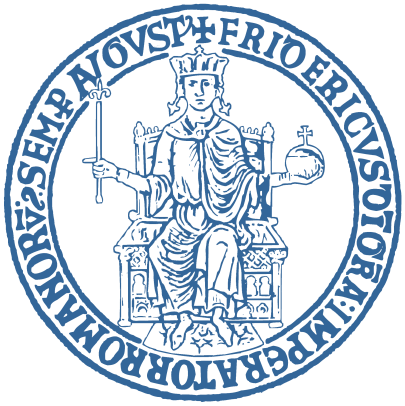
Università degli Studi di Napoli
"Federico II"
Scuola Politecnica e delle Scienze di Base
Area Didattica di Scienze Matematiche Fisiche e Naturali
Dipartimento di Fisica



Identification of photo-z outliers with ML methods

Relatori:
Prof. Giuseppe Longo
Dott. Michele Delli Veneri
Dott. Stefano Cavuoti

Candidato:
Matteo Rossi
Matricola:
N85000247

A.A. 2020/2021

But I also know that "the powers of instruction are of very little efficacy except in those happy circumstances in which they are practically superfluous"
- Richard Feynman

Contents

List of Figures

List of Tables

Introduction

In order to understand the intrinsic properties of many astronomical objects, it is necessary to measure their distances. Since at this scale direct measures are impractical, the estimation of the distance presents itself as a complex task. While many techniques were developed to measure distances for a specific scale, none of them could be adopted at all ranges encountered in astronomy.

This lets us introduce the Distance Ladder, a collection of methods developed to measure distances based on the luminous source type and the distance range. Stellar Parallax, which relays on principles of trigonometry, was the first used to measure distances of stars close to the solar system. As the Earth orbits the Sun, nearby stars appear to slightly shift in position when compared to the more distant ones. Thus, by measuring the shift entity when the Earth is in two different orbit positions the distance of the nearby star can be calculated. However, for remote stars, the parallax angle is too small to accurately measure distance and non-direct methods of distance determination are required. Further developed techniques to measure distances will then rely on the magnitude of the source observed. A standard candle is an astronomical object that has a known absolute magnitude, thus we can determine its distance by measuring its apparent magnitude. Two examples of this kind of object are the Cepheid variable stars and the Type Ia supernovae. Cepheids are horizontal branch stars that lie in the instability strip of the Hertzsprung-Russell diagram. Instabilities cause size and temperature periodic variation, also reflecting on their luminosity. Henrietta Leavitt established a relationship between the period of a star's pulsation and its average apparent magnitude, and later Hertzsprung used properties of the Cepheids' light curves and statistical parallax to arrive at an estimated distance to the Magellanic clouds. The furthest Cepheid measured so far, with the use of the Hubble Space Telescope, is up to a distance of 10^8 light-years. Greater distances make individual stars undetectable, so astronomers switch the focus to one of the brightest events in the universe: the supernovae. In particular, Type Ia occurs when a white dwarf belonging to a binary system accretes matter from its companion to the point of instability, which happens when the Chandrasekar limit is exceeded.

Type Ia supernovae therefore all start at essentially the same mass and produce the same light curve with a known absolute magnitude at various stages of the event. Thus by examining its light curve and measuring the supernova's maximum apparent magnitude, because their maximum absolute magnitude is known, the distance to the supernova can be determined.

However, for objects far over the billion light-years, the previously mentioned techniques are inefficient and observations must be used in conjunction with the theory of cosmological expansion, in particular the Hubble Law. The Hubble constant reflects the rate at which the universe is expanding, thus, to determine an object's distance we only need to know its velocity, which is measurable due to the shift of the source's spectral lines also known as redshift. Redshifts lay at the base of almost all studies of the extragalactic universe and in many scientific research fields such as astronomical sources classification or to understand the cosmic large scale structure. Historically, redshifts have been measured with spectroscopy and several spectroscopic surveys have been done in the past with some being still active nowadays (zCOSMOS [1], VANDELS [2]). These surveys, however, are very time consuming and cannot follow up today modern precision cosmology which is based on samples of many millions of galaxies, with the redshift estimation through multi-band photometry (hereafter photometric redshift or photo-z [3]) becoming an indispensable tool [4]. However, this greatly increased redshift estimation capability comes at a price, namely their much lower accuracy with respect to spectroscopic measurements. Many methods and techniques for photo-z estimation have been tested on a large variety of all-sky multi-band surveys (see for istance KiDS [5, 6], DES [7]). These methods are broadly split into two large groups: physical template models fitting the Spectral Energy Distributions or the empirical exploration of the photometric parameter space (defined mainly by fluxes and derived colors). At the core, they create a mapping between photometrical parameters such as magnitudes, fluxes or colors and the redshift in order to obtain the redshift solution and its associated Probability Distribution Function (PDF), but the way in which they build this mapping is very different. Template methods operate by comparing the observed features with template models and selecting the model which best fit the observed features. For example a redshift estimation could be taken by properly shifting a model spectral energy distribution over the observed one. Their main advantage is the understendability of the results but they are bound by the physical assumptions about the observed galaxies. Machine Learning methods, on the other side, are not based on any physical assumption and the mapping is obtained by learning it directly from the data. Among the several factors which influence the quality of the produced photometric redshifts, the algorithmic and physical choices are for sure of

great importance. For such a reason, to understand the impact of these choices on the photo-z quality, over the years, many data challenges have been designed [8, 9, 10]. To cite one remarkable example we briefly introduce the Euclid survey and data challenge. The Euclid survey is a photometric and spectroscopic survey, that during its nominal mission of 6 years, will survey 15,000 deg^2 of extragalactic sky with a 1.2 m-diameter space telescope. The main scientific goal is to investigate the Universe's accelerating expansion through two main probes, baryonic acoustic oscillations and weak-lensing tomography. The Euclid photo-z challenge [10], was designed to test the performances of 13 different models on the simulated data product of the Euclid survey [11] and to evaluate if said methods will be capable of providing redshifts with the required precision. Given that the Euclid data product is not yet available, the challenge designers have modified images provided by COSMOS [12] in order to resemble the EUCLID data product. The data was then split equally into two subsets, with the first being used for the calibration of the different methods and the second one being used to assess their performances. The results of the challenge showed that each method has its advantages and disadvantages, and thus performs efficiently in different regimes. Machine-learning methods are based on a training sample and their results depend strongly on the quality of this training. Template-fitting methods do not have this problem and perform relatively well for sources in regions of the color space with a sparse redshift coverage. This work was inspired by the results of this challenge and its scope is to study the possibility of combining some models in an automatic way in order to increase the overall accuracy of the photometric redshift estimation. Given that the direct optimization of the PDFs was a problem too complex to fit in the time frame of a Bachelor Thesis, we decided to tackle a much easier but needed problem of determining if the point estimations quality could be optimized. With point estimation, or photometric redshift, we indicate the highest probability value for each PDF. The initial idea on how to improve the estimation was to try and see if specific parts of the photometric parameter space exist in which one or a combination of models performs better than others. If one could individuate such sections, for each galaxy, in an automatic way and using only the galaxy photometry, then, the most suitable model could be assigned to said galaxy in order to receive the best possible estimation of its redshift. On the other side, in reality, the redshift for most galaxies tends to be correctly classified by most models, so, in order to improve the overall accuracy of the estimations, we decided to turn the logic around and see if we could find the regions of the photometric parameter space were model performs badly. This requires, first, to distinguish outliers (galaxies that received incorrect photometric redshift estimations) from galaxies for which the redshifts were correctly regressed on the basis of their photometry and then to

recognize which model produced the outliers (from one specific to all models) in order to substitute them with correct predictions from other models. We thus investigated the possibility of solving this problem with two Multi-Layer Perceptrons.
The presented work is organized as follows:

in Chapter 1, we introduce the basic concepts of Machine Learning necessary to develop our pipeline, we describe the Multi-Layer Perceptron architecture, the concept of activation and loss functions, the concept of gradient descent, and the Error backpropagation algorithm. The chapter is concluded by the section Other Methods (Sec. 1.2) in which we make a brief overview of the Machine Learning and Template Fitting methods that produced the photometric redshift estimations analyzed in this work.

In Chapter 2, we outline all the astrophysics concept required to understand the data: the photometric redshift (Sec.2.1), what it is and why its estimate's optimization is a major objective, as well as the concept of magnitudes (Sec. 2.2) and color index (Sec. 2.3). The astronomical surveys that gathered the data used by both us and by the methods to produced the redshift estimations are described in Sec. 2.4. We show the data properties in the Data Collection section (Sec. 2.5) and in Data Pruning (Sec. 2.6) we outline the data cleaning procedures applied before feeding it to our pipeline. In Chapter 3, we outline the experiments: in the Data Preparation section (Sec. 3.1), we present the subset of features selected to train our model, along with the criteria adopted to define the two sets of classes for the classifications tasks and the evaluation metrics needed to assess the models' performances over the photometric redshift estimations. Moreover we outline the Multi-Layer Perceptrons' training procedures. In the Experiments section (Sec. 3.2), we overview the models' setup and their training phase, and we analyze the results obtained on the validation and test sets. This section also describes the labels aggregation, introduced to overcome the class imbalance affecting our data, and the implementation of the outliers hierarchical replacement strategy developed to substitute outliers with correct predictions made by other models. This thesis ends with the Conclusion chapter, in which we recapitulate the workflow of our experimental procedure, briefly summarize our findings and we outline future prospects.

Chapter 1

Methods

In this section, we are first going to make a brief review of the main concepts of machine learning needed to explain the implementational choices of the models employed to solve the scientific problem discussed in this work, i.e. two Multi-Layer Perceptrons [13, 14]. Furthermore, a brief explanation of the other techniques used in this work for exploratory analysis is also provided.

The main focus of machine learning is making decisions or predictions based on data. As machine learning (ML) methods have improved in their capability and scope almost any application that involves understanding data or signals that come from the real world can be addressed using ML. Great examples are remote sensing [15], speech recognition [16], and many kinds of language-processing tasks [17].
Machine learning models are named after the type of problems which they tackle, in particular, supervised models use a set of examples to learn the functional mappings between a set of inputs variable and a set of target variables (discrete for classification and continuous for regression). When this functional mapping is learnt, through the minimization of a loss function, the model can be used to predict the target variable for unseen instances. In particular, in this work, we focus on the problem of classification.

1.1 The Multi-Layer Perceptron

The basic idea for a ML model inspired by neurons goes way back to 1943 [13]. There were good training methods (e.g., perceptron) for linear functions, and interesting examples of non-linear functions, but no good way to train non-linear functions from data. Interest in the model was lightened up in the 1980s when several people came up with a way to train neural networks with "back-propagation", a specific form of

gradient descent [18]. The Multi-Layer Perceptron architecture is one of the most typical feed-forward neural network models. The perceptron (depicted in Figure 1.1), also known as neuron since was inspired by its biological counterpart, is the basic block of the MLP and represents a mathematical function that takes the weighted sum of the inputs and passes it through a nonlinear activation function.

$$a = f(z) = f\left(\sum_{j=1}^{n} w_j x_j + w_0\right) \tag{1.1}$$

where $j = 1, ..., n$, w_j are the *weights* and w_0 is the *bias*. The quantities a is known as *activations* and f is a differentiable, non linear *activation function*. Neurons are organized into layers, where each neuron can be connected to the neurons of the next layer but not to the neurons of the same layer. The expression feed-forward identifies the fact that in this neural network model, the impulse is always propagated in the same direction, e.g. from the input layer to the output layer, passing through one or more hidden layers, by combining the sum of weights associated to all neurons except the input ones. To "learn" how to classify instances, the network must be trained by comparing iteratively the outputs of the network with "truth" (target variables). The error between the output of the MLP and the target variable is generally referred to as loss. To train the network, the weights are changed to minimize a loss function and the way the minimization is actually performed is by propagating back the loss through the network (backpropagation), changing the network's weights in the direction that maximizes the gradient of the loss function.

