Forside

Eksamensinformation
NFYK10020E - Physics Thesis 60 ECTS, Niels Bohr Institute - Kontrakt:129788 (Marina Koukouvaou, Ioannis Mageiras)

Besvarelsen afleveres af
Marina Koukouvaou
nzt639@alumni.ku.dk

Ioannis Mageiras
dvl605@alumni.ku.dk

Eksamensadministratorer
Eksamensteam, tel 35 33 64 57
eksamen@science.ku.dk

Bedømmere
Johan Peter Uldall Fynbo
Eksaminator
jfynbo@nbi.ku.dk
+4535325983

Frank Grundahl
Censor
fgj@phys.au.dk

Besvarelsesinformationer

Titel: Applying Machine Learning on Quasar Selection
Titel, engelsk: Applying Machine Learning on Quasar Selection
Tro og love-erklæring: Ja
Indeholder besvarelsen fortroligt materiale: Nej

Applying Machine Learning on Quasar Selection

MSc Thesis in Astrophysics

Niels Bohr Institute

University of Copenhagen

Written by

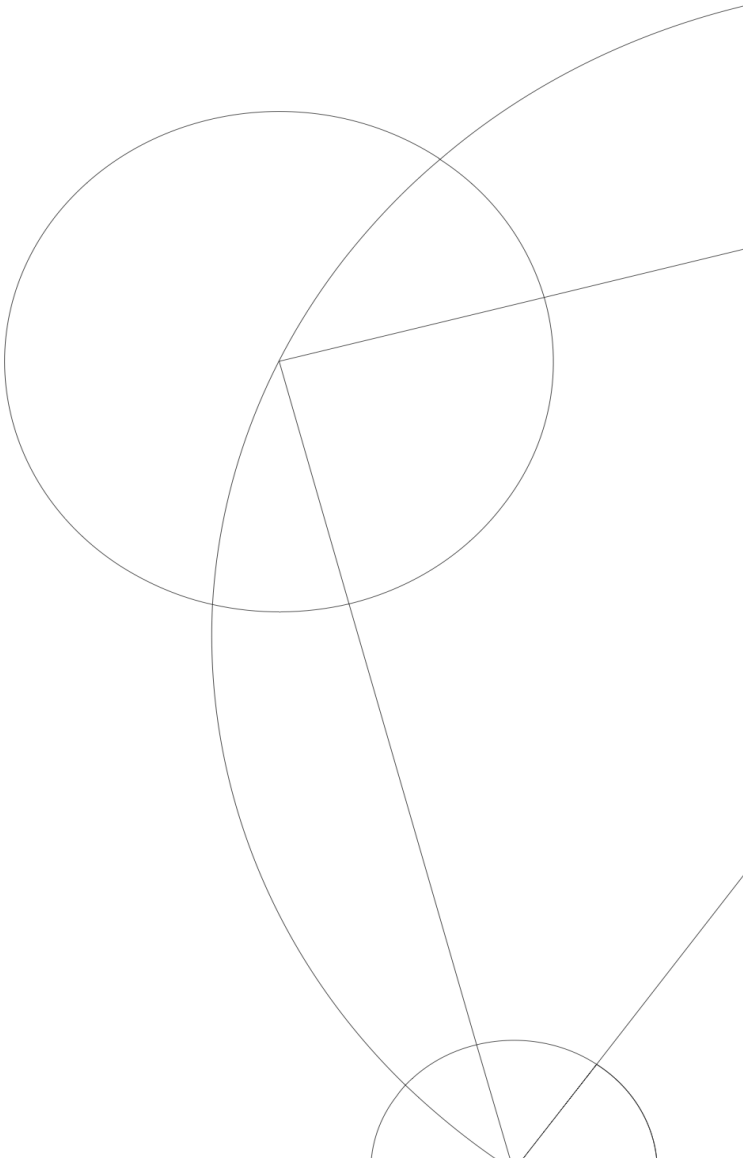
Ioannis Mageiras

Marina Koukouvaou

Supervisor

Johan P. U. Fynbo

August 2022



Contents

Abstract

How galaxies form, develop and die is a longstanding problem in astrophysics. This knowledge is necessary to uncloak how the Universe evolves and also to gain insight into the origin of our own Milky Way Galaxy. Quasi Stellar Objects (QSOs), one of the most violent astrophysical phenomena, helped shape our Universe and in fact without them we might not be here at all. The first quasars were born in a very young state of the Universe that was still filled with the neutral hydrogen generated from Big Bang and it is shown that they played a role in re-ionizing the interstellar and intergalactic medium. Quasars host the most massive black holes in the Universe and have the highest accretion rates - that is the rates at which material falls onto a central object due to gravity. The deep gravitational well of the galactic core pulls the surrounding gas which is swept up into a whirlpool of super-heated plasma, gaining incredible speeds. The material's energy of motion is turned into heat, or the material's mass is directly converted into radiation, leading to the extraordinary luminosities of quasars. These enormous amounts of energy can push the surrounding gas outwards, generating strong outflows that tear across interstellar space like a tsunami. In some cases quasars launch powerful beams of charged particles (jets) that erupt from the quasar's poles, with lengths that range between some light years up to several hundred thousand light years [Lukasz Stawarz, 2004]. These jets are bright radio sources, producing the radio signals that attracted the attention of the first quasar researchers. It is interesting, though, that not all quasars have jets and consequently extreme radio emissions. However, the name quasar - "quasi stellar radio source" - still remains, reminding us the history of how these bright sources were discovered. Moreover, by studying very high-redshift quasars we can acquire information about the young Universe, as well as information about intervening galaxies, such as Damped Lyman Alpha Systems (DLAs), which is very difficult to acquire with other methods. Quasars can therefore be the bright lighthouses we can use to study the Universe.

In order to unravel the interesting physics of quasars we first have to find them. Modern surveys contain billions of astronomical sources and these catalogues will only keep increasing in size. Due to their big size and the time consuming acquisition/extraction of the sources' spectra, an automatic quasar selection process would be very helpful on the research on quasars. Through the years, many different selection techniques were proposed. Unfortunately, each of them entails to biases. At first, the radio emission excess of quasars gave rise to the first QSO catalogues. However, pretty soon it was realized that only about 10% of the total quasar population are radio loud [A. Sandage, 1965]. The second approach was based on the UV excess of quasars compared to stars. Although this method is good in selecting quasars, leaving out stars that would contaminate the catalogue, it is unable to select all kinds of quasars. Dust reddened and high-redshift quasars are systematically overlooked by this method. A better selection technique would include more optical colors, resulting in a multicolor method. Followingly, with the development of surveys that observe in longer wavelengths, such as near and mid-infrared, the selection of these obscured quasars became less biased. However, these methods still fail to select a fair amount of quasars. Astrometric criteria have also been imposed in the effort to lift the color biases from the selection process [K. E. Heintz, J. P. U. Fynbo et al., 2020]. In this work we explore a

different way of selecting quasars. We exploit the computational power of machine learning models, "feeding" them with photometric and/or astrometric observations and letting them make predictions on the nature of unseen point sources. This allows us to define more complex separations in the multi-color space than it is possible with simple empirical color criteria. Some of the interesting quasar candidates that arise from the ML predictions are selected and spectroscopically observed at the Nordic Optical Telescope (NOT) in La Palma, Spain. The first observing run was part of IDA (Instrument Center for Danish Astrophysics) summer school, in August 2021. The second was conducted during December 2021 by our supervisor, Johan Fynbo, based on a purely photometric model's predictions. Finally, on the end of May 2022 an observing run took place at the NOT telescope after an approved proposal we submitted to IDA - young researcher's program. For the observing run we travelled to La Palma in order to obtain the data. Specifically, we obtained 11 low-resolution spectrums of 11 quasar candidates using the ALFOSC instrument. Out of them, one particular target turned out to be a relatively high-redshift quasar with apparent $Ly\alpha$ forest at $z = 3.8$. Those quasar candidates were selected based on the classifier trained on the Gaia and WISE colors as well as on the astrometric features. The observations in many of the cases verified the machine learning predictions. In other cases, they revealed the weaknesses of the models and pointed out the modifications we should apply on them. Based on this work and with the use of Gaia's observations alone, we make a catalogue of 372.000 quasar candidates, with high completeness but with a low purity of $\sim 75\%$, for galactic latitudes $b > 40^{\circ}$. Applying a more robust machine learning model that we trained, we provide a purer sub-catalogue of 42.300 quasar candidates with a very high estimated completeness and purity ($\sim 99\%$). Critical cut-offs on the S/N_{pm} are applied through the process in order to guarantee a low stellar contamination in our catalogues. After all those considerations, we re-evaluate the quasar surface density. We propose a mean surface density of ~ 37 quasars per deg^2 for $G_{mag} < 20$. This number is somewhat higher than previous surveys have shown. It is however justified, considering the known fact that biases in the traditional selection processes entail to an underestimated quasar population. The fact that we imposed many multi-color and astrometric criteria and cut-offs to prevent the stellar contamination, leads us to the conclusion that this raise is due to a higher level of quasars inclusion, generated by the machine learning's high completeness. In other words, the quasars that would have been overlooked with any other selection process are contained in our quasar candidate catalogue and at the same time, as many stars as possible have been left out.

Acknowledgements

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (https://www.cosmos.esa.int/gaia), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

Part I

Background

Chapter 1

Quasars in a nutshell

"The story so far: In the beginning the Universe was created. This has made a lot of people very angry and been widely regarded as a bad move."

- Douglas Adams, Hitchhikers guide to the galaxy

1.1 What Quasars look like

Astronomers since 1963 [Schmidt M., 1963] know that quasars are extragalactic objects and specifically that they are the very active centers of galaxies, where a super-massive black hole accretes gas from the surrounding area. They also fall into a wider category of objects called Active Galactic Nuclei or AGNs. Quasars in particular are found billions of light years away from the Earth and they shine more than a hundred or a thousand galaxies combined. In this chapter we are going to discuss some of the main questions that puzzled astronomers through the years, concerning these extremely luminous objects. Some of these questions are the following.

What is the size of a quasar or an AGN? Could it be light seconds or thousands of light years across? How massive is the central black hole and how fast does it accrete mass from the surrounding area? Why do some quasars give off radio waves while some are radio quiet? How many types of quasars are there and what is the difference among them? How long do quasars last and finally how do we select quasars out of the billions of point sources on the night sky?

The very last question addressed is the main question this thesis is dedicated to give an insight to. But before we unravel the journey towards finding better selection techniques, let us start from the basics: from narrowing down the size of a quasar. As it is known, optical observations from Earth suffer from what is called seeing. This basically describes the blurring an astronomical image will suffer from, due to atmospheric turbulence. A result of this phenomenon is that star like objects, measured by ground based telescopes, are smeared by about 0.5 arcseconds. To overcome this intrinsic turbulence from the atmosphere, space telescopes can be used. For example, Hubble Space Telescope can resolve images as low as 0.05 arcseconds. As a reminder, an arcsecond is 1/3600 of a degree and since there are 57.32 degrees in a radian, one arcsecond is equal to $2.4\text{x}10^{-7}$ rad.

For such small angles the physical size of an object is related to the distance of an observer by the following equation: $l = d \times \theta$, where θ is the angular size of the object measured in radians, d is the distance from the observer and l is the physical size of the object. Using this formula to the nearest AGN found by HST, named NGC 4395, one can have an actual upper limit of an AGN's size. Specifically, since HST can resolve things as small as 0.05 arcsec and at the same time it was not able to resolve the AGN in the middle of NGC 4395, then the size of the AGN should be smaller that 0.05" (or 2.4x10^{-7} radians). With this AGN lying 4.3 Mpc away from the Earth, we find: $l = (4.3 \cdot 10^6 pc) \times (2.4 \cdot 10^{-7} rad) \sim 1pc$. So at least for the nearest AGN found, an upper limit of 1 pc can be placed (of course for more distant quasars this upper limit would be significantly larger). But even this approximation provides an astonishing result. A parsec is a tiny distance compared to galactic scales. Even Proxima Centauri, the closest star to Earth apart from the Sun, is 1.3 pc (or 4.2 light years) away. That means that enormous amounts of energy come from an area that is even smaller than the distance to our nearest neighbor star.

As it turns out there is also another way to constrain the size of an AGN. This comes from the fact that quasars are not just active, but also variable.So the question here is how an observer can translate changes in the quasar's brightness into physical size of the region where the radiation came from. Imagine an object with diameter $d = 1$ light hour. At some point the entire object emits a brief flash of light. Photons emitted from the part of the object that is closer to Earth will arrive first. Light from the middle part of the object will arrive at some point later and finally, radiation from the most distant parts of the object will arrive last. So, although the object emitted a sudden flash of light from its entire volume at once, telescopes on Earth will observe a gradual increase in brightness that will last a full hour, counting from the first recorded incident. Inverting this idea we could say that if an observer sees a significant change in brightness of an object in a time Δt, then the size of the source can be no larger than $R \sim c \cdot \Delta t$. Returning now to the AGNs, it has been found that for optical wavelengths the brightness can vary on timescales as short as days, while for X-ray wavelengths brightness variability oscillates every few hours [Klimek et al., 2000] [Stalin et al., 2004]. So in the case of the X-ray variability one can have: $R = 3 \cdot 10^8 (m/s) \cdot 10^4 (s) = 3 \cdot 10^{12} = 10^{-4} pc$. This value is extremely small and is in fact 10.000 times smaller than the approximation using simple trigonometry. This number corresponds to approximately 2 light-hours distance and as a result it could easily fit inside our solar system, as it would reach around the trajectory of Uranus (19.8 AU). One important note is that the variability of AGNs usually depends on which wavelength they are observed. As we already discussed, the X-rays vary within a few hours time-scale while the optical radiation can vary within a few days time-scale (and thus the size would be some light-days across or 0.01pc). Does this mean that the size of an AGN depends on the wavelength we observe? In a sense, yes, as we can argue that we are observing different wavelength emission from different areas of the AGN. So, X-rays seem to come from a much smaller region than the optical emission region.

With unprecedented angular resolution (of $20 \mu as$), astrophysicists at the Event Horizon Telescope (EHT) have imaged for the first time the blazar J1924-2914 (Fig. 1.1) [Sara Issaoun et al., 2022]. This image provided details of the source structure with high analysis, allowing for the study of quasars morphology. A helically bent jet is shown emerging from a compact quasar core. Simultaneous observations across the radio frequency band from the EHT, operating at 230 GHz, the Global Millimeter VLBI Array, operating at 86 GHz, and the Very Long Baseline Array operating at 2.3 and 8.7 GHz made possible this astonishing achievement. The physical size length unit of the leftmost box in Fig. 1.1 is $0.25pc$. The accretion disk is still not resolved and that indicates that its physical size is even smaller.

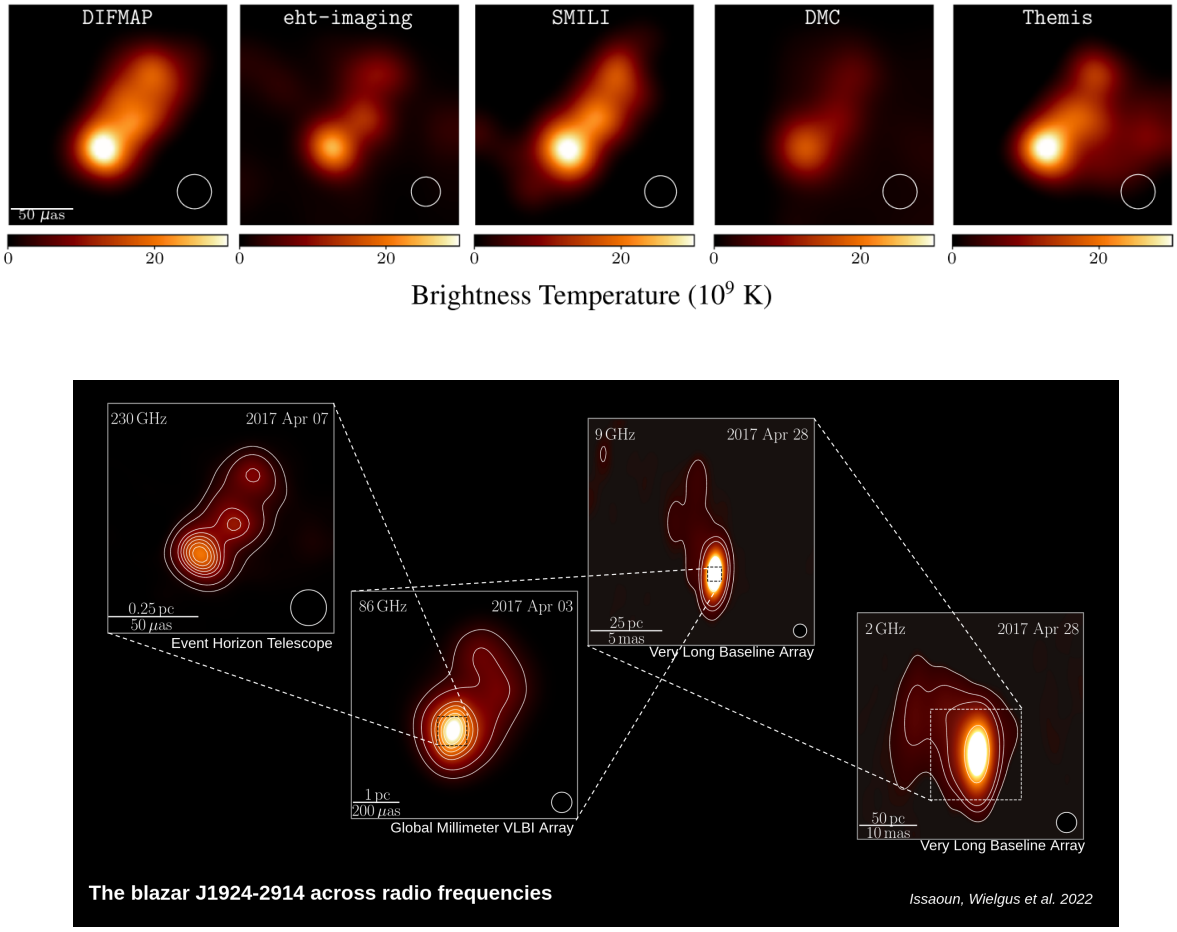


Figure 1.1: Image with an extemely high resolution of the QSO J1924-2914. The EHT array achieves an angular resolution of $\sim 20\mu as$. From right to left the box subfigures start from a physical size scale of 50pc reaching 0.25 pc. Image Credit: [Sara Issaoun et al., 2022]

A proposed morphology and the geometrical features (regions) related to different emissions from a quasar, are described through the AGN unified model that follows below.

1.1.1 The Unified Model of Active Galactic Nuclei

An AGN consists of several components, illustrated in Fig.1.2. In the central region resides the active SMBH, powered and surrounded by the accreting material that forms a rotating, geometrically thin accretion disk at 0.01-0.1 pc from the center. At a distance of ∼0.01-1 pc from the SMBH is the Broad Line Region, composed of gas clouds with high densities and velocities. The assumption of an optically thick, dusty, clumpy torus surrounding the inner parts at ∼1-10 pc was included to explain the obscuration of the inner disk and the absence of broad lines, if the AGN is seen edge-on. Further outside above the opening of the torus lies the Narrow Line Region, a low density and low velocity region of cold gas that produces the narrow lines found in the AGN spectra. All those constituents lead to AGN emission covering the whole electromagnetic spectrum, meaning that AGNs have been detected and studied at a wide range of wavelengths. The emission characteristics, along with the orientation effect of AGNs became the reasons why active galactic nuclei has been divided in so many - sometimes vaguely discriminated - types and subclasses.

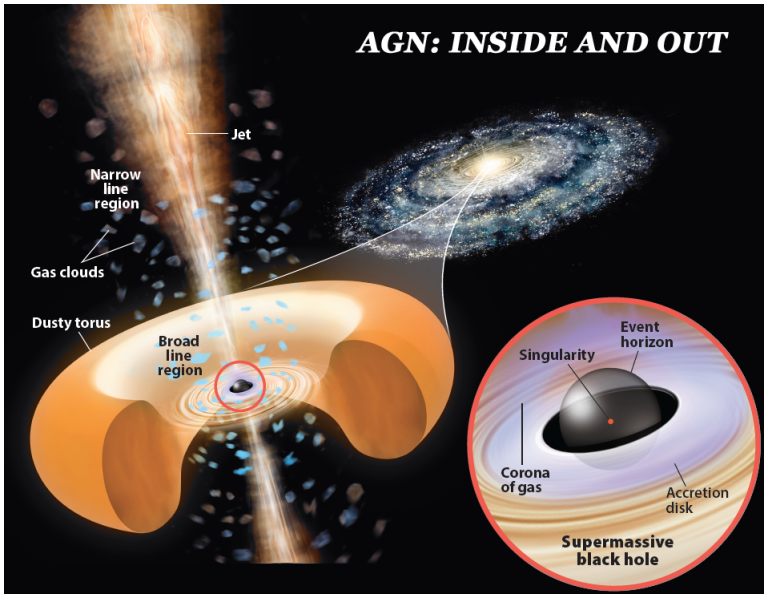


Figure 1.2: Schematic representation of the Unified Model of Active Galactic Nuclei. In type I AGNs or type I Seyfert galaxies the nucleus of the galaxy is directly visible and in the optical spectrum one can observe lines both from the Broad and Narrow Line Region (BLR and NLR). In Type II AGNs the observer sees the galaxy edge on and as a result the broad line features are gone and only narrow lines are seen in the spectrum. Blazars is the case where the relativistic jet erupting from the very center of the accretion disc happens to be pointing directly at the observer.

The AGN Unified Model was first proposed by [Antonucci, 1993], in an attempt to explain the observed properties of the AGNs by a set of a limited amount of physical parameters. According to the unified model, the main classifying parameters are the AGN's luminosity and the inclination to the line of sight. There are various AGN types such as LINERS (Low ionization nuclear emission regions), the Seyferts type I and II, radio-loud and quiet AGNs, broad and narrow emission line quasars, BAL (Broad Absorption Line) quasars with or without low ionization elements (LoBALs, HiBALs respectively), extremely luminous blazars and BL Lac objects and many more. In a simple way, we can cluster all these objects into two large groups: i) The Seyfert galaxies and ii) the quasars or blazars. The first ones are nearby and the less luminous objects among the active nuclei ($L \sim 10^{41} - 10^{44}$ erg s^{-1}). They hold supermassive black holes surrounded by accretion disks in their centers, emmiting enormous amounts of radiation, which is however comparable to the total radiation of the host's galaxy's stars. This fact, along with the Seyfert galaxies' proximity allows the observation of the host galaxy as well. Secondly, we have the quasars or blazars, with no apparent difference from one another, apart from their angle with respect to the observer's line of sight. In the case of a blazar, the relativistic jet is pointing directly towards the observer. Objects of these groups are far more luminous and distant than the Seyfert galaxies ($L > 10^{45} ergs^{-1}$). Unlike the latter, the emission from their active nuclei is at least 2 orders of magnitude more luminous than the host galaxy's constituent stellar population that is thus not detectable. AGNs can also be divided into subgroups according to the nature of their spectral lines, and both Seyfert galaxies or quasars can be found in type I and type II subgroups. According to the standard unified model of Active Galactic Nuclei these different phenomenological subclasses arise due to orientation effects with respect to the line of sight to the observer [Antonucci, 1993], [Urry & Padovani, 1995]. The angle is measured with respect to the direction perpendicular to the central accretion disk (the z- axis of the system). A smaller inclination (face-on view) results in clearer observation of the central region, while inclinations closer to 90° (edge-on view) face the intervening torus that obscures the central source's light.

Type I refers to AGNs in which the nucleus, the Broad Line Region (BLR) and Narrow Line Region (NLR) are directly visible by an observer (Fig.1.2). As a result, their optical and UV spectra appear with broad permitted or semi-forbidden lines, originating from the fast rotating central BLR (FWHM 1000-20.000 km s^{-1}) and possibly also narrow emission lines.

In type II AGNs one can see only narrow emission lines in the UV, optical and near IR spectrum. The lines originate from the NLR region rotating with lower velocities (FWHM 300-1000 km s^{-1}). This, according to the unified model, can be explained if an observer sees the AGN edge on, where there is the dusty/molecular torus around the supermassive black hole that can hide the sub-parsec size BLR but not the kilo-parsec size NLR. The dust there consists of amorphous silicates and carbonaceous grains of typical sizes $\sim \mu$m, that absorb the corresponding wavelengths of radiation and re-emit it in the infrared. This AGN unified model was developed when a spectropolarimetric observation was made on a Seyfert 2 galaxy (NGC 1068) revealing the hidden broad line region [Antonucci & Miller, 2985]. The strength of the spectropolarimetry relies on the fact that the polarization state of scattered light carries the imprint of the scattering geometry [Smith et al., 2005]. The hidden region is revealed because the light emitted from the BLR scatters towards the line of sight of the observer, most probably, in an electron scattering region far away from the accretion disc and BLR [Yoshiaki & Anabuki, 1999]. Although this model is successful in explaining the existence of hidden broad lines in the spectrum of many narrow line AGNs (type II), in only half of the type II cases are broad lines revealed via spectropolarimetry [Smith, Paul S, 2003]. Those are the so called real type II AGNs, that show no broad lines in their spectrum even when unobscured. A proposed explanation is that more parameters than just the orientation of the AGN with respect to the observer should be taken into account. Those physical parameters include the existence of different geometric features of the torus, the modification of the thin disk approximation, as well as the implementation of different accretion rates and black hole characteristics (mass, spin, charge).

Mathematical Formulation of the AGN accretion

While accreting mass, the AGN has a strong impact on the interstellar medium of its host galaxy. The redeposition of momentum and energy through the interaction with the ISM strongly affects the star formation, a mechanism that is known as the AGN feedback. Winds and jets of outflowing material emerge speeding outwards from the poles of the black hole. Jets extend to hundreds of thousands light years across, beyond the galaxy's plane, while winds travel shorter distances from the black hole. The emission from these highly collimated jets originates from relativistic electrons accelerated in a spiral path following the magnetic field. This is the synchrotron radiation, from which the radio signals we detect from the radio-loud AGNs are produced.

At a first level, the process of the accretion around a SMBH, can be approached by the standard spherical accretion model. For a free falling mass m, travelling from infinity to a distance r from the central object of mass M, its potential energy is converted to kinetic:

$$\frac{1}{2}mv^2 = \frac{GMm}{r} \tag{1.1}$$

For an accretion rate $\dot{m}$, the differentiation of eq. 1.1 provides the rate that gravitational energy is released in the form of heat when a mass falls onto the surface of a central object

$$L_{acc}(r) = \frac{GM\dot{m}}{r} \tag{1.2}$$

This is the accretion luminosity. A temperature profile of the accretion disk is easily inferred from eq. 1.2, if we assume that accretion luminosity is produced as black body radiation. According to Stefan-Bolztmann law, the power output of a black body over a surface area of radius r, in terms of the temperature is

$$L = 4\pi r^2 \sigma T^4 \tag{1.3}$$

which together with eq. 1.2 leads to a temperature profile T(r) of a power law

$$T \sim r^{-3/4} \tag{1.4}$$

so that the temperature is not uniform across the accretion disk. Eddington introduced the L_{Edd}, an upper limit to the accretion luminosity that can be reached. His arguments included the assumption of spherical symmetry, the act of the inward force of gravity and the outward radiative pressure. In equilibrium, the gravitational attraction would balance out the repulsive radiation field. The two forces per unit volume f_{rad} and f_{grav} are equal and steady state spherical accretion is achieved. High luminosities therefore indicate high masses such that gravity can overcome the radiation pressure. For a luminosity higher that the Eddington, the accreting matter would be pushed away by the radiation field and the system would be unstable, blowing away the accretion disk.

$$\frac{L\kappa\rho}{4\pi r^2 c} = \frac{GM\rho}{r^2} \Rightarrow L_{Edd} = \frac{4\pi c G M}{\kappa} = 1.5 \times 10^{38} \frac{M_{BH}}{M_\odot} ergs^{-1} \tag{1.5}$$

κ here is the opacity. Typical values for the accretion luminosity of AGNs range from 10^{44}-10^{47} erg/s, allowing us to estimate the mass ranges for the SMBH that reside in their centers (that is 10^6-$10^9 M_\odot$). As discussed above, another interesting observable of the ANG luminosities is that they are variable on timescales that range from years to months and even days or hours. Assuming a daily variability, photons travelling with the speed of light c would have to cross a distance of 1 day $\times$ c , which is $\sim$100 AUs. That is of a solar system order of magnitude, which implies that the region where the accretion takes place is spatially fairly small. Those two arguments lead to the conclusion that the AGNs have a very high mass density.

Luminosity is by definition the total amount of energy emitted in a given time, $L_{bol} = \frac{dE}{dt}$. On the other hand, the accretion process converts the rest mass into energy, but we assume that only a fraction can be radiated away. This fraction is expressed by the introduction of the accretion efficiency η in the rest-mass energy relation $E = \eta mc^2$, which typically takes the value $\eta \sim 10 - 40\%$, making the AGN accretion the most effective known process of mass-energy conversion. For comparison, nuclear fusion releases 0.07 % of the mass-energy and constitutes one of the reasons why the luminosity observed from active galactic nuclei cannot be associated with stellar activity. The combination of the above equations leads to

$$L_{bol} = \frac{d(\eta mc^2)}{dt} = \eta \dot{m} c^2 \tag{1.6}$$

If this bolometric luminosity originates from the complete conversion of the infalling matter's kinetic energy to radiative output, then it can be replaced by L_{acc} and the accretion efficiency can be written as

$$\eta = \frac{GM}{rc^2} \tag{1.7}$$

In the case of a BH where there is no hard surface for the mass to fall onto, the Schwarzschild radius $R_{sch} = \frac{2GM}{c^2}$ is a convenient way to express the accretion luminosity and efficiency

$$L_{acc} = 2\eta \frac{GM\dot{m}}{R_{sch}} \tag{1.8}$$

$$\eta = \frac{R_{sch}}{2r} \tag{1.9}$$

Using eq. 1.7 and the Eddington luminosity, we can also define the Eddington accretion rate,

$$\dot{M}_{Edd} = \frac{4\pi GM}{\eta\kappa c} \tag{1.10}$$

1.2 Selection Techniques

1.2.1 Historical overview: A star-like object with large redshift

In 1963 Maarten Schmidt published a paper in Nature with the title '3C273: A Star-Like object with large redshift' [Schmidt M., 1963]. The name 3C273 comes from the fact that this object was the 273rd object in the 3rd Cambridge catalogue of radio sources. It is worth mentioning that these first radio telescopes had quite poor angular resolution and as a result many radio sources were unable to be associated with specific optical ones. Astronomers could point down some big blobs on the night sky where radio light was coming from, but not always their optical counterparts. In order to overcome this problem a group of Australian astronomers [Hazard et al. ,1963] used a lunar occultation[1] to pin down the location of 3C273 by the precise timings of the appearance and disappearance of the radio signal when the moon was passing in front of it (Fig.1.5). Maarten Schmidt was then able to identify two optical counterparts of 3C273, finding a fairly faint blue star-like object (Component B) and a faint "wisp" or jet (Component A, Fig.1.4). To investigate the object even further Maarten Schmidt took the spectrum of it using the 200-inch (5m) Mount Palomar telescope (Fig.1.3). The spectrum of 3C273 appeared to be nothing like a star, a galaxy or anything ever seen before. It showed some strong broad emission lines that could not be matched with any known emission lines. Schmidt identified that some of them were the well known Balmer lines but just redshifted. The amount of redshift necessary for this displacement had to be z=0.158 entailing that this object should be receding away from us with a velocity of around 50.000 km/s. In the 1960s, most astronomical objects measured had way lower redshifts. This high velocity meant that the object should be very distant. Schmidt states in his paper that if one considers this redshift as cosmological in nature (due to the expanding universe), then it would indicate a distance of 500 Mpc or around 1.5 Gly.

[1] Occultation occurs when one object is hidden by another object that passes between it and an observer. The Parkes radio telescope in Australia in 1962 registered the exact instant that the radio signal of 3C273 vanished behind the moon (Fig.1.5). That timing allowed astronomers to find a point on the sky - a star like object - as the source of the radio waves

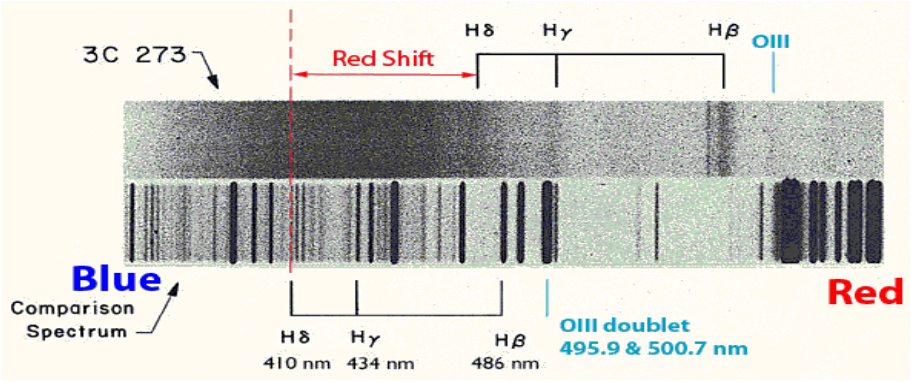


Figure 1.3: The spectrum used by Maarten Schmidt to determine the redshift of 3C 273 at z=0.16. The spectrum was acquired using the Palomar 5m telescope.

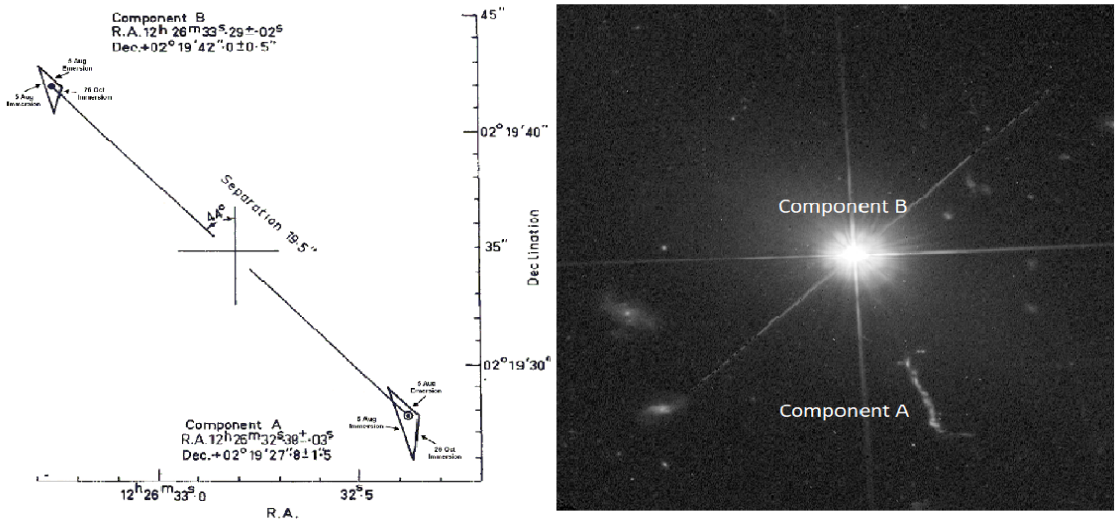


Figure 1.4: Left: The occultation configuration and relative positions of the 3C 273 components, from [Hazard et al. ,1963]. Right: Image from Hubble's Wide Field and Planetary Camera 2 (WFPC2) of the 3C273 quasar which resides in a giant elliptical galaxy in the constellation of Virgo. One can observe the two components (A and B) with component B being the quasar and component A a jet originating from it.

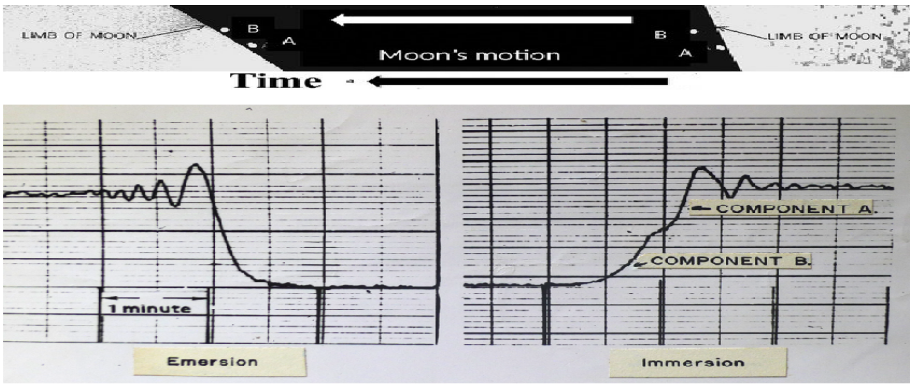


Figure 1.5: The August 5 1962 disappearance and reappearance of the 2 optical counterparts of the 3C273 radio source, measured at 136 MHz and 410 MHz. Note that time increases from right to left, and that the Moon is also moving from right to left. The top panel shows the positions of source components A and B relative to the limb of the Moon.

1.2.2 Photometry

Contemplating the 60 year research on active galaxies and quasars, it is explicit that their photometry has strongly influenced the methods of finding them. One of the firstly proposed methods for selecting quasars is the ultraviolet excess (UVX). By definition, emission from a quasar does not resemble that of a star or an ensemble of stars. It has a different broadband spectral shape (the Spectral Energy Distribution - SED) than that of stellar types which have the well-defined blackbody SEDs. Quasars are separating themselves because of their relatively blue colors ($U - B < 0$). Specifically, the Bright Quasar Survey (BQS) was able to find 114 quasar candidates based on the color criterion $U - B < -0.44$ [Schmidt and Green, 1983]. The first problem with this method arises for redshifts $z > 2.2$. In this redshift range the Ly-alpha line (121.6 nm) starts falling into the Blue band filter while the Ly break and Ly forest enters in the UV band, lifting the characteristic ultraviolet excess. Other problems include dust-reddened quasars where it is known that dust affects the blue more than the red light. There are various ways a quasar can be reddened e.g, due to dust in its host galaxy, due to the intergalactic medium or due to an extra intervening system (e.g, Damped Lyman Absorber - DLA). Many major spectroscopic and photometric surveys, such as SDSS, contain systematic biases towards red quasars that end up being overlooked [Strauss et al., 2002, Krogager, Fynbo et al., 2016]. To overcome these biases, infrared selection techniques were proposed. Since quasar spectra have a power law nature, while stars follow a blackbody radiation, they do not only have higher intensities in lower wavelengths (UVX) but also in longer wavelengths. Based on the brightness excess in the K band (2.17 μm), Warren et al. introduced a new selection method for quasars [Warren et al., 2000]. This method has a big advantage over the UV excess method since it works well for dust-obscured quasars too. Warren et al. also suggested as a a tentative selection boundary the relation:

$$J - K > 0.36(V - J) + 0.18 \tag{1.11}$$

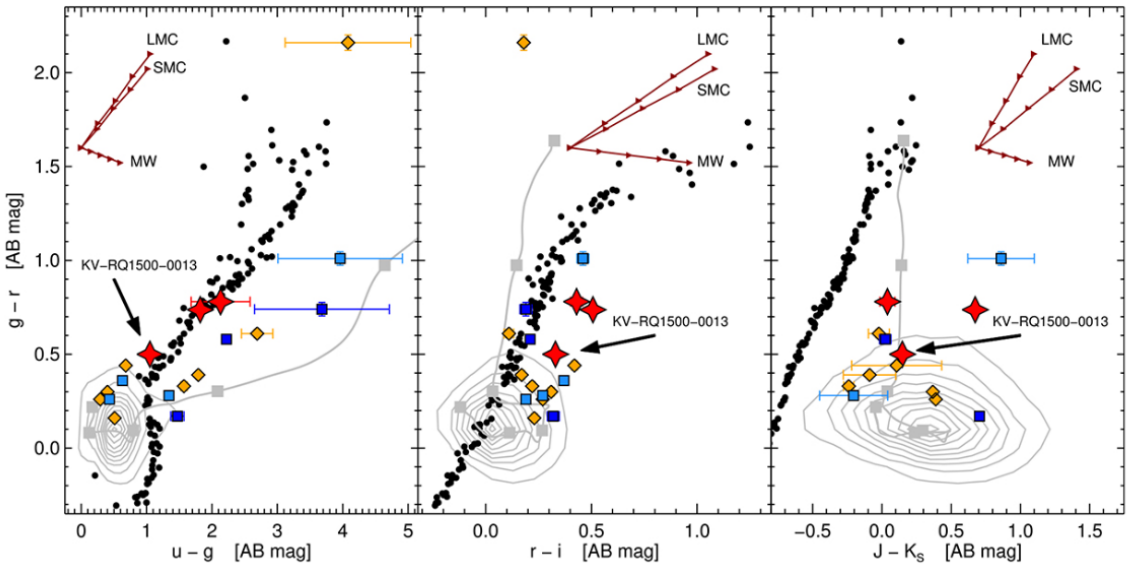


Figure 1.6: Color-color plots in optical/NIR wavelengths. Black dots show the colors of standard, main sequence stars. The grey contours represent the bulk SDSS DR12 quasar population at $1.5 < z < 4.0$. Also, the grey line denotes the redshift track of a composite quasar template in steps $\Delta z = 0.5$ (starting from 1.5 to 4). Red stars represent quasars that were mis-classified as stars in SDSS but identified in the HAQ and KV-RQ surveys. The blue squares show the sources photometrically selected to be quasars in the SDSS (light) and BOSS (dark) surveys. [Heintz et al., 2018]

The two micron all sky survey (2MASS) scanned the whole sky in three NIR bands (J, H, K) to detect and characterize point sources brighter than approximately 1 mJy. This means that the 2MASS was only able to detect the brightest of the quasars. The successor of this survey, the United Kingdom Infrared Telescope (UKIRT) Deep Sky Survey (UKIDSS) has a three times fainter magnitude limit in the K-band. One of the UKIDSS objectives was to detect the highest redshift quasars and it provided the ground for creating a large NIR selected quasar sample. It is also evident, in Fig.1.6, that dust reddened quasars fall into the stellar cluster in optical colors while there is a stronger separation in J-K color. However, it can be seen from the grey line track that as the redshift increases even the J-K colors will resemble that of stars. Although it cannot be seen in these Figures, this overlap in high redshifts happens mostly between early or late M stars (red or brown stars respectively).

Another infrared selection technique is based on the mid-infrared WISE survey. WISE stands for Wide-field Infrared Survey Explorer, a satellite launched by NASA in 2009. The WISE survey includes four photometric bands, the W1, W2, W3, W4. AGNs have been selected based on a MIR colour criterion $W1 - W2 > 0.8$ (Vega), that can identify 78 % of the AGN candidates with 95% reliability [Stern et al., 2012]. MIR selection is less biased and more effective in selecting dust reddened and high redshift quasars. Although MIR is very good in selecting all kinds of quasars resulting in a very good completeness, it suffers from a relatively low purity. In particular, it is shown that many X-ray sources have similar MIR colors with star-forming galaxies which contaminate the quasar catalogues [Lacy et al., 2004].

1.2.3 Astrometry

One of the methods selecting quasars without using light is by using their kinematics. It is introduced by [G. Kron, 1981], with the aim to lift all the former biases induced by color dependent selection methods. Since quasars are distant, extragalactic objects, their parallax and proper motion are negligible and as a result the sources appear to be stationary. On the contrary, galactic stars are characterised by higher values of astrometric measurements. A simple signal to noise cut-off for the proper motion was proposed by [K. E. Heintz, J. P. U. Fynbo et al., 2020], where the authors suggested that a complete quasar candidate catalogue can be made by selecting all the sources with proper motions consistent with zero within 2σ uncertainty. That translates to a signal to noise $S/N_{pm} < 2$. Based on that criterion the authors state that the selection efficiency at high galactic latitudes - far from the galactic plane - is 60% (purity).

Hipparcos (High Precision Parallax Collecting Satellite) was the first space mission devoted to astrometic measurements with high precisions, launched by ESA in 1989 and in operation until 1993. The accuracy of its measurements was 2 $mas\ yr^{-1}$ Its successor is the ESA's Gaia Satellite, launched in 2013. It performs with the even lower parallax error limit of 0.5 mas and a proper motion error limit of 0.3 mas at $G = 20\ mag$ [Gaia DR3: The extragalactic content]. Still, this accuracy is not enough to reliably identify and classify quasars through a simple cut on parallaxes and proper motions. It can be calculated that a quasar at redshift 1 (7.8 Gly), moving with the speed of light perpendicular to the line of sight, will have a proper motion less than 0.1 mas (Fig. 1.7).

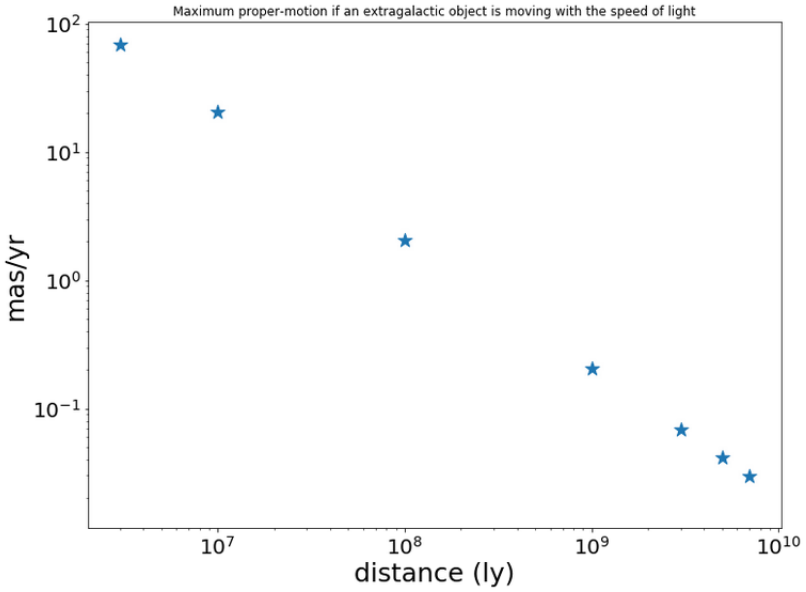


Figure 1.7: Plot of of maximum proper motion vs. distance. Maximum proper motion is calculated for an extragalactic object if it is moving with the speed of light perpendicular to the line of sight.

That does not mean, though, that astrometric observations are not useful and we indeed use them in our machine learning models, as we show in following chapters. What we basically aim to do is to explore the quasar selection process using machine learning models, trained on either purely astrometric measurements, either on a combination of astrometry and photometry.

Chapter 2

Machine Learning

2.1 Historical overview

One major difference between humans and computers is that humans have the ability to improve their performance on solving a specific problem. Humans acquire knowledge from their mistakes and use that knowledge to tackle problems in a different way or understand something about their environment. We can imagine a silly case of an infant trying to learn about gravity just by looking at objects falling down. Automatically the young one realizes that if he lets something from his hands it will most probably fall down, since this is a repetitive result 100%. After all, no one ever saw an infant looking up after letting something from his hands.

On the other hand, traditional computer programs do not learn from outcomes and hence are unable to improve their performance. This is where the field of machine learning plays an important role by creating algorithms that are able to learn in the same sense that the young human learns about gravity; by gathering more data and experience. In 1943, neurophysiologist Warren McCulloch and mathematician Walter Pitts wrote a paper on how brain neurons might work [Warren & Pitts, 1943] and it was mankind's first mathematical model of a biological neuron. The paper described every neuron in the brain as a simple digital processor and compared the whole brain to a computing machine. Almost 10 years later, in 1952, the first self-learning program was made by Arthur Samuel in IBM [Arthur Samuel, 1956]. It was programmed to play the game checkers while it had the potential to become better at it with the number of games played. In order to learn how to play, the program was using the first alpha-beta pruning algorithm in the IBM 704 computer as well as the idea of the minimax strategy. Similar to what is happening in Reinforcement Learning algorithms, the program implemented a loss function that could calculate the probability of winning the game based on the board's current state. In order to make predictions the program was using various features such as the number of pieces on each side, the proximity of the pieces and the number of kings. In 1957, Frank Rosenblatt, an American neurobiologist, designed the first artificial neural network inspired by the biological principles of a brain neuron. Rosenblatt's perceptron, as this work was named, is a single layer neural network and can be seen as a set of inputs, that are weighted and to which we apply an activation function. The details of how these machines work is not relevant to this work and so they are not going to be discussed in depth. Some technical information can be found in the literature [Rosenblatt, 1958, Andinda et al.,].

2.2 Tree based Algorithms

Machine learning aims to learn from existing data during the training part. There are different kinds of ML models that can be classified into three groups: the supervised, unsupervised and the reinforcement learning Fig. 2.1. Supervised learning algorithms investigate relationships among variables by using labeled datasets to train algorithms. After this step, the algorithm can classify new, unseen data and predict outcomes. Classification and regression are the two main subgroups in supervised learning. Classification aims to predict categorical outcomes (e.g., assign the class "star", or "quasar" etc.), while regression aims in predicting continuous outcomes (e.g., the redshift of an object). On the other hand, unsupervised learning is applied in order to uncover hidden patterns inside data that are not labeled. Gaussian mixture models, t-distributed stochastic neighbor embedding (t-SNE) and K-means clustering are three of the most popular approaches for grouping unlabeled data into clusters. Finally, reinforcement learning is a machine learning type that is based on the idea of sequentially rewarding desired behaviours and/or punishing undesired ones. As it was discussed in the historical overview section, algorithms made for playing games is the most common usage of reinforcement learning. It is proved capable of achieving superhuman performance in numerous games, like chess, go, checkers and many more. This type of learning as it is obvious does not need labeled data while it can be seen as a way of redirecting into correct outcomes unsupervised learning techniques.

