Spectra classification for Quasar detection in the Sloan Digital Sky Survey DR16

Chihab Khnifass, Jaime A. Silva, Angela M. Montoya, Carlos U. Porras

***Abstract*— In this project, we took a subset of 110.000 spectra from the Sloan Digital Sky Survey (SDSS), and tested different approaches for dimensionality reduction and classification models to identify the kind of object based on the spectra it emits. For dimensionality reduction, the tested approaches were: Linear Discriminant Analysis, Principal Component Analysis and K-Best. While for the classification we used a Support Vector Machine, K-Nearest Neighbours, a Random Forest, Naive Bayes Classification, a Multilayer Perceptron and Decision Trees generated with Gradient Boosting. Our results favored clearly the k-best algorithm over LDA and PCA for dimensionality reduction, and the chosen 30 features (from over 30.000 in the initial dataset) point to interesting wavelength ranges to watch in the further study of quasars. Between the classification methods, the Gradient Boosting generated trees were slightly ahead of the perceptron, followed by KNN and Random Forest; the SVM and the Naive Bayes were the worst performing.**

***Index Terms*—**Active Galaxy, Active Galaxy Nuclei, AGN, Astronomy, Data Science, Data Analytics, Exploratory Analysis, Classification, Dimension reduction.

I. Introduction

Of all the galaxies observed to date, a small group of them are distinguished by strange behaviours in their center. They are called Galaxies with Active Nuclei, and their main characteristic is having a supermassive black hole (an Active Galaxy Nucleus - AGN) in their center which, strangelly, doesn't retain all the matter and energy at its grasp, but instead expels part of it away from the galaxy in a strong and very fast radiation beam. With a mass greater than $1 * 10^8$ solar masses in a physical space between $3 * 10^{10}$m and $3 * 10^{11}$ m radius (our sun has a radius of $7 * 10^8$ m) these supermassive black holes are some of the most dense objects we know in space, and as such is supposed to be the cause of their unique behavior. Usually, Black holes absorb all forms of matter and energy in the vicinity of their area of influence (including light), the vicinity of the black hole where no particle has possibility of escape is called the Accretion Disk. One of the most accepted explanations for the counterintuitive radiation beam expelled from AGNs is that when matter enters the Accretion Disk, the speed at which all particles are absorbed and the amount of absorbed material generates excessive friction forces and energy from collisions, to the point that a small part of it gets enough energy to escape the accretion disk. The speed of these particles as they leave the galaxy is huge, and have been observed with jet-shaped x-ray filters reaching great distances outside the galaxy, perpendicular to the center of the galaxy and taking different shapes.

Unlike the usual spectrography of astronomical objects, where the radiation captured from it has a narrow distinct pattern that allows to analyse the composition, temperature, speed and other properties of the object, the center of AGNs has a vicinity that emits an unusually broad spectrum.This region is called the Broad Line Region, and is currently spatially unresolved, this means we are currently unable to predict the size and shape (even the existence) of this region given other properties from the AGN, only through direct observation we can determine this in a case by case basis.

For this reason, automatically detecting AGNs in a big sample of candidates can be of great help in searching for patterns that allow the construction of a consistent model to explain the physical properties of the BLR. Also, the use of dimensionality reduction techniques can provide future researchers with insights about wavelength ranges that could be of special importance in the study of quasars.

In this document, after presenting our main objective, we give a short review of key concepts that the general reader can find useful in understanding the process we went through and the data we dealt with. Next, we describe the process we followed to collect, clean and organize the data, and the shape of our working dataset. Then, we explain the techniques used for dimensionality reduction and classification, and close the report with our final results and conclusions.

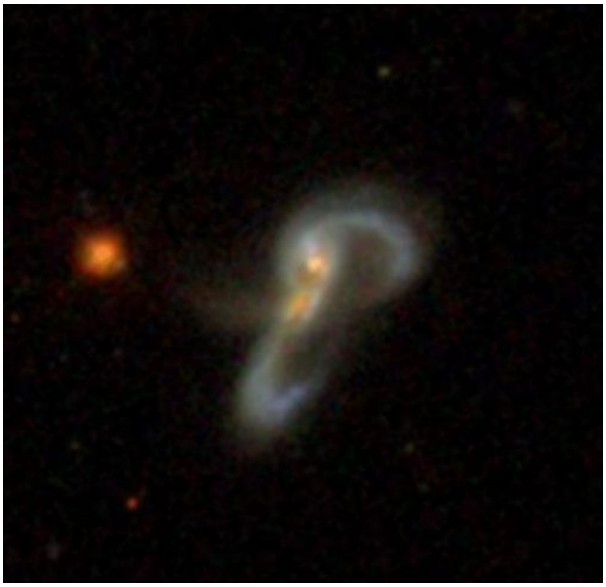


Fig. 1. Real image of AGN Mrk 848A (SDSS, observed 2003)

II. Project Goal And Scope

The main objective of this project is to apply different dimensionality reduction and classification algorithms to identify if an astronomical object is a Star, a Galaxy or a Quasar (AGN), given its continuum spectra obtained from the SDSS Data Release 16.

III. Literature Review

A Galaxy is called an Active Galaxy, or a Galaxy with an Active Galaxy Nuclei (AGN), when it has a large supermassive black hole that emits large amounts of radiation in its center, spreading luminous material and maintaining high temperatures usually attributed to friction between the particles it absorbs. From Earth, we are able to see this radiation in the form of photons which are collected and

[1]Submited on July 8th 2020, final project report for Introduction to Data Science and Data Visualization, graduate class in Universidad Nacional de Colombia. Code is available in Colaboratory for reference [12], [13]

A. M. Montoya - Msc. Computer Science (ammontoyaca@unal.edu.co)
C. U. Porras - Msc. Astronomy (cporrasd@unal.edu.co)
C. Khnifassis - Beng. Generalist (chihab.khnifass@mines-ales.org)
J. A. Silva - Bsc. Mathematics (jaiasilvavel@unal.edu.co)

counted with the aid of telescopes and special cameras, which measure the wavelength of the light and classify it in known ranges corresponding to the spectrum of absorption or emission of different elements.