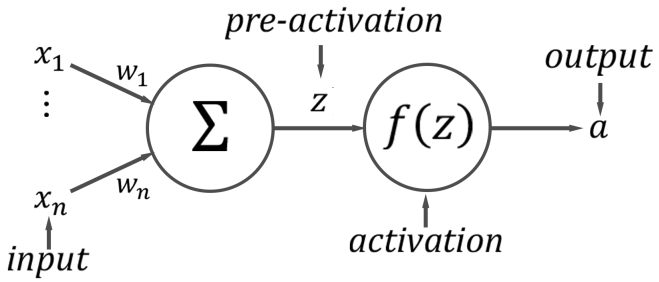


Figure 1.1: Graphical representation of Eq. 1.1 showing the flow of data through a neuron.

With matrix notation, we can rewrite the latter equation for the l-th layer as

$$A^l = f^l(Z^l) = f^l({W^l}^T A^{l-1} + W_0^l) \tag{1.2}$$

where $l \in (1, ..., L)$ is the layer index, W is the weight matrix, and A is the activation for said layer. In order to reduce the loss, the gradient of the loss function with respect to the weights is computed and thereafter the weights are modified accordingly. First, we want to display how the loss depends on the weights of the final layer W^L. Since the loss function takes as input the activation of the last layer, $A^L = f^L(Z^L)$ and $Z^L = W^{L^T} A^{L-1}$ through the application of the chain rule, we get:

$$\frac{\partial loss}{\partial W^L} = \frac{\partial loss}{\partial A^L} \frac{\partial A^L}{\partial Z^L} \frac{\partial Z^L}{\partial W^L} \tag{1.3}$$

Which, for the generic layer l, can be written as:

$$\frac{\partial loss}{\partial W^l} = A^{l-1} \left(\frac{\partial loss}{\partial Z^l} \right)^T \tag{1.4}$$

In order to find the gradient of the loss with respect to the weights in the subsequent layers of the network, we need to compute $\partial loss / \partial Z^l$ which, again, can be evaluated with the chain rule. This chain continues up to the input layer.

$$\frac{\partial loss}{\partial Z^1} = \frac{\partial loss}{\partial A^L} \frac{\partial A^L}{\partial Z^L} \frac{\partial Z^L}{\partial A^{L-1}} \frac{\partial A^{L-1}}{\partial Z^{L-1}} \cdots \frac{\partial Z^2}{\partial A^1} \frac{\partial A^1}{\partial Z^1} \tag{1.5}$$

$$\frac{\partial loss}{\partial Z^1} = \frac{\partial loss}{\partial A^1} \frac{\partial A^1}{\partial Z^1} \tag{1.6}$$

For the generic layer l the relation between the loss and the pre-activation of the layer is expressed by:

$$\frac{\partial loss}{\partial Z^l} = \frac{\partial A^l}{\partial Z^l} \cdot W^{l+1} \frac{\partial A^{l+1}}{\partial Z^{l+1}} \cdots W^{L-1} \cdot \frac{\partial A^{L-1}}{\partial Z^{L-1}} \cdot W^L \cdot \frac{\partial A^L}{\partial Z^L} \cdot \frac{\partial loss}{\partial A^L} \tag{1.7}$$

The gradient of the loss with respect to the weights is then obtained by combining Eq. 1.7 and 1.4. We are now able to express how the weights affect the loss and use this information to update the weights of the network in order to minimize the loss. The update criteria is showed below:

$$W_{new} = W - \eta \nabla loss \tag{1.8}$$

or for a specific layer:

$$W^l = W^l - \eta \frac{\partial loss}{\partial W^l} \tag{1.9}$$

where the η parameter is a scalar value defined as *learning rate* that determines the step size at each iteration while moving toward a minimum of a loss function.

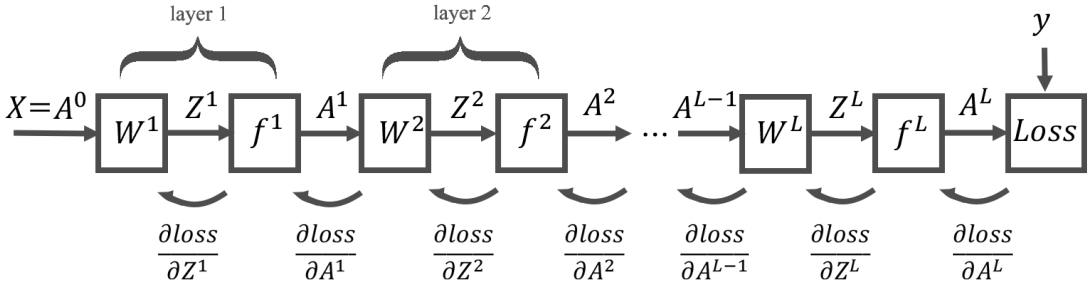


Figure 1.2: Error back-propagation example

For all the models used in this work, the nodes in the hidden layers share a common activation function, the Rectified Linear Unit (ReLU, [19]) which has been proven effective in many ML problems and is considered a standard in the field [20]:

$$\mathrm{ReLU}(z) = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{otherwise} \end{cases} = \max(0, z) \tag{1.10}$$

As activation function of the output layer of the MLP for the binary classification problem (detection of outliers), we used the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{1.11}$$

The output of the sigmoid, also known as the logistic function, lies in the [0, 1] range and can be interpreted as a class probability.

The recognition of the model related to the outlier photo-z's estimation can be designed as a multi-class classification task. In this case, we utilized the softmax function which is an extension of the sigmoid in case of multiple classes:

$$\mathrm{softmax}(z) = \begin{bmatrix} \exp(z_1) / \sum_i \exp(z_i) \\ \vdots \\ \exp(z_n) / \sum_i \exp(z_i) \end{bmatrix} \tag{1.12}$$

As loss function, we have chosen the Negative Log-Likelihood (NLL) which can be written as:

$$L(y) = -log(y) \tag{1.13}$$

where y is the feedforward output. Supposing that the classification problem contains j classes, if we define f as the vector of class scores for a given input, then we can rewrite the softmax output as:

$$p_k = \frac{e^{f_k}}{\sum_j e^{f_j}} \tag{1.14}$$

where f_k is an element for a given class k in all j classes, and the negative log-likelihood as:

$$L_i = -log(p_{y_i}) \tag{1.15}$$

where L_i is the loss for the i-th input and p_{y_i} is the output of the model. To perform backpropagation, we compute how the loss changes with respect to the output of the network and through the chain rule we get:

$$\frac{\partial L_i}{\partial f_k} = \frac{\partial L_i}{\partial p_k}\frac{\partial p_k}{\partial f_k} = -\frac{1}{p_k}(p_k * (1 - p_k)) = (p_k - 1) \tag{1.16}$$

which is the derivative of the negative layer with respect to the softmax output. Further derivatives are computed with the chain rule and network weights are updated.

The last piece needed to train the network regards the step size value η for the gradient descent. So far we have decided in which direction to update the weights but we have still to decide how big this change should be. If it's too small, then convergence is slow (chances are to get stuck in a local minimum) and if it's too large, then we risk divergence or slow convergence due to oscillation. This choice is crucial in order to reach our goal, to find the global minima of the loss and avoid possible local minima. The choice of a single global step size may be dangerous in a multi-layered network such as the MLPs employed in our experiments. This is because the magnitude of the gradient $\frac{\partial \text{loss}}{\partial W^L}$ may differ (by a considerable amount) between the first and the last layer. Adam [21] solves this problem by selecting an independent step size parameter for each weight matrix which is proportional to the mean and variance of the gradient in said layer. In particular, the weights at layer j are updated through the following equation:

$$W_{t,j} = W_{t-1,j} - \frac{\eta}{\sqrt{\hat{v}_{t,j} + \epsilon}}\hat{m}_{t,j} \tag{1.17}$$

Where ϵ is very small value introduced to avoid dividing by zero, t is the index of the iteration, W are the weights matrices and m_{tj} and v_{tj} are the estimated mean and variance of the gradient at layer j.

$$\hat{m}_{t,j} = \frac{m_{t,j}}{1 - B_1^t} \tag{1.18}$$

$$\hat{v}_{t,j} = \frac{v_{t,j}}{1 - B_2^t} \tag{1.19}$$

In the original work authors proposed the following values for the parameters: $B_1 = 0.9, B_2 = 0.99, \epsilon = 10^{-8}$.

1.2 Other Methods

Hereafter we briefly describe all the photometric redshift producing methods, divided into SED fitting method, like LePhare and Phosphoros, and ML methods, like METAPHOR, Random Forest and AdaBoost.

1.2.1 AdaBoost

AdaBoost [22] is a ML method based on ensemble learning, which combines several base algorithms to form one optimized predictive algorithm. Ensemble Learning methods are distinguished mostly by the way they combine weak learners. Several combination strategies such as bagging, boosting and stacking have been developed over the years [23, 24, 25]. AdaBoost is based on the boosting techniques in which different models are generated sequentially and the mistakes of previous models are learned by their successors. This aims at exploiting the dependency between models by giving the mislabeled examples higher weights. The models are defined as weak classifier, a definition which identifies models performing better at random guessing than correctly predicting objects. These weak classifiers are grouped in a way that each model learn from the previous model miss-classified object in order to build a better performing model. In the first step a weak classifier (e.g. a decision stump) is made on top of the training data based on the weighted samples. Here, the weights of each sample indicate how important it is to be correctly classified. Initially, for the first stump, we give all the samples equal weights. In the second step, after a weak classifier is created for each variable, the algorithm check how well each model classifies samples to their target classes. More weight is then assigned to the incorrectly classified samples so that they're classified correctly in the next model. Weight is also assigned to each classifier based on its accuracy, meaning that high accuracy it's translated into higher weight. Lastly the process come back to the second step until all the data points have been correctly classified, or the maximum iteration level has been reached.

1.2.2 LePhare

LePhare [26] is a template fitting method for photo-z estimation. The LePhare code is based on the χ^2 method described in [27]. The χ^2 merit function is defined as the sum over the different filters for the squared difference between the observed and the predicted fluxes divided by the error. The photometric redshift is then estimated from the minimization of χ^2 as a function of the parameters z (redshift), T (Template at redshift z) and a normalization factor A.

1.2.3 METAPHOR

The Machine-learning Estimation Tool for Accurate PHOtometric Redshifts (METAPHOR, [28]) is based on the Multi Layer Perceptron with Quasi Newton Algorithm model (MLPQNA, [29, 30]) and designed to produce the redshift point estimations and the PDFs. Given a data sample, it performs a random shuffle-split into a train and test set, with the photometry of the latter also being perturbed in order to return an arbitrary number N of modified test sets (by variable photometric noise contamination). The model then produces N+1 estimations for a specific photo-z, and after that, computes the probability that a given photo-z belongs to each bin with the resulting PDF being the set of all the probabilities obtained.

1.2.4 Random Forest

Random Forest [31] is another example of Ensemble Learning model based on bagging. Bagging operates by sampling from the training dataset uniformly and with repetitions in order to create m training datasets out of the original one. As weak learners, the Random Forest utilizes Decision Trees [32]. Decision Trees (DTs) are a non-parametric models composed of two elements: nodes and branches. Each node represent a question made about a data feature, and each branch the outcome of that decision. An instance is passed through the three generating a path from the input node to a final leaf which represent the target variable. A decision tree is trained by optimizing the paths in order to capture patterns in the data. In Random Forest, several subsets of data and features are created from the given dataset, so each DT has its own set of features allocated to it. The randomly split dataset is distributed among all the trees with each tree focusing on the data that it has been provided. In classification, votes are collected from each tree, and the most popular class is chosen as the final output, whereas in regression an average is taken over all the outputs and is considered as the final result. Unlike Decision Trees, where the best performing features are taken as the split nodes, in Random Forest, these features

are selected randomly. Only a selected bag of features are taken into consideration, and a randomized threshold is used to create the Decision tree. After training, predictions for unseen samples can be made by averaging the predictions from all the individual regression trees on the new data or by taking the majority of vote in the case of classification trees.

1.2.5 Phosphoros

Phosphoros [33] is a Bayesian template fitting tool. Bayesian inference derives the posterior probability as a consequence of a prior probability and a "likelihood function" derived from a statistical model for the observed data.