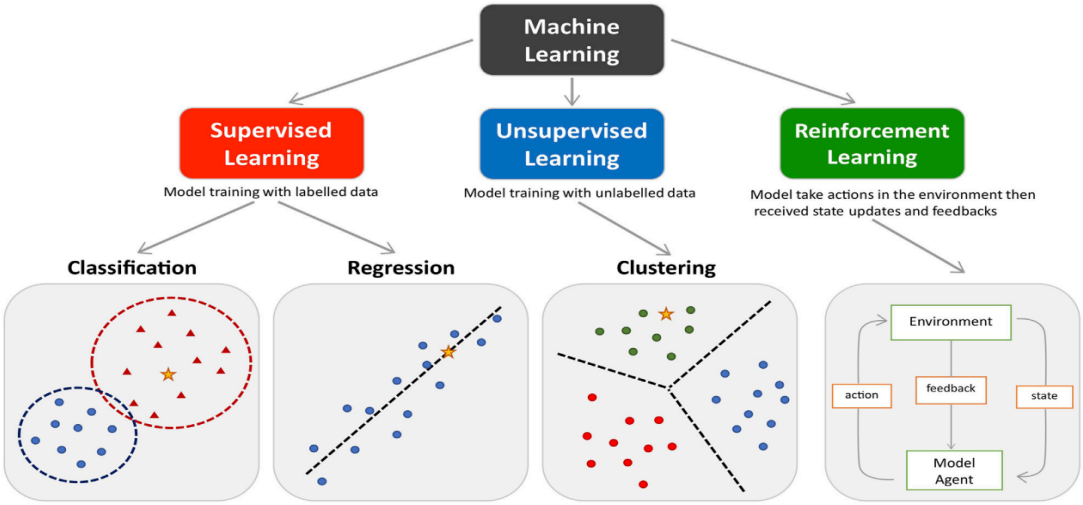


Figure 2.1: Different Machine Learning types. Under the supervised learning approach lie the classification and regression techniques. Clustering is part of the unsupervised approach where the true labels of the classes are unknown. Coloured dots and triangles represent the training data, while yellow stars represent the new unseen data which can be predicted by the new model. Credit: doi:10.3389/fphar.2021.720694

In general, the process of any kind of machine learning type consists of the following stages:

1. the gathering of the data
2. data processing
3. the training of the machine learning model
4. the testing on new data.

In the following sections the machine learning algorithms and their evaluation metrics that are used in the

Methodology part of this work (Part II) will be discussed in some depth. More machine learing systems are discussed in the Appendix E

2.2.1 Decision Tree

The Decision Tree Classifier is a simple algorithm that also belongs to the family of the supervised machine learning models. It is one of the most popular models that is used in various fields for both classification and regression problems. Decision trees classify the objects by sorting them down the tree through a series of decisions, from the root to a terminal node that corresponds to a specific class label. The algorithm operates structured like a flowchart, built on carefully selected questions about the attributes of the classes involved. These questions form a hierarchically built decision tree, consisting of nodes that expand and eventually lead to terminal ones (Fig. 2.2). Initially, the entire population of the training set is represented by the root node, which then results to two or more branches. This process is referred to as splitting, while a subsection of the entire decision tree is called a branch. The nodes who are divided into sub-nodes are called decision nodes. On each node attributes of the population are put into test. The different outcomes descending from a decision node, namely the sub-nodes of the parent one, will become themselves the new root nodes for the recursive process of splitting. Once a node is no further split, it is called a leaf or terminal node, in the sense that it provides the final decision about the class of an object.

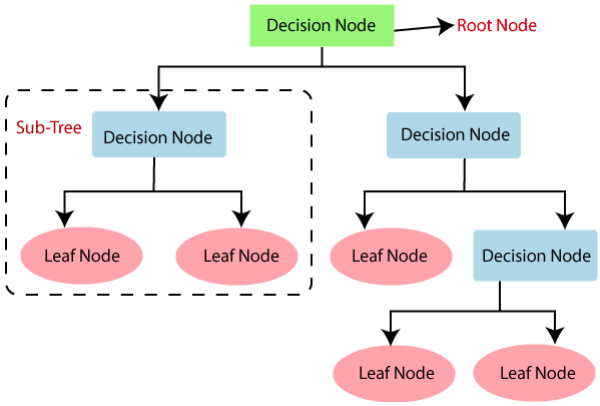


Figure 2.2: Visualization of a decision tree, from the root node, splitting down to decision nodes, to terminal/leaf nodes.

It is evident that the attributes whose features are used as local splitting criteria on each node play a crucial role for the efficiency of the decision tree classifier. This is a basic challenge that the decision tree faces, and concerns the manipulation of the attributes and the splitting strategics. The algorithm also has to find the optimal test conditions and most homogeneous subdivisions. At the same time, it has to decide when the termination of the tree growth is necessary. An objective metric to quantify the efficiency of the procedure is used by the classifier on each step, evaluating the distribution right before and right after the splitting. Some of those metrics are the entropy, information gain, gain ratio, gini index.

2.2.2 Random Forest

Random Forests are great statistical machine learning models. The random forest algorithm was proposed by L.Breiman [Breiman, 2001] and the idea behind it is the combination of many decision trees. The model aggregates its predictions by averaging or by voting and it has shown excellent performance also in cases where

the number of variables is much larger than the number of observations. Another advantage of random forests is the fact that they can be applied to a wide range of prediction problems and also that they have relatively small number of parameters. Before diving into the working mechanism of random forests it is useful to clear out what the ensemble technique is. Ensemble simply means combining multiple models. Ensemble training uses two types of methods. Firstly, the bagging method in which the training dataset is being split into subsets and the final output is based on the majority voting of all those trained models (Fig. 2.3). This randomization of the splitting minimizes the variance of a single decision tree estimator. The second method, namely the boosting method, combines many weak learners to create a strong model. An example of the boosting method is the XGBoost classifier which will be discussed in depth in the following section.

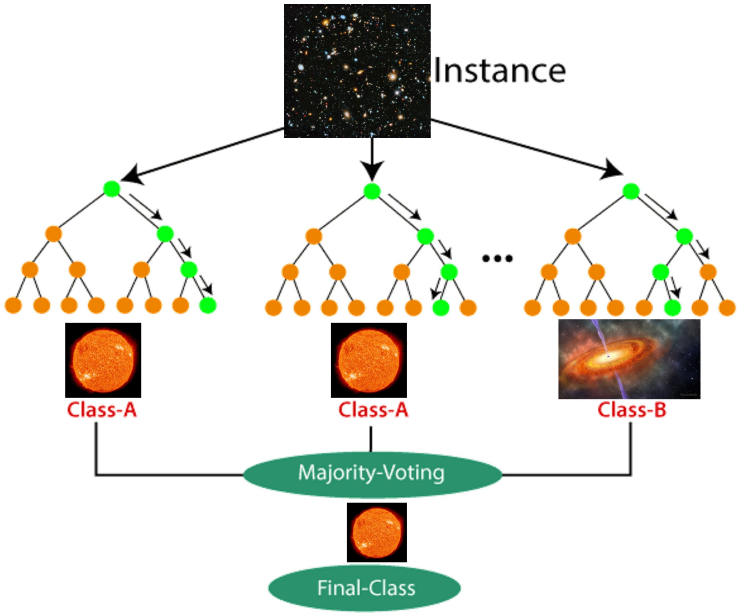


Figure 2.3: An example of how a Random Forest algorithm make a prediction based on the voting method (bagging)

2.2.3 XGBoost

XGBoost is a scalable and highly accurate implementation of gradient boosting [Tianqi, Guestrin, 2016]. The name stands for eXtreme Gradient Boosting and it is also a tree based, ensemble learning algorithm. In general, prediction problems that involve unstructured data (images or text), are best tackled with artificial neural networks. However, for medium structured tabular data, decision tree algorithms and especially ensemble methods like XGBoost, are considered the best ones. The algorithm owes its success on the weighting technique it uses. These weights are assigned to all the independent variables which are then fed into the individual decision trees which are the ones making the predictions. Followingly, the weights of variables that led to wrong predictions are boosted and the new weighted variables are then fed to a second decision tree. All these weak individual learners are then combined to provide a strong, more precise and more accurate model. As it was discussed above, a decision tree divides the parameter space, starting with the maximal separation possible based on the training data. An important question here would be how to decide where the parameter space should be divided or in other words, where to make the split. There are several ways in which this can be done, and it is also dependent whether the problem is a classification or a regression one. But in general, a rule of thumb would be to make the split where the information gain maximises. This term can be understood through

entropy,

$$H(x) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i) \tag{2.1}$$

In information theory, the expected information gain is the reduction in information entropy H, such that $Information\ Gain = 1 - H$. Entropy is correlated to the average level of uncertainty that follows a possible outcome. In classification problems, the average binary cross entropy is often used. It is also known as "log-loss" and describes how close the prediction probability is to the respective actual value.

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} [y_n log \hat{y_n} + (1 - y_n) log(1 - \hat{y_n})] \tag{2.2}$$

where y_n is the truth label, N the number of observations and $\hat{y_n}$ is the predicted estimate from the model. In an ideal case, log-loss is as close to zero as possible. This constitutes the most common metric used in tree based algorithms. Alternatives include gini coefficients, variance reduction and chiSquare.

Since XGBoost is the model that is exploited throughout this work, it is important to also explain some of the hyperparameters of the model. In machine learning, hyperparameter is a parameter whose value is used to control the learning process. In order to acquire the best predictive power of a model based on a specific training set, a step called 'tuning of the hyperparameters' has to take place. XGBoost offers a wide range of hyperparameters. Five of the most important ones are: 1.$n_{estimators}$, 2. the learning rate, 3. the maximum tree depth, 4. evaluation metric and 5. objective. The number of estimators defines how many different decision trees the XGBoost will utilize. Basically, for $n_{estimators} = 1$ we have a single Decision Tree classifier. Followingly, the learning rate simply describes how fast the model learns. The advantage of a slower learning rate is that the model becomes more robust and efficient. Thirdly, the max-depth parameter shows how deep each estimator is permitted to make a tree. Typically, increasing tree depth can lead to overfitting if not other mitigating steps will be taken into account to prevent it. The Evaluation metric parameter is the metric that will be used to validate how good the training is. There are various choices such as logloss, mlogloss for multiclass classification, auc (area under the curve) and many more. With the objective parameter one can specify if the training is going to be binary or multiclass classification, or regression.

2.3 Evaluation Metrics of Classification

In this section we will describe the statistical metrics used to evaluate the performance of a machine learning algorithm. In order to make the definitions stand out clearly, let us first describe our statistical ensemble of the input data. Let us consider the 2 class problem; each data point is either a positive, or a negative one. The trained model will make predictions and will assign a class to every data point. All the possible outcomes are four: it can be a TP (true positive), a TN (true negative), a FP (false positive), or a FN (false negative). Only in the first two cases, has the program identified correctly the class of the given object (Fig 2.4). Through these labels we can define the basic statistical metrics, namely the accuracy, recall, precision, and f1-score, that all take values between 0 and 1.

The most common and well-known metric is the accuracy, defined as the ratio between all the successful

predictions of the model, to the total number of the entries:

$$Accuracy = \frac{TP + TF}{TP + FP + TN + FN} \tag{2.3}$$

Table 2.1: Confusion matrix of a binary classification problem

		Predicted Class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

In the case of a multiclass classification problem the confusion matrix is naturally extended as shown in Table 2.2. The closest to unity the diagonal elements are and the closest to zero the non-diagonal elements are, the better the metrics of the model will occur. Despite its usefulness in balanced classification problems, accuracy can be a misleading metric, especially in cases of a highly imbalanced training set or a multiclass classification problem. For example, in an ensemble of 100 objects where 97 of them are positive and the model predicts all of the 100 as positive, it is clear that it totally fails to identify the negative cases, while it still hits an accuracy of 97%. This dictates the introduction and use of quantities that can measure the loss of information for each of the classes involved in the training, such as the ones described in eq. 2.4-2.5.

Recall is the ratio of the successful positive predictions, to the total number of positive entries. This metric is referred to as completeness and we will use this terminology throughout the thesis.

$$Recall = \frac{TP}{TP + FN} \tag{2.4}$$

In probabilistic terms, recall answers the question 'how probable is to pick a true positive out of the actual positives ensemble'. Based on that definition, recall is also referred to as sensitivity, or true positive rate (TPR). It is related to the false negative rate (FNR) as $TPR + FNR = 1$. Simirarly, the true negative rate (TNR), or else specificity, is expressed by the fraction:

$$True\ Negative\ Rate = \frac{TN}{TN + FP} \tag{2.5}$$

Precision refers to the ratio of the successful identifications, to the total number of positive predictions, those not identified correctly included. This ratio is also known as purity. Throughout the thesis we will use this terminology to quantify the effectiveness of our machine learning models.

$$Precision = \frac{TP}{TP + FP} \tag{2.6}$$

In other words, precision attempts to answer the question 'what proportion of the positive predictions is actually correct'. Lastly, the F1-score metric is a combination of recall and precision, as the harmonic mean of the two:

$$F1_score = 2 \cdot \frac{Recall \cdot Precision}{Recall + Precision} = \frac{2TP}{2TP + FP + FN} \tag{2.7}$$

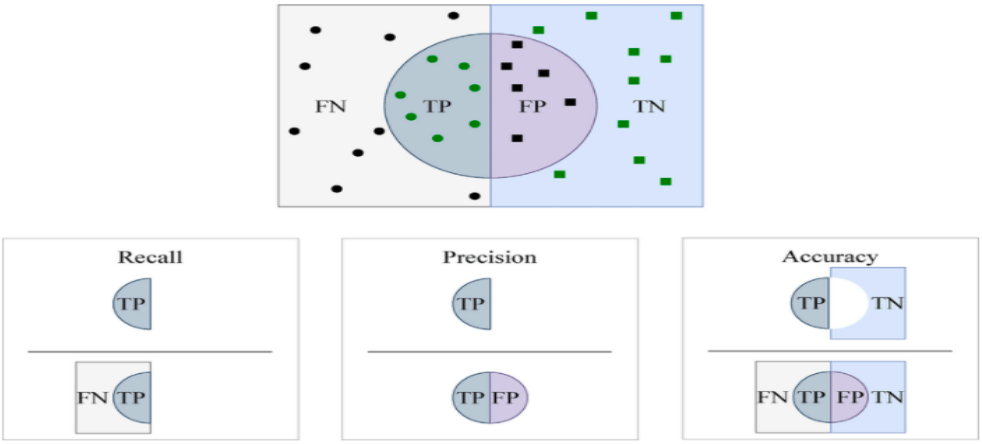


Figure 2.4: Upper channel: The sample space of an ensemble of two populations. The samples can belong to two classes (positive or negative samples). The left half of the rectangle represents the space where the positives lie, while the right one represents the negative samples' space. The green dots/squares are the samples whose class was predicted correctly by the model. The circle encloses all the positive predictions, whether they were correct or false. Lower channel: Schematic visualization of recall's, precision's and accuracy's defining ratios.

The situation becomes more clear through the confusion matrix: a visualization tool to express the performance of a model in assigning classes to objects (Table 2.1). Here the contamination from objects of the mislabeled class is more obvious and different classification problems should focus more on different elements of the confusion matrix. For example, if one wants to identify as many objects of the 'positive' class as possible, he should construct a model that minimizes the false negatives and care less about the false positives. If so, it would mean that the positive class would surely be contaminated by negative objects, but there would be minimum loss of positive objects, mislabeled as negative. This situation describes well a classification problem that focuses on the outliers of a class, as in our case. We may attempt the best fitting classification of point sources into stars, quasars or galaxies, but our main concern remains the inclusion of as many quasars as possible, even if their photometry resembles that of another class. This is the ingredient that will push our classification method to a more complete census of the quasar population, since outliers will not be excluded by the process. The contamination induced by the false positive stars, will be eliminated by the introduction of extra astrometric features that will be discussed later on. By its definition (eq. 2.4), recall is correlated to the false negatives. The lower the FN, the higher the recall value. Therefore, throughout our machine learning analysis presented in the methodology section, the completeness will play the most important evaluation criterion.

Table 2.2: Confusion matrix of a multiclass classification problem

		Predicted Class			
		A1	A2	···	An
Actual class	A1	TP	FN	···	FN
	A2	FP	$A_{2,2}$	···	$A_{2,n}$
	···	···	···	···	···
	An	FP	$A_{n,2}$	···	$A_{n,n}$

2.4 Regression in Machine Learning

Regression is a technique for finding the relationship between independent variables or features (as they normally called in ML) and a dependent variable or outcome. Since this relationship is established, outcomes can be predicted on new unseen data. In contrast to the classification methods, regression is making predictions on continuous outcomes. The final stage of a regression draws a line of the best fit through the training data points. The distance between each point and the line is minimized in order to achieve the best fit. As with all supervised machine learning techniques, the labelled training data should be representative of the overall population in order to produce a trustworthy model. If the data are not representative, the model will produce inaccurate predictions. Moreover, special care should also be taken to include the right selection of features. That is because regression finds the correlation between these features and the target outcome. Regression can be achieved by different kind of models. Based on the estimated function and error chosen, the following types of regression can occur.

Linear Regression

In this type the goal is to fit a line by minimizing the sum of the mean squared error for every data point. Basically this method is the same as the least squares method. In mathematical terms this can be achieved by the following formula:

$$\hat{y}_i = f(x_{ij}) = \beta_0 + \sum_{j=1}^{n} \beta_j x_{ij} \tag{2.8}$$

where n is the number of features or independent variables, β_j is the coefficient of the regression for each feature and i is a specific observation. Mean squared error minimization of the β coefficients is expressed as:

$$min_\beta \sum_{i=1}^{p} ||y_i - \hat{y}_i||^2 \tag{2.9}$$

Polynomial Regression

The previous kind of regression makes the assumption that the dependent and independent variables are linked through a linear relationship. As it is obvious this kind of model will fail to fit data points that the correlation between them is not linear. Polynomial regression succeeds in overcoming this problem by fitting a polynomial

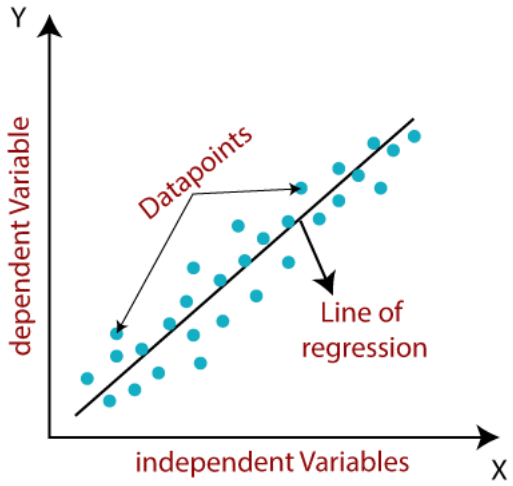


Figure 2.5: Plot that demonstrates the linear correlation between the dependent variable and the independent ones. The optimal line is the one that minimises the mean square error among the data points.

of degree m to the data points. Once more in mathematical terms polynomial regression looks like this:

$$\hat{y}_i = f(x_{ij}) = \beta_0 + \sum_{i=1}^{m}\sum_{j=1}^{n}\beta_j x_{ij}^m \tag{2.10}$$

where m is the degree of the equation and n is the number of features. Once again the mean squared error minimization can be calculated as:

$$min_\beta \sum_{i=1}^{p} ||y_i - \hat{y}_i||^2 \tag{2.11}$$

which is the same expression used in linear regression too, but now $\hat{y}_i$ is defined by eq. 2.10.

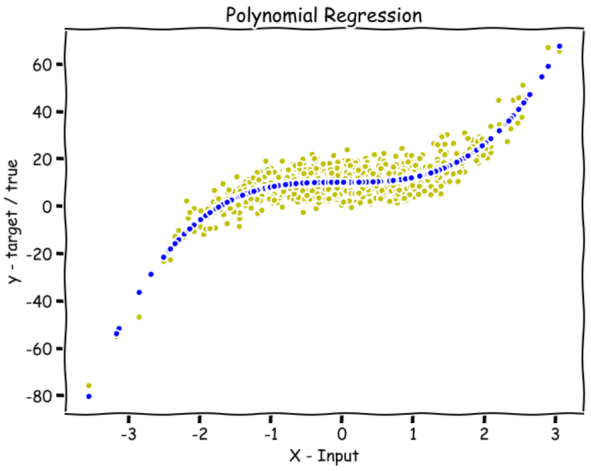
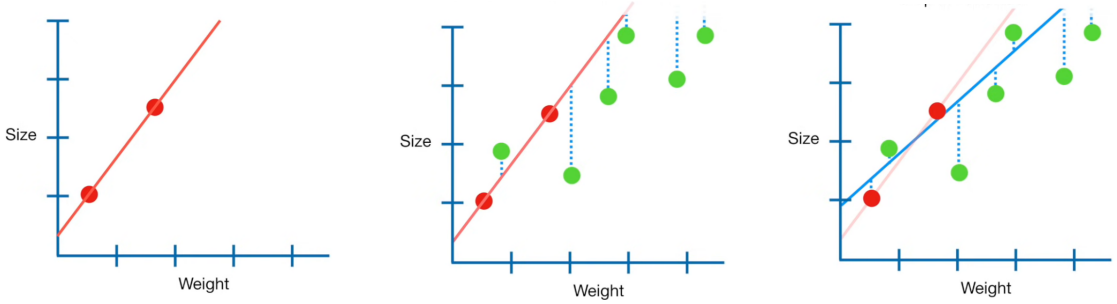


Figure 2.6: Plot demonstrating the polynomial relation between the dependent and the independent variables. A simple line could not optimally fit the data points and thus a more complex relation is correlationg the variables.

Before moving to the next type of regression it would be useful to discuss about the notions of bias and variance in machine learning regression. Let us consider the case of Fig. 2.7.



(a) (b) (c)

Figure 2.7: Bias and Variance balance using the Ridge regression method. (a) Least squares method for finding the best fit between the two red training points. The bias in this case is extremely small since both the data points fall exactly on the estimated line. (b) Green dots represent the test set. As can be seen the model is overfitted on the training set and generalizations are hard to occur, resulting in bad predictions on the test set. The variance in this case is high. (c) Using the Ridge or Lasso regression method the overfitted model can be improved by finding the line that balances between the values of bias and variance.

The red dots in the first image represent the training set and so the red solid line represents the least squares best fit between these two points. The fit may look excellent for the training data points but it may cannot generalise to any given unseen dataset. In this case we say that the bias is super low. As can be seen in Fig. 2.7b, where the green dots represent the test set, this calculated least squares line is way off the average data point of the test set. Therefore, in this case the variance is very high, even if the initial bias is very low. In machine learning language that translates to a model that is overfitted on the training data. The optimal model would be made based on a balance between the the bias and the variance. By introducing a small amount of bias a better variance value can be found.

Ridge Regression

This new line described above, that balances the bias and the variance, is what ridge regression is made to calculate. This type of regression addresses exactly the issue of overfitting. In a case where the polynomial function, for example, has a very high degree (let us say m=25 for 10 training points) then all the training points would be fitted correctly but with the cost of having very weak generalization for other unseen data. So what ridge regression does is to minimize the generalization error by compromising with the intoduction of bias. Mathematically this is achieved by modifying the mean square error or loss function.

$$min_\beta \sum_{i=1}^{p} ||y_i - \hat{y}_i||^2 + \alpha||\beta||^2 \tag{2.12}$$

where $\alpha \in \mathbf{R}$ is a positive scaling factor.

In this type of regression the function can be either a linear or a polynomial one. When ridge regression is not included the weights of the function tend to be pretty high. By using ridge regression overfitting is avoided by limiting the value of the weights with the introduction of the term $\alpha||\beta||^2$. This term is also called L2 norm. A small value of α would potentially lead to overfitting while a large value would result in a more conservative model.

LASSO Regression

This kind of regression is very similar to the Ridge one discussed above. Both of these models use regularizations against overfitting. In the Ridge regression the regularization factor was the L2 normalizer which was the addition of the squared magnitudes of the coefficients, $\alpha||\beta||^2$. In the LASSO regression case the regularization factor is called L1 and is simply the addition of the absolute value of the magnitude of the coefficients, $\alpha||\beta||$. So the expression for the mean square error becomes:

$$min_\beta \sum_{i=1}^{p} ||y_i - \sum_{j=1}^{n} f(x_{ij})||^2 + \alpha||\beta|| \tag{2.13}$$

where once more $\alpha \in \mathbf{R}$ is the scaling factor. Another main difference between ridge and lasso regression is that the L1 norm tends to induce sparcity to the weights. This means that there are going to be elements/features with weights equal to 0. Using the L2 norm weights can become really small but never zero.

ElasticNet Regression

The elasticNet method is a combination of the previous two ones, meaning the ridge and lasso regression. The loss term now includes both the L1 norm and L2 norm giving rize to the mean square error that can be seen below:

$$min_\beta \sum_{i=1}^{p} ||y_i - \sum_{j=1}^{n} f(x_{ij})||^2 + \alpha_1||\beta|| + \alpha_2||\beta||^2 \tag{2.14}$$

Part II

Methodology

Overview

In part II we discuss all the stages of handling the data available from different surveys. From quering them, to pre-processing them, to using them to build different machine learning models in order to treat the classification problem.

The process begins with gathering the data. This step is about combining the photometric and astrometric observations of different surveys in a wide range of wavelenghts. In a few words, what we are attempting to do is to train machine learning models in distinguishing quasars from galaxies and stars, using spectroscopically known objects from the SDSS DR16 survey. For higher accuracy, we enrich the information on these objects by cross-matching them with other surveys (focused on measurements in the infrared, or on the kinematics of the sources). To achieve that we utilize the online tool : CDS-XMatch. The gathering of the data is analytically presented in the Appendix A. Before the data are used as training sets for any machine learning model, their assessment is required. For instance, we need to account for NaN values in our training catalogues, or for extremely high values, way beyond the typical ranges of a variable. The latter could correspond to a bad measurement, or a non existing measurement that is usually substituted with the number 9999. Secondly, the bias that every survey may induce has to be given some thought. Such considerations are discussed in chapter 3. There, the features selected for the training are also discussed.

For the automatised classification we use the XGBoost Classifier (eXtreme Gradient Boosting), a tree-based algorithm which sits under the supervised branch of machine learning (section 2.2.3). In order to examine the behaviour of the classifier we run a number of tests and evaluate them using the completeness, purity and accuracy metrics. We build 3 purely photometric ML models trained on the three main classes: **stars**, **quasars** and **galaxies**. We have also examined the binary classification problem where only stars and quasars are included. Moreover, an investigation of the effect of adding astrometric features on ML algorithms takes place. We create a purely astrometric model and 4 more ML models that use both astrometric and photometric features. Finally, a model that is trained to distinguish different stellar classes and quasars based on photometric and astrometric features is presented. All the models and their performance are presented in chapter 4. Depending on the available observations one may have on an object, they can choose a different trained model to make predictions. As a final use of machine learning on quasars, we present the performance of the XGBoost-Regression model on predicting the redshift of quasars, based on their photometry. If developed well, this could be a very useful tool in locating and observing very distant quasars.

Chapter 3

Data Analysis

In this chapter we present the process of cleaning all the data that were used for training the various machine learning models. In the first part of the chapter we discuss the pre-processing of the raw data that needs to take place before we use them for the training. It is reasonable that any algorithm's accuracy and validity in the predictions it makes, relies first and foremost on the quality of the data used to build the training set. It is therefore very crucial to create datasets that are complete, carefully inspected and sensibly filtered out. In the last section of this chapter we present all the acquired observations and determine the important colors for separating the different astronomical objects through color color plots, in order to have a preliminary idea of what features we should use for the training. This exploration led us to revised versions of empirical color criteria found in the literature. For comparison, we test our new estimations and the former relations on the same sets and present each one's accuracy.

3.1 Pre-processing

3.1.1 Classification

As it is discussed in Appendix A, many surveys were cross-matched and their measurements were combined to create a catalogue as complete as possible in the photometric magnitudes and astrometric features it includes. But this comes with a cost; each cross-matching successively reduces the length of the catalogue more and more. That is because for some patches of the sky there may be no overlap in the sky area that each survey has measured, and so, objects listed in one survey may not be recognised by another one.

Dealing with NaN values

Another important step of the dataset cleaning was to account for the non-existing (NaN) values on the important columns that would later serve as features for the training. Inspection of the catalogues revealed the existence of many rows with NaN magnitude values that had to be removed, such that our training set was as trustworthy as possible. Erasing the rows that have NaN values in general is not the only way to deal with this problem. One different approach would be to impute these cells with specific values such as the mean or median value of the whole column. In this work we could afford the method of erasing the whole rows, since the initial datasets were already big enough. This method was preferred as the one inducing less errors and leading to the most robust model. In the optical bands (SDSS), there were no NaN values. However, due to the non-measured WISE magnitudes, there is a loss of 424 stars, 263 quasars and 1224 galaxies. Also, the

non-existing near-infrared colors (Y,H,J,K) caused an additional loss of 53.474 stars, 56.761 quasars and 180.861 galaxies. The overall data processing suggests that the inclusion of near-infrared photometry is the source of the most dramatic decrease of the training datasets. This may not be so harmful for the accuracy of the algorithms in this case, where the datasets are very large, but could bring up an issue in the following situation. One can imagine having a short list of unknown objects that he wants to classify, but he misses the near-infrared colors. A potential cross-match with the UKIDSS survey could possibly result in the depletion of the candidate list and the prediction would be untrustworthy or even impossible. In such a situation a solution could be to use different surveys to obtain these magnitudes, such as the 2MASS, however these measurements are not as reliable. Another solution could be to use models that are not trained on the near-infrared magnitudes. The accuracy would drop in this case, but at least predictions could be made on a larger population.

Losses from cross-matches

From the SDSS query a total number of ∼1 million stars, ∼2 million galaxies and ∼750 thousand quasars was acquired. The cross-match with the AllWISE sky survey reduced the length of those datasets, but the most important impact came from the cross-match with the UKIDSS survey. The latter caused a drop to the dataset size by about one order of magnitude. The exact total numbers for every class after each cross-matching are gathered and shown in Fig.3.1.

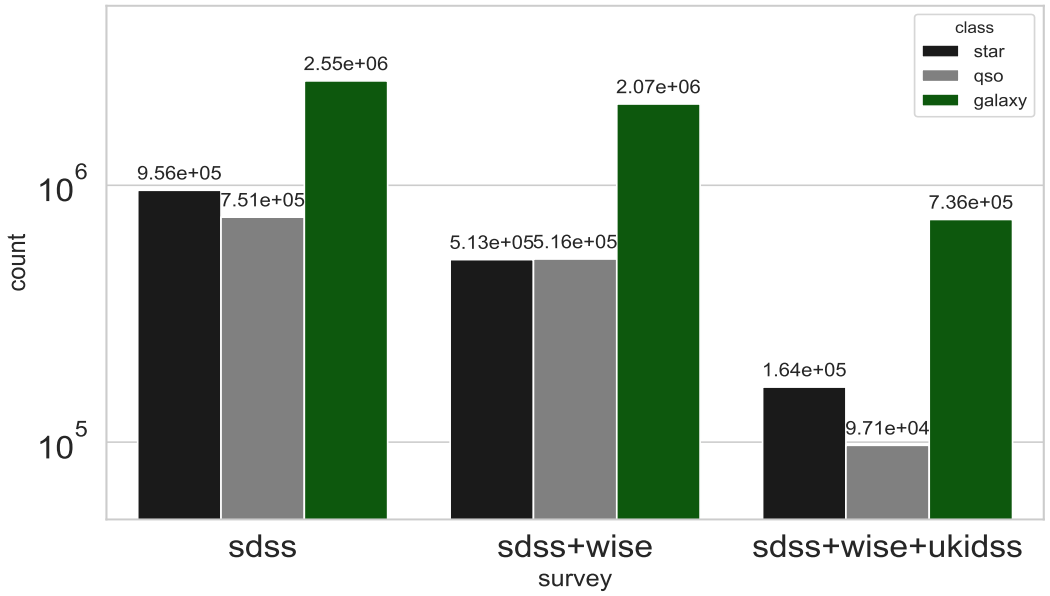


Figure 3.1: Histogram showing the size of the stellar, quasar and the galaxy catalogues that are used as training sets, after each cross-matching with a different photometric survey and after the deletion of rows with NaN values.

Similar process was followed in the case where the astrometry of the sources was taken into account. It is evident from Fig.3.2 that the GAIA survey aims for point sources, and as a result its inclusion massively cuts off the majority of the galaxies, as extended sources. The contribution of the NaN values to the loss of objects accounts for 234 galaxies.

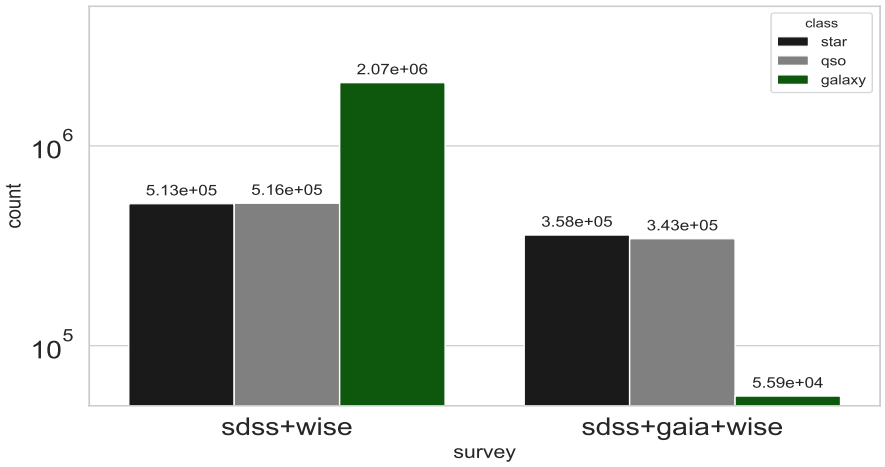


Figure 3.2: Histogram showing the size of the star, the quasar and the galaxy catalogues that are used as training sets, after the cross-matching with AIIWISE and GAIA survey and after the deletion of rows with NaN values.

Redshift distribution

For the completeness of the datasets, the whole range of redshifts between the spectroscopically known quasars and galaxies is included (Fig.3.3). It can be seen that the quasar population's redshift follows a gaussian-like distribution centered around z = 1.5. Many extremely distant quasars are also included, but due to their rareness are underrepresented.

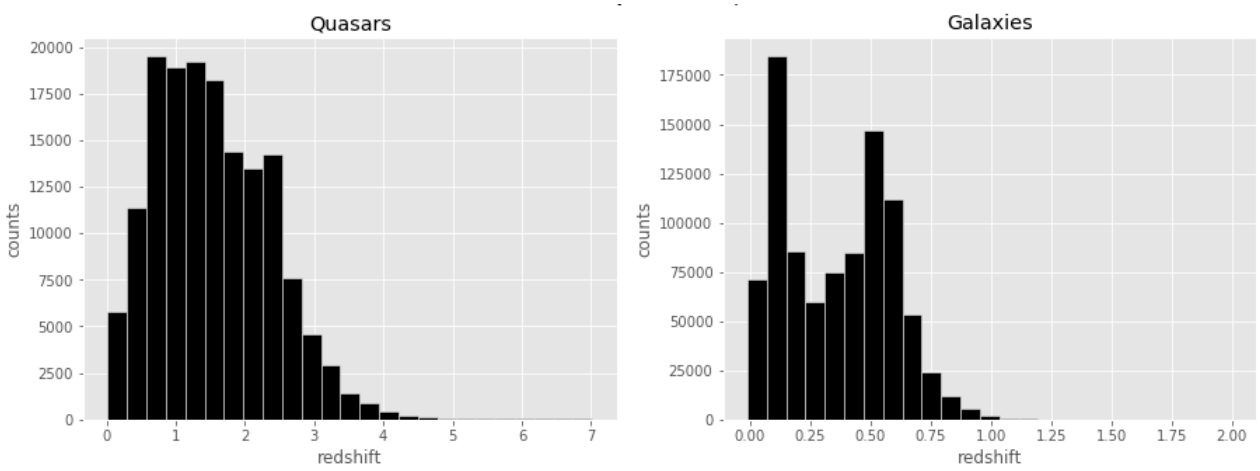


Figure 3.3: Histrograms of the redshift distributions for data points in the training dataset. Left: Redshift distribution for quasars peaks at around 1.5. A small number of quasars is found in high redhsifts (for $z > 4$). Right: Redshift distribution of the galaxy population ranges from $\sim$ 0to $\sim$ 1, showing two peaks, one at around 0.15 and another at around 0.5.

3.1.2 Regression

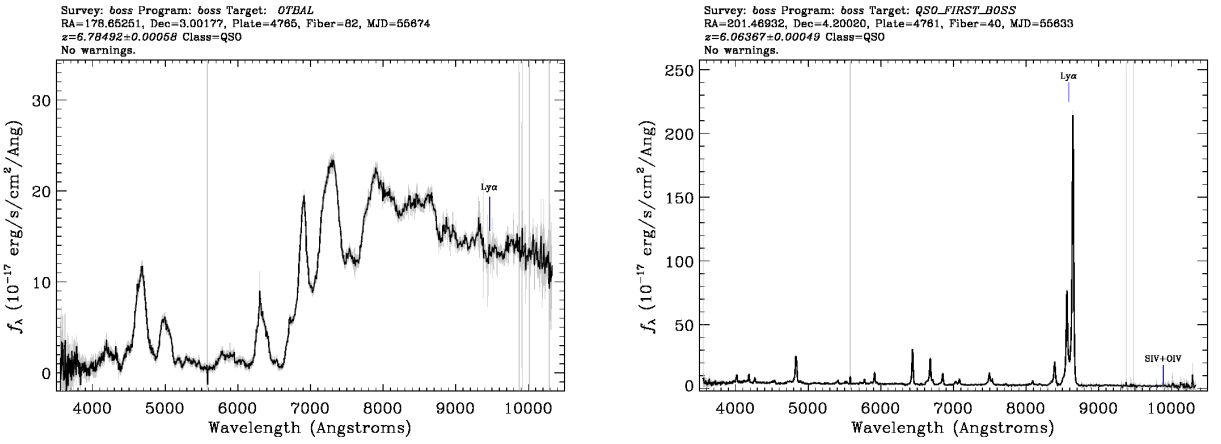
For the regression part (as will be discussed in Section 4.4) the features used for the training are the SDSS, WISE and UKIDSS photometric magnitudes. The spectroscopic redshift from the SDSS DR16 survey is used as the dependent target variable of the regression model. The initial dataset is comprised of 153.910 spectroscopically classified quasars with known redshifts (Fig. 3.3).

NaN values

After removing the NaN values, the final catalogue has a number of 96.736 quasars. Once again, optical photometry from SDSS had no NaN values, WISE survey's NaN values reduced the quasar catalogue by 263 and finally UKIDSS had the major loss contibution of 56.761 NaN values.

Untrustworthy SDSS redshifts

The first regression training we made included all available redshifts up to $z = 7.017$. It should be noted here that redshifts $z > 5$ represent only 0.28% of the total training set and thus this redshift range is highly underrepresented (34.5% of objects have $0 < z < 1$, 36.7% of objects have $1 < z < 2$, 23.2% of objects have $2 < z < 3$, 4.6% of objects have $3 < z < 4$, 0.57% of objects have $4 < z < 5$). Due to that, it is safe to say that the regression model will behave badly for redshifts higher than 5. Additionally, we encountered another problem concerning these high redshifts. We visually examined some of the SDSS spectra for redshifts higher than 6. What we have found (Fig. 3.4) is that many SDSS spectra have wrong redshift identifications. Thus SDSS redshifts, specifically in high redshift ranges, cannot always be trusted. As a result, for the regression part of the training only, we dropped out all the observations for redshifts higher than 5, and made a model which is trained on the rest, more trustworthy redshifts.



(a) Spectrum of a quasar from SDSS DR16. Predicted redshifts by the SDSS pipeline is 6.78. It is obvious though that the $Ly\alpha$ line does not lie where it is marked.

(b) Spectrum of a quasar from SDSS DR16. Predicted redshifts by the SDSS pipeline is 6.06. It can be seen that the $Ly\alpha$ line in reality is the $OIII$ doublet lines. So this object's redshift is highly overestimated by the SDSS, as its real redshift is around 0.7.

Figure 3.4: Examples of bad SDSS assigned redshifts to quasar observations. For both cases the redshift seems to be significantly overestimated.

3.2 Feature Selection - Empirical relations revisited

The final and most complete dataset is filled with 4 coordinates, the equatorial and the galactic ones, 13 photometric features, 4 astrometric, the redshift and the class/subclass of each object. The number of the photometric/astrometric features used depends on the individual needs of a training and differs from ML model to ML model. For example, as we show in the next chapter, a purely astrometric training is attempted, where all the photometric features are neglected. In this case, we sacrifise the completeness of the feature collection in order to ensure the completeness in data points (no cross-matches with other surveys, ans thus longer training

dataset). In other cases we made models based on all the available features and trained on the shrunken datasets. Table 3.1 shows all the available features used for training purposes.

Table 3.1: Photometric and astrometric measurements obtained from SDSS, AllWISE, UKIDSS, GAIA surveys that will serve as features for the machine learning training.

Feature	Description
ra	Right Ascention
dec	Declination
b	Galactic Latitude
l	Galactic Longitude
redshift	Redshift as measured from the SDSS
class	Best spectroscopic classification (Star, QSO, Galaxy)
subclass	Stellar subclasses, Starburst/starforming galaxies, AGNs, Broadline QSOs
u	u band magnitude (SDSS measurement)
g	g band magnitude (SDSS measurement)
r	r band magnitude (SDSS measurement)
i	i band magnitude (SDSS measurement)
z	z band magnitude (SDSS measurement)
W1mag	W1 band magnitude (AllWISE measurement)
W2mag	W2 band magnitude (AllWISE measurement)
W3mag	W3 band magnitude (AllWISE measurement)
W4mag	W4 band magnitude (AllWISE measurement)
Y	Y band magnitude (UKIDSS measurement)
J	J band magnitude (UKIDSS measurement)
H	H band magnitude (UKIDSS measurement)
K	K band magnitude (UKIDSS measurement)
parallax	Parallax as measured by GAIA mission
S/N_{par}	Signal to noise ratio for parallax
pm	Proper motion as measured by GAIA mission
S/N_{pm}	Signal to noise ratio for proper motion
phot g mean mag	g band mangitude as measured by GAIA mission
phot bp mean mag	Proper motion as measured by GAIA mission
phot rp mean mag	Proper motion as measured by GAIA mission

The visualisation of the data can be made using color-color plots. Through them, one can get a preliminary idea of the colors that will play the most important role as features of the machine learning. In specific color plots, such as the W1W2-JK, JK-iY, YK-gz, the separation is remarkable, while for others such as the W3-W4 color, the typical value range is the same, both for quasars and stars. There is an overall trend for the highly redshifted quasars ($z > 2$) to fall very close to the stellar cluster, which explains well how they could have been overlooked with the previous selection techniques. The situation is best demonstrated in subfigures 3.5(c) where it is clear that the UV excess method would fail in retrieving most of the quasars with redshift higher that 2.

For all the subplots of Fig. 3.5 we find a linear relation between the colors, shown with a dashed red line, that represents our best estimation for separating the stars from the quasars. Our best fitting results provide the following empirical criteria that should be fullfilled when a source is a quasar:

$$(g - r) < 2.1(J - K)[AB] + 1.2 \tag{3.1}$$

$$(W1 - W2)[AB] > -(J - K)[AB] - 0.3 \tag{3.2}$$

$$(g - r) > -5.5(u - g)[AB] + 4.79 \tag{3.3}$$

$$(W1 - W2)[AB] = -0.4 \tag{3.4}$$

$$(r - z) > (u - g) - 0.6 \tag{3.5}$$

$$(J - K)[AB] > 0.06 \tag{3.6}$$

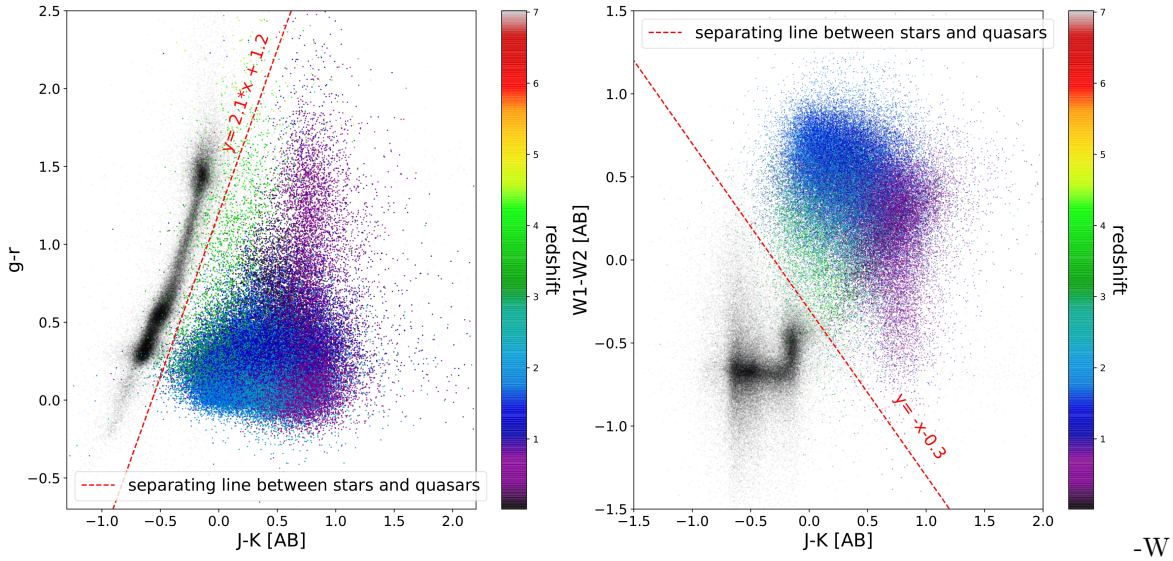
$$(J - K) > 0.24(g - J) + 0.15 \tag{3.7}$$

$$(J - K) > 0.45(i - Y) + 0.37 \tag{3.8}$$

$$(Y - K) > 0.46(g - z) + 0.53 \tag{3.9}$$

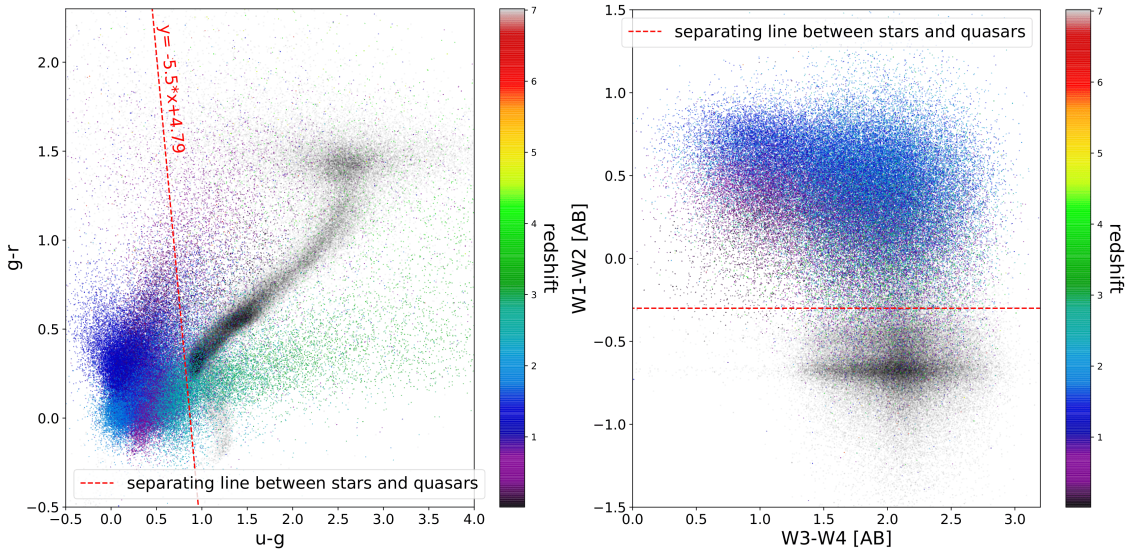
It should be noted that eq. 3.7-3.9 make use of the Vega photometric system and are revisited versions of criteria found in the literature, also formulated in the Vega system. The latter are marked with the green dashed lines in the three last subplots, where our current estimations are overplotted as well. Wherever the AB magnitude system is used for the near and mid-IR colors, it is stated in the corresponding equation.

Eq. 3.9 is in agreement with the criterion found in the work of [Wu & Zhendong, 2010]. They introduce it to separate the different sources using their optical and near-infrared colors. This relation applies better to quasars with redshift z<4. Despite its efficiency, this criterion fails to recover fainter objects in the z-band and for such cases the authors propose a Y-K versus g-Y criterion. For quasars with redshift higher than 4 they propose a second criterion, $(J - K) > 0.45(i - Y) + 0.64$, for which eq. 3.8 is the modified version, proposed in our work. The change is not dramatic, since the two equations differ only by a coefficient. Yet, as can be seen in subfigure (i), this small change leads to a significantly larger inclusion of quasars that were excluded before. This is also reflected in the accuracies the different criteria reach, when applied to the test datasets. As for eq. 3.7, it is the revised version of the criterion proposed by [Warren et al., 2000], $(J - K) > 0.36(g - J) + 0.18$. Both the coefficients of this first order equation are different in our estimation. This alternation aims to lift the inability of Warren's proposal to include highly redshifted quasars.



