

# Clasificación y estimación de *Redshift* de espectros astrofísicos mediante algoritmos de *Machine Learning*

Jairo Andres Saavedra Alfonso  
Asesor: Jaime Ernesto Forero-Romero  
Física  
Avance de Monografía 30%  
Universidad de los Andes  
2019

# Introducción

- La exploración activa de espectros astrofísicos requiere de precisión para determinar clasificación espectral y estimación de *Redshift* del objeto observado. [1]

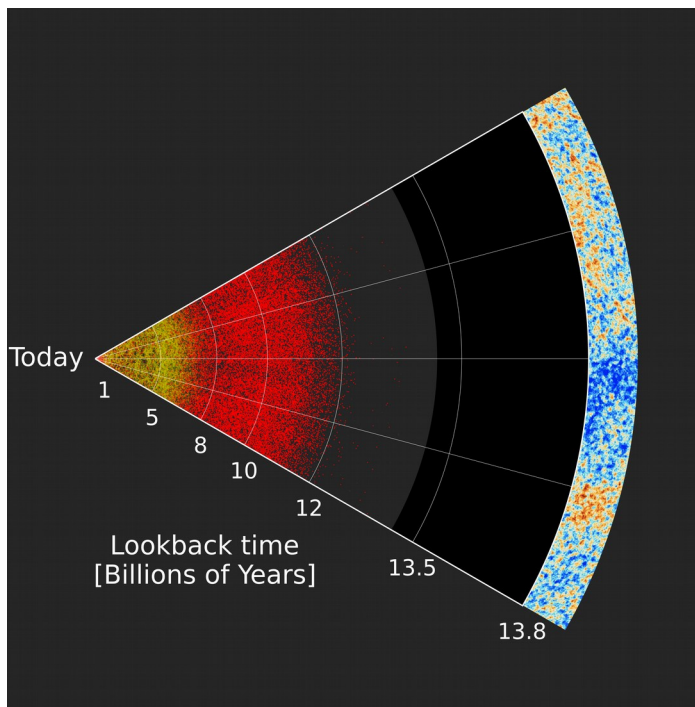


Figura 1: Porción del mapa 3D de estructura a gran escala de SDSS

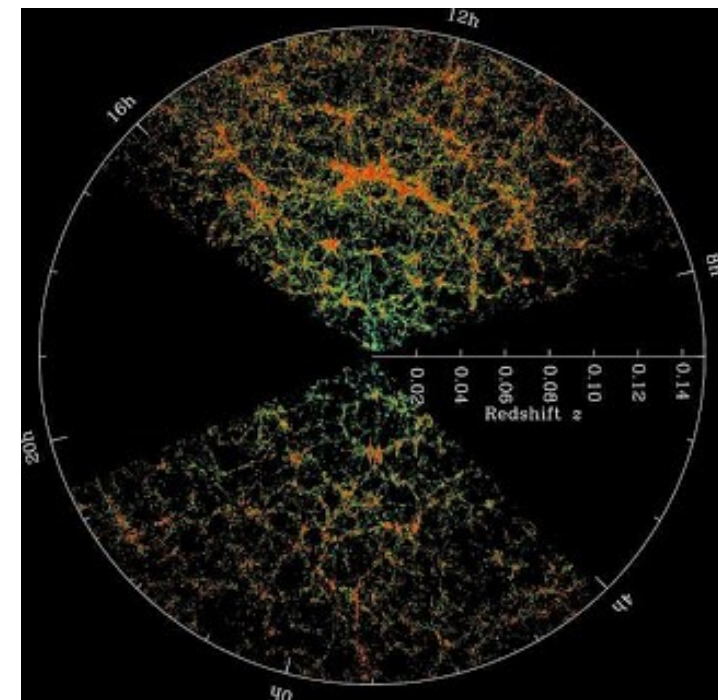


Figura 2: Porción del mapa de Galaxias de SDSS

# Introducción

- Métodos estándar automatizados *REDMOSTER Software (eBOSS)*. [2]

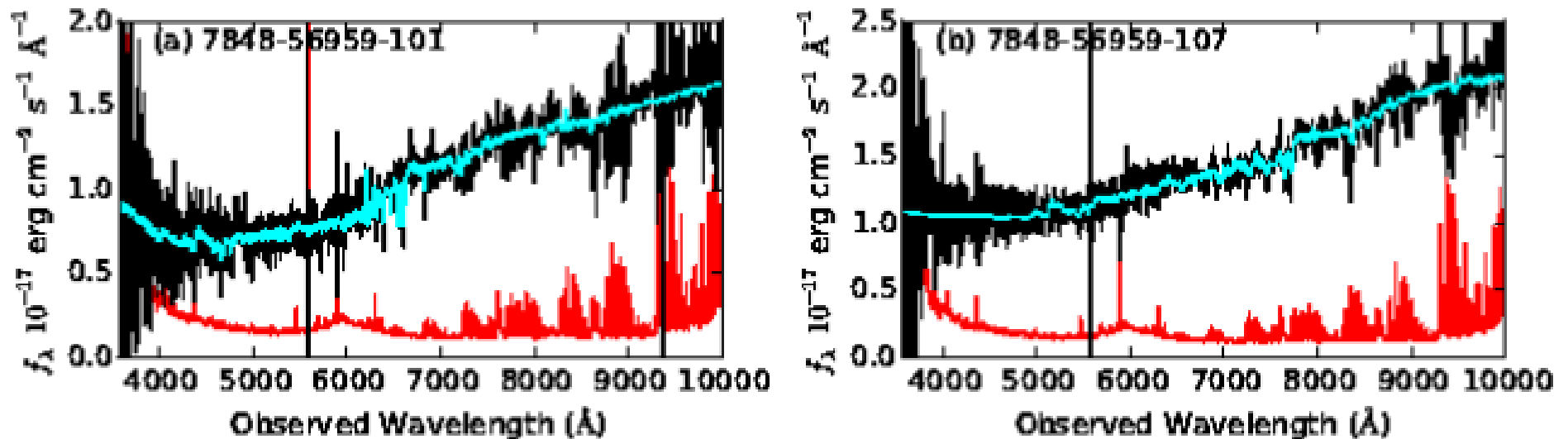


Figura 3: REDMOSTER para espectros de LRG de eBOSS.