Chapter 2

Astrophysical Background and Data

In this chapter we lay down the basic astrophysical concepts which are needed in order to understand our analysis, we briefly review the data sources (surveys) from which the data was extracted, and outline the data preparation for the experiments performed in this thesis.

2.1 Photometric Redshifts

Electromagnetic radiation consists of waves of the electromagnetic field, propagating through space, carrying radiant energy. While a wave travels through the space, if the relative distance between observer and source of emissions changes, also a wavelength variation is observed and with it a variation of frequency and photon energy. In astronomy and cosmology the main causes of electromagnetic redshift are:

- relativistic Doppler effect
- gravitational redshift
- cosmological redshift

The latter occurs because, due to the expansion of the universe, the distances of galaxies from us is increasing. In particular the Hubble Law states that galaxies are receding away from Earth with a speed proportional to their distance. By observing the emissions of galaxies, the shifts of the emissions and absorption lines of known

elements can be directly connected to the velocities of their sources and thus, through the Hubble law, to their distances. If one of said lines can be detected on the spectrum, its shift from its theoretical counterpart consists in a multiplicative factor $(1 + z)$, with z being the redshift:

$$z = \frac{\lambda_{obs} - \lambda_{emit}}{\lambda_{emit}} \rightarrow 1 + z = \frac{\lambda_{obs}}{\lambda_{emit}}$$

Modern precision cosmology is however based on samples of many millions of galaxies and Spectroscopic surveys are very time expensive (in terms of telescope observing time and data reduction). In response to these needs an alternative is provided by Photometric redshifts (or photo-zs) which are based on photometry, and while they measure distance with higher uncertainty, photo-zs offer several advantages over their spectroscopic counterparts with the main one being the observational time effectiveness. While Spectrographs need long observational times and have small field of view, Photometric redshifts are derived from broadband imaging. Nowadays, photo-zs are used in many different research fields such as distance calibration measurements, the study of the cosmic time evolution of galaxy properties [34], the search for primordial galaxies [35], and the study of the relation between galaxy properties and their dark matter halos [36].

2.2 Magnitudes

The study of the light emitted by stars, galaxies and objects beyond the Solar System represents a large part of the accessible information about the universe, and quantitative measurements of the intensity and polarization of light in each part of the electromagnetic spectrum is a fundamental part for developing modern theories. Apparent magnitude is a unitless measure of the brightness of a star or other astronomical object observed from Earth. An object's apparent magnitude depends on its intrinsic luminosity and its distance from the observer. Given that the luminosity is the energy emitted by the source per second, we can define the flux F as the ratio of the luminosity and the unit area oriented perpendicular to the light. Taking the assumption that light emitted by the source, travels through space as a spherical shell of radius r, luminosity and flux are connected by the inverse square law for light:

$$F = \frac{L}{4\pi r^2} \tag{2.1}$$

Magnitude is a unitless measurement for historic reasons and nowadays is formulated as the ratio of the flux of a source with respect to a reference source. Stars, infact,

were extensively observed before the advent of the scientific method and the first magnitude scale was developed by Hipparchus. His scale has been extended by far in both directions thanks to the advent of generations of telescopes with increasing resolving power and today it goes from $m = -26.83$ for the Sun to approximately $m = 30$ for the faintest object detectable (so that the lower is m, the brightest is the source). A difference of 1 magnitude corresponds to a brightness ratio of $100^{1/5} \simeq 2.512$. Thus a first-magnitude star appears 2.512 times brighter than a second-magnitude star, $2.512^2 = 6.310$ times brighter than a third-magnitude star, and 100 times brighter than a sixth-magnitude star. The total range of magnitudes corresponds to over $100^{57/5} = (10^2)^{11.4} \simeq 10^{23}$ for the ratio of the apparent brightness of the Sun to that of the faintest star or galaxy yet observed. Given two sources with apparent magnitudes m_1 and m_2, their difference in magnitudes is empirically connected to the ratios of their fluxes through the equation:

$$100^{m_1 - m_2} = \frac{F_2}{F_1} \tag{2.2}$$

By applying the logarithm in base 10 to both sides, and rearranging we get:

$$m_1 - m_2 = -2.5 \log_{10} \left(\frac{F_1}{F_2} \right) \tag{2.3}$$

By using a reference source magnitude m_2, then we can define the magnitude as:

$$m = m_{ref} - 2.5 \log_{10} \left(\frac{F_1}{F_{ref}} \right) \tag{2.4}$$

The absolute magnitude M is defined as the apparent magnitude that a source would have if positioned at a distance of 10 pc from the observer and it is connected to the apparent magnitude through the following equation:

$$d = 10^{(m-M+5)/5} pc \tag{2.5}$$

So, in order to estimate the absolute magnitude and fairly compare different sources, is required to derive their distances.

2.3 Color Index

In order to characterize astronomical sources, it is useful to limit the observed light to specific regions. This is achieved by the utilization of photometric filters. The range

in frequency of a filter is known in astrophysics as photometric band. The difference between magnitudes related to two different bands is known as color index. Utilizing equation 2.3, we can define the color index between band U and B as:

$$C_{U-B} = m_U - m_B = -2.5 \log_{10} \left(\frac{\int_0^\infty F_U(\lambda) S_U(\lambda)\, d\lambda}{\int_0^\infty F_B(\lambda) S_B(\lambda)\, d\lambda} \right) \tag{2.6}$$

where m_U and m_B are the apparent magnitudes in the U and B bands, $F_U(\lambda)$ and $F_B(\lambda)$ are the fluxes emitted at the wavelength λ and $S_U(\lambda)$ and $S_B(\lambda)$ are the two filters response functions. By capturing the ratios between the fluxes in two bands, colors have been proven useful for astrophysicists in defining source properties given that the amount of light emitted in specific bands can be used as a tracer for source properties such as chemical composition and, in case of galaxies, of gas and stellar abundances.

2.4 Surveys

An astronomical survey is a collection of astronomical observations that share common features such as position in the sky or frequency of observation. The data product of a survey, a catalog, represents the fundamental data basis for astronomy. We may classify surveys in regard to their scientific motivation and strategy, their wavelength regime, ground-based vs. space-based, the type of observations (e.g. imaging, spectroscopy, etc.), their area coverage and depth, and their temporal character (one-time vs. multi-epoch). Hereafter, we describe the astronomical surveys from which the data, needed to carry out the proposed research, was extracted. We, then, outline the procedural steps to clean the extracted data and finally we describe how the data was divided into several sets in order to receive fair redshift prediction from all the models tested during the challenge.

2.4.1 The Kilo-Degree Survey

The Kilo-Degree Survey (KiDS, [6]), is an ongoing optical wide field imaging survey at the the European Southern Observatory (ESO) Very Large Telescope (VLT) survey telescope, the VST. KiDS was designed as a cosmological survey in order to study galaxies population up to $z \sim 1$, and among other scientific goals, measure weak gravitational lensing events. The KiDS images were processed with two independent pipelines to produce stacked images in the four bands, from which the photometric properties of sources were extracted in order to produce a catalogue. KiDS covers about 1350 deg^2 of extragalactic sky, within in the four bands: u, g, r and i.

2.4.2 The VISTA Kilo-degree Infrared Galaxy survey

The VISTA Kilo-degree Infrared Galaxy Public Survey (VIKING, [37]) is one of the six survey conducted at the Visible and Infrared Survey Telescope for Astronomy (VISTA), which is part of the ESO's Paranal observatory. It is a wide area (area of 1350 deg^2), intermediate-depth near-infrared imaging survey, in the five broadband filters Z, Y, J, H, Ks. The sky coverage has maximum overlap with KiDS in the optical bands.

2.4.3 The Galaxy And Mass Assembly survey

The Galaxy And Mass Assembly (GAMA, [38, 39]) survey is a spectroscopic redshift and multi-wavelength photometrict survey. GAMA was originally designed to survey regions of the space that match the KiDS and VIKING survey.

2.5 Data Collection

The original dataset is composed by objects gathered from the images provided by the fourth Data Release (DR) of the KiDS and VIKING surveys, and thereafter cross-matched with the GAMA DR3 survey in order to retrieve the spectroscopic redshift of the galaxies. To compare the PDF producing method performances in a fair manner, the dataset of $150,000$ galaxies was divided into three subsets. While the first contained both photometric and spectroscopic information in order for the models to be trained and tested, the second was a blind test set containing only the photometry. Each group was asked to produce photometric redshifts for the blind dataset and send the predictions back. In this thesis we will work only with the $50,000$ objects belonging to the blind test set. In the next section, we explain how the received predictions were selected in order to create a common dataset on which predictions for all sources from all the models can be trusted, i.e. the predictions are not flagged as untrustworthy from any of the models.

2.6 Data Pruning and Splitting

The predictions received from the participating models (AdaBoost, LePhare, METAPHOR, Random Forest, Phosphoros, introduced in Sec.1.2) contained $50,000$ sources and the following features:

- ID: an unique identifier describing each object;

- Quality Flag: given by each method team in order to indicate the quality of the produced PDFs (0: superior quality, 1: standard quality, 2: lower quality)
- Photometric Redshift: the point-like estimation of the redshift
- Bins: each one pointing the probability that the true redshift value falls into that particular bin. 350 bins of width (0.02), with $z \in [0, 7]$

All objects for which at least one method provided a value of 2 in the flag column, were excluded. The number of surviving objects went from $50,000$ to $45,971$.
LePhare method produced an ulterior flag named *Type FLAG* separating galaxies from stars: all objects marked as *STAR* in the LePhare *Type FLAG* were excluded bringing the number of surviving number of objects from $45,971$ to $44,464$.
Figure 3.3, which shows the scatter plots between the spectroscopic redshifts and the photometric ones produced by all methods, shows that all methods have poor performances on galaxies with spectroscopic redshift higher than 4. For that reason, all the objects outside the closed spectroscopic redshift interval $[0.01, 4]$ were excluded from the analysis. The number of surviving objects went from $44,464$ to $43,878$; The final spectroscopic redshifts distribution is shown in Figure 2.1 with the high redshift cut highlighted in purple.

In order to asses the model performances in estimating the photometric redshifts, the following metric was introduced:

$$|\Delta z_{norm}| = \frac{|z_{spec} - z_{phot}|}{1 + z_{spec}} \tag{2.7}$$

with z_{spec} being the spectroscopic redshift and z_{phot} being the photometric one. This is the normalized difference between the two redshifts.
The final cleaned dataset contained, for each galaxy, the following features:

- ID: unique identifier for the object;
- MAG_GAAP: The Gaussian aperture and Photometry (GAaP) magnitudes have been measured on Gaussian-weighted apertures, which are modified per-source and per-image, in the following bands: u, g, r, i, z, Y, J, H, K_s;
- Fluxes_GAAP: fluxes for the following bads: u, g, r, i, z, Y, J, H, K_s;
- *Errors*: photometric errors for all magnitudes and fluxes listed above;
- *Spectroscopic Redshifts*: the spectroscopic redshifts;

- *Photometric redshifts*: the photometric redshifts produced by the methods (0:AdaBoost, 1:Lephare, 2:Metaphor, 3:Phosphoros, 4:Random Forest).