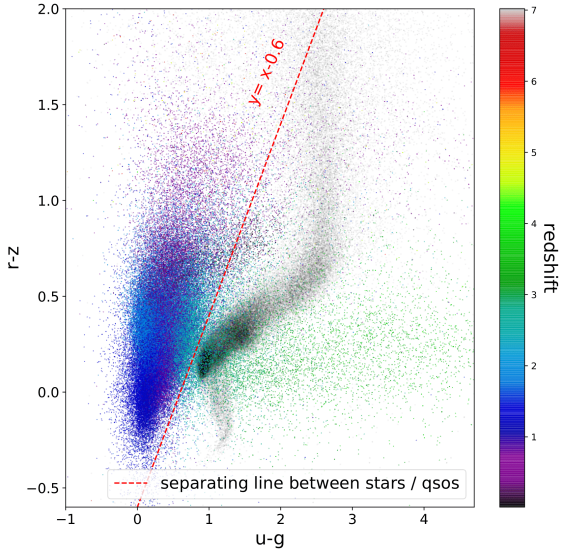
(a) g-r vs. J-K in AB magnitudes

(b) W1-W2 vs. J-K in AB magnitudes

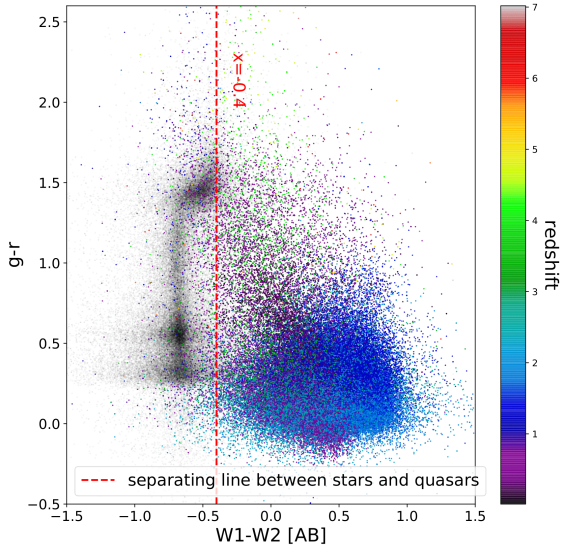


(c) g-r vs. u-g in Vega magnitudes

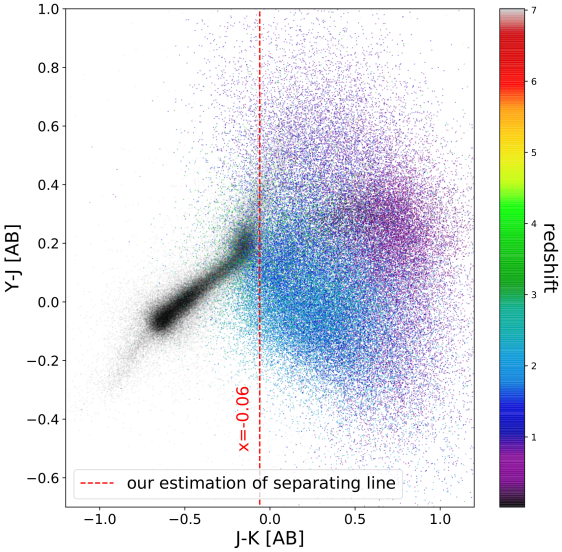
(d) W1-W2 vs. W3-W4 in AB magnitudes



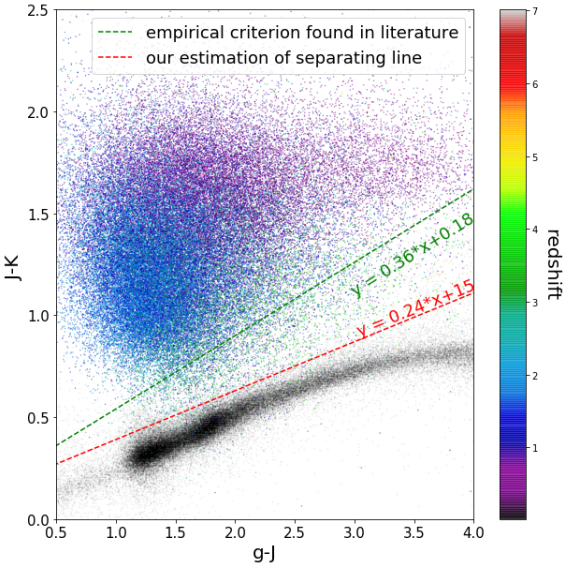
(e) r-z vs. u-g in Vega magnitudes



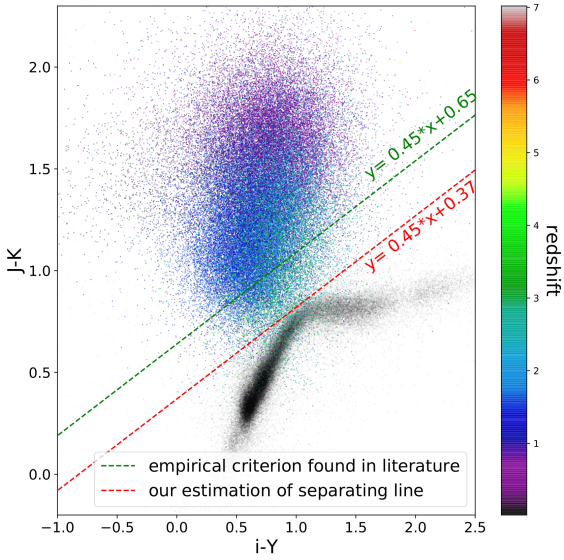
(f) g-r vs. W1-W2 in AB magnitudes



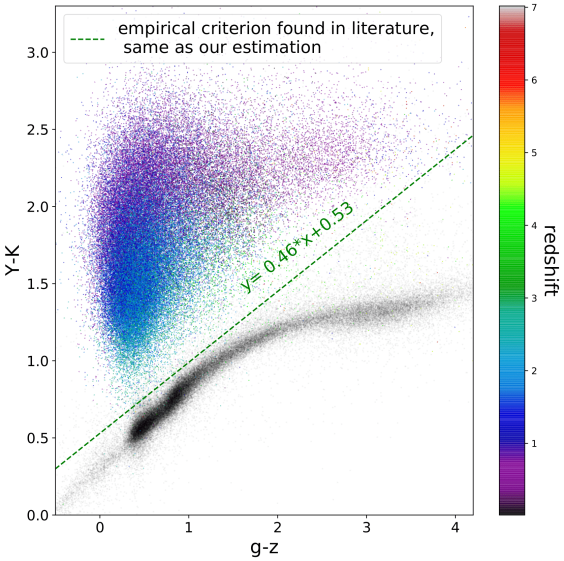
(g) Y-J vs. J-K in AB magnitudes.



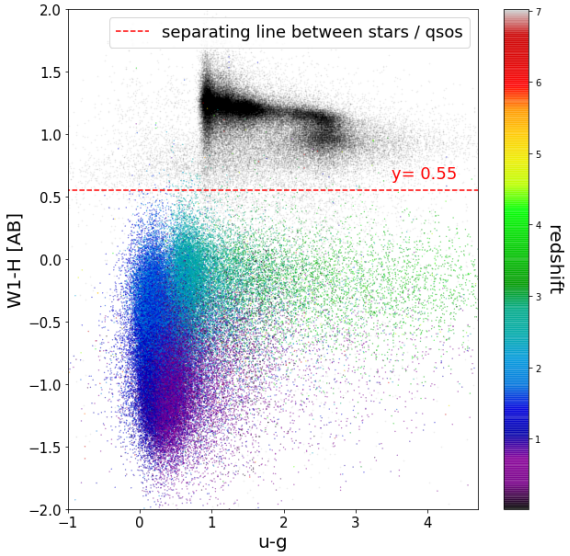
(h) J-K vs. g-J in Vega magnitudes



(i) J-K vs. i-Y in Vega magnitudes



(j) Y-K vs. g-z in Vega magnitudes



(k) W1-H vs. u-g in AB magnitudes

Figure 3.5: Visualisation of the training datasets according to their class, through color color plots. In all the subplots, only 100.000 objects of each class are randomly selected to be shown. The black dots correspond to stars, while the quasars are differentiated according to their redshift, appearing to move in groups along the color-color plots while their redshift evolves. Galaxies are not included. In each plot the formation of class-clusters is evident, but depending on the color selection there might be overlaps, or the clusters might appear distinctly separable. The reader should pay attention to the axes units. In some subplots the magnitude system is the AB and in some is the Vega. We kept the Vega system in subfigures (H), (I) and (J) for our linear criteria estimations to be compatible with the ones found in literature.

In this work we handle larget datasets of sources and this gives the advantage of creating a bigger picture of where the boundaries in the color ranges lie. However, in Fig. 3.5 only 100.000, randomly selected objects of each class are plotted. If the proportions were realistic and we had an enormous amount of stars, they would have almost filled the x-y surface and the discrimination of different classes using empirical relations would be a much

harder thing to do. It should also be noted that the lines we draw lie closer to the stellar locus, such that less quasars are left out. At the same time though, more stars cross the line and contaminate the quasar catalogues. In other words, the modifications we made on the empirical criteria can result in a better completeness but lower purity of the quasar catalogue. What is anticipated from the machine learning is to overcome such obstacles that color selection techniques would have failed. To hopefully be able to predict correctly the different objects, even if they fall outside their class's standard color range. Apart from the traditional color criteria, in what follows we also explore the astrometric cut-offs between stars and quasars. Subfigure 3.6(a) shows that the majority of quasars are clustering for S/N_{pm} lower than ~ 3. Proposed cut-offs in the literature set the limit to $S/N_{pm} < 2$, achieving to downsize the stellar contamination. Quasars with values higher 2 are shown in bigger brown dots, showing that such a cut-off leaves aside a notable percentage ($\sim 4\%$) of the total quasar population. There are also some quasars with extremely high S/N_{pm} values, but maybe their astrometric measurements are not so trustworthy. Driven by these ambiguities, we drop such observations from the training dataset, as they could strongly influence the results. Finally, an empirical relation to determine whether a source is quasar is found between the S/N_{par} and W1-W2 color, reading:

$$S/N_{par} > 20(W1 - W2) - 3.3 \tag{3.10}$$

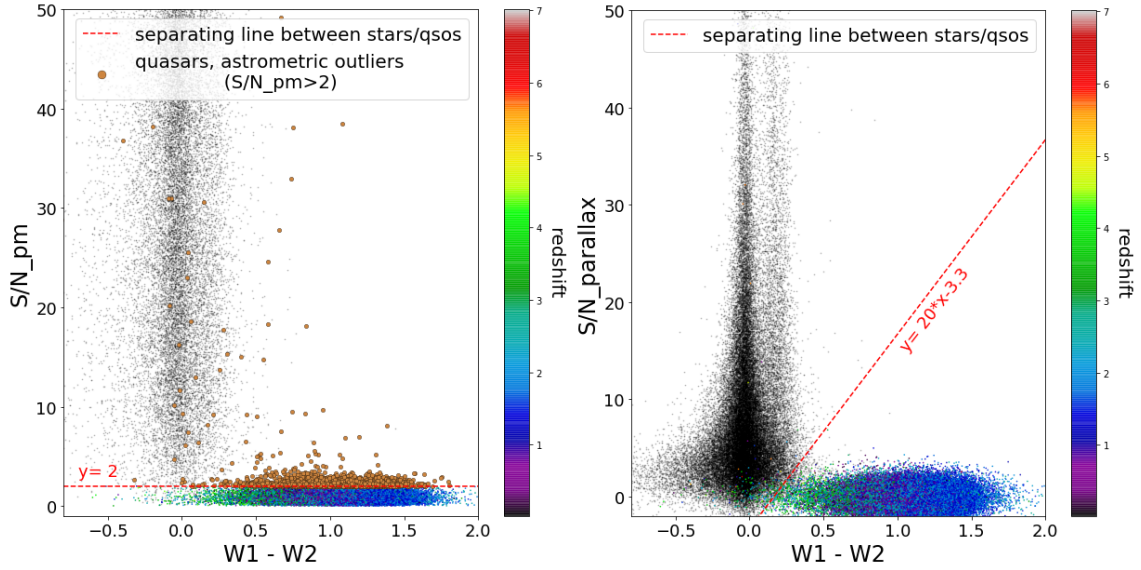


Figure 3.6: Visualisation of the training datasets according to their class, through astrometric - color plots. Left: S/N_{pm} vs. W1-W2 color. Brown dots are quasars with signal to noise ratios higher that 2. High redshift quasars are located very close to the tip of the stellar cluster. Right: S/N_{plx} vs. W1-W2 color. Red dashed line is a proposed separating criterion.

Chapter 4

Machine Learning Models

Up to this point the first stages of the machine learning process have been accomplished. These include the extraction, preparation and pre-processing of the raw datasets. In this chapter we discuss the training on those data. Different approaches have been explored, from the application of various algorithms, to numerous feature selections. In the beginning of the chapter a purely photometric classification is presented. The training features in this case extend from the optical magnitudes of the SDSS survey to the near and mid infrared photometric bands. All the differences among magnitudes of the same survey (colors), are also included as features of the training but can be partially neglected afterwards, based on their feature importance. Later on the chapter we present the results of a purely astrometric model. Diving deeper into the quasar selection process, we investigate the addition of astrometric properties to the list of photometric features. The analysis revealed the importance of astrometric features, as the new model's accuracy climbed to even higher values.

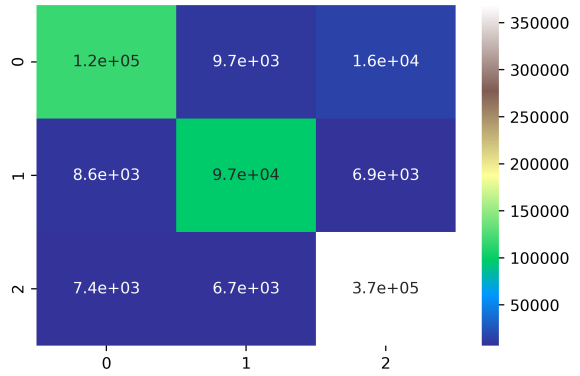
4.1 Purely Photometric Classification

Table 4.1: Performance of all the photometric XGBoost models on the unseen **test set**. Number of samples (qso+star+galaxy) vary depending on how many cross matches with other surveys have been made. Results for both the balanced and unbalanced trainings are shown.

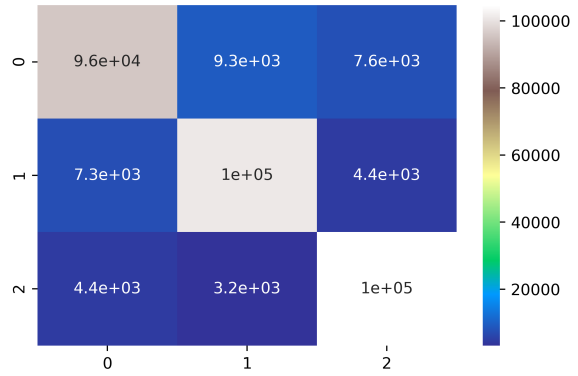
MODEL	DATASET SIZE	TRUE QSO PREDICTIONS	ACCURACY
Three Classes			
SDSS $_{unbalanced}$	638424	96718	0.914
SDSS $_{balanced}$	337674	100921	0.894
SDSS|WISE $_{unbalanced}$	308024	75148	0.979
SDSS|WISE $_{balanced}$	231222	75061	0.980
SDSS|WISE|UKIDSS $_{unbalanced}$	93013	13971	0.989
SDSS|WISE|UKIDSS $_{balanced}$	43653	14134	**0.987**
Two Classes			
SDSS|WISE|UKIDSS $_{balanced}$	43653	14568	**0.997**

4.1.1 Optical features

The minimum number of magnitudes comes from the SDSS survey and includes the five optical u,g,r,i,z bands. Subtracted in pairs, they provide ten extra features, the respective color indices u-g, u-r, u-i, u-z, g-r, g-i, g-z, r-i, r-z, i-z. Since there is no beforehand knowledge of every color's and magnitude's significance on the training's accuracy, all the 15 photometric measurements are used as features. The impact of an unbalanced in contrast to a balanced training is studied, revealing a small drop of the accuracy from the initial 0.914 value, to 0.893 for the balanced training. Although this looks in favor of the unbalanced case, a deeper investigation through the confusion matrix shows better overall results concerning the purity and the completeness of the balanced training. The numerical superiority of the galaxy class inherited a tendency to the model to predict more objects as galaxies. This defect appears to be lifted by balancing out the three classes, since this way less stars and quasars are falsely predicted as galaxies. The analytical classification report of the training is shown below.



(a) Confusion matrix of the unbalanced training on the test data. Only optical SDSS magnitudes are used for the training. Accuracy score on the test data : 91.4 %



(b) Confusion matrix of the balanced training on the test data. Only optical SDSS magnitudes are used for the training. Accuracy score on the test data : 89.3 %

Table 4.2: Unbalanced training, SDSS optical photometric features

Class	Purity	Completeness	F1score	Accuracy
Stars	0.881	0.823	0.851	
Quasars	0.855	0.862	0.859	**91.4** %
Galaxies	0.942	0.963	0.952	

Table 4.3: Balanced training, SDSS optical photometric features

Class	Purity	Completeness	F1score	Accuracy
Stars	0.891	0.850	0.870	
Quasars	0.890	0.896	0.893	**89.3** %
Galaxies	0.897	0.932	0.914	

From Fig.4.2 it is evident that the colors are of higher importance for the classification than the magnitudes, with u-i, r-i, r-z, u-g, g-i serving as the best photometric features for distinguishing the different types of sources. Nevertheless, the use of only optical colors cannot provide a model that can surpass the accuracy of 91.4% on

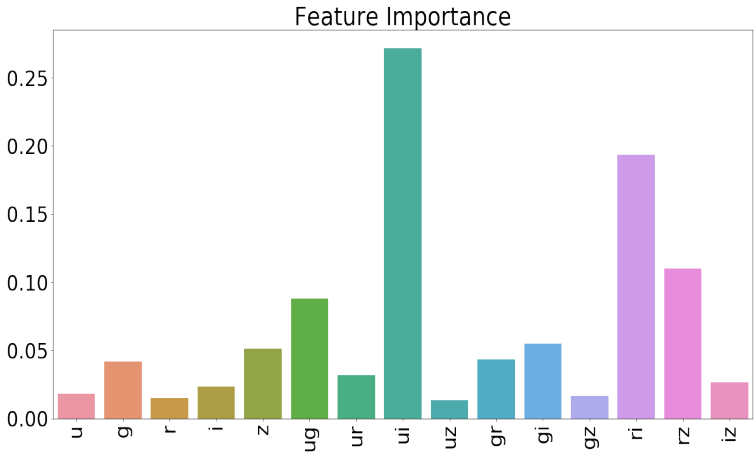
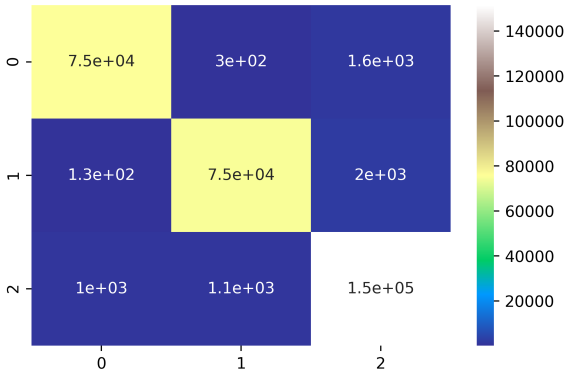


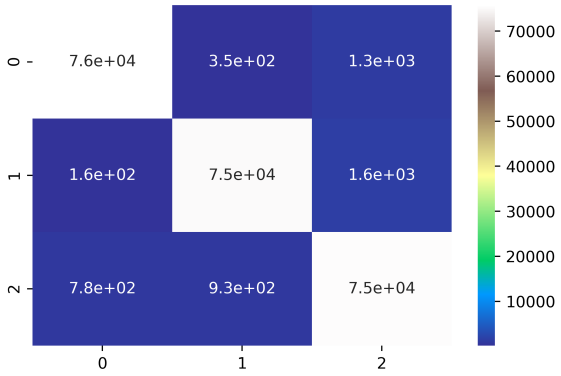
Figure 4.2: Feature importance for the training based only on the optical magnitudes from SDSS. u-i color plays the most important role for the classification, followed by r-i and r-z colors. The magnitudes alone do seem to make a great contribution in distinguish different kind of sources.

the test set, which means that ~ 1 out of 10 predictions is expected to be incorrect. In what follows, we present the impact of the WISE colors inclusion to the training.

4.1.2 Optical and mid-IR features



(a) Confusion matrix of the unbalanced training on the test data. Optical and mid-infrared colors are used for the training. Accuracy score on the test data : 98.0%



(b) Confusion matrix of the balanced training on the test data. Optical and mid-infrared colors are used for the training. Accuracy score on the test data : 97.8%

Table 4.4: Unbalanced training, SDSS optical and AllWISE mid-IR photometric features

Class	Purity	Completeness	F1score	Accuracy
Stars	0.985	0.975	0.980	
Quasars	0.981	0.973	0.977	**98.0**%
Galaxies	0.977	0.986	0.981	

Once again, the size reduction of the galaxy class, required for the balanced training, led to a slightly worse performance on the galaxy class, while it improved the performance on the 2 remaining classes. The situation is demonstrated through the evaluation metrics of tables 4.4 and 4.5 of the imbalanced and balanced training. The feature importance plot shows that W1-W2 color is the most significant feature for the discrimination among the different classes. This result constitutes a strong confirmation of former selection techniques that utilized

Table 4.5: Balanced training, SDSS optical and AllWISE mid-IR photometric features

Class	Purity	Completeness	F1score	Accuracy
Stars	0.988	0.979	0.983	
Quasars	0.983	0.977	0.980	**97.8**%
Galaxies	0.963	0.978	0.970	

the mid-infrared photometry of AllWISE in the search of quasars [Stern et al., 2012]. The r-i and r-z colors continue to play an important role as discriminative features, while in this training the optical magnitudes g and z appear to be more significant than the rest of the SDSS colors.

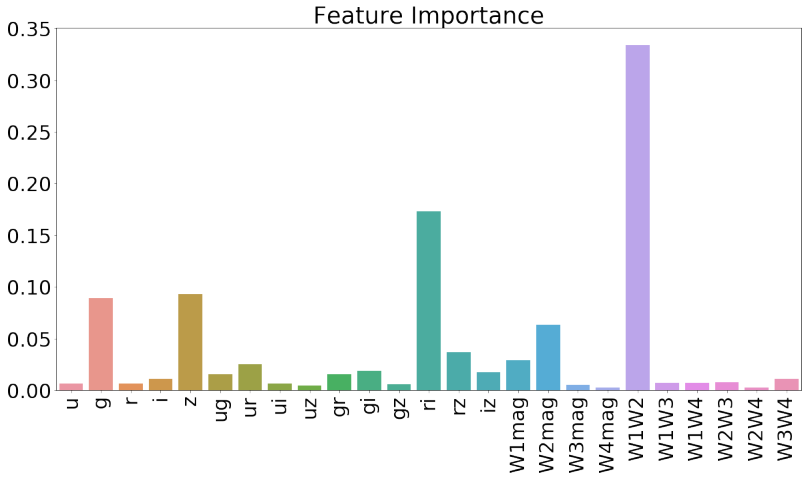
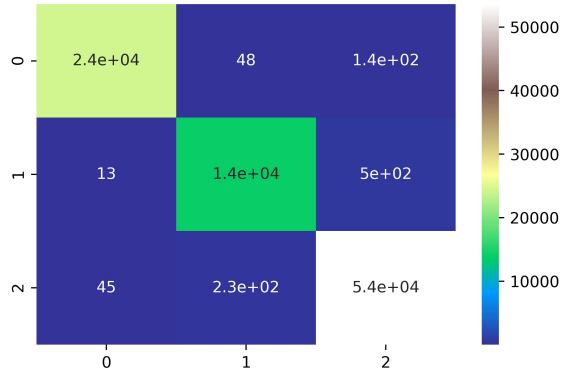


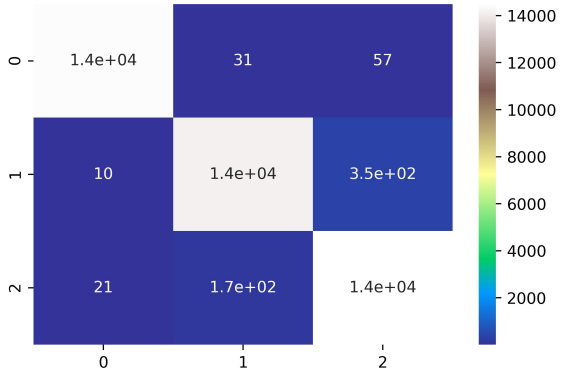
Figure 4.4: Feature importance of the training based on optical and mid-IR magnitudes. W1-W2 color is by far the most important feature of thes training in distinguishing the different types of sources. Second most important is the r-i color, followed by the equally significant g and z magnitudes.

4.1.3 Optical, near and mid-IR features

For the sake of the completeness of the analysis, the UKIDSS near-infrared magnitudes are subsequently taken into account as observations of the training set. The new model for the imbalanced, 3 class classification gives an accuracy of 99.0% for all the available features. By evening the size of the different classes the accuracy drops to 98.5%, but in general the confusion matrix demonstrates a better predicting behavior.



(a) Confusion matrix of the unbalanced training on the test data. Optical, near and mid-infrared colors are used for the training. Accuracy score on the test data :99.0%



(b) Confusion matrix of the balanced training on the test data. Optical, near and mid-infrared colors are used for the training. Accuracy score on the test data : 98.5%

Table 4.6: Unbalanced training, SDSS optical, UKIDSS near-IR and AllWISE mid-IR photometric features

Class	Purity	Completeness	F1score	Accuracy
Stars	0.998	0.992	0.995	
Quasars	0.981	0.965	0.973	**99.0**%
Galaxies	0.988	0.995	0.992	

Table 4.7: Balanced training, SDSS optical, UKIDSS near-IR and AllWISE mid-IR photometric features

Class	Purity	Completeness	F1score	Accuracy
Stars	0.998	0.994	0.996	
Quasars	0.986	0.975	0.980	**98.5**%
Galaxies	0.973	0.987	0.980	

This model's most significant feature turns out to be the J-K color. Previous studies [Warren et al., 2000] have already shown that colors that involve the K-band magnitude, such as the J-K, Y-K, H-K, are crucial in separating the star from the quasar population. The machine learning model we have trained seems to come in agreement with those conclusions. The next most important ones in the feature importance plot are the Y-H, W1-W2 and J-H colors. This comes with no surprise. A visualization of the training data through the corresponding color-color plots has revealed a high degree of separation among the different classes (see section 3.2). The importance of those colors was thus theoretically anticipated and the feature scores shown in Fig. 4.6 are fairly justified. The SDSS colors u-i, r-z and g-i continue to play a important role, but it should be noted that their ranking is not kept the same when photometric features from more surveys are added. With this

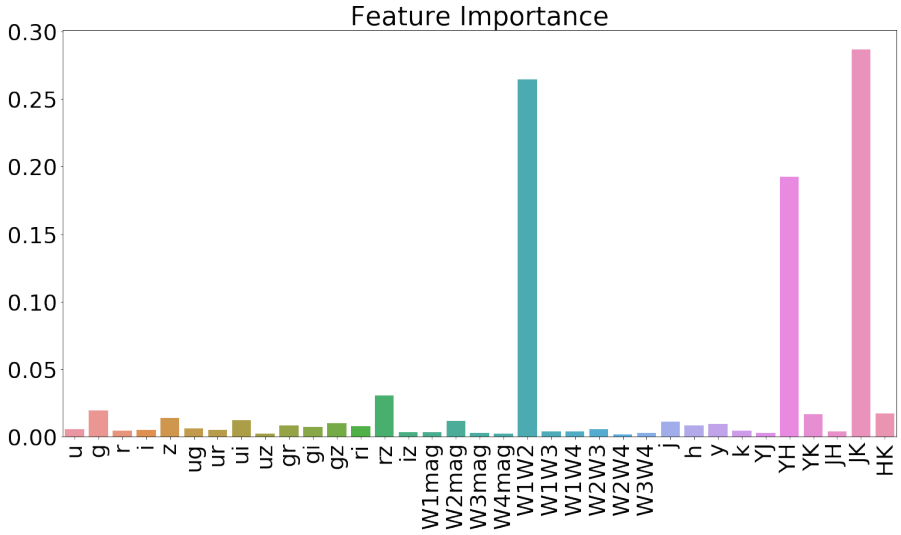


Figure 4.6: Feature importance for the classification based on all the optical, near and mid infrared magnitudes. r-z and W1-W2 colors continue to play the most important roles among the optical and mid-IR colors, respectively. However this training highlihgts the J-K color as the most important of all. Y-H also semms to be crucial for the classification.

final training the attempt to perform a purely photometric classification is concluded. One remark that should be underlined concerns the number of features used, which naturally increases the accuracy as itself increases. The cost is that the need for more computational time rises with every new feature inclusion. Dimensionality reduction can however take place, and only the highest ranked features can be selected, with a notable but not terrible consequent loss of predicting power. For example, the training with the SDSS magnitudes and all their possible permutations gives an accuracy of 91.4%. On the other hand, the selection of the 5 most significant features provides a model with an accuracy of 89.2%, which translates to an accuracy loss of 2%.

4.2 Adding Astrometric features

Table 4.8: Performance of all the astrometric models on the unseen **split test set**.

Model	Number of samples	QSOs retrieved	Accuracy
Three classes			
Purely Astrometric	113.505	50.128/51.260	92.55
Astrometry-Gaia-colors	113.505	50.764/51.260	98.84
Astrometry-AllWISE	113.505	50.939/51.260	98.96
Astrometry/Gaia-colors-AllWISE	113.505	51.035/51.260	99.78
Astrometry/SDSS-AllWISE/UKIDSS	35.624	13.932/13.980	99.45
Astrometry/Gaia-Colors/SDSS-AllWISE/UKIDSS	35.624	11.785/11.835	99.56
Two classes			
Purely Astrometric	105.084	51.127/51.271	99.67
Astrometry/Gaia-colors-AllWISE	105.084	51.229/51.271	99.94
Astrometry/Gaia-Colors/SDSS-AllWISE/UKIDSS	35.624	11.829/11.835	99.92

4.2.1 Purely astrometric

In this paragraph the potential of astrometric characteristics as effective discriminative factors for quasars is explored and presented. The features on which the model is trained on are six in total. The parallax, proper motion, their corresponding errors and signal to noise ratios. Among them, the proper motion signal to noise ratio demonstrates an outstanding precedence in the feature importance plot (Fig. 4.7) and is by far the most critical characteristic for selecting quasars based on purely astrometric measurements.

The training dataset consists of a balanced number of stars and quasars (each class contains $\sim 3.5 \cdot 10^5$ observations). Galaxies are found in a proportion 1:7 compared to the stars or quasars observations, as an immediate result of Gaia mission's focus on point sources rather than extended ones. Due to this high imbalance of the galaxy class, a binary classification between only stars and quasars is also performed, besides the multiclass one. In Fig. 4.8b the resulting predictions on the test dataset for both ML classification cases are displayed. The consequences of the imbalance in the galaxy class is imprinted there, with the number of true galaxy predictions being of the same order of magnitude as the false negative galaxies. Things start getting better when the model is trained only on stars and quasars, with the accuracy climbing from 92.5% to 99,7% on the test set. The completeness of the quasar class is exceedingly high in both classifications, namely 99.0% in the two-class and 99.7% in the 3 class training. That means that almost all the quasars were recovered and in fact, only

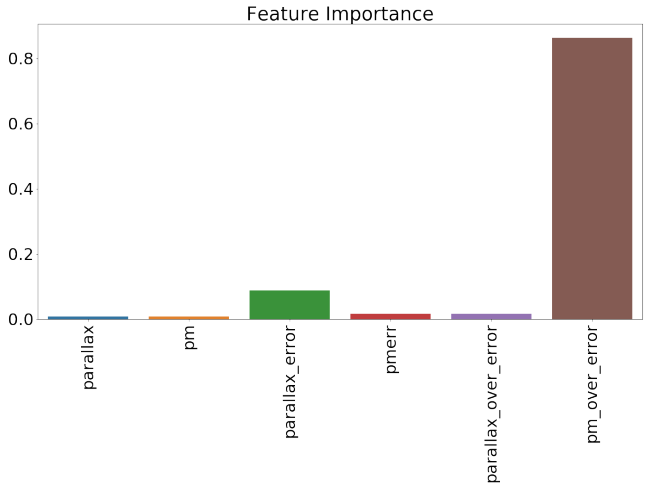
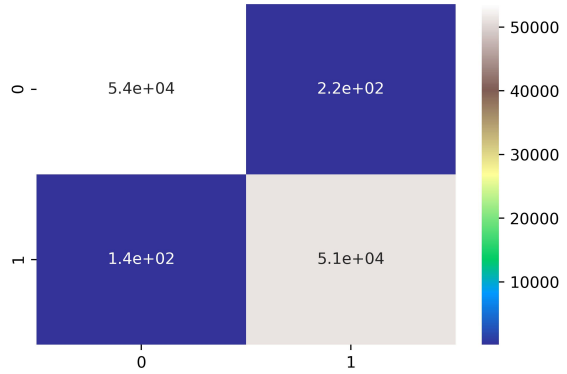
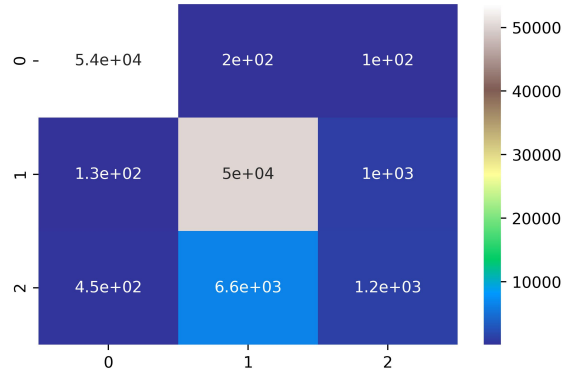


Figure 4.7: Feature importance for the purely astrometric classification. S/N_{pm} is by far the most important feature, followed by the S/N_{plx}. The S/N ratios of the astrometric features are far more important than the astrometric measurements themselves.

140 out of the total 51.000 were misclassified using the binary classification. This constitutes a big leap forward in the effort to identify and correctly estimate the quasar population, with the minimum amount of knowledge on the candidates. Sure enough, the objects can be discriminated only with the use of the noticeably low number of 4 measurements: the parallax, proper motion and their errors. No color and its respective inevitable bias is involved, and if the galaxy training set could be made more complete, the improvement of the purely astrometric model would might not be necessary at all. Comparing the two models, the two-class model appears to provide better predictions and classification reports. But its drawback is that it cannot be applied on datasets that potentially contain galaxies. If used to make predictions on a random sample of stars, quasars and galaxies, the contamination from galaxies misclassified as quasars would result in some misleading outcomes. The necessity to modify and meliorate the multiclass training brings this study to the next phase: the combination of photometry and astrometry in machine learning algorithms, which will be discussed in the following paragraph.



(a) Confusion matrix of the binary astrometric classification on the test data.



(b) Confusion matrix of multiclass astrometric classification on the test data.

Table 4.9: Purely astrometric, 3 class (multiclass) ML classification

Class	Purity	Completeness	F1score	Accuracy
Stars	0.989	0.994	0.992	
Quasars	0.881	0.978	0.927	92.5%
Galaxies	0.527	0.150	0.233	

Table 4.10: Purely astrometric, 2 class (binary) ML classification

Class	Purity	Completeness	F1score	Accuracy
Stars	0.997	0.996	0.997	
Quasars	0.996	0.997	0.996	99.65%

4.2.2 Combination of photometry and astrometry

Astrometry | Gaia Colors

Before the inclusion of any other survey's measured magnitudes, we exploit to the fullest the measurements that the Gaia EDR3 survey provides. The three optical bands, phot-g-mean-mag, phot-bp-mean-mag and phot-rp-

mean-mag and the colors b-g, b-r, g-r are included as features of the new training. The result is a modified, better version of the mutliclass purely astrometric model, with a much higher accuracy and better separation among the different objects. All the non-diagonal elements of the confusion matrix (Fig.4.9) are at least by one order of magnitude lower. Stars and quasars are very effectively distinguished when the color criteria is added to the existing astrometric ones. The most important colors are the b-g and g-r, and then comes the g magnitude in the feature importance plot (Fig. 4.10). The biggest weakness remains the mixing of the quasars recognised as galaxies and vice versa.

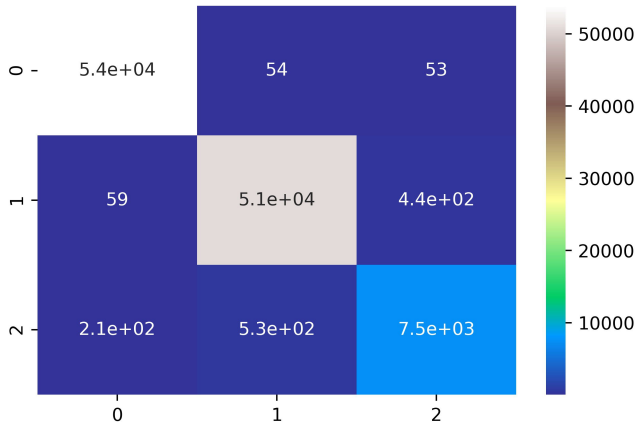


Figure 4.9: Confusion matrix for the multiclass classification based on Gaia's astrometry and photometry

Table 4.11: Astrometric and Gaia colors, 3 class (multiclass) ML classification.

Class	Purity	Completeness	F1score	Accuracy
Stars	0.995	0.998	0.997	
Quasars	0.989	0.990	0.989	98.8%
Galaxies	0.939	0.910	0.924	

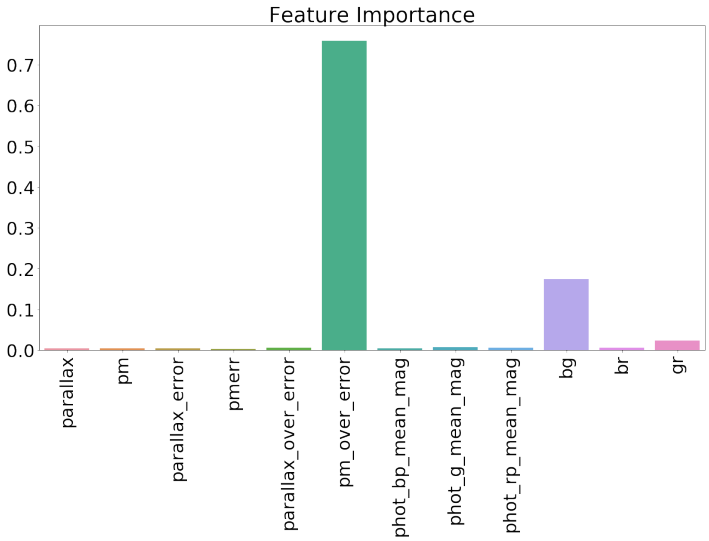


Figure 4.10: Feature importance for the classification based on Gaia's astrometry and photometry. THe proper motion signal to noise ratio remains the most crucial feature, but the inclusion of gaia's magnitudes highlighted the importance og b-g color as well. All the other features, with the excemption of g-r color, have little impact on the classification results

The advantage of having a trained ML model on only Gaia features is that in order to make predictions on unseen Gaia data there is no need for cross matches with other surveys that would result in shrinking the initial catalogue.

Astrometry | Gaia colors | AllWISE

In order to improve even more the classification accuracy the mid-IR observations from the AllWISE survey are added as features for the new training. These features include the 4 magnitude bands W1, W2, W3, W4 and their respective colors. This addition works incredibly well for the stellar class, as only 61 stars out of the 540.000 are misclassified as of other type (Fig. 4.11). Mid-infrared colors also enhance the performance of the model on the other two classes, with the galaxy class still having the least good classification reports, yet significantly better when only astrometry is involved in the training process. As for the significance among the WISE features, the results of the photometric training having taken place in the previous section are repeating themselves. W1-W2 color is the most crucial of all the features, after S/N-pm, and right before the Gaia's b-g optical color.

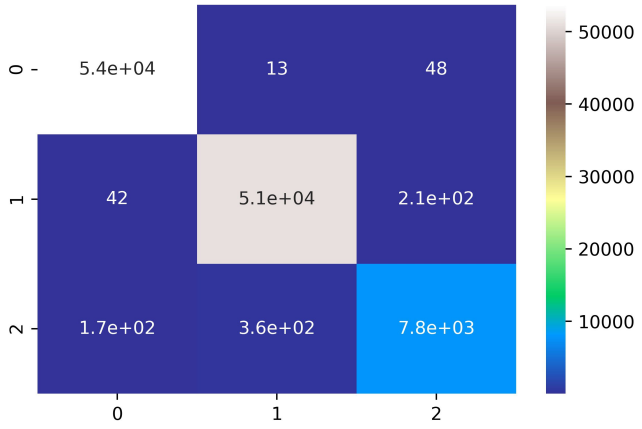


Figure 4.11: Confusion matrix for the 3 class classification based on Gaia's astrometry and photometry. M-d infrared photometry is also included in the training.

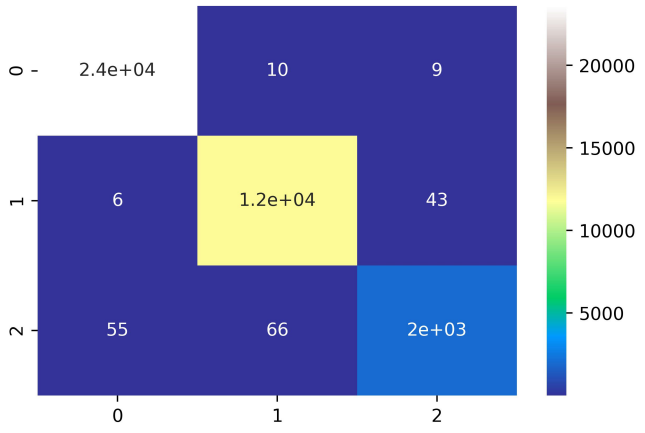
Table 4.12: Astrometric, Gaia and mid-IR colors, 3 class (multiclass) ML classification.

Class	Purity	Completeness	F1score	Accuracy
Stars	0.996	0.999	0.998	
Quasars	0.993	0.995	0.994	99.2%
Galaxies	0.968	0.937	0.952	

Astrometry | SDSS | AllWISE | UKIDSS

One last model is made, including astrometry as well as all the available magnitude bands that are used throughout this chapter. It is the most complete model, in the sense that it uses the full collection of photometric and astrometric observations. The accuracy increases to the value 99.5% and the recall and precision values are higher than in any other model tried before. Again, it should be stated that this rise comes with the cost of an increased need for successive cross-matches, when a set has to be put into test. In other words, the

Figure 4.12: Confusion matrix for the classification based on Gaia's astrometry and all the available photometry



knowledge we need to acquire on the set has to be much deeper, and range from the optical to the mid-infrared colors and the astrometric properties. Finally, the performance on the galaxy class is significantly improved, with the losses there being two orders of magnitudes lower than the total galaxy population (Fig. 4.12). Another important result that should be underlined refers to the importance of each feature, now that all of them are used and put in comparison. By far, S/N_{pm} provides the largest information gain about the type of the object and is a feature that should not be missing from any ML classification training. In general, it is shown that the signal to noise ratio of proper motion is far more important than the pm itself. Another interesting thing is that for the parallax measurement more information is gained throught the parallax-error than the S/N of the parallax. Among the photometric magnitudes and colors, W1-W2 remains the most crucial one followed by UKIDSS Y magnitude band and Y-H color index (Fig. 4.13).

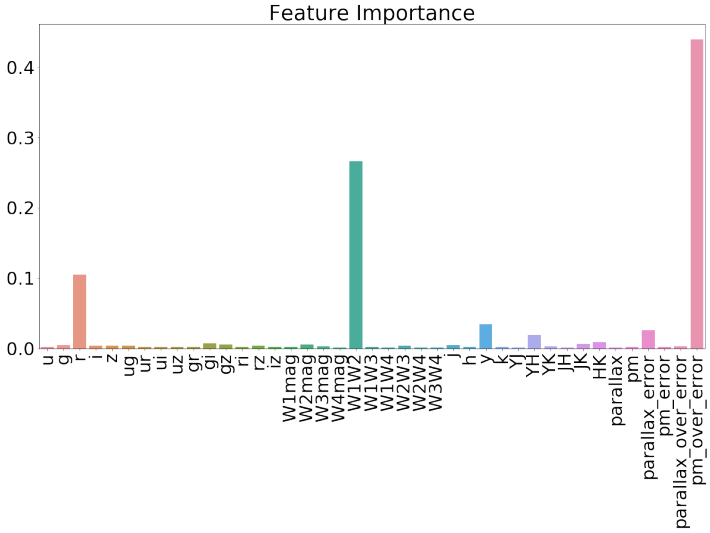


Figure 4.13: Feature importance for the classification based on Gaia's astrometry and all the available photometry, from optical to mid infrared wavelengths.

4.3 The multiclass stellar classification

Now that we have on our disposal all the tools for running the algorithms, we tried to push the classification to the limit where it can also distinguish the different stellar subclasses. To do that, we had to split our training dataset that contained the stars and gather the observations into 9 different groups. The O, OB, B, A, F, G, K, M, L, T subclasses (Fig. 4.15). The brighter classes O, OB, B are populated by 79, 22 and 206 respectively, while

each of the other classes contains some tens of thousands of stars. This comes with no surprise, since the massive and hot stars of the main sequence burn their fuel quickly and are led to their death within a timescale of 1-10 million years, compared to the coldest stars that spend $\sim$ 10 billion years in the main sequence. Hot stars are hence intrinsically more rare and underrepresented, and for that reason we decided to leave them out of the classification process. The red dwarf class (noted as RD in Fig.4.14) is created by the concatenation of the M1, M2, M3 and M4 subdivisions of the M class. In the same manner we create the brown dwarf class (BD), concatenating the L,T and the late M types. In the end, we run the classification using the XGBoost algorithm, after setting the parameters to $objective =' multiclass : softmax'$ and $numclass = 7$. Both astrometric and photometric measurements are used as features of the training, with the photometry ranging from the optical to the mid-IR bands. The results show a very effective quasar identification, where only 6 quasars are misclassified as stars.

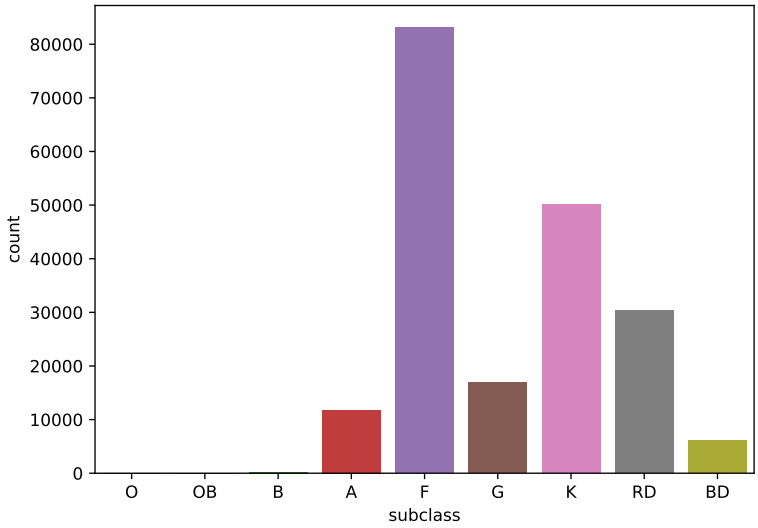


Figure 4.14: Histogram showing the stellar population of each subclass. O, OB, and B stars are underrepresented, which is expected as it is consistent with their rareness (only 1 out of approximately 3 million stars is found to be of type O).

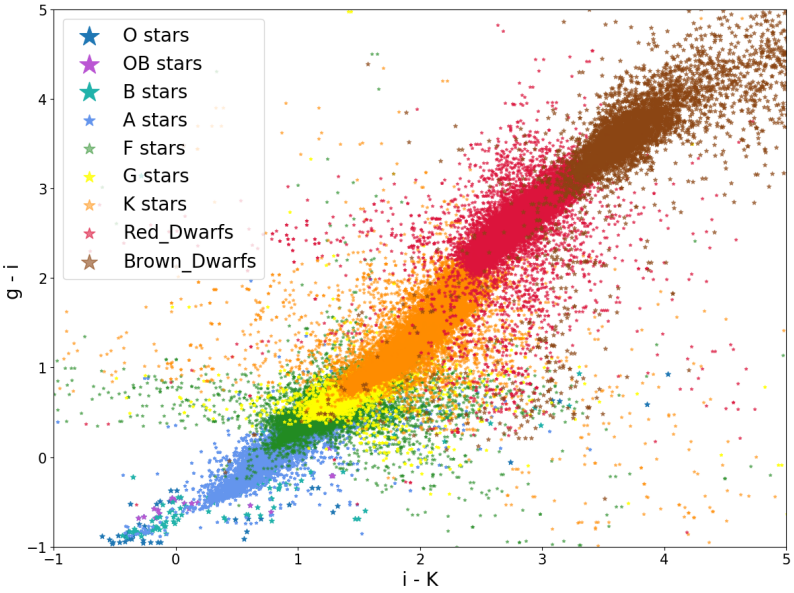


Figure 4.15: Color color plot for all the stellar classes, showing the distribution of the stars from the bluest (O, OB, B located at the lower left corner) to the reddest (red and brown dwarfs, at the upper right corner).

The overall accuracy is pretty low, 90%, but this is due to the mixing up between the F and G subclasses (see confusion matrix 4.16 and Fig. 4.17). Even in a color color plot, these two subclasses appear to be overlapping and we assume this is the reason behind the confusion.

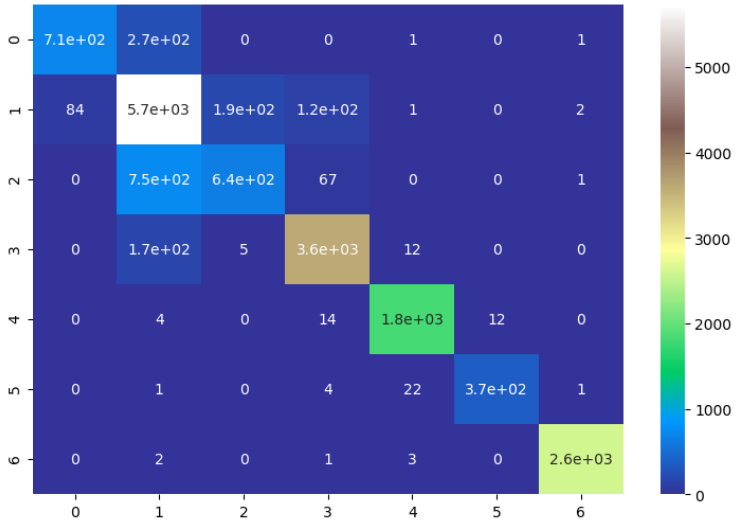


Figure 4.16: Confusion matrix of the multiclass classification. Rows and columns labeled with the numbers 0-3 correspond to stellar subclasses A to K respectively, number 4 refers to the red dwarfs, 5 to the brown dwarfs and lastly 6 refers to the quasars. Subclasses A-K are the problematic ones, seeming to be mixed up with their neighbouring subclasses. On the other hand, the model seems to well separate the red, brown dwafs and the quasars from the rest objects.