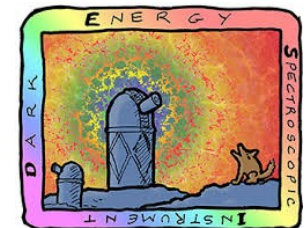
- Clasificación por expertos en área mediante inspección visual del espectro. [3]

# Motivación

- Automatizar los procesos de clasificación y estimación del *Redshift* de espectros para futuros *Surveys (DESI)*



- *Medir el efecto de la materia oscura en la expansión del universo.*
- *~60 millones de espectros de galaxias y QSO [4].*



# Objetivo General

Implementar y evaluar algoritmos de *Machine Learning* (Convolutional Neural Networks, Random Forest) para clasificación y estimación del *Redshift* de espectros astrofísicos de futuros *surveys*.

# Objetivos Específicos

- Construir un conjunto de entrenamiento, test y validación de al menos 100.000 espectros a partir del DR 12 del Baryon Oscillation Spectroscopic Survey (BOSS).
- Implementar una arquitectura de RNC para predecir clase objeto observado, a partir de espectros astrofísicos.
- Implementar una arquitectura de RNC para estimar Redshift del objeto observado.
- Evaluar el rendimiento de los algoritmos propuestos en el reconocimiento de Cuasares y Galaxias.

# Preprocesamiento de los datos

- *SDSS Data Release 12 [4]*
- *Baryon Oscillations Spectroscopic Survey (BOSS)*
- 546856 Objetos (*Stars, Galaxies, QSO & QSO BAL*)
- 639464 Espectros (*Stars, Galaxies, QSO & QSO BAL*)
- 537677 Objetos/Espectros correlacionados
- Remoción de espectros sin ID, sin etiqueta de clase y con Z\_CONF\_PERSON=0
- Estandarización de los espectros 
$$z_i = \frac{x_i - \bar{x}}{\sigma}$$



# Espectros

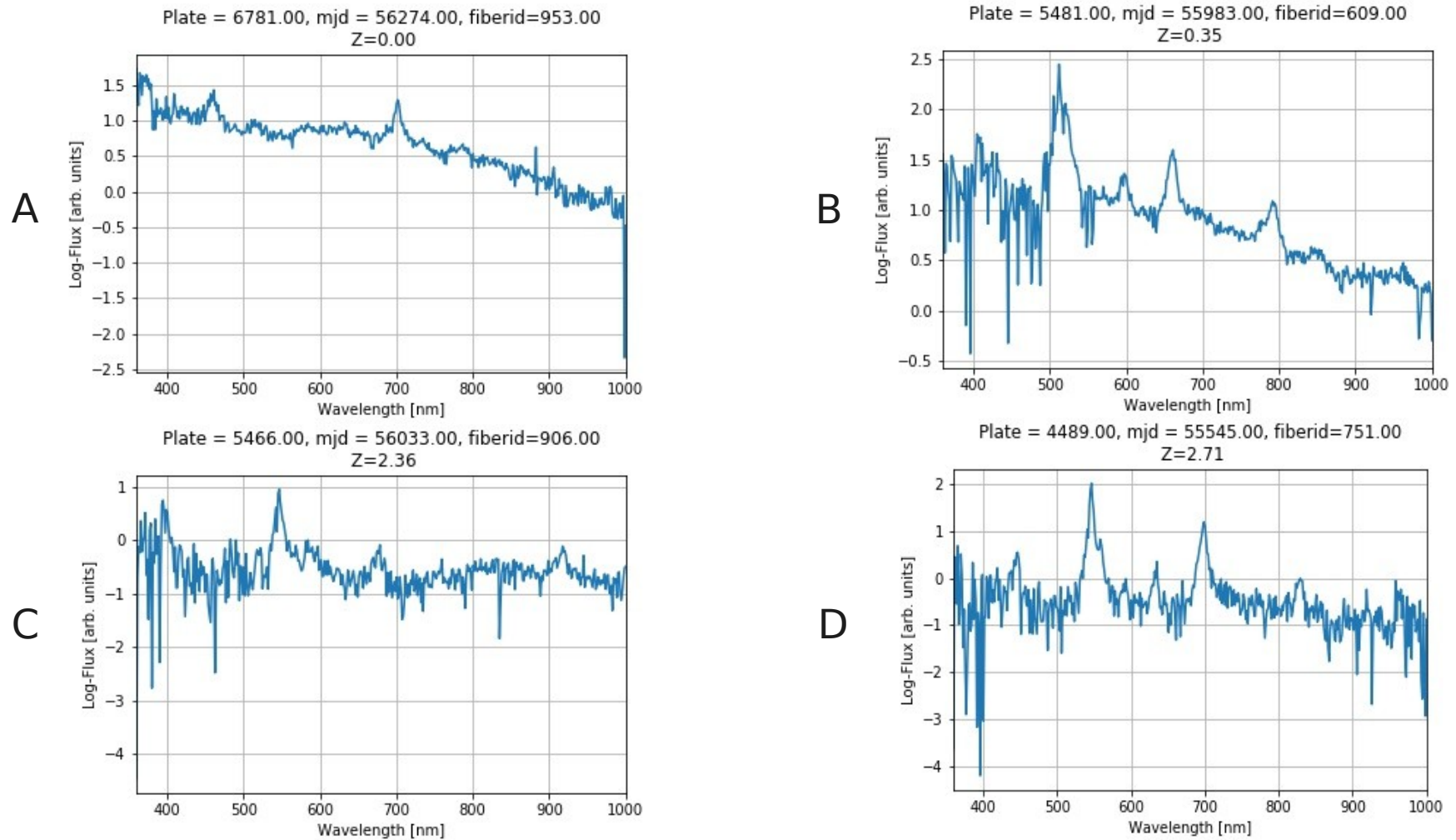


Figura 4: Espectro de (A) Estrellas, (B) Galaxias, (C) QSO y (D) QSO BAL



# Metodologia

Evaluar diferentes estructuras de Redes Neuronales Convolucionales (RNC) para:

# Metodologia

Evaluar diferentes estructuras de Redes Neuronales Convolucionales (RNC) para:

- Clasificación espectral (Estrellas, Galaxias QSO y QSO *BAL*).

# Metodologia

Evaluar diferentes estructuras de Redes Neuronales Convolucionales (RNC) para:

- Clasificación espectral (Estrellas, Galaxias, QSO y QSO *BAL*).
- Estimar *Redshift* de los objetos observados (Regresión).

# Clasificación Espectral RNC 1.0

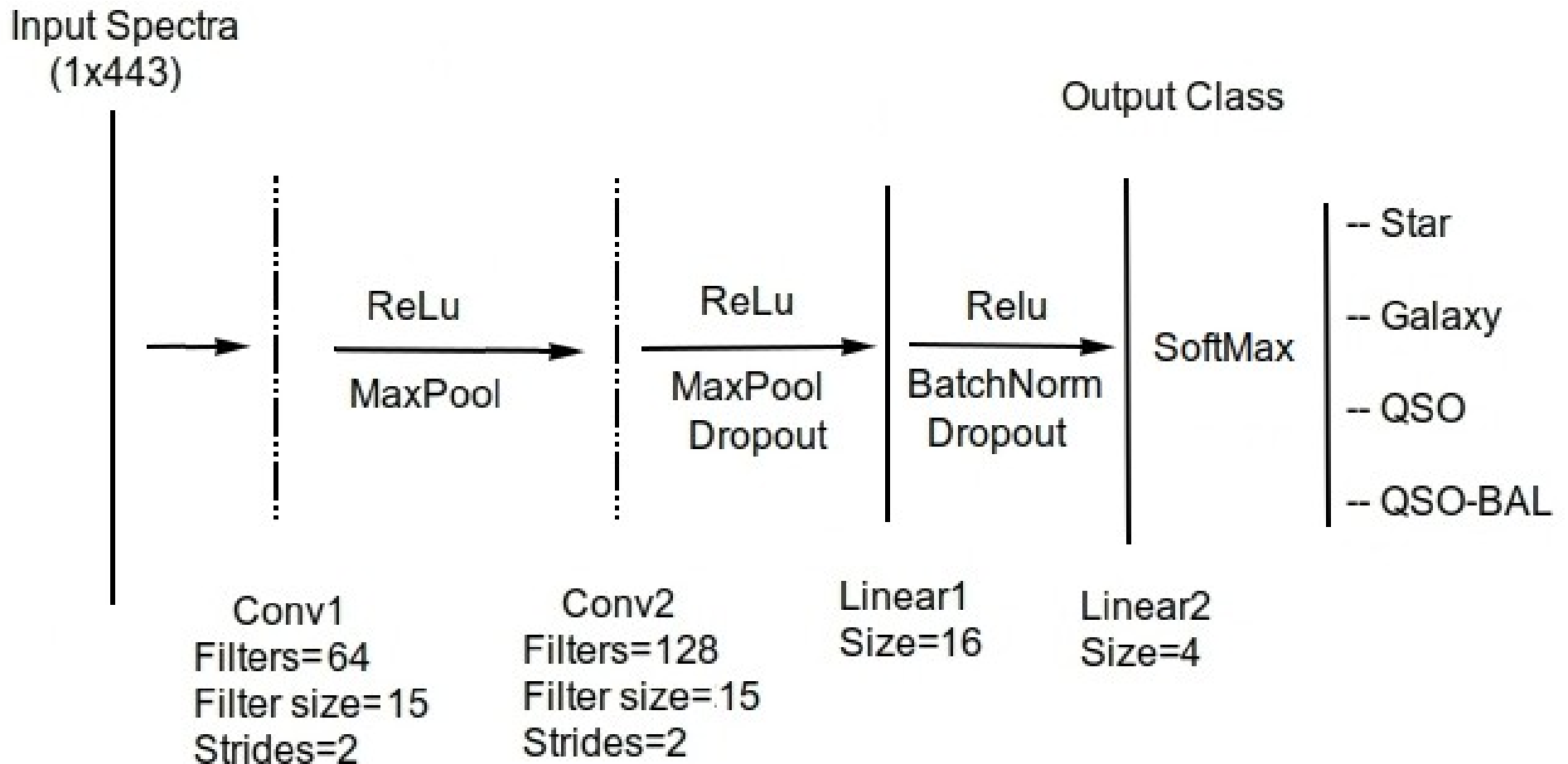


Figura 5: Primera estructura tentativa de RNC para clasificación espectral

# Clasificación Espectral RNC 2.0

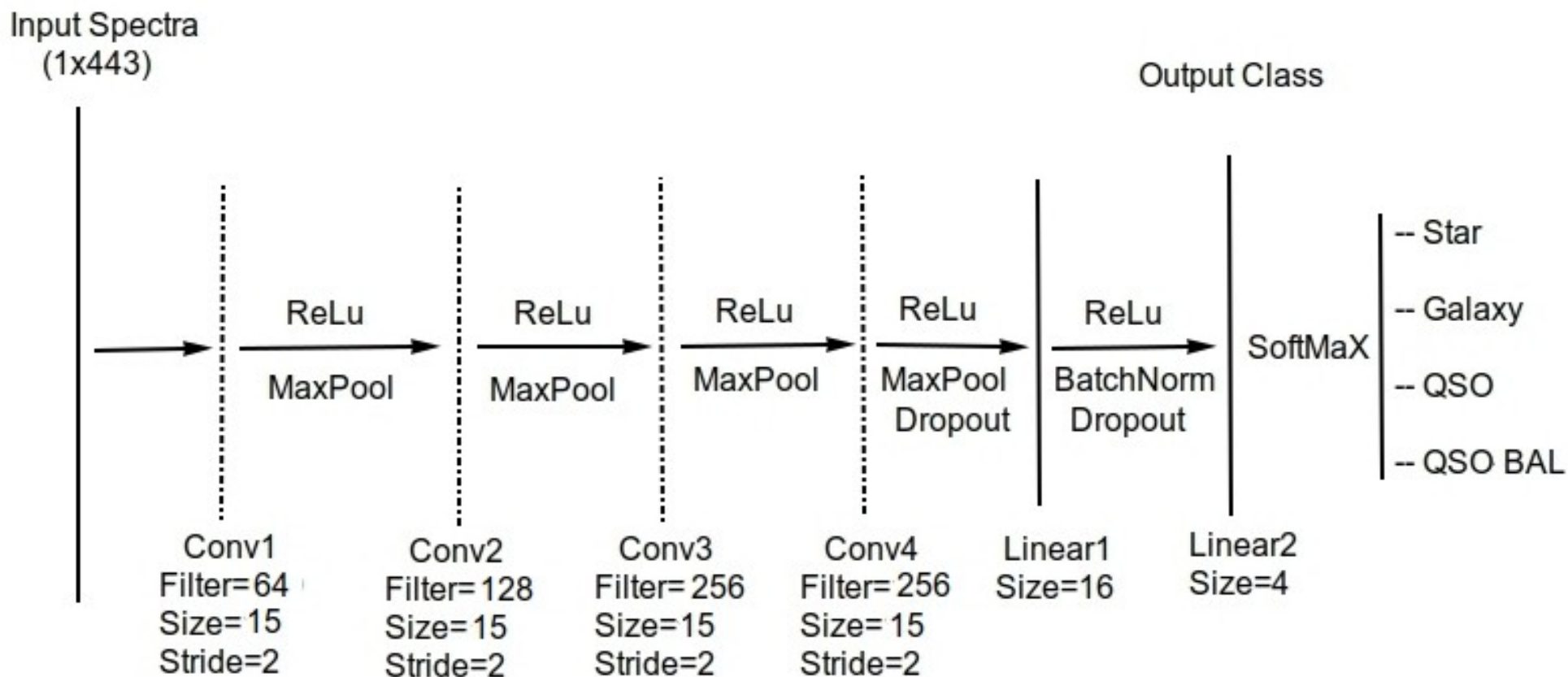


Figura 6: Segunda estructura tentativa de RNC para clasificación espectral

# Entrenamiento

- 80.000 Espectros (240 espectros por *batch*)
- 50 épocas - 10.000 Iteraciones.
- 60/20/20 - *Train/Validation/Test*.
- Espectros con ID y CLASS\_PERSON valida

Objeto	Cant. de espectros
Estrellas	207905
Galaxias	20699
QSO	270534
QSO BAL	29652

Tabla 1: Datos disponible para entrenamiento

# Matriz de confusión test Clasificación Espectral

	Actual Positives	Actual Negatives
Positive Predictions	True Positives (TP)	False Positives (FP)
Negative Predictions	False Negatives (FN)	True Negatives (TN)

