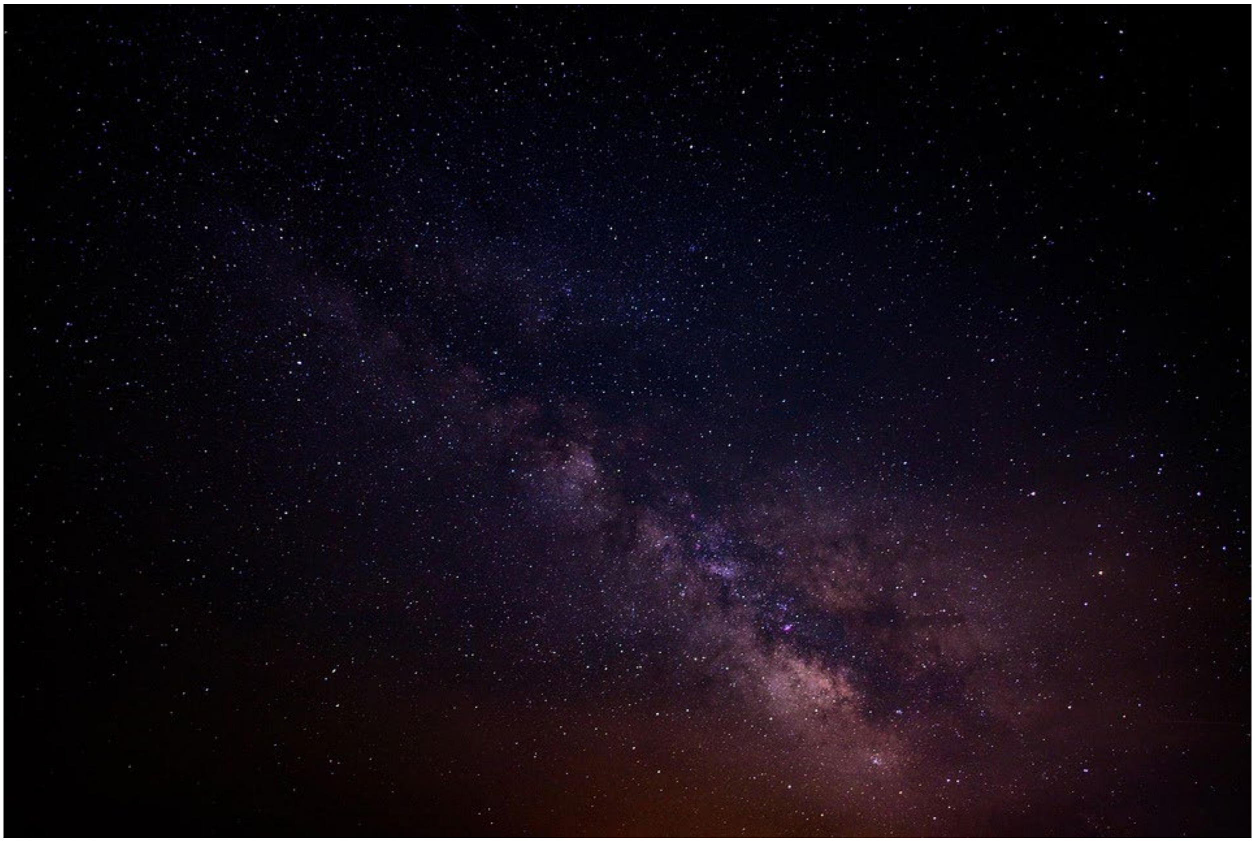
Celestial Object Detection

Multi-class classification of astronomical objects into Stars, Galaxies or Quasars,

based on spectroscopic & photometric features made available as a tabular dataset.



Note: While we have seen machine learning techniques applied to various problem contexts from the business domain as part of the case studies so far, this case study aims to demonstrate their applicability to problem statements from the natural sciences as well.

Astronomy is one such discipline with an abundance of vast datasets, and machine learning algorithms are sometimes a necessity due to the labor intensity of analyzing all this data and deriving insights and conclusions in a more manual fashion.

Problem Context

The detection of celestial objects observed through telescopes as being either a star, a galaxy or a quasar, is an important classification scheme in astronomy. Stars have been known to humanity since time immemorial, but the idea of the existence of whole galaxies of stars outside our own galaxy (The Milky Way), was first theorized by the philosopher Immanuel Kant in 1755, and conclusively observed in 1925 by the American astronomer Edwin Hubble. Quasars have been a more recent discovery made possible significantly by the emergence of radio astronomy in the 1950s.

