Cross-matching between Optical and Radio catalogues for classification of Elliptical Galaxies using Machine Learning techniques

Cosmic Data Crew team, SSERD

11th October, 2020 – 28th November, 2020

Mentor: Mr. Sundar M.N.

Team Member

(1) Biswajit Jana; 12017002002018; University Of Engineering and Management, Jaipur

(2) Joshua Dsilva; 218; St Xaviers College, Mumbai

(3) Merin John; 1740420; CHRIST Deemed to be University, Bangalore

(4) Sreevishnu T; 4sf18cs157; Sahyadri College of Engineering and Management, Mangalore

(5) M Sri Sehsa Sai Manoj; 18BEC7034; VIT-AP, Vijayawada

(6) Sudeep Gopavaram; 35813302808; Guru Gobind Singh Indraprashta University, Delhi

(7) Suman Dey; VB-1930; Visva-Bharati university, Santiniketan, Bolpur, West Bengal

(8) Sruthi G Panicker; MAC18EC106; Mar Athanasius College of Engineeing, Ernakulam

(9) Manya Patel; 180320117028; LJ Institute of Engineering and Management, Ahmedabad

(10) Yashwini Thakur; 180320111054;LJ Institute of Engineering and Management, Ahmedabad

(11) Neha P; 20125612921; Reva University, Bangalore

(12) Vikram; 199303067; Manipal University, Jaipur

(13) Bhanvi Menghani; 27; Jaipur Engineering College and Research centre, Jaipur

(14) Mahesh Kumar; 1970144; Kirorimal College, Delhi

Acknowledgments

It is our earnest endeavor to express our gratitude for being assisted to corroborate our cause in the brevity of time. Working together as a team has been a wonderful and wholesome experience. Being benefited by the esteemed guidance of our mentor, M N Sundar, Jain University, Bangalore, and coordinators, Prateek Boga, has instilled the requisite vigor and dedication in us, to make our vision a reality. We would like to express our heartfelt gratitude towards Mahesh P for arranging the technical aids and technical talks by Pavan Kumar which en-compassed topics from training sessions on how to determine the authenticity of paper, writing papers in Latex to important cues while making career choices, has indeed helped us immensely in entirety. Concise and apt inputs from SSERD IPD B4 members Rutuja Attal has also been an important conduit for our work on Data-Driven Astronomy. Most importantly, we would like to thank SSERD, especially Nikhitha C and Sujay Sreedhar for providing us with this opportunity of cohorts with like-minded individuals.

Abstract

Galaxies? The word deeply fascinates one of how one can know that the star which you see in the night sky is just a celestial object or a galaxy which is billions of light years away. Here's a smart and efficient way to find out.
The report is about cross matching between optical and radio catalogs for classification of elliptical galaxies using machine learning techniques. Till now, millions of galaxies has been discovered by various space programs. Here we have studied that how galaxies are evolved and classified different types of galaxies by their size, shapes, redshifts, temperature etc. We have taken catalogues having datasets containing various parameters such as RA, DEC, different bands like u, g, r, i, z etc. which can be separated and used to calculate various other parameters by machine learning using python programming. The shapes of galaxies are classified on the basis of ellipticity which is calculated using Stoke's theorem. Hubble made divisions on the basis of ellipticity of the galaxy and Hubble's - De Vaucouleurs diagram is being used.

Contents

1 Introduction

The universe is boundless. In this fast-growing technology, discoveries are made every day. To cope up with this speed of advancement, one must acknowledge themselves with modern technology. The most adapted and widely used technology in the current scenario is artificial intelligence and machine learning.

1.1 Aims, Objectives and Motivation

Our project aims to cross-match between optical and radio catalogs for the classification of elliptical galaxies using machine learning techniques.
Till now, millions of galaxies have been discovered by various space programs. In addition to that, other space elements like quasars, supernova explosions, nebula, and even black hole has been discovered. Until about a few 100 years ago, the Milky Way galaxy was thought to be the only galaxy existing and it was considered as the entire universe. In the 17^{th} Century, French Astronomer Charles Messier discovered a few 'spiral nebulae' which were later termed and classified as galaxies. Edwin Hubble first discovered that our galaxy isn't the only one in this universe.
The motivation behind our project is to learn and explore more about galaxies. As we go deep into its formation, galaxy's constituents, and its classification, our curiosity grew stronger. We studied about galaxy's shape, size, temperatures, redshifts, and several other parameters. We challenged ourselves into finding more about these fascinating elements of our universe and started collecting data from various sources like Sloan Digital Sky Survey (SDSS), NASA, Vizier, GMRT, Kaggle, and a few more. We faced a lot of difficulties in understanding the data that we obtained. But that was the exact thing that kept us moving forward. We did not give up and kept challenging ourselves till the very end. We cross-matched between the optical and radio catalogs using one of the finest algorithms and we were able to detect and classify elliptical galaxies. Also, to stay up to date, we worked on the 16^{th} data release of the SDSS server which was released in 2020 with more updated catalogs.
In the machine learning area, we got a chance to work with real-time data and that took on a large amount of data. We faced lots and lots of errors but we rectified every one of them and in the end, obtained our desired result.
The only motivation that our team kept was not to give up no matter what. After a lot of errors, rectifications, mistakes, and a true amount of teamwork, we completed our project and fulfilled the expectation of our problem statement.

2 Stellar Evolution

Any star will begin as a cloud of gas and dust at least a few light-years across[1]. Things get so hot that nuclear fusion begins establishing equilibrium and generating a yellow or red main-sequence star.
The process by which star changes with time, depending on the mass of the star, its lifetime can range from a few million years for the most massive to trillions of years for the least massive, which is considerably longer than the age of the universe. The primary factor determining how a star evolves is its mass as it reaches the main sequence. The following is a brief outline tracing the evolution of a low-mass and a high-mass star.

2.1 The Life of a star

Stars are born out of the gravitational collapse of cool, dense molecular clouds. As the cloud collapses, it fragments into smaller regions, which themselves contract to form stellar cores. These protostars rotate faster and increase in temperature as they condense, and are surrounded by a protoplanetary disk out of which planets may form later. The central temperature of the contracting protostar increases to the point where nuclear reactions begin. At this point, hydrogen is converted into helium in the core and the star is born onto the main sequence. For about 90% of its life, the star will continue to burn hydrogen into helium and will remain a main-sequence star.
Once the hydrogen in the core has all been burned to helium, energy generation stops and the core begins to contract. This raises the internal temperature of the star and ignites a shell of hydrogen burning around the inert core. Meanwhile, the helium core continues to contract and increase in temperature, which leads to an increased energy generation rate in the hydrogen shell. This causes the star to expand enormously and increase in luminosity – the star becomes a red giant.
Eventually, the core reaches temperatures high enough to burn helium into carbon. If the mass of the star is less than about 2.2 solar masses, the entire core ignites suddenly in a helium core flash. If the star is more massive than this, the ignition of the core is gentler. At the same time, the star continues to burn hydrogen in a shell around the core. The star burns helium into carbon in its core for a much shorter time than it burned hydrogen. Once the helium has all been converted, the inert carbon core begins to contract and increase in temperature. This ignites a helium burning shell just above the core, which in turn is surrounded by a hydrogen-burning shell.

What happens next depends on the mass of the star**:**

1. ***Stars less than 8 solar masses:***

(a) The inert carbon core continues to contract but never reaches temperatures sufficient to initiate carbon burning. However, the existence of two burning shells leads to a thermally unstable situation in which hydrogen and helium burning occur out of phase with each other. This thermal pulsing is characteristic of asymptotic giant branch stars.

(b) The carbon core continues to contract until it is supported by electron degeneracy pressure. No further contraction is possible (the core is now supported by the pressure of electrons, not gas pressure), and the core has formed a white dwarf. Meanwhile, each thermal pulse causes the outer layers of the star to expand, resulting in a period of mass loss. Eventually, the outer layers of the star are ejected completely and ionized by the white dwarf to form a planetary nebula.

2. ***Stars greater than 8 solar masses:***

(a) The contracting core will reach the temperature for carbon ignition, and begin to burn to neon. This process of core burning followed by core contraction and shell burning is repeated in a series of nuclear reactions producing successively heavier elements until the iron is formed in the core.

(b) Iron cannot be burned to heavier elements as this reaction does not generate energy – it requires an input of energy to proceed. The star has therefore finally run out of fuel and collapses under its gravity.

(c) The mass of the core of the star dictates what happens next. If the core has a mass less than about 3 times that of our Sun, the collapse of the core may be halted by the pressure of neutrons (this is an even more extreme state than the electron pressure that supports white dwarfs!). In this case, the core becomes a neutron star. The sudden halt in the contraction of the core produces a shock wave which propagates back out through the outer layers of the star, blowing it apart in a core-collapse supernova explosion. If the core has a mass greater than about 3 solar masses, even neutron pressure is not sufficient to withstand gravity, and it will collapse further into a stellar black hole.

(d) The ejected gas expands into the interstellar medium, enriching it with all the elements synthesized during the star's lifetime and in the explosion itself. These supernova remnants are the chemical distribution centers of the Universe.

(e) An important tool in the study of stellar evolution is the Hertzsprung - Russell diagram (HR diagram), which plots the absolute magnitudes of stars against their spectral type (or stellar luminosity versus effective temperature). As a star evolves, it moves to specific regions in the HR diagram, following a characteristic path that depends on the star's mass and chemical composition.

(f) Stars are classified by their spectra (the elements that they absorb) and their temperature. There are seven main types of stars. In order of decreasing temperature, O, B, A, F, G, K, and M. O and B stars are uncommon but very bright; M stars are common but dim.

Mass (solar masses)	Time (years)	Spectral type
60	3 million	O3
30	11 million	O7
10	32 million	B4
3	370 million	A5
1.5	3 billion	F5
1	10 billion	G2 (Sun)
0.1	1000s billions	M7

Further, the star will swell up into a giant star just like we saw for the low mass star while the core of a high-mass star continues to compress, it gets much hotter than the core of a low mass star and it can fuse helium nuclei to form carbon, oxygen, neon, silicon and then iron.
Iron nuclei are so stable that further fusion will not release any more energy. At this point gravity wins the fight and the star collapses within a single second. The outer layers bouncing off the core and triggering an explosion, thus ejecting all of the heavy nuclei the star has created out into space. This event is called a supernova.
A supernova generates such an unbelievable burst of energy that in this brief moment dozens of elements heavier than iron such as nickel, copper-zinc, silver, gold, and any element with an atomic number greater than 26. These heavier elements are made in a supernova or a rare event like the collision of two neutron stars or a neutron star and a black hole. Stars like the sun take 10+ billion years to form and 100+ billion years to orbit their galaxies. Colliding galaxies takes billions of years to merge.
Before any stars had formed, everything was just spread out gas and a lot of dark matter. Gravity would have caused slightly denser areas of dark matter to attract into clumps, pulling in bits of gas until they were dense enough on their own to gravitationally collapse and starts thermonuclear fusion. A cluster of stars and their associated dark matter attracted together

and merged and then those clusters clustered together eventually forming a galaxy. This happens after the big bang.