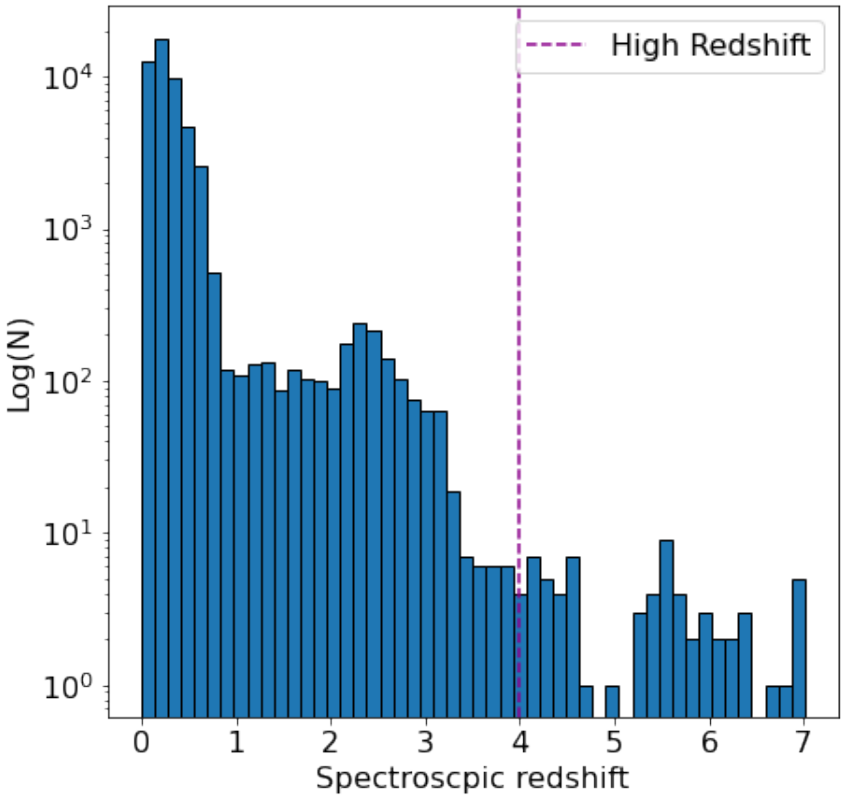


Figure 2.1: Spectroscopic redshift distribution of the sample, in purple there is the highlight for the cut performed to separate high redshift objects

Chapter 3

Experimental Procedure

This project aims to build a pipeline of MLPs trained to identify model outliers and substitute them with better predictions from other models. The overall process can be broken into two main steps: the first MLP identifies outliers from the rest of the sources, then feeds these outliers to a second MLP which identifies the methods that produced said outliers. This identification lets us substitute the outliers with better predictions made by other models. To achieve this goal and train the MLPs we divided the process into two experiments: in the first one, the Outlier Identifier MLP will be fed with the raw data discovering possible outliers related to any unspecified method, while, in the second, a MLP will be trained to predict the method to which an outlier belongs to. In the latter, we also explore the possibility to improve the performance of the MLP by aggregating all the less populated classes into macro-classes. This modification tries to solve the problem of class imbalance in the data. In the Test phase, the two MLPs are chained one after the other. The first one identifies the outliers which are then fed to the second MLP in order to recognize the method to which they belong and substitute them with better predictions from the remaining models.

3.1 Data Preparation and Discovery

In order to train the MLPs, guided by our previous knowledge about the physics of the problem and the nature of the photometrical features, we selected the following set of features among the available ones in the cleaned $43,878$ sources dataset (see Sec. 2.6):

- **Photometric Redshifts**: produced by the five methods: photo_0, photo_1, photo_2, photo_3, photo_4

- **Magnitudes**: MAG_GAAP_u, MAG_GAAP_g, MAG_GAAP_r, MAG_GAAP_i, MAG_GAAP_Z, MAG_GAAP_Y, MAG_GAAP_J, MAG_GAAP_H, and MAG_GAAP_Ks;
- **Colors**: col_U-G, col_G-R, col_R-I, col_I-Z, col_Z-Y, col_Y-J, col_J-H and col_H-Ks

for a total of 22 photometric features. The index near the photo-z feature is related to the method that produced it. In particular:
0:AdaBoost, 1:LePhare, 2:METAPHOR, 3:Phosphoros, 4:Random Forest.

The dataset was divided into a Train, Validation and Test set. The train set is used to train the model, the validation set to check for overfitting during training and the test set, as the name suggests, to assess the performance of the model on unseen instances. Data is split using the following criteria: 55% for the Train set, 20% for the Validation set, and 25% for the Test set.

Different sets of labels have been prepared, binary labels for the first experiment (outliers vs non-outliers) and multi-class labels for the second.

Binary labels:

The binary labels for the first experiments are created using Eq. 2.7. Given a source, if the $|\Delta z_{norm}|$ between any of the estimated photo-zs (by the five participating models) for that source and its spectroscopic redshift, is higher than 0.15, then it is labeled with a 1 (outlier) otherwise with 0 (non-outlier).

Multi-class labels:

In this case, each instance is labeled by an acronym made up of the capital letters related to all the methods for which $|\Delta z_{norm}| > 0.15$ (e.g. an outlier for all the methods is identified by $ALMPR$[1], or an outlier for A and R as AR). The set of unique labels is transformed into a set of integers for the purpose of training the MLP. The initial class distribution, shown in Figure 3.4, contains 31 unique classes. As it can be seen, there is a high variance in the size of the classes' supports, a phenomenon which is called *class imbalance*. Class imbalance has a deep impact on classificators' performances [40, 41, 42], and several techniques have been proposed to address this problem such as undersampling, oversampling and synthetic data generation with algorithms such as SMOTE [43] or AdaBoost [22]. Several of these techniques have been tried to improve the MLP performance on those low support

[1]A - AdaBoost; L - LePhare; M - METAPHOR; P - Phosphoros; R - Random Forest

classes, and, in the end, we settled for an aggregation strategy that groups under the same label all the objects related to a specific class of interest. Figure 3.4 displays the aggregation for the classes related to the AdaBoost method, with the total number of classes decreasing from 31 to 17. We focus on AdaBoost because from our preliminary analysis shown in Table 3.2, it is recognized as the best performing method among the participants, and thus, successfully predicting its outliers, would bring the highest boost in combined classification performance. To check the gain in performance due to class aggregation, the second experiment is repeated on the aggregated classes.

Principal Component Analysis

In order to visualize the data in lower dimensions, we performed a Principal Component Analysis (PCA). This technique is implemented to reduce the dimensionality of a dataset and to gain insights on the class separability. PCA calculates the covariance matrix of all the features and then generates the eigenvectors and eigenvalues from the matrix. Then, the covariance matrix is multiplied by the eigenvectors to create principal components. These principal components are the new features based on our original features and their importance in terms of explaining the variability in the dataset is given by its eigenvalues. The dimensionality of the problem can be thus reduced by ranking the principal components by their explained variance and selecting only the top n. By retaining a high explained variance, the new synthetic set of features should encompass the same information contained in the original features at the expense of interpretability of the results. In fact, the principal components have no physical meaning (to the contrary of the photometric features selected to train our MLPs). From our experiments, we can see that 98% of the variation within the dataset can be captured using only the first six principal components which seems to suggest that some of the chosen features are redundant. Two scatter plots using the first three components and with samples colored by their assigned labels (for both the binary and multi-class experiments) are shown respectively in Figure 3.1 and Figure 3.2. In the latter only the seven most populated classes (representing $\sim 80\%$ of the overall outliers population) are displayed. These figures show that, inside the parameter space of the principal components, both these sets of labels are separated which suggests that the MLPs should be able to find suitable hyper-planes to separate the classes.

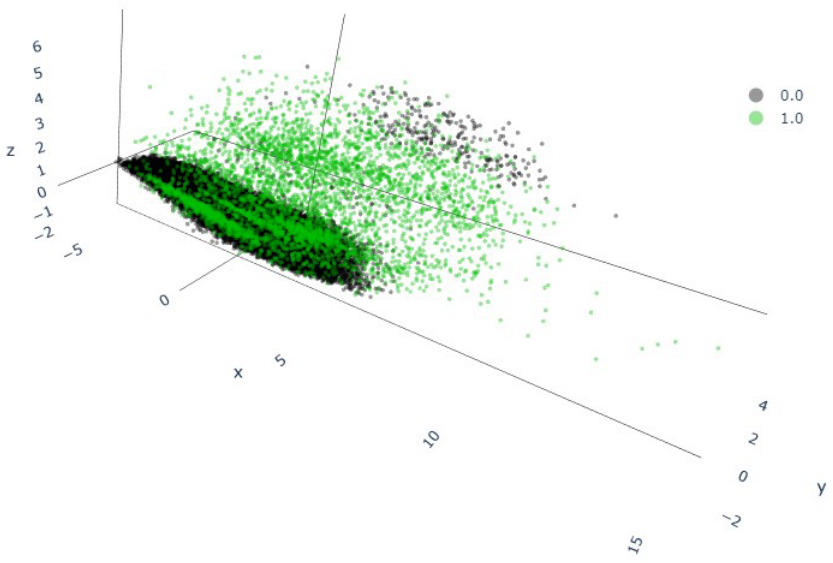


Figure 3.1: 3D scatterplot over the first three principal components for binary labeled data

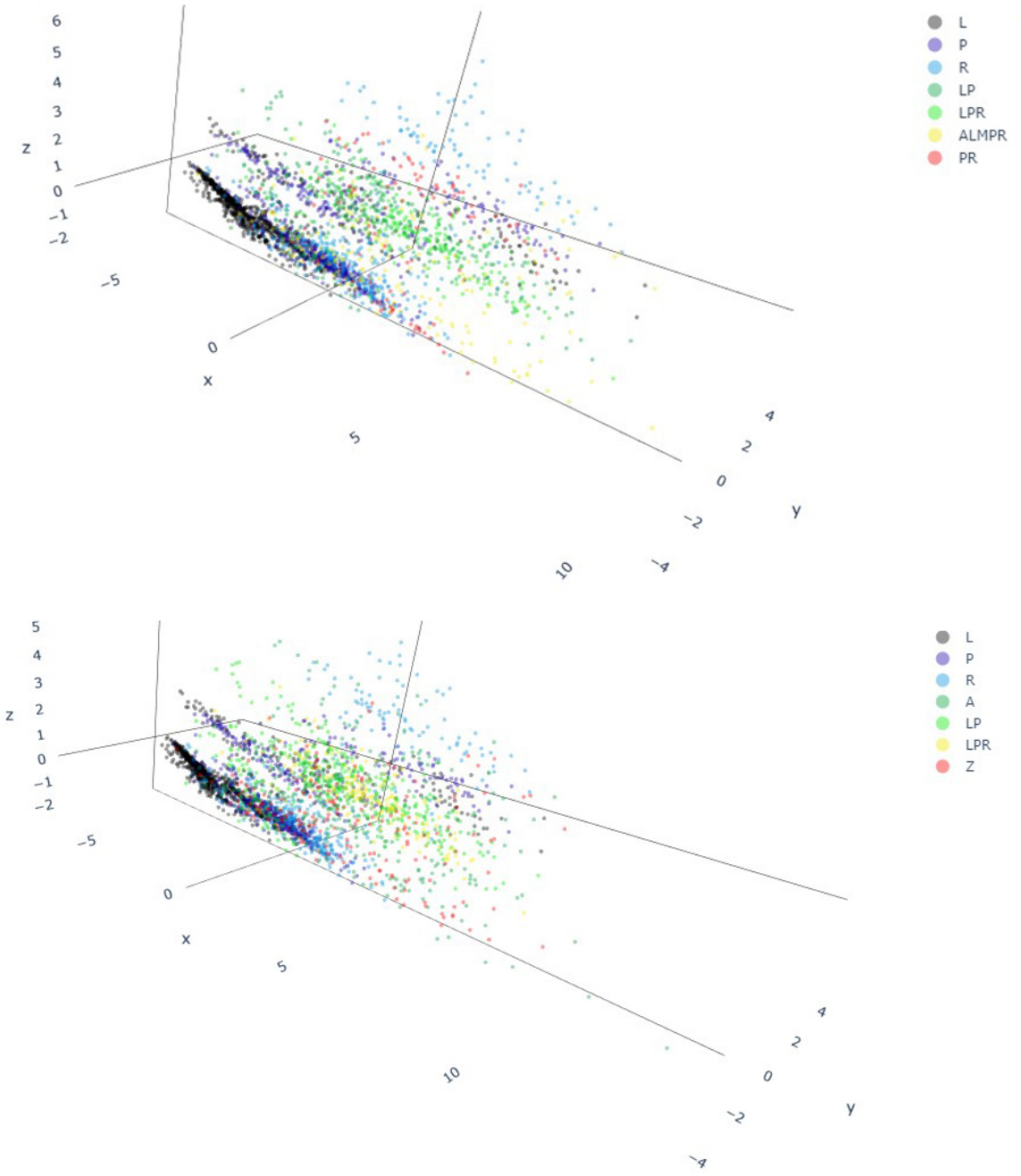


Figure 3.2: 3D scatterplot over the first three principal components for the seven most populated labels (before and after aggregation) for the multi-class set of labels

Statistical Estimators

Here we introduce all the statistical estimators used to assess models' performances in estimating photometric redshifts and to evaluate the MLPs' performances.