Figura 7: Matriz de confusión

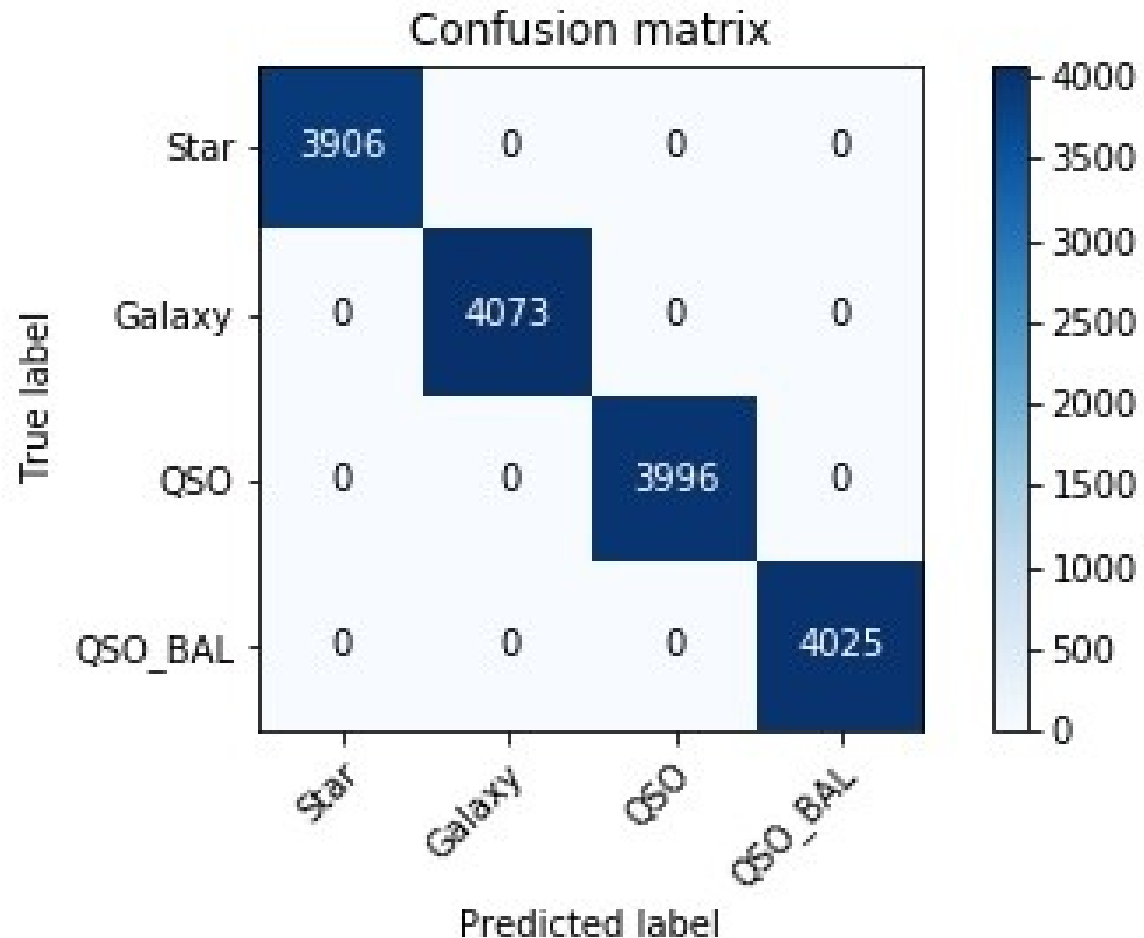


Figura 8: Matriz de confusión Test



# Matriz de confusión test

## RNC 1.0

## RNC 2.0

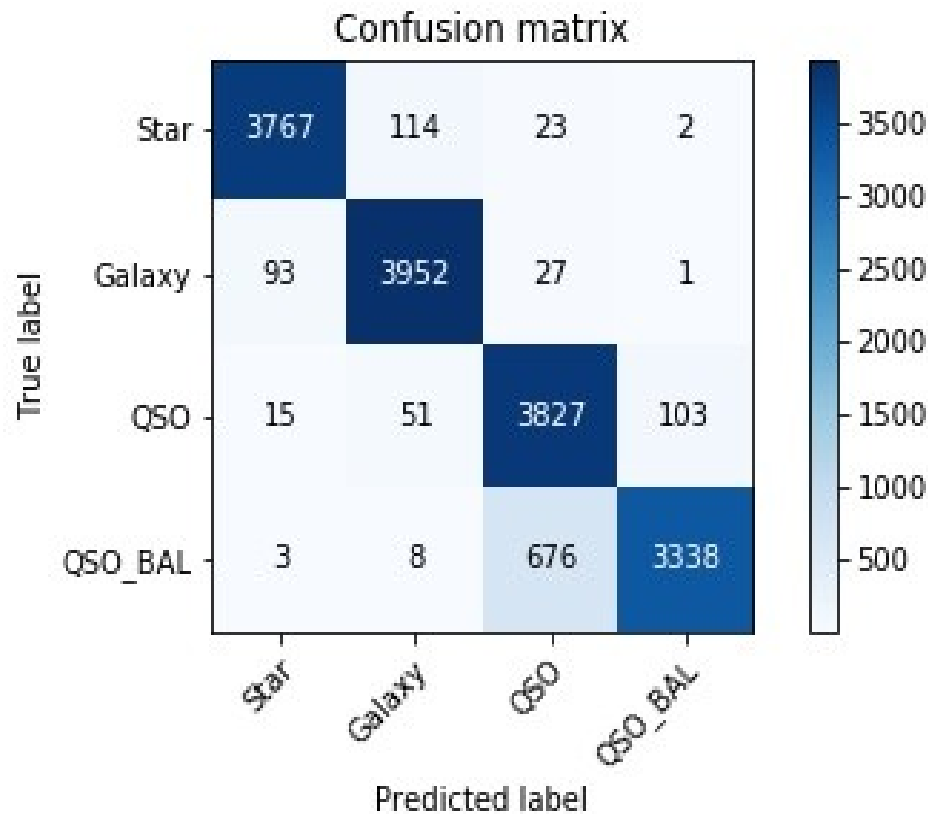


Figura 9: Matriz de confusión RNC 1.0

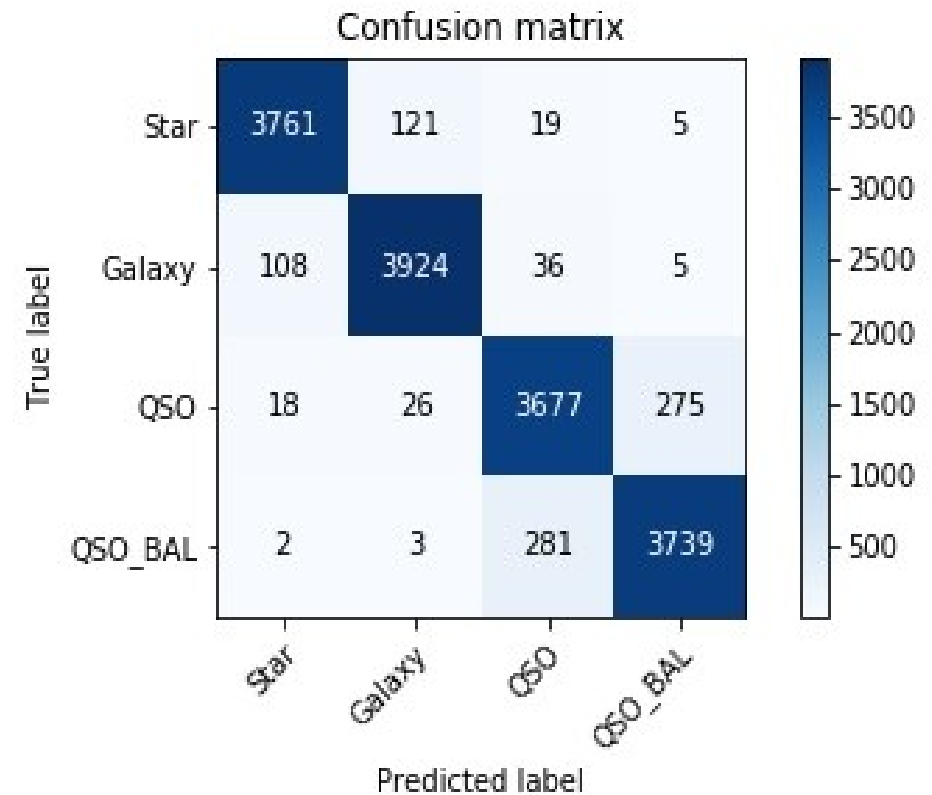


Figura 10: Matriz de confusión RNC 2.0

# Metricas para Clasificación Espectral

	Actual Positives	Actual Negatives
Positive Predictions	True Positives (TP)	False Positives (FP)
Negative Predictions	False Negatives (FN)	True Negatives (TN)

Figura 11: Matriz de confusión Test

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

# Metricas de los modelos RNC 1.0 & RNC 2.0

Values/Obj	Stars	Galaxies	QSO	QSO BAL
Support	3906	4073	3996	4025
Precision	0.97	0.96	0.84	0.97
Recall	0.96	0.97	0.96	0.83
F1	0.97	0.96	0.90	0.89
Accuracy	96.44 %	97.03 %	95.77 %	82.93 %

Tabla 2: Valores de relevancia RNC 1.0

Values/Obj	Stars	Galaxies	QSO	QSO BAL
Support	3906	4073	3996	4025
Precision	0.97	0.96	0.92	0.93
Recall	0.96	0.96	0.92	0.93
F1	0.96	0.96	0.92	0.93
Accuracy	98.22 %	98.06 %	95.84 %	96.36 %

Tabla 3: Valores de relevancia RNC 2.0

# QuasarNET

## QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks

Nicolas G. Busca,<sup>1★</sup> and Christophe Balland,<sup>1</sup>

<sup>1</sup>*Sorbonne Université, Université Paris Diderot, CNRS/IN2P3,  
Laboratoire de Physique Nucléaire et de Hautes Energies, LPNHE, 4 Place Jussieu, F-75252 Paris, France*