3 Formation of Galaxies

3.1 Galaxy Evolution

Before we know how galaxies are formed, what exactly galaxies are? Galaxies are known as a collection of Interstellar dust, different gases, dark matter, and the key constituent i.e. million to trillions of stars which are held together by the gravitational force of each other.
There are more than a million galaxies in our universe and it is thought that every galaxy contains a supermassive black hole(SBM) at the center, Like in our Milky way galaxy i.e. Sagittarius A*. Another term that is related to the galaxy is known as clusters or mergers so basically galaxies are categorized under clusters or superclusters when there is a group of galaxies existing together, one such example is the Virgo cluster which exists within the Virgo supercluster. These galaxies are spread across more than a million light-years across. And mergers are formed when more than one galaxy collide with each other during which the gases present within each galaxy tend to move toward the galactic center of each galaxy due to the gravitational pull which can lead to different actions one of them can be rapid star formation and more.
Now when it comes to how these galaxies form this is still a big question.
A basic overview of galaxy formation can be given as after the big bang happened the energy condensed into the matter and as the universe started to cool down, atomic nuclei formed eventually started interacting with electrons to make neutral hydrogen and helium atoms. According to Einstein's general theory of relativity objects with the mass warp, the space-time i.e. the space-time fabric tend to bend which attracts the other massive objects towards each other even though the matter in the universe consisted only of small atoms but they still exert gravitational force due to which all this hydrogen and helium in the universe started to accumulate forming regions of higher density as a some of the regions become so dense particle within the gas exert a certain pressure or outward force and gravitational force will be acting in the opposite direction which will make this gas cloud remain in equilibrium unless and until one of these two forces become more than other this situation is also called hydrostatic equilibrium. But in case the mass of gas cloud having helium and hydrogen becomes such that it exceeds jeans

mass

$$M_J = \left(\frac{5lkT}{G\mu m_H}\right)^{\frac{3}{2}}\left(\frac{3}{4\pi\rho}\right)^{\frac{1}{2}} \tag{1}$$

which can be considered as a threshold value as a result space-time fabric will bend more and more which will result in more gravitational force and minimum particle force leading to a gravitational collapse increasing the mass at the core and rise of temperature which is too hot for a neutral atom to exist this inner region is so hot that the outward pressure supports the gas against further collapse forming a protostar an object which is in temporary hydrostatic equilibrium. As more material starts to collect on protostar increasing the gravitational potential energy with additional mass collapsing then continues again so that it causes the temperature to rise to millions of degrees until things are hot enough for nuclear fusion to occur and when fusion starts a star is born. The formation of this star will trigger reionization in the surrounding nebula, striping these gas particles of their electrons which will lead to star formation over hundreds of millions of years. A massive star exerts a tremendous amount of gravity; the star began to collect in a dense region just as the gas-particle did. This same period that accounts for star formation is also when the galaxies began to form some of these are dwarf and some of these are huge. But gravity didn't stop there these huge galaxies along with the dark matter exert even greater gravitational influence such that they collect to form groups, clusters, and superclusters.
Some theories suggest that galaxies are formed by the combination of small star clusters called globular cluster and other suggest that first galaxies are formed then these splits into several globular clusters but one common thing in these theories is a dark matter every galaxy is assumed to be present at the haloes of dark matter. So before we study these two theories we need to understand some basic classifications of dark matter and what dark matter is.

$$Masscloud > MJ \rightarrow Collapse$$

3.2 Dark Matter

This topic comes under one of the biggest mysteries of the universe. A non-visible matter which is interacting with the visible matter with the help of gravity is supposed to provide some mass to galaxies and star which is helping the galaxies and the star to remain intact.
Dark matter is classified as Cold dark matter (CDM), hot dark matter (HDM), Warm Dark matter (WDM) all these are the hypothetical classification of

dark matter, and out of these three CDM is more dominant, but our galaxy formation theories depend on the cold and dark matter only.
The distinguishing characteristic in determining the temperature of matter is how fast it moves. In thermodynamics, the relation between kinetic energy and temperature of the gas is given as

$$E_k = \frac{3}{2}kT \tag{2}$$

Where E_k is kinetic energy and T is the temperature in Kelvin, k is Boltzmann constant
$k = 1.38 \times 10^{-23} \ \ m^2\ kg\ s^{-2}\ K^{-1}$
And kinetic energy of something moving with velocity v, and having mass m is given as

$$E_k = \frac{mv^2}{2} \tag{3}$$

When we manipulate the above two equations we get a relation between temperature and velocity given as

$$V_{rms} = \sqrt{\frac{3kT}{m}} \tag{4}$$

Where, V_{rms} is defined as the root mean square value
By referring to this equation we can say that if something is hot then it is moving fast and anything cold can be considered as moving slow reason for this can be atoms within any hotter body will not be stable because of higher thermal energy rather they will be in random motion i.e. higher kinetic energy which will make the body to move fast and when it comes to the colder body there will be less kinetic energy within the atoms because of less thermal energy which will lead to a lesser velocity of the material.
Name of the two theories are :-

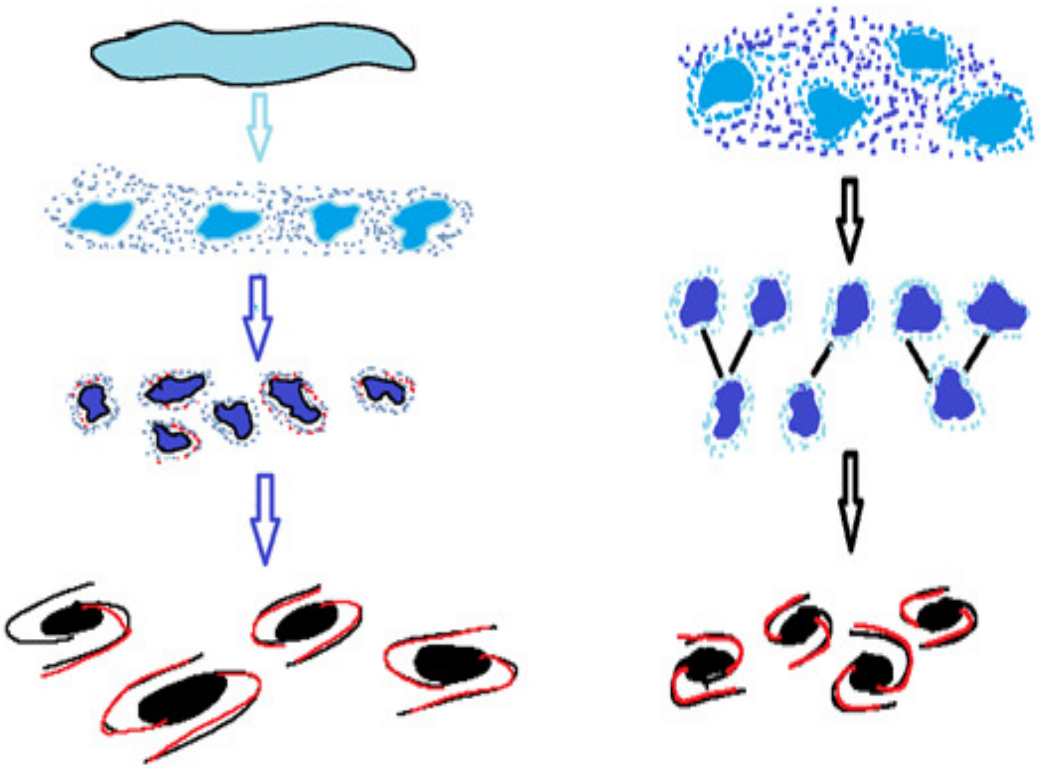
3.2.1 Top-down structure formation

This theory suggests that large superclusters of galaxies form first and these then separate to form smaller structures like galaxies [2].
This theory is favored by hot dark matter as it consists of a particle that moves at velocities comparable to that of light. However, calculations indicate that this does not permit structure to form initially on scales smaller than superclusters because of the high velocity if the material will destroy all structure on smaller scales such as that of galaxies; therefore this theory assumes that structure formation in the early universe was dominated by hot

dark matter. In a top-down scenario, large pancakes of matter form first, and then fragment into galaxy-sized lumps

3.2.2 Bottom-up structure formation

This theory suggests that masses of the size of star clusters collapse first and these due to collision with each other merge into galaxies and superclusters[3]. This theory is favored by the cold dark matter because to form small structures velocity of matter must be small otherwise collision will happen and because dark matter has high velocity it would destroy these small-scale structures. In a bottom-up scenario, small, a dwarf galaxy-sized lump form first, and then merges to make galaxies and clusters of galaxies



(a) Top-down Formation

(b) Bottom-up Formation

4 Constituents of Galaxies

For the formation as well as evolution of galaxies local environment has a profound impact on this. For understanding the distribution of different types of galaxy constituents, in the large-scale structure it is necessary to have a clear notion of the formation of galaxy. Highly-dense environments are very much expected to compose of diverse galaxy constituents, including star forming galaxies, X-ray sources, Active Galactic Nuclei. In order to classify galaxies can be grouped into four types: spiral galaxies, lenticular galaxies, elliptical galaxies, and irregular galaxies. At low redshifts, dense cluster galaxies environments have experienced to be more massive, contain stellar populations which are much older as well as have very low star formation rates and few dust contents also there is a higher fraction to be elliptical[4].

Typically, Galaxy composed of following constituents: Stars, Interstellar gases, Dust particles, Cosmic Rays, Dark matter and Dark Matter (DM) halo[5].

Stars:

Stars or star clusters are a bunch of huge celestial bodies that are made of hydrogen and helium. This is the reason for churning nuclear forges inside the cores of the star. Typically, stars can occur in galaxy as: field stars (i.e. isolated stars), stars contained in open (galactic) clusters, stars contained in globular cluster. Globular clusters are denser than open clusters.

Interstellar gas:

Interstellar gases are mostly in the galactic disk region, typically in a thin layer. Much of the gases are in discrete cloud form, especially atomic neutral hydrogen (HI) and molecular dense clouds (H_2). Diffused gaseous nebulae (HII), low dense inter-cloud medium, highly ionized coronal gas also present there. Examining warm ionized interstellar gasses have been shown using optical emission lines with a large field of view with extreme high accuracy. Both primary and secondary set of images are been taken including ionization as well as excitation condition of diffused gaseous nebulae (HII).Hot i.e. typically around 10^6 K, interstellar gas is been studied in X-rays. Stellar coronal emission as well as X-ray binaries both contributes to the observed X-ray flux[6].

Dust Particles:

Interstellar dust cloud consists of a large number of small and extremely

solid particles, mainly silicates were found and also some amount of water ice, in interstellar space region, it was found typically in ∼100 nm sizes. Dust grains generally occur in a thin layer in the galactic disk region, in the dark clouds (the Coal sack dark nebula), and also in small and very dense dark globules. Dust can also be seen as "reflection" nebulae near some extreme hot stars (Pleiades).The evolution of each and every dust particle can be affected by

1. The gravitational force from other components of galaxy and from other dust particles present in the galaxy
2. Radiation Pressure of stars on gas
3. Force due to drag of surrounding gas particles.

The destruction of dust particles by hot haloes of galaxies, in the detailed processes this would depends on mass density as well as temperature of such hot halos[7].

Cosmic Rays:

Galactic cosmic rays are typically high energy particles $E > 10^9$ eV are in plasmonic form and may be occasionally high as 10^{19} eV. They are mainly protons, as well as helium nuclei (called α particles), and also some nuclei of other light particles, like Carbon(C), Nitrogen(N) and Oxygen(O). They are been trapped magnetic field in the galactic nuclei region. The origin of those may be from exploding of massive stars such as supernovae.

To avoid the problems associated with confinement in the galactic disk, the further suggestion has been made that the galactic cosmic rays are confined in a spherical volume of radius within 15 kpc, which is referred as the Galactic Halo region. A halo is dynamically stable with long-lived cosmic-rays trapped inside the volume. It allows for free passage of cosmic rays between the disk region and halo regions[8].

Dark Matter:

Typically, Dark matter emits no light and is a form of unknown mass, and not even baryonic (composed of protons, neutrons), But kind a fermionic have spin integer[9]. It has its gravitational field and this cause the Galaxy can rotate faster than can otherwise like light. 90 per cent of the mass of the Galaxy may be made of "dark matter". It is a major problem of modern astrophysics. Using a combined observation with rotational curves as well as gravitational lensing method, showed this can be used to implement the equation of state of a dark matter[10].

5 Classification of Galaxies

5.1 Hubble's classification of galaxy :

In understanding any new collection of objects, it is necessary to classify them according to their intrinsic characteristics and their properties. Hubble played a major role in classifying galaxies. Hubble in his book "The realm of nebulae" proposed that galaxies can be grouped into three primary categories based on their complete appearance. This morphological classification is now known as the Hubble sequence. This sequence then divides galaxies into ellipticals (E's), spirals, and irregulars (Irr's). The spirals are further sectioned into two, the normal spirals (S's), and the barred spirals (SB's).

