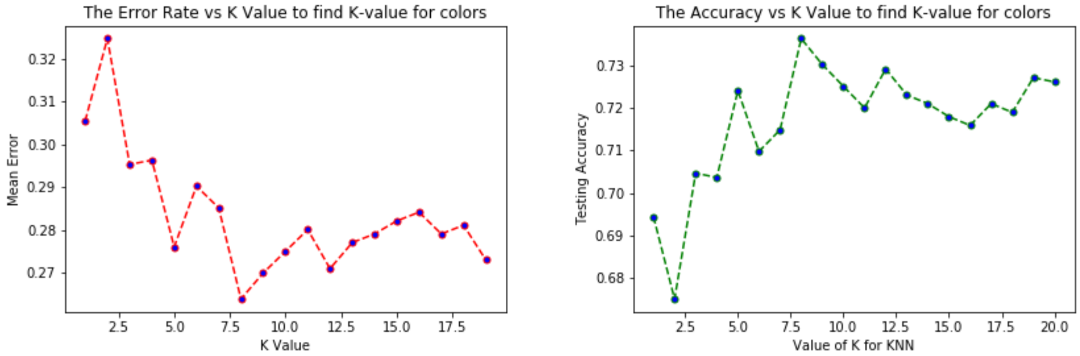

Figure 9: The Error Rate vs K Value and Accuracy vs K Value for u-g, g-r, r-i, i-z

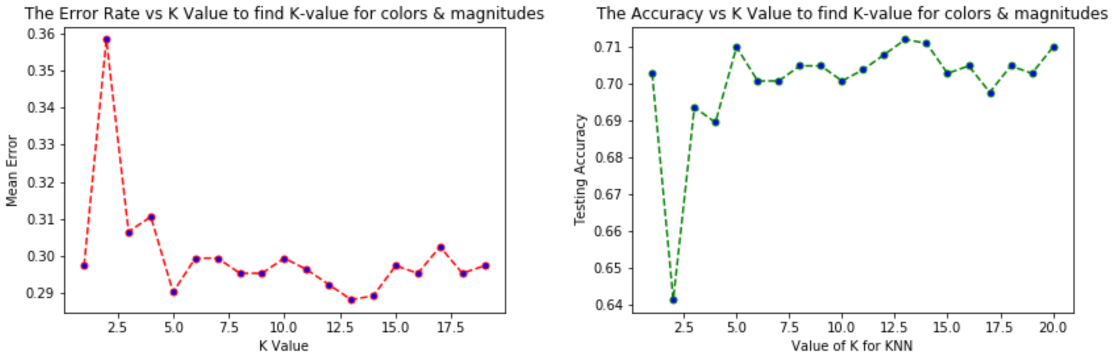

Figure 10: The Error Rate vs K Value and Accuracy vs K Value for u-g, g-r, r-i, i-z, u, g, r, i, z

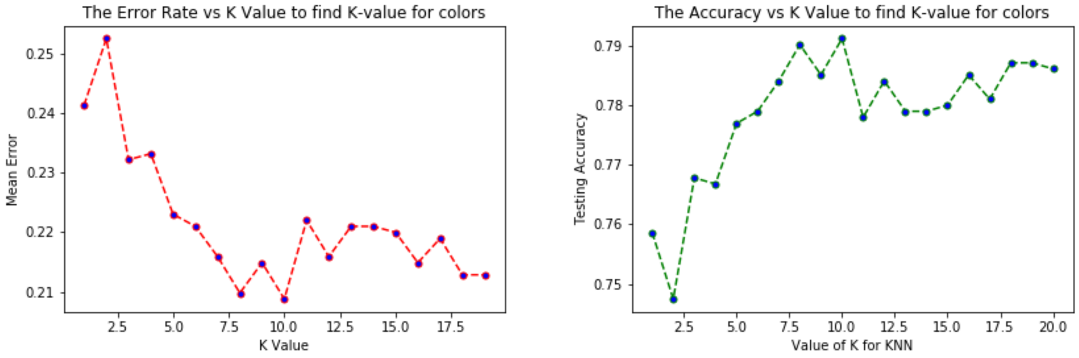

Figure 11: The Error Rate vs K Value and Accuracy vs K Value for u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K