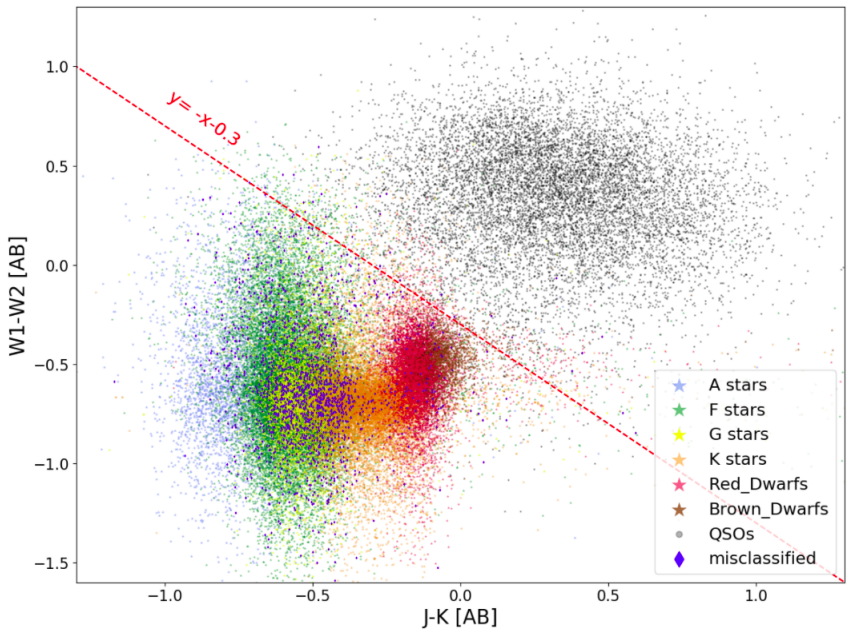


Figure 4.17: Color color plot (in AB magnitude system) where the stellar subclasses are denoted in different colors. A clear overlap between the F and G class is the cause of the model's bad perfornance in distinguishing those classes. Most of the misclassified objects fall on this part of the JK-W1W2 plot, while almost no misclassified objects are found in the quasar cluster. The red line is our estimation of separating limit between stars and quasars.

A careful look at the confusion matrix 4.18 reveals the problematic subclasses, but also a good performance of the model between the dwarf classes and the quasars. With this realization, we attempt to run a 3 class

model that classifies objects in one of the following classes: red, brown dwarf or quasar. None of the input features has changed, we are still using all the available photometry and astrometry, as before. The accuracy now climbs to 99.4%. The lost quasars are only 17 out of the 11 thousand, meaning that the completeness of the quasar class is remarkably high. The number of the false positives are equally optimistic, since no red dwarfs and only 11 brown dwarfs were predicted as quasars. The highest level of confusion is found in the separation between the early and late M type stars, a fact that is also demonstrated in Fig. 4.19. The motivation behind this final multiclass classification was an already known tendency of the quasars to be misclassified as M (especially of late type) stars. First of all, the M stars are the more abundant type of stars. In that sense, it is statistically more probable for a quasar to be mixed with an M star than with any other stellar type. Besides that, M stars are cold and less bright in the optical, such that their photometry resembles that of a highly redshifted or reddened quasar. As we will see in the following chapter, many of the targets fall into that category and were proved to be M stars, even if the model predicted them as quasars. Thus, we wanted to build a model that could focus only on those two type of objects: M dwarfs and quasars. This model does not stand on its own, in the sense that it cannot have the extended application to galaxies or every kind of star. However, we presume that it could be the final step of a ML classification pipeline, that would take as inputs all the outputs of the 3class (star, quasar, galaxy) algorithm that were predicted as quasars, and test them under the M dwarf classification. If the prediction for an object persists on the quasar label, then it should be more probable that it actually is indeed a quasar.

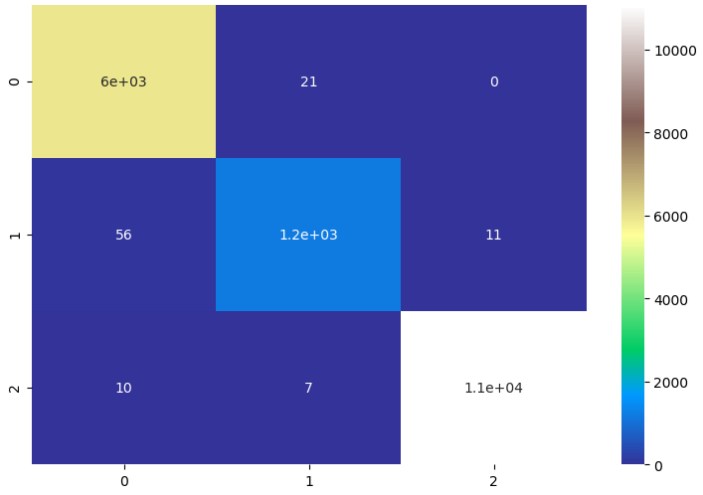


Figure 4.18: Confusion matrix for the red (0), brown dwarf (1) and quasar (2) classification.

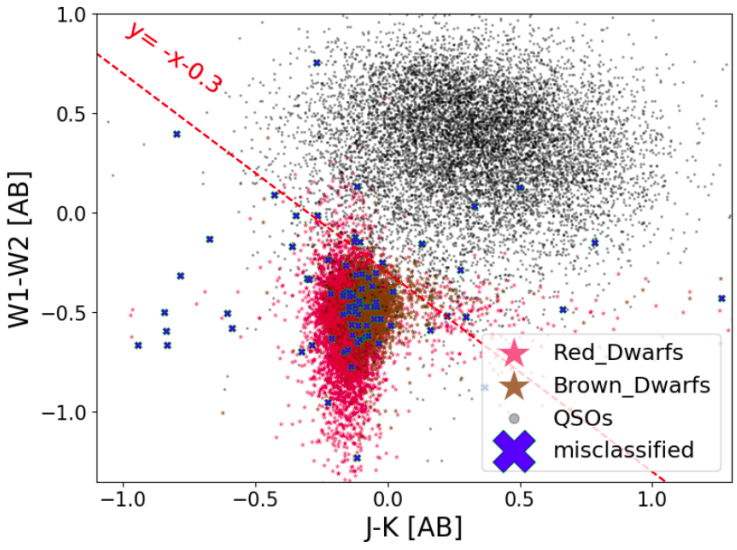


Figure 4.19: JK-W1W2 color color plot in AB magnitudes demonstrating the concentration of the misclassified objects on the overlaps of the red and brown dwarf classes. The quasars that the model failed to retrieve on the other hand are only 17.

4.4 Regression- Predicting the photometric redshift of quasars

A first approach on predicting the photometric redshift of quasars comes from a simple Linear Regression model. This model is trained on the SDSS | WISE | UKIDSS magnitudes as well as on their respective colors. The final result can be seen in Fig. 4.20 - left subfigure - where the spectroscopic redshift from SDSS DR16 is plotted versus the predicted linear regression redshift. As can be seen from the plot, the root mean square error is 0.6, which basically means that the mean predicted redshift is deviating from the true spectroscopic one by a factor of 0.6. Roughly speaking this is not a completely bad result. However, one main problem concerns the high redshift quasars.

Figure 4.20: Left: SDSS spectroscopic redshift vs. Linear Regression redshift). The training dataset contains data from SDSS, AllWISE and UKIDSS. In this plot we present the perfomance of the Linear Regression on the unseen test set. Right: SDSS spectroscopic redshift vs. XGBoost predicted redshift). The root mean square error is 0.28. In this plot we present the perfomance of the Linear Regression on the unseen test set.

It is clear that for redshifts $z > 4$ the regression behaves poorly and has a very low predicting power. One simple explanation is that quasars with high redshifts ($z > 4$) are underrepresented in the training and test set. A possible inclusion of high redshift targets in the training set is expected to provide far better results.To get better results we utilize the power of the XGBoost algorithm. As in the classification process, XGBoost Regressor provides a wide range of hyperparameters one can use to enhance the regression model. Some of these are the objective (reg:logistic, reg:squaredlogerror, reg:squarederror) , max-depth, n-estimators. The model is made more conservative (Ridge and Lasso Regression) by also tuning the min-child-weight, reg-alpha and reg-lambda values. The XGBoost Regression model constructed and presented in this chapter used the following hyperparameters:

- base-score=0.5
- booster='gbtree'

- eval-metric='rmsle'
- gamma=0
- learning-rate=0.3
- max-depth=9
- min-child-weight=1
- n-estimators=200
- reg-alpha=0, reg-lambda=1
- tree-method='exact', random-state=42

The results of the XGBoost model are presented in Fig. 4.20, right panel, where the SDSS DR16 spectroscopic redshift is plotted versus the predicted one. XGBoost regression perfomrs better than the Simple Linear Regression. The root mean square error has dropped to 0.28. One problem that remains is the predictions on high redshift targets. In the XGBoost case we dropped all the observations with redhsift higher than 5. Two are the main reasons; 1. underrepresented high redshift population and 2. untrustworthy SDSS high redshift identifications (Appendix C). An improvement in these higher photometric redshift predictions is extremely important for making new interesting quasar catalogues that could provide new insights on how galaxies evolve, how SMBH form and how accretion discs are formed in the early Universe.

Part III

Results

Overview

In the previous part we introduced all the machine learning models we built and presented a preliminary evaluation for them, through the metrics of machine learning: the accuracy, the completeness and purity. Part III is focused on the evaluation of the models through 3 different approaches:

The first one, is the cross-validation of ML predictions with already known objects. Objects spectroscopically observed by our supervisor, Johan Fynbo, form a long list of 575 objects that serve as the validation set. It mainly consists out outliers, BALs and reddened quasars, or stars with photometry similar to a quasar, and is thus a strong evaluation metric on how well each model can retrieve peculiar quasars.

The second validation comes from the comparison between the models and known empirical relations that are widely used in the literature as quasar selection criteria. Through this comparison we quantify the improvement in the selection techniques that is induced by the use of our ML models, instead of the color based methods.

The third evaluation metric is direct observation of quasar candidates that are 'born' through the predictions of 2 ML models we constructed. The first model is a purely photometric one, based on which we prepared an observing run at the NOT Telescope during December 2021. The second model, on which May's observing run relied, is a combination of the photometric + astrometric training. The details of the target selection and spectral reduction are presented in chapter 6 and in the Appendix.

Chapter 5

Validation of the ML models

5.1 An unusual quasar catalogue

The set on which the model's accuracy is calculated is another factor that requires to be given some thought. In general, if the accuracy is calculated on the same data the model was trained on, the result is expected to be a value close to 100%. The reason behind this is pretty obvious and relies on the fact that the model has already "seen" these data during the training phase. The most commonly used and trustworthy accuracy is performed on unseen data, namely the test set. But this accuracy is not an absolute metric on its own. It is highly dependent on the type of objects one is making predictions on. A test dataset full of outliers is for instance expected to provide a much much lower accuracy. In Figure 5.1 we present the objects of such a dataset.

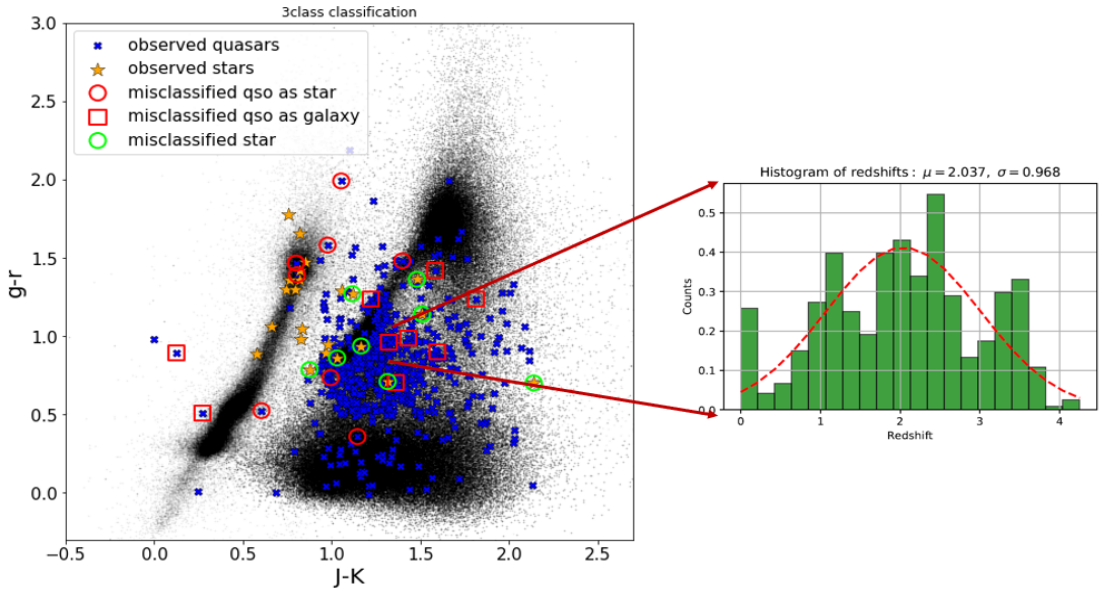


Figure 5.1: gr vs. JK color-color plot that shows the stellar, galaxy and quasar clusters at the background (black dots). The objects of mastercatalogue are overplotted to show their relative position with respect to the three forementioned classes. In green circles are shown the stars that were falsely recognised as quasars by the ML model, and in red the quasars that were lost (red circles: recognised falsely as stars, red squares: falsely recognised as galaxies). Histogram of the redshifts for all the spectroscopically validated objects found in the master catalogue is shown in the right plot. Mean redshift value is around z=2.

In order to gain the highest trust possible on our models, it is crucial to test them on such extreme cases. Based on previous works by ([Geier, Heintz, Fynbo et al., 2019]) an outliers catalogue is made, containing spectroscopically classified BALs and dust reddened quasars. The catalogue also contains stars with photometry that resembles that of a quasar. The objects of the catalogue are plotted in a color-color plot in Fig. 5.1

Photometric Models

Table 5.1: Performance of all the photometric models on the **Master Catalogue** by [Geier, Heintz, Fynbo et al., 2019], 575 in total observation of which 550 quasars and 25 stars. Out of the 550 quasars 184 are Broad absorption line quasars, 397 have r-z>0.5 and so can be labeled as red quasars and finally, 119 are high redshift quasars with z>3

MODEL	TRUE QSO PREDICTIONS	QSO LOST AS STARS	QSO LOST AS GALAXIES	ACCURACY (%)
Three Classes				
SDSS (unbalanced)	172	134	244	33.0
SDSS (balanced)	238	162	150	45.0
SDSS+WISE (unbalanced)	521	9	20	93.0
SDSS+WISE (balanced)	530	10	13	94.0
SDSS+WISE+UKIDSS (unbalanced)	522	7	18	94.0
SDSS+WISE+UKIDSS (balanced)	532	9	9	**95.0**
Two Classes				
SDSS+WISE+UKIDSS	537	12	-	**96.3**

In Table 5.1 we present the performance of all the purely photometric models described in Part II. The incompetency of the purely optical machine learning model on outliers is notable. The accuracy dropped from the 91% on the test set to 33% on these outliers. This also justifies the inclusion of infrared colors for retrieving reddened and BAL quasars. The best accuracy for this catalogue is achieved with the use of the all-color, balanced ML model. Moreover, the 2class model performs even better but it should be noted that this validation catalogue does not contain any galaxies.

Astrometric Models

Table 5.2: Performance of all the astrometric models on **Master Catalogue**. Total number of samples are 575 where 550 are quasars and 25 are stars.

MODEL	QSOs RETRIEVED	ACCURACY
Three classes		
Purely Astrometric	543	97.93
Astrometry/Gaia-colors	532	94.88
Astrometry/AllWISE	537	94.97
Astrometry/Gaia-colors/AllWISE	538	95.86
Astrometry/SDSS/AllWISE-UKIDSS	536	96.58
Astrometry/GaiaColors-SDSS/AllWISE/UKIDSS	**537**	**97.02**
Two classes		
Purely Astrometric	545	98.46
Astrometry/Gaia-colors/AllWISE	545	96.75
Astrometry/Gaia-colors/SDSS/AllWISE/UKIDSS	542	98.15

In Table 5.2 we present the performance of all the models that also include astrometric measurements as features. The first aspect that should be noted is that the purely astrometric model performs better, by around 1% than the combined all photometric+astrometric model. In particular, when we include to the purely astrometric model only the Gaia-colors the accuracy drops by about 3%. This is a strong verification that colors (especially optical ones) include biases towards dust-reddened, redshifted and in general unusual quasars. At first, a machine learning model, through the ability to make complex divisions in the multi-color space, can make generalizations and include a larger number of outliers than simple color empirical relations do. However, a model trained on biased data, towards a specific subgroup, will copy some of these biases. This is evident by our results and an indication that a purely astrometric classification, if it was free of errors, it would be the best choice for making quasar catalogues of high completeness and high purity. Nevertheless, it should be stated again that our data do come with errors. If the master catalogue contained more stars and galaxies, our purely astrometric classification would not perform better than the one that includes Gaia-colors as features. With this catalogue we only test a specific capability. How the different models perform on retrieving outlier quasars and not how robust they are. With this in mind, although the purely astrometric model is able to select most of the outliers it would perform bad in a case where the catalogue is contaminated by a huge number of stars and galaxies (4.2.1). Finally, the all-included model (all photometric + astrometric features) is the best in terms of both including outliers and being robust.

5.2 Comparison with empirical criteria

Many criteria on how to discriminate stars from quasars based on their position in color-color plots have been suggested and can be found in literature. As was mentioned in Part II, stars and quasars can be empirically separated using only their SDSS optical and UKIDSS near-infrared colors [Wu & Zhendong, 2010]. In the paper

they introduce relations such as:

$$(Y - K) > 0.46 \cdot (g - z) + 0.53 \tag{5.1}$$

which is the criterion to better separate stars from quasars with redshift z<4. For higher redshift quasars, they propose the criterion:

$$(J - K) > 0.45 \cdot (i - Y) + 0.64 \tag{5.2}$$

To test the predicting power of the machine learning models that each use different magnitude bands as features, we compare their accuracy with the accuracy of the above empirical criteria, on the same validation sets. Firstly, the master catalogue of red and BAL quasars is used as a validation set. The inspection of the objects' characteristics that are contained in the master catalogue reveals that the quasar redshift distribution ranges from 0.09 to 4.25, with a peak around z=2 . (Fig.5.1). It is therefore expected that eq. 5.1, is appropriate for this redshift range and can achieve a higher accuracy, retrieving more quasars than eq. 5.2 can. Indeed, applying 5.1 criterion to the master catalogue, we retrieve 531 quasars out of the 550 and 15 stars out of the total 25. In accuracy terms, this corresponds to 94,95% success. On the other hand, the application of eq. 5.2 to the master catalogue leads to a successful recovery of 470 quasars out of the 550, and 19 stars out of the 25. The accuracy in this cut-off is 85%, significantly lower than in the former case, as it was anticipated. It should be noted though that the combination of the two above relations provides a very strong separation criterion, being able to recover 540 quasars and reaching the accuracy of 96.5% when applied on the same validation set. Comparing these accuracy values with the ones in table 5.2, we conclude that the empirical criterion Y-K gives the same results as the all-color, 3 class model (95% accuracy). At first glance, the role of machine learning seems to be of equal efficiency with the empirical relation in this example, and no significant improvement emerges from the ML field. One should have in mind though that the machine learning models are trained to also distinguish galaxies, a possibility that from the empirical relations is elusive. A fair comparison would concern a binary machine learning model trained only on quasars and stars. The performance of this kind of model on the master catalogue can be seen in the last row of table 5.2. From there it is concluded that even though the accuracies differ only on the first decimal digit, the combined empirical criteria can retrieve 3 extra missing quasars than the 537 the 2 class model can.

Due to the fact that the master catalogue that is used for the comparison between the ML models and the empirical relations is highly unbalanced (550 qsos for 25 stars), we provide the performance of the the two methods on the 15% split test set we originally use to evaluate our models. Both the Y-K, J-K and the combined methods provide an accuracy of around 97.9%, while the ML 3 class model reach an accuracy of 98.7% and the 2 class an accuracy of 99.7%. So, in the most general test set the ML performs nearly 2% better than the empirical methods previously used.

Finally, the all-color astrometric model surpasses the best empirical criterion's accuracy by $0.5 - 2\%$.
The analysis performed in this section leads to the following overall remarks:

- The optical colors from the SDSS DR16 survey are not sufficient to recover outlying quasars, while the contamination level from stars and galaxies is very high. The first statement means that the completeness (recall) of the quasar class is low, and the latter statement translates to an equally low purity (precision).
- The inclusion of infrared photometric observations (WISE/UKIDSS) results in a tremendous rise of the accuracy from 33% to 95% on the master catalogue validation set, and from 89% to 98.5% on the split test set. On a purely photometric machine learning classification, we deduce that the infrared photometry

is essential in order to get precise predictions.

- The purely photometric machine learning training is very effective on the binary classification problem. In the case where a test set consists only of stars or quasars (for example if extended sources are cut off), the 2 class model trained on both optical and infrared bands can separate the different objects with remarkable accuracy.

- Previous selection techniques involve empirical relations using the SDSS and UKIDSS colors. Their accuracy on a dataset of outliers is similar to the ML models' accuracy, however the empirical criteria perform increasingly worse when the validation catalogue contains more and more stars. They are also incapable of being applied to datasets that contain galaxies, a drawback that is lifted with the use of the 3 class ML model.

Chapter 6

Observations

The observations we conducted are separated in 3 parts. The first observing run was performed in August 2021, during the IDA summer school. The data were taken remotely from the NOT telesope in La Palma, Spain. At the time of the observing block's preparation, this thesis had not started yet, and as a result, the targets selected then were not the predicted outcome of a machine learning process. They were only targets that our supervisor, Johan P. U. Fynbo suggested as interesting quasar candidates that were at the same time bright enough (G < 20) to be observed. However, we include the results here, since the data reduction was a product of our work, and this list can serve as an extra validation set of unusual quasars. It should be noted that all of their g-r or J-K colors lied outside the typical range for quasars, making them a strong evaluation metric for our models.

The second observing run was scheduled for the end of December. This time, the targets were selected from a long list of sources predicted to be quasars by a photometric machine learning classification. All the available optical, near and mid-infrared colors, for those unseen objects queried from GAIA EDR3 survey, were used as features for the training. Once again, outlying candidates were selected and observed in order to confirm or reject the prediction. Driven by the results of this attempt, we afterwards decided to include astrometric features to our ML training. Surprisingly, a lot of the predicted quasars exhibited a unexpectedly high S/N_{pm} ratio, many of them with even higher than 2, which is an empirical cut-off to eliminate the stellar contamination in the quasar population. For some of such objects, their spectral data were already existing in the SDSS catalogue, confirming that our revisited model could indeed find quasars with both photometry and astrometry that resembles that of a star. With the astrometry's inclusion, we could therefore infer that we pushed further the boundaries of a complete quasar selection, overcoming the previous ML algorithms or empirical relations.

This brought us to the third and final observing run, taking place at the end of May 2022. To test the model we constructed, we made a proposal for IDA's young researcher's program, which was approved and allowed us to visit the Nordic Optical Telescope on top of the Roques de los Muchachos mountain in La Palma, and conduct the observations ourselves. The targets for this observing run where predicted quasars that are extreme astrometric, and/or photometric outliers. The targets' description and overall results are discussed below.

6.1 Selecting candidates

Synopsis

Before we present analytically the selection processes for every observing run individually, we present color plots of all the observed targets.

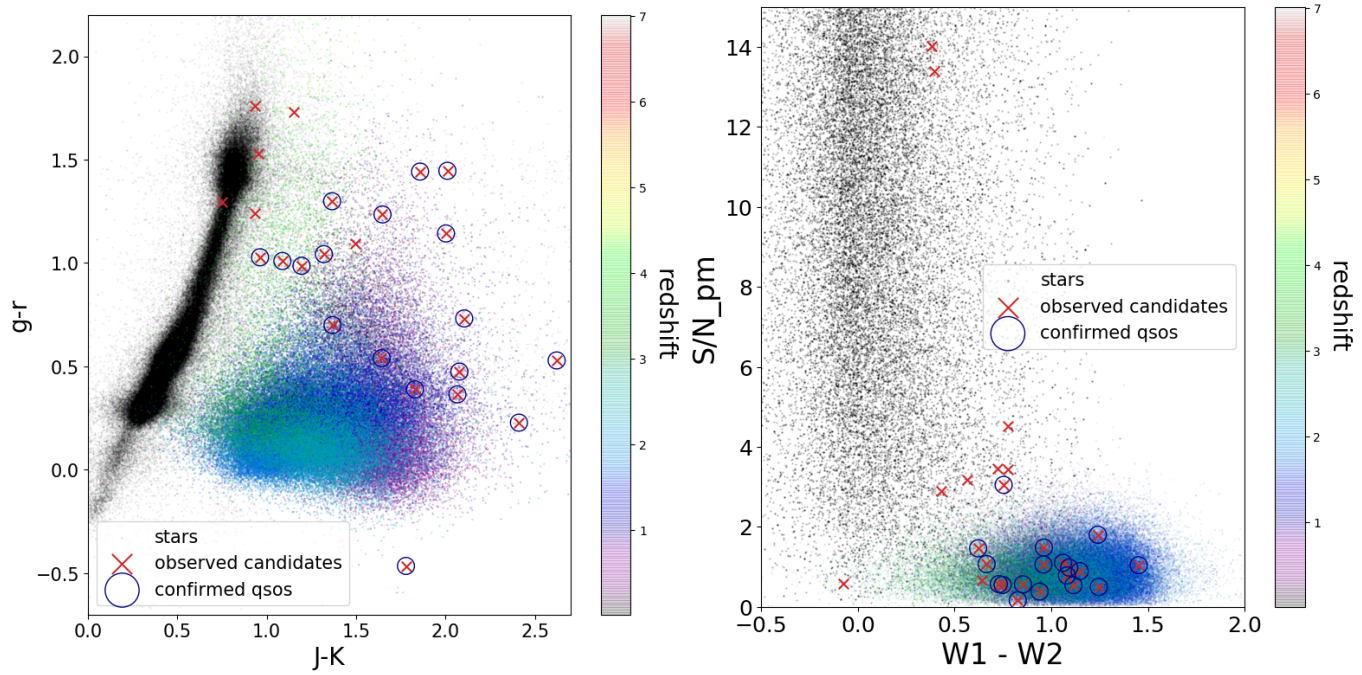


Figure 6.1: Left: gr - JK color color plot where the observing targets are marked, to emphasize the fact that they are all photometric outliers. Some of them clearly fall in the stellar regime, many are very red in the mid-IR (J-K>2), while most of them fall on the vague space between the quasar/stellar or quasar/galaxy clusters. The training sets of stars and quasars are showed in the background, forming the corresponding separated groups. The galaxy class is not illustrated, but lies right above the quasars, where the upper right targets are shown. Right: S/N_{pm} - W1W2 plot, where the astrometric outliers observed are shown near the stellar cluster.

Subplot 6.1(a) shows that none of the targets falls in the main area of the quasar clusters. They are either extremely red in the near-infrared colors, either fall onto the high redshift regime, or they are right on the M subclass tip of the stellar cluster. Astrometrically and in mid-IR colors, the targets are more clustered on the main quasar locus (subfigure 6.1(b)). However, 7 of them are intentionally selected to have S/N_{pm} ratio much higher than the typical value for a quasar, in order to verify or reject the astrometric model's predictions. These constraints and limitations are very important to be explored, since a combined photometric/astrometric model has not been built before and its weaknesses are yet to be discovered. A data overview of the selected candidates and their observations are gathered in Tables 6.6 and 6.4.

August 2021 - IDA Summer School

The list of targets for the summer school consisted of objects that our supervisor, Johan Fynbo, suggested as interesting ones (8 in total). Out of them, 3 are very red in the $g-r$ and $J-K$ at the same time, falling in the regime of the low redshift quasars (purple part of the colorbar, $z < 1$). Basically, this is the area where the galaxy cluster, if it was shown, would fall. They are the targets J022742.93-173121.5, J234530.36-135743.3 and J000404.29-135905.43. Another, J234859.52-193324.89, is extremely blue in the $g-r$ (with negative gr color

value), but red in the $J-K$ color. Finally, 3 of the targets, namely J003634.80-140924.9, J010339.41-132238.91 and J012813.63-120319.79 have both $g-r$ and $J-K$ values close to unity, falling right in the high redshift quasars zone (green part of the color bar, $3 < z < 4$) (Table 6.4). Apart from the target J010013.02+280225.8, which has a magnitude in the G-band 25.254 and a very high redshift of $z = 5.8$, all the other candidates are pretty bright, with magnitude no higher than 20.8. At the same time, none of the summer school's targets is astrometric outliers, since their signal to noise ratios for the proper motion is well below 2 (Table 6.6). Lastly, they are all well set within the traditional $W1-W2 > 0.8$ quasar limit for the mid-infrared colors and no outlier in this color was chosen.

Table 6.1: Table of quasar candidates, targets of August 2021

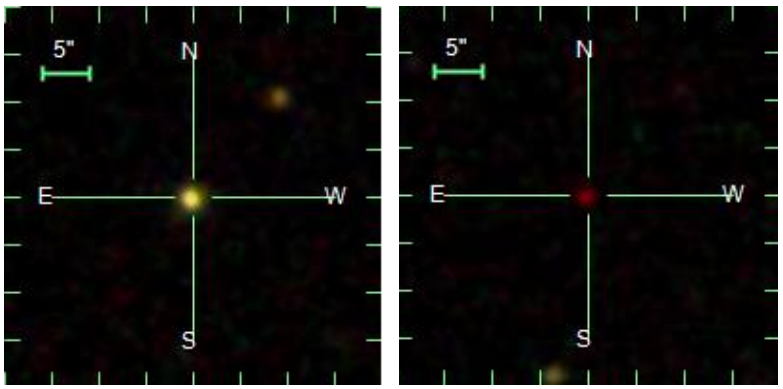
Target	RA	DEC	parallax	S/N$_{par}$	PM	S/N$_{PM}$
J022742.93-173121.5	36.9288333	-17.52263889	0.045	0.299	0.195	1.086
J010013.02+280225.8	15.0542083	28.0405	-	-	-	-
J234530.36-135743.3	356.37651	-13.962034	-0.054	-0.150	0.486	0.785
J000404.29-135905.43	1.017115	-13.98715	0.446	1.278	0.39	0.891
J234859.52-193324.89	357.247816	-19.556916	-0.246	-2.276	0.082	0.496
J003634.80-140924.9	9.145232	-14.156932	-0.159	-0.692	0.673	1.801
J010339.41-132238.91	15.913797	-13.379394	0.226	1.911	0.203	0.979
J012813.63-120319.79	22.056879	-12.057607	-0.287	-1.145	0.294	1.047

Table 6.2: Quasar candidates for observations on August 2021 and their coordinates. Astrometric criteria for selecting those targets imposed constraints on the parallax and proper motion.

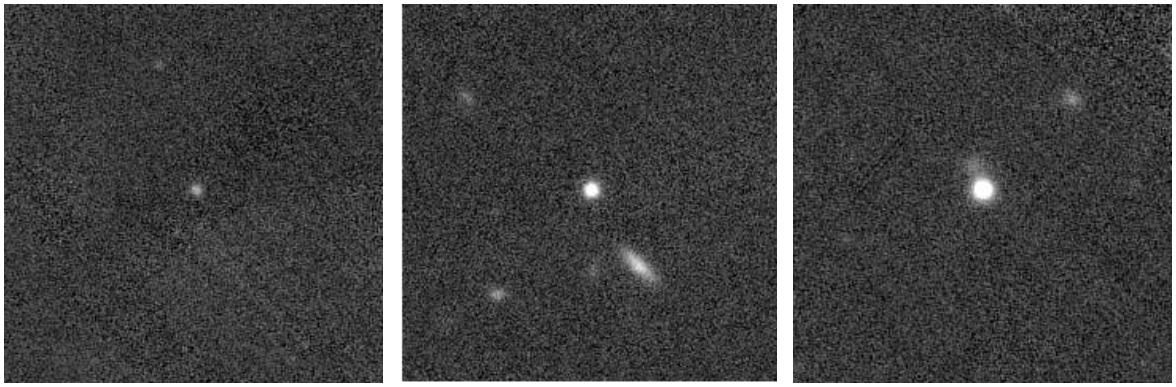
Table 6.3: Table of quasar candidates continued

Target	z$_{this\ work}$	AB	G$_{SDSS}$	g-r	J-K	W1-W2	EXPTIME
J022742.93-173121.5	2.31	0.56	19.247	1.235	1.648	0.96	2x600
J010013.02+280225.8	5.8	0.2	25.254	0.389	1.83	0.823	2x600
J234530.36-135743.3	2.325	0.82	20.8186	1.442	1.858	1.081	2x500
J000404.29-135905.43	1.44	1.23	20.67	1.298	1.366	1.147	2x500
J234859.52-193324.89	0.44	0.0	17.4875	-0.4648	1.779	1.146	2x500
J003634.80-140924.9	1.54	2.06	19.860	0.9866	1.195	1.241	2x500
J010339.41-132238.91	1.75	0 / 1.64	18.8779	1.0278	0.963	1.093	2x500
J012813.63-120319.79	1.447	0.53 / 1.32	20.0989	1.01	1.088	1.452	1x500, 1x300

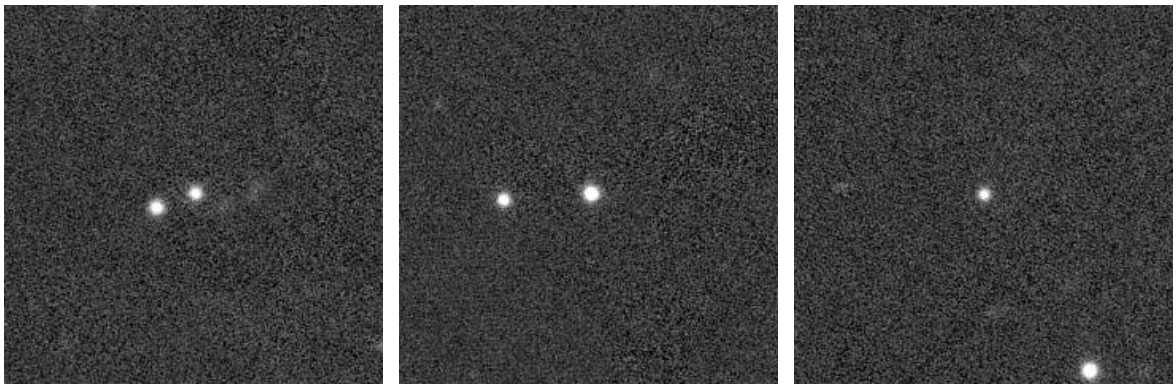
Table 6.4: Each target's redshift is determined by spectral analysis in this current work. No previous redshift data were available. Colours in the optical were taken from SDSS survey, while the J, H, K and the WISE W1, W2, W3, W4 colours were provided by the UKIDSS and ALLWISE catalogues repsectively.



(a) $J022742.93 - 173121.5$ (b) $J010013.02 + 280225.8$



(c) $j234530.36 - 135743.3$ (d) $J000404.29 - 135905.43$ (e) $J234859.52 - 193324.89$



(f) $J003634.80 - 140924.9$ (g) $J010339.41 - 132238.91$ (h) $J012813.63 - 120319.79$

Figure 6.2: Targets of August's summer school, as seen in Object Explorer of the SDSS (http://cas.sdss.org/dr16/en/tools/chart/navi.aspx) or the Pan-STARRS Image Access (https://ps1images.stsci.edu/cgi-bin/ps1cutouts?). From the latter we show the targets as they look in the g band.

The spectral analysis revealed that all these observed targets were indeed quasars, with relatively high redshift (mostly above 1.4, with 2 targets having redshift higher than 2.3 and even one target with redshift 5.8). Their estimated redshift according to their position in the color-color plot and the redshift colorbar was not compatible with their calculated redshift. Again though, this is just a statistical thing, and targets that for example fall in the high redshift area of the colorbar are not necessarily but only statistically probable to be highly redshifted quasars.

December 2021

December's targets were chosen from a long list of predicted quasars by the 3 class, all color, purely photometric ML model. In the $J-K$ color, almost all the targets are very red ($J-K>2$), while their $g-r$ color varies over the whole range (from 0.228 to 1.758). As seen in subfigure 6.1(a), this is the color range where quasars with mostly $z<2$ are found. Indeed, the spectral analysis of these targets showed that all of them, besides J015455.79+203623.8 (that is found in redshift $z=2.291$), have redshifts $0.27<z<1.95$. Three of them, namely J013121.61+444513.1, J013232.61+444123.0 and J025111.38+321221.5 were extreme astrometric outliers, with $S/N_{pm}>13$ and in one case, even 34.364. All of them also have $W1-W2<0.8$ and lie very close to the M dwarfs tip of the stellar cluster. Such characteristics should ring a bell that the predictions are probably wrong but we were curious to find out the source of the model's confusion. We already knew that this model's purity on the quasar class is 98.6%, so it was expected that around 1.4% of the quasar predictions would be wrong. We also wanted to determine what kind of stars - if there was a systematic pattern - were mistakenly taken for quasars. Indeed, it turned out that these three candidates were stars, and more specifically of type M, as expected.

Table 6.5: Table of quasar candidates, targets of December 2021

Target	RA	DEC	parallax	S/N$_{par}$	PM	S/N$_{PM}$
J012534.07+351344.0	21.39196415	35.22888345	0.360	0.639	0.445	0.565
J012848.44+341704.9	22.20188617	34.28472194	0.07	0.196	0.237	0.546
J013121.61+444513.1	22.84001362	44.75365457	1.473	4.433	5.399	13.396
J013232.61+444123.0	23.13584069	44.68975148	3.407	7.507	20.24	34.364
J015455.79+203623.8	28.73247668	20.60662766	-0.098	-0.575	0.139	0.547
J015748.65+284752.6	29.4527354	28.79796798	-0.047	-0.446	0.182	1.099
J025004.61+324039.7	42.51924199	32.67771608	-0.333	-1.257	0.551	1.487
J025111.38+321221.5	42.7974727	32.20596862	0.597	1.654	8.08	14.927
J025832.77+354254.6	44.63653379	35.71521151	-1.403	-0.922	2.889	1.068
J130703.91+251415.5	196.766341985744	25.2376388883	-0.8214	-0.8193	0.295	0.16

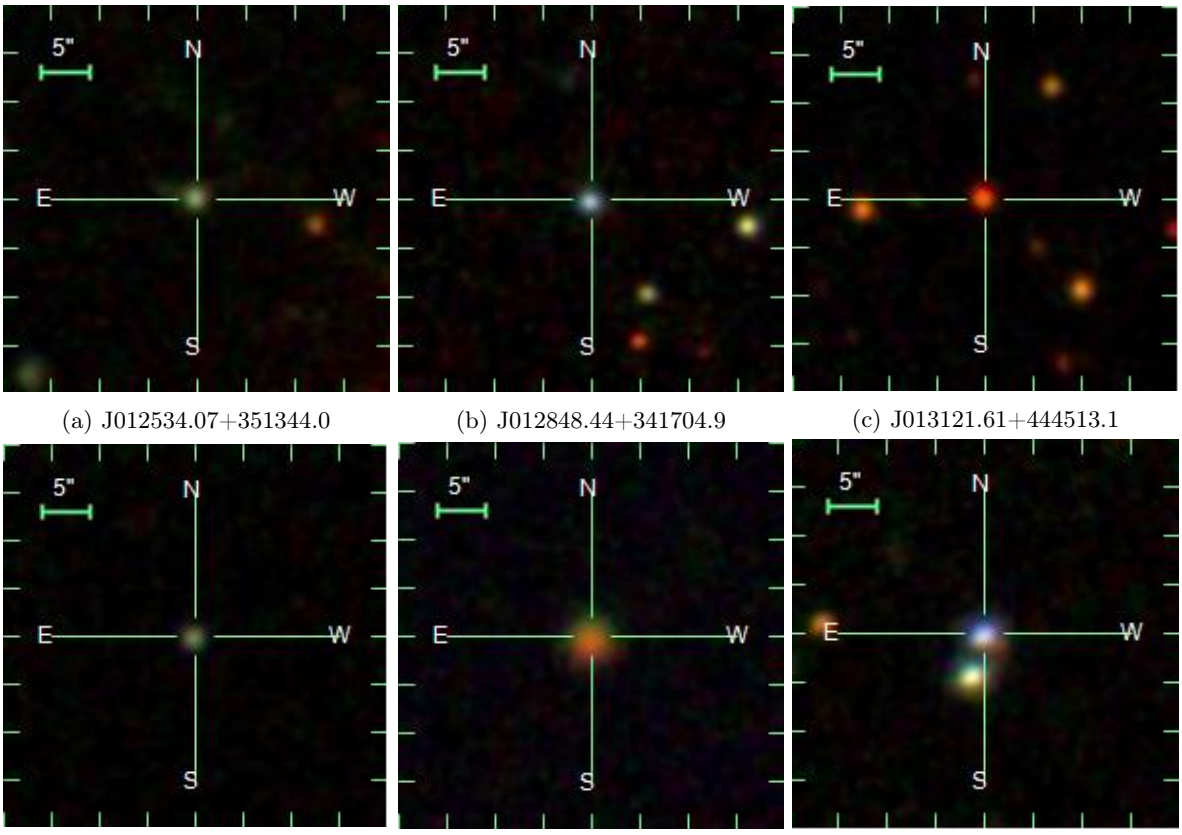
Table 6.6: Quasar candidates for observations on December 2021 and their galactic coordinates. Astrometric criteria for selecting those targets imposed constraints on the parallax and proper motion.

Table 6.7: Table of quasar candidates continued

Target	$z_{this\ work}$	AB	G_{SDSS}	g-r	J-K	W1-W2	EXPTIME
J012534.07+351344.0	0.312	2.74	20.257	0.732	2.105	0.854	2x900
J012848.44+341704.9	0.346	2.19	19.165	0.228	2.411	1.115	2x900
J013121.61+444513.1	-	-	21.389	1.53	0.956	0.397	2x900
J013232.61+444123.0	-	-	21.958	1.73	1.151	0.435	2x900
J015455.79+203623.8	2.291	1.78	19.607	1.143	2.003	0.749	2x900
J015748.65+284752.6	0.444	1.64	18.034	0.475	2.078	1.061	2x900
J025004.61+324039.7	0.3685	2.47	18.949	0.365	2.067	0.962	2x900
J025111.38+321221.5	-	-	21.303	1.758	0.934	0.384	2x900
J025832.77+354254.6	0.26	4.25	20.144	1.446	2.011	0.666	2x900
J130703.91+251415.5	0.273	2.33	19.936	0.529	2.622	0.827	2x900

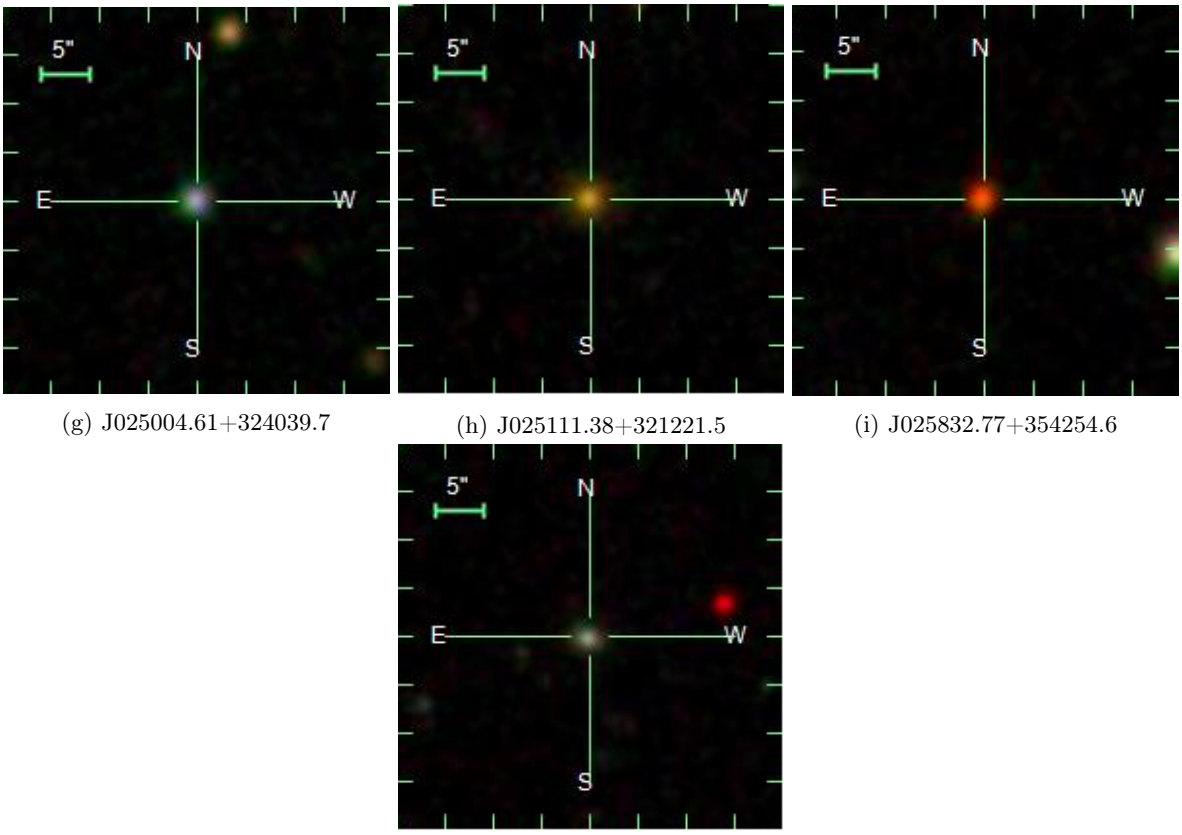
Table 6.8: Each target's redshift is determined by spectral analysis in this current work. No previous redshift data were available. The targets for whom the redshift and the A_B extinction is missing are found to be stars. Magnitudes were taken from SDSS, the UKIDSS and the ALLWISE catalogues.

Figure 6.7 shows how the targets of December's observing run look through the SDSS Object Explorer.



(a) J012534.07+351344.0 (b) J012848.44+341704.9 (c) J013121.61+444513.1

(d) J013232.61+444123.0 (e) J015455.79+203623.8 (f) J015748.65+284752.6



(g) J025004.61+324039.7 (h) J025111.38+321221.5 (i) J025832.77+354254.6

(j) J130703.91+251415.5

Figure 6.3: Targets of December's observing run, as seen in Object Explorer of the SDSS.

This observing run revealed the purely photometric model's successes and failures. On one hand, it could predict not only usual quasars but also objects that are very red in the infrared and lie outside the main quasar locus. However, such objects would have also been selected via the traditional selection techniques and more specifically, the K-excess method. In addition, the output quasar candidate list it produces is not completely pure, and this fact was also imprinted on the observations' results. Three of the proposed quasars were actually M stars, probably in a binary with a white dwarf. The spectrum fitting was made possible with the use of Pyhammer, a tool that contains templates of all the stellar subtypes and finds the best match for a given spectrum. In our cases, the best fit assumed a binary comprised of the two companions: an M type primary stellar type, and a DA white dwarf. The Pyhammer fitting for all the three targets is shown in Fig.6.6. For these M stars we detect the characteristic TiO absorption lines in their spectra, at 6651Å, 7053Å, 7666Å, 8206Å, 8432Å. For the later type M stars in subfigures 6.6 the most prominent atomic feature of Hα line at 6564Å is stronger. Pyhammer tool can only be used on stars. For the extraction of the quasar spectra we use a different method, which is discussed in a following section.

It seems that the model's confusion is due to the W1-W2 color which was in the typical range for stars. For that we cannot do much to improve the performance of the model, besides applying an artificial cut-off on the predicted outcomes. However, when we cross-matched those observations with GaiaEDR3, we realised that their S/N_{pm} values were extremely high. We then thought that we could eliminate the bad predictions by using the astrometric characteristics as features of a new model. This was the motivation behind the idea of constructing an astrometric/photometric model, and the reason why we had a third observing run on May.

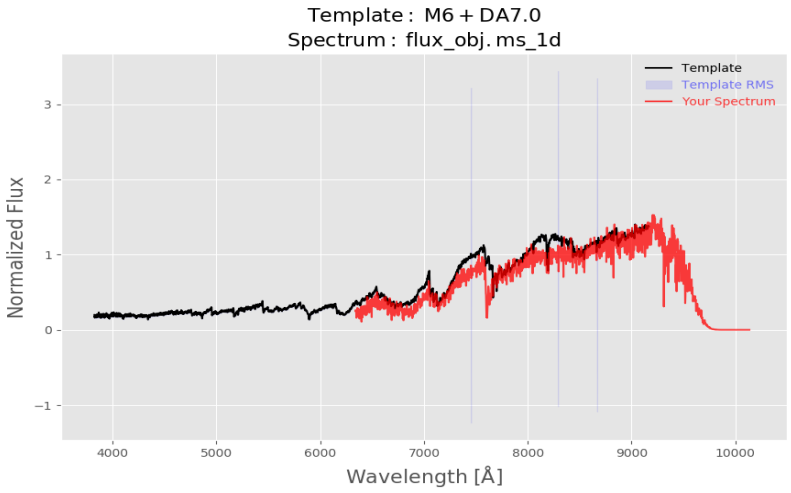


Figure 6.4: Pyhammer template fit on the extracted spectrum of target J013121.61+444513.1. Best fit occurs for the case of a binary, consisting of a late type M6 star and a DA7 white dwarf.