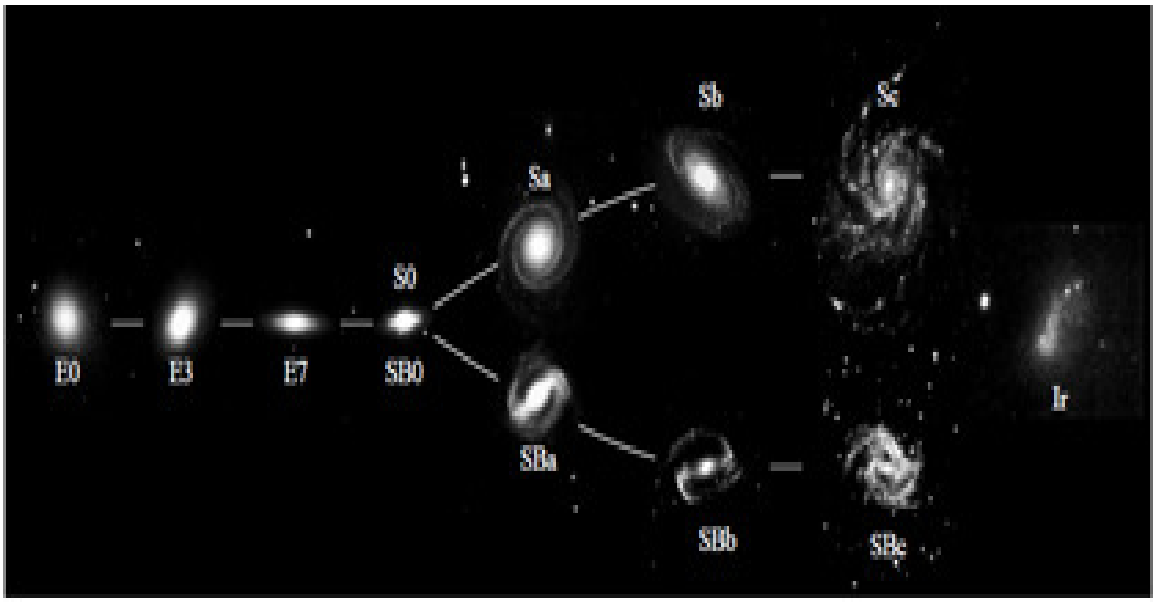


Figure 2: Hubble Classification Of Galaxy

Large Elliptical galaxies

In the category of ellipticals, Hubble made divisions based on the ellipticity of the galaxy which is defined by

$$\epsilon = 1 - \frac{b}{a} \tag{5}$$

Where α and β are the apparent major and minor axes of the ellipse, respectively, projected onto the plane of the sky. If the ellipticity of the galaxy is 0.7, then 0.7 is multiplied by 10 giving us 7. The letter E stands for elliptical and the type of that galaxy becomes E7.
Eg. M110

Note:- The apparent ellipticity may not be the actual ellipticity since the orientation of the spheroid to our line of sight plays a crucial role in our observations. A spheroidal galaxy that has lengths a=b and c<a. A prolate spheroidal galaxy has axis lengths b=c and a>b.

Lenticular galaxy

A transitional class of galaxies between ellipticals and spirals is known as lenticulars. It can be either normal (S0's) or barred (SB0's). Hubble arranged his sequence in the form of a tuning-fork diagram. As a further enhancement to the system, the lenticular galaxies are also subdivided according to the amount of dust absorption in their disks. S01 galaxies have no dust on their disks, while S03 galaxies have significant amounts of dust. Similarly, it is for SB01 through SB03.
Eg. NGC 2787

Spiral galaxies

The spiral galaxies are been subdivided by Hubble into Sa, Sab, Sb, Sbc, Sc, and SBa, SBab, SBb, SBbc, SBc. The Spiral galaxy is divided into two parts. S stands for **normal spiral** and a, ab, b, bc, and c indicates how close the spiral arms are arranged. a tells us that spiral arms are arranged extremally closely and c tells us that the arms are highly spread out. Whereas, SB stands for the **barred galaxy.** Similarly, in the case of the barred spiral a, ab, b, bc, and c indicates how close the spiral arms are arranged.
Normal spiral galaxy - Eg. NGC1232
Barred spiral galaxy - Eg. NGC1300

Irregular galaxy

The remaining category of galaxies were been split by Hubble into irregulars and then they were subdivided into Irr I if there was at least some hint of an organized structure and Irr II for the most extremely disorganized structures.
Irregular type of galaxy - Eg. NGC1427

5.2 De Vaucouleurs classification of galaxy:

To make much better distinctions between normal and barred spirals, De Vaucouleurs recommended that normal spirals should be referred to as SA rather than simply S. This change in the Hubble's classification helped in classifying intermediate types of galaxies. Intermediate types of galaxies

with weak bars are then been denoted as SAB, and strongly barred galaxies are SB. The overall picture of classification looks like this

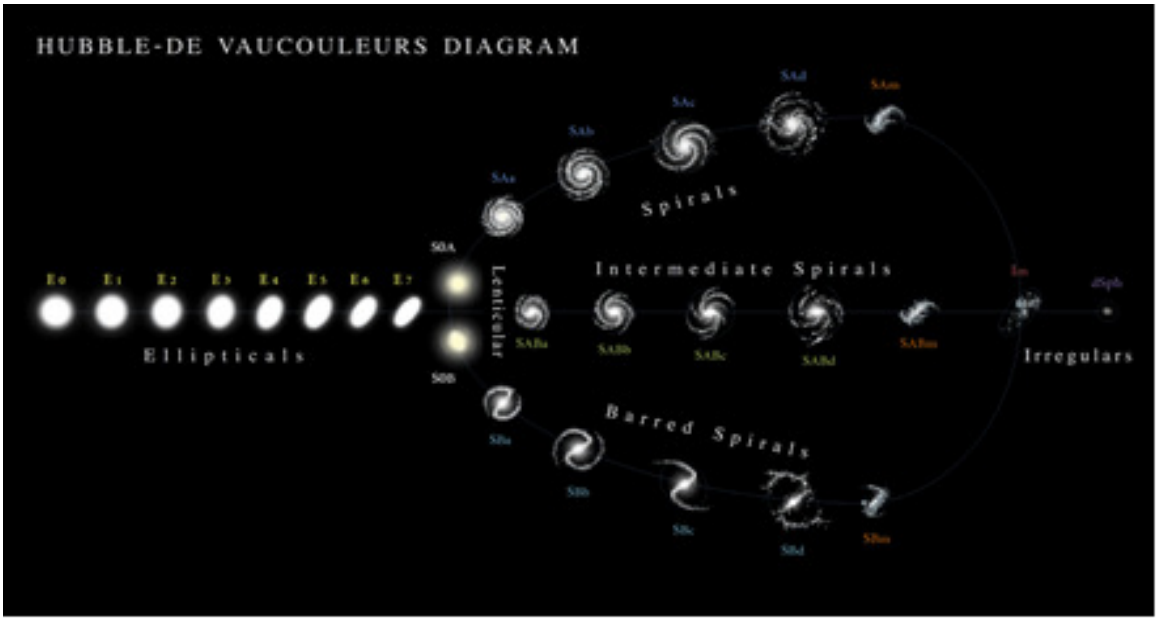


Figure 3: De Vaucouleurs Classification Of Galaxy

6 Properties

Hubble type:

Hubble invented classification of galaxies and it is represented in the tuning fork diag. He classified them based on the size of the central bulge and the arms of the galaxies[11].
According to Hubble's classification, the ellipiticals are classified based on the ellipticity and it is given as

$$E = 10\left(1 - \frac{b}{a}\right) \tag{6}$$

And it ranges from E0 – E7 depending on the semi major and semi minor axes.
Hubble classified the spiral galaxies into two types: Regular spirals and Barred spirals. Regular spirals are placed at the top and barred spirals are placed at the bottom. Both are further classified into

1. Sa (SBa) – They have tightly wound smooth arms and bright central bulge.
2. Sb (SBb) – They have less tightly wound arms and a little faint central bulge.

3. Sc (SBc) – They have loosely wound arms and very faint central bulge.

Irregular galaxies are extended classification of Hubble and he divided it into two types:
Irr 1 – They are slightly organised and Irr 2 – They are completely disorganised.

Size:

To measure the size of the galaxies there are various methods. The Holmberg radius is one of the methods to measure the radius of galaxy. It is measured using the surface brightness. It is the radius at which the surface brightness has an apparent magnitude of 26.5 per square arc second. The surface brightness doesn't decline with the distance so the radius doesn't change and can be used to measure the size of the galaxy.
The conventional method used for measuring the diameter involves summing up of the total light from the galaxy through a large aperture so that the total light doesn't depend on the particular size of the aperture used. The giant ellipticals have diameters upto 600,000 ly, the disc of spirals range from 30,000 to 150,000 ly, the dwarfs and irregulars have diameters as small as few thousand light years.

Colour:

The colour of the galaxy depends on the stars in it. The galaxy will appear blue if most of the stars in that galaxy are blue in colour and if most of the stars are red then the galaxy appears red. The colour of galaxy tells astronomers the type of stars present in the galaxy. The blue region contains new stars and the red region contains old stars whereas the light red/ pink region is a cloud of ionized hydrogen. These are the features that can be observed in visible wavelength range but when we observe radio, infrared, X-ray, UV, gamma rays wavelength it shows more galactic characteristics. The colour of galaxies is affected by a few factors like red-shift.

Surface Brightness:

The surface brightness is set by the density of stars and it is an intrinsic property of the galaxy. Galaxies are not point sources like stars therefore the galaxy is harder to see than the star. So, we describe the galaxy's light distribution in terms of surface brightness, which describes the flux density per unit angular area. Measurement of surface brightness of celestial objects in astronomy is called photometry. The surface brightness of a source of total magnitude m over an area A square arc seconds is given by

$$S = m + 2.5log_{10}A \tag{7}$$

The surface brightness taken in physical units of solar luminosity per square parsec is given by

$$S(mag/arcsec^2) = M + 21.572 - 2.5log_{10}S(L/pc) \tag{8}$$

Where, M and L are the absolute magnitude and luminosity of the sun[12].

Galaxy stellar mass:

The galaxy's mass depends upon the orbital motion of the stars. With the help of the size of star's orbit we can derive galaxy's mass. The spiral galaxy mass is calculated using the rotation curve, which shows how the orbital speeds in the galaxy depends on the distance from the center of the galaxy. The elliptical galaxy mass is calculated by blending together the width of absorption lines from all the stars.

$$Elliptical\ galaxy\ mass = \left(k \times V^2 \times d\right)/G \tag{9}$$

Where, k is a factor which depends on the angle of galaxy from earth and shape of the galaxy, V is the velocity dispersion and d is the distance of the stars from galaxy center[13].

Rotational velocity:

The maximum rotational velocity of the galaxy is measured from the HI and Hα rotational curve profiles. The size of the galaxy disc V^2 times can be used to calculate the total mass of the galaxy.
Some of the most basic properties of galaxy include spectral line properties, sizes, internal velocities, spectral energy distributions, structural features and morphologies, metallicities, the current and past star formation rates.
The properties that we used in our project are adaptive moments, colour indices, Petrosian flux and colour index[14].

7 Application

We now turn to the use of machine learning algorithms and stellar population models in astronomical applications, and their track record in addressing some common problems. Given that there is no exact definition of what constitutes a data mining tool, it would not be possible to provide a complete overview of their application. This section, therefore, illustrates the wide variety of actual uses to date, with actual or implied further possibilities.Most of the applications in this section are made by astronomers utilizing data mining algorithms. However, several projects and studies have also been made by data mining experts utilizing astronomical data, because, along with other fields such as high energy physics and medicine, astronomy has produced many large datasets that are amenable to the approach. In general, the data miner is likely to employ more appropriate, modern, and sophisticated algorithms than the domain scientist, but will require collaboration with the domain scientist to acquire knowledge as to which aspects of the problem are the most important.