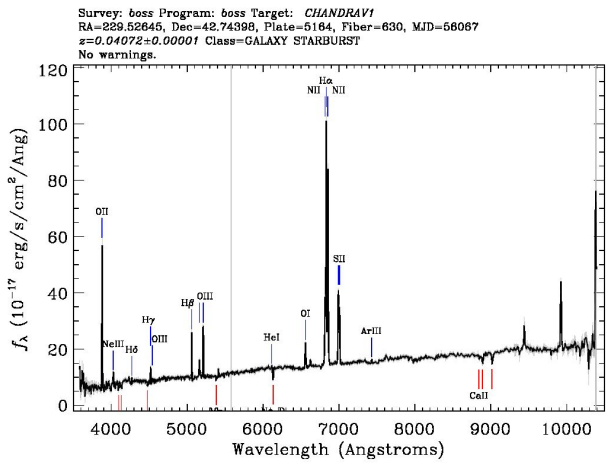


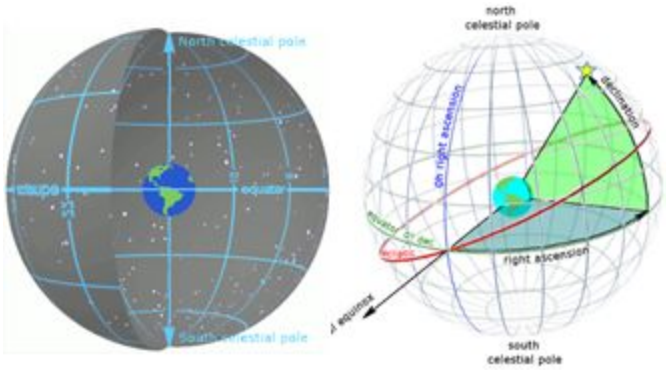
Fig. 2. Calculated Spectra of AGN Mrk 848A (SDSS, observed 2003)

Spectrophotometric. Spectrophotometry is a method to measure how much a chemical substance absorbs light by measuring the intensity of light as a beam of white light passes through a sample solution. The basic principle is that each compound absorbs or transmits light over a certain range of wavelengths. This measurement can also be used to measure the amount of a known chemical substance. Spectrophotometry is one of the most useful methods of quantitative analysis in various fields such as chemistry, physics, biochemistry, material and chemical engineering and clinical applications. [2]

Quasars. By the 1950s, astronomers already used radio telescopes to probe the heavens, and pairing their signals with visible examinations of the night sky. However, some of the smaller point-source objects of radio emission didn't match with their visual counterpart (being much stronger in their radio signals than in their visible brightness). Astronomers called them "quasi-stellar radio sources" or "quasars" because the signals came from a singular point in space, like star light. However, the name is a misnomer; according to the National Astronomical Observatory of Japan, only about 10 percent of quasars emit strong radio waves. "Quasars are among the brightest and most distant known celestial objects and are crucial to understanding the early universe," astronomer Bram Venemans of the Max Planck Institute for Astronomy in Germany said. [3]

Flux. Is the apparent brightness of a galactic object. In the spectrometric measures, it's the photon count for a given wavelength as measured by the instrument.

Celestial Coordinates. RA (right ascension) and Dec (declination) are the coordinates on the sky that correspond to longitude and latitude on Earth. RA measures east and west on the celestial sphere and is like longitude on the Earth. Dec measures north and south on the celestial sphere and is like latitude on the Earth.[6]



6

Fig. 3. Celestial sphere. Right Ascension (ra) and Declination (dec) [7].

Redshift (z). Redshift is an example of the Doppler effect. In Astronomy you can learn about the movement of cosmic objects by observing how their color changes over time or how it varies from what is expected. If an object is redder than expected, it can be concluded that it is moving away, and if it is more blue, it can be said to be approaching. The most accurate way to measure redshift is by spectroscopy. When a white light beam strikes a triangular prism, it separates into its various components (ROYGBIV). This is known as spectrum (plural: spectra). The spectra created by different elements can be observed and compared with the spectra of the stars. If the absorption or emission lines seen in the star's spectra shift, it is known where the object is moving. Red shift is referred to in terms of the red shift parameter z. This is calculated with an equation, where λ observed is the observed wavelength of a spectral line, and λ rest is the wavelength that line would have if its source were not in motion.

$z = (\lambda \text{ observed} - \lambda \text{ rest}) / \lambda \text{ rest}$ [8].

SDSS. Sloan Digital Sky Survey. SDSS has a database of millions of galaxies, quasars and stars, including spectrographic measures for almost all the celestial North hemisphere. Spectrographic information is captured using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. [9]

IV. Reference Studies

Our main base study is [10], this project applied a convolutional neural network (CNN) and a random forest model to classify and detect quasars in the Sloan Digital Sky Survey Stripe 82 and also to predict the photometric redshifts of quasars. Also, we take into account [1] for a deeper understanding of the astrophysical meaning of the information in the dataset, and to check the consistency of our approach and conclusions; they worked with measurements of 17 quasars during 7.5 years, trying to determine and explain cinematic properties of their BLR, according to the authors, a strong point of

their research is the amount of data collected, which allegedly doubles the amount of data previously available related to AGNs.