Descriptions of these three celestial objects are provided below:

- **Star:** A star is an astronomical object consisting of a luminous plasma spheroid held together by the force of its own gravity. The nuclear fusion reactions taking place at a star's core are exoergic (there is a net release of energy) and are hence responsible for the light emitted by the star. The closest star to Earth is, of course, the Sun. The next nearest star is Proxima Centauri, which is around 4.25 light years away (a light year refers to the unit of distance travelled by light in one year, around 9.46 trillion kilometers). Several stars are visible to us in the night sky, however they are so far away they appear as mere points of light to us here on Earth. There are an estimated 10^{22} to 10^{24} stars in the observable universe, but the only ones visible to the unaided eye are those in the Milky Way, our home galaxy.
- **Galaxy:** Galaxies are gravitationally bound groupings or systems of stars that additionally contain other matter such as stellar remnants, interstellar gas, cosmic dust and even dark matter. Galaxies may contain anywhere between the order of 10^8 to 10^{14} stars, which orbit the center of mass of the galaxy.
- **Quasar:** Quasars, also called Quasi-stellar objects (abbv. QSO) are a kind of highly luminous "Active Galactic Nucleus". Quasars emit an enormous amount of energy, because they have supermassive black holes at their center. (A black hole is an astronomical object whose gravitational pull is so strong that not even light can escape from it if closer than a certain distance from it) The gravitational pull of the black holes causes gas to spiral and fall into "accretion discs" around the black hole, hence emitting energy in the form of electromagnetic radiation.

Quasars were understood to be different from other stars and galaxies, because their spectral measurements (which indicated their chemical composition) and their luminosity changes were strange and initially defied explanation based on conventional knowledge - they were observed to be far more luminous than galaxies, but also far more compact, indicating tremendous power density. However, also crucially, it was the extreme "redshift" observed in the spectral readings of Quasars that stood out and gave rise to the realization that they were separate entities from other, less luminous stars and galaxies.

Note: In astronomy, **redshift** refers to an increase in wavelength, and hence decrease in energy/frequency of any observed electromagnetic radiation, such as light. The loss of energy of the radiation due to some factor is the key reason behind the observed redshift of that radiation. Redshift is a specific example of what's called the **Doppler Effect** in Physics.

An everyday example of the Doppler Effect is the change in the wailing sound of the siren of an ambulance as it drives further away from us - when the ambulance is driving away from us, it feels as if the sound of the siren falls in pitch, in comparison to the higher pitched sound when the ambulance was initially driving towards us before passing our position.

While redshift may occur for relativistic or gravitational reasons, the most significant reason for redshift of any sufficiently-far astronomical object is that **the universe is expanding** - this causes the radiation to travel a greater distance through the expanding space and hence lose energy.

For cosmological reasons, quasars are more common in the early universe, which is the part of the observable universe that is furthest away from us here on earth. It is also known from astrophysics (and attributed to the existence of "dark energy" in the universe) that **not only is the universe expanding, but the further an astronomical object is, the faster it appears to be receding away from Earth** (similar to points on an expanding balloon), and this causes the redshift of far-away galaxies and quasars to be much higher than that of galaxies closer to Earth.

This high redshift is one of the defining traits of Quasars, as we will see from the insights in this case study.

Problem Statement

The objective of the problem is to use the tabular features available to us about every astronomical object, to **predict whether the object is a star, a galaxy or a quasar**, through the use of supervised machine learning methods.

In this notebook, we will use simple non-linear methods such as **Decision Trees** to perform this classification.

Data Description

The source for this dataset is the **Sloan Digital Sky Survey (SDSS)**, one of the most comprehensive public sources of astronomical datasets available on the web today. SDSS has been one of the most successful surveys in astronomy history, having created highly detailed three-dimensional maps of the universe and curated spectroscopic and photometric information on over three million astronomical objects in the night sky. SDSS uses a dedicated 2.5 m wide-angle optical telescope which is located at the **Apache Point Observatory** in New Mexico, USA.

The survey was named after the Alfred P. Sloan Foundation (established by Alfred P. Sloan, ex-president of General Motors), a major donor to this initiative and among others, the MIT Sloan School of Management.

The following dataset consists of 250,000 celestial object observations taken by SDSS. Each observation is described by 17 feature columns and 1 class column that identifies the real object to be one of **a star, a galaxy or a quasar.**

- **objid** = Object Identifier, the unique value that identifies the object in the image catalog used by the CAS
- **u** = Ultraviolet filter in the photometric system
- **ra** = Right Ascension angle (at J2000 epoch)
- **dec** = Declination angle (at J2000 epoch)
- **g** = Green filter in the photometric system

- **r** = Red filter in the photometric system
- **i** = Near-Infrared filter in the photometric system
- **z** = Infrared filter in the photometric system
- **run** = Run Number used to identify the specific scan
- **rerun** = Rerun Number to specify how the image was processed
- **camcol** = Camera column to identify the scanline within the run
- **field** = Field number to identify each field
- **specobjid** = Unique ID used for optical spectroscopic objects (this means that 2 different observations with the same spec_obj_ID must share the output class)
- **class** = object class (galaxy, star, or quasar object)
- **redshift** = redshift value based on the increase in wavelength
- **plate** = plate ID, identifies each plate in SDSS
- **mjd** = Modified Julian Date used to indicate when a given piece of SDSS data was taken
- **fiberid** = fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

Submission Instructions :

1. Run all the cells of your final code. Download code file as html. Submit the .html file.

2. The dataset to use is skyServer250k.csv

3. Due date : 3/31/2023. Friday after the class.

4. I have also uploaded the solution notebook file that we discussed in last class involving post and pre pruning. You will find it in unit 4 decision tree folder.