Object classification:

Classification is often an important initial step in the scientific process, as it provides a method for organizing information in a way that can be used to make hypotheses and to compare with models. Two useful concepts in object classification are *completeness* and *efficiency*, also known as recall and precision. [15]They are defined in terms of true and false positives (TP and FP) and true and false negatives (TN and FN). The completeness is the fraction of objects that are true of a given type that is classified as that type:

$$Completeness = \frac{TP}{TP + FN} \tag{10}$$

and the efficiency is the fraction of objects classified as a given type that are true of that type

$$Efficiency = \frac{TP}{TP + FP} \tag{11}$$

These two quantities are astrophysically interesting because, while one wants both higher completeness and efficiency, there is generally a tradeoff involved. The importance of each often depends on the application, for example, an investigation of rare objects generally requires high completeness while allowing some contamination (lower efficiency), but statistical clustering of cosmological objects requires high efficiency, even at the expense of completeness.

Star-Galaxy Separation:

Due to their small physical size compared to their distance from us, almost all stars are unresolved in photometric datasets, and thus appear as point sources. Galaxies, however, despite being further away, generally subtend a larger angle, and thus appear as extended sources. However, other astrophysical objects such as quasars and supernovae, also appear as point sources. Thus, the separation of photometric catalogs into stars and galaxies, or more generally, stars, galaxies, and other objects, is an important problem. The sheer number of galaxies and stars in typical surveys (of order 10^8 or above) requires that such separation be automated.

Galaxy Morphology:

Galaxies come in a range of different sizes and shapes, or more collectively, morphology. The most well-known system for the morphological classification of galaxies is the Hubble Sequence of elliptical, spiral, barred spiral, and irregular, along with various subclasses. This system correlates too many physical properties known to be important in the formation and evolution of galaxies. Because galaxy morphology is a complex phenomenon that correlates to the underlying physics but is not unique to anyone given process, the Hubble sequence has endured, despite it being rather subjective and based on visible-light morphology originally derived from blue-biased photographic plates. The Hubble sequence has been extended in various ways, and for data mining purposes the T system has been extensively used. This system maps the categorical Hubble types E, S0, Sa, Sb, Sc, Sd, and Irr onto the numerical values -5 to 10.
One can, therefore, train a supervised algorithm to assign T types to images for which measured parameters are available. Such parameters can be purely morphological or include other information such as color.

Quasars/AGN:

Most of the emitted electromagnetic radiation in the universe is either from stars, or the accretion disks surrounding supermassive black holes in active galactic nuclei (AGN). The latter phenomenon is particularly dramatic in the case of quasars, where the light from the central region can outshine the rest of the galaxy. Because supermassive black holes are thought to be fairly ubiquitous in large galaxies, and their fueling, and thus their intrinsic brightness, can be influenced by the environment surrounding the host galaxy, quasars

and other AGN are important for understanding the formation and evolution of structure in the universe.
Similarly, one can select and classify candidates of all types of AGN. If multi-wavelength data are available, the characteristic data mining algorithm's ability to form a model of the required complexity to extract the information could enable it to use the full information to extract more complete AGN samples. More generally, one can classify both normal and active galaxies in one system, differentiating between star formation and AGN.

Photometric redshifts:

An area of astrophysics that has greatly increased in popularity in the last few years is the estimation of redshifts from photometric data (photo-*z*s). This is because, although the distances are less accurate than those obtained with spectra, the sheer number of objects with photometric measurements can often make up for the reduction in individual accuracy by suppressing the statistical noise of an ensemble calculation.
Photo-zs was first demonstrated in the mid 20th century, and later in the 1980s. In the 1990s, the advent of the Hubble Space Telescope Deep fields resulted in numerous approaches. In the past decade, the advent of wide-field CCD surveys and multifiber spectroscopy have revolutionized the study of photo-*z*s to the point where they are indispensable for the upcoming next-generation surveys, and a large number of studies have been made. The two common approaches to photo-*z*s are the template method and the empirical training set method. The template approach has many complicating issues, including calibration, zero-points, priors, multi-wavelength performance (e.g., poor in the mid-infrared), and difficulty handling missing or incomplete training data. We focus in this review on the empirical approach, as it is an implementation of supervised learning. In the future, it is likely that a hybrid method incorporating both templates and the empirical approach will be used, and that the use of full probability density functions will become increasingly important. For many applications, knowing the error distribution in the redshifts is at least as important as the accuracy of the redshifts themselves, further motivating the calculation of PDFs.

Galaxies:

At low redshifts, the calculation of photometric redshifts for normal galaxies is quite straightforward due to the break in the typical galaxy spectrum

at 4000A. Thus, as a galaxy is redshifted with increasing distance, the color (measured as a difference in magnitudes) changes relatively smoothly. As a result, both template and empirical photo-z approaches obtain similar results, a root-mean-square deviation of ~ 0.02 in redshift, which is close to the best possible result given the intrinsic spread in the properties. Another issue at higher redshift is that the available numbers of objects can become quite small (in the hundreds or fewer), thus reintroducing the curse of dimensionality by a simple lack of objects compared to measured wavebands. The methods of dimension reduction can help to mitigate this effect.

Finally, data mining can be performed on astronomical simulations, as well as real datasets. Modern simulations can rival or even exceed real datasets in size and complexity, and as such the data mining approach can be appropriate. An example is the incorporation of theory into the Virtual Observatory. Mining simulation data will present extra challenges compared to observations because in general there are fewer constraints on the type of data presented, e.g., observations are of the same universe, but simulations are not, simulations can probe many astrophysical processes that are not directly observable, such as stellar interiors, and they provide direct physical quantities as well as observational ones. Most of the largest simulations are cosmological, but they span many areas of astrophysics. A prominent cosmological simulation is the Millennium Run, and over 200 papers have utilized its data.

8 Detection of Galaxies

When it comes to the detection of galaxies there are several methods to detect galaxies, but as we were working with something related to the datasets so there are some method which are:

8.1 Detection of Galaxy using image retrieval method

The basic approach of this method is to determine the type of galaxy but also to find a similar image to our selected one. This method uses a nature-inspired algorithm which can be defined as the set of problem-solving methodologies which are derived from natural process some popular examples are genetic algorithm, particle swarm optimization, an algorithm which is used in this process is known as a sine cosine algorithm (SCA).
Computer-based image retrieval is one of the most used which avoids textual descriptions and instead retrieves images based on the similarities that can

be color, texture, shape, and location by using feature extracting method i.e. similarity between pre-saved information with that of queried image.
As we can note that this image retrieval method can be a hectic process because of the comparison of extracted data with pre-existing data. So to make it faster correlation and symmetry functions can be used and one drawback of this is that we can consider either we can select color/texture, color/shape, shape/texture. So to avoid this, what we can do is to extract every data and use a feature selection performed by the sin-cosine algorithm which selects the most relevant features[16].

Sin-cosine algorithm:

$$X_{i,t+1} = \begin{cases} X_{i,t} + r_1 \times \sin(r_2) \times \left| r_3 P_{i,t} - X_{i,t} \right| ; & r_4 < 0.5 \\ X_{i,t} + r_1 \times \cos(r_2) \times \left| r_3 P_{i,t} - X_{i,t} \right| ; & r_4 \geq 0.5 \end{cases} \quad (12)$$

$$i = 1, 2, 3 \ldots PS$$

$$r_1 = a - \frac{a \times t}{T}$$

It initializes the search process randomly and then each candidate solution is updated using the search equation.
Here r_1, r_3, r_2 are known as the random variable which is used to signify a rule by which a real number is assigned to each possible outcome of an experiment.
Here 'i' represents the number of feasible solution
'P' is an elite candidate solution or destination point obtained
r_4 is a uniformly distributed random number in the interval (0,1) which helps in the transition from sin to cos
X_i^t and X_i^{t+r}represents the i^{th} solution vector at t^{th} and $(t+1)^{th}$ iteration.
r_1 random variable that controls the exploration and exploitation during the search process using eq
Random number r_2, j lies in the interval (0, 2) and decides the direction of the moment either towards the current solution or away.
Random number r_3 provides the weight to the 'p' which emphasizes the exploration (r_3>1) and exploitation (r_3<1).
Similar to other nature-inspired algorithms the SCA adopts two conflicting features during the search process exploration/Diversification and exploitation/intensification.
Exploration is the feature where new promising regions of the search space is explored and exploitation refers to the local search where neighborhood

search is performed around the allocated space, decide the moment either towards the current solution(exploitation) or outside the current solution (exploration)

8.2 Red Sequence Method

This method is used for early stages galaxy detection elliptical and lenticular that fall under red sequence giving the relationship between their color and magnitude. It is used for detecting galaxy clusters.
Virgo cluster is one of the largest nearby clusters so within this if we want to find the small clusters it will difficult to detect. [17]George Abell was the first one who observed such a large part of the sky in different regions and detected several galaxies within that particular region i.e. counting the galaxies by hand using photographs from Palomar observatory finding over 4,000 clusters and publishing his data in 1958 called Abell Data.
But the drawback with this method can be that it can lead to false detection of a bunch of galaxies that might be lined up on the line of sight creating a galaxy over density in the sky even if they are not part of a continuous structure. So to overcome this red sequence method was used.
This method focuses not just on visual over densities but also the over density of galaxies at the same redshift. Clusters contain a lot of hot gases that radiates X-ray in a process known as Bremsstrahlung radiation so they can be detected using their X-ray luminosity.

Bremsstrahlung radiation[18]- is also known as decelerating radiation which is electromagnetic radiation produced by the deceleration of a charged particle because of the loss of kinetic energy by particle this loss is released in the form of radiation. It has a continuous spectrum becoming intense and peak intensity shift moves towards higher frequencies as decelerated particle change energy
Every galaxy cluster observed to date contains a well-defined red sequence of galaxies this means that we can potentially find clusters by looking at over densities of galaxies that form a red sequence.
This method starts by simulating what red sequence will look like at various redshift and these simulated reds sequences are compared against data which is pre-existing and all the galaxy in entire sky survey and assign to each other of them a probability they belong to a red sequence of specific redshift which helps in determining how close they are falling exactly on the red sequence line if a bunch of them happen to cluster in the region then we are successfully improving that a compact region of sky contains a believable

population of red sequence galaxies at specific redshift.

9 Challenges in Astronomy

General

- **Statistical inference and visualization with very-large-N datasets**

With the installation of gigantic telescopes in different continents across the world huge-sized data sets are now available to astronomers. Astronomy like nearly every other field of science is now facing exponential growth in the volume, complexity, and even quality of data, both from actual measurements and from numerical simulation processes and phenomena which cannot be addressed in a simple analytical fashion. This scientific and technological challenge to cope with data flow is the foremost and biggest challenge in astronomy. This explosive growth of information has led that the richness of the newly available data can be managed, explored, and analyzed effectively. Improvements in computing and storage will track the growth in data volume Investment in software is critical, and growing. The astronomical community has responded to these challenges with the concept of a Virtual Observatory (VO): a geographically and institutionally distributed, web-based research environment for astronomy with massive and complex data sets, which unifies data archives and other information infrastructure, and computational and data analysis tools for their exploration and analysis. Many considerable works are being conducted by computer scientists and applied mathematicians in other applied fields so that independent development by Astro-statisticians might not be necessary to achieve certain goals. Aside from the computational challenges with large numbers of data vectors and a large dimensionality, this poses some highly non-trivial statistical problems. The problems are driven not just by the size of the data sets, but mainly by the heterogeneity and intrinsic complexity of the data[19].

- **Multivariate analysis with measurement errors and censoring**

The measurement error problem is everywhere in astronomy. Most methods of data mining developed for analyzing massive data sets assume that the data are free from measurement errors, which is far from true for astronomical data. The astronomers measure the physical values of astronomical objects, such as stars, galaxies, galaxy clusters, and black holes, as well as the uncertainty of the measurements. The measurements of uncertainty often

differ between the data points leading to heteroscedasticity and are determined independently of the actual measurements of scientific interest. In astrophysics, the measures of uncertainty are data input rather than model output. Measurement error problems with this type of data are not well treated. Developing techniques that incorporate measurement errors is one of the outstanding issues in measurement error modeling of astronomical data. Astrophysicists often devote as much effort to the precise determination of their errors as they devote to the measurements of the quantities of interest. The instruments are carefully calibrated to reduce systematic uncertainties, and background levels and random fluctuations are carefully evaluated to determine random errors. In astronomy, often, the data are derived quantities from an astrophysical model. Such data are subject to errors that could be large, skewed, and exhibit multiple modes[20].