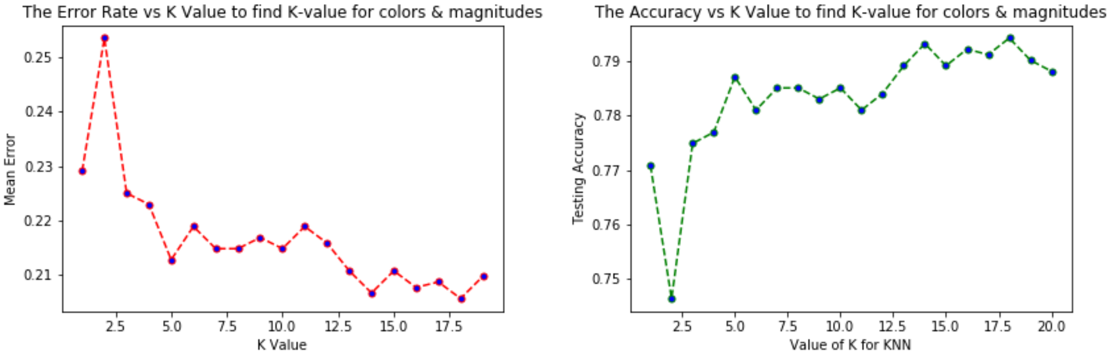

Figure 12: The Error Rate vs K Value and Accuracy vs K Value for u-g, g-r, r-i, i-z, W1-W2, W2-W3, W3-W4, J-H, H-K, W1-J, W2-H, W3-K, u, g, r, i, z, W1, W2, W3, W4, J, H, K

4 Conclusion

The Random forest approach is more accurate than KNN algorithm. Both Random forest and KNN have distinguished SFG with a higher accuracy (or more SFG), independent of the decision of the input parameters. Both Random forest and KNN algorithms perform better with color parameters contradicted to the situation where both (magnitude and color) are utilized.

Our process of classification between SFG & AGN was a success (we see by comparing figure3&4 and figure 5&6). Also (Zhang et al., 2019) support my findings.

To improve the accuracy we must have a lot of data. Also in our data we must take a certain range in the magnitudes from SDDS (optical data) when performing classification, just because of outliers in data, and this reduce the accuracy (see table 8,9,10,11 &12). We need to study deep in a multi-wavelength surveys by cross-matching more catalog like GALEX, FIRST, NVSS, etc.

References

Ahumada, R., Prieto, C. A., Almeida, A., Anders, F., Anderson, S. F., Andrews, B. H., Anguiano, B., Arcodia, R., Armengaud, E., Aubert, M., et al. (2019). The sixteenth data release of the sloan digital sky surveys: First release from the apogee-2 southern survey and full release of eboss spectra. *arXiv preprint arXiv:1912.02905*.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185.

Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. (2006). Robust machine learning applied to astronomical data sets. i. star-galaxy classification of the sloan digital sky survey dr3 using decision trees. *The Astrophysical Journal*, 650(1), 497.

Borne, K. (2013). Virtual observatories, data mining, and astroinformatics. *Planets, Stars and Stellar Systems*, 2, 403–443.

Clarke, A., Scaife, A., Greenhalgh, R., & Griguta, V. (2020a). Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million sdss sources without spectra. *Astronomy & Astrophysics*, 639, A84.

Clarke, A., Scaife, A., Greenhalgh, R., & Griguta, V. (2020b). Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million sdss sources without spectra. *Astronomy & Astrophysics*, 639, A84.