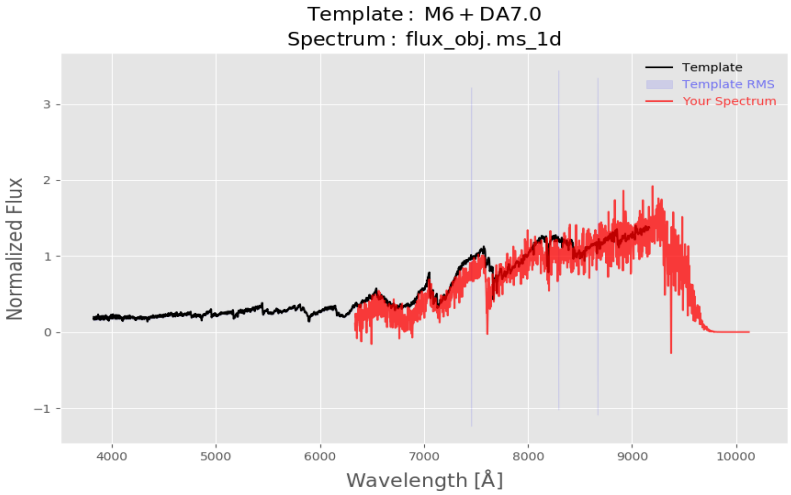


Figure 6.5: Pyhammer template fit on the spectrum of target J013232.61+444123.0. Best fit occurs for the case of a binary, consisting of a late type M6 star and a DA7 white dwarf. The fitting appears to be good, with the excemption of noise excess for wavelengths longer than 8000Å .

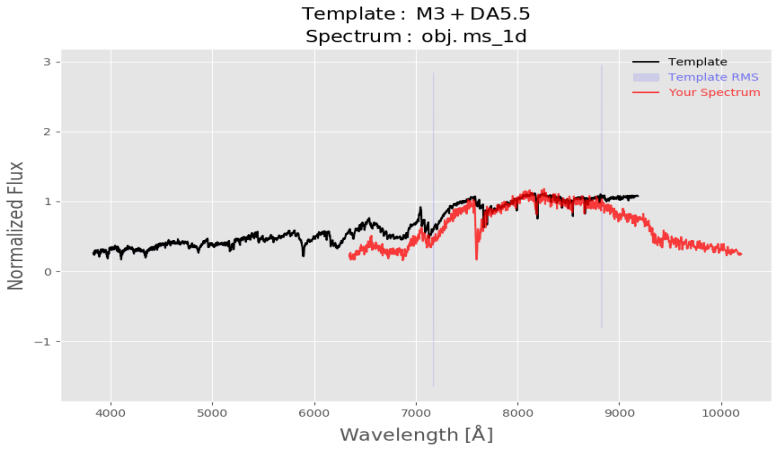


Figure 6.6: Pyhammer template fit on the extracted spectrum of target J025111.38+321221.5. Best fit occurs for the case of a binary, consisting of an early type M3 star and a DA5.5 white dwarf. The fit is quite satisfactory, apart from the lower strength of $H\alpha$ line and the steeper spectral curve above 9000Å.

May 2022 - IDA Young Researcher's Programme

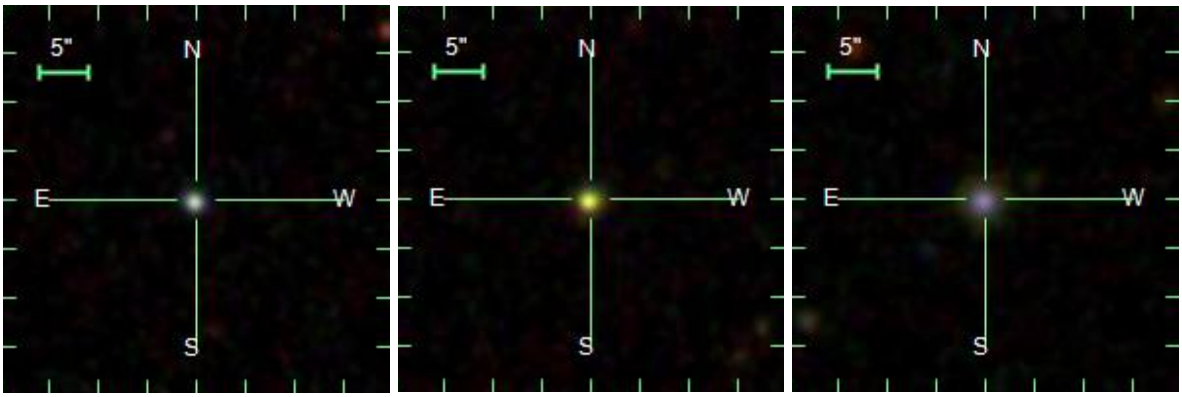
The considerations of December's observational results gave birth to the idea of combine astrometric and photometric data to make ML predictions about the nature of astronomical objects. This lead to a series of models, all described in section *Adding Astrometry* of the *Machine Learning Results*. Each model is the appropriate tool for different applications. For example, it was shown that the one with the best accuracy was the model that combined all the available photometric measurements, in all the optical, near and mid-IR bands, together with the astrometric measurements. So, if someone wants the most accurate predictions about a random population of unseen data that could be either stars, quasars or galaxies, this is the model that he should use. However, we have already underlined the biggest drawback of this model: it is trained upon shorter training datasets, because every magnitude's addition requires another cross-match with some survey, and not all surveys have available data for the same regions of the sky. Now think about the case that someone wants to make a complete quasar candidate catalogue and find the quasar density distribution. Firstly, he would like to acquire as many observations available as possible. Secondly, he would like to construct an ML model that does not require many features as inputs from different surveys. Because the W1-W2 color is the most important photometric feature and at the same time the cross much with AllWISE does not eliminate many observations, we used a model trained only on Gaia's astrometry and AllWISE photometry. Therefore, even if we knew that the all-color model was more effective, we chose to use the simpler model trained on the astrometry and just the WISE colors. This way, we knew that we would insert a higher uncertainty to the predicting outcomes, but at least we could talk in more realistic population numbers instead of approximations. The observing run was designed in order to detect and quantify the limitations of this model. To determine an upper limit for the S/N_{pm} above which the predictions become unreliable. Such estimations will be crucial when we later on the thesis calculate the quasar population for an area that covers the whole sky from $b = 40°$ all the way up to the North Galactic Pole. Here we present the science justification and the candidate selection criteria as described on the proposal we sent to IDA.

Part of the NOT IDA Application

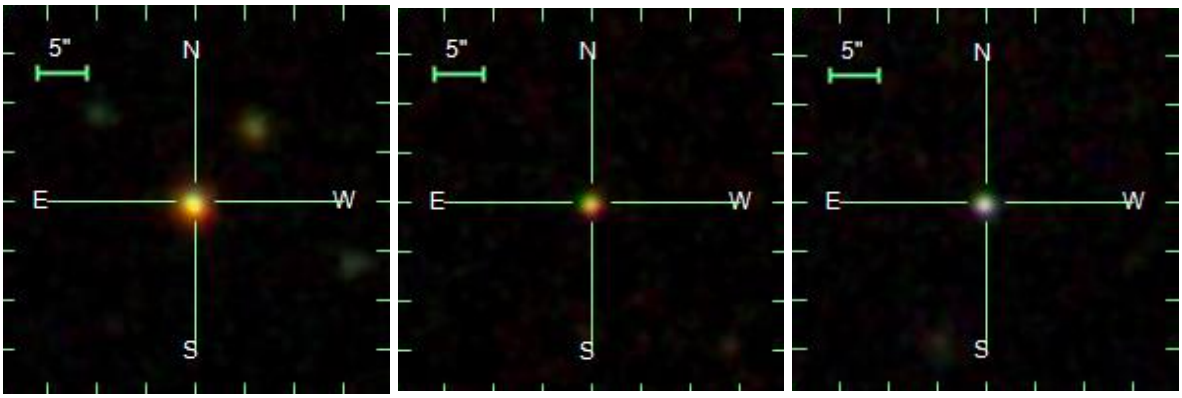
9/5/2022

'The selection criteria of quasars around the north galactic pole are based on XGBoost, a machine learning algorithm trained to solve the 3 class (star, quasar, galaxy) classification problem. The features required for a precise classification are ranged from the optical (g-r) to mid-infrared photometry (W1-W2, W3-W4 AllWISE) and also include the astrometric measurements from Gaia EDR3 survey (parallax and proper motion). The resulting performance of the model reaches an accuracy of 99.2%. The source dataset of the proposed candidates is queried from the Gaia EDR3 catalogue for b>80 degrees and $G < 20$, cross-matched with AllWISE catalogue. These 386.940 objects are classified by our model, predicting 11.088 QSOs, 374.419 stars and 1433 galaxies, allowing therefore the estimation of the quasar density population. Sources with known spectrum are excluded and our interest is focused on the predicted QSOs with $S/N < 2$, to ensure a lower contamination by stellar population. Out of this final cut-off, we carefully select interesting quasar candidates using color criteria. It is known that $u - g > 4$ is a common factor for quasars that are redshifted beyond z 3, since the $Ly\alpha$ emission line in such case is redshifted to ~ 5000Å, resulting in an excess on the G band. Similarly, quasars with $r - z > 0.5$ are highly probable to be dust-reddened quasars. However, quasars in the majority of cases are selected based on their blue optical colors, and so candidates who satisfy the above u-g and r-z conditions are systematically overlooked in the past selection techniques. With this proposal we wish to observe 10 machine learning predicted

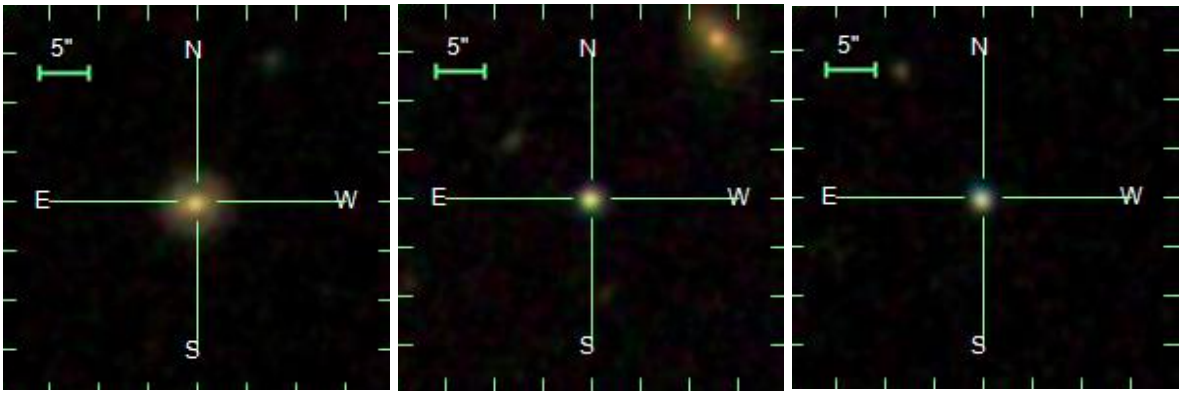
quasars that are not classified as such by the SDSS survey. The list includes 2 bright candidates ($G < 18$) that also appear red in the r-z color. Out of the total, 3 candidates have $u - g > 3.4$, while only 3 more are blue and selected in the typical $W1 - W2 > 0.8$ range of quasars.'



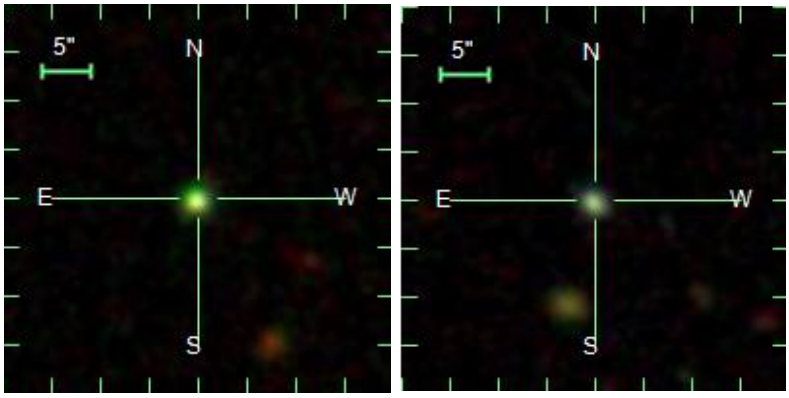
(a) J123022.68+191340.0 (b) J122153.19+235324.3 (c) J123654.12+321711.6



(d) J130550.86+230841.4 (e) J131925.63+260504.6 (f) J123504.03+202301.6



(g) J125323.0+215717.6 (h) 121222.63+272609.8 (i) J123430.94+213600.7



(j) J130504.74+301206.8 (k) J130448.30+354845.7

Figure 6.7: Targets of May's observing run, as seen in Object Explorer of the SDSS.

Table 6.9: Table of quasar candidates, targets of May 2022

Target	RA	DEC	parallax	S/N$_{par}$	PM	S/N$_{PM}$
J123022.68+191340.0	187.594521	19.22777778	-0.3471	-1.1213	2.594	3.4366
J122153.19+235324.3	185.4716667	23.89008333	-0.0078	-0.035	0.193	0.6632
J123654.12+321711.6	189.2255369	32.28655894	0.6785	2.5989	1.143	3.048
J130550.86+230841.4	196.461927	23.144828	0.0855	0.3473	0.272	0.5731
J131925.63+260504.6	199.856822	26.084639	0.7238	2.4321	1.44	2.901
J123504.03+202301.6	188.766827	20.383753	0.2128	0.7242	1.968	4.5084
J125323.0+215717.6	193.3458469	21.95490184	0.3518	0.2634	1.174	0.5677
J121222.63+272609.8	183.094295	27.436077	0.1095	0.5762	0.81	3.1682
J123430.94+213600.7	188.628909	21.600209	-0.5961	-2.345	1.203	3.459
J130504.74+301206.8	196.269773	30.201911	0.091	0.7057	0.197	1.4613
J130448.3+354845.7	196.2008333	35.8126944436	1.292	3.071	1.52	0.384

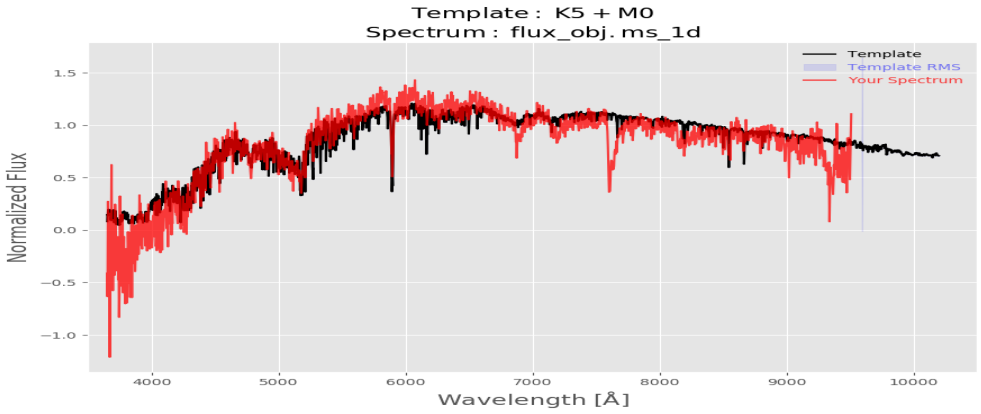
Table 6.10: Quasar candidates for observations on May 2022 and their coordinates. Astrometric criteria for selecting those targets imposed constraints on the parallax and proper motion.

On the nights of 25^{st} and 26^{th} of May, we had the pleasure to conduct ourselves the astronomical observations at the NOT telescope. As it turned out, all of the astrometric outliers with $S/N_{pm} > 3$ apart from one were all stars. This suggests that in order to have more pure quasar catalogues, we should reject the quasar predictions with proper motion signal to noise ratio higher than 3. Of course this way we lose some true candidates, but the stellar contamination we eliminate is more significant in count numbers. Another important result of this observing run concerns the mid-infrared colors. As seen in Table6.4 all of the candidates were deliberately chosen to fall in the stellar regime, with $W1 - W2 < 0.8$. Such quasars would have evaded the empirical criteria, but the ML was able to successfully retrieve them.

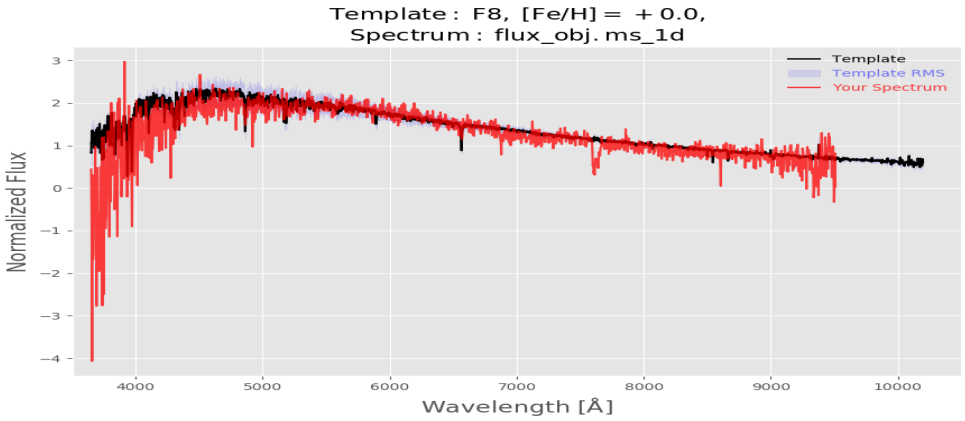
Table 6.11: Table of quasar candidates continued

Target	$z_{this\ work}$	AB	G_{SDSS}	r-z	u-g	W1-W2	EXPTIME
J123022.68+191340.0	-	-	19.456	0.205	0.863	0.777	2x500
J122153.19+235324.3	-	-	19.53	0.428	3.22	0.644	1x500, 1x600
J123654.12+321711.6	0.424	0.0	19.149	0.414	0.39	0.754	2x400
J130550.86+230841.4	-	-	18.9	0.856	2.399	-0.073	2x300
J131925.63+260504.6	-	-	20.74	0.768	4.058	0.433	2x1200
J123504.03+202301.6	-	-	19.322	0.256	1.23	0.778	2x500
J125323.0+215717.6	0.2	1.64	18.417	0.605	0.881	0.729	2x300
J121222.63+272609.8	-	-	19.306	0.432	1.659	0.564	2x500
J123430.94+213600.7	-	-	19.416	0.292	1.344	0.724	2x500
J130504.74+301206.8	3.82	0.0	18.96 4	0.22	4.91	0.622	2x300
J130448.3+354845.7	0.314	1.1	19.34	0.78	0.47	0.94	2x500

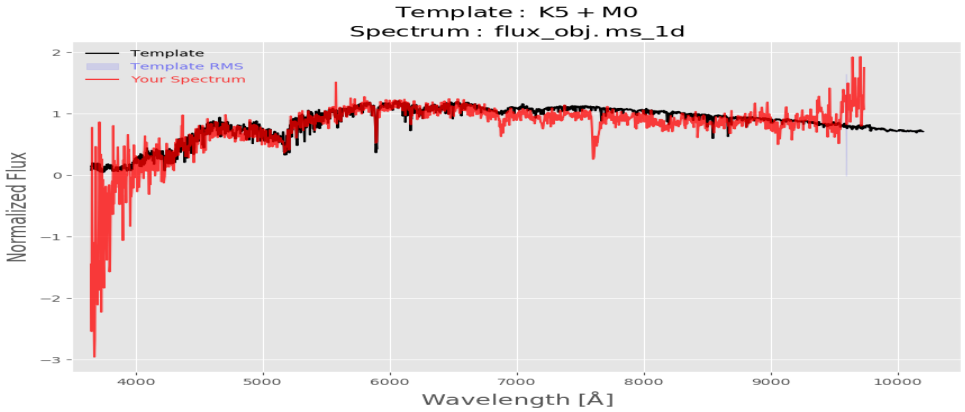
Table 6.12: Each target's redshift is determined by spectral analysis in this current work. No previous redshift data were available. Wherever the redshift value is missing, it is a star.



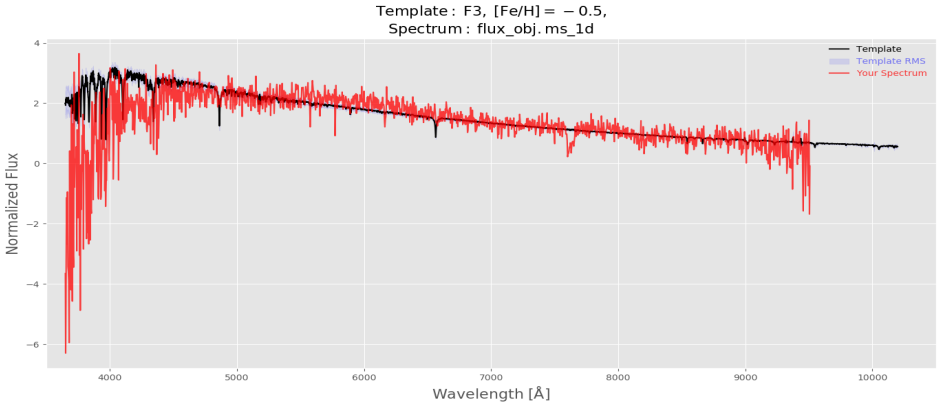
(b) Pyhammer template fit of J130550.86+230841.4. Best fitting occurs for the case of a binary system, with a K5 primary stellar type and a M0 companion. Strong Na at 5809Å and TiO lines around 7000Å.



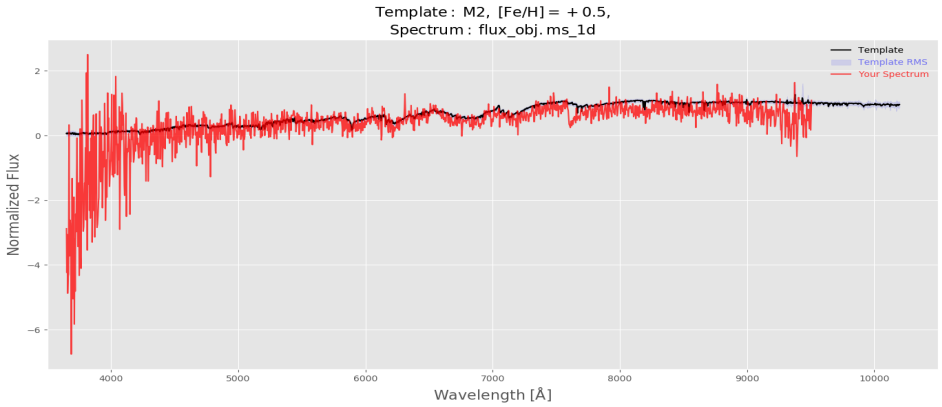
(d) Pyhammer template fit of J123430.94+213600.7. Fitted well with a late F type (F8) star template. $H\alpha$ absorption line is evident at 6564Å



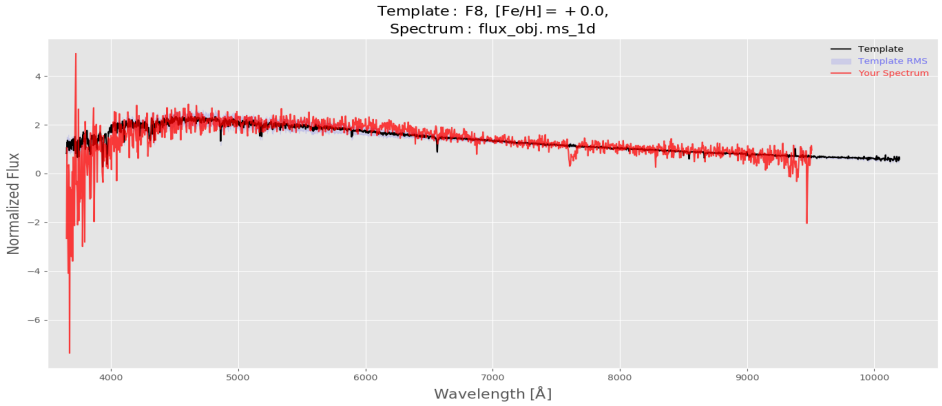
(f) Pyhammer template fit of J131925.63+260504.6. Another proposed binary by the Pyhammer fit. Primary stellar type is K5, accompanied by a secondary M0 star. Sodium line at 5890Å appears to be strong



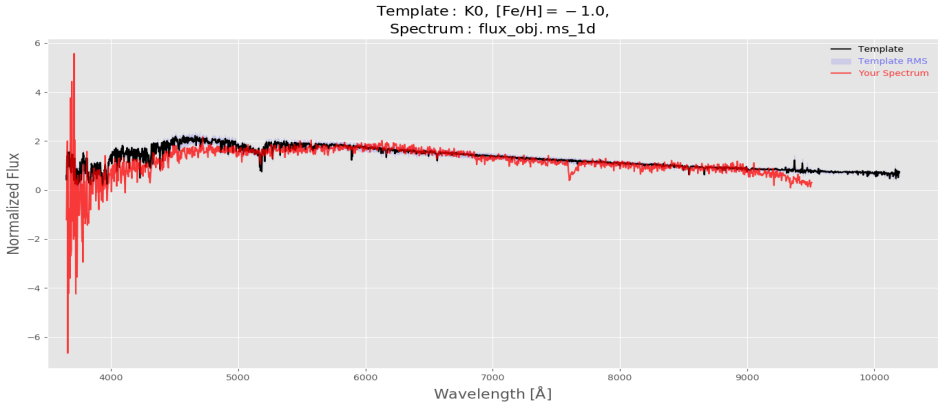
(h) Pyhammer template fit of J123022.68+191340.0 with an early type F (F3) star.



(j) Pyhammer template fit of 122153.19+235324.3 with an early type M (M2) star.



(l) Pyhammer template fit of J123504.03+202301.6 with a late type F (F8) star.



(n) Pyhammer template fit of J121222.63+272609.8 with an early type K (K0) star.

For the candidates of this observing run we didn't acquire the UKIDDS colors, so in Table 6.12 we demonstrate the $r-z$ and $u-g$ colors instead. Typical quasar values in the $r-z$ are below 0.5. This range covers the blue and normal quasars. Values higher than 0.5 characterize the dust reddened quasar population, so those that appear red independently of their redshift. Candidates of this category are also included in the observing plan. The $u-g$ color was another candidate selection criterion for this observing run. Candidates with high ($u-g>4$) values are expected to be very redshifted, with a distinguishable $Ly\alpha$ forest. For the spectral extraction of all the quasars we observed we have followed a different method which is described in the following paragraph.

6.2 Quasar spectrum reductions

The data reduction followed in order to obtain the spectra of the observed targets made use of a python scipt pipeline written by Fynbo and Krogager, and will not be discussed here. The detailed steps of the science frames reduction, the wavelength and the flux calibration can be found in Appendix B. Here we only discuss the basic parameters that need to be taken into account when we apply the final step of the spectrum reduction - the template fitting - and the corresponding equations.

6.2.1 Quasar templates

As a final step of the data reduction, we had to fit our extracted spectrum to a quasar template and find the redshift and the extinction A_B. For the redshift determination, the quasar template is really helpful. Characteristic emission lines are already shown in the quasar spectrum, and by identifying them and matching them to the ones in our extracted spectrum, the redshift is automatically determined. The addition of extinction on the other hand, lets the quasar template fit the continuum level of the extracted spectrum. In this work we used one of the two available quasar templates depending on the redshift of the quasar, (Selsing or Vanden Berk template) and extinction like that of the SMC (Small Magellanic Cloud).

The first template we are using throughout the chapter is a composite quasar spectrum constructed by Vanden Berk et al. [Vanden Berk, 2001] in August 2001, using an SDSS dataset that contained more than 2200 quasar spectra, measured in the wavelength range 3800 - 9200 Å. The redshift interval of those SDSS quasars was 0.044 $<$ z $<$ 4.789. The median composite is made from all the available quasar components and its contimuum is fit by two power laws; one for shorter and one for longer wavelengths. The UV continuum bump is considered to be the result of friction on the accretion disk, leading to thermal emission. The lines in this wavelength range are, as discussed in Part I, produced by a broad and a narrow line region around the central SMBH. The addition of a second power law to describe better the change of the continuum's gradient at redder wavelengths is due to the contribution of the host galaxy. For smaller redshifts, the host galaxy's dust absorbs the light in the μm range and re-emmits it in the near infrared, causing an increase of the continuum level at wavelengths above $\sim$ 5000 Å. Apart from that, the stars in the host galaxy contribute with stellar absorption lines in the quasar spectrum. It has been found that the lower the redshift, the stonger those stellar absorption lines are, as a result of the reduced quasar luminosity to host galaxy luminosity ratio, due to the proximity of the quasar.

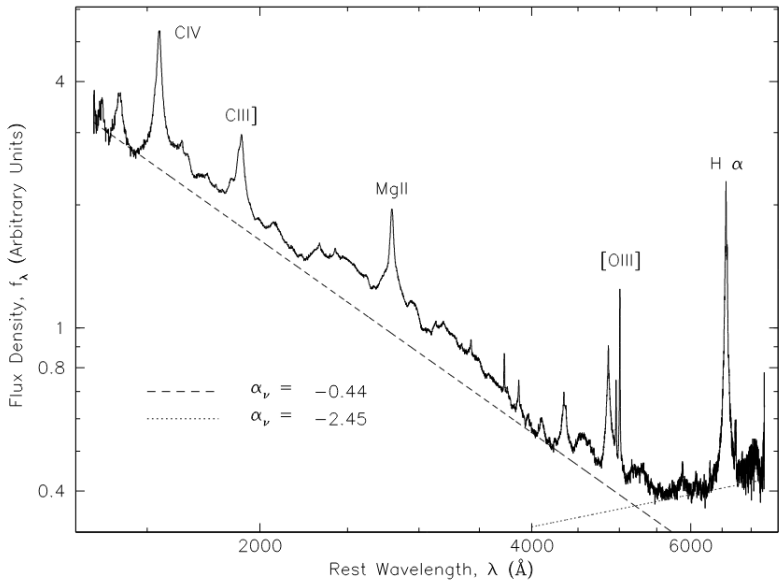


Figure 6.9: Median composite quasar spectrum for low redshifted quasars, with evident host galaxy's light contanimation. Figure taken from :[Vanden Berk, 2001].

Of course, the Vanden Berk template cannot fit every quasar spectrum, especially the very bright and redshifted quasars, for which the host galaxy contribution is insignificant. To bypass the issue of host galaxy contamination, [Selsing, 2015] selected a sample of quasi-stellar objects from the SDSS catalogue with higher redshifts ($1 < z < 2.1$) and very bright in the r band ($r < 17$). The targets were observed with X-Shooter that has a very wide wavelength range, extracting the quasar spectra in the λ range ∼300 - 2500 nm. The resulting composite spectrum of quasi-stellar objects (QSO) is shown in the Figure 6.10.

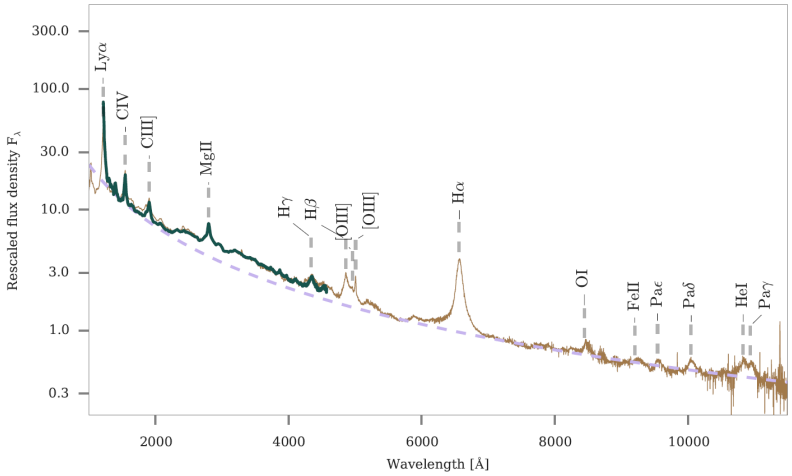


Figure 6.10: Quasar composite spectrum for redshifts $z > 1$, constructed by J. Selsing et al. in 2015 (brown continuous line). The purple dashed line shows the fitting of a power low to the continuum level. Figure taken from : [Selsing, 2015].

6.2.2 Extinction

The light from the targets that we want to observe has been impacted by intervening gas and dust clouds, and the overall extinction it experienced until it reached the observer is a factor needing to be accounted for. Up to sixty percent of some galaxies' total radiation is due to dust, while the scale of attenuation it causes depends on the grains' properties, like the size, the geometry or the type of the grains. Absorption is efficient on grains with physical sizes that are of the same order of magnitude as the incident radiation, or higher. But the number density of the dust grains in the interstellar medium as a function of their size follows a law $n(\alpha) \sim \alpha^{-3.5}$, suggesting that the vast majority of the grains are small. With sizes of about 0.1-1 μm, those silicates, metal oxides, amorphous or graphite carbonaceous grains are highly efficient in scattering and absorbing the shorter wavelengths (UV, optical) and thermally re-emitting it in the *submm* scale. For ground-based observations, apart from the interstellar dust, the Earth's atmosphere is another contributor in the astronomical object's extinction. Analytic derivation of dust's role in scattering, absorbing and re-emitting the light from astronomical sources is presented in Appendix B.2.

Extinction $A(\lambda)$ caused by dust is not the same for all wavelengths. Bluer parts of the spectrum are more prone to extinction, leading to the effect known as *reddening* of radiation. It is measured through the color excess E_{B-V} that represents the selective extinction. If B_0 and V_0 are the intrinsic magnitudes in the 450 nm and 550 nm respectively, then the extinction at each band is

$$\begin{aligned} B &= B_0 + A_B \\ V &= V_0 + A_V \end{aligned} \tag{6.1}$$

and the color excess can be expressed as

$$E_{B-V} = A_B - A_V = (B-V) - (B-V)_0 \tag{6.2}$$

The parameter R_V is defined as $\frac{A_V}{E_{B-V}}$ and can approximate the galactic extinction curves. This empirical factor changes with repsect to the grain size distributions and how these vary in dense or more diffuse enivironments. From its definition, R_V is equal to $\frac{A_V}{A_B - A_V}$, leading to the relation between the extinction in the V and the extinction in the B band:

$$A_B = \left(\frac{1+R_V}{R_V}\right) A_V \tag{6.3}$$

Typical values are $R_V \sim 3.1$ for the Milky Way, $R_V \sim 2.7$ for the SMC and ~ 5 for dense molecular clouds. Denser regions have higher R_V values and a less steep rise in the far-UV range, which denotes a larger in size grain distribution. The situation is shown in Fig. 6.11, where the normalised A_λ is plotted as a function of $\frac{1}{\lambda}$.

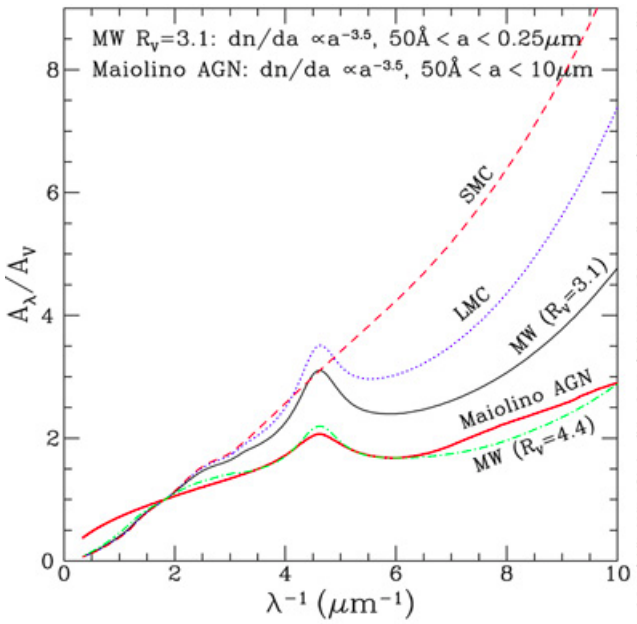


Figure 6.11: Galactic extinction curves of MW, SMC and LMC. Less dense regions with smaller dust grains show lower R_V values. The SMC extinction curve can be approximated by a linear relation with λ^{-1}. A broad absorption feature appears at about $\lambda^{-1} \sim 4.6\mu m^{-1}$ ($\lambda \sim 2175$Å) for the LMC and MW extinction curves, which is attributed to dust grains rich in carbon. This bump at 2175Å is absent in the SMC extinction curve.

The extinction curve we use in the script for the spectrum reduction is the SMC's and is expressed by a fitting function proposed by Pei [Pei, 1992]. Using four free parameters and graphite-silicate models for the grains, the function can reproduce the approximate $1/\lambda$ dependence of the SMC extinction, with a proper strength of the 2175Å and other features (feature at 9.7 μm and 18 μm, see Table 6.13). Three of the terms represent the background (BKG), the far-UV and far-IR extinctions. The fitting function sums the contributions of all the six parameters according to the equation :

$$\frac{A_\lambda}{A_B} = \sum_{i=1}^{6} \frac{a_i}{\left(\frac{\lambda}{\lambda_i}\right)^{n_i} + \left(\frac{\lambda_i}{\lambda}\right)^{n_i} + b_i} \tag{6.4}$$

Table 6.13: Analytic extinction curve parameters for the Small Magellanic Cloud

Parameter	α_i	$\lambda_i(\mu m)$	b_i	n_i	K_i
BKG	185	0.042	90	2.0	2.89
FUV	27	0.08	5.50	4.0	0.91
2175 Å	0.005	0.22	-1.95	2.0	0.02
9.7 μm	0.010	9.7	-1.95	2.0	1.55
18 μm	0.012	18	-1.80	2.0	1.72
FIR	0.030	25	0.00	2.0	1.89

6.2.3 Adding Photometry

On top of the spectrum of each quasar, the photometric measurements in the various pass bands are shown. To do so, we had to convert the AB magnitude to flux. When the spectral flux density f_ν is measured in CGS

units of $ergs^{-1}cm^{-2}Hz^{-1}$, the definition of the magnitude is expressed in equation 6.5:

$$m_{AB} = -2.5\ log_{10} f_\nu - 28.6 \tag{6.5}$$

Solved with respect to the flux density per unit frequency, the former equation leads to:

$$f_\nu = 10^{-0.4(48.6+m_{AB})} \tag{6.6}$$

The final step requires the transformation of the flux density from its expression per unit frequency to its expression in terms of λ through the equation:

$$f_\lambda = \frac{c}{\lambda^2} f_\nu \tag{6.7}$$

where the speed of light c should be in CGS units as well. The SDSS magnitudes are already in AB system, but the AllWISE and UKIDDS are in the Vega system and an AB offset, dependant on the specific wavelength band needs to be added to each magnitude. The exact values for every band can be found in table 6.14.

Table 6.14: Magnitudes in near and mid infrared that are measured in Vega system, their effective wavelengths and their AB offset. Optical colors are not included because as measured by the SDSS, they don't have significant difference in different magnitude systems.

Magnitude	$\lambda(\mu m)$	AB offset
Y	1.0305	0.634
H	1.6313	1.379
J	1.2483	0.938
K	2.2010	1.9
W1	3.368	2.699
W2	4.618	3.339
W3	12.082	5.174
W4	22.194	6.62

6.2.4 Extracted spectra

After all the considerations discussed in the previous paragraphs, the proper quasar template is fit, the redshift and reddening is determined and the photometry is added individually for each one of the observed targets. In Fig. 6.12-6.30 all the extracted quasar spectra are shown. For some of the quasars a gaussian fit of their emission lines is done and the rotational velocites are calculated (Appendix B.9). Some remarks about every observation's spectral fitting are discussed below:

J012534.07+351344.0

Observed during December with the use of Grism 20, this is a relatively low redshift (z=0.312) quasar, fitted with the Vanden Berk template. The extinction for J012534.07+351344.0 is noticeably high, $A_B = 2.74$, making this target an example of a strongly reddened quasar. While the template can fit well the continuum level and is in agreement with the photometric points, it cannot fit the $H\alpha + NII$ lines which appear in excess. The oxygen doublet lines and $H\beta$ line are also a bit underestimated by the fitting. Significant SII emission is also evident at ~8800 Å but is not included in the quasar template.

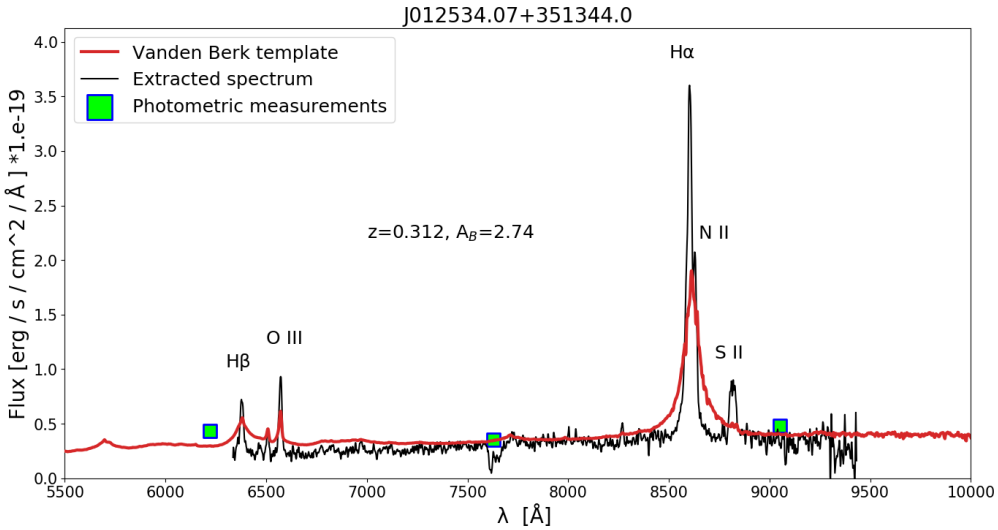


Figure 6.12: Quasar J012534.07+351344.0 observed in December with grism 20

J012848.44+341704.9

Similarly to target J012534.07+351344.0, this quasar is very reddened, with $A_B = 2.19$. It has been observed with Grism 20, fit with the Vanden Berk profile, and its redshift is estimated at $z = 0.346$. The photometric measurements are well fitting the spectrum's flux. Excess in the flux of OII 5007 Å emission line is observed.

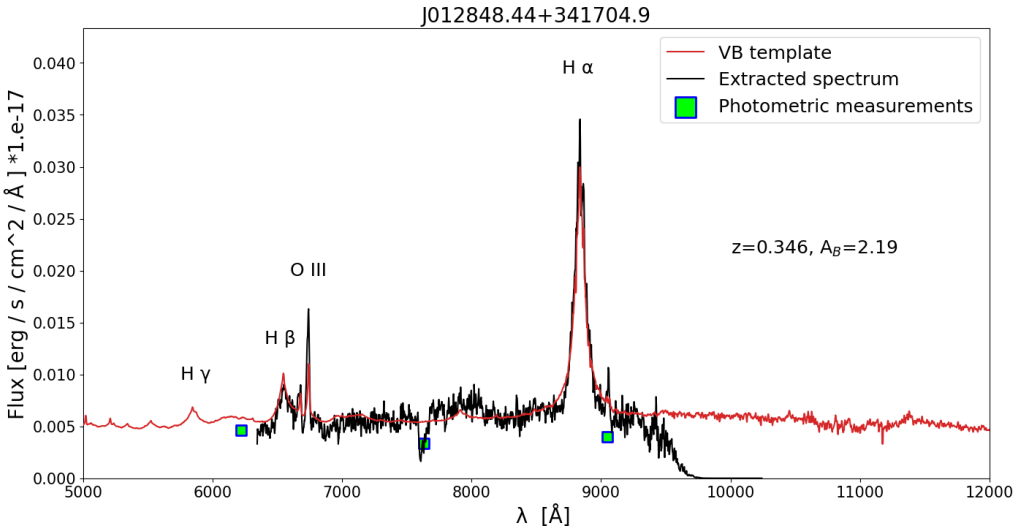


Figure 6.13: Quasar J012848.44+341704.9 observed in December with grism 20

J015455.79+203623.8

In the spectrum of this target we estimate the redshift by matching the only evident emission line with $MgII$, finding $z = 2.291$. Unfortunately the range of Grism 20 that was used for this observations ends right after the wavelength where $CIII$ line was supposed to be seen. The photometric point in the r band though coincides with this line on the template spectrum, providing another indication that our redshift estimation must be true. The troughs that exist blueward to the magnesium line suggest that this is a LoBAL quasar. The spectrum is fitted with the Selsing template in combination with extinction $A_B = 1.78$.

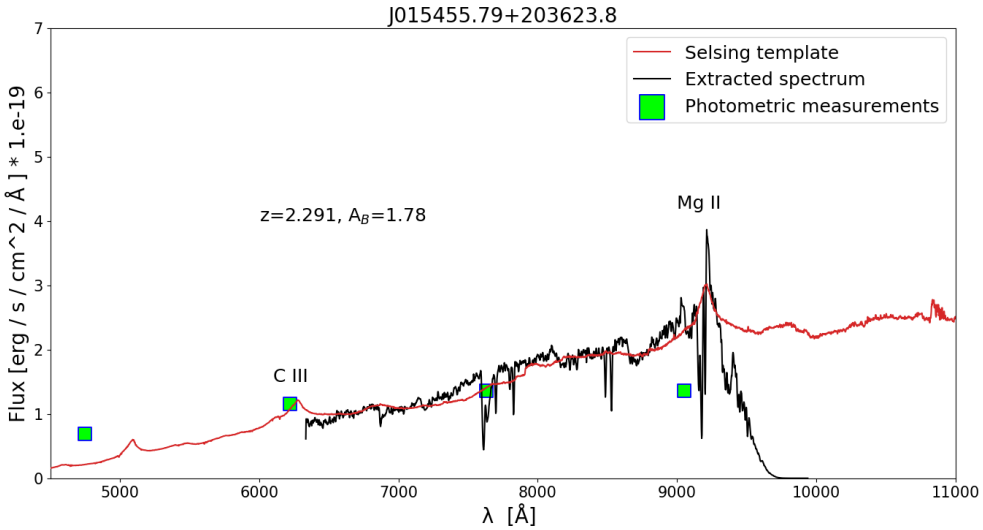


Figure 6.14: Quasar J015455.79+203623.8 observed in December with grism 20

J015748.65+284752.6

Quasar measured with Grism 20. $H\alpha$ appears much stronger than the template fit's shape, but this could be because it falls on the edge of the grism, where the signal is not so good. Its redshift is found to be $z = 0.444$, justifying the use of the Vanden Berk quasar profile and its extinction is significant ($A_B = 1.64$).

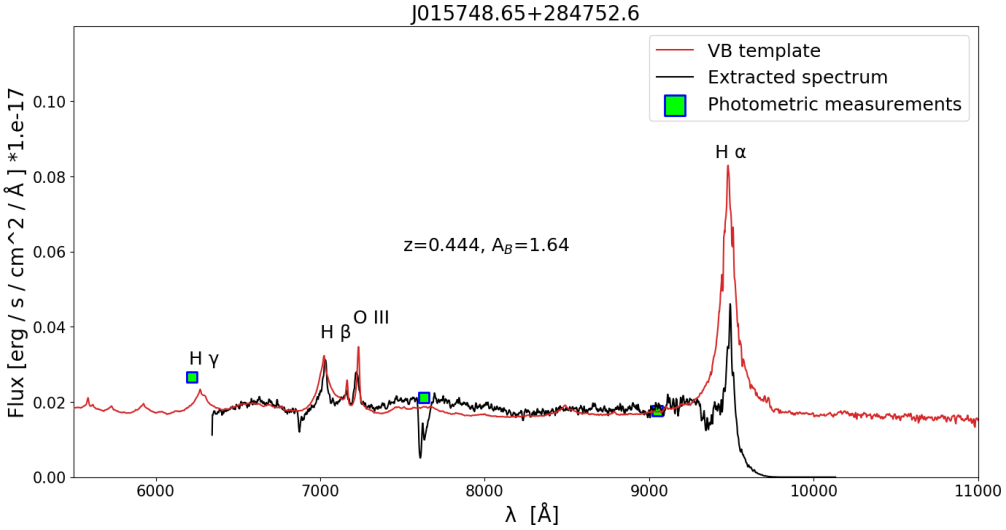


Figure 6.15: Quasar J015748.65+284752.6 observed in December with grism 20

J025004.61+324039.7

This quasar is very similar to J012848.44+341704.9 regarding their redshift, fluxes and extinction values. However for this target the extinction is more significant, while the $O\ III$ 5007 Å line has much less strength. With $z = 0.3685$, it has been fit with the Vanden Berk profile.

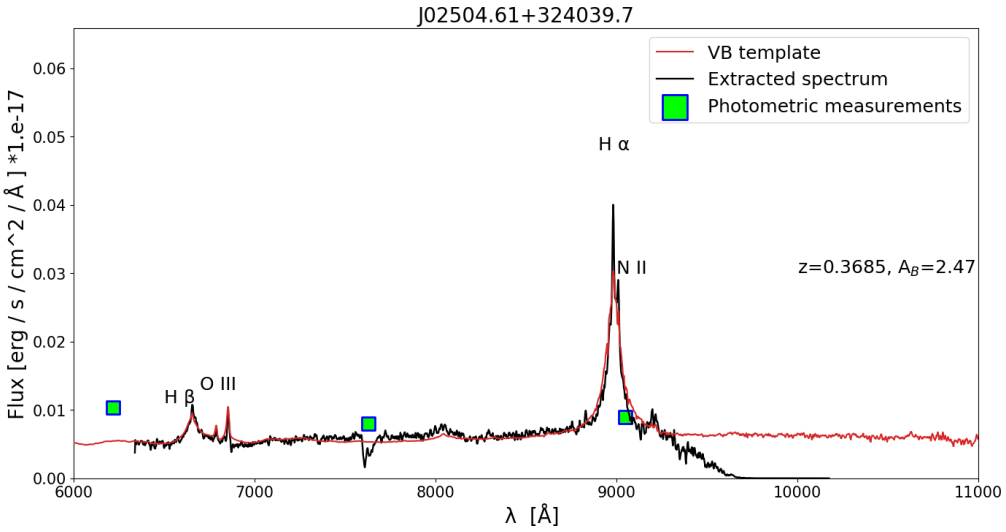


Figure 6.16: Quasar J025004.61+324039.7 observed in December with grism 20

J025832.77+354253.6

Example of a low redshift ($z = 0.26$), yet extremely reddened quasar ($A_B = 4.25$). Vanden Berk template can fit the extracted spectrum, implying that partially the reddening is due to the host galaxy's contamination. Measured with Grism 20, the spectrum does not extend to the wavelengths where the $OIII$doublet could be observed, but there are OI and SII lines next to the $H\alpha + NII$ lines.