- **Bayesian computation**

Bayes Theorem and Bayes factors are becoming increasingly well known in astronomical research. Approximate Bayesian Computation represents a powerful methodology for the analysis of complex stochastic systems for which the likelihood of the observed data under an arbitrary set of input parameters may be entirely intractable. Part of the problem is conceptual astronomers need training in how to construct likelihoods for familiar parametric situations. astronomers need methods and software for the oft-complex computations. Many such methods, such as Markov chain Monte Carlo, are already well-established and can be directly adopted for astronomy.

- **Dynamical, Real-Time Classification of Astronomical Transient Events in Synoptic Sky Surveys**

Most scientific measurement and discovery process and method traditionally follows the pattern of theory followed by experiment then analysis of results, and then follow-up experiments, often on time scales from days to decades after the original measurements, feeding back to a new theoretical understanding. For a rapid change that occurs on time scales shorter than what it takes to set up the new round of measurement, there is a need for dynamical, real-time scientific measurement systems, consisting of discovery instruments or sensors, real-time computational analysis and decision engine, and optimized follow-up instruments which can be deployed selectively in real-time and where measurements feedback into the analysis immediately. The time domain is rapidly becoming one of the most exciting new research frontiers in astronomy. The sky is no longer seen as a slow and orderly changing; important physical phenomena are occurring on

scales as short as seconds, whose rapid and appropriate follow-ups promises to broaden substantially our understanding of the physical universe, and perhaps lead to a discovery of previously unknown phenomena. There is a growing number of autonomous robotic telescopes geared to the discovery and follow-up of transient events. Yet, most systems rely on a delayed human judgment in decision making and follow-up of events.

Galaxy classification challenges

- **Surveys of Galaxy Redshifts**:

Measuring galaxy redshift is fundamental for inferring luminosity, but also clustering, gravitational lensing potential, probes of cosmology. The difficulty to construct a truly three-dimensional picture of the universe arises from our ability to first interpret the observed projected distribution of galaxies in the sky and ultimately measure the true location in space of large numbers of objects which depends on our ability to measure the Doppler shift or "redshift" of absorption or emission lines observed in the spectra of galaxies, the observation of which has been greatly advanced in the past decade through technological developments at both radio and optical wavelengths. Application of the redshift-distance relation (Hubble's law) allows the analysis of the large-scale distribution of galaxies. Comparison of the observed redshifts with those expected based on other distance estimates allows mapping of the gravitational field and the underlying density distribution. Estimation of the many inherent selection biases and instrumental limitations is critical in understanding how our view of the universe is affected by our observational perspective and by the way information is received by current technologies. The challenging task was to measure three unknown input parameters used in the simulation: the Hubble constant, the matter density fraction, and the clustering amplitude[21].

- **Analysis of Patterns in Galaxy Clustering**

The current state of play about correlation function analysis, both in terms of computational and sampling problems and concerning the fundamental limitations of such an approach in giving a complete statistical description of the pattern. cosmic "voids" that are large-scale under dense regions have great importance in the analysis of clustered patterns and the *Void Probability Function* is a good example of a statistical characteristic containing more information than the hierarchy of correlation functions. The existence of large

voids also has important consequences for theories of galaxy formation. Various techniques have been devised to attempt to quantify the topology of the clustering pattern with varying degrees of success. For the Assessment of Sub clustering in Clusters of Galaxies, the lack of adequate cluster samples to provide an unbiased look for the problem, the lack of supplemental data in most clusters to confirm or reject possible small-scale structures is also a challenge[22].

- **Star-Galaxy Image Classification**

A classical problem in the analysis of astronomical panoramic imagery. The characteristic resolution of an astronomical image is given through a combination of the instrumental resolution and atmospheric turbulence for the ground-based optical and IR surveys. The accuracy and completeness of the morphological source classification is often the limiting factor in the scientific applications of such data, more stringent than the detection limits. In modern times, the same astronomical source is being imaged many times in different conditions, different filters, etc., and it can get many independent classifications. Alternatively, one could try to perform a classification process using all measured parameters from all imaging passes at once. Another problem of optimized source detection given multiple images in different filters, from different instruments, at different times, etc. A source may be detected with varying degrees of statistical significance in some of them, but not in others; this could be due to the variations in the data quality and/or the intrinsic variability or the spectral energy distribution; in many cases, non-detections also provide useful information. One has to evaluate a joint significance for multiple detections.

Challenges in radio astronomy

- **Gravitational lensing**

Gravitational lensing occurs when the light emitted by distant galaxies passes by massive objects in the universe, the gravitational pull from these objects might distort or bend the light. Strong gravitational lensing can result in such strongly bent light that multiple images of the light-emitting galaxy are formed. Weak gravitational lensing results in galaxies appearing distorted, stretched, or magnified. Although difficult to measure for an individual galaxy, galaxies clustered close together will exhibit similar lensing

patterns. Analyzing the nature of gravitational lensing patterns tells astronomers about the way dark matter is distributed within galaxies and their distance from Earth. For weak lensing, the main challenge is that the measurements are very difficult to acquire and analyze. There is an increasing need for computer science, machine learning, signal processing, and image processing to bring enormous advantage if channeled into lensing studies.

- **Electronic Data Challenges**

Large telescope surveys will generate data volumes that are too large to transport and process. To keep up with the data flow from the telescope, data processing must be done in near-real-time, which requires that it must be automated. Some major electronic major challenges are

1. Large bandwidth from telescope to processor: ∼10 Tb/s from antennas to correlator (< 6 km), 40 Gb/s from correlator to processor (∼ 600 km)
2. Large processing power: 750 Tflop/s expected/budgeted, 1 Pflop desired
3. Pipeline processing essential: Including data validation, source extraction,

   cross-identification, etc
4. Power consumption of processors: 1 MW at the site, 10 MW for the processor, Power bill ∼3M p.a
5. Storage and curation of data: 70 PB/yr. if all products are kept, 5 PB/yr. with current funding, 8 h to write 12 h of data to disk at 10GB/s
6. Retrieval of data by users: All data in the public domain accessed using VO tools & services
7. Data-intensive research: Data mining, stacking, cross-correlation, etc.

- **DATA ANALYSIS AND CROSS-IDENTIFICATION**

As data volumes become unmanageable for the computing resources of an individual astronomer, volumes of processed information become unmanageable. Comparison of radio and optical images has traditionally been done by eye, but this will become impossible when the catalog size is so huge as in the range of millions. The science will only flow from these catalogs when we cross-identify them with optical/infrared objects from surveys. We expect that 70% of the cross-matchings will be performed automatically with a simple algorithm which compares catalogs for simple isolated objects, and

assigns probabilities to cross-matches based on their separation, their properties (such as brightness, color, or polarisation), and our prior knowledge of astrophysical objects. But for extremely large data sets, the time constraint is an important factor that we need to look into. Also, for extended or multiple objects, the process will be much harder, and at present, we need the strongest algorithms that can handle all these tens of millions of cross-identifications. Next-generation radio telescopes offer us a profound change in the way that we do astronomy but they will also require a level of computing power. No doubt these are enormous challenges ahead.
Our aim from the project was to work on the biggest challenge in astronomy i.e. challenges in data-rich astronomy. We were obvious to find a better faster algorithm that could be implied in astronomy and then to use the algorithm to crossmatch between any two large sets of catalogs and to further study those galaxies.

10 Data Analysis

Data

Factual information such as measurement, used as a basis for reasoning, discussion, or calculation. Data is nothing but information output by a sensing device like telescopes that are sensitive to a particular range of frequency.

Astronomical Catalogues

An astronomical data consist of information about astronomical objects that are grouped together since they share common origin or morphology.
When we talk about data, we are limiting ourselves to one particular perspective. Data is a vast division available as an open source all around the internet. Data has been a crucial entity in almost every professional sector. Be it the marketing sector, public sector, hospitals, schools, colleges, local as well as international banking and security, agriculture and farming, communication sector and last but not the least, the Space sector.
In Astronomy, we theorize as per the data available. Scientists rely on data in order to understand scientific phenomena so that ultimately they can draw conclusions about scientific occurrences. According to Geoffrey Moore, "without big data, you are blind and deaf and in the middle of a free way". But, processing that huge data takes a lot of time as well as meticulous work. In the past, it was difficult to keep tracks of all the observations made by the scientists. Thanks to the Supercomputers and modern technology, we are able to do it faster and more efficiently.
There are 2 types of data, one is quantitative and the other is qualitative.

- Quantitative data is the sort of data that involves numerical quantities, such as the magnitude of an earthquake, the amount of rainfall an area receives, or the average height of professional basketball players.
- Qualitative data is any form of data that does not involve numbers. It is data that can easily be observed with the human eye. This includes observations of properties like shape, color, texture, location, or other non-numerical details of a subject. Common qualitative earth science observations include data found on maps, the shape of a stream's channel, or the color of a particular mineral.

The cosmos is huge, and so is its data. We need a faster algorithm to keep track of all the data sent at its input end. To do so, we came up with the idea of kd-tree algorithm. We planned to do cross-matching between optical and radio catalogues of data for a particular galaxy in order to get information about its redshift, its location in the cosmos and to identify its various constituents.

Purpose of using Radio and Optical

Radio waves are emitted by clouds of dust and molecules of gas. Due to this property of radio waves we can come to know the amounts of gas or dust present in the galaxy. Since we know the amount of gas and dust present in the galaxy we can classify the galaxy based on the Hubble's classification.In our project we used Radio data from Giant Meterwave Radio Telescope(GMRT), taken from VizieR[23].

Optical data for galaxies is widely available on the internet and even in huge amount from various sources like the SkyServer, SDSS(Sloan Digital Sky Survey), Vizier, Kaggle, NASA and many more. In our project, we particularly used the optical data set of galaxies from SDSS i.e. Sloan Digital Sky Survey. The catalogue we used was uploaded in 2020 so that our research would be quite updated. We also used a radio catalogue which we got from Vizier. Using SQL query, we created a catalogue of around 4.5 lakhs from the Sloan Digital Sky Survey's 16^{th} data release i.e., DR16. The SQL query is added in the section below[24]:

```
-- This query does a table JOIN between the imaging (PhotoObj) and spectra
-- (SpecObj) tables and includes the necessary columns in the SELECT to upload
-- the results to the SAS (Science Archive Server) for FITS file retrieval.
SELECT TOP 500000
   p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,zErr,
   q_u,q_i,q_r,q_g,q_z,u_u,u_i,u_r,u_g,u_z,
   petroR50_u, petroR50_r, petroR50_z, petroR90_u, petroR90_r, petroR90_z,
   mCr4_u,mCr4_g,mCr4_r, mCr4_i,mCr4_z,
   p.run, p.rerun, p.camcol, p.field,
```

```
    s.specobjid, s.class,s.subclass, s.z as redshift,
    s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
    JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
    class = 'GALAXY'
    AND s.z BETWEEN 0 AND 2
    AND p.u BETWEEN 0 AND 19.6
    AND g BETWEEN 0 AND 20
```

Implementation of Data

Problem Statement: Cross-matching between optical and radio catalogues for classification of Elliptical Galaxies using Machine Learning Techniques. To find the solution to our problem statement we divided into 3 parts to achieve the output:

1. Cross Matching Catalogues
2. Plotting the Red shift
3. Classification of elliptical galaxies

We use *positional cross-matching* to find the closest counterpart within a given radius on the sky.

1. This is done through radio and optical catalogues, radio one lists the brightest sources while optical one lists the galaxies observed by visible light surveys.
2. Compare the angular distance between objects on the celestial sphere to crossmatch

If we have an object on the celestial sphere with right ascension and declination is:

$$d = 2arcsin\sqrt{sin^2\frac{|\delta_1 - \delta_2|}{2} + cos\delta_1 cos\delta_2 sin^2\frac{|\alpha_1 - \alpha_2|}{2}} \tag{13}$$

Or,

$$d = 2arcsin\sqrt{a+b} \tag{14}$$

where, $a = sin^2\frac{|\delta_1-\delta_2|}{2}$ and $b = cos\delta_1 cos\delta_2 sin^2\frac{|\alpha_1-\alpha_2|}{2}$
Implying faster algorithms to match large catalogues.

