Celestial Object Classification

Xiaoyang Wang Ziang Zeng

Outline

Section 1

Astronomical Challenge

Astronomical Challenge

Classifying celestial objects into stars, galaxies or quasars.

- Stars: a luminous sphere of plasma held together by its own gravity.
- Galaxy: a massive, gravitationally bound system that consists of stars, stellar remnants, interstellar gas, dust, and dark matter.
- Quasars: a very energetic and distant active galactic nucleus, with its energy output sometimes surpassing that of the rest of the galaxy combined.

Idea

- There are a lot of classification models: KNN,Tree,Logistic Regression, Neural Networks
- Different models may perform differently on same data
- Can we combine them together to make more accurate classification?
- Voting Classifier

Section 2

Data & Preprocessing

Images

Figure 1: Galaxy

Figure 2: Star

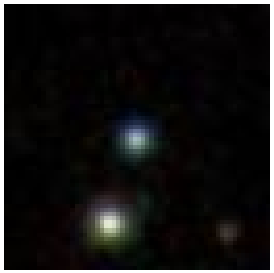


Figure 3: Qusar

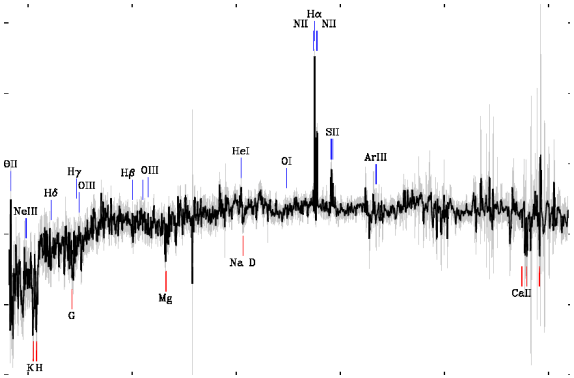


Figure 4: Galaxy Spec

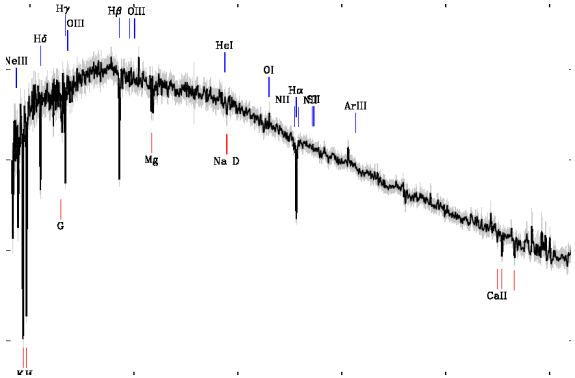


Figure 5: Star Spec

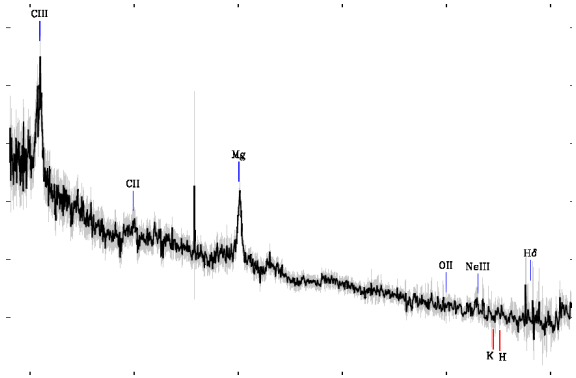


Figure 6: Qusar Spec

Metadata

Table 1: Metadata of the celestial objects

vars	explanations
ra	Right Ascension angle (at J2000 epoch)
dec	Declination angle (at J2000 epoch)
u	Ultraviolet filter
g	Green filter
r	Red filter
i	Near Infrared filter
z	Infrared filter
run	Run Number
rerun	Rerun Number
camcol	Camera column
field	Field number
specobjid	Unique ID used for optical spectroscopic objects
class	Object class
redshift	Redshift value based on the increase in wavelength
plate	Plate
mjd	Modified Julian Date

EDA

- Missing Values:
  - Metadata: 3, Regression Imputation
  - Image of Spectra: 14115
- Samples for each category: 33333
- Correlationship

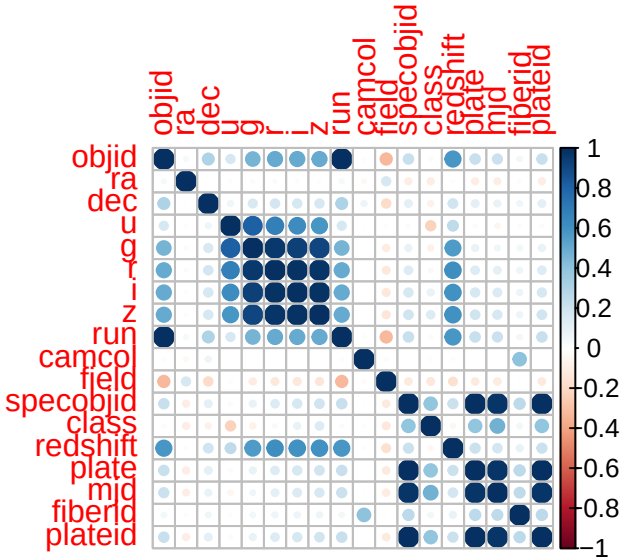


Figure 7: Correlationship of Variables