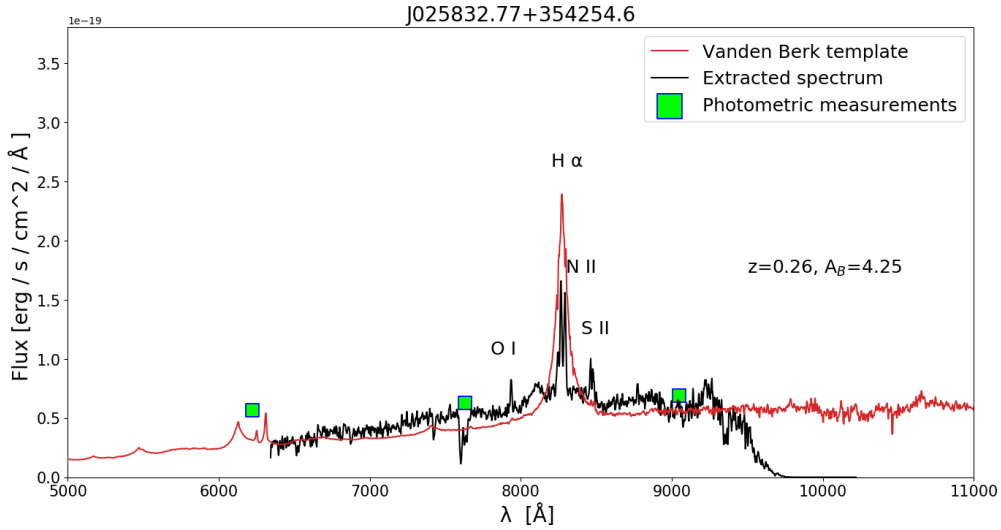


Figure 6.17: Quasar J025832.77+354254.6 observed in December with grism 20

J130703.91+251415.5

Another example of relatively nearby but very reddened quasar, target J130703.91+251415.5 has $z = 0.273$ and $A_B = 2.33$. It appears with strong $O/III/5007$Å and $H\alpha$ emission, and there are also hints of SII emission, a bit indiscriminate due to the high noise level. The vertical line at ~ 9200Å is noise.

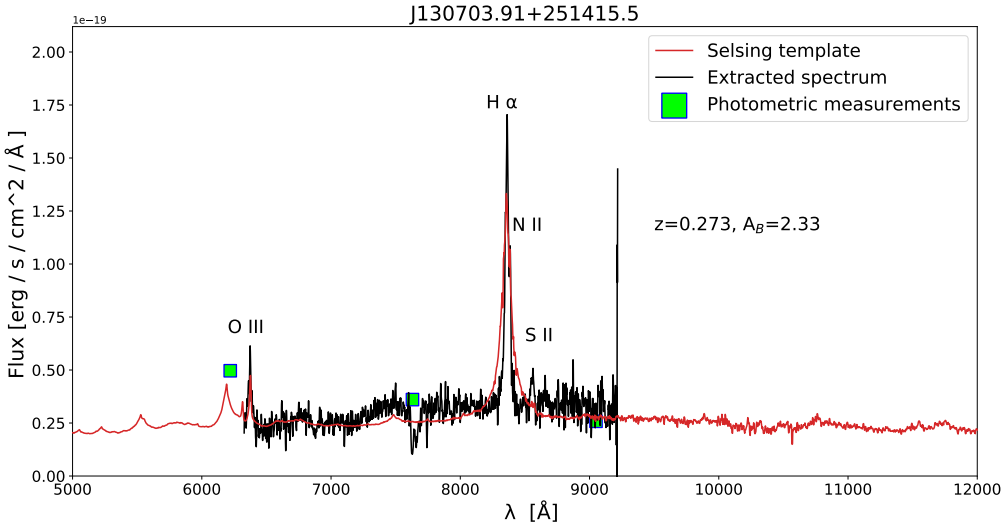


Figure 6.18: Quasar J130703.91+251415.5 observed in December with grism 20

J022742.93-173121.5

Quasar J022742.93-173121.5 has much higher redshift ($z = 2.31$) than the quasars described before, allowing the observation of the most characteristic $Ly\alpha$ line, as well as the rest of the characteristic $Si\ IV, C\ IV, C\ III, Mg\ II$ lines. This is a LoBAL (Low ionization Broad Absorption Line) quasar with low ionization species such as $Mg\ II$, $Al\ III$ emission lines in its spectrum. Many abrupt drops in the flux, especially blueward and redward of the $SiIV$ emission exist. The photometric measurements seem to match the template, except for the r band measurement which falls in the atmosphere's Telluric absorption line and for which we have not corrected the extracted spectrum. There are indications for a $Ly\alpha$ forest for wavelengths smaller than 4000Å, but the Grism 4 that was used for this observation does not extend there. The fitting for this $z > 1$ quasar is achieved with the use of the Selsing template.

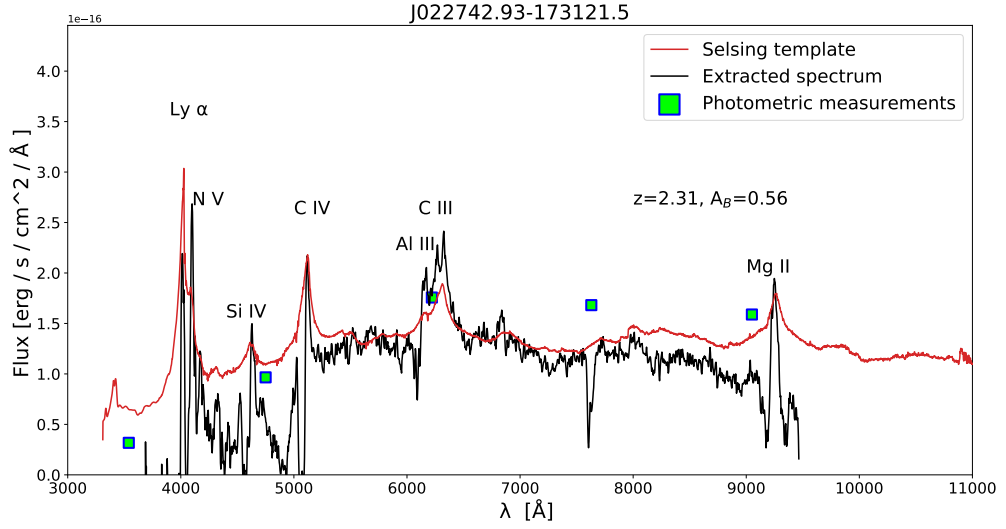


Figure 6.19: Quasar J022742.93-173121.5 observed in August with grism 4

J234859.52-193324.89

This is a normal quasar with no extinction and redshift $z = 0.44$. Grism 4 is used to obtain this spectrum and the several main emission lines can be seen, $Mg\ II, H\gamma, O\ III$. $H\alpha$ appears understimated in the spectrum, probably due to the bad signal at the end of Grism 4.

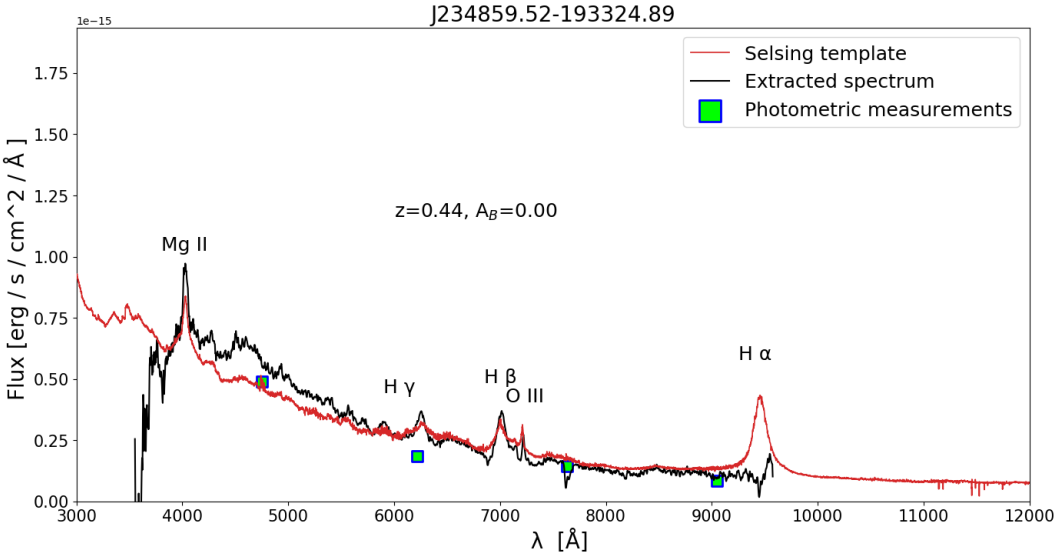


Figure 6.20: Quasar J234859.52-193324.89 observed in August with grism 4

J003634.82-140924.9

The redshift estimation for this target is quite tricky, due to the fact that we can hardly see any lines. We assume that the spectral feature at around 7000Å is the $MgII$ line and from that we estimate the redshift to be $z = 1.54$. The existence of consecutive troughs blueward to $MgII$ suggests that this is a LoBAL quasar. For the fitting of this quasar two Selsing templates with different A_B extinction are used. The blue template assumes no extinction at all, and is used to fit the redward to $Mg\ II$ part of the spectrum. For shorter wavelengths, a Selsing template with $A_B = 2.06$ (red line) seems to be the appropriate fitting.

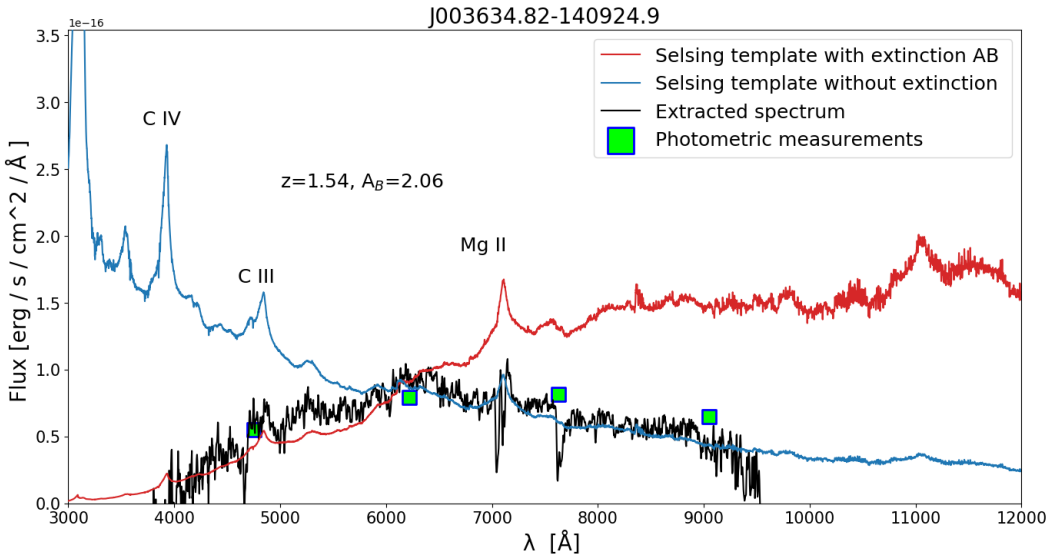


Figure 6.21: Quasar J003634.82-140924.9 observed in August with grism 4

J234530.36-135743.3

The observation of this target with Grism 4 could not give a clear spectrum from which a redshift estimation can be made. From a second observation performed in longer wavelengths (with Grism 19), Johan Fynbo was able to determine the redshift at $z = 2.325$. With this knowledge, we match the $Mg\ II$ line at about 9000Å. For shorter wavelengths we can detect the $C\ III$and $Al\ III$ lines, while the continuum is almost absent and has been replaced by big troughs. Based on these observations, we characterise J234530.36-135743.3 as a LoBal quasar.

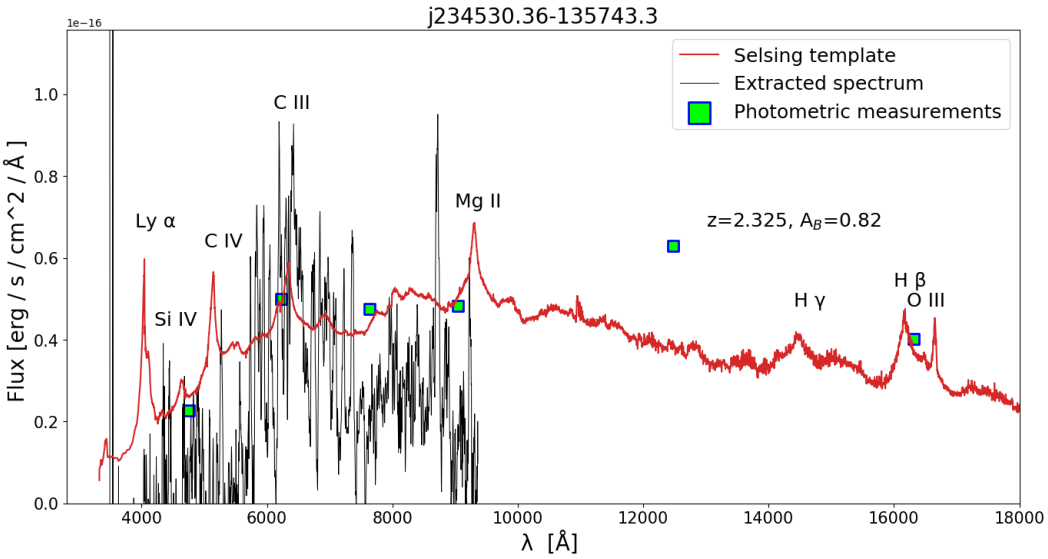


Figure 6.22: Quasar J234530.36-135743.3 observed in August with grism 4

J010013.02+280225.8

Quasar J010013.02+280225.8 is the most redshifted target among all our observations, with $z = 5.8$ and a very strong $Ly\alpha$ emission. $Ly\alpha$ is redshifted beyond 8000Å and is followed by a Gunn-Peterson trough [Gunn, Peterson 1965] for shorter wavelengths than that. Around 7000Å, we can see the appearance of $Ly\beta$ line. It is interesting to compare this spectrum with the spectrum of J130504.74+301206.8 that can be seen in Figure (@@@). For the latter, found at redshift $z = 3.8$, blueward to the $Ly\alpha$ we observe successive absorptions. This is the $Ly\alpha$ forest, created by intervening clumps of HI. But for target J010013.02+280225.8 the situation is different. Here we observe wide $Ly\alpha$ emission and a deep trough rather than a forest. This is an indication that the universe, as we observe it at $z = 5.8$, was uniformly filled with neutral hydrogen, that absorved the redshifted, HI sensitive $Ly\alpha$ emission. Thus, the reinonization epoch seems to not have been complete yet, at redshift $z = 5.8$. All the photometric measurements are in good fit with the extracted spectrum, but the depth of the Gunn-Peterson trough is underestimated by the Selsing template which we used for the fitting.

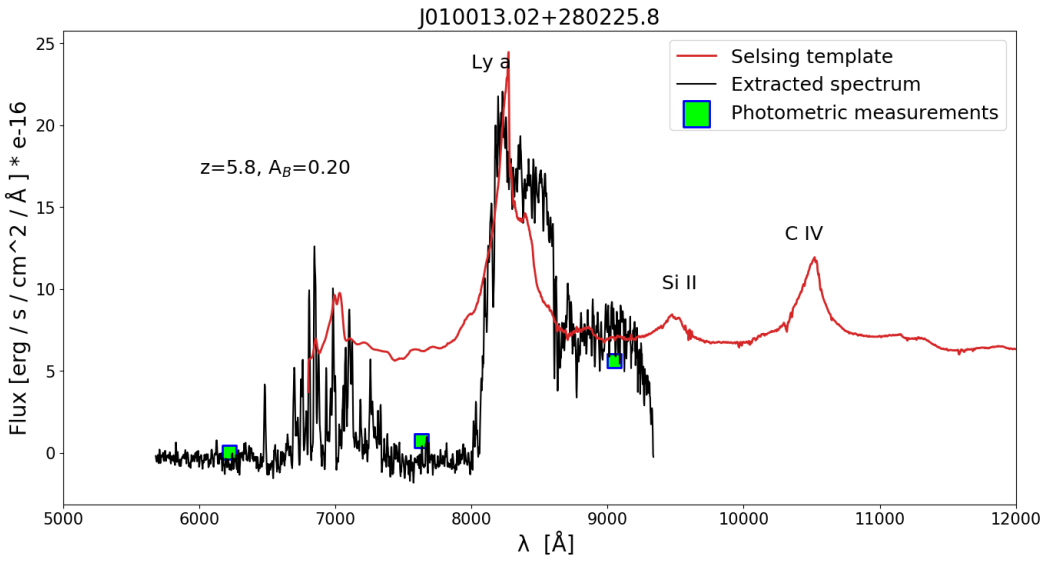


Figure 6.23: Quasar J010013.02+280225.8 observed in August with grism 4

J010339.41-132238.91

The absence of many emission lines in the spectrum of this target makes its redshift estimation difficult. We estimate that the emission at around 7800Å is the $MgII$ emission line, but unfortunately it falls right on top of the Telluric absorption line and is thus hard to discriminate. Bsed on this match, we find its redshift to be $z = 1.75$. Target J010339.41-132238.91 is another example for which the fitting required two templates with different extinction values. The blue part of the spectrum is fit with the Selsing template without extinction, while the red part requires the Selsing template in combination with $A_B = 1.75$ in the B band.

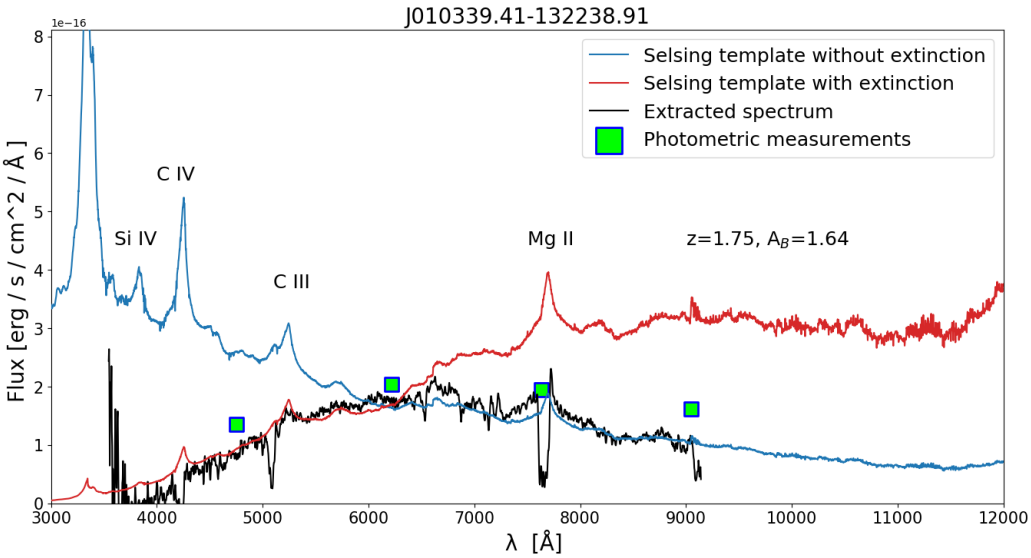


Figure 6.24: Quasar J010339.41-132238.91 observed in August with grism 4

J012813.63-120319.79

We suggest that the feature at about 7000Å is the $MgII$ redshifted line, which leads to a redshift of $z = 1.447$. The existence of this line in combination with the deep troughs that exist for wavelenghts bluer than $MgII$, suggest that this is a LoBAL quasar. The flux of target J012813.63-120319.79 drops considerably for both short and near infrared wavelenghts, and for that we suggest a double fitting for its spectrum. The short wavelengths match a Selsing profile with an extinction of $A_B = 1.32$. This UV extinction would leads to an increasing flux in the infrared, while the extracted spectrum shows a downward trend in those wavelengths, and the difference from the template (red line) gradually increases. So to fit the longer wavelengths we use a zero extinction Selsing template. The photometric measurements match the flux values.

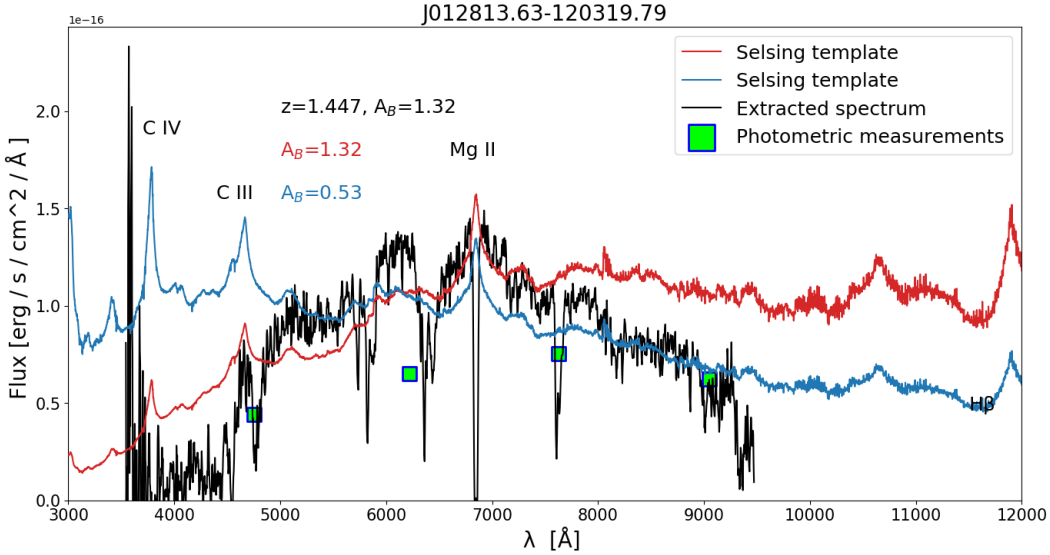


Figure 6.25: Quasar J012813.63-120319.79 observed in August with grism 4

J000404.29-135905.43

Although it looks like it is a BAL quasar, the bad signal of this observations makes its nature and redshift determination hard. We assume that the situation here is similar to the case described above(J012813.63-120319.79). The feature at about 7000Å is matched with $MgII$ line, resulting in a redshift $z = 1.44$. For the fitting we use the Selsing template with $A_B = 1.23$ extinction. THe photometric measurements are in good agreement with the observed fluxes.

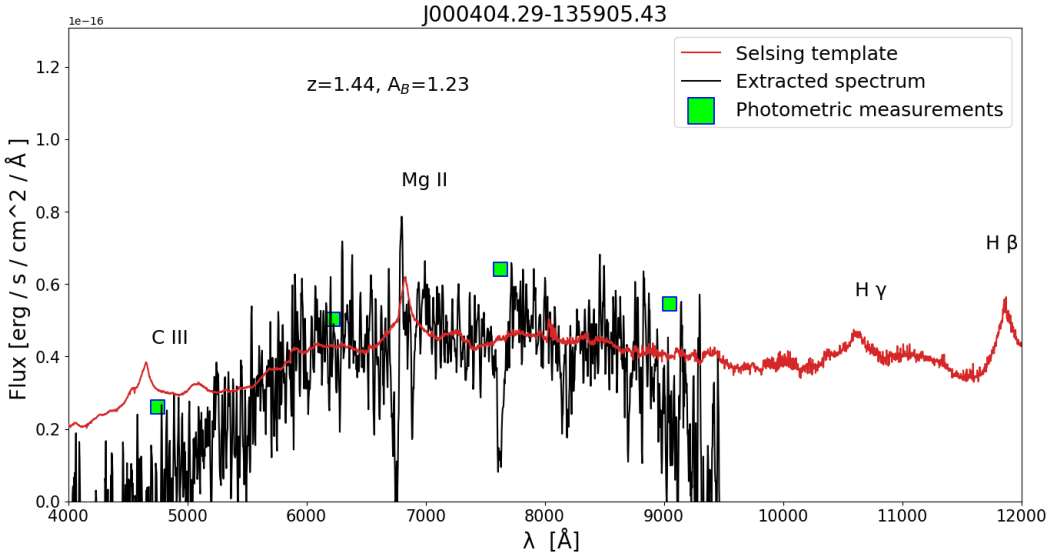


Figure 6.26: Quasar J000404.29-135905.43 observed in August with grism 4

J123654.12+321711.6

Very similar to J234859.52-193324.89, this target measured on May with the use of Grism 4 is another example of a normal quasar. It is found at redshift $z = 0.424$ and all the lines from $MgII$ to $H\alpha$ are visble in its spectrum. To fit it, the Vanden Berk template is used, in combination with no extinction in the B band. The photometric points are all uniformly shifted to higher flux values, which could be an indication of the source's variability. That means that the flux coming outwards from the quasar was higher at the time of the photometric measurements than the time of the spectroscopic observation.

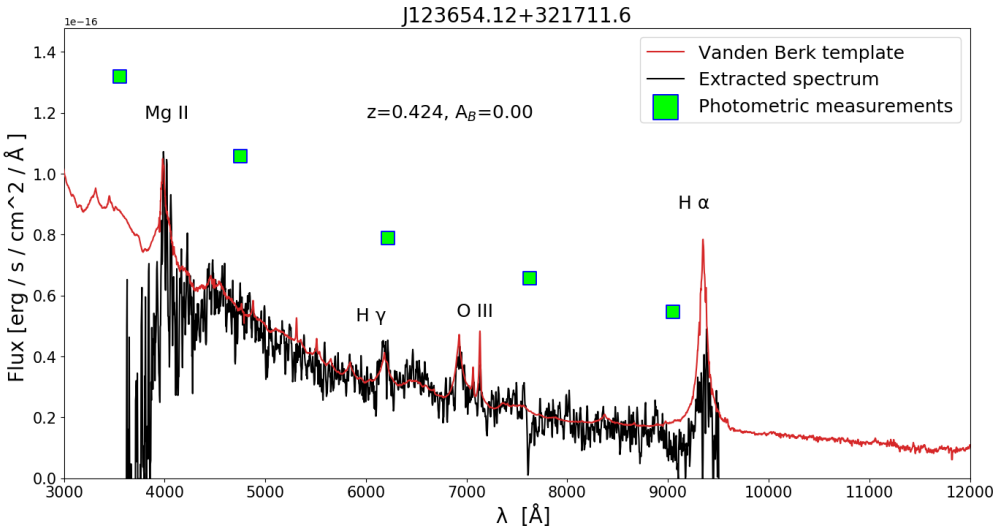


Figure 6.27: Quasar J123654.12+321711.6 observed in May with grism 4

J125323.0+215717.6

J125323.0+215717.6 is a $z = 0.2$ emission line and probably AGN-like galaxy. It is fit with the Vanben Berk profile and an extinction $A_B = 1.64$. There seems to be an excess of $OIII$ emission and the $H\alpha$ line appears narrower and less strong than the template fit expects it to be. Due to the low FWHM of the emission lines and its proximity ($z = 0.2$), we suggest that this target is a Seyfert galaxy. This can also be infered with naked eye in Fig. 6.7g, where the host galaxy is observable. To confirm this we made a Gaussian fit on the stronger lines to find their intensity. By calculating the ratios of nearby lines, namely the $NII/H\alpha$ and $OIII/H\beta$ ratio, this extragalactic source can find its place in a Baldwin, Phillips, Terlevich (BPT) diagram. As the authors showed in [Baldwin, J. A. ; Phillips, M. M. ; Terlevich, 1981], different kind of sources (H II regions, AGNs) tend to cluster in a BPT diagram, depending on their emission line ratios. That is because different excitation mechanisms probe these emissions in each case. The details of the characterisation of J125323.0+215717.6 as a galaxy that hosts an active galactic nucleus are found in Appendix B.3.

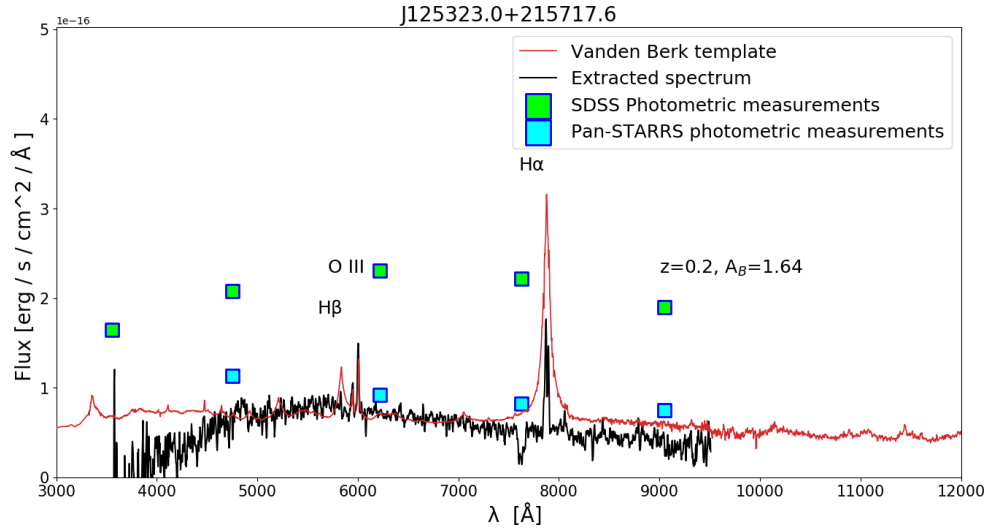


Figure 6.28: Quasar J125323.0+215717.6 observed in May with grism 4

J130504.74+301206.8

Target J130504.74+301206.8 with redshift $z = 3.8$ constitutes the most redshifted quasar that the machine learning could retrieve. The object has not been observed spectroscopically by previous surveys, which means that it could not have been part of the training set. In that sense, the machine learning could not have 'memorized' the characteristics of the source, but has rather made a genuine prediction of a very far away quasar. For wavelengths blueward to $Ly\alpha$ emission line at around 5800Å the Ly forest is clearly observable, with successive absorptions of the $Ly\alpha$ line at different redshifts. $Ly\beta$ line is also apparent at 4900Å. No extinction has been considered in the Selsing fitting of this quasar. The photomertis measurements are slightly shifted upwards to both the extracted spectrum and the template, while the carbon lines also appear weaker than the template suggests.

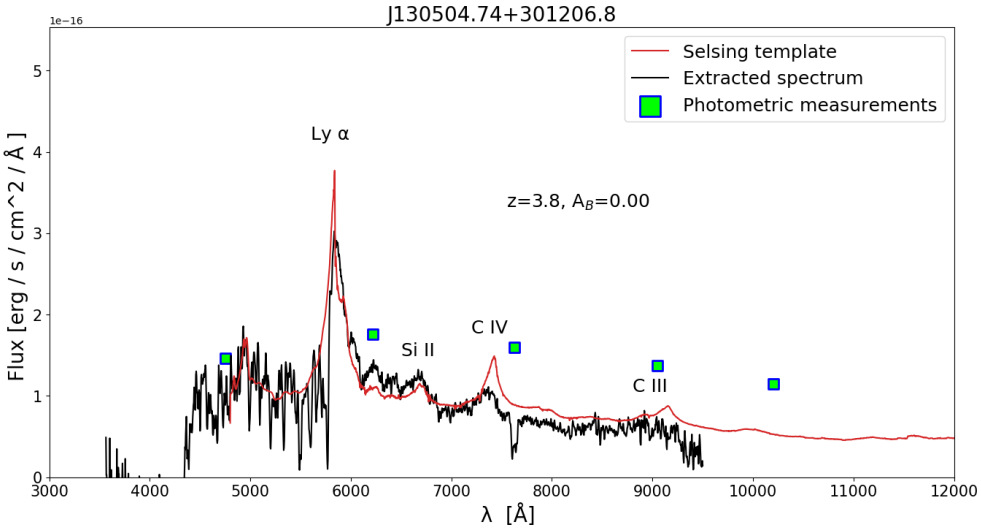


Figure 6.29: Quasar J130504.74+301206.8 observed in May with grism 4

J130448.3+354845.7

The last observation is a $z = 0.314$ quasar, for which a Selsing profile with extinction $A_B = 1.1$ seems to be a good fit, even if the redshift is smaller than 1. The fact that the Vanden Berk profile would overestimate the infrared flux is an indication that the host galaxy of this quasar does not contribute to its infrared light. The photometry seems to be off for this target, which could be attributed either to bad measurements or to a bad flux calibration of the extracted spectrum. On the left of the $H\alpha$ emission line we can see OI emission, and SII on the right side of it.

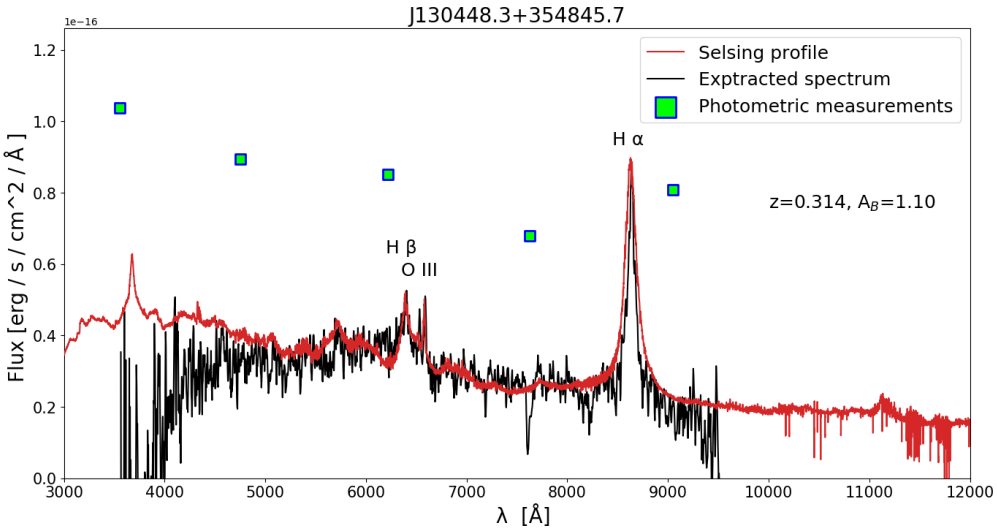


Figure 6.30: Quasar J130448.3+354845.7 observed in May with grism 4

All targets summary

The targets of all the three observing runs and their class/subclass are shown in an $(W1-W2)$ - $(g-r)$ color color plot (Fig. 6.31). All of the BAL quasars that we observed lie in the upper right corner of the plot, and so are red in the g-r ($g-r>0.98$) and in the W1-W2 ($W1-W2>0.75$). In the main quasar locus, apart from the normal and reddened quasars we also have contamination from stars (F, in this case), that were falsely predicted as quasars. Predicted targets whose positions in the color color plot are on the boundaries between the two classes, seem to carry the biggest uncertainty. Even if they were predicted as quasars, they truely are M, K or F stars. However, a few outlying quasars with $W1-W2<0.8$ are correctly predicted as such. Figure 6.32 shows how the dust extinction A_B is correlated to redshift for different types of quasars. Based on our extinction estimations by fitting the template to the spectra, we find that our most reddened targets are found in lower redshifts. That could be explained as the effect of the host galaxy's dust that absorbs effectively the bluer part of the electromagnetic radiation and re-emits it in the infrared. The nearest a quasar is to the observer, the more likely this re-emitted light to be evident in the quasar spectrum. Apart from that, the galaxy is largely populated by cooler M stars, that contribute with their light to the redder parts of the spectrum. The Seyfert galaxy that we observed in May is also found in low redshift and with pretty high dust extinction, probably for the forementioned reasons. In our very far away quasars ($z=3.8, z=5.8$), the host galaxy's light is many orders of magnitude fainter, and we fit the spectra without the need to add extinction. As a result there the reddening comes mostly as an intrinsic result of the redshift.

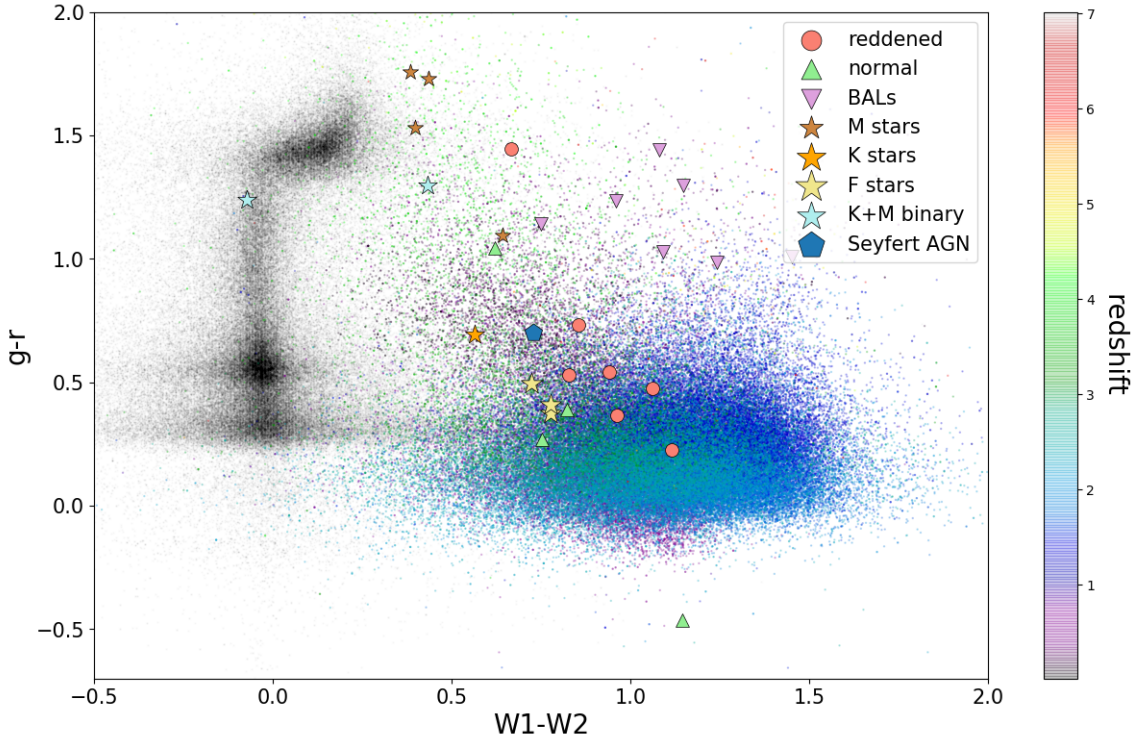


Figure 6.31: Positions of all the observed targets in a (W1-W2)-(g-r) color color plot. Stellar subclasses and quasar types, as they were determined from the data reduction, are denoted. The stellar 'tail' (black points) meets the high redshift quasar region at high y-axis values($g - r > 1$).

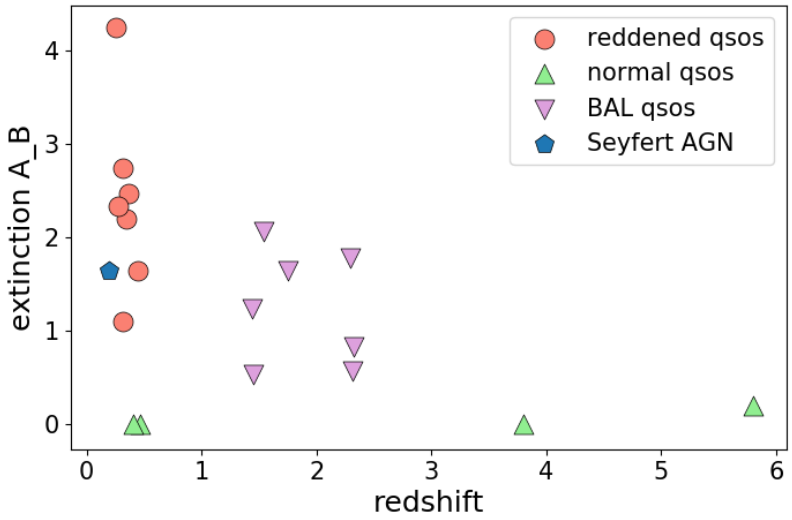


Figure 6.32: Dust extinction A_B od all the observed quasars as a function of redshift. Normal quasars appear with no or very little A_B independently of their redshift. BAL quasars are found for redshifts $1.44 < z < 2.91$ and for them we estimate high dust extinction values. The dust reddened quasars on the other hand are found in lower redshifts ($z < 0.45$) and for them we estimate very high A_B values.

In the following table we have gathered the spectral class and subclass of all our 29 observed targets. We also show how the most complete photometric/astrometric ML model and the multiclass stellar model perform when they make predictions on the targets.

Target	Observing Period	Real Class	Real Subclass	Combined model's ML Predictions	Stellar Multiclass ML Predictions
J022742.93-173121.5	August	QSO	LoBAL	QSO	QSO
J010013.02+280225.8	August	QSO	normal	QSO	LATE M STAR
J234530.36-135743.3	August	QSO	LoBAL	QSO	QSO
J000404.29-135905.43	August	QSO	BAL	QSO	QSO
J234859.52-193324.89	August	QSO	normal	QSO	QSO
J003634.80-140924.9	August	QSO	BAL	QSO	QSO
J010339.41-132238.91	August	QSO	BAL	QSO	QSO
J012813.63-120319.79	August	QSO	LoBAL	QSO	QSO
J012534.07+351344.0	December	QSO	red	QSO	QSO
J012848.44+341704.9	December	QSO	red	QSO	QSO
J013121.61+444513.1	December	STAR	M6+WD	STAR	EARLY M STAR
J013232.61+444123.0	December	STAR	M6+WD	STAR	EARLY M STAR
J015455.79+203623.8	December	QSO	BAL	QSO	QSO
J015748.65+284752.6	December	QSO	red	QSO	QSO
J025004.61+324039.7	December	QSO	red	QSO	QSO
J025111.38+321221.5	December	STAR	M3+WD STAR	STAR	EARLY M STAR
J025832.77+354254.6	December	QSO	red	QSO	QSO
J130703.91+251415.5	December	QSO	red	QSO	A STAR
J123022.68+191340.0	May	STAR	F3	STAR	F STAR
J122153.19+235324.3	May	STAR	M2	QSO	QSO
J123654.12+321711.6	May	QSO	normal	STAR	A STAR
J130550.86+230841.4	May	STAR	K5+M0	STAR	K STAR
J131925.63+260504.6	May	STAR	K5+M0	STAR	K STAR
J123504.03+202301.6	May	STAR	F8	STAR	F STAR
J125323.0+215717.6	May	GALAXY	Seyfert AGN	GALAXY	QSO
J121222.63+272609.8	May	STAR	K0	STAR	F STAR
J123430.94+213600.7	May	STAR	F8	STAR	F STAR
J130504.74+301206.8	May	QSO	normal	QSO	QSO
J130448.3+354845.7	May	QSO	red	QSO	QSO

Chapter 7

Exploring the Gaia Survey

7.1 Early Data Release (EDR3) Database - Unseen Data

All the observed sources from Gaia EDR3 (that satisfy the condition $g_{mag} < 20$, within 10 degrees around the North Galactic Pole are SQL queried[1] as shown below:

```
SELECT  gaia_source.ra,gaia_source.dec,gaia_source.parallax,gaia_source.parallax_over_error,gaia_source.pm,
gaia_source.pmra_error,gaia_source.pmdec_error,gaia_source.phot_g_mean_mag,gaia_source.phot_bp_mean_mag,
gaia_source.phot_rp_mean_mag,gaia_source.l,gaia_source.b

FROM gaiaedr3.gaia_source

WHERE (gaiadr3.gaia_source.b BETWEEN 80 AND 90 AND gaiadr3.gaia_source.phot_g_mean_mag<=20)
```

Figure 7.1: SQL query for downloading all the observed objects from Gaia EDR3 survey, 10 degrees around the NGP, with g< 20.

The total number of observations is 494.777 sources. After the cross-match with AllWISE survey we are left with 393.085 objects. Another cross-match with the SDSS DR16 further reduces the catalogue by around 10 thousand observations. In the following plots we present the distribution of sources that are found 10° around the NGP. Their division in groups is according to the predictions of the model that uses atrometric and photometric observations from Gaia. Using the original catalogue from Gaia (494.777) we make a quasar candidate list of 31.665 objects and a galaxy candidate list of 69.351. The rest of the objects, (393.761 in number) are predicted as stars (left subfigure 7.2). However, we noticed that an excess amount of quasars and galaxies have $S/N_{pm} > 3$ (10.974 predicted quasars and 50.952 galaxies), and thus we do not consider them as trustworthy predictions. The final catalogues, after removing some extra NaN values, are comprised of 16.270 quasars and 15.512 galaxies. Right subfigure 7.2 shows the more reliable cut-offs for the quasar and galaxy populations. It should be noted that stars are also found with low signal to noise ratios (932 objects for $S/N_{pm} < 3$). Those are the stars that would be mislabeled as quasars by a simple empirical proper motion cut-off. The SDSS cross-match provides the known spectral class for some of these sources. Namely, 2719 quasars, 418 galaxies and 4354 stars have already been observed and can act as cross-validation for the ML predictions. To be specific, 2700/2719 quasars, 375/418 galaxies and 3993/4354 stars are indeed correctly predicted (lower panel of figure 7.2c). The remaining

[1] https://gea.esac.esa.int/archive/

19 known quasars, are misclassified and predicted as 15 galaxies and 4 stars. This cross-validation also informs us that

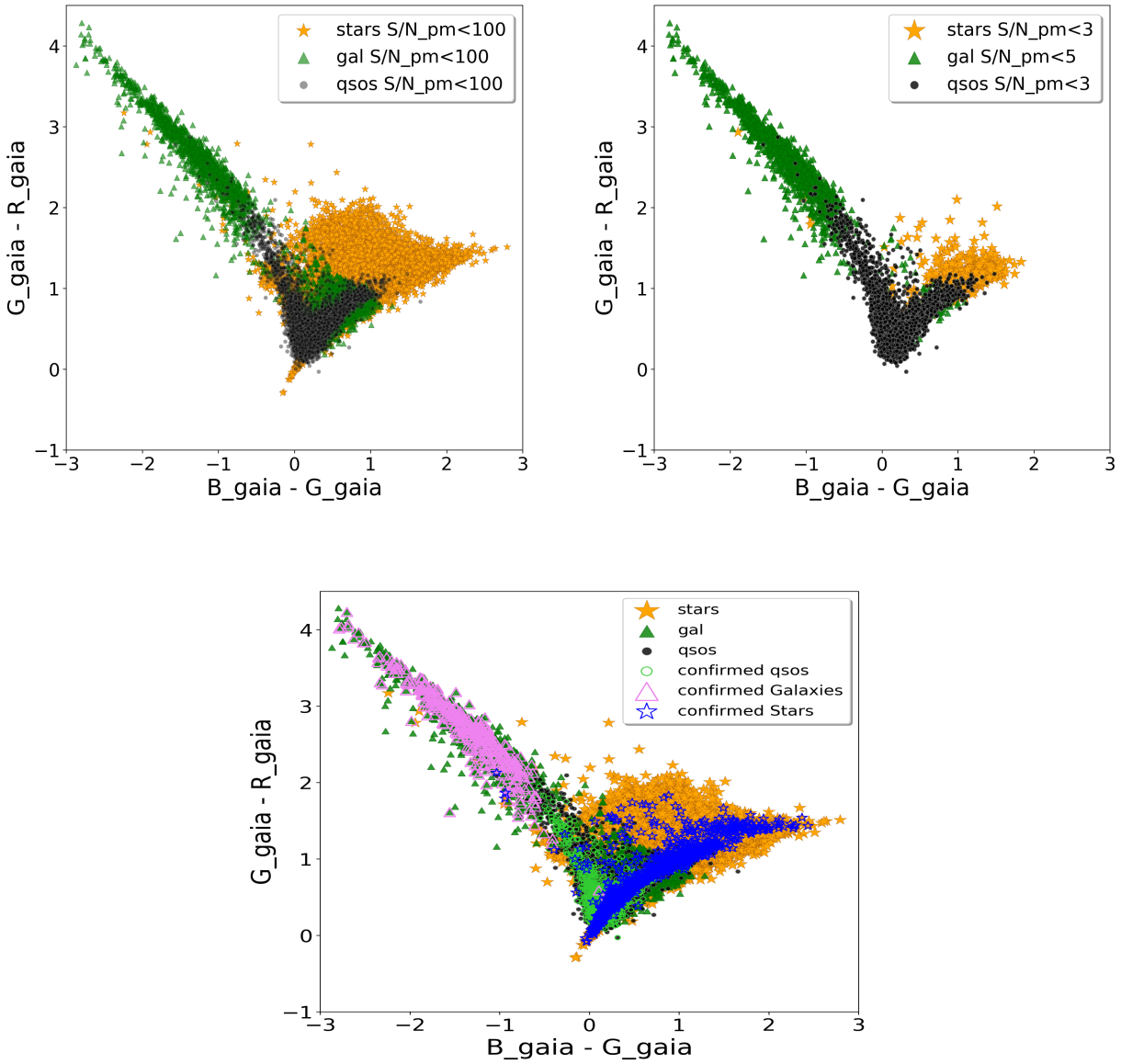


Figure 7.2: Left: g-r vs. b-g color color plot of the sources 10° around the NGP. Class label is assigned by the 3 classes, astrometric + photometric ML model based only on Gaia measurements. Right: Subgroups of the candidate lists after the application of the S/N_{pm} cut-off. Lower panel: Spectroscopically known sources from the SDSS DR16 survey which are correctly classified by the ML model are overplotted on the overall predictions. The majority of known galaxies forms the galaxy "tail" on the upper left corner of the plot. In the overlapping region for b-g > 0 the ability of the model to discriminate stars and quasars is evident from the amount of confirmed sources.

the quasar candidate list is contaminated by 57 stars predicted as quasars. If we make an assumption that these numbers from the cross-validation are representative of the whole population, then the purity of the quasar class is 97.19% and the completeness is 99.30%. However, if we include the $S/N_{pm} < 3$ cut-off, the cross-validation

from SDSS shows that the contamination from stars drops to 14, reaching a purity of 98.7%. As for the contamination by galaxies, it accounts for 21 galaxies predicted wrongly as QSOs.

7.1.1 Surface Density of QSOs, Galaxies and Stars

Followingly, in order to make bigger quasar catalogues and to make an estimation of the quasar surface density in different galactic latitudes, we query from Gaia EDR3 all the available data for 4 more galactic latitude bins, of 10 degrees each. For $70° < b < 80°$ there are 1.601.149 sources, for $60° < b < 70°$ there are 2.914.974 sources, for $50° < b < 60°$ there are 5.886.165 sources and for the last bin $40° < b < 50°$, we find and query 8.118.647 objects. The predictions of our model can be seen in Table 7.1.