The about optical and radio catalogues and each of its parameters. Also, about time complexity challenges while using a naive cross matching algorithm.

Studied about a better faster algorithm named 'kd-tree' and implied the algorithm into our large radio and optical catalogues and obtained the cross matched results.
1. Using Astropy module and KD-tree Data Structure to finalize and efficient algorithm
2. Increased the time complexity of the program to a great extent.
3. Landed up with the best match
4. The Matched data obtained was used to train over ML model.

KD Tree

A K-D Tree (also called as K-Dimensional Tree) is a binary search tree where data in each node is a K-Dimensional point in space. In short, it is a space partitioning (details below) data structure for organizing points in a K-Dimensional space.
A non-leaf node in K-D tree divides the space into two parts, called as half-spaces.
Points to the let of this space are represented by the left sub-tree of that node and points to the right of the space are represented by the right sub-tree. Repeat 2-3 until each partition only has one.

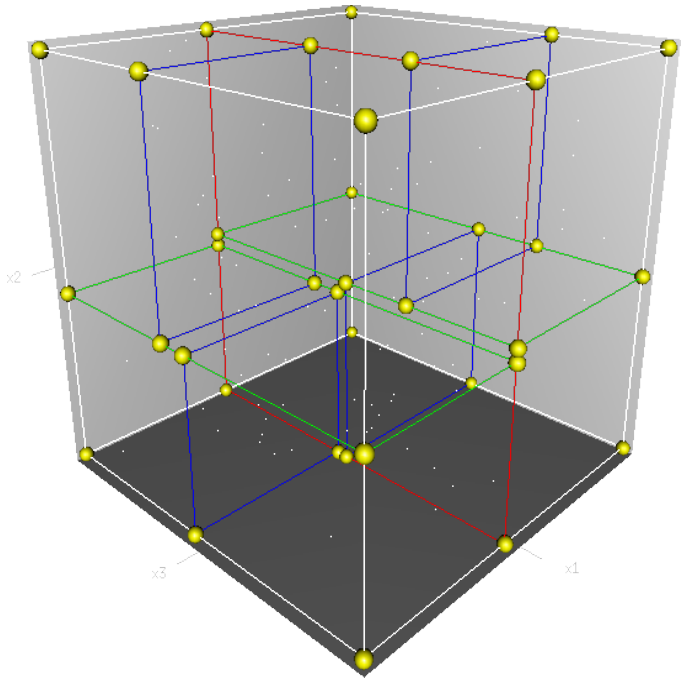


Figure 4: kd-tree[18]

Plotting the Red shift

Backdrop:

We made use of decision trees to determine the redshifts of galaxies from their photometric colours

According to the Big Bang theory, the fabric of space has been expanding ever since the big bang occured, so that distant galaxies appear to be moving away from us at very fast speed. This results in the continuous expansion of the universe and because of this distant Galaxies appears to be red - shifted which means that their photons are shifted to lower frequencies.

To measure redshift, according to cosmological theory, we use a parameter z with observed wavelength λ_{obs} and emitted wavelength λ_{em}.

$$\lambda_{obs} = (1 + z)\,\lambda_{cm} \tag{15}$$

Why estimating Redshift:

The photometric redshift problem is very important. To view of the universe as a whole: its history, its geometry, and its fate we need to obtaining an accurate estimate of the redshift of galaxies. As it even gives images of extremely faint galaxies, for which it's difficult to obtain a spectrum.
Hence, to advance our understanding of the Universe and the dark energy that is currently accelerating the cosmic expansion photometric redshift coding is needed.

Dataset

We used flux magnitude in five frequency bands *u*, *g*, *r*, i and z scraped from Sloan Digital Sky Survey to get color index or astronomical color which is the difference between the magnitudes of two filters, i.e. *u* - *g* or *i* - *z*. This is a way to categorize the galaxies.

u	g	r	I	z	...	redshift
19.84	19.53	19.47	19.18	19.11	...	0.54
19.86	18.66	17.84	17.39	17.14	...	0.16
...	...	...	...	...	...	...
18.00	17.81	17.77	17.73	17.73	...	0.47

Code Insights

Inputs/Features: Colour Indices From Photometric Imaging
ML Algorithm: Decision Tree Regressor
Output/Targets: Photometric Redshift.
Training Set: Spectroscopic Redshifts

Estimating Accuracy of Model

We use the median of the differences between our predicted and actual values to get to know how well our model performs and this is given by:
Med_diff = median (| Y(i, pred) - Y(i, actu)|) (**16**)
Median of differences gives a fair representation of the errors especially when the distribution of errors is skewed.

Classifying the Elliptical galaxies using machine learning algorithm

Backdrop:
In the next iteration for model improvement we used Classification to approach our problem rather than regression, the only difference here is our targets are classes rather than real values.

Model:
Features

- **colours:** u-g, g-r, r-i, and i-z;
- **4th adaptive moments:** mCr4_u, mCr4_g, mCr4_r, mCr4_i, and mCr4_z;
- **50% Petrosian:** petroR50_u, petroR50_r, petroR50_z;
- **90% Petrosian:** petroR90_u, petroR90_r, petroR90_z.

Parameter description:

1. Colour Indices : u-g, g-r, r-i, and i-z

2. Ellipticity: E0, E1, E2, E3, E4, E5, E6, E7 according to Hubble type classification. The ellipticity by stoke parameters is given by

$$e = 1 - \frac{b}{a} = 1 - \frac{1 - \sqrt{Q^2 + U^2}}{1 - \sqrt{Q^2 - U^2}} \tag{16}$$

3. Adaptive moments: It describes the shape of a galaxy. They are used in image analysis to detect similar objects at different sizes and orientations. We use the fourth moment here for each band.

4. Petrosian flux: The Petrosian method allows us to compare the radial profiles of galaxies at different distances

5. Concentrations: It is the luminosity profile of the galaxy, which measures what proportion of a galaxy's total light is emitted in a radius.

$$conc = \frac{petro_{R50}}{petro_{R90}} \tag{17}$$

Accuracy:

The Accuracy of the classification is simpler than for regression problems. The simplest measure is the fraction of objects that are correctly classified. That is

$$accuracy = \frac{correct\ predictions}{predictions} \tag{18}$$

$$accuracy = \frac{\sum_{i=1}^{n} predicted_i = actual_i}{n} \tag{19}$$

The accuracy measure is often called the model score. Accuracy is the important parameter in classification techniques which tells the probability of the predicted value is equivalent to the actual value.

Codes used for classification of elliptical galaxies

1. **Code for Cross matching the**

```
import numpy as np
import statistics
import time
from astropy.coordinates import SkyCoord
from astropy import units as u
import pandas as pd
def crossmatch(cat1, cat2, max_dist):
matches = []
nomatches = []
start = time.perf\_counter()
skycat1 = SkyCoord(cat1*u.degree, frame='icrs')
skycat2 = SkyCoord(cat2*u.degree, frame='icrs')
closest_ids, closest_dists, closest_dists3d =   skycat1.match_to_catalog_sky(
    skycat2)
closest_dists_deg = closest_dists.value
for cat1idx in range(len(cat1)):
if closest_dists_deg[cat1idx] > max_dist:
nomatches.append(cat1idx)
else:
matches.append((cat1idx,closest_ids[cat1idx],closest_dists_deg[cat1idx]))
#closest_dist.value returns an array of degrees
#print("vals", closest_ids)
```

```
#print("dists", closest_dists)
#print("dists.val", closest_dists.value)
return (matches, nomatches, time.perf_counter() - start)
#to create a text file
def conv(match):
match_1 = matches
f = open('newfiles.txt', 'w')
for t in match_1:
line = ' '.join(str(x) for x in t)
f.write(line + '{\textbackslash}n')

f.close()
if __name__ == '__main__':
# The example in the question
cat1 = np.genfromtxt('gmrt.csv', delimiter=',',skip\_header=55,usecols=[5,6],max
    \_rows=5434)
cat2 = np.genfromtxt('opticaldata.csv', delimiter=',',skip\_header=1,usecols
    =[1,2],max\_rows=252198)
matches, no\_matches, time\_taken = crossmatch(cat1, cat2, 5)
text = conv(matches)
print('matches:', matches)
print('unmatched:', no\_matches)
print('time taken:', time\_taken)
```

Output : We find the optical dataset that could be used to find redshift

2. Plotting Red shift

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt

# Complete the following to make the plot
if __name__ == "__main__":
    data = pd.read_csv('opticaldatafinal_SDSS.csv')
    # Get a colour map
    cmap = plt.get_cmap('YlOrRd')

    # Define our colour indexes u-g and r-i
    u_g = data['u'] - data['g']
    r_i = data['r'] - data['i']

    # Make a redshift array
    redshift = data['redshift']

    # Create the plot with plt.scatter
    plot = plt.scatter(u_g, r_i, s=0.5, lw=0, c=redshift, cmap=cmap)

    cb = plt.colorbar(plot)
    cb.set_label('Redshift')

    # Define your axis labels and plot title
    plt.xlabel('Colour index  u-g')
    plt.ylabel('Colour index  r-i')
    plt.title('Redshift (colour) u-g versus r-i')

    # Set any axis limits
```

```
    plt.xlim(-0.5, 2.5)
    plt.ylim(-0.5, 1)
    plt.show()
```

Output :

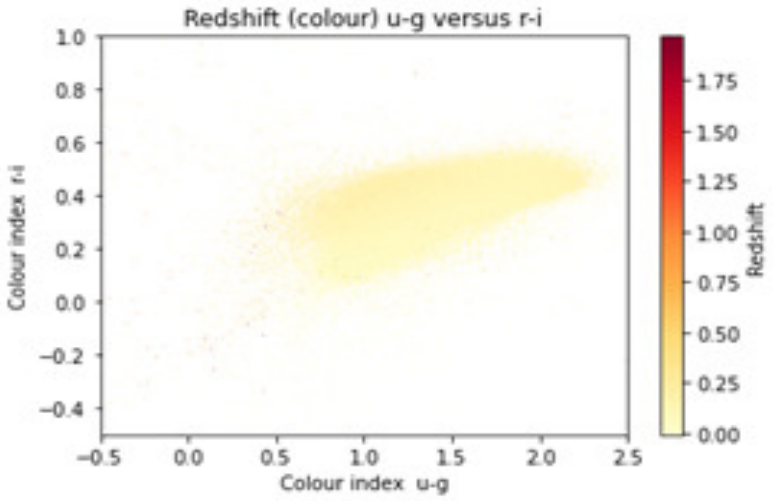


Figure 5: Colour-Colour redshift plot

3. To classify the Elliptical galaxies using machine learning algorithm :

3.1. Overfitting Trees

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from sklearn.tree import DecisionTreeRegressor

# paste your get_features_targets function here
def get_features_targets(data):
  features = np.zeros((data.shape[0], 4))
  features[:, 0] = data['u'] - data['g']
  features[:, 1] = data['g'] - data['r']
  features[:, 2] = data['r'] - data['i']
  features[:, 3] = data['i'] - data['z']
  targets = data['redshift']
  return features, targets

# paste your median_diff function here
def median_diff(predicted, actual):
  return np.median(np.abs(predicted - actual))

# Complete the following function
def accuracy_by_treedepth(features, targets, depths):
  # split the data into testing and training sets
  split = features.shape[0]//2
  train_features, test_features = features[:split], features[split:]
  train_targets, test_targets = targets[:split], targets[split:]

  # Initialise arrays or lists to store the accuracies for the below loop
  train_diffs = []
  test_diffs = []
```

```
  # Loop through depths
  for depth in depths:
    # initialize model with the maximum depth.
    dtr = DecisionTreeRegressor(max_depth=depth)

    # train the model using the training set
    dtr.fit(train_features, train_targets)

    # Get the predictions for the training set and calculate their med_diff
    predictions = dtr.predict(train_features)
    train_diffs.append(median_diff(train_targets, predictions))

    # Get the predictions for the testing set and calculate their med_diff
    predictions = dtr.predict(test_features)
    test_diffs.append(median_diff(test_targets, predictions))

  # Return the accuracies for the training and testing sets
  return train_diffs, test_diffs

if __name__ == "__main__":
    data = pd.read_csv('opticaldatafinal_SDSS.csv')
    features, targets = get_features_targets(data)
  # Generate several depths to test
    tree_depths = [i for i in range(1, 36, 2)]
  # Call the function
    train_med_diffs, test_med_diffs = accuracy_by_treedepth(features, targets,
   tree_depths)
    print("Depth with lowest median difference : {}".format(tree_depths[
   test_med_diffs.index(min(test_med_diffs))]))
  # Plot the results
    train_plot = plt.plot(tree_depths, train_med_diffs, label='Training set')
    test_plot = plt.plot(tree_depths, test_med_diffs, label='Validation set')
    plt.xlabel("Maximum Tree Depth")
    plt.ylabel("Median of Differences")
    plt.legend()
    plt.show()
```