Cutri, R. M., Wright, E. L., Conrow, T., Bauer, J., Benford, D., Brandenburg, H., Dailey, J., Eisenhardt, P. R. M., Evans, T., Fajardo-Acosta, S., Fowler, J., Gelino, C., Grillmair, C., Harbut, M., Hoffman, D., Jarrett, T., Kirkpatrick, J. D., Leisawitz, D., Liu, W., Mainzer, A., Marsh, K., Masci, F., McCallon, H., Padgett, D., Ressler, M. E., Royer, D., Skrutskie, M. F., Stanford, S. A., Wyatt, P. L., Tholen, D., Tsai, C. W., Wachter, S., Wheelock, S. L., Yan, L., Alles, R., Beck, R., Grav, T., Masiero, J., McCollum, B., McGehee, P., Papin, M., & Wittman, M. (2012). Explanatory Supplement to the WISE All-Sky Data Release Products. Explanatory Supplement to the WISE All-Sky Data Release Products.

Ebeling, H., Ma, C.-J., & Barrett, E. (2014). Spectroscopic redshifts of galaxies within the frontier fields. *The Astrophysical Journal Supplement Series*, 211(2), 21.

Gao, D., Zhang, Y.-X., & Zhao, Y.-H. (2009). Random forest algorithm for classification of multiwavelength data. *Research in Astronomy and Astrophysics*, 9(2), 220.

Guo, Y., Giavalisco, M., Ferguson, H. C., Cassata, P., & Koekemoer, A. M. (2012). Multi-wavelength view of kiloparsec-scale clumps in star-forming galaxies at z 2. *The Astrophysical Journal*, 757(2), 120.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Heinis, S., Kumar, S., Gezari, S., Burgett, W., Chambers, K., Draper, P., Flewelling, H., Kaiser, N., Magnier, E., Metcalfe, N., et al. (2016). Of genes and machines: Application of a combination of machine learning tools to astronomy data sets. *The Astrophysical Journal*, 821(2), 86.

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. (2019). *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data*. Princeton University Press.

Kauffmann, G., Heckman, T. M., Tremonti, C., Brinchmann, J., Charlot, S., White, S. D., Ridgway, S. E., Brinkmann, J., Fukugita, M., Hall, P. B., et al. (2003). The host galaxies of active galactic nuclei. *Monthly Notices of the Royal Astronomical Society*, 346(4), 1055–1077.

Koenig, X. & Leisawitz, D. (2014). A classification scheme for young stellar objects using the wide-field infrared survey explorer allwise catalog: revealing low-density star formation in the outer galaxy. *The Astrophysical Journal*, 791(2), 131.

Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S., & Igel, C. (2017). Big universe, big data: machine learning and image analysis for astronomy. *IEEE Intelligent Systems*, 32(2), 16–22.

Kurcz, A., Bilicki, M., Solarz, A., Krupa, M., Pollo, A., & Małek, K. (2016). Towards automatic classification of all wise sources. *Astronomy & Astrophysics*, 592, A25.

Loh, E. & Spillar, E. (1986). Photometric redshifts of galaxies. *The Astrophysical Journal*, 303, 154–161.

Mainzer, A., Bauer, J., Grav, T., Masiero, J., Cutri, R., Dailey, J., Eisenhardt, P., McMillan, R., Wright, E., Walker, R., et al. (2011). Preliminary results from neowise: an enhancement to the wide-field infrared survey explorer for solar system science. *The Astrophysical Journal*, 731(1), 53.

Mobasher, B., Capak, P., Scoville, N., Dahlen, T., Salvato, M., Aussel, H., Thompson, D., Feldmann, R., Tasca, L., Lefevre, O., et al. (2007a). Photometric redshifts of galaxies in cosmos. *The Astrophysical Journal Supplement Series*, 172(1), 117.

Mobasher, B., Capak, P., Scoville, N., Dahlen, T., Salvato, M., Aussel, H., Thompson, D., Feldmann, R., Tasca, L., Lefevre, O., et al. (2007b). Photometric redshifts of galaxies in cosmos. *The Astrophysical Journal Supplement Series*, 172(1), 117.

Osterbrock, D. E. & Ferland, G. J. (2006). *Astrophysics Of Gas Nebulae and Active Galactic Nuclei*. University science books.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Édouard Duchesnay (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85), 2825–2830.

Polsterer, K. L., Gieseke, F., Igel, C., & Goto, T. (2013). Improving the performance of photometric regression models via massive parallel feature selection. In *Proceedings of the 23rd Annual Astronomical Data Analysis Software & Systems conference (ADASS)*.