Table 7.1: Left column: number of available data from Gaia EDR3 survey for each galactic latitude bin, and for G<20. The rest of the columns show the division of the whole queried datasets into class according to the ML predictions.

b	DATASET	QSOs	**S/Npm<3**	GALAXIEs	**S/Npm<3**	STARs	**S/Npm<3**
$90° < b < 80°$	494.777	31.665	**12.215**	69.351	**1.463**	393.761	**1120**
$80° < b < 70°$	1.601.149	117.461	**40.061**	226.231	**7.367**	1.257.457	**3064**
$70° < b < 60°$	2.914.974	185.980	**64.765**	393.823	**8.853**	2.335.171	**6481**
$60° < b < 50°$	5.886.165	330.284	**114.652**	359.270	**14.650**	4.742.566	**16.107**
$50° < b < 40°$	8.118.646	359.270	**140.899**	1.065.044	**11.776**	6.534.184	**24.498**
Total	19.072.341	1.021.660	**372.592**	2.113.719	**44.109**	15.263.139	**98.379**

This increase in the total amount of objects for lower and lower latitudes is the product of two factors. Firstly, the area 10° around the North Galactic Pole is not as extended as the the surface area 10 degrees near the equator (galactic plane). In fact, as can be seen from eq. 7.1, the surface area (in square degrees) starts from 313 for 10° around the North Galactic Pole and reaches 2.542 for galactic latitude $40° < b < 50°$. The second reason for the increase of the total number of sources is that the stellar population increases as one moves towards the galactic plane. This is reflected in Fig. 7.3, where the quasar surface density remains approximately the same as we move to smaller galactic latitudes, while the stellar surface density increases. The galaxy surface density also appears to have a significant increase in lower latitudes, but in physical interpretation this result could not stand. It is expected for an isotropic and homogeneous universe to have a uniform distribution of quasars and galaxies, but the stellar population in our data is indeed affected by the intervening Milky Way. The reason why we have this upward trend also in the galaxy class, must be an overestimation of the galaxy population by the ML model. To make the situation more clear, we have added errorbars for the galaxy class in Fig. 7.3. They are calculated by counting the predicted galaxies with $S/N_{pm} > 3$, which we do not trust as valid predictions. The subtractions of the errorbars from the overall galaxy population, would give a constant galaxy surface density. The integral that calculates the surface of a sphere in spherical coordinates r, θ, ϕ, where $0 < \theta < \pi$ and $0 < \phi < 2\pi$, is:

$$S = \int_0^{\pi} \int_0^{2\pi} r^2 sin\theta \; d\theta \; d\phi \tag{7.1}$$

For the surface area of this hypothetical sphere to be in square degrees, the radius r has to be in degrees, $r = 57.29577951308232° = 180/\pi = 1/rad$. This comes from the simple fact that $2\pi r = 360° \rightarrow r = 180°/\pi$. Applying eq. 7.1, we find that the total area of the celestial sphere is 41252.96124941927 deg^2. Solving the same integral for each galactic latitude bin, we find the surface area to be:

- $80° - 90°$ $(170\pi/180 - \pi)$:

$$S = \int_{17\pi/18}^{\pi} \int_{0}^{2\pi} r^2 sin\theta \; d\theta \; d\phi \; = 313.36 \; deg^2$$

- $70° - 80°$ $(160\pi/180 - 170\pi/180)$:

$$S = \int_{8\pi/9}^{17\pi/18} \int_{0}^{2\pi} r^2 sin\theta \; d\theta \; d\phi = \; 930.56 \; deg^2$$

- $60° - 70°$ $(150\pi/180 - 160\pi/180)$:

$$S = \int_{5\pi/6}^{8\pi/9} \int_{0}^{2\pi} r^2 sin\theta \; d\theta \; d\phi = \; 1519.49 \; deg^2$$

- $50° - 60°$ $(140\pi/180 - 150\pi/180)$:

$$S = \int_{7\pi/9}^{5\pi/6} \int_{0}^{2\pi} r^2 sin\theta \; d\theta \; d\phi = \; 2062.25 \; deg^2$$

- $40° - 50°$ $(130\pi/180 - 140\pi/180)$:

$$S = \int_{13\pi/18}^{7\pi/9} \int_{0}^{2\pi} r^2 sin\theta \; d\theta \; d\phi = \; 2542.35 \; deg^2$$

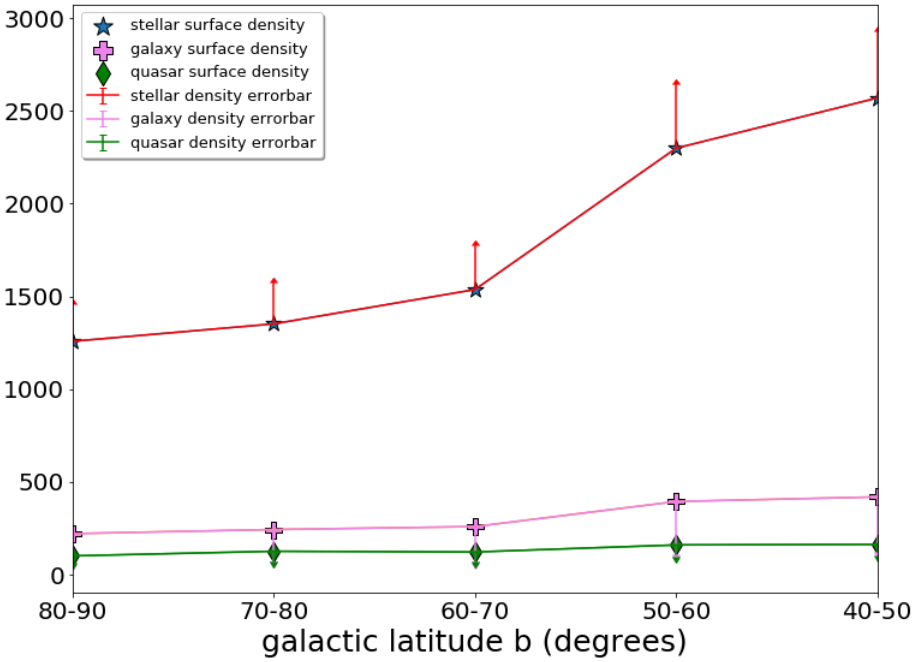


Figure 7.3: Surface density of stars, galaxies and quasars as function of the galactic latitude bin. Stellar number density demonstrates a more rapid increase as we move towards the galactic plane. It is expected that for even lower latitudes the density will increase more abruptly. Quasars and Galaxies appear also with a small increase but we argue that this is due to an overestimation of their populations.

Table 7.2: Surface Density of QSOs, Galaxies and Stars

b	QSOs	**S/Npm<3**	GALAXIEs	S/Npm<3	STARs
$90° < b < 80°$	101 deg^{-2}	**39** deg^{-2}	222 deg^{-2}	4.7 deg^{-2}	1258 - 1450 deg^{-2}
$80° < b < 70°$	126 deg^{-2}	**43** deg^{-2}	243 deg^{-2}	7.9 deg^{-2}	1352 - 1600 deg^{-2}
$70° < b < 60°$	122 deg^{-2}	**42** deg^{-2}	259 deg^{-2}	5.8 deg^{-2}	1537 - 1800 deg^{-2}
$60° < b < 50°$	160 deg^{-2}	**55** deg^{-2}	394 deg^{-2}	7.1 deg^{-2}	2300 - 2663 deg^{-2}
$50° < b < 40°$	141 deg^{-2}	**55** deg^{-2}	418 deg^{-2}	4.6 deg^{-2}	2570 - 2900 deg^{-2}

In Fig. 7.4 it is shown the surface density of quasars at given magnitudes in the $G-band$ from Gaia. For $G < 16$ the surface density of quasars is around 0.02 QSO per deg^2 while for $G < 18$ the density becomes $1QSO/deg^2$. Including the full range of the G-band the density is $50QSO/deg^2$ or $38QSO/deg^2$ depending whether we consider the qso list with 100% or 75% purity respectively.

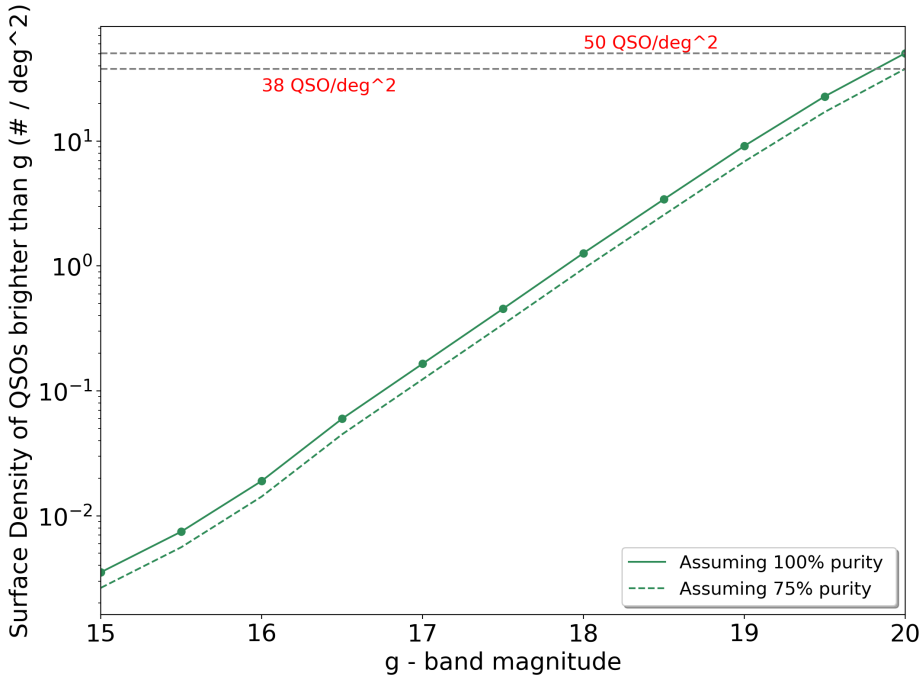


Figure 7.4: Surface density of quasars versus the given cut-off in the G-band from the Gaia survey. Including the full range of the G-band the surface density is 50 QSO/deg^2 or 38 QSO/deg^2 depending whether we consider the qso list with 100% or 75% purity respectively. The justification for the 75% purity comes from an extra quasar estimation by using the all photometric+astrometric model on a subset of the initial qso list.

The quasar surface density demonstrates a slight upward trend as we move towards the galactic plane, from 39 qsos deg^{-2} to 55 qsos deg^{-2}, which corresponds to a 41 % increase. There is no physical reason for this trend to be observed, thus we therefore conclude that even with the S/N_{pm} cut-off we still get contamination from stars in the predicted quasar catalogues. As a result, we suggest to classify those objects with predictions based on the full photometric ML training. By cross-matching the $\sim$ 372.592 predicted quasars with SDSS, AllWISE and UKIDSS surveys, their number is reduced to 57.200 objects ($\sim$ 15% of the initial set). A (g-r) vs (J-K) color color plot indeed reveals that a percentage of them clearly forms the well known stellar locus. Subfigure

7.5(a) shows the predicted quasars from the ML predictions based only on Gaia features, and highlights the weakness to correctly discriminate the sources. In subfigure 7.5(b) the ability of the infrared ML model to identify different kind of sources is shown. Based on this classification, we find that 74% of the 57.200 objects are again predicted quasars and 1% as galaxies. The rest 25% are now predicted to be stars. Our most reliable quasar candidate catalogue therefore contains 42.328 objects.

To continue with our estimation, we make the assumption that these 25% and 1% analogies also hold for the quasar candidate catalogue before it suffered the cross-match reductions (372.592 objects). Thus, 93.148 predicted quasars should have been labeled as stars and 3.725 as galaxies.

That brings us to the conclusion that for :

- $40° < b < 90°$
- $G < 20$
- $S/N_{pm} < 3$

the revised population percentages are 53.6% quasars, 37.2% stars and 9.2% galaxies. The corresponding quasar (mean) surface density is: **37 qsos per square degree** for an area 50 degrees around the NGP.

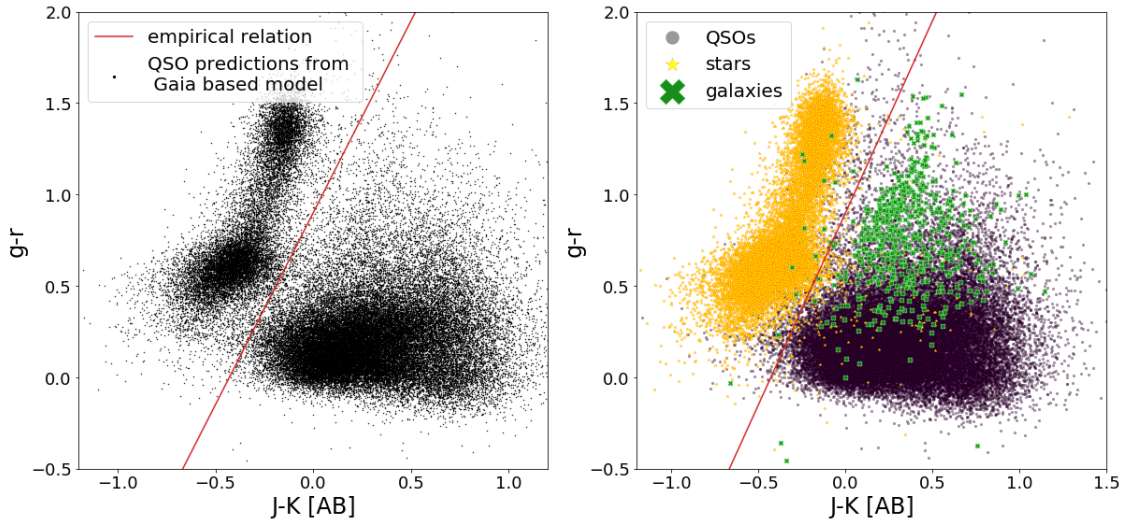


Figure 7.5: The initial Astrometry+Gaia-colors model selected 372.592 objects as quasar candidates for $40° < b < 90°$. A cross-match of this catalogue with AllWISE and UKIDSS reduced it to 57.200 objects. Left: Color-color plot (g-r vs. J-K) of the diminished catalogue. It is shown that some of the predicted quasars clearly compose the stellar locus. Right: Predicted objects of the all photometric+astrometric ML model. This model is able to identify and distinguish the objects into stars and galaxies that the initial one predicted as quasars. In particular, 14.300 objects are now predicted as stars (25%) and 572 (1%) as galaxies.

7.2 Redshift predictions

In this section we show how we predict photometric-redshifts using the XGBoost regression model described in chap. 4.4). The redshift predictions are made our final quasar candidate catalogue from the Gaia EDR3 (referring to the quasar locus of subfigure 7.5). We remind that, the regression model is trained only on photometric features and specifically on SDSS, WISE and UKIDSS photometry.

Table 7.3:

Galactic latitude (b)	QSOs (S/Npm<3)	After Cross-matches	After drop-duplicates	After drop-NaN	Mean Predicted Redshift
$90° < b < 70°$	52.276	57.835	15.508	13.423	1.71 ± 0.36
$70° < b < 60°$	64.765	102.057	20.591	17.625	1.83 ± 0.36
$60° < b < 50°$	114.652	109.112	14.879	11.355	1.96 ± 0.36
$50° < b < 40°$	140.899	90.853	18.053	14.797	2.04 ± 0.36
Total	372.592	359.857	69.031	57.200	

In total, we make redshift estimations for 57.200 predicted quasars. After the cross matches and dropping the duplicates we lose about 81% of the initial observations. Followingly, we remove the NaN values and the catalogue is reduced by another 4% of its initial length. The results are pretty satisfactory; the known pattern of quasar redshift distribution in a gr - JK plot appears. The main body of the quasar cluster is characterised by intermediate redshifts ($0.5 < z < 1.5$). The low redshift region ($z < 0.5$) is correctly predicted to frame the right, red in th JK part of the main quasar locus. The high redhift quasars, on the other hand, are predicted to gradually climb the upper left side of the plot, a bit rightward to the expected position of the stellar locus (which is not plotted in Fig. 7.6). Many interesting quasar candidates are brought up this way, as the regression model assigns redshifts higher than 3 to 1550 objects, out of which 624 are already observed and classified by the SDSS.

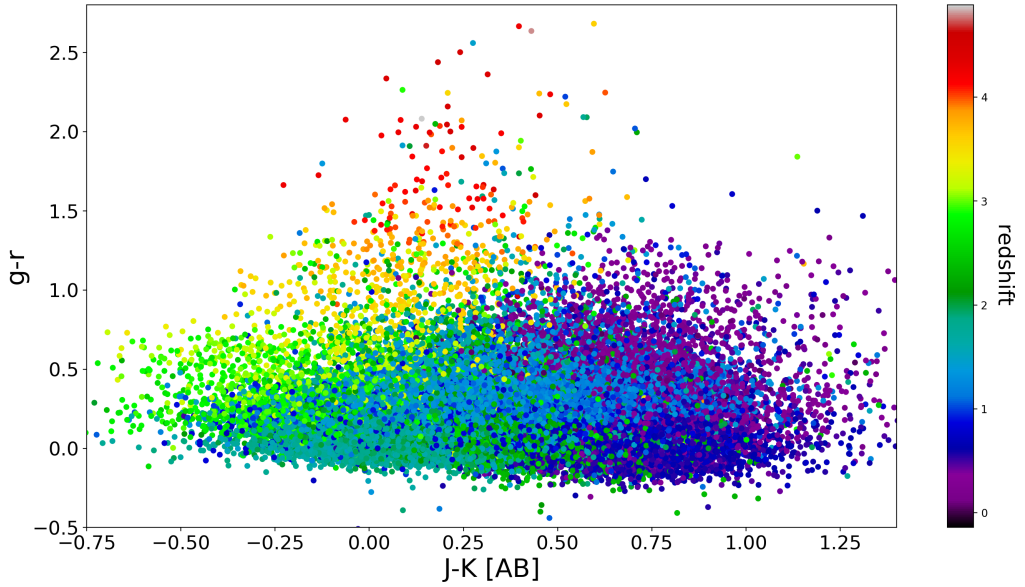


Figure 7.6: Color- color plot of the photometric-redshift predictions from the XGBoost Regression model. The model predicts 78 high redshift quasars ($z > 4$), out of which 42 are cross-validated with the SDSS DR16. A slight overestimation of the redshifts in the blue part of the diagram is noticed, but could be neutralised if the margin of errors (± 0.28 for this regression) is considered.

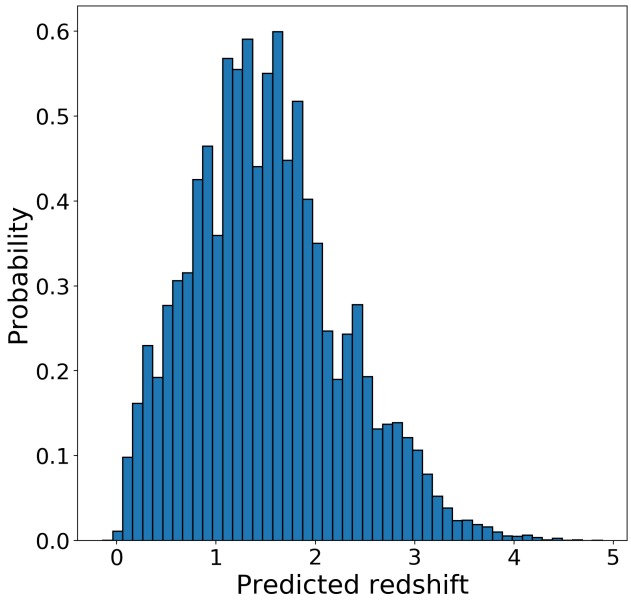


Figure 7.7: Histogram of the predicted redshifts distributions. The peak occurs at $z = 1.51$, in agreement with the redshift distributions of the training dataset's quasars. The probability of having a quasar with redshift $z > 3$ follows a reasonable drop, given the fact that they are harder to detect and observe. The XGBoost Resgression model is trained to predict up to redshift $z = 4$, as the results are ambiguous for higher values. Maximum predicted value is $z = 4.88$. The quasar that was assigned this value happens to be spectroscopically observed by the SDSS, where we get its real redshift from :$z = 4.48$. So if we also take the error into account, the prediction is very close to the real value. What is for sure therefore is that with the XGBoost regression we may not be able to accquire the exact redshift, but we can get an idea of the redshift bin where the object lies. That could provide a good selection tool for distant quasars that ones want to detect and investigate.

Part IV

Conclusion and Future Work

This work constitutes an investigation of how machine learning can be effectively used in the quasar selection process. This exploration gave birth to a plethora of XGBoost ML models trained on different survey's features, from optical to near and mid-IR, as well as on astrometric. Each model can be used based on the available data one has and wants to make a prediction on an object. Our work specified the weights of the features used for each training, with respect to their importance for the classification. Among them, the models that utilize both astrometric and photometric observations for the training stand out. Until very recently, all the models that were trained in order to solve the classification problem for quasars / stars relied only on the photometry of the sources. Astrometric machine learning training was a field that had not been investigated before, even though empirical criteria had been proposed. On the 13_{th} of June 2022, though, ESA's Gaia Collaboration published the DR3 Data Release. In their paper they describe the results of using Gaia Satellite's observations in a machine learning model, in order to classify extragalactic objects they had measured. In our work we followed similar process, with the difference that we also used additional magnitudes in longer wavelengths. To our knowledge, and as it turns out from the work we did in this thesis, this addition only improves the predictions and generates purer quasar catalogues. The ability of machine learning to also predict outliers, such as high-redshift, dust-reddened and Broad Absorption Line quasars, is also tested and quantified through direct observations. Based on these observations, we argue that the field of machine learning can automatize the quasar selection and even improve it, incorporating the former selection techniques in its core, but using them in a more dynamical way. With that, it is meant that even though we 'feed' the model with colors, magnitudes, parallaxes and proper motions, we avoid passing simultaneously all the biases that come with them. This is possible because, during the training the model recognises patterns and relations between the variables in a more flexible way, than the stiff formulations of empirical cut-offs and criteria can describe. Throughout the different validation approaches, we observe this trend; the ML models retrieve quasars that would have evaded other selection techniques, increasing the completeness, without necessarily decreasing the purity.

In no case does that mean that we discovered the holy grail for selecting quasars. As we showed, the models actually suffer from inevitable errors that are induced to them through the training data; although the surveys become more and more accurate in their measurements with time, the instruments are characterised by a limited accuracy. We believe that more accurate kinematic measurements in the future could lead to a training catalogue that would be as error-free as possible. Ultimately, leading to an even better quasar selection process. The same arguments hold for the color surveys as well. The reasons why we prioritise the creation of a more astrometrically acccurate training dataset are based on simple getaway messages we took from our experience while working on the topic this year. Firstly, the proper motion signal to noise ratio is outstandingly the most crucial feature among all the photometric and astrometric ones, as we show in the feature importance plot **??**. Proper motion is thus desired to be as a precise measurement as it can be. In fact, if the proper motion could be completely errorless, then distinguishing extragalactic objects from stars would require just a simple cut-off. Secondly, by only using Gaia's observations, the length of lists that one wants to test is not sacrificed. As an all-sky survey, with the exception of galaxies that do not appear as point sources, Gaia has available observations for most galactic stars and bright enough quasars. Lastly, as Gaia Mission has scanned a huge area of the sky, allows also for surveying the least explored, southern hemisphere. The cut-off on the brightness of selected objects is another aspect of the work that could take place in the future. For the purposes of this work, we focused only on sources that are brighter that 20 in the G band. This choice is made to limit the errors in the measurements, but left out fainter objects that could prove to be interesting quasars.

The construction and application of a Regression model, in order to find the photometric redshifts of the

predicted quasars, is another important outcome of our work. With an upper limit of $z = 5$ in the training set, this model can predict a quasar's redshift within a ± 0.28 margin of error. For this thesis' purposes we did not dive deep into the details of the regression process, we however spotted errors in the SDSS redshift estimation, with a tendency of the pipeline to assign higher redshifts than the real ones to quasars with $z > 5$. Left for future work, we encourage a more profound cleaning of the spectroscopic data used for the training, since not all of them are to be trusted. The classification and regression tools combined, can provide a handy way to easily classify astronomical sources, detect the quasars among them, and sort them into redshift bins. In that scheme, quasars in different redshifts can be insightfully pinpointed and be selected for observations. This would allow their distant epochs to be studied and the interesting physics that characterize them to be unraveled.

Finally, in order to eliminate the errors in the training sets while providing a plethora of different quasar photometric characteristics, we propose a future machine learning training on simulated data. Using the SEDs of any quasar type, we could generate artificial photometry at any redshift and for any extinction. The fluxes in optical and infrared bands could then be integrated and translated to magnitudes, that could serve as input variables of a machine learning training. The problem in applying such a method could perhaps - and ironically enough - be its intrinsic flawlessness. How good a model can be, if it is trained on errorless data and asked to predict on real observations with errors? This is why a hybrid method of both simulated and real data could possibly provide the optimal solution.

All the above are the modifications we propose for future machine learning models that rely on photometric and astrometric data, for classification and regression purposes. With the operation of ESO's spectroscopic facility, the 4-meter Multi-Object Spectroscopic Telescope (4MOST) in a couple of years, the field of machine learning will find another useful application in quasar astrophysics. 4MOST will be designed to simultaneously acquire spectra of ~ 2400 objects, reducing significantly the data collection time and largely enriching our quasar spectral catalogues. Hopefully, we will then reach a point where after 60 years of contemplating the quasar selection techniques, terms as color bias for selecting interesting quasars could turn obsolete and astronomers will be puzzled to select the interesting ones out of a continuously growing spectral database. In such a scenario, machine learning could play a supportive role to the observations. For example, a neural network trained on quasar spectra could be then used to automatically extract important information and characteristic features of these, that would otherwise require a time consuming manual inspection.

New and more efficient ways into exploring the Universe can be achieved with the use of Machine Learning. This work is relevant for many future surveys instead of 4MOST, such as the Euclid, MOONS and many more. Euclid is a cosmology survey mission, optimised to determine the properties of dark energy and dark matter on universal scales. Euclid will take images in optical and near-infrared light; these images will eventually cover more than one-third of the night sky, and depict billions of cosmic targets out to a distance up to 10 billion light years. MOONS stands for Multi-Object Optical and Near-IR Spectrograph and is going to be the new next generation spectrograph for the Very Large Telescope (VLT).

Part V

Appendices

Appendix A

Observational Surveys - Data Extraction

Appendix A presents the surveys from which the data were collected and the process of gathering all the data that were used for training the various machine learning models.

A.1 Sloan Digital Sky Survey: SDSS

The Sloan Digital Sky Survey (SDSS) [1] is an optical survey dedicated to collect both photometric and spectroscopic data, having covered more than 1/3 of the entire sky. It utilizes the Sloan Foundation 2.5 m wide-angle telescope built at the Apache Point Observatory, in New Mexico. The imaging camera collects photometric data in five filters and consists of thirty, 2048 by 2048 pixel CCDs, arranged in six columns [Gunn et al., 1998]. Each array corresponds to a different filter, namely the u, g, r, i, z with effective wavelengths 3550 Å, 4770Å, 6230Å, 7620Å, , 9130Å and respectively (Fig A.1). A follow-up spectroscopic survey used those photometric data, with the SDSS spectrographs being able to extract multiple spectra in a single observation (640 spectra at a time with the SDSS spectrograph and 1000 with the updated BOSS). The spectrograph uses an aluminum plate with 1.000 pre-drilled holes, with each hole corresponding to a specific object such as a quasar, a galaxy or a standard star. Followingly, 1.000 optical fibers with 2 arcsec diameters are plugged into the holes. The fibers send the light from each object through a beamsplitter with a coating that reflects the blue part of the spectrum while allowing the red part through. It has a wide wavelength coverage between 3600Å and 10,400 Å, while its spectral resolution, defined by the fraction $R = \frac{\lambda}{\Delta\lambda}$, takes the value range 1560-2270 in the blue channel and 1850-2650 in the red channel.

SDSS SQL Search Tool

For our aims, we queried and used data from the full photometric (PhotObjAll) as well as the full spectroscopic (SpecObjAll) catalogues. To do so, we used structured query language (SQL) and the online SQL Search Tool [2]. The joined catalogues were matched based on each object's ID, with the requirement that the zWarning equals zero - an indication that no problems were detected to secure the data quality. Out of the many quantities listed in the photometric catalogue, we chose to download the galactic coordinates b (galactic latitude) and l (galactic longitude), the magnitudes in each of the five SDSS filtres and their corresponding errors. As for the spectroscopic catalogue, it was used to obtain the equatorial coordinates ra and dec, the redshift with its error, but most importantly - the spectroscopic class of all the objects. This feature will serve as the 'truth'/target

[1] The SDSS survey web page: https://www.sdss.org
[2] The SDSS SQL Search Tool found at http://skyserver.sdss.org/dr16/en/tools/search/sql.aspx

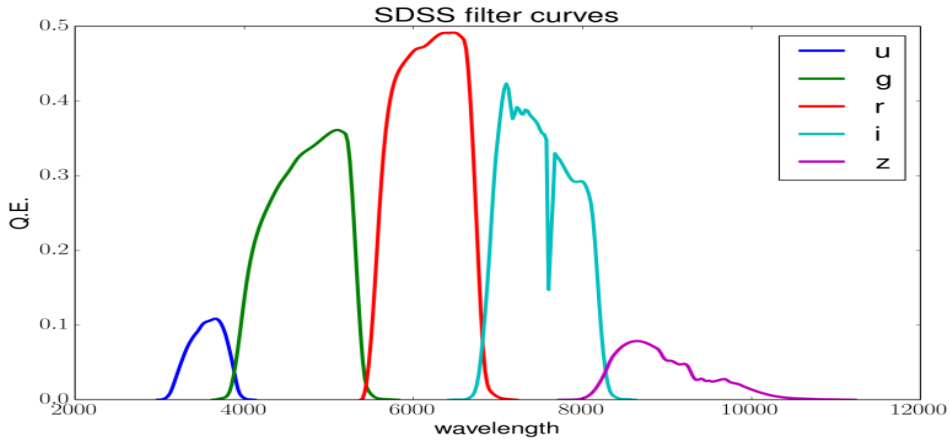


Figure A.1: Filter response curves for the u g r i z bands of the SDSS camera

variable for the supervised machine learning. With this procedure we managed to obtain 751.160 quasars with redshifts between 0.056 and 7.64, 2.548.984 galaxies and 956.010 stars.

A.2 Wide-Field Infrared Survey Explorer: WISE

The Wide Field Infrared Survey Explorer is a near Earth space telescope, with a diameter of 40 cm. Constructed to perform mid-infrared astronomical survey, the instrument scans the sky in four bands, W1, W2, W3, W4, with effective wavelengths 3.4 μm, 4.6 μm, 12 μm and 22 μm respectively (Fig A.2). The angular resolution at each of the wavelengths mentioned is 6.1", 6.4", 6.5" and 12.0". The scientific importance of the survey in the infrared, is underlined by the wide range of objects that are studied in those bands and the beautiful physics that emerges. From faint, cool brown dwarfs emitting in the μm, to the detection of thermal radiation from asteroids. From the role of interstellar dust and dust in star-forming regions, to young stars and debris disks [Wright et al., 2010]. Of great value for the purposes of this thesis is the contribution of the survey to the

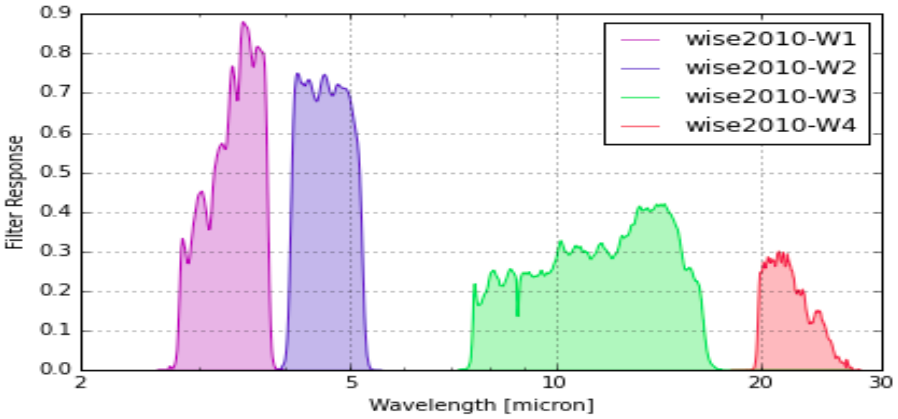


Figure A.2: Filter response curves for the W1 W2 W3 W4 infrared bands of the WISE

detection of the obsured type II AGN/QSO population that would evade the surveys in the optical emission lines and UV excess. It is therefore a significant leap of progress towards making a complete census of the quasar population. As a result, it was necessary to include this set of infrared photometry to our initial SDSS

catalogues.

CDS-XMatch : SDSS × AIIWISE

One useful tool to achieve this extend is the CDS-XMatch Service [3]. Once the .csv files containing the SDSS stars, quasars and galaxies are uploaded, the web page offers a variety of surveys available for cross-matching. The cross-match criteria are set by the position of the objects (the coordinates ra and dec). More specifically, our SDSS catalogue inputs' coordinates were matched with those of the AIIWISE catalogue, within a margin of error of 1 arcsec. This small value of margin of error was selected on purpose to eliminate the appearance of duplicate entries. An attentive inspection of the merged dataset that was created this way indeed revealed the absence of duplicate rows with respect to their coordinates. This means that for the objects, there were no more than one nearest neighbour, within the radius of 1 arcsec.

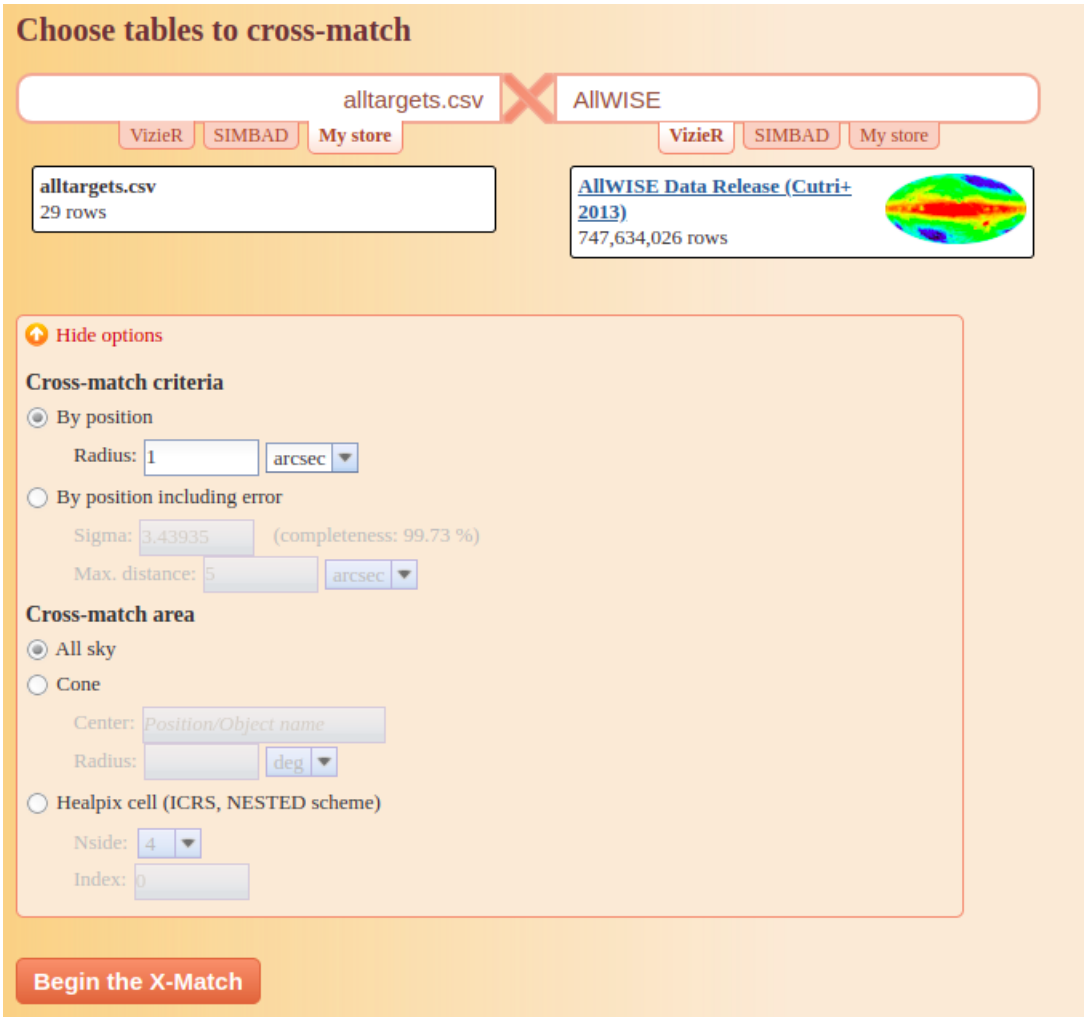
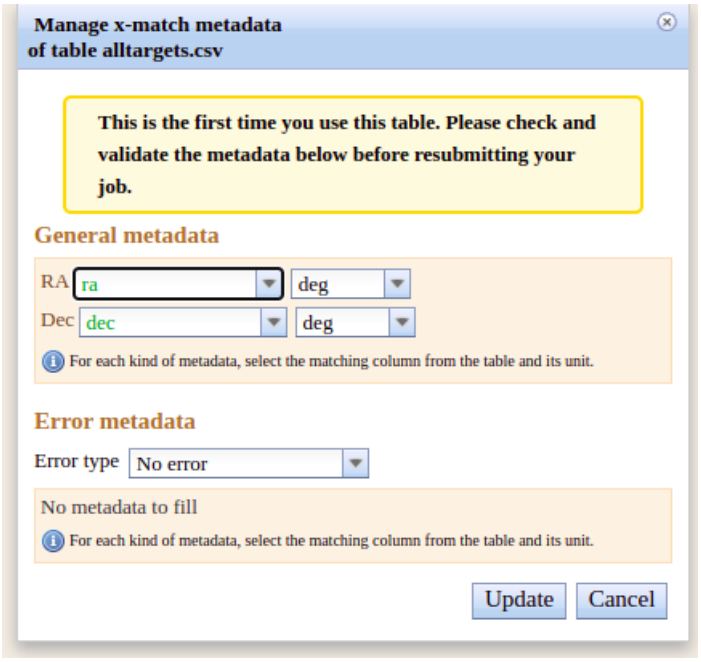


Figure A.3: CDSX-Match tool, where two any surveys can be cross-matched to enrich the measurements on the given astronomical objects. There is also the option to upload '.csv' files one has made to 'My store' and combine them to any other survey. As cross-match criteria we chose the position, within a radius of 1 arcsec.

[3] CDS-XMatch Service web page : http://cdsxmatch.u-strasbg.fr



A.3 UKIRT Infrared Deep Sky Survey: UKIDSS

The UKIRT Deep Sky Survey is a deep near-IR survey carried out by the United Kingdon Infrared Telescope (UKIRT) [4], based on Mauna Kea in Hawaii. The scientific goals of the survey mainly focus on the study of near brown dwarfs and galaxy evolutionary stages, by looking back to the local extragalactic universe. It also explores high-redshift galaxies and galaxy clusters, as well as the observation of the highest-redshift quasars, at $z \sim 7$, with long exposure times in the infrared [Lawrence et al., 2007]. The inclusion of these near-infrared magnitudes aims to not leave obscured, dust reddened or highly red-shifted quasars out of the selection, such that their population is not underrepresented.

CDS-XMatch : SDSS / AIIWISE × UKIDSS

For our analysis we used the UKIDSS DR9 LAS (Large Area Survey) data release, which is the most recent data release available on the CDS-XMatch for cross-match. It is also the one compatible with the SDSS areas, as it covers a large section of the Northern sky, for which spectroscopic and photometric data were already obtained. The Large Area Survey covers an area of 4028 square degrees, in four near-infrared bands. The survey is conducted with the use of the Wide Field Camera (WFCAM) which has a field of view of 0.21 square degrees. Those broad band filters are the Y (1.0μm), J(1.2μm), H(1.6μm) and the K(2.2μm) to a depth in the K band of 18.4 (Vega system). Their response curves are shown in (Fig A.4). The merged and processed SDSS/AIIWISE catalogue is uploaded to the CDS-XMatch Service and is cross-matched, under the same conditions (SDSS ra, dec coordinates, within a 1 arcsec mismatch) with the UKIDSS DR9 LAS catalogue.

[4] UKIRT Survey web page: https://about.ifa.hawaii.edu/ukirt/

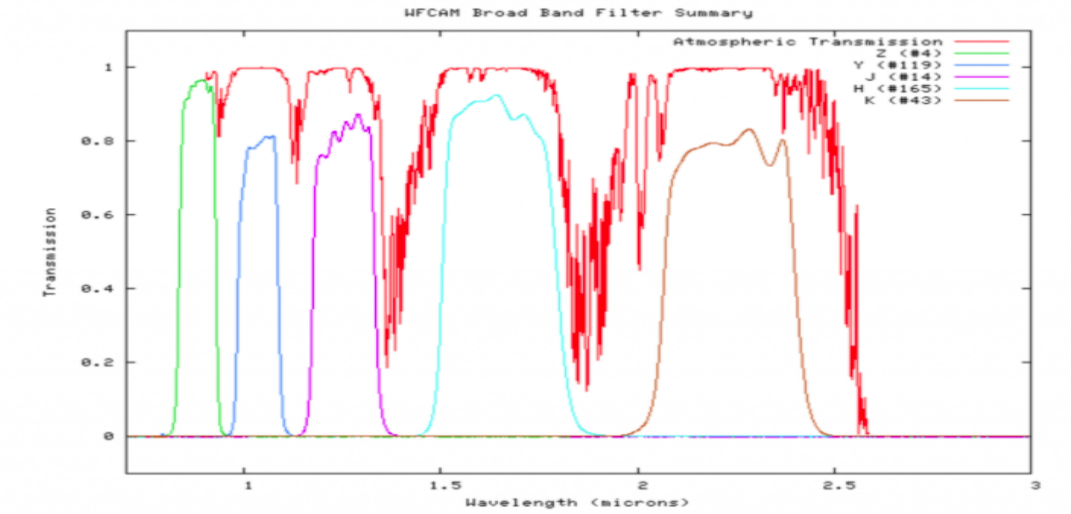


Figure A.4: Filter response curves for the Y J H K infrared bands of the UKIRT WFCAM. The red curve shoes the 'windows' of the atmospheric transmission in the corresponding wavelength ranges. https://about.ifa.hawaii.edu/ukirt/instruments/wfcam/description-of-wfcam/

A.4 GAIA

Gaia[5] (Global Astrometric Interferometer for Astrophysics) is a space telescope, launched to head to the Lagrangian L2 point of the Earth-Sun system, aiming to create a large catalogue of astronomical sources and their astrometry and kinematics. While mapping the stars in the Galaxy, it measures the distances and the exact positions of stars on a timescale of years in order to calculate the parallax and the proper motion. These observations can provide useful information for selecting extra-galactic objects since they are very far away and should have low values of parallax and proper motions. Former surveys have shown that parallaxes and proper motions consistent with zero within 2σ in the majority of cases describes a quasar [Heintz, Fynbo, Geier et al., 2020]. In other words, whenever photometry might not be able to provide with clear conclusions about the nature of an object, astrometry can help finding a solution.

Photometric catalogues $\times$ Gaia EDR3

The star, quasar and galaxy catalogues created through the method described in the previous sections are complete regarding the magnitude range they contain, from the ultra-violet to mid-infrared. At the same time, they provide the redshift and the spectroscopic class of all the sources. They, however, need to be enriched with information about the astrometric features, that is the proper motion, the parallax and their corresponding signal to noise ratios. The first two are found in the Gaia Early Release 3 (GAIA EDR3), which is cross-matched with the previous catalogues to provide the extra features. The error in the proper motion is not directly provided by the Gaia EDR3, but is rather calculated manually from the available columns as:

$$pm_{error} = \sqrt{(pmra_{err})^2 + (pmdec_{err})^2} \tag{A.1}$$

Having calculated the total error of the proper motion, the signal to noise ratio is then calculated as

$$S/N_{pm} = \frac{pm}{pm_{error}} \tag{A.2}$$

[5]https://sci.esa.int/web/gaia/

and serves, along with the signal to noise ratio for the parallax, as a great tool for distinguishing different kind of sources. Finally, the 3 broadband magnitudes (gmag, bmag, rmag) that GAIA measures were queried and included as features in the training. It should be noted though, that the g and r magnitudes are different from the corresponding SDSS magnitudes.

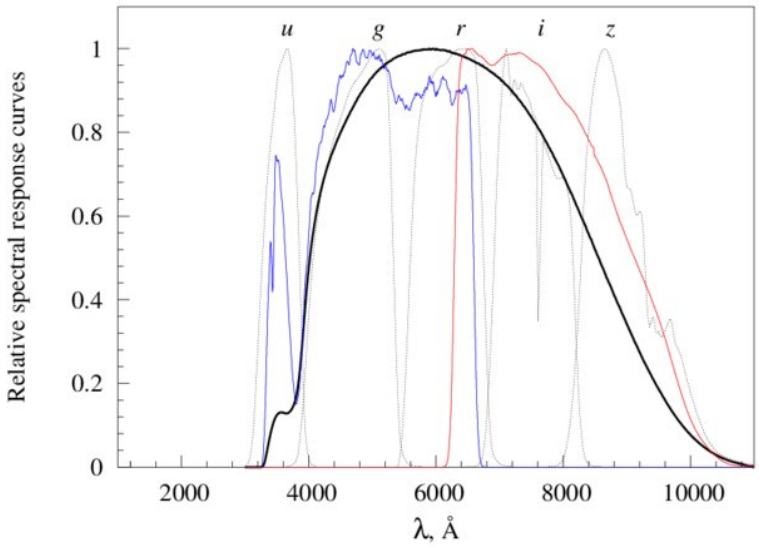


Figure A.5: Caption

An important drawback regarding the use of the Gaia mission as a source of astrometric characteristics is the fact that the telescope measures point-like sources and neglects the extended sources. This implies that only galaxies that are compact enough can be included in Gaia's catalogues. The result after the cross-match with the Gaia EDR3 is therefore a galaxy catalogue with an insufficient number of representatives.

Appendix B

Data Reduction

B.1 Long Slit Spectroscopy

This chapter discusses the steps of astronomical CCD spectroscopic data reduction. The spectra of the quasar candidates is derived and quantitative measurements of their spectral features are made. The main information extracted from these spectra will be those of the flux as a function of wavelength and the differentiation and strength of spectral lines. The latter will unravel the redshift of the astronomical objects, their type (whether they are quasars, galaxies or stars) and also the extinction. Meanwhile, absence of expected lines in the spectrum can provide crucial information for the intervening medium along specific lines of sight, such as DLAs. The data were collected at the Nordic Optical Telescope in La Palma, Spain, with long-slit spectroscopy. For the measurements the grism 20 of the ALFOSC instrument was used, with a range between 5650Å and 10150Å and with resolution $R = \frac{\lambda}{\Delta\lambda} = 770$ for a 1.0 arcsec slit. The observing schedules were made according to the interval of time our targets would be best observable, depending on their coordinates. This means that the target's altitude should exceed the NOT lowest limit for observing, while the possibility of a closed lower hatch that would alter the observing window should be taken into account. Moon gives significant rise in brightness in UV and optical range thus produces disturbance when it is not the target object. The moon coordinate is also shown in the visibility plots, one must make sure the object is not close to the moon. Our observation nights are during the period of new moon and crescent. Therefore the disturbance from the moon is minimized. Each observing night a standard star was used for the flux calibration. A visibility plot for GD71 standard star (of spectral type DAw) that was used for August's observations is shown in B.1. Apart from that and the science images, multiple spectral bias, flat and arclamps calibrating frames were also taken.

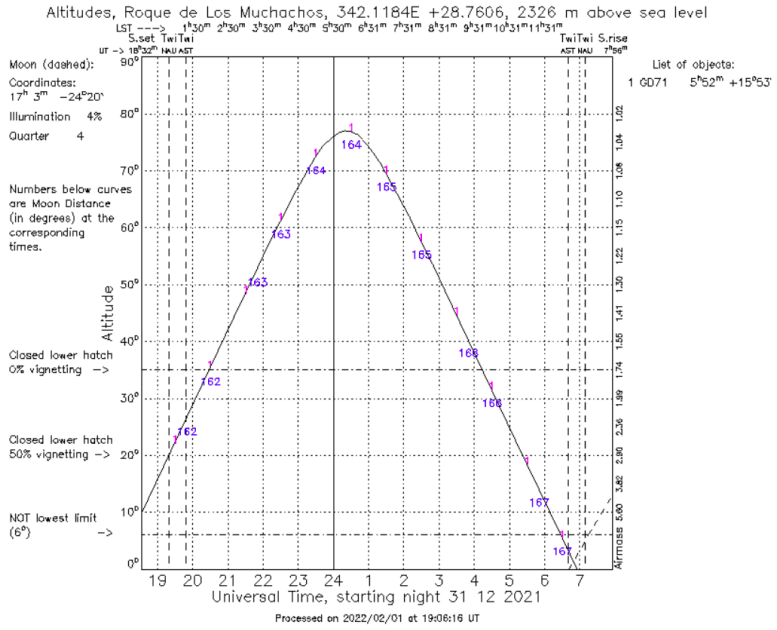


Figure B.1: Visibility plot of the standard star GD71 for the date of the first observing run.