V. Dataset Collection

The main dataset for our study is the Sloan Digital Sky Survey (SDSS) Data Release 16. It consists of a catalogue of spectra, including millions of objects classified in three types: stars, galaxies and quasars, including their location in the sky (right ascension and declination) and their redshift index. It also includes a gallery of images from a wide angle telescope, including visible spectrum and specific emission lines. Catalog data is available online, we used the SciServer python package to connect to SDSS CasJobs (an asynchronous service that enables long running SQL queries) and execute SQL queries.

```
SELECT s.specobjid,
                    s.ra,
                    s.dec,
                    s.class,
                    s.z,
                    s.plate,
                    s.mjd,
                    s.fiberid,
                    s.deredSN2
                    FROM SpecObj AS s
                    WHERE z >= {}
                    AND z < {}
                    AND run2d ='v5_13_0'
                    AND zWarning = 0
```

SpecObj is the main view available in CasJobs to access the best quality spectra. Spectra are the result of passing the raw measurements from the spectrometer through a processing pipeline, and the format in which spectra are written to files changes with the version of this pipeline, for this reason we limit our results to spectra from the latest pipeline (i.e. 'v5_13_0') in order to ensure uniformity and make data management easier. zWarning is a bitmask with flags for potential issues in the spectra, we limit the search to entries without any bit on in this bitmask.

There are roughly 2.8 million spectra in the dataset that fit our requirements, but due to memory constraints in the SQLServer instance we must fraction the catalog query and locally join the resulting tables, we used z to do this segmentation, previously knowing that this value is between -1 and 10 for all entries in the dataset and adjusting the ranges to make sure all have a similar rowcount. The ranges used where [-1, 0.01), [0.01, 0.3), [0.3, 0.56), [0.56, 1) and [1,10).

The SDSS spectrometers were designed to be able to measure hundreds of spectra at the same time, this is achieved with the use of precision cut aluminium plates, with holes corresponding to the specific location of objects of interest in a field of view. Consequently with this design, any spectra in the dataset is uniquely identified by three numbers: plate (the id of the aluminium plate used during the measure), mjd (date of the observation) and fiberid (the id of the spectrometer fiber receptor connected to the hole in the aluminium plate).

The columns obtained are:

- specobjid: a unique id that identifies the spectra, hash computed from plate, mjd and fiberid.
- ra: right ascension of the measured object in degrees.
- dec: declination of the measured object in degrees.
- class: type of object, either GALAXY, QSO or STAR.
- z: calculated redshift of the object.
- plate: ID of the aluminium plate.
- mjd: modified julian date of observation.
- fiberid: id of the spectrometer receptor.
- deredSN2: calculated Signal to Noise ratio squared.

A first glance at the catalog reveals there are about 1.77 million galaxies, 666 thousand quasars and 385 thousand stars, for a ratio of approximately 6:2:1.

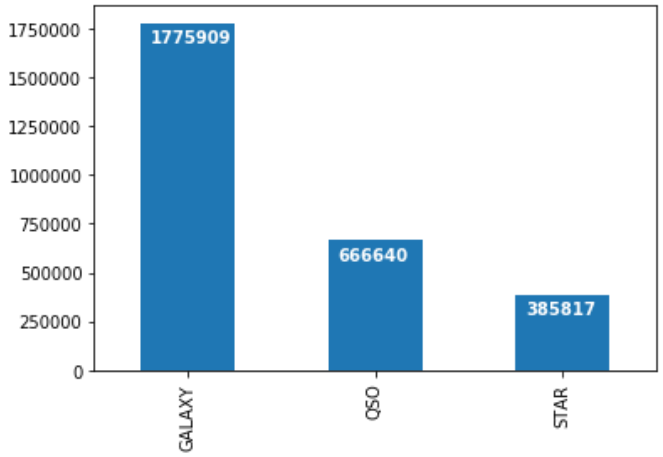


Fig. 4. Distribution of objects by class in the full dataset

Also, we get to see why the partition we used for z was not uniform, because data distribution is very skewed to the left (note the log scale), also this shows a lot of the spectra have a negative value for z, being this the reason for the very small upper limit in the first range.

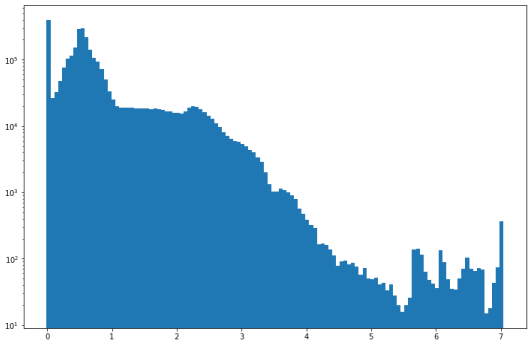


Fig. 5. Distribution of objects by class in the full dataset

Up to this point, we've only obtained the spectra catalog, but the actual spectra is a much bigger dataset weighing about 220kb for each spectra (almost 600GB for the whole 2.8 million observations). For our use case, we don't have the computing power, nor the storage, to work with such dataset.

For this reason, we're going to work with a tiny fraction of it, made of about 110000 spectra (totalling a much more manageable 22 GB). As our main class of interest is Quasars, we're going to choose our sample with a ratio of 5:4:2, to be sure quasars are well represented without altering too much the actual proportion. For our sample, we're going to use the cleanest available spectra, and the plots ahead will show that most of the entries have a very good Signal to Noise ratio (i.e. less than 20%) which is expected because the used view only contains the best observations from the SDSS dataset.

VI. Exploratory Analysis and Sample Selection

Since we're going to perform classification, we're going to see the behaviour of SNR for each class, this is important in order to make sure our small sample is representative of the whole dataset.