Output :

```
Depth with lowest median difference : 13
```

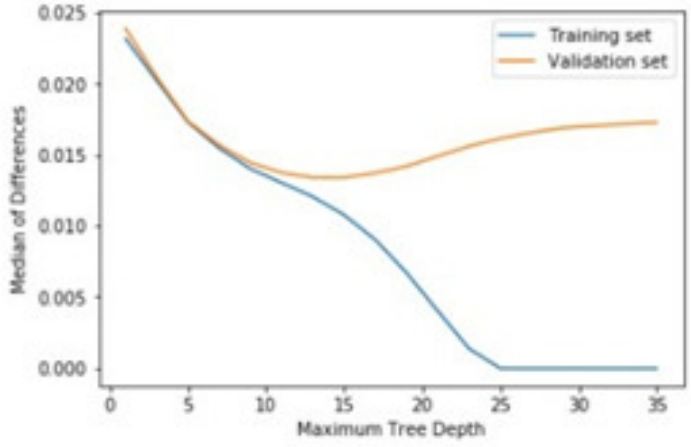


Figure 6: Overfitting Trees

3.2. KFold Cross Validation

```
import numpy as np
import pandas as pd
from sklearn.model_selection import KFold
from sklearn.tree import DecisionTreeRegressor
from matplotlib import pyplot as plt

def get_features_targets(data):
  features = np.zeros((data.shape[0], 4)) #n lines, 4 columns
  features[:,0] = data['u'] - data['g']
  features[:,1] = data['g'] - data['r']
  features[:,2] = data['r'] - data['i']
  features[:,3] = data['i'] - data['z']
  targets = data['redshift']
  return (features, targets)

def median_diff(predicted, actual):
  diff = np.median(np.absolute(predicted - actual))
  return diff

# complete this function
def cross_validate_model(model, features, targets, k):
  kf = KFold(n_splits=k, shuffle=True)

  # initialise a list to collect median_diffs for each iteration of the loop
   below
  mediandiffs = []

  for train_indices, test_indices in kf.split(features):
    train_features, test_features = features[train_indices], features[
   test_indices]
    train_targets, test_targets = targets[train_indices], targets[test_indices]

    # fit the model for the current set
    model.fit(train_features, train_targets)

    # predict using the model
    predictions = model.predict(test_features)

    # calculate the median_diff from predicted values and append to results
   array
    mediandiffs.append(median_diff(test_targets, predictions))

  # return the list with your median difference values
  return mediandiffs

if __name__ == "__main__":
  data = pd.read_csv('opticaldatafinal_SDSS.csv')
  features, targets = get_features_targets(data)

  # initialize model with a maximum depth of 19
  dtr = DecisionTreeRegressor(max_depth=19)

  # call your cross validation function
  diffs = cross_validate_model(dtr, features, targets, 10)

  # Print the values
  print('Differences: {}'.format(', '.join(['{:.3f}'.format(val) for val in
   diffs])))
  print('Mean difference: {:.3f}'.format(np.mean(diffs)))
```

Output :

Differences: 0.014, 0.014, 0.014, 0.014, 0.014, 0.014, 0.013, 0.014, 0.014, 0.014

Mean difference: 0.014

3.3. KFold Cross Validated Predictions

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
from sklearn.model_selection import KFold
from sklearn.tree import DecisionTreeRegressor

# paste your get_features_targets function here
def get_features_targets(data):
  features = np.zeros((data.shape[0], 4))
  features[:, 0] = data['u'] - data['g']
  features[:, 1] = data['g'] - data['r']
  features[:, 2] = data['r'] - data['i']
  features[:, 3] = data['i'] - data['z']
  targets = data['redshift']
  return features, targets

# paste your median_diff function here
def median_diff(predicted, actual):
  return np.median(np.abs(predicted - actual))

# complete this function
def cross_validate_predictions(model, features, targets, k):
  kf = KFold(n_splits=k, shuffle=True)

  # declare an array for predicted redshifts from each iteration
  all_predictions = np.zeros_like(targets)

  for train_indices, test_indices in kf.split(features):
    # split the data into training and testing
    train_features, test_features = features[train_indices], features[
    test_indices]
    train_targets, test_targets = targets[train_indices], targets[test_indices]

    # fit the model for the current set
    model.fit(train_features, train_targets)

    # predict using the model
    predictions = model.predict(test_features)

    # put the predicted values in the all_predictions array defined above
    all_predictions[test_indices] = predictions

  # return the predictions
  return all_predictions


if __name__ == "__main__":
  data = pd.read_csv('opticaldatafinal_SDSS.csv')
  features, targets = get_features_targets(data)
```

```
# initialize model
dtr = DecisionTreeRegressor(max_depth=19)

# call your cross validation function
predictions = cross_validate_predictions(dtr, features, targets, 10)

# calculate and print the rmsd as a sanity check
diffs = median_diff(predictions, targets)
print('Median difference: {:.3f}'.format(diffs))

# plot the results to see how well our model looks
plt.scatter(targets, predictions, s=0.4)

plt.xlim((0, targets.max()))
plt.ylim((0, predictions.max()))

plt.xlabel('Measured Redshift')
plt.ylabel('Predicted Redshift')

plt.show()
```

Output :

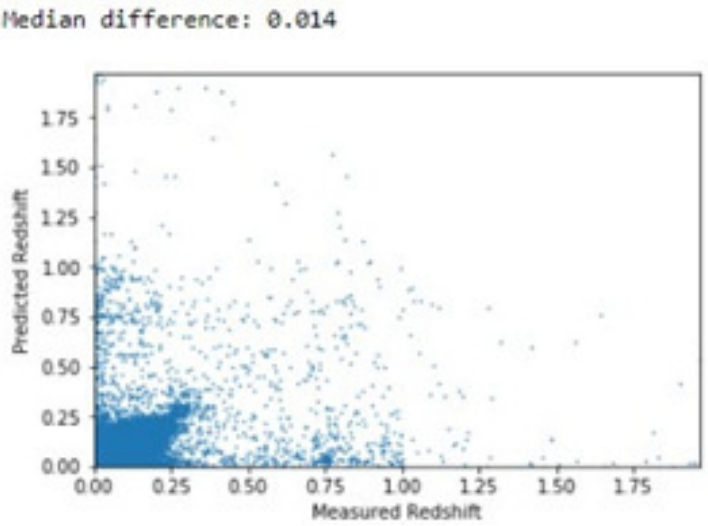


Figure 7: KFold Cross Validation Predictions

3.4 Splitting the Train and Test Sets

```
import numpy as np

import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split

data = pd.read_csv('opticaldatafinal_SDSS.csv',delimiter=",")


fraction_training = 0.7
  # split the data using your function
training,testing = train_test_split(data, test_size = 0.7, random_state = 0)

  # print the key values
print('Number data galaxies:', len(data))
print('Train fraction:', fraction_training)
print('Number of galaxies in training set:', len(training))
print('Number of galaxies in testing set:', len(testing))
```

Output :

Number data galaxies: 370082
Train fraction: 0.7
Number of galaxies in training set: 111024
Number of galaxies in testing set:259058

3.5. Generating Features and Targets

```
import numpy as np
import pandas as pd

def generate_features_targets(data):
  # complete the function by calculating the concentrations

  targets = data['class']

  features = np.empty(shape=(len(data), 13))
  features[:, 0] = data['u-g']
  features[:, 1] = data['g-r']
  features[:, 2] = data['r-i']
  features[:, 3] = data['i-z']
  features[:, 4] = data['ellipticity']
  features[:, 5] = data['mCr4_u']
  features[:, 6] = data['mCr4_g']
  features[:, 7] = data['mCr4_r']
  features[:, 8] = data['mCr4_i']
  features[:, 9] = data['mCr4_z']

  # fill the remaining 3 columns with concentrations in the u, r and z filters
  # concentration in u filter
  features[:, 10] = data['petroR50_u']/data['petroR90_u']
```

```
  # concentration in r filter
  features[:, 11] = data['petroR50_r']/data['petroR90_r']
  # concentration in z filter
  features[:, 12] = data['petroR50_z']/data['petroR90_z']

  return features, targets

if __name__ == "__main__":
  data = pd.read_csv('opticaldatafinal_SDSS.csv')

  features, targets = generate_features_targets(data)

  # Print the shape of each array to check the arrays are the correct dimensions
    .
  print("Features shape:", features.shape)
  print("Targets shape:", targets.shape)
```

Output : Features shape: (370082, 13)
Targets shape: (370082)

3.6. Train the Decision Tree Classifier

```
import numpy as np
import pandas as pd
import math
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split

def generate_features_targets(data):
    targets = data['class1']

    features = np.empty(shape=(len(data), 13))
    features[:, 0] = data['u-g']
    features[:, 1] = data['g-r']
    features[:, 2] = data['r-i']
    features[:, 3] = data['i-z']
    features[:, 4] = data['ellipticity']
    features[:, 5] = data['mCr4_u']
    features[:, 6] = data['mCr4_g']
    features[:, 7] = data['mCr4_r']
    features[:, 8] = data['mCr4_i']
    features[:, 9] = data['mCr4_z']
    features[:, 10] = data['petroR50_u']/data['petroR90_u']
    features[:, 11] = data['petroR50_r']/data['petroR90_r']
    features[:, 12] = data['petroR50_z']/data['petroR90_z']
    return features, targets

# complete this function by splitting the data set and training a decision tree
    classifier
def dtc_predict_actual(data):
    training,testing = train_test_split(data, test_size = 0.7,random_state = 0)
    features_training, targets_training = generate_features_targets(training)
    features_testing, targets_testing = generate_features_targets(testing)
    dtc = DecisionTreeClassifier()
    dtc.fit(features_training, targets_training)
    predictions = dtc.predict(features_testing)
    return predictions, targets_testing
```

```
35
36
37 if __name__ == '__main__':
38     data = pd.read_csv('opticaldatafinal_SDSS.csv')
39     predicted_class, actual_class = dtc_predict_actual(data)
40
41   # Print some of the initial results
42     print("Some initial results...\n   predicted,  actual")
43     for i in range(1):
44         print("{}. {}, {}".format(i, predicted_class[i], actual_class[i]))
```

Output :

Some initial results...
predicted, actual
0. E2, E4

3.7. Accuracy in Classification