Rahman, M., Ménard, B., Scranton, R., Schmidt, S. J., & Morrison, C. B. (2015). Clustering-based redshift estimation: comparison to spectroscopic redshifts. *Monthly Notices of the Royal Astronomical Society*, 447(4), 3500–3511.

Rossum, G. (1995). Python reference manual.

Shajib, A. J. & Wright, E. L. (2016). Measurement of the integrated sachs–wolfe effect using the allwise data release. *The Astrophysical Journal*, 827(2), 116.

Sharma, K., Kembhavi, A., Kembhavi, A., Sivarani, T., & Abraham, S. (2019). Detecting outliers in sdss using convolutional neural network. .

Sparke, L. S. & Gallagher, John S., I. (2006). *Galaxies in the Universe - 2nd Edition*. Publishing.

Subhashini, V. (2009a). Data mining techniques to classify astronomy objects. *NITK Surathkal*.

Subhashini, V. (2009b). Data mining techniques to classify astronomy objects. *NITK Surathkal*.

Tacconi, L., Genzel, R., Neri, R., Cox, P., Cooper, M., Shapiro, K., Bolatto, A., Bouché, N., Bournaud, F., Burkert, A., et al. (2010). High molecular gas fractions in normal massive star-forming galaxies in the young universe. *Nature*, 463(7282), 781–784.

Taylor, M. B. (2005). Topcat & stil: Starlink table/votable processing software. In *Astronomical data analysis software and systems XIV*, volume 347 (pp.2̃9).

VanderPlas, J., Connolly, A. J., Ivezic, Z., & Gray, A. (2012). Introduction to astroML: Machine learning for astrophysics. In *Proceedings of Conference on Intelligent Data Understanding (CIDU* (pp. 47–54).

VanderPlas, J., Fouesneau, M., & Taylor, J. (2014). AstroML: Machine learning and data mining in astronomy.

Wang, X. (2011). A fast exact k-nearest neighbors algorithm for high dimensional search using k-means clustering and triangle inequality. In *The 2011 International Joint Conference on Neural Networks* (pp. 1293–1299).: IEEE.

Wright, E. L., Eisenhardt, P. R., Mainzer, A. K., Ressler, M. E., Cutri, R. M., Jarrett, T., Kirkpatrick, J. D., Padgett, D., McMillan, R. S., Skrutskie, M., et al. (2010). The wide-field infrared survey explorer (wise): mission description and initial on-orbit performance. *The Astronomical Journal*, 140(6), 1868.

Ye, J. (2013). Multiple closed-form local metric learning for k-nearest neighbor classifier. *arXiv preprint arXiv:1311.3157*.

Zhang, K., Schlegel, D. J., Andrews, B. H., Comparat, J., Schäfer, C., Mata, J. A. V., Kneib, J.-P., & Yan, R. (2019). Machine-learning classifiers for intermediate redshift emission-line galaxies. *The Astrophysical Journal*, 883(1), 63.

A Appendix A: Abbreviations

ML	Machine Learning
SFG	Star Forming Galaxy
AGN	Active Galactic Nuclei
KNN	K-Nearest Neighbour
RF	Random Forest
SDSS	Sloan Digital Sky Survey
WISE	Wide-field Infrared Survey Explorer
2MASS	Two Micron All-Sky Survey
URC	Unified Radio Catalogue
AllWISE	All Wide-field Infrared Survey Explorer
SDSS	Sloan Digital Sky Survey
WISE	Wide-field Infrared Survey Explorer
2MASS	Two Micron All-Sky Survey
GB6(6cm)	Green bank 6 Centimetres
FIRST(20cm)	Faint Images of the Radio Sky at 20 Centimetres
NVSS(20cm)	NRAO VLASky Survey 20 Centimetres
WENS(92cm)	Westerbork Northern Sky Survey92 Centimeters
VLSSr(4m)	Very Large Array (VLA) Low-Frequency SkySurvey redux 4 Metres
GALEX	Galaxy Evolution Explorer