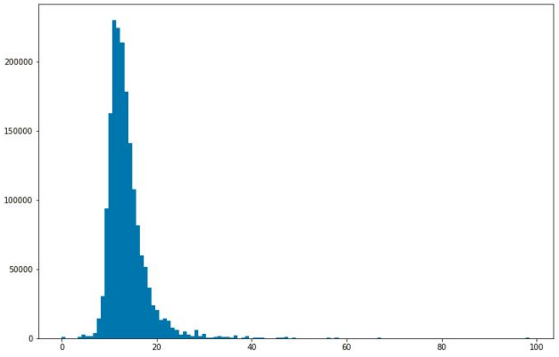


Fig. 6. Signal to Noise Level SNR for Galaxies

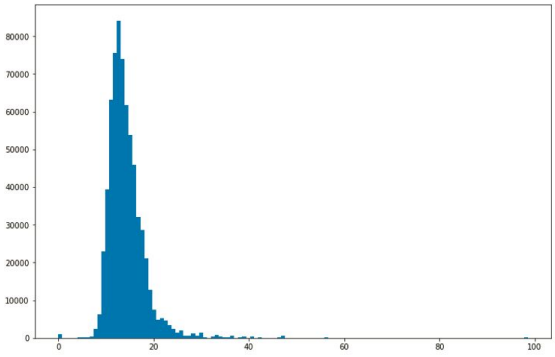


Fig. 7. Signal to Noise Level SNR for Quasars

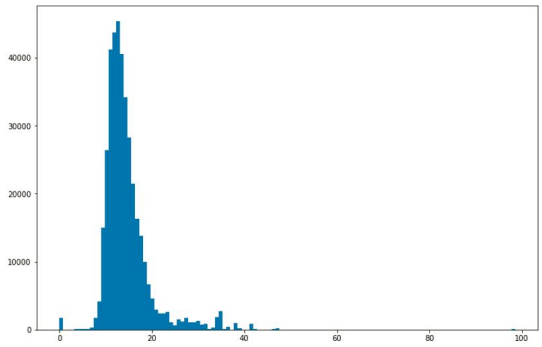


Fig. 8. Signal to Noise Level SNR for Stars

This shows that for all the classes the low SNR trend holds, and that we can reach our target sample size by using only entries with desired SNR2 below 9% for galaxies and 10% for quasars and stars. But we also want to make sure the ra, dec and z spaces are well represented in this sample.

In Fig 7, we can see the distribution of all kinds of objects is not uniform across all the sky, with a big patch in the middle of the plot. The region of the sky that isn't observed in the dataset corresponds to the Milky Way, and this happens because it obstructs our view, so as the targets of the survey lie beyond our galaxy they can only be visible in other directions. Outside this region, the distribution appears to be uniform.

Also, we can see that the distribution for z varies a lot between classes, being very small for stars, varying from 0 to 2 for galaxies and going up to 7 for quasars.

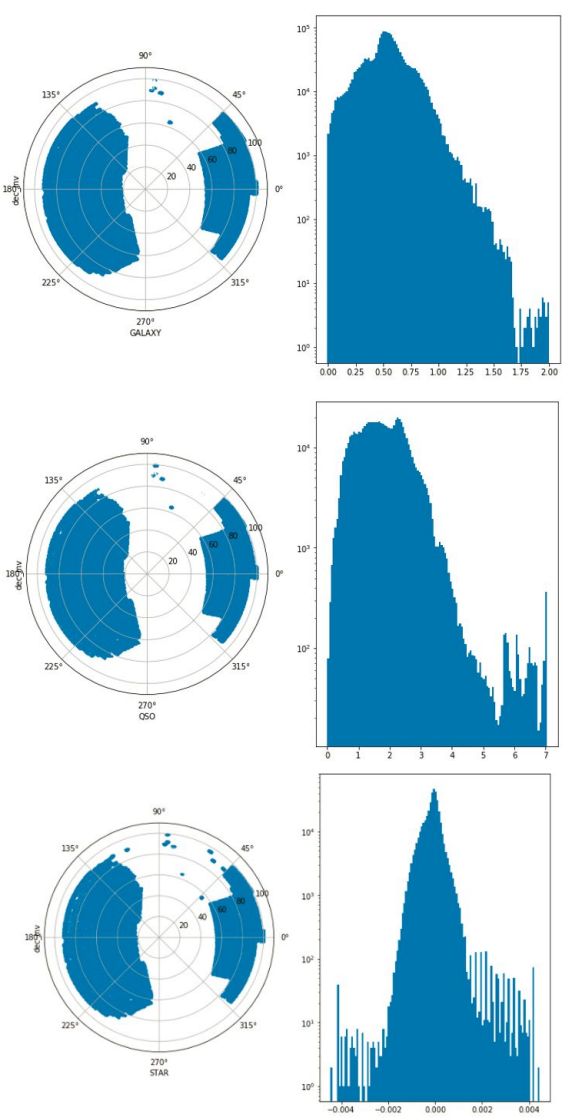


Fig. 9. Left: Distribution of observed Galaxys, QSOs and Stars in the northern hemisphere night sky. Right: Redshift z for Galaxy, QSO and Star

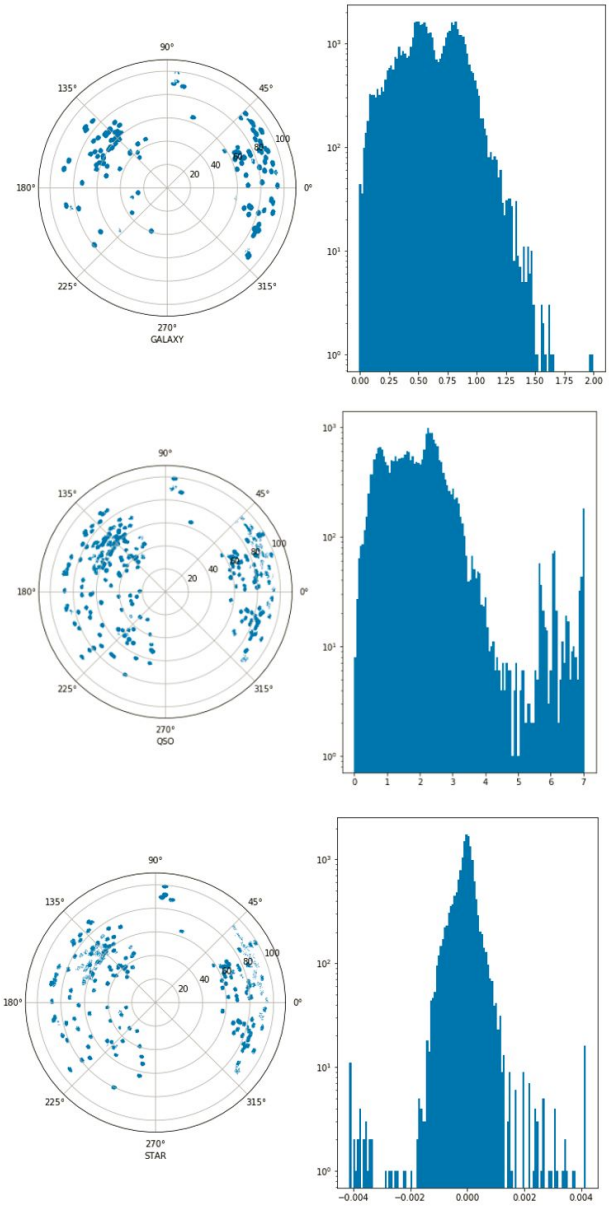


Fig. 10. Distribution of chosen sample

The Fig 7. show the distribution of objects in the full dataset and our reduced sample Fig 8. For z, the distribution is kept pretty well, while the distribution of locations has clusters in our sample but is “uniform” in the full dataset. This is probably because SNR2 has the same value for all spectra from a plate observed in a given date, this means that most of the best 110000 spectra probably come from the best 200 to 400 (plate, mjd) combinations, this results in a clustered sample, with a cluster corresponding to each of this best plates and most of these having 250 to 500 spectra belonging to it.

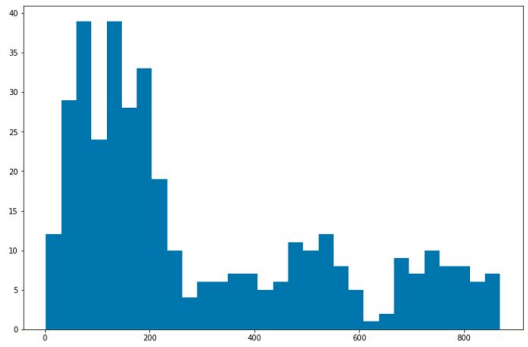


Fig. 9 Spectra number distribution respect to the plate

This shows that the hypothesis about clustering relation to plates is most likely true, with many plates of over 500 entries and a big majority having more than 100. In total, the 112.000 chosen spectra come from only 378 plates, each of which creates a cluster of observations.

Spectra download. Each spectra is stored in a FITS file in the SDSS servers, and is downloadable from a service called SAS. We generated the download url for each file and stored it in a text file, which was passed as input to wget to perform a batch download.

FITS files description. To read and process the FITS files, we used the astropy package, which has methods to generate a pandas DataFrame from the tables inside a FITS file.

Each FITS file is composed of 4 tables. The first one PRIMARY is proper of the FITS format, and is not used. The second one has most of the data, having always 8 columns and a varying amount of rows, this is the main piece of data we're going to use, containing the measured light flux for each wavelength, as well as the calculated portion of this flux corresponding to the sky and the observed object respectively and the parameters of a fitted model (according to the 'v5_13_0' algorithm). The third is a wide table with one row and 236 columns, this contains the same information available in the specObj table in the SQLServer, so as we already collected the information we need we're not going to use it. Finally, the last table contains a lighter summary of the spectra, for 32 wavelength regions, since some of our classification strategies take advantage of greater detail, we preferred to use the more detailed information in the second table.

This table contains the next columns:

FLUX: Calculated amount of light coming from the observed object (instead of the atmosphere/sky) for this wavelength.

LOGLAM: Base 10 logarithm of the current wavelength, each row has a 0.0001 increment respect to the previous one.

IVAR: Inverse of the calculated variance of the measured flux with respect to the best fit model, this can be understood as the confidence of the measurement, with IVAR=0 meaning no confidence at all (meaningless value) and IVAR$\rightarrow\infty$ being almost certain.

AND_MASK: The measured spectra is the result of 3 observations, this bitmask stores warning flags that were present for the current wavelength in ALL of the observations.

OR_MASK: The measured spectra is the result of 3 observations, this bitmask stores warning flags that were present for the current wavelength in ANY of the observations.

WDISP: The fitting model uses a wavelength dispersion parameter to weight the influence a given measured wavelength has on its neighbours best-fit model.

SKY: Calculated amount of light coming from the atmosphere/sky (instead of the observed object) for this wavelength.

MODEL: Expected flux from the object according to the best fit model.

Joining the downloaded files into a single dataset. A total of 112.683 files were successfully downloaded, but having that many separated files introduces an extra level of complexity, and can become time and resource consuming. Also, the FITS files contain header information that repeats for each file and increases disk space, and some other information about the spectra that we're not going to use in our classification problem. For this reason we decided to join them in a single dataframe with a row for each observation, doing this in a file per file basis and then concatenating them in a single operation proved to be an expensive and RAM heavy process beyond the server capacity, so the data was fractionated in 113 buckets, each with 1.000 files (683 for the last one).

For each of the resulting folders a single DataFrame was created with the same format of the table in the FITS file, with an extra column for the file name. The resulting dataframe list was dumped in separate joblib files.

The process was repeated to create 4 dataframes with 30.000 spectra each (22.683 for the last one), these in turn where joined in pairs (resulting in two DFs one with 60.000 spectra and the other with 52.683) and these were subsequently joined in a single final Data Frame. The reason this was needed (having enough memory to fit the final dataset three times) is not clear, but probably is due to implementation details of the pandas library that make the concat() method with an iterable of DataFrames as input very memory hungry.

Determining optimal range for wavelength. Although the step between LOGLAM values is fixed in all files, and there are complete ranges for each observation, this range varies from one

observation to the other. If we attempt to pivot the table as is, without first filtering for an optimal range of wavelengths, a single outlier with a too high or too low value for LOGLAM will hugely increase the amount of columns for the whole dataset, most of which would be filled with NaN anyway. This would unnecessarily increase the memory requirements for the dataset, the storage space, the complexity of the classification algorithms to use, and all this without a foreseeable improvement for the generated model.

For this reason, we'll look at the distribution of the start and end of the range for all files to determine an optimal range that is completely present in most of the observations.

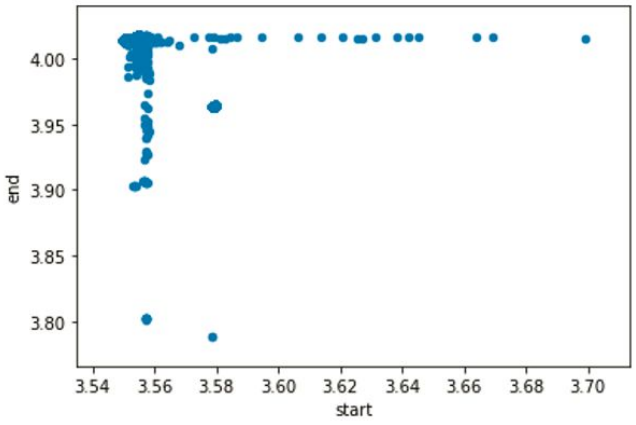


Fig 10. Distribution of LOGLAM (Base 10 logarithm of the current wavelength)

This shows that most of the spectra (over 99%) in the sample has the whole range between 3.5600 and 4.000. This range also contains the most important emission and absorption lines in astronomy, so we used it as our wavelength range and filtered out entries that are outside.

Done this, we finally proceeded to pivot the table by LOGLAM, using FileName as index. Getting 112683 rows × 30807 columns

Joining catalog data and spectra. We then proceeded to join the spectra and catalog tables, using the filename as key. This generated a single dataframe with all the spectral data for each observation (first 30807 columns) as well as the general information of the object (redshift, correct classification, SNR, etc.).

Creating reduced datasets for experimentation. Given the huge size of the complete dataset, we also exported a small (5%) and a medium (20%) dataset for use during experimentation, leaving the full dataset to be used only in the end when the model training code had been tested with the smaller datasets.

I. Data Cleansing

We perform the deletion of ID type characteristics and rows with empty data or NaN and encode class labels (Quasar,Galaxy and star) to integer (0,1,2). Due to the important number of features and Data we normalize the dataset for a fastest computation.

```
#normalize the data between 0 and 1 for a  fastest computation
scaler = MinMaxScaler()
sdss = scaler.fit_transform(df_fe.drop('class', axis=1))
```

```
#Verification that there is no NaN
df.isnull().sum()
```

```
FLUX_3.5600    0
FLUX_3.5601    0
FLUX_3.5602    0
FLUX_3.5603    0
FLUX_3.5604    0
              ..
class          0
z              0
plate          0
mjd            0
deredSN2       0
Length: 30814, dtype: int64
```

Fig 11. Normalize the data and verification there is no NaN

VII. Dimensionality Reduction

Due to the size of our dataset 112683 rows × 30807 columns and the high number of characteristics we perform a dimensionality reduction. This allows us to reduce the resources necessary to pass said data through a supervised algorithm, finally facilitating the interpretation of the data. For this we used three different approaches: Linear Discriminant Analysis, Principal Component Analysis and SelectKBest.

Reduction of the number of features

LDA. Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. This technique for dimensional reduction could be good for our project because we have a classification problema.

	0	1
0	-6.139914	-1.261268
1	-4.592996	-0.777152
2	-4.906928	-2.972774
3	7.947040	-1.173472
4	7.785666	-1.185137
...	...	...
14960	7.634269	-0.503582
14961	6.897308	0.598619
14962	-5.325301	-3.571202
14963	-4.128613	-3.936954
14964	5.070271	-1.949887

14965 rows × 2 columns

Table 1. Results of the reduction of number of Features using LDA

PCA. The main linear technique for dimensionality reduction, principal component analysis, performs a linear mapping of the data to a lower-dimensional space in such a way that the variance of the data in the low-dimensional representation is maximized. In practice, the covariance (and sometimes the correlation) matrix of the data is constructed and the eigenvectors on this matrix are computed. Contrary to LDA it's unsupervised but we can have more features on the final dataset. We reduce the space to 30 features because we know that the key emission and absorption lines are about 15, so we assumed a 30D space would give us enough information.

	0	1	2	3	4	5	6	7	8	9	...
0	-1.706060	-6.773781	-0.026301	0.889020	1.206097	3.502681	0.351938	-1.039640	0.738321	-4.507248	...
1	-6.600835	2.627868	1.872530	-0.153218	1.843666	2.199340	-1.762125	-0.138497	-0.023463	-1.951324	...
2	3.106851	-1.435163	-2.670352	-2.033955	-4.763781	0.737731	4.280189	2.426484	-2.991748	0.954137	...
3	1.274700	4.725868	0.036895	1.758481	-2.269102	-4.738896	2.729188	1.399094	-2.146195	-1.697120	...
4	-6.943780	3.614102	0.500150	2.086964	1.651064	-5.464588	0.091410	-0.199750	0.290330	-0.045911	...
...	...	...	...	...	...	...	...	...	...	...	...
22331	-2.713739	1.661335	4.942405	0.585471	-7.050896	-3.488987	2.501322	1.317013	1.750856	0.303643	...
22332	1.997167	-4.854724	-2.746761	1.284289	-2.717720	2.723163	-0.965520	-1.095721	-0.840725	-0.713381	...
22333	-2.389830	-1.381036	-1.326966	-2.056764	-4.036133	1.012858	1.603253	0.576943	-1.515933	2.522334	...
22334	-5.658409	-3.702827	-2.693620	0.295591	-2.606961	1.072154	-2.787840	-1.165192	-1.178341	-0.104087	...
22335	-5.917265	-1.264396	-4.999173	-1.050401	1.443005	1.380734	0.082855	-0.696315	0.070575	0.458118	...