- $Mean(x) = \frac{\sum_i^n x_i}{n}$
- $Median(x) = x_{\frac{n}{2}}$
- $StandardDeviation(x) = \sigma(x) = \sqrt{\frac{\sum_i^n [x_i - \frac{\sum_i^n x_i}{n}]^2}{n}}$
- $SEM = \frac{\sigma}{\sqrt{n}}$
- $nMAD(x) = 1.48 \times Median(|x|)$
- $Outliers\% = \frac{n_{out}}{n_{tot}} \cdot 100, n_{out} = |\Delta z_{norm}| > 0.15$

where n is the total number of samples, SEM is the Standard Error of the Mean and the $nMAD$ is the normalized Median Absolute Deviation.

The confusion matrix is a $N \times N$ matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values (represented by the columns) with those predicted by the ML model (represented by the rows). If we consider a binary classification problem, with a positive class and a negative class, we can define the following possibilities for the prediction of a given instance:

True Positive (TP): the instance is correctly predicted as positive;

False Positive (FP): the correct class is negative, but the model predicted the instance as positive;

True Negative (TN): the instance is correctly predicted as negative;

False Negative (FN): the correct class is positive, but the model predicted the instance as negative.

These values are summarized in Table 3.1, which refers to the binary classification but can then be easily extended to the multi-class case. We can use these definitions to count the number of TPs, TNs, FPs and FNs predictions and produce the following classification metrics:

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Table 3.1: Example of a binary confusion matrix.

Accuracy: is the fraction of correct predictions over the total number of predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

Precision: is the ability of a classifier not to label as positive an instance that is actually negative.

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

Recall: express the classifier capability to find all the positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

F_1 **score:** is the weighted harmonic mean of Precision and Recall.

$$F_1 score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{3.4}$$

Preliminary Models' Performance Analysis

In this subsection, we want to measure the models' performances in estimating the photo-zs and understand the maximum gain in performance we would gain if we were able to eliminate all the outliers for each model. To estimate the models' performances, we computed the $|\Delta z_{norm}|$ (see Eq. 2.7) for all the instances of each model and produced descriptive statistics which are shown in Table 3.2. We also produced, for each model, the scatter plots of the predicted photo-zs against the true spectroscopic redshifts. The plots are shown in Figure 2.1. All the points falling outside the orange highlighted area are deemed as outliers. The plots show the majority of the spectroscopic redshift belongs to the range $z \in [0.01, 4]$ and that all models have poor performances on instances characterized by $z > 4$. For that reason, all sources with $z > 4$ are excluded from the analysis. From Table 3.2, we can see that the metrics for all the methods are comparable, with the models based

on ML performing slightly better than those based on template fitting. We also can see, by comparing, for each model, the metrics before and after outliers removal, that there is a potential ample gain in performance if outliers are correctly identified and removed.

Model	Outliers	Objects	Mean	σ	SEM	Median	nMAD	Outliers %
AdaBoost	included	4.39e+04	2.54e-02	7.73e-02	3.69e-04	1.35e-02	1.28e-02	1.60
	not included	4.32e+04	1.79e-02	1.75e-02	8.41e-05	1.32e-02	1.24e-02	0.00
LePhare	included	4.39e+04	6.64e-02	1.76e-01	8.39e-04	3.28e-02	3.12e-02	6.33
	not included	4.11e+04	3.81e-02	3.14e-02	1.55e-04	3.03e-02	2.78e-02	0.00
METAPHOR	included	4.39e+04	2.83e-02	7.41e-02	3.54e-04	1.58e-02	1.48e-02	1.86
	not included	4.31e+04	2.11e-02	2.07e-02	9.96e-05	1.55e-02	1.43e-02	0.00
Phosphoros	included	4.39e+04	5.96e-02	1.98e-01	9.44e-04	2.38e-02	2.30e-02	4.98
	not included	4.17e+04	3.08e-02	2.87e-02	1.40e-04	2.23e-02	2.11e-02	0.00
Random Forest	included	4.39e+04	3.31e-02	8.47e-02	4.05e-04	1.40e-02	1.35e-02	4.03
	not included	4.21e+04	1.97e-02	2.19e-02	1.07e-04	1.32e-02	1.25e-02	0.00

Table 3.2: Models' statistical estimators comparison for the Test set. Each row shows, for a given model, the statistical estimators (columns) of the distribution of $|\Delta z_{norm}|$ for the prediction of that model.

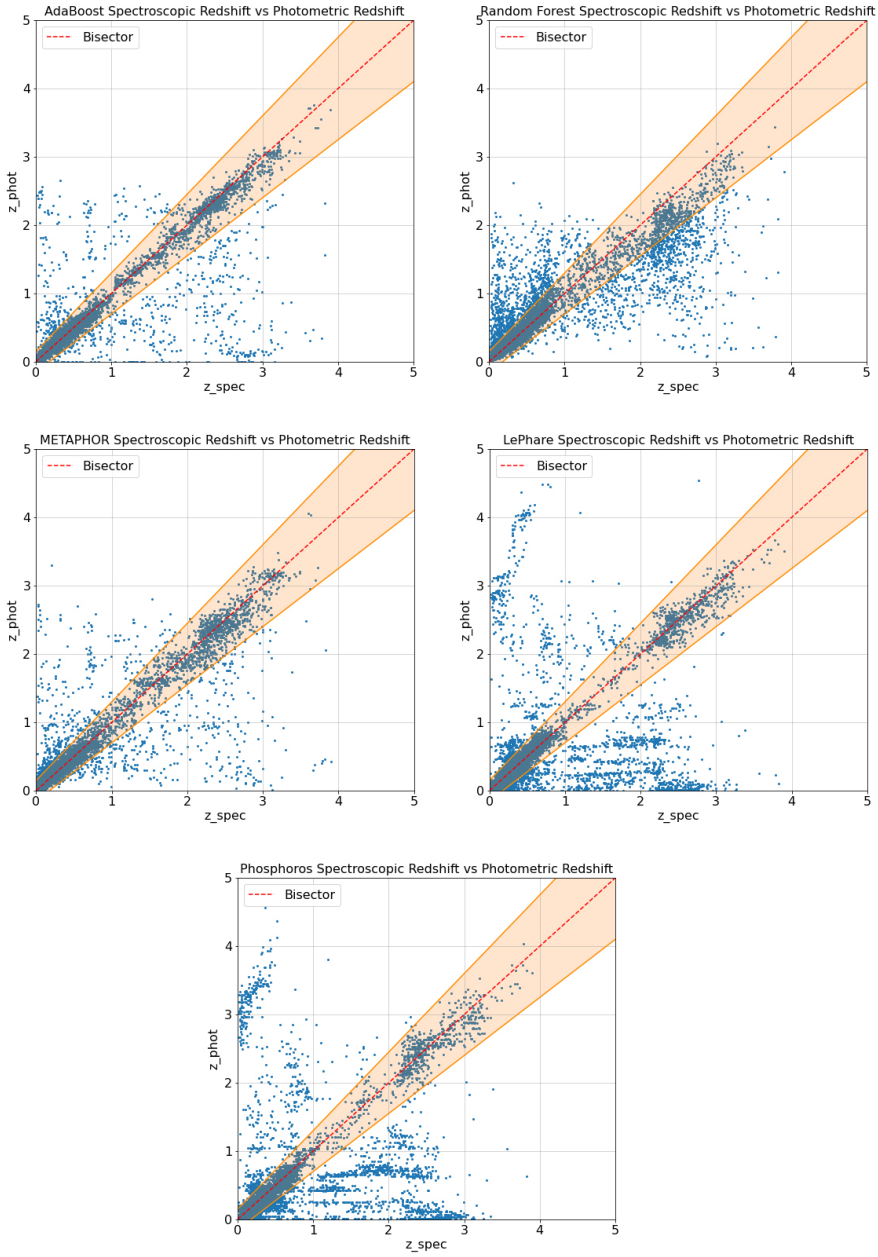


Figure 3.3: Spectroscopic redshift versus Photometric redshift scatter plots for the photo-z estimation related to each model. On the X axis are shown the spectroscopic redshifts, on the Y the photometric ones. The red dotted line is the bisector of the quadrant, points lying on the line are perfect predictions. Points outside the orange area, are deemed as outliers by Eq. 2.7.

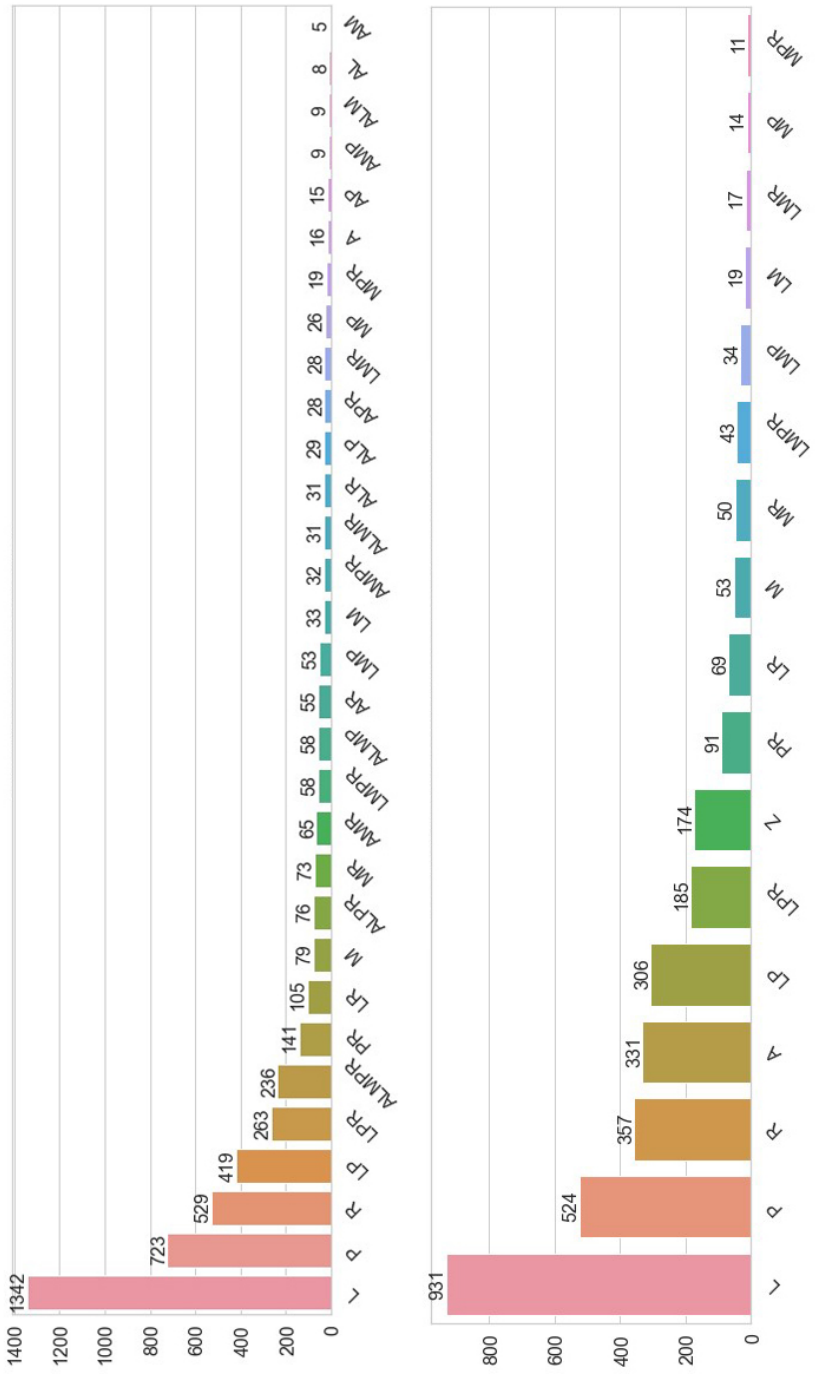


Figure 3.4: Outliers labels distribution (pre-post aggregation).