B.1.1 Calibration frames

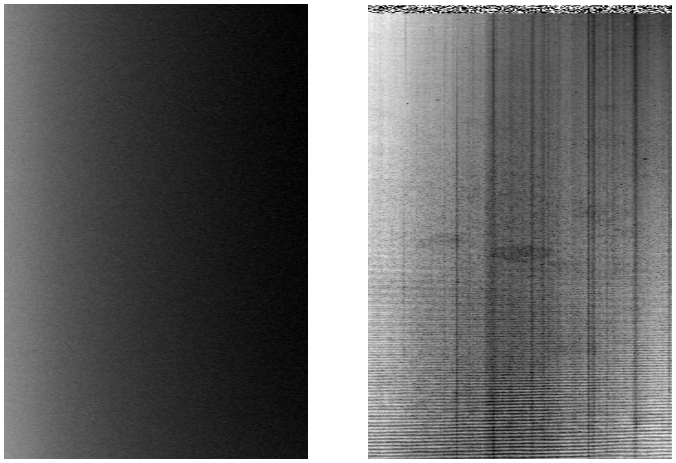
Long slit spectroscopy utilizes a narrow slit where the light from the source enters and refracts due to the existence of a prism or a diffraction grating. Followingly, the dispersed light is detected by the charge coupled device (CCD camera). When the CCD is exposed to light from a source, each pixel collects a number of impacting photons that excite electrons out of the silicon layer and add to the charge on a capacitor connected with that pixel. The voltage induced by this accumulated charge is then turned into analog to digital units (ADUs). This process comes with an inherent noise, the read out noise. Due to physical limitations of the instrumentation, one needs to add an electronic offset (a DC level) that is to be subtracted to get the raw signal. This offset is called the "bias", and it can be seen as a constant value or something with a spatial structure, depending on the CCD. To remove this bias level, one needs to take preferably multiple bias frames of so called zero second exposures with a closed shutter. This allows one to identify the added electronic offset and correct for it. A second calibration process is the flat field, that can reveal pixel to-pixel sensitivity variations to the incoming photons in order to eliminate such systematic errors. The flat field image is taken of an evenly illuminated surface, such as a twilight sky. Different filters that are later to be used for the science image are used, to account for wavelength dependency. The final science image is in the end based on the following processing: the derivation of the master bias and flat frame which together are used to create the normalised flat image. For each calibration image 10 frames of the same type were combined and stacked together, to lower the noise that is eventually added.

$$Normalised\ Flat = \frac{Master\ Flat - Master\ Bias}{Master\ Flat's\ mean\ value} \tag{B.1}$$

The final image that has undergone all the possible corrections and is valid for scientific measurements is then the product of the following subtraction:

$$Science\ Image = \frac{Raw\ Frame - Master\ Bias}{Normalized\ Flat} \tag{B.2}$$

The calibration images reduction is automated with corresponding python scripts that can be found in the Appendix.

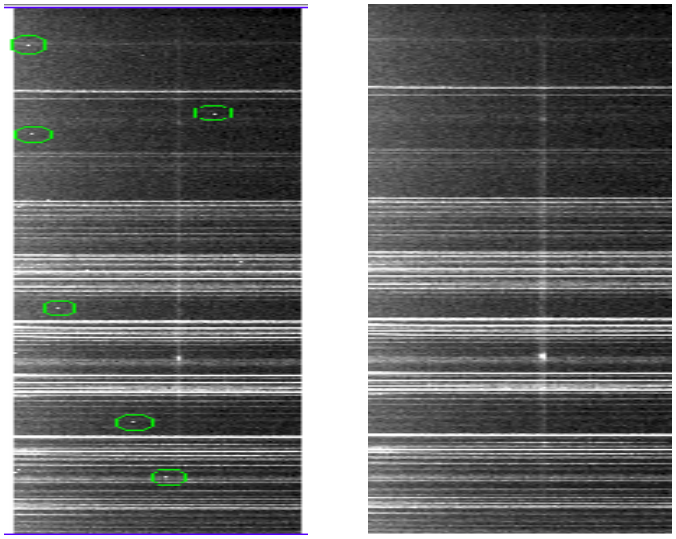


(a) Master bias frame (b) Normalized flat frame

Figure B.2: Calibration frames. Spatial sensitivity variances, structures and overscan regions are evident. Ds9 inspection of the mormalized flat reveals the legitimate mean pixel value around 1.

B.1.2 Cosmic ray correction

Further processing steps of the raw science images manipulations include the combination of the two exposures and the removal of any incident cosmic rays. This process takes place by running the astro-scrappy package, designed to recognise cosmic rays in images, originally based on Pieter van Dokkum's L.A.Cosmic algorithm. In the result of the detection and cosmic ray subtraction is demonstrated.



(a) Raw science image. Cosmic rays are evident as saturated pixel throughout the image

(b) Science image after the cosmic ray subtraction

Figure B.3: Calibration frames. Spatial sensitivity variances, structures and overscan regions are evident. Ds9 inspection of the mormalized flat reveals the legitimate mean pixel value around 1.

B.1.3 Wavelength calibration

THe process that follows the reduction of the scinece frame is the correlation of the pixel number with wavelength values. This calibration process utilizes the spectra of two halogen arclamp frames. The images are taken by illuminating the CCD with the HeNe lamp or ThAr lamp. One frame is taken right before and one right after the science exposures, at the same Grism. Spectral features of those lamps (He-Ne) identified at different pixel coordinates help to match those pixels with the corresponding wavelength where the line is produced, at the rest frame of the gas. The recognition of a fair amount of lines lets the extrapolation to the full axis range take place, and after this procedure of fitting a nth-degree polynomial on known intensity peaks, we end up with a x-axis that has units of wavelength in

Å

.

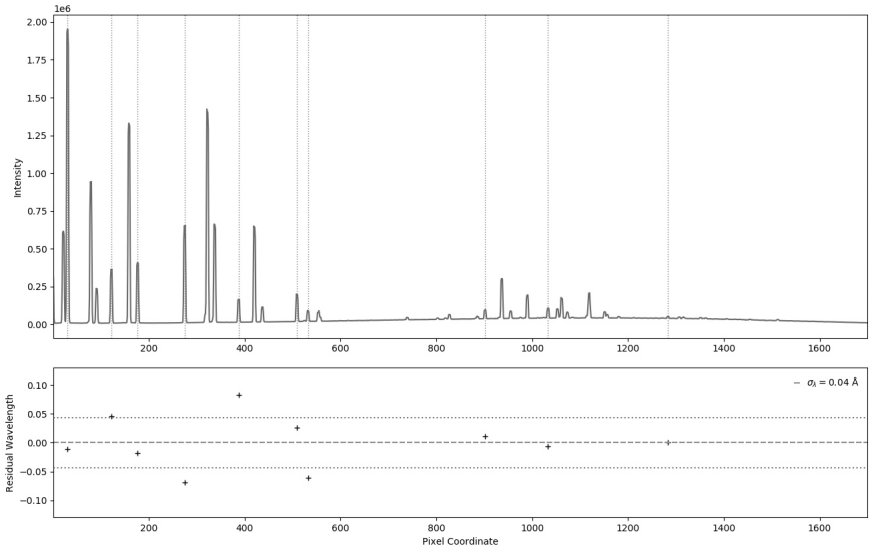


Figure B.4: Upper panel: Wavelength calibration using the He-Ne arclamps and the recognition of their spectral features, in Grism 20. Each pixel of the image is assigned to the corresponding wavelength where the lines are found. Lower panel: Wavelength residuals for each pixel coordinate, spread closely around zero, after a fitting of Chebyshev polynomial of order 6

B.1.4 Sky subtraction

Ground based telescopes such as the NOT telescope where the data were taken from need to account for the atmospheric absorptions and emissions. Especially in spectroscopy, that would significantly impact the spectral analysis leaving sky imprints to the object's specrtum and alter the scientific measurements and results. Running the extract 1d spectrum python script, a stripe of background before and after the line is manually selected, the average of which is subtracted by the image as seen in B.6.

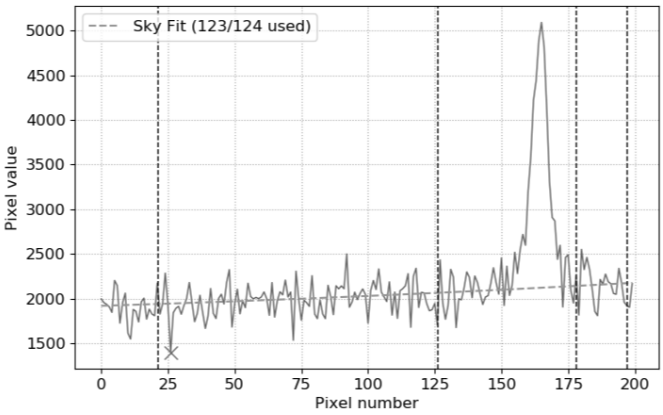
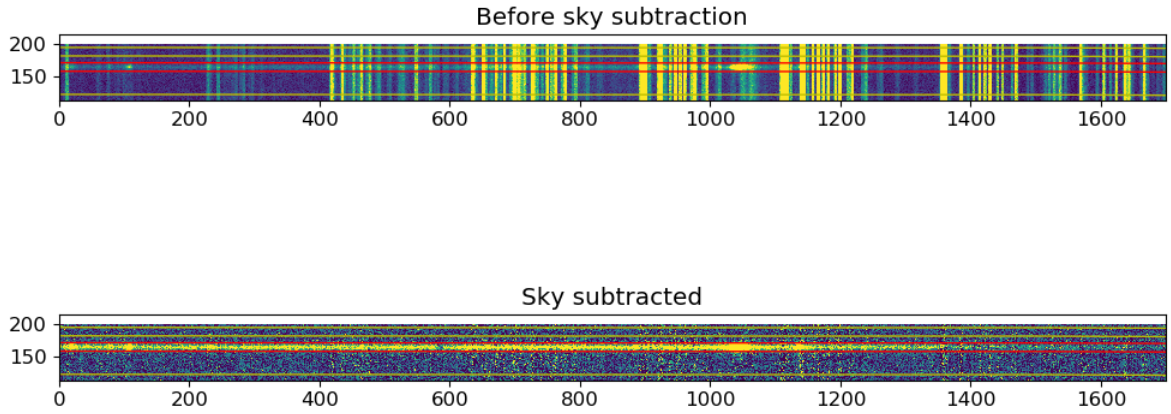


Figure B.5: Selection of background signal, to be discriminated by the observed object

Figure B.6: 1-D spectrum of image of quasar candidate J012534.07+351344.0. Upper panel: the spectrum before the sky subtraction. Lower panel: the same spectrum after the sky emissions are subtracted.



The same procedure is held for the spectrophotometric standard star, GD71.

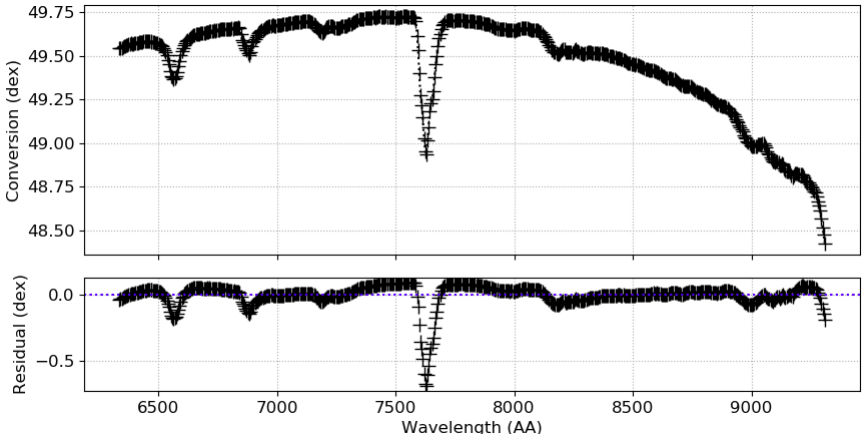


Figure B.7: Graphic window of sensitivity function script's output

B.2 Extinction

We can consider the dust grains as modified black bodies of radius α, with a SED described by the black body's radiation but multiplied by a factor Q, the absorption coefficient (opacity):

$$L_\lambda(\lambda, T) = 4\pi\alpha^2\pi Q B(\lambda)(\lambda, T) \qquad (ergs^{-1}cm^{-1}) \tag{B.3}$$

where $B(\lambda)$ is the Planck function:

$$B(\lambda)(\lambda, T) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{\frac{hc}{\lambda k_B T}} - 1} \tag{B.4}$$

and Q can be expressed in terms of the mass absorption coefficient,

$$k(\alpha, \lambda) = \frac{3Q(\alpha, \lambda)}{4\alpha\rho} \qquad (cm^2g^-1) \tag{B.5}$$

.

k's values range between 0 and 1, for a perfect reflector(mirror) and a perfect black body, respectively. Using eq. B.5 we can express the mass of an ensemble of grains (same radius α) as

$$M_d = \rho V = \frac{\pi\alpha^2 Q}{k} \tag{B.6}$$

Substitution of Q in eq. B.3, leads to the description of the dust radiation in terms of the total dust mass M_d:

$$L_\lambda(\lambda, T) = 4\pi k(\alpha, \lambda) M_d B(\lambda, T_d) \tag{B.7}$$

The emissivity of the intervening dust is therefore a black body distribution modified by the specific characteristics of the dust type, eg. its total mass M_d, number density and mass absorption coefficient. It is thus dependent on location and different regions, such as the Milky Way or the Small and Large Magellanic Clouds that have different dust consistency, generate different extinction curves. For our spectra reduction we use the SMC extinction coefficients.

Besides the light absorption, dust grains also account for scattering of the light, with the combination of the two effects leading to the extinction of the sources' light. The extinction efficiency is therefore comprised by two factors,

$$Q_{ext} = Q_{abs} + Q_{scat} \tag{B.8}$$

For small dust partiles ($2\pi\alpha << \lambda$), the scattering mechanism is described by the Rayleigh scattering. Defining $x = \frac{2\pi\alpha}{\lambda}$, the scattering efficiency is then described by the equation

$$Q_{sc} = \frac{8}{3}x^4\left|\frac{m^2 - 1}{m^2 + 2}\right| \propto \frac{1}{\lambda^4} \tag{B.9}$$

There, m is the complex index of refraction that describes the properties of the material and consists of a real and an imaginary part that are functions of the wavelength : $m(\lambda = n(\lambda) - ik(\lambda)$. The larger the real part, the better the grain acts as a scatterer- while the larger the imaginary part, the more effective absorber the grain is.

In general, the extinction by dust that leads to the attenuation of the radiation can be expressed through

an exponential decay,

$$I = I_0 e^{-\tau} \tag{B.10}$$

In this equation τ is the optical depth along the line of sight, defined as

$$d\tau(\lambda) = n\sigma_d(\lambda)dl \tag{B.11}$$

,n is the number density of the grains and σ_d is the absorption cross-section for a single grain,

$$\sigma_d(\lambda) = m_d k(\alpha, \lambda) \tag{B.12}$$

As seen in Fig.B.8, the incident radiation travels through the absorbing medium consisting of dust grains with absorption cross-section σ and its intensity I_{in} is reduced by a factor that is dependent on the opacity on that medium, resulting to the outgoing I_{out}. Converted in terms of magnitudes, the incident and final intensities are related as:

$$m_{out} - m_{in} = -2.5 log_{10} \frac{I_{out}}{I_{in}} \tag{B.13}$$

which with the use of eq. B.10 becomes

$$m_{out} - m_{in} = -2.5 log_{10} e^{-\tau} = 2.5\tau log_{10} e = 1.086\tau \tag{B.14}$$

So for more opaque mediums τ has a larger value and m_{out} increases, meaning that the astronomical object appears fainter. The quantity $1.086\tau(\lambda)$ is defined as extinction A.

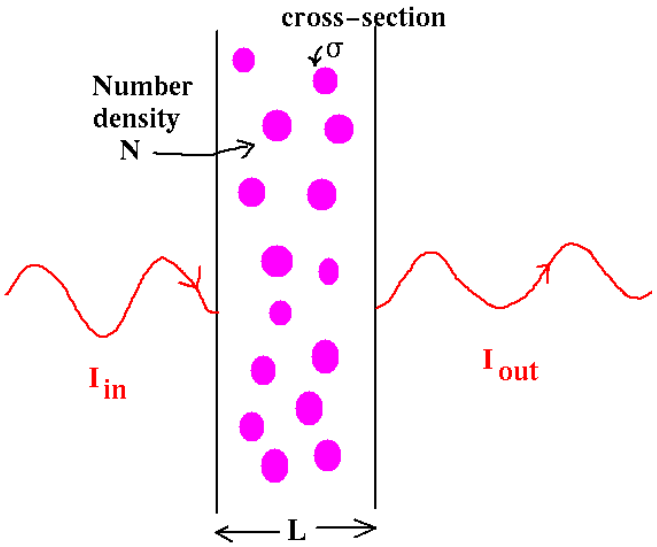


Figure B.8: Representation of radiation's attenuation, due to its propagation through a dusty medium

B.3 BPT Diagram

Target J125323.0+215717.6 is estimated to have redshift $z = 0.2$. Its emission lines are much less wide than any other observation's. From the SDSS Object Explorer, it also has observable host galaxy. To verify that it indeed hosts an active galactic nuclei, we refer to the BPT diagnostic diagram and calculate the $OIII5007Å/H\beta$ and $NII6583Å/H\alpha$ flux ratios [Baldwin, J. A. ; Phillips, M. M. ; Terlevich, 1981]. Lines that are close in wave-

length, and thus nearly reddening independent are used. Astronomical sources with emission lines in their spectra can be classified into groups, based on the relative intensities of those strong lines (Fig. B.9). Each group's (H II regions, AGNs) lines represent different excitation mechanisms (excitation by photoionisation by hot O, B stars in the case of H II regions, shock wave heating in LINERS or power-law photoionisation).

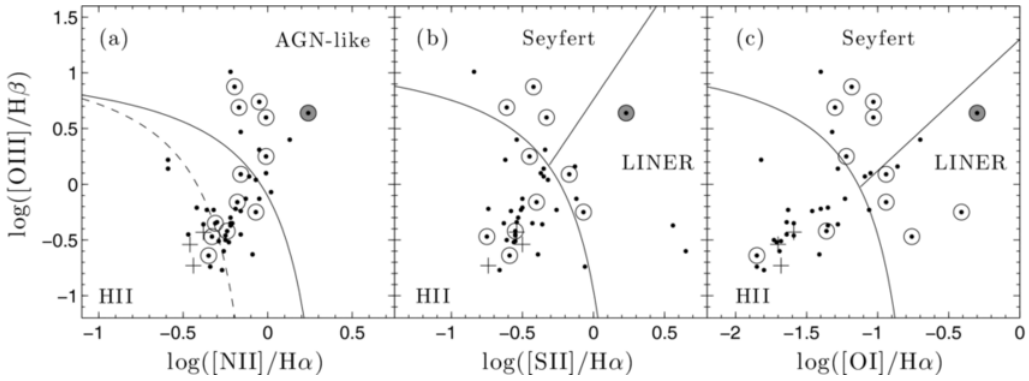


Figure B.9: BPT diagram for emission-line galaxies. (a)$NII/H\alpha$ vs. $OIII/H\beta$ diagram, (b)$SII/H\alpha$ vs. $OIII/H\beta$ diagram (c)$OI/H\alpha$ vs. $OIII/H\beta$ diagram. The open circles are for H II regions and similar sources that are clearly ionized by hot stars. The closed circles are narrow-line AGNs (Seyfert 2s and NLRGs) which are ionized by power-law continua. The division line between the different classes is empirical. LINERs are characterised by higher values of $NII/H\alpha$ ratio than H II regions, while they can be distinguished from Seyfert galaxies by lower values of $OIII/H\beta$.

By a simple gaussian fitting (Fig. B.10), the strength and FWHM of those are calculated and shown below. After we subtract the continuum level and divide the flux peaks, we find

$$log_{10}(\frac{OII}{H\beta}) = -0.09$$
$$log_{10}(\frac{NII}{H\alpha}) = 0.36$$

Target J125323.0+215717.6 is therefore placed on the AGN-like region of the BPT diagram but very close to the separating line. In that scheme, more investigation of the source's lines needs to be done in order to decide whether it is an AGN-like galaxy or not.

Fitted lines of J125323.0+215717.6:

wavelength (Angstrom)	Flux Peak (erg/s/cm^2/A)	FWHM	u_rot [km/s]
5947.00000	1.04E-16	47.93	2417.31
5983.00000	5.28E-17	18.06	905.18
6002.00000	1.54E-16	22.20	1108.18
7850.00000	8.76E-17	23.43	812.49
7870.00000	1.86E-16	71.50	645.81
7895.00000	1.60E-16	-16.95	515.60

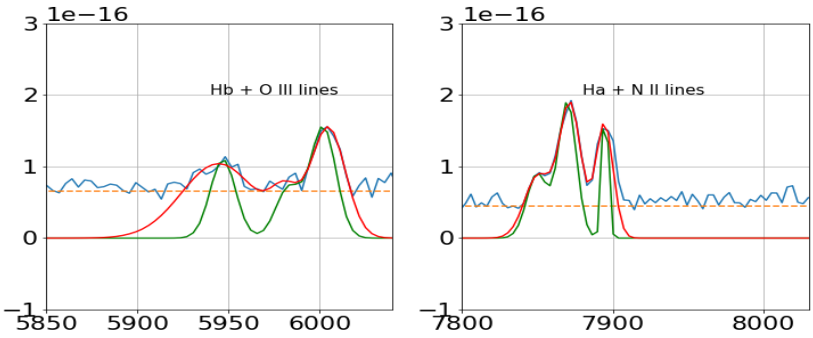
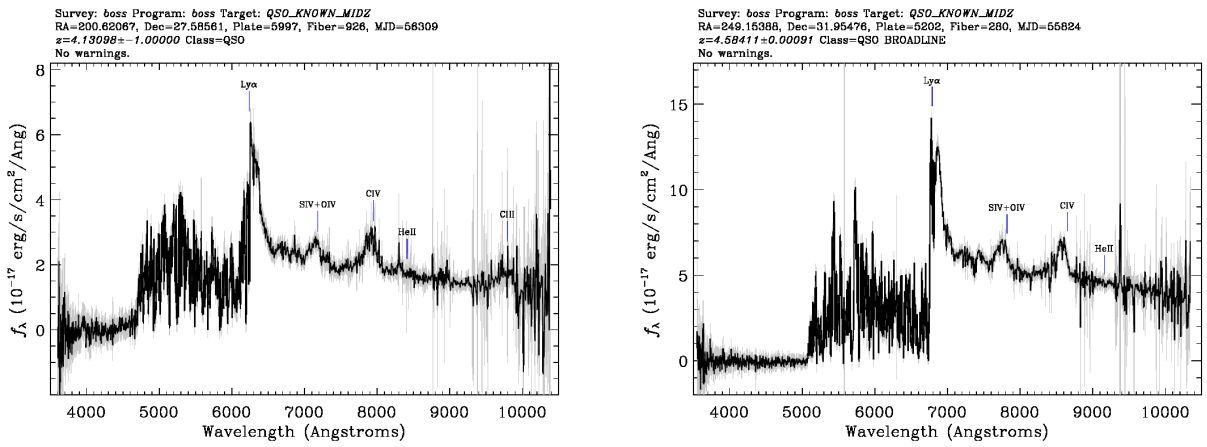


Figure B.10: Fitted lines of J125323.0+215717.6 with models.Gaussian.1D. Continuum level is $0.65 \cdot 10^{-16}$ and $0.45 \cdot 10^{-16}$ for the $H\beta + OIII$ and $H\alpha + NII$ lines,respectively.

Appendix C

Redshift Predictions - SDSS verification and deviation

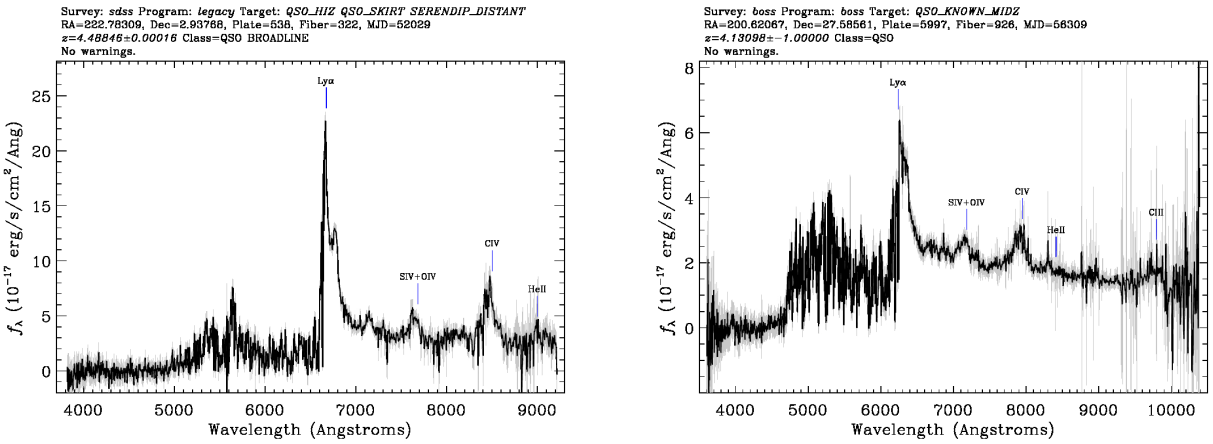
In chapter 7.2 we make photometric-redshift predictions, using the machine learning XGBoost Regression model, on 42.296 predicted QSOs. Here we present some spectra (SDSS DR16) predicted with a correct redshift (Figure C.1 C.2). Specifically, 78 out of the 42.296 are predicted with high redshifts ($z > 4$) up to redshift $z = 4.8$. It is fortunate that 42 out of the 78 quasars are also found in the SDSS Data Release 16 and so we could have a partial verification of our model. The spectroscopic redshifts found have the same values as the ones predicted with a ± 0.4 margin of error.



(a) A

(b) B

Figure C.1



(a) A (b) B

Figure C.2

But it should also be noted that some of the redshifts (especially ones higher than 6) although they were predicted correct compare to the SDSS ones, even from the SDSS were very oversestimated. The pipeline that SDSS uses in order to put redshift identifications to spectra is not always correct and should not be trusted blindly.

Appendix D

Gaia's Data Release 3 machine learning approach

Gaia Data Release 3 (DR3) has been released on 13^{th} of June 2022, right after the final observing run we conducted on the end of May. DR3 is an extension of the Early Data Release 3 (EDR3) (December 2020)[Gaia Collaboration et al. 2020a], that has been used a source of observating targets for this thesis. Briefly, EDR3 comprises of a full astrometric catalogue (positions, parallaxes, proper motions) of about 1.4 billion sources with a limiting magnitude of G $\sim$ 21. It also includes the G magnitudes of around 1.8 billion sources.

The full Data Release 3 contains an astonishing amount of information concerning physical parameters as well as object classifications for stars, quasars and galaxies. Although not all of this information is relevant to this thesis, it is worth mentioning some of the main aspects of the new released data. There are spectral types for 217 million stars within Milky Way galaxy. Spectroscopic parameters for 2.3 million hot stars and 94,000 ultra-cool stars. Evolutionary parameters (mass and age) for 128 million stars. It also contains astrophysical parameters such as effective temperature, [M/H], [X/M] for 5.5 million objects. It includes mean radial velocities of 33 million stars as well as rotational velocities of 3.5 million sources. DR3 includes a huge list of variable objects such as: Cepheids (15.021 objects), Compact companions (6306 objects), Eclipsing binaries (2,184,477 objects), Long-period variables (1,720,588 objects), Microlensing events (363 objects), Planetary transits (214 objects), RR Lyrae stars (271,779 objects), Short-timescale variables (471,679 objects), Solar-like rotational modulation variables (474,026 objects), Upper-main-sequence oscillators (54,476 objects) and lastly the most important for our work Active galactic nuclei (872,228 objects).

Concerning the AGNs, it contains a candidate list of about 6.6 million quasars with redshift estimates for most of them (it is stated though that the purity is low at values $50-70\%$). About 1.1 million quasars are analysed and 60.000 of them are found with a host galaxy. Moreover, it includes candidate list of about 4.8 million galaxies and redshift estimations for more than 1 million of them (again with the same low purity). Finally, more than 900.000 galaxies are analysed with two surface brightness profiles. Gaia made the predictions for the extragalacitc sources based on a machine learning model, namely the Discrete Source Classifier (DSC).

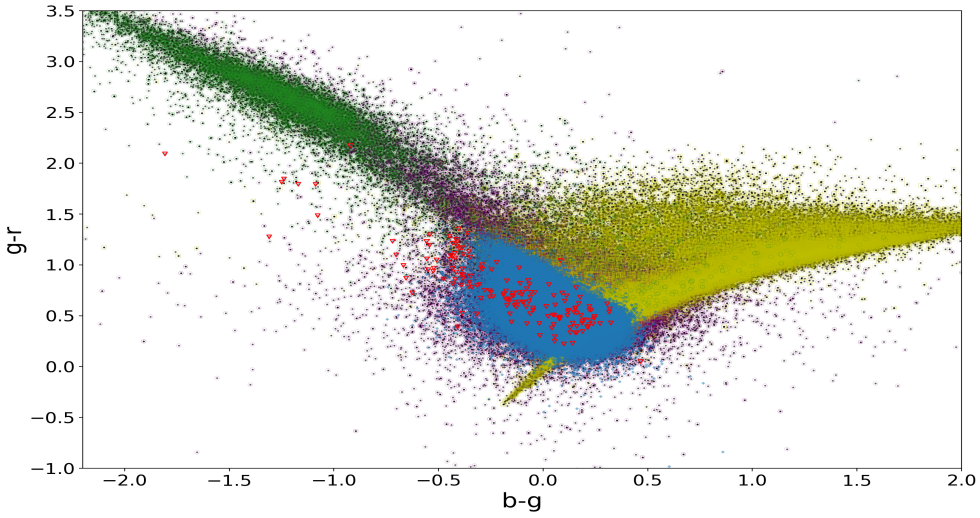


Figure D.1: Color-color plot of the qso-candidate list (371.708 sources) from the [Gaia DR3: The extragalactic content]. The predicted quasars from the Gaia DR3 Survey are shown with the blue color. The green, purple and yellow colors are for the galaxy, quasar and stellar locus respectively. We tested our XGBoost model on this qso candidates finding that 99.9 percent of the objects are also classified as quasars. The deviations from our model predictions are shown with the red triangles. In total 171 sources has been classified differently from our model. Specifically, 84 objects are classified as stars and 89 as galaxies

They also have made a purer qso-candidate list containing 371.708 sources (Fig. D.1 - blue dots). In order to investigate the performance of our purely Gaia model (astrometry+gaia photometry) we also made predictions on this qso-candidate list. What we find is a match between their and our predictions of 99 %. However, there are 89 stars and 84 galaxies that our model predicts as such on this qso-candidate list. In Fig. D.2 we show one example that our model predicted as a galaxy and the Gaia's model as a quasar.

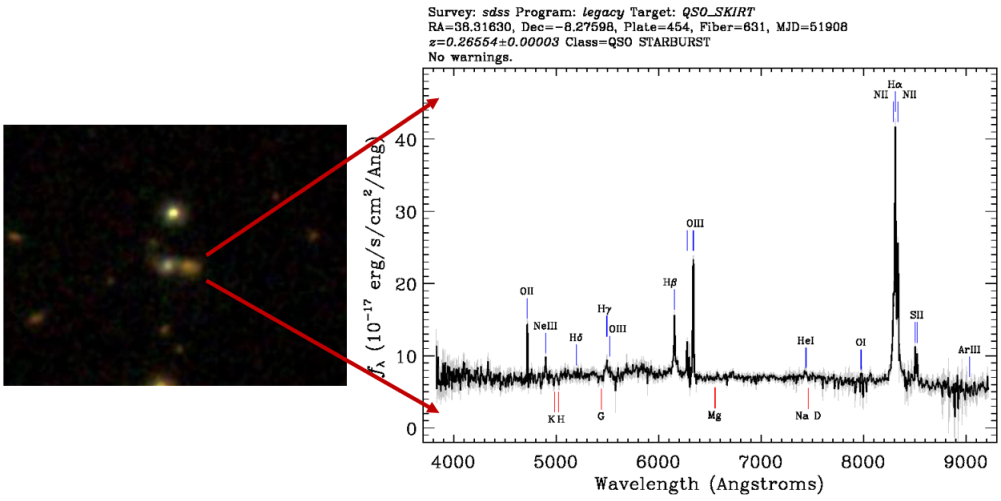


Figure D.2: This is an example of a Seyfert Galaxy that was found in the Gaia's purer qso-candidate list but predicted correctly as galaxy by our model. Left: Image from Object Explorer SDSS DR16. Right: Spectrum of the object.

Appendix E

Other supervised machine learning systems

E.0.1 kNN - Nearest - Neighbor Methods

The k-nearest neighbors algorithm is a data classification method that is used in predicting whether a data point belongs to one group or another. The model's decision is based on the group to which the nearest neighbours belong to. It is a supervised machine learning technique used to solve classification or regression problems and is sometimes described as a lazy non-parametric learning. In fact, the only parameter of the algorithm is the "k" that appears in the name of the method. It is used to describe the number of the nearest neighbours that will influence the decision process. It is called a lazy learning because it does not perform any training based on the input data and thus it is unable to make generalizations and find patterns through the data. Instead, it just stores the data during the training time so they can be used on future unseen data based on the nearest neighbors technique. Concerning the optimal value of "k", literature states that there is not a specific way to determine it. This means that one has to experiment with a few values before he decides which is the best one. However, there is a fast way of choosing the k parameter by calculating the $\sqrt{N}$, where N denotes the number of samples in the training data set. One advantage of the kNN method is that it is easy to understand compared to other methods and is relatively simple to implement. On the other hand, it requires high memory storage as it stores all the training data and the prediction is slow if N is large. Another disadvantage of this kind of model is that it is very susceptible to overfitting due to the curse of dimensionality (Fig. E.1). What this term refers to is a situation in which the dataset has too many features. It might be counter-intuitive, but as the number of features increases observations become harder to cluster. Because clustering (or classifying), at least in the case of the kNN models, uses a distance measure (such as Euclidean distance) to quantify the similarity between observations. So, by increasing the number of features, i.e. the dimensions, you risk every observation in your dataset to appear equidistant from all the others in the input space.

E.0.2 SVM - Support Vector Machines

Support vector machines are also a supervised machine learning approach, meaning that the true labels (target values [y]) of the different observations ([X]) are needed for the training part. A support vector machine takes these observations and creates a hyperplane that best separates them into the different classes (Fig. E.2). This line is also called a decision boundary. Most of the times this line is straight and so we end up with a linear classifier. But as can be seen in Fig. **??**, the two classes shown are impossible to be directly separated

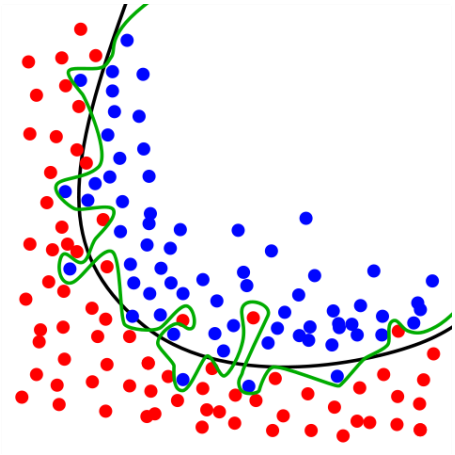


Figure E.1: An example of an kNN model in which k parameter was set very low and as a result the model is over-fitted. This means that the model became very good in classifying the training set but cannot generalize well on unseen data.

by a linear hyper-plane. That is why support vector machine introduces a new additional feature which is a combination of the two previous feature vectors: $z = x^2 + y^2$. This new hyperplane can now properly separate the two classes. Building an SVM which is linear is a fairly easy process, whereas finding the correct additional feature for a non-linear classification manually can be computationally expensive and in general non-trivial. Luckily, this procedure is done automatically by an SVM technique called kernel trick. This method is based on functions that take data that are not linearly separated in a low dimensional space and transform it into a higher dimensional space where they can be linearly separated.

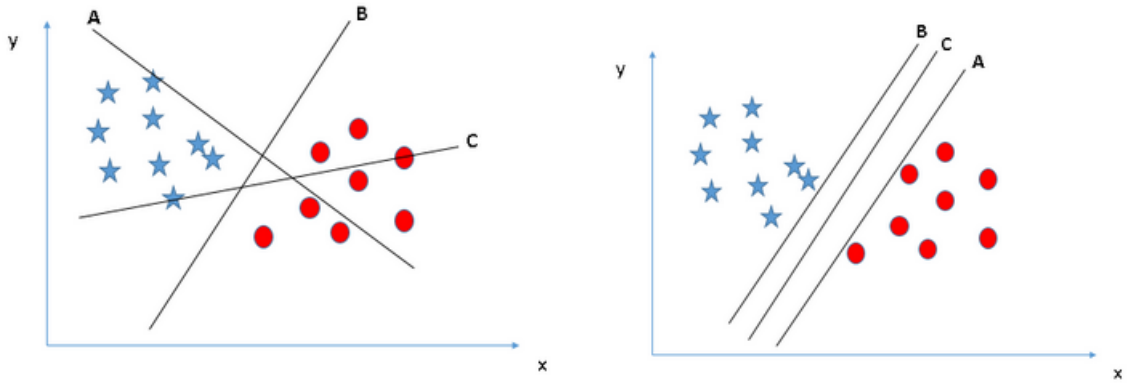


Figure E.2: Left panel: A, B and C represents three hyperplanes. The hyperplane that separates the classes better is the one that will be selected by the SVM mehtod. As the figure shows, B is the best choise as it separates perfectly all the data points. Right panel: The three hyperplanes are equally good at separating the classes and the one with the highest distance will be chosen. This distance is called margin.

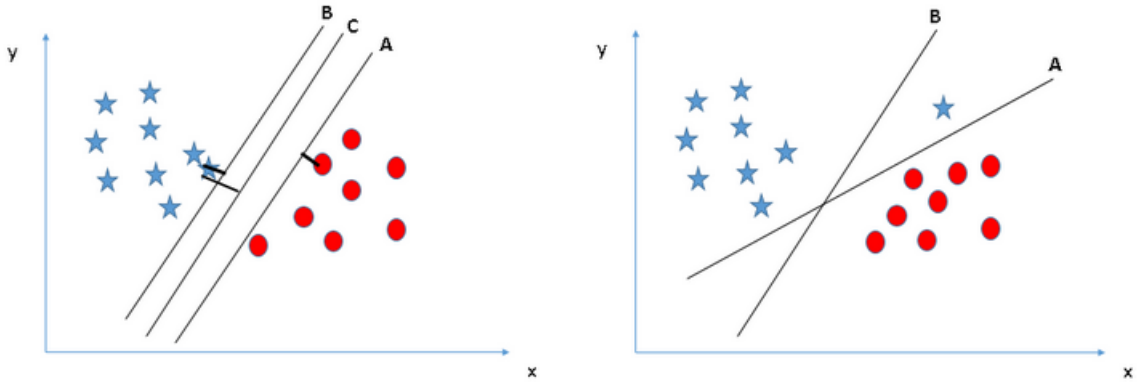


Figure E.3: Left panel: In this sub-figure the procedure of selecting equally good hyperplanes is presented. As it can be seen the hyperplane C will be chosen as the optimal one since it has the highest margin. Right panel:In this scenario the model will prioritize the correct classification rather than the highest margin and so hyperplane A will be chosen.

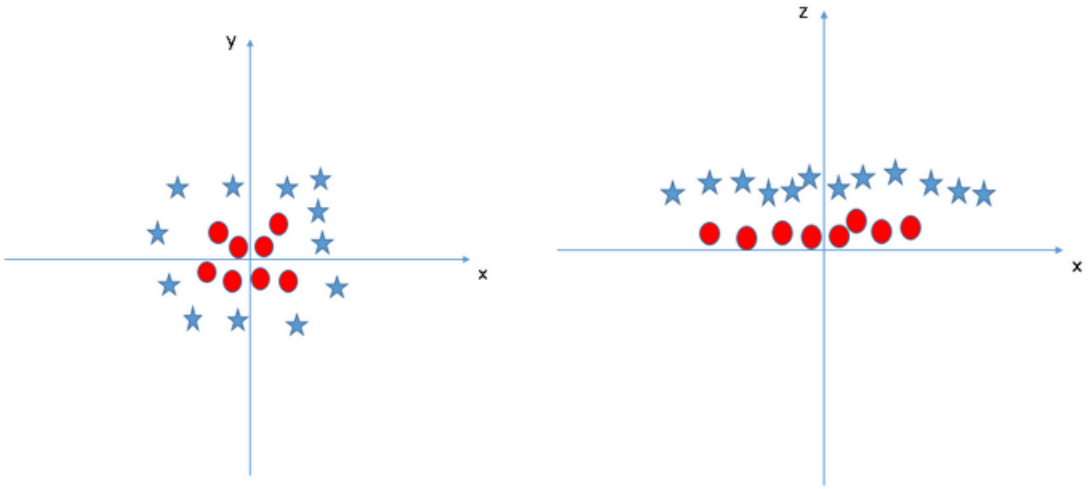


Figure E.4: Left panel:Scenario where two classes are impossible to get separated by a linear hyperplane. Right panel: The solution can be found by introducing a new feature vector which is some kind of combination among the initial features. In the initial input space this hyperplane appears as a non-linear one.

Bibliography

[Schmidt M., 1963] 3C 273 : A Star-Like Object with Large Red-Shift. Nature 197,1040 (1963). https://doi.org/10.1038/1971040a0

[Hazard et al. ,1963] Investigation of the Radio Source 3C 273 By The Method of Lunar Occultations. Nature, Volume 197, Issue 4872, pp. 1037-1039 (1963).

[Gunn et al., 1998] Gunn et al., Dec 1998, The Sloan Digital Sky Survey Photometric Camera, The Astronomical Journal, Volume 116, Issue 6 arXiv:astro-ph/9809085

[Wright et al., 2010] Wright E. L. et al. , 2010, The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance, The Astronomical Journal arXiv:1008.0031

[Lawrence et al., 2007] A. Lawrence et al., May 2007, The UKIRT Infrared Deep Sky Survey (UKIDSS), doi:10.1111/j.1365-2966.2007.12040.x

[Antonucci, 1993] Unified models for Active Galactic Nuclei and quasars. 10.1146/annurev.aa.31.090193.002353

[Urry & Padovani, 1995] Urry, C. M., & Padovani, P. 1995, PASP, 107, 803 10.1086/133630

[Kauffmann et al, 2003] Kauffmann, G., Heckman, T. M., Tremonti, C., et al. 2003,MNRAS,346,1055

[Antonucci & Miller, 2985] Antonucci, R. & Miller, J. S. Spectropolarimetry and the nature of NGC 1068. Astronom. J. 297, 621–632 (1985)

[Smith et al., 2005] J. E. Smith et al. , Equatorial scattering and the structure of the broad-line region in Seyfert nuclei: evidence for a rotating disc (2005). https://doi.org/10.1111/j.1365-2966.2005.08895.x

[Yoshiaki & Anabuki, 1999] Yoshiaki Taniguchi, Naohisa Anabuki, The Electron Scattering Region in Seyfert Nuclei (1999) https://doi.org/10.48550/arXiv.astro-ph/9906430

[Heintz, Fynbo, Geier et al., 2020] K. E. Heintz, J. P. U. Fynbo, S. J. Geier et al., Spectroscopic classification of a complete sample of astrometrically-selected quasar candidates using Gaia DR2 (2020) arXiv:2010.05934

[Wu & Zhendong, 2010] Xue-Bing Wu, Zhendong Jia, Quasar candidate selection and photometric redshift estimation based on SDSS and UKIDSS data (2010) doi:10.1111/j.1365-2966.2010.16807.x

[Warren & Pitts, 1943] Warren S. Mcculloch & Walter Pitts, A Logical Calculus of the Ideas Immanent in Nervous Activity (1943), 10.2307/2268029

[Rosenblatt, 1958] Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, 65, 386. https://doi.org/10.1037/h0042519

[Andinda et al.,] D. Andina, A. Vega-Corona, J. I. Seijas and J. Torres-Garcìa., Neural Networks Historical Review DOI:10.1007/0-387-37452-3_2

[Tianqi, Guestrin, 2016] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System (2016). arXiv:1603.02754

[Russell et al., 2010] Russell, S. J., Norvig, P., and Davis, E. Artificial Intelligence : A Modern Approach. 3rd ed. Upper Saddle River: Prentice-Hall (2010). xviii, 1132 p. p. https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf

[Warren, 2000] Warren, S. J., P. C. Hewett, C. B. Foltz,The KX method for producing K-band flux-limited samples of quasars, 2000 https://doi.org/10.1046/j.1365-8711.2000.03206.xx

[Vanden Berk, 2001] Vanden Berk et al., Composite Quasar Spectra From the Sloan Digital Sky Survey1, 2001 arXiv : arXiv:astro-ph/0105231

[Selsing, 2015] J. Selsing, J. P. U. Fynbo, L. Christensen, and J.-K. Krogager An X-Shooter composite of bright $1 < z < 2$ quasars from UV to infrared, 2015 doi : https://doi.org/10.1051/0004-6361/201527096

[Arthur Samuel, 1956] Arthur L. Samuel, Some Studies in Machine Learning Using the Game of Checkers. 10.1007/978-1-4613-8716-9_14

[Breiman, 2001] Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[Pei, 1992] Interstellar dust from the Milky Way to the Magellanic Clouds https://doi.org/10.1086/171637

[Wills et al., 1993] Statistics of QSO Broad Emission-Line Profiles. I. The C IV lambda 1549 Line and the lambda 1400 Feature. https://doi.org/10.1086/173186

[Giveon et al., 1999] Long-term optical variability properties of the Palomar-Green quasars arXiv:astro-ph/9902254

[Vanden Berk et al. 2004] The Ensemble Photometric Variability of 25,000 Quasars in the Sloan Digital Sky Survey arXiv:astro-ph/0310336

[Rumbaugh et al. 2018] Extreme Variability Quasars from the Sloan Digital Sky Survey and the Dark Energy Survey arXiv:1706.07875

[Sartori et al. 2019] A Forward Modeling Approach to AGN Variability–Method Description and Early Applications arXiv:1909.06374

[Klimek et al., 2000] Optical Variability of Narrow-Line Seyfert 1 Galaxies arXiv:astro-ph/0403334

[Stalin et al., 2004] Intranight optical variability of radio-quiet and radio lobe-dominated quasars arXiv:astro-ph/0306394

[Gaia Collaboration et al. 2020a] Gaia Early Data Release 3. Acceleration of the Solar System from Gaia astrometry arXiv:2012.02036

[Stern et al., 2000] Discovery of a Color-selected Quasar at z = 5.50*. 10.1086/312614

[Gaia DR3: The extragalactic content] Gaia Collaboration et al.:Gaia Data Release 3: The extragalactic content.

[Gunn, Peterson 1965] On the Density of Neutral Hydrogen in Intergalactic Space. 10.1086/148444

[Lukasz Stawarz, 2004] ON THE JET ACTIVITY IN 3C 273.

[E. Bañados et al., 2018] E. Bañados, B. P. Venemans, C. Mazzucchelli, E. P. Farina, F. Walter, F. Wang, R. Decarli, D. Stern, X. Fan, F. B. Davies, J. F. Hennawi, R. A. Simcoe, M. L. Turner, H.-W. Rix, J. Yang, D. D. Kelson, G. C. Rudie, and J. M. Winters. An 800-million-solar-mass black hole in a significantly neutral Universe at a redshift of 7.5. , 553:473–476, January 2018. 10.1038/nature25180.

[P.C. Hewett et al., 2003] The Frequency and Radio Properties of Broad Absorption Line Quasars. , 125:1784–1794, April 2003. 10.1086/368392.

[A. Sandage, 1965] The Existence of a Major New Constituent of the Universe: the Quasistellar Galaxies. May 1965. 10.1086/148245.

[Schmidt and Green, 1983] M. Schmidt and R. F. Green. Quasar evolution derived from the Palomar bright quasar survey and other complete quasar surveys. June 1983. 10.1086/161048.

[K. E. Heintz, J. P. U. Fynbo et al., 2020] Spectroscopic classification of a complete sample of astrometrically-selected quasar candidates using Gaia DR2. arXiv:2010.05934

[Strauss et al., 2002] Spectroscopic target selection in the Sloan Digital Sky Survey: the main galaxy sample.

[Krogager, Fynbo et al., 2016] The Extended High A (V) Quasar Survey: Searching for Dusty Absorbers toward Mid-infraredselected Quasars. arXiv:1608.08404

[Heintz et al., 2018] A quasar hiding behind two dusty absorbers-Quantifying the selection bias of metal-rich, damped Lyα absorption systems. arXiv:1803.09805

[Warren et al., 2000] The KX method for producing K-band flux-limited samples of quasars. arXiv:astro-ph/9911064

[Baldwin, J. A. ; Phillips, M. M. ; Terlevich, 1981] Classification parameters for the emission-line spectra of extragalactic objects. doi:10.1086/130766

[Stern et al., 2012] Mid-infrared Selection of Active Galactic Nuclei with the Wide-Field Infrared Survey Explorer. I. Characterizing WISE-selected Active Galactic Nuclei in COSMOS. arXiv:1205.0811

[Lacy et al., 2004] Obscured and unobscured active galactic nuclei in the Spitzer Space Telescope First Look Survey. arXiv:astro-ph/0405604

[Goldschmidt, Pippa et al., 1992] The high surface density of bright ultraviolet-excess quasars. 10.1093/mnras/256.1.65P

[Smith, Paul S, 2003] Optical Spectropolarimetry of Quasi-stellar Objects Discovered by the Two-Micron All Sky Survey. arXiv:astro-ph/0305098

[Tian Qiu et al., 2020] Proper motion measurements for stars up to 100 kpc with Subaru HSC and SDSS Stripe 82 arXiv:2004.12899

[G. Kron, 1981] Stars with zero proper motion and the number of QSOs.

[Krogager et al., 2015] The High AV Quasar Survey: Reddened Quasi-Stellar Objects Selected from Optical/Near-Infrared Photometry—II. arXiv:1410.7783

[Krogager et al., 2016] The Extended High A(V) Quasar Survey: Searching for Dusty Absorbers toward Mid-infrared-selected Quasars. arXiv:1608.08404

[Geier, Heintz, Fynbo et al., 2019] Gaia-assisted selection of a quasar reddened by dust in an extremely strong damped Lyman-α absorber at z = 2.226. arXiv:1904.01686

[Sara Issaoun et al., 2022] Resolving the Inner Parsec of the Blazar J1924–2914 with the Event Horizon Telescope. https://iopscience.iop.org/article/10.3847/1538-4357/ac7a40