Accepted XXX. Received YYY; in original form ZZZ

### ABSTRACT

We introduce QuasarNET, a deep convolutional neural network that performs classification and redshift estimation of astrophysical spectra with human-expert accuracy. We pose these two tasks as a *feature detection* problem: presence or absence of spectral features determines the class, and their wavelength determines the redshift, very much like human-experts proceed. When ran on BOSS data to identify quasars through their emission lines, QuasarNET defines a sample  $99.51 \pm 0.03\%$  pure and  $99.52 \pm 0.03\%$  complete, well above the requirements of many analyses using these data. QuasarNET significantly reduces the problem of line-confusion that induces catastrophic redshift failures to below 0.2%. We also extend QuasarNET to classify spectra with broad absorption line (BAL) features, achieving an accuracy of  $98.0 \pm 0.4\%$  for recognizing BAL and  $97.0 \pm 0.2\%$  for rejecting non-BAL quasars. QuasarNET is trained on data of low signal-to-noise and medium resolution, typical of current and future astrophysical surveys, and could be easily applied to classify spectra from current and upcoming surveys such as eBOSS, DESI and 4MOST.

**Key words:** cosmology: observations – quasars: emission lines – quasar: absorption lines

# QuasarNET vs RNC 2.0

	QuasarNET	RNC 2.0
Metrica/Obj	QSO	QSO
Precision	99.51	92.2
Recall	99.52	92.1
F1	99.51	92.1