3.2 Experiments

The MLPs for the two Experiments share the same structure, which consists of five layers and differs for the number of units in the output layer. The input layer consisting of N_f neurons where N_f is the number of input features ($N_f = 22$), the first hidden layer consisting of $2N_f + 1$ neurons, the second hidden layer consisting of $N_f - 1$ neurons and the third hidden layer consisting of $N_f/2$ neurons. For the binary classification problem, the output layer consists of a single neuron (scalar), while for the multi-class classification problem, the number of neurons equals the number of classes in the problem, i.e. 31 for the original outliers' class distribution, 17 after the class aggregation. A dropout layer is inserted within any couple of hidden layers.

Experiment 1: Outlier Identifier (OI)

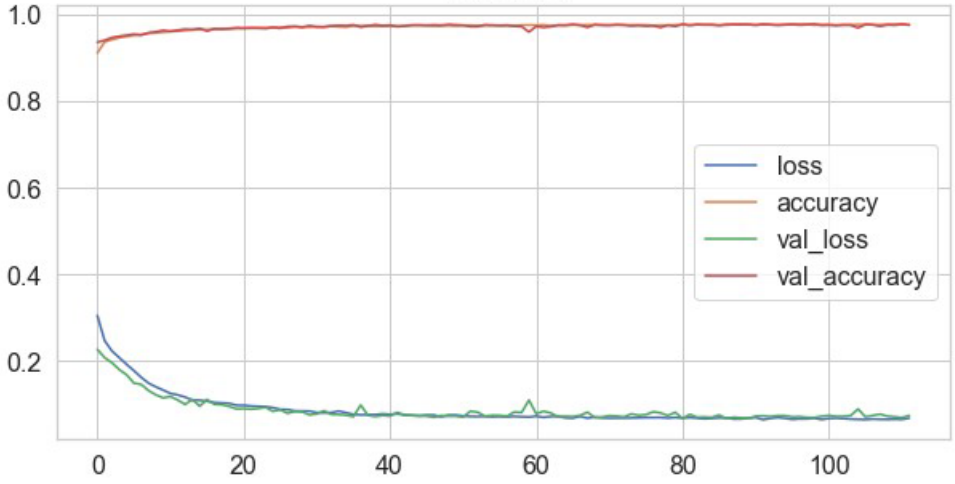


Figure 3.5: Training metrics for the OI MLP over the validation Set. In this plot the X axis shows the number of epochs, and the Y axis is the loss/accuracy value

The binary classifier training results are reported in Figure 4.5, which display that loss values, over the training and validation sets, decrease rapidly in the first epochs to a point of stability with a minimal gap between the two final values. These curves can also be used to address two different types of problems that arise during the training process: overfitting and unrepresentativeness of the validation set, which means validation data are not very representative of the training data due to their insufficient numbers. Overfitting can be spotted by the loss curves reaching a

	precision	recall	f1-score	support
0.0	0.98	0.99	0.99	8270
1.0	0.89	0.86	0.88	945
accuracy	0.98			9215

Table 3.3: Classification metrics for the OI MLP over the Validation Set; the table displays the precision, recall, f1 score and support size for the two classes (see Section 3.1) and the total classification accuracy.

minimum value and then raising again, which doesn't occur at any time during the training and validation process, while the good representativeness of the validation set is identified by both the curves being almost similar, without any offset between them and a stable val_loss curve. To assess the classification performance of the MLP over the validation/Test Set, we use to the Statistical estimators introduced in Sec. 3.1 and reported in Table 3.3 and Table 3.4. These tables show almost identical estimators' values, thus the following considerations can be extended to both sets: a high precision score related to the 0 class describes the classifier's capability to rarely label a good photo-z estimation as a potential outlier, while the recall score related to the 1 class highlight that it underperforms in terms of capability to detect all the positive instances (actual outlier for any method).

	precision	recall	f1-score	support
0.0	0.98	0.99	0.99	11779
1.0	0.95	0.84	0.89	1385
accuracy	0.98			13164

Table 3.4: Classification metrics for the OI MLP over the Test Set; the table displays the precision, recall, f1 score and support size for the two classes (see Section 3.1) and the total classification accuracy.

Experiment 2: Outlier Class Identifier (OCI)

The OCI MLP is trained on the instances labeled as outliers in the OI MLP training set. This is done to follow the project workflow in which the outliers found by the first MLP are then classified by the second. As previously discussed, the MLP is trained first with all 31 labels, and then with the aggregated 17 in order to mitigate class imbalance.

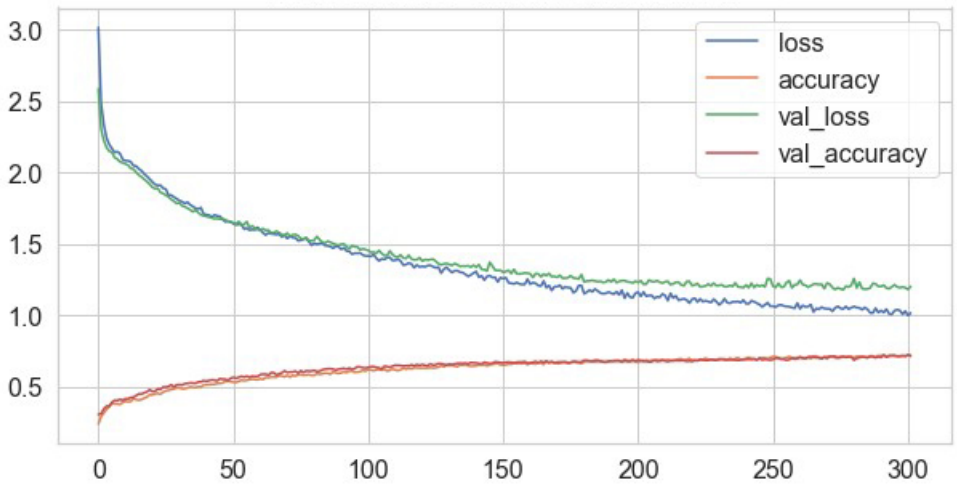


Figure 3.6: Training metrics for the OCI MLP over the validation Set. In this plot the X axis shows the number of epochs, and the Y axis is the lossand accuracy value

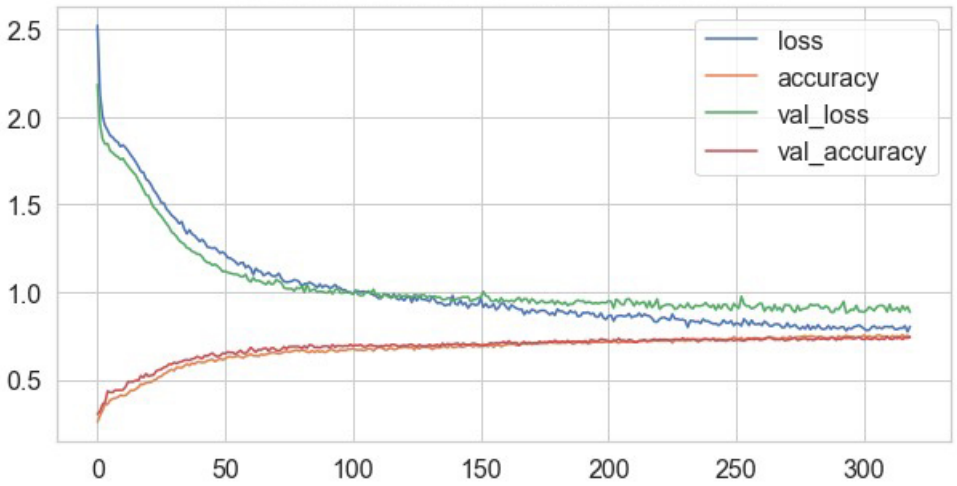


Figure 3.7: Training metrics for the OCI MLP over the validation Set after label aggregation. In this plot the X axis shows the number of epochs, and the Y axis is the loss and accuracy value.

Figure 3.6 and Figure 3.7 show the trends with number of training epochs, of the training loss, validation loss, training accuracy and validation accuracy using

respectively 31 and 17 labels (labels aggregation). The curves described in these Figures, as per the training results of the binary classifier, do not show overfitting and that both validation sets are representatives of the related training sets. While the accuracy does not improve remarkably after class aggregation, the technique brought further minimization of the loss values which is reflected in the betterment of the classification metrics outlined in Table 3.6 and Table 3.7. Still, the MLPs' performances on the multi-class classification do not reach the same levels obtained on the binary classification. Table 3.8 and Table 3.9 report the classification performance metrics for the Test set, and display comparable values with the ones previously reported in Table 3.6 and Table 3.7. In particular, if one compares the precision, recall and f-score on the A outliers which in the left table are split among several classes, and on the right are combined in a single class, one can see a substantial increase in all metrics. That being said, the final performance shows that the MLP is not able to reliably identify outliers for A and M, while it shows good performances on L, P and R outliers. After outliers are classified, to improve the total classification performance, we need to substitute outliers with the predictions made by the best performing model for which the instances are not outliers. For that reason, we introduced a hierarchical replacement approach based on models performances on the photometric redshifts estimation (see Table. 3.2)

$$A \rightarrow M \rightarrow R \rightarrow P \rightarrow L$$

An instance which is identified as outlier for some models is replaced with the photo-z prediction of the first method (in the ranking) for which it is not an outlier. e.g. an instance which is an outlier for A is replaced by the estimation made by M if the instance is not also classified as an outlier for M, otherwise by R and so on.

Table 3.5 shows the photometric redshift estimation metrics for the AdaBoost model. The first row reports the performances on the original dataset, the second reports the best possible performances obtainable if perfect outlier classification was obtained, the third shows the performances after the instances classified as outliers by OI MLP are removed, the third shows the performances after the instances classified as AdaBoost outliers by OCI MLP are removed and, the fourth shows the performances after the instances classified as AdaBoost outliers are substituted, through the hierarchical substitution (hereafter HS), with the photo-z estimation of the best model for which that instance is a non-outlier. The performances in all but the first row are reported as percentile variation with respect to the original values. As it can be seen, the best increase is obtained through the OI outlier removal which removes 88% of outliers (174 out of 198) but also 1050 correctly predicted instances. The further improvement of metrics over the theoretical best removal (second row) could

	objects	mean	σ	SEM	Median	nMAD	Outliers %
original	13164	2.52e-02	7.67e-02	6.69e-04	1.35e-02	1.28e-02	1.50
No Outliers	-1.50%	-28.49%	-76.77%	-76.53%	-1.83%	-2.56%	-100.00
OI	-9.30%	-30.30%	-71.68%	-70.25%	-3.97%	-5.22%	-88.27
OCI	-1.49%	-17.84%	-35.85%	-35.43%	-1.23%	-1.68%	-52.70
HS	-0.93%	-13.17%	-23.39%	-23.02%	-0.79%	-1.05%	-40.70

Table 3.5: Comparison of AdaBoost $|\Delta z_{norm}|$ statistical estimators over the test set: the first row is related to the unaltered original performance (see Table 3.2), while all the other estimators are expressed in terms of percentage variation with respect to the original values. The second is related to the theoretical best, the third is related to OI outlier removal, the forth to the OCI AdaBoost outlier removal and the last to the HS substitution.

be explained by the fact that the outlier selection role ($|\Delta z_{norm}| > 0.15$) is based on a hard threshold, and many instances could be close to that threshold but still not identified as outliers. However, as can be seen in Figure 3.8, while the number of sources that have a $|\Delta z_{norm}| \sim 0.15$ is non-negligible, this observation alone seems not sufficient to justify the improvements. Regarding the OCI and HS performances, they still can remove around half the outliers but with a much lower cost on the number of correctly removed instances. The hierarchical substitution still removes 0.93% of instances because they are classified as outliers for all models.