```
1  import numpy as np
2  import pandas as pd
3  from matplotlib import pyplot as plt
4  from sklearn.metrics import confusion_matrix
5  from sklearn.model_selection import cross_val_predict
6  from sklearn.ensemble import RandomForestClassifier
7  import itertools
8
9
10
11 def calculate_accuracy(predicted_classes, actual_classes, ):
12     return sum(actual_classes[:] == predicted_classes[:]) / len(actual_classes)
13
14
15 def generate_features_targets(data):
16     output_targets = np.empty(shape=(len(data)), dtype='<U20')
17     output_targets[:] = data['class1']
18
19     input_features = np.empty(shape=(len(data), 13))
20     input_features[:, 0] = data['u-g']
21     input_features[:, 1] = data['g-r']
22     input_features[:, 2] = data['r-i']
23     input_features[:, 3] = data['i-z']
24     input_features[:, 4] = data['ellipticity']
25     input_features[:, 5] = data['mCr4_u']
26     input_features[:, 6] = data['mCr4_g']
27     input_features[:, 7] = data['mCr4_r']
28     input_features[:, 8] = data['mCr4_i']
29     input_features[:, 9] = data['mCr4_z']
30     input_features[:, 10] = data['petroR50_u'] / data['petroR90_u']
31     input_features[:, 11] = data['petroR50_r'] / data['petroR90_r']
32     input_features[:, 12] = data['petroR50_z'] / data['petroR90_z']
33
34     return input_features, output_targets
35
36
```

```
def plot_confusion_matrix(cm, classes, normalize=False,title='Confusion matrix',
    cmap=plt.cm.Blues):
    """
    This function prints and plots the confusion matrix.Normalization can be
    applied by setting `normalize=True`.
    """
    plt.imshow(cm, interpolation='nearest', cmap=cmap)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)

    if normalize:
        cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        print("Normalized confusion matrix")
    else:
        print('Confusion matrix, without normalization')

    print(cm)

    thresh = cm.max() / 2.
    for i, j in itertools.product(range(cm.shape[0]), range(cm.shape[1])):
        plt.text(j, i, "{}".format(cm[i, j]),horizontalalignment="center",color=
    "white" if cm[i, j] > thresh else "black")
    plt.tight_layout()
    plt.ylabel('True Class')
    plt.xlabel('Predicted Class')
#     return cm,classes


# complete this function to get predictions from a random forest classifier
def rf_predict_actual(data, n_estimators):
  # generate the features and targets
    features, targets = generate_features_targets(data)

  # instantiate a random forest classifier using n estimators
    rfc = RandomForestClassifier(n_estimators=n_estimators)

  # get predictions using 10-fold cross validation with cross_val_predict
    predicted = cross_val_predict(rfc, features, targets, cv=10)

      # return the predictions and their actual classes
    return predicted, data['class1']

if __name__ == "__main__":
    data = pd.read_csv('opticaldatafinal_SDSS.csv')

  # get the predicted and actual classes
    number_estimators = 50              # Number of trees
    predicted, actual = rf_predict_actual(data, number_estimators)

      # calculate the model score using your function
    accuracy = calculate_accuracy(predicted, actual)
    print("Accuracy score:", accuracy)

      # calculate the models confusion matrix using sklearns confusion_matrix
    function
    class_labels = list(set(actual))
    model_cm = confusion_matrix(y_true=actual, y_pred=predicted, labels=
    class_labels)
```

```
    # plot the confusion matrix using the provided functions.
plt.figure()
plot_confusion_matrix(model_cm, classes=class_labels, normalize=False)
plt.show()
```

Output :

Accuracy Score : 0.8712

```
[[ 51756      4      0      0      0      0      6      0]
 [     2  74820      0      0      0      0      0      1]
 [     1      0   5518      2      3    388     31      0]
 [     0      0    275    553     10     53     11      0]
 [     5      0     21     16  82949     18     15      0]
 [     6      0      2      0      4  19164    122      0]
 [    10      0      0      0      1      5  33969      0]
 [     0      1      0      0      1      0      0 100339]]
```

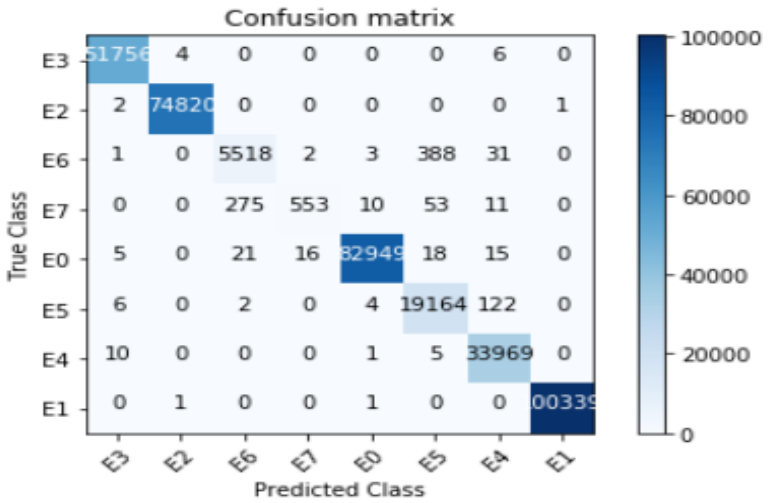


Figure 8: Confusion Matrix

Types	No. of Galaxies
E0	82826
E1	100339
E2	74823
E3	51765
E4	33985
E5	19298
E6	5943
E7	902
FALSE	107

We have learnt that assessing the models performance is a lot simpler with classification than it is with regression.Confusion matrices can be a useful

tool to help understand where our model is over and under preforming with respect to each class. Above is the following table of the output we get from the conclusion table. ”E1” type of galaxies are more in our dataset.

11 Conclusion

In this internship we investigated the cross-match between optical and radio catalogs for the classification of elliptical galaxies using machine learning techniques. So, the first aim for the project was to work on the biggest challenge in astronomy i.e. challenges in data rich astronomy. We were obvious to find a better faster algorithm that could be implied in astronomy and then to use the algorithm to crossmatch between any two large sets of catalogues and to further study those galaxies.

Crossmatching basically means searching i.e. to find a counterpart of each object in 2nd catalogue. By surveying data of different wavelength telescopes like optical, radio, UV or x-rays, we get different characteristics about the galaxies. But only by combining all this information together, we will have a complete idea of the galaxy we are studying.

As crossmatching means to search so, we studied about the challenges in using a naive crossmatching algorithm. It might seem pretty simple to crossmatch the telescope catalogues but if we start with naive matching algorithm, as we have a million of galaxies, it might take months to complete crossmatching. So, we worked on finding better algorithms that can work with very large dataset and resolve the time complexity problem.

We sorted it out with an algorithm called KD Tree (K-Dimensional Tree). KD tree is a space partitioning data structure for organizing points in a k-dimensional space. This data structure algorithm due to its k-dimensional characteristics has substantial application in astronomy. The algorithm was implemented to the radio and optical catalogue taking two input datasets namely, ‘right ascension’, ‘declination’ and the crossmatched results were obtained.

Galaxies can be classified on the basis of their morphology, color and shape. In 1936, Edward Hubble developed a classification scheme to divide galaxies into distinct categories based upon morphological features. This system, which would eventually become known as the Hubble Sequence, consists of 4 major categories of galaxies: Elliptical, Lenticular, Barred Spiral, and Spiral.

For classification of galaxies we will use the below given features:

1) Color
2) Luminosity profile
3) Ellipticity

In classification, the predictions are from a fixed set of classes, whereas in regression the prediction typically corresponds to a continuum of possible values.
In regression, we measure accuracy by looking at the size of the differences between the predicted values and the actual values. In contrast, in classification problems a prediction can either be correct or incorrect. This makes measuring the accuracy of our model a lot simpler.
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. Also, Random forests help to mitigate overfitting in decision trees.
One common way to represent the results from machine learning and answer questions like these is a confusion matrix.
A confusion matrix visualizes the relationship between the true labels from our gold standard and the predicted labels from our classifier.
A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.So we took about 360532 Radio catalogue, 450548 Optical catalogue and the total number of matches is 370328.The accuracy that we obtained using our machine learning algorithm was 87.12%

12 Future Scope

Using Random Forest: Individual decision trees are trained with different subsets of features. So in our current problem, one tree might be trained using eccentricity and another using concentration and the 4th adaptive moment. By using different combinations of input features you create expert trees that are can better identify classes by a given feature.

1. Increasing the amount of data can improve the results
2. Adding more features
3. Using improved ML models

References

[1] https://en.wikipedia.org/wiki/Stellar_evolution#:~:text=Stellar%20evolution%20is%20the%20process,the%20age%20of%20the%20universe.

[2] Wikipedia. Top-down formation. http://csep10.phys.utk.edu/0JTA2dev/ojta/course2/early/formation/topdown_tl.html, 2020.

[3] Bottom-up formation. http://csep10.phys.utk.edu/0JTA2dev/ojta/c2c/early/formaion/bottomup_tl.html.

[4] Ke Shi, Kyoung-Soo Lee, Arjun Dey, Yun Huang, Nicola Malavasi, Chao-Ling Hung, Hanae Inami, Matthew Ashby, Kenneth Duncan, Rui Xue, et al. A census of galaxy constituents in a coma progenitor observed at z¿ 3. *The Astrophysical Journal*, 871(1):83, 2019.

[5] George R Blumenthal, SM Faber, Joel R Primack, and Martin J Rees. Formation of galaxies and large-scale structure with cold dark matter. *Nature*, 311(5986):517–525, 1984.

[6] Dominik J Bomans, Y-H Chu, and Ulrich Hopp. Hot interstellar gas in the irregular galaxy ngc 4449. *arXiv preprint astro-ph/9702121*, 1997.

[7] Kenji Bekki. Dust-regulated galaxy formation and evolution: a new chemodynamical model with live dust particles. *Monthly Notices of the Royal Astronomical Society*, 449(2):1625–1649, 2015.

[8] K Brecher and GR Burbidge. Extragalactic cosmic rays. *The Astrophysical Journal*, 174:253, 1972.

[9] Somnath Bharadwaj and Sayan Kar. Modeling galaxy halos using dark matter with pressure. *Physical Review D*, 68(2):023516, 2003.

[10] Rachel M Reddick, Risa H Wechsler, Jeremy L Tinker, and Peter S Behroozi. The connection between galaxies and dark matter structures in the local universe. *The Astrophysical Journal*, 771(1):30, 2013.

[11] Classification . https://astronomy.swin.edu.au/cosmos/H/Hubble+Classification#:~:text=The%20Hubble%20classification%20of%20galaxies%2C%20also%20referred%20to,galaxies.%202%20Spiral%20galaxies.%203%20Barred%20Spiral%20Galaxies.

[12] Surface Brightness . https://en.wikipedia.org/wiki/Surface_brightness#Calculating_surface_brightness, 2020.

[13] Galaxy . https://www.astronomynotes.com/galaxy/s8.htm.

[14] Christopher J Conselice. The fundamental properties of galaxies and a new galaxy classification system. *Monthly Notices of the Royal Astronomical Society*, 373(4):1389–1408, 2006.

[15] https://ned.ipac.caltech.edu/level5/March11/Ball/Ball3.html.

[16] Mohamed Abd El Aziz, IM Selim, and Shengwu Xiong. Automatic detection of galaxy type from datasets of galaxies image based on image retrieval approach. *Scientific Reports*, 7(1):1–9, 2017.

[17] H.K.C. Yee Michael D. Gladders. a new method for galaxy cluster detection i: the algorithm. https://astrobites.org/2012/03/27/the-red-sequence-method-for-galaxy-cluster-detection/, 2020.

[18] Wikipedia. Bremsstrahlung radiation. https://en.wikipedia.org/wiki/Bremsstrahlung, 2020.

[19] Eric D Feigelson and GJ Babu. Statistical challenges in modern astronomy. *arXiv preprint astro-ph/0401404*, 2004.

[20] S George Djorgovski, Ciro Donalek, Ashish Mahabal, R Williams, Andrew J Drake, Matthew J Graham, and E Glikman. Some pattern recognition challenges in data-intensive astronomy. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 856–863. IEEE, 2006.

[21] Andrea Ferrara. The most distant galaxies: Theoretical challenges. In *AIP Conference Proceedings*, volume 1294, pages 148–157. American Institute of Physics, 2010.

[22] SG Djorgovski, R Brunner, A Mahabal, R Williams, R Granat, and P Stolorz. Challenges for cluster analysis in a virtual observatory. In *Statistical Challenges in Astronomy*, pages 127–141. Springer, 2003.

[23] https://vizier.u-strasbg.fr/viz-bin/VizieR-3?-source=IX/52/xxl_gmrt.

[24] https://www.sdss.org.