Tabla 4: Metricas de QuasarNET vs RNC 2.0

# Estimación del Redshift con RNC 2.0

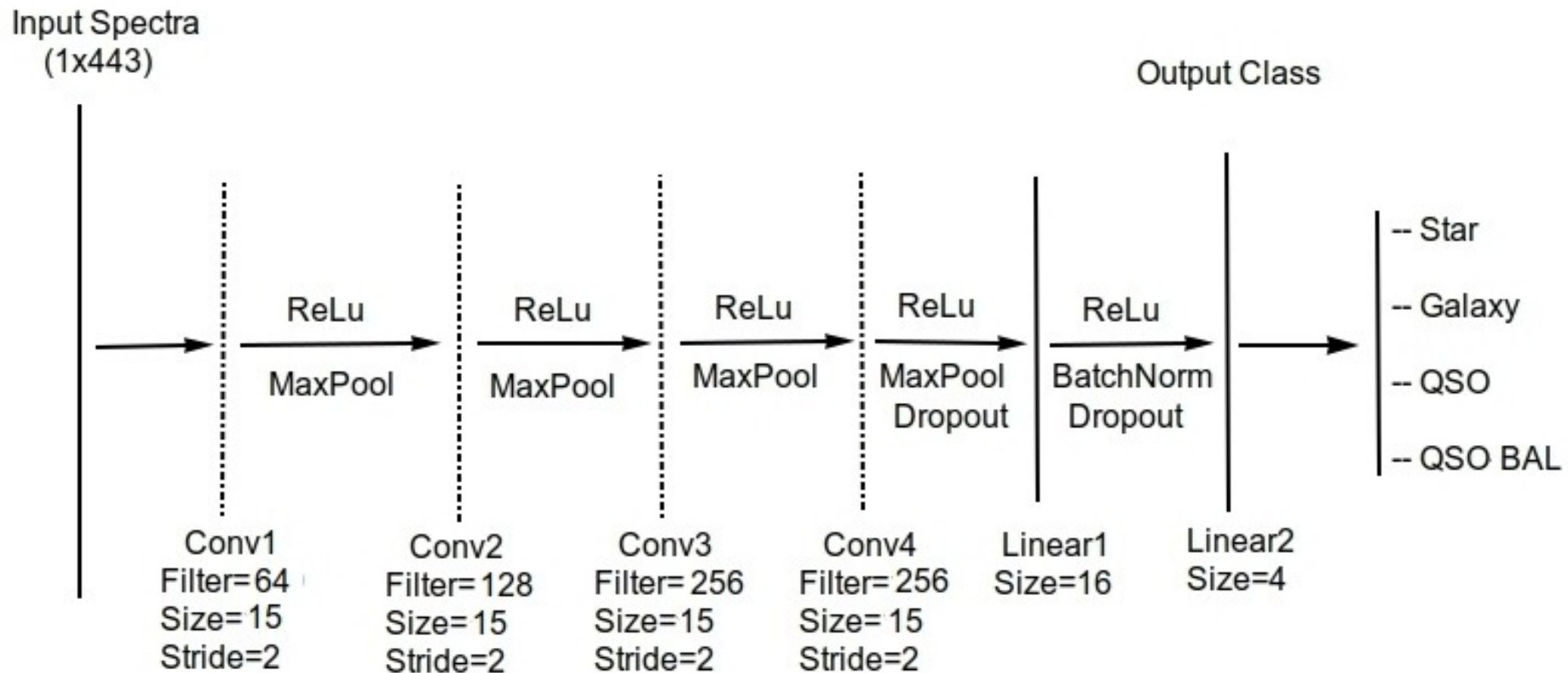


Figura 11: RNC 2.0 para estimación del Redshift de QSO

# Entrenamiento

- 80.000 Espectros (240 espectros por *batch*)
- 50 épocas - 10.000 Iteraciones.
- 60/20/20 – *Train/Validation/Test*.
- Espectros con  $Z\_CONF\_PERSON=3$

Objeto	Cant. de espectros
Estrellas	207905
Galaxias	20699
QSO	270534
QSO BAL	29652