The removal of the outliers, for the test set, at each step of the pipeline is summarized in Figure 3.9 and Table 3.10. The latter reports the total number of true outliers (first column, "P"), the number of instances predicted as outliers (second column, "TP+FP") and the number and fraction of said predictions which are TPs (third column). As it can be seen, OI MLP is capable of identifying 95% of the test set outliers. The second row shows the total number of outliers for AdaBoost in the test set after being processed by the OI MLP and thus passed as input to the OCI MLP, the number of outliers classified as A by OCI MLP, the number of correctly predicted AdaBoost outliers and their fraction over the number of actual AdaBoost outliers. From the low number of TPs and the similar number of FPs, we confirm that the OCI MLP is not able to reliably identify AdaBoost outliers. The third row shows the results of the hierarchical substitution, starting from 74 of the 196 predicted outliers that are not recognized as outliers for at least one of the remaining model. Of these 74 roughly only 50% is correctly substituted with non-outlier. This behavior is to be connected to the poor performance of the OCI MLP classifier.

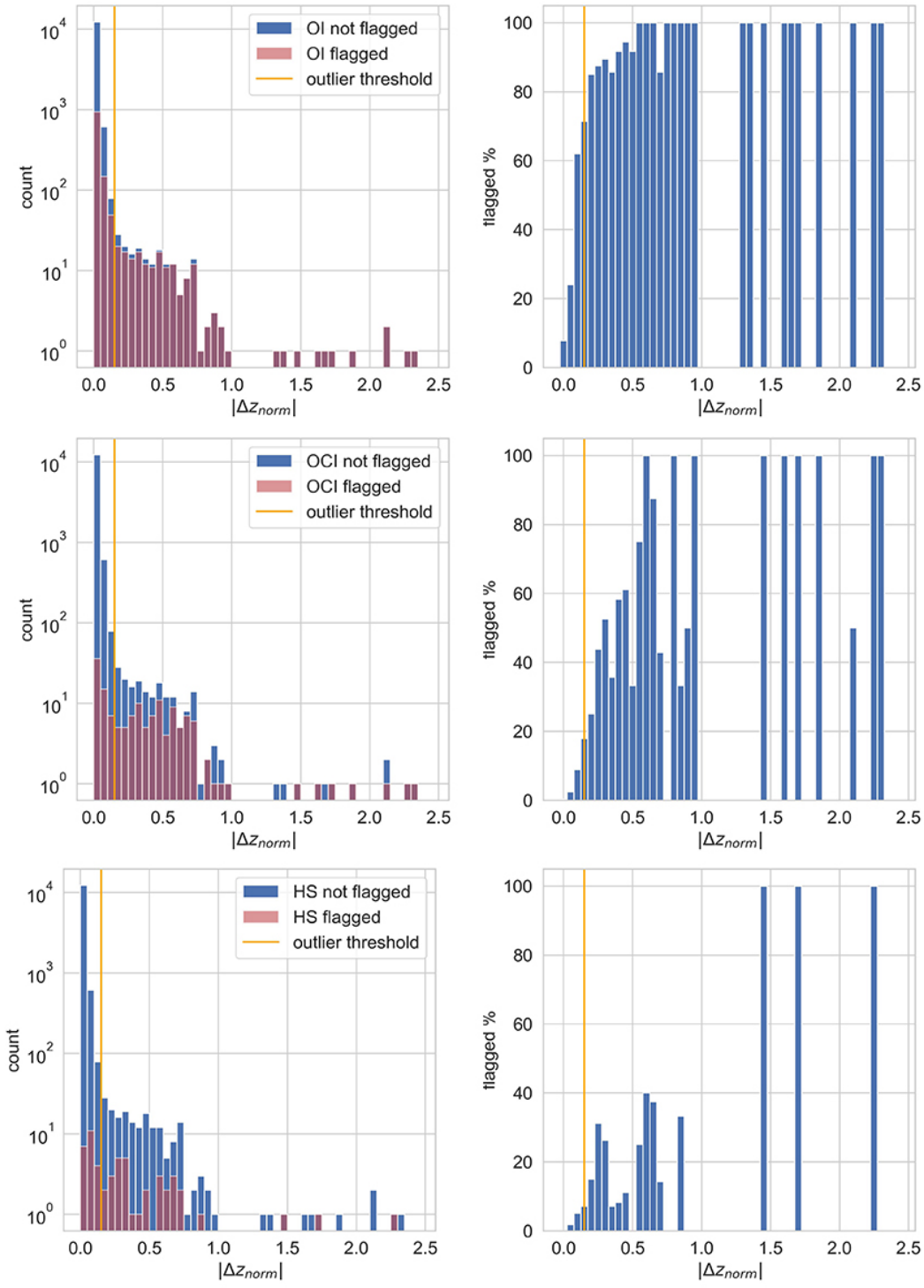


Figure 3.8: AdaBoost removed outliers $|\Delta z_{norm}|$ values

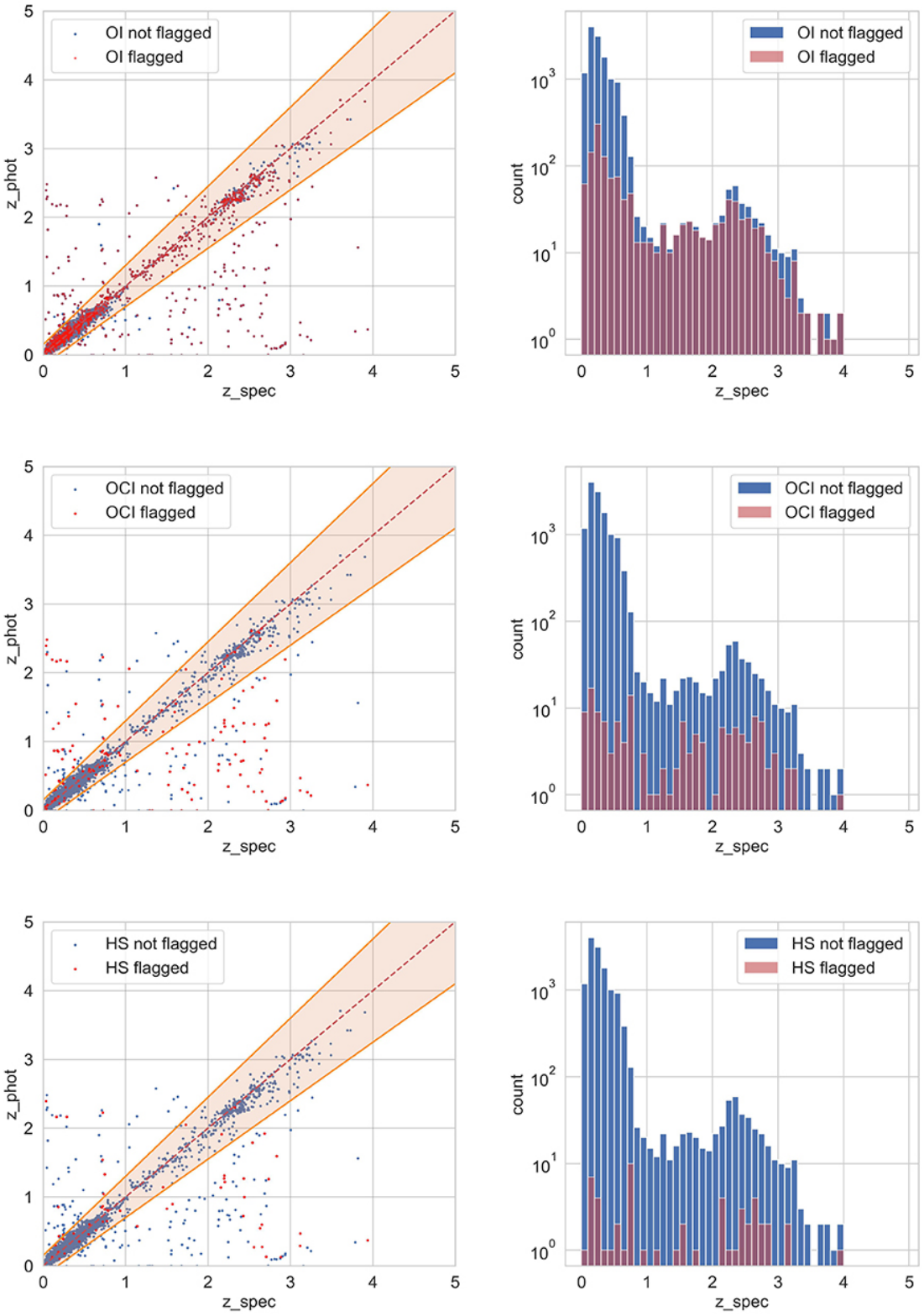


Figure 3.9: AdaBoost identified outliers for each stage of the workflow. In these plots, the instances flagged as outlier are reported in red, while sources deemed as non-outliers are colored in blue.

	precision	recall	f1-score	support
A	0.00	0.00	0.00	6.00
AL	0.00	0.00	0.00	1.00
ALM	0.00	0.00	0.00	1.00
ALMP	0.00	0.00	0.00	14.00
ALMPR	0.20	0.22	0.21	50.00
ALMR	0.00	0.00	0.00	10.00
ALP	0.00	0.00	0.00	8.00
ALPR	0.50	0.05	0.09	21.00
ALR	0.31	0.67	0.42	6.00
AM	0.00	0.00	0.00	1.00
AMP	0.00	0.00	0.00	2.00
AMPR	0.00	0.00	0.00	4.00
AMR	0.00	0.00	0.00	13.00
AP	0.00	0.00	0.00	5.00
APR	0.25	0.17	0.20	6.00
AR	0.50	0.46	0.48	13.00
L	0.88	0.97	0.92	292.00
LM	0.00	0.00	0.00	6.00
LMP	0.00	0.00	0.00	10.00
LMPR	0.00	0.00	0.00	14.00
LMR	0.00	0.00	0.00	4.00
LP	0.58	0.80	0.68	86.00
LPR	0.61	0.80	0.69	59.00
LR	0.50	0.48	0.49	23.00
M	0.00	0.00	0.00	18.00
MP	0.00	0.00	0.00	3.00
MPR	0.00	0.00	0.00	3.00
MR	0.75	0.21	0.33	14.00
P	0.84	0.99	0.91	143.00
PR	0.64	0.50	0.56	28.00
R	0.68	0.96	0.79	99.00
accuracy	0.71			963.00

Table 3.6: Classification metrics for the OCI MLP over the validation set

	precision	recall	f1-score	support
A	0.57	0.69	0.62	111.00
L	0.89	0.96	0.92	292.00
LM	0.00	0.00	0.00	6.00
LMP	0.00	0.00	0.00	10.00
LMPR	0.00	0.00	0.00	14.00
LMR	0.00	0.00	0.00	4.00
LP	0.64	0.76	0.69	86.00
LPR	0.68	0.58	0.62	59.00
LR	0.38	0.39	0.38	23.00
M	0.33	0.06	0.10	18.00
MP	0.00	0.00	0.00	3.00
MPR	0.00	0.00	0.00	3.00
MR	1.00	0.07	0.13	14.00
P	0.89	0.94	0.91	143.00
PR	0.59	0.61	0.60	28.00
R	0.70	0.97	0.81	99.00
Z	0.39	0.14	0.21	50.00
accuracy	0.75			963.00

Table 3.7: Classification metrics for the OCI MLP over the validation set (post A label aggregation)

	precision	recall	f1-score	support
A	0.00	0.00	0.00	3.00
AL	0.00	0.00	0.00	3.00
ALM	0.00	0.00	0.00	0.00
ALMP	0.00	0.00	0.00	18.00
ALMPR	0.23	0.23	0.23	52.00
ALMR	0.00	0.00	0.00	7.00
ALP	0.00	0.00	0.00	7.00
ALPR	0.00	0.00	0.00	22.00
ALR	0.20	0.17	0.18	6.00
AM	0.00	0.00	0.00	1.00
AMP	0.00	0.00	0.00	3.00
AMPR	0.00	0.00	0.00	18.00
AMR	0.00	0.00	0.00	16.00
AP	0.00	0.00	0.00	1.00
APR	0.25	0.33	0.29	3.00
AR	0.00	0.00	0.00	17.00
L	0.79	0.97	0.87	327.00
LM	0.00	0.00	0.00	12.00
LMP	0.00	0.00	0.00	19.00
LMPR	0.00	0.00	0.00	15.00
LMR	0.00	0.00	0.00	10.00
LP	0.49	0.78	0.60	109.00
LPR	0.47	0.36	0.41	78.00
LR	0.36	0.56	0.43	36.00
M	0.00	0.00	0.00	7.00
MP	0.00	0.00	0.00	12.00
MPR	0.00	0.00	0.00	8.00
MR	0.00	0.00	0.00	23.00
P	0.67	0.96	0.79	137.00
PR	0.82	0.36	0.50	50.00
R	0.57	0.99	0.72	144.00
accuracy	0.62			1164.00

Table 3.8: Classification metrics for the OCI MLP over the test set

	precision	recall	f1-score	support
A	0.41	0.65	0.50	125.00
L	0.84	0.96	0.89	327.00
LM	0.00	0.00	0.00	12.00
LMP	0.00	0.00	0.00	19.00
LMPR	0.00	0.00	0.00	15.00
LMR	0.00	0.00	0.00	10.00
LP	0.51	0.53	0.52	109.00
LPR	0.40	0.49	0.44	78.00
LR	0.64	0.50	0.56	36.00
M	0.00	0.00	0.00	7.00
MP	0.00	0.00	0.00	12.00
MPR	0.00	0.00	0.00	8.00
MR	0.00	0.00	0.00	23.00
P	0.66	0.95	0.78	137.00
PR	0.56	0.48	0.52	50.00
R	0.74	0.92	0.82	144.00
Z	0.00	0.00	0.00	52.00
accuracy	0.65			1164.00

Table 3.9: Classification metrics for the OCI MLP over the test set (post A label aggregation)

	P	TP+FP	TP	Precision %
OI	1385	1224	1164	95.0%
OCI	125	196	81	41.3%
HS	125	74	39	52.8%

Table 3.10: Detection of AdaBoost Outlier at various stages in the workflow.