22336 rows × 30 columns

TABLE 2. RESULTS OF THE REDUCTION OF NUMBER OF FEATURES USING PCA

Feature Selection SelectKBest. For the feature selection we will select the 30 best features for the same reasons as the PCA methods.To choose the best feature we will use the python function which is SelectKBest. We will use Chi2 as a score function because we have a classification problem, SelectKBest will compute the Chi2 statistic between each feature of X and y (assumed to be class labels). A small value will mean the feature is independent of y. A large value will mean the feature is non-randomly related to y, and so likely to not provide extra information. Only 30 features will be retained.

	SCORE OF THE FEATURE
AND_MASK_3.6131	9333
AND_MASK_3.6376	9578
AND_MASK_3.6377	9579
AND_MASK_3.6378	9580
AND_MASK_3.6393	9595
AND_MASK_3.6394	9596
AND_MASK_3.6395	9597
AND_MASK_3.6868	10070
AND_MASK_3.6869	10071
AND_MASK_3.6870	10072
AND_MASK_3.9793	12995
OR_MASK_3.5713	13316
OR_MASK_3.5714	13317
OR_MASK_3.5715	13318
OR_MASK_3.5716	13319

	SCORE OF THE FEATURE
OR_MASK_3.5717	13320
OR_MASK_3.6376	13979
OR_MASK_3.6377	13980
OR_MASK_3.6378	13981
OR_MASK_3.6867	14470
OR_MASK_3.6868	14471
OR_MASK_3.6869	14472
OR_MASK_3.6870	14473
OR_MASK_3.6871	14474
OR_MASK_3.6872	14475
OR_MASK_3.8272	15875
OR_MASK_3.8273	15876
OR_MASK_3.8274	15877
OR_MASK_3.8284	15887
z	30809

TABLE 3. RESULTS OF THE REDUCTION OF NUMBER OF FEATURES USING SELECTKBEST

Here is an interesting result because these characteristics chosen by the model are not directly values of the fluxes at the different wavelengths, which we expected. The redshift appears as the characteristic with the highest score, this result is consistent with our exploratory plots, where we saw the distribution of z varies a lot between classes, so this feature is a great first step in classification. The surprising result is that all the rest of the columns point to flag columns, where there are warnings about specific circumstances presented during the collection or processing of some information about certain wavelengths associated to emission lines, for example, the 3 features with the highest score after z correspond to the emission lines of Sii, one of the most commonly observed emission lines in astronomy and astrophotography.

VIII. CLASSIFICATION

Our main objective is to assign the correct class to the observed objects, this translates into a classification problem. Classification problems are characterized by having a qualitative variable Y as an answer (Galaxy, Quasar or Star). These qualitative variables are also called categorical variables.

Predicting a qualitative response to an observation is called classifying that observation, that is, predicting the category or class of that observation. Often the methods in charge of classifying what they do is predicting the probability of an observation belonging to each of the categories. In a way they also behave like regression algorithms but without being.

To carry out this classification, 7 classification methods were tested with all methods used for the reduction of the dataset. The tested methods were: K-Nearest Neighbors, Naive Bayes classifier, Random Forest, Support Vector Machine SVM , two flavors of gradient boosting for decision trees, and a multilayer perceptron.

As they are supervised learning methods, we split in train and test (33% of the dataset for the test) the 3 dataset which are the result of 3 methods to reduce the dataset. Separation for SelectKbest, Separation for PCA the Separation for LDA is done when we reduce the dimensión.

For all methods, to find the best parameters we performed a grid search.

KNeighborsClassifier. KNN works by finding the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification)

TABLE 4. KNEIGBORS CLASSIFIER PREDICTION ACCURACY

KNeighborsClassifier		
KBEST	**PCA**	**LDA**
88,63%	76,66%	77,4%

Naive Bayes classifier. Primarily Naïve Bayes is a linear classifier, which is a supervised machine learning method and works as a probabilistic classifier as well. When handling real-time data with continuous distribution, Naïve Bayes classifier considers that the big data is generated through a Gaussian process with normal distribution.

Table 5. Naive Bayes Classifier Prediction Accuracy

Naive Bayes Classifier		
KBEST	**PCA**	**LDA**
36.89%	55,52%	51,97%

Random Forest. The random forest is a classification algorithm consisting of many decision trees, for aur case 15 trees. It uses bagging and features randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

Table 6. Random Forest Classifier Prediction Accuracy

Random Forest		
KBEST	**PCA**	**LDA**
88,97%	74,64%	76,65%

Support Vector Machine. SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

Table 7. Support Vector Machine SVM Classifier Prediction Accuracy

SVM Support Vector Machine		
KBEST	**PCA**	**LDA**
73,53%	40,25%	68,65%

Catboost. CatBoost is an algorithm for gradient boosting on decision trees. Developed by Yandex researchers and engineers.

Table 8. Catboost Classifier Prediction Accuracy

Catboost		
KBEST	**PCA**	**LDA**
91,18%	79,55%	73,31%

XGboost. Like CatBoost is an algorithm for gradient boosting on decision trees.

Table 9. Xgboost Classifier Prediction Accuracy

XGboost		
KBEST	**PCA**	**LDA**
91,07%	79,21%	76,91%

Neural Network. Using tensorflow, we implemented a multi layer perceptron for the classification. After trying different architectures, we chose a neural network of 5 layers:

- The number of neurons of the first layer is the numbers of column of the dataset
- The number of neurons of the second and third layers are 10 and for the fourth the number is 6
- The number of neurons for the ultimate layer is 3, one for each class

Table 10. Multi Layer Perceptron Classifier Prediction Accuracy

NeuralNetwork		
KBEST	**PCA**	**LDA**
90,66	76,88	76,54

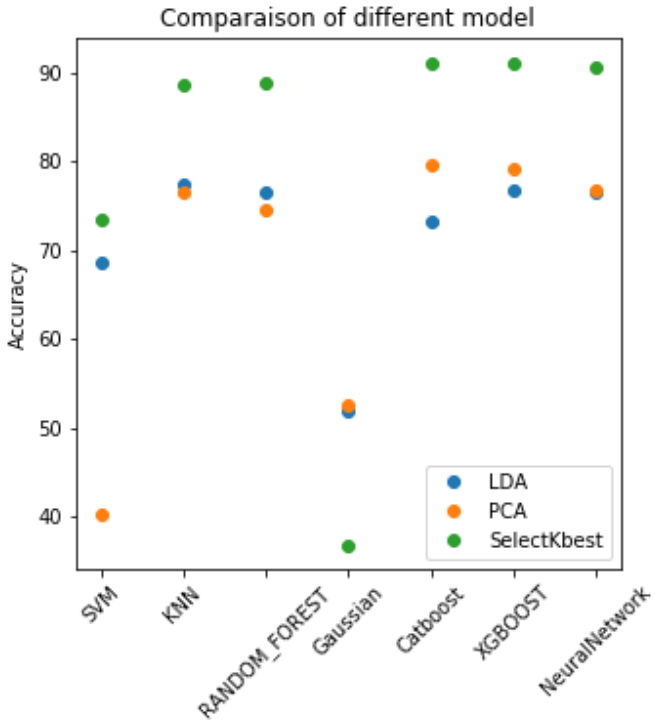


Fig 11. Comparison of classifiers with respect to the reduction of features methods

Comparison of classification models

As we can see, SelectKbest is the best method for the reduction of the dataset and the best classifier with this method is the Catboost with an accuracy of 0,9118.

Gaussian (Naive Bayes Classifier) doesn't work very well, and this could be explained by applying a normality test to each feature, our guess is they probably are not normally distributed.

As for the other classification approaches, they are not too far from each other, but the gradient boosted decision trees came on top by a small margin. This makes a lot of sense watching the columns chosen by the k-best algorithm, as the only one that is not discrete is z, and the rest are bitmasks, that is, a lot of yes/no conditions, great to build decision trees.

VII. Conclusion and Future Work

The k-best algorithm proved to be a really interesting approach to dimensionality reduction when there is little domain knowledge, and the dataset size is too large to analyze completely. In this case, not only it provided us with a reduced set of features (from over 30.000 to just 30, it discarded 99.9% of the columns) that where enough to achieve a decent classification, but it also gave us insights about what to look in a conceptual manner from our data, for example, it pointed us to the importance of Sii, H-Betha and H-Gamma emission lines in the analysis of astronomical objects, as the 12 features with highest scores after z were related to these.

One of the main criticisms of the supervised classification models is their difficulty in extrapolating beyond the limits of the data used. It would be interesting how well this approach works with noisier spectra, or spectra from different pipelines.

References

[1] S. Kaspi, P. Smith, H. Netzer, D. Maoz, B. Jannuzi and U. Giveon, "Reverberation Measurements for 17 Quasars and The Size-Mass Luminosity Relations in Active Galactic Nuclei", The Astrophysical Journal, 533:631-649, April 2000.

[2] Department of Education Open Textbook Pilot Project. (2019, Sep 29). Spectrophotometry [Online]. Available: https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Kinetics/Reaction_Rates/Experimental_Determination_of_Kinetcs/Spectrophotometry

[3] N. Taylor. (2018, Feb 24). Quasars: Brightest Objects in the Universe [Online]. Available: https://www.space.com/17262-quasar-definition.html

[4] Wikimedia Foundation (2019, Dec). Reverberation Mapping [Online] Available: https://en.wikipedia.org/wiki/Reverberation_mapping

[5]The University of Alabama Huntsville (2020, Jan) Tuesday Physics Seminar: Active Galactic Nuclei [Online]. Available: https://www.uah.edu/events/icalrepeat.detail/2020/01/07/11778/-/tuesday-physics-seminar-active-galactic-nuclei

[6] Celestron.com. 2020. What Are RA And DEC? | Celestron - Telescopes, Telescope Accessories, Outdoor And Scientific Products. [online] Available at: <https://www.celestron.com/blogs/knowledgebase/what-are-ra-and-dec> [Accessed 22 May 2020].

[7] En.wikipedia.org. 2020. Right Ascension. [online] Available at: <https://en.wikipedia.org/wiki/Right_ascension> [Accessed 22 May 2020].

[9] Sloan Digital Sky Survey(2020, Jan) The SDSS Telescopes [Online]. Available: http://cas.sdss.org/stripe82/en/sdss/telescope/telescope.asp

[10] J. Pasquet-Itam,and J. Pasquet, "Deep learning approach for classifying, detecting and predicting photometric redshifts of quasars in the Sloan Digital Sky Survey Stripe 82", Astronomy Astrophysics A&A Volume 611, March 2018. https://doi.org/10.1051/0004-6361/201731106

[11]Skyserver.sdss.org. 2020. SDSS Skyserver DR16. [online] Available at: <http://skyserver.sdss.org/dr16/en/help/howto/search/practice2.aspx> [Accessed 22 May 2020].

[12] Khnifassi, C., 2020. Google Colaboratory. [online] Colab.research.google.com. Available at: <https://colab.research.google.com/drive/1mnTyPmXp-R7bt_gs1GayQJ0VRQkBsaVI?usp=sharing#scrollTo=DKFKlFKK6m6R> [Accessed 22 May 2020].

[13] Silva, J., 2020. Google Colaboratory. [online] Colab.research.google.com. Available at: <https://colab.research.google.com/drive/12485Hi2JlW2eDL6YOL-RUk4BQtbFMsdh> [Accessed 22 May 2020].

[14]Montoya, A., 2020. Google Colaboratory. [online] Colab.research.google.com. Available at: <https://colab.research.google.com/drive/1mnTyPmXp-R7bt_gs1GayQJ0VRQkBsaVI?usp=sharing#scrollTo=DKFKlFKK6m6R> [Accessed 22 May 2020].

[15] SQL Tutorial. Sloan Digital Sky Survey(2020, Jan) The SDSS Telescopes [Online]. Available: http://skyserver.sdss.org/dr16/en/help/howto/search/introduction.aspx