Tabla 1: Datos disponible para entrenamiento



# Evaluación de RNC 2.0 en regresión de Redshift

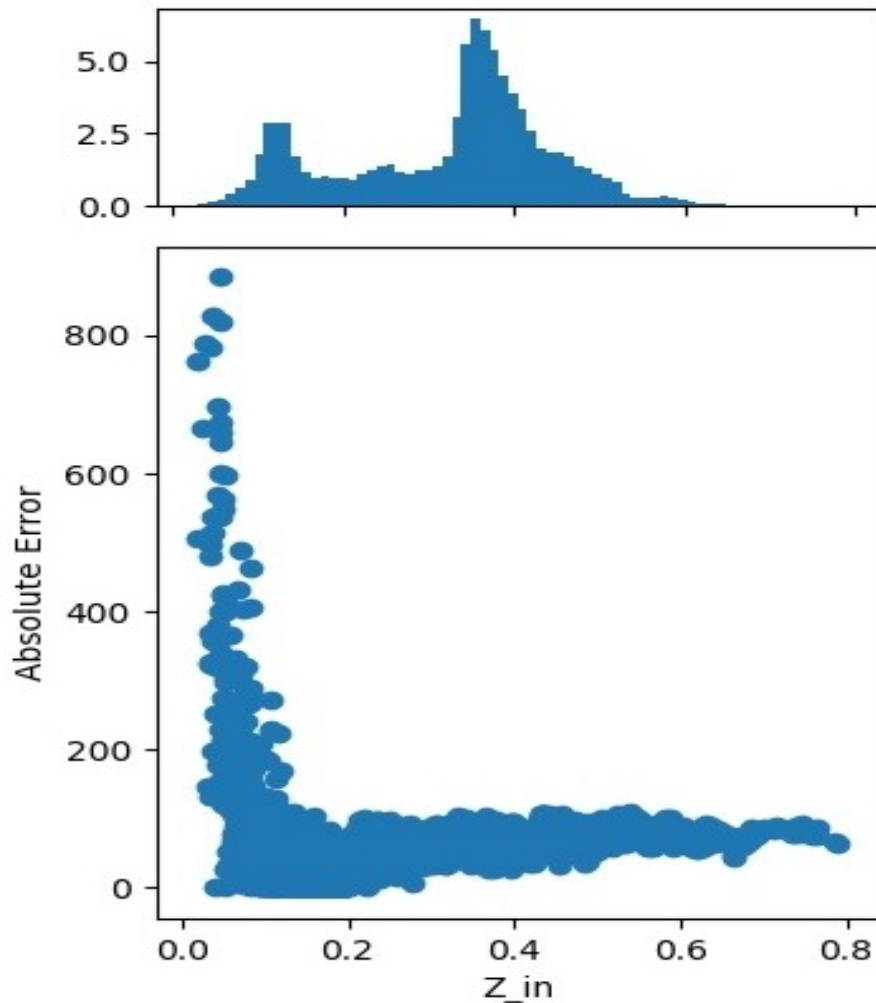


Figura 12: Error Abs. Vs  $Z_{in}$

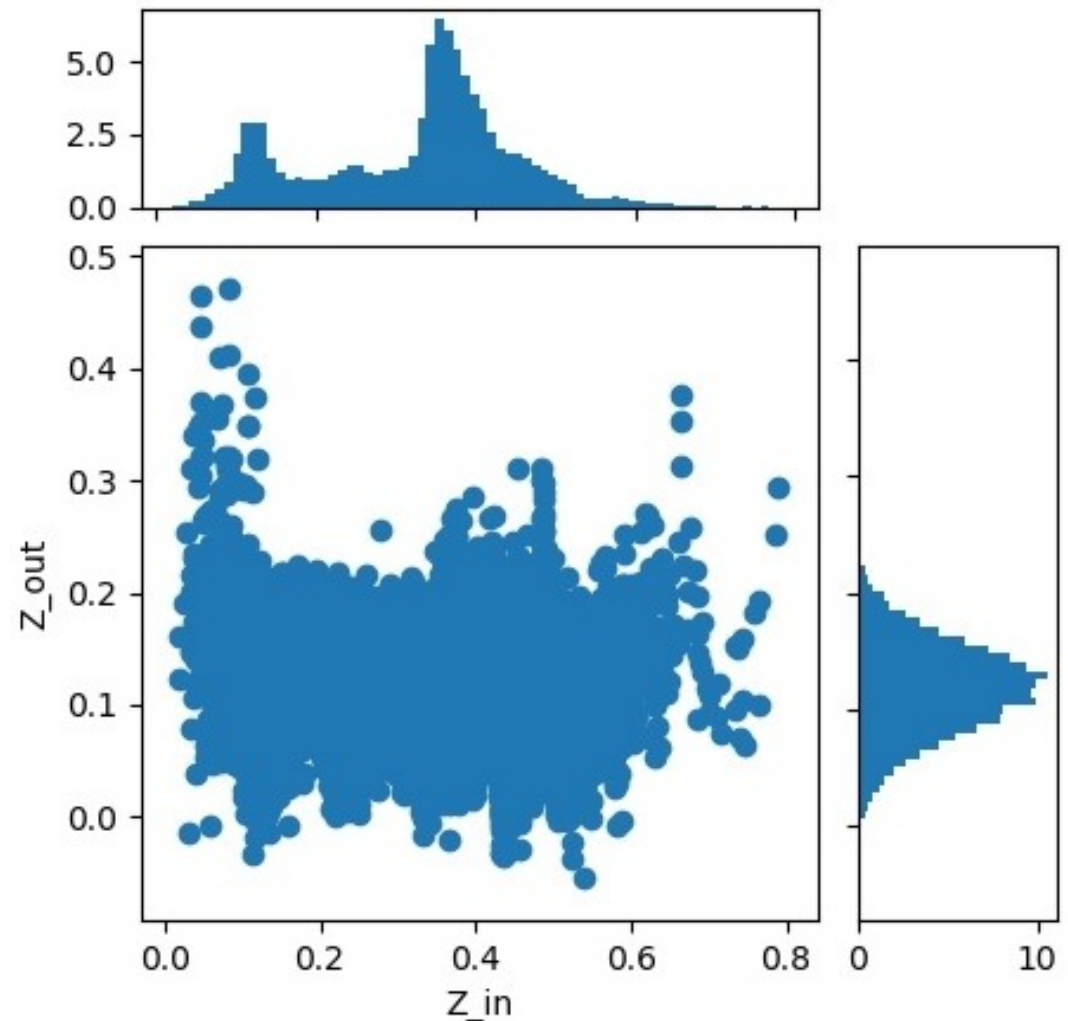


Figura 13:  $Z_{out}$  vs  $Z_{in}$

# Conclusiones Preliminares

- Se demostro que es posible usar RNC para clasificación espectral de Estrellas, Galaxias y Cuasares.
- Se observo como la elección de los hiperaparametros afecta los resultados de las RNC.
- Se pudo observar como las RNC tiene bajo rendimiento cuando la distribución de datos es baja.
- Es necesario explorar otros algoritmos de Machine Learning para la regresión del Redshift (*Random Forest, SVM...*).

# Cronograma

- Tarea 1: Revisar bibliografía
- Tarea 2: Cargar datos de espectros lineales.
- Tarea 3: Explorar los datos disponibles.
- Tarea 4: Correlacionar los objetos astrofísicos con sus respectivos espectros lineales para crear una base de datos de entrenamiento.
- Tarea 5: Implementar y evaluar una Red Neuronal simple para observar el rendimiento con un conjunto reducido de datos.
- Tarea 6: Presentar avance correspondiente al 30 %

Tareas \ Semanas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	X	X														
2		X	X													
3			X	X												
4				X	X	X										
5						X	X									
6								X								
7									X	X	X					
8												X	X	X		
9															X	X
10								X	X	X	X	X	X	X	X	X

# Objetivos Cumplidos

- Construir un conjunto de entrenamiento, prueba y validación de al menos 100.000 espectros a partir del *Data Release 12* de *Baryon Oscillation Spectroscopic Survey (BOSS)*.
- Implementar una arquitectura de RNC para predecir clase objeto observado, a partir de espectros astrofísicos.
- Implementar algoritmos de *Machine Learning* para estimar *Redshift* del objeto observado.
- Evaluar el rendimiento de los algoritmos propuestos en el reconocimiento de Cuasares y Galaxias.

# Trabajo Futuro

- Explorar hiperparámetros (*epochs*, *batch size*, *learning rate*, *hidden layers*....) con el fin de mejorar el rendimiento de la RNC sobre los datos de test para clasificación espectral.
- Explorar otros algoritmos de *ML* (*Random Forest*, *SVM*...) para estimar *Redshift* a partir de los espectros.
- Extender la estimación de *Redshift* (Regresión) para QSO a Galaxias.

# Referencias

- [1] Alam, S. et al. (2017). *The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample*. Monthly Notices of the Royal Astronomical Society, 470(3), 2617-2652.
- [2] Hutchinson, T. et al. (2016). Redshift measurement and spectral classification for eboss galaxies with the redmonster software. The Astronomical Journal, 152(6), 205.
- [3] Dawson, K. et al. (2013). *The Baryon Oscillation Spectroscopic Survey of SDSS-III*. arXiv preprint arXiv:1208.0022, 145(10).
- [4] Aghamousa, A. et al. (2016). *The DESI experiment Part I: science, targeting, and survey design*. arXiv preprint arXiv:1611.00036.