Chapter 4

Conclusion

In this work we built a pipeline of MLPs to identify outliers found among the photometric redshift estimations produced by five different models over a set of common sources. These models' estimations are based on the sole photometry and to assess their performances we introduced a quality measure that takes into account the normalized difference between the spectroscopic and photometric redshift prediction of each source, also used to formalize the outlier definition. We noted that, for all these models, a certain percentage of the photometric redshift point estimations could be defined as outliers and designed two different classification tasks approached by two separate MLPs. The first identifies sources for which at least one of the models performed a poor estimation, and the second identifies the specific method/s for which the photometric estimation is an outlier. To mitigate the class imbalance in the second multi-class classification, we implemented an aggregation strategy for the labels, and evaluated the performances on both classification problems (original 31 classes and the aggregated 17). Given the better performances of the AdaBoost model over the other models, we focus our analysis on the AdaBoost outliers. The aggregation of all the classes containing AdaBoost outliers improves the detection of the aforesaid class but still doesn't affect the MLP capability to predict other classes, even if their total number is almost halved. The Z class (namely $ALMPR$ before the labels aggregation, containing outliers for all methods is never recognized properly regardless of the support size. By comparing performances on the photometric estimation of the AdaBoost model after outliers removal with different strategies (see Table 3.5), it can be seen that the best improvement is obtained by the removal of all outliers by the OI MLP along $\sim 8\%$ of non-outliers, but still significant improvements can be obtained through the OCI MLP removal and the hierarchical substitution at a much lower price in terms of number of non-outliers removed. We think it would be

worth to inspect the removed instances in order to understand if they could be also deemed as outliers if a relaxation or a modification of the outliers assignment criterion ($|\Delta z_{norm}| > 0.15$). Furthermore, the classification metrics shown in Table 3.9 show that the model is correctly classifying pure outliers for L, P and R, a strategy could be implemented for removing these outliers first and then try and classify the remaining instances. This work was deemed outside the scope of this thesis, which has to be concluded at some point, but we intend to carry out these trials and after further improvements try to tackle, in the future, the same approach on the PDFs.

Bibliography

[1] S. Lilly, O. Le Fèvre, A. Renzini, G. Zamorani, M. Scodeggio, T. Contini, C. M. Carollo, G. Hasinger, J.-P. Kneib, A. Iovino, *et al.*, "zcosmos: a large vlt/vimos redshift survey covering 0¡ z¡ 3 in the cosmos field," *The Astrophysical Journal Supplement Series*, vol. 172, no. 1, p. 70, 2007.

[2] B. Garilli, R. McLure, L. Pentericci, P. Franzetti, A. Gargiulo, A. Carnall, O. Cucciati, A. Iovino, R. Amorin, M. Bolzonella, *et al.*, "The vandels eso public spectroscopic survey-final data release of 2087 spectra and spectroscopic measurements," *Astronomy & Astrophysics*, vol. 647, p. A150, 2021.

[3] W. A. Baum, "Photoelectric magnitudes and red-shifts," in *Problems of Extra-Galactic Research*, vol. 15, p. 390, 1962.

[4] M. Salvato, O. Ilbert, and B. Hoyle, "The many flavours of photometric redshifts," *Nature Astronomy*, vol. 3, no. 3, pp. 212–222, 2019.

[5] J. T. De Jong, G. A. V. Kleijn, D. R. Boxhoorn, H. Buddelmeijer, M. Capaccioli, F. Getman, A. Grado, E. Helmich, Z. Huang, N. Irisarri, *et al.*, "The first and second data releases of the kilo-degree survey," *Astronomy & Astrophysics*, vol. 582, p. A62, 2015.

[6] K. Kuijken, C. Heymans, A. Dvornik, H. Hildebrandt, J. de Jong, A. Wright, T. Erben, M. Bilicki, B. Giblin, H.-Y. Shan, *et al.*, "The fourth data release of the kilo-degree survey: ugri imaging and nine-band optical-ir photometry over 1000 square degrees," *Astronomy & Astrophysics*, vol. 625, p. A2, 2019.

[7] D. E. S. Collaboration *et al.*, "The dark energy survey," *arXiv preprint astro-ph/0510346*, 2005.

[8] H. Hildebrandt, S. Arnouts, P. Capak, L. Moustakas, C. Wolf, F. B. Abdalla, R. Assef, M. Banerji, N. Benítez, G. Brammer, *et al.*, "Phat: Photo-z accuracy testing," *Astronomy & Astrophysics*, vol. 523, p. A31, 2010.

[9] S. Schmidt, A. Malz, J. Soo, I. Almosallam, M. Brescia, S. Cavuoti, J. Cohen-Tanugi, A. Connolly, J. DeRose, P. Freeman, *et al.*, "Evaluation of probabilistic photometric redshift estimation approaches for the rubin observatory legacy survey of space and time (lsst)," *Monthly Notices of the Royal Astronomical Society*, vol. 499, no. 2, pp. 1587–1606, 2020.

[10] G. Desprez, S. Paltani, J. Coupon, I. Almosallam, A. Alvarez-Ayllon, V. Amaro, M. Brescia, M. Brodwin, S. Cavuoti, J. de Vicente-Albendea, *et al.*, "Euclid preparation-x. the euclid photometric-redshift challenge," *Astronomy & Astrophysics*, vol. 644, p. A31, 2020.

[11] R. Laureijs, J. Amiaux, S. Arduini, J.-L. Augueres, J. Brinchmann, R. Cole, M. Cropper, C. Dabin, L. Duvet, A. Ealet, *et al.*, "Euclid definition study report," *arXiv preprint arXiv:1110.3193*, 2011.

[12] N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, *et al.*, "The cosmic evolution survey (cosmos): overview," *The Astrophysical Journal Supplement Series*, vol. 172, no. 1, p. 1, 2007.

[13] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[14] F. Rosenblatt, "Perceptron simulation experiments," *Proceedings of the IRE*, vol. 48, no. 3, pp. 301–309, 1960.

[15] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, 2016.

[16] M. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv preprint arXiv:1001.2267*, 2010.

[17] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *ieee Computational intelligenCe magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[18] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[19] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[23] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[24] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[25] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.

[26] S. Arnouts, L. Moscardini, E. Vanzella, S. Colombi, S. Cristiani, A. Fontana, E. Giallongo, S. Matarrese, and P. Saracco, "Measuring the redshift evolution of clustering: the hubble deep field south," *Monthly Notices of the Royal Astronomical Society*, vol. 329, no. 2, pp. 355–366, 2002.

[27] S. Arnouts, S. Cristiani, L. Moscardini, S. Matarrese, F. Lucchin, A. Fontana, and E. Giallongo, "Measuring and modelling the redshift evolution of clustering: the hubble deep field north," *Monthly Notices of the Royal Astronomical Society*, vol. 310, no. 2, pp. 540–556, 1999.

[28] S. Cavuoti, V. Amaro, M. Brescia, C. Vellucci, C. Tortora, and G. Longo, "Metaphor: a machine-learning-based method for the probability density estimation of photometric redshifts," *Monthly Notices of the Royal Astronomical Society*, vol. 465, no. 2, pp. 1959–1973, 2017.

[29] M. Brescia, S. Cavuoti, R. D'Abrusco, G. Longo, and A. Mercurio, "Photometric redshifts for quasars in multi-band surveys," *The Astrophysical Journal*, vol. 772, no. 2, p. 140, 2013.

[30] S. Cavuoti, M. Brescia, V. De Stefano, and G. Longo, "Photometric redshift estimation based on data mining with photoraptor," *Experimental Astronomy*, vol. 39, no. 1, pp. 45–71, 2015.

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[32] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[33] N. Apostolakos, H. Degaudenzi, F. Dubath, P. Dubath, N. Morisset, S. Paltani, and M. Schefer, "Designing Modular Software for Template Fitting Photo-z Estimation," in *Astronomical Data Analysis Software and Systems XXVI* (M. Molinaro, K. Shortridge, and F. Pasian, eds.), vol. 521 of *Astronomical Society of the Pacific Conference Series*, p. 169, Oct. 2019.

[34] A. Fontana, S. D'Odorico, F. Poli, E. Giallongo, S. Arnouts, S. Cristiani, A. Moorwood, and P. Saracco, "photometric redshifts and selection of high-redshift galaxies in the ntt and hubble deep fields," *The Astronomical Journal*, vol. 120, no. 5, p. 2206, 2000.

[35] J. Dunlop, R. McLure, B. Robertson, R. Ellis, D. Stark, M. Cirasuolo, and L. De Ravel, "A critical analysis of the ultraviolet continuum slopes (β) of high-redshift galaxies: no evidence (yet) for extreme stellar populations at z¿ 6," *Monthly Notices of the Royal Astronomical Society*, vol. 420, no. 1, pp. 901–912, 2012.

[36] J. Coupon, S. Arnouts, L. van Waerbeke, T. Moutard, O. Ilbert, E. van Uitert, T. Erben, B. Garilli, L. Guzzo, C. Heymans, *et al.*, "The galaxy–halo connection from a joint lensing, clustering and abundance analysis in the cfhtlens/vipers field," *Monthly Notices of the Royal Astronomical Society*, vol. 449, no. 2, pp. 1352–1379, 2015.

[37] A. Edge, W. Sutherland, K. Kuijken, S. Driver, R. McMahon, S. Eales, and J. P. Emerson, "The vista kilo-degree infrared galaxy (viking) survey: bridging the gap between low and high redshift," *The Messenger*, vol. 154, pp. 32–34, 2013.

[38] S. P. Driver, D. T. Hill, L. S. Kelvin, A. S. Robotham, J. Liske, P. Norberg, I. K. Baldry, S. P. Bamford, A. M. Hopkins, J. Loveday, *et al.*, "Galaxy and mass assembly (gama): survey diagnostics and core data release," *Monthly Notices of the Royal Astronomical Society*, vol. 413, no. 2, pp. 971–995, 2011.

[39] I. K. Baldry, J. Liske, M. Brown, A. Robotham, S. Driver, L. Dunne, M. Alpaslan, S. Brough, M. Cluver, E. Eardley, *et al.*, "Galaxy and mass assembly: the g02 field, herschel–atlas target selection and data release 3," *Monthly Notices of the Royal Astronomical Society*, vol. 474, no. 3, pp. 3875–3888, 2018.

[40] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, July 2019.

[41] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, pp. 20–29, June 2004.

[42] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, pp. 42–47, 01 2012.

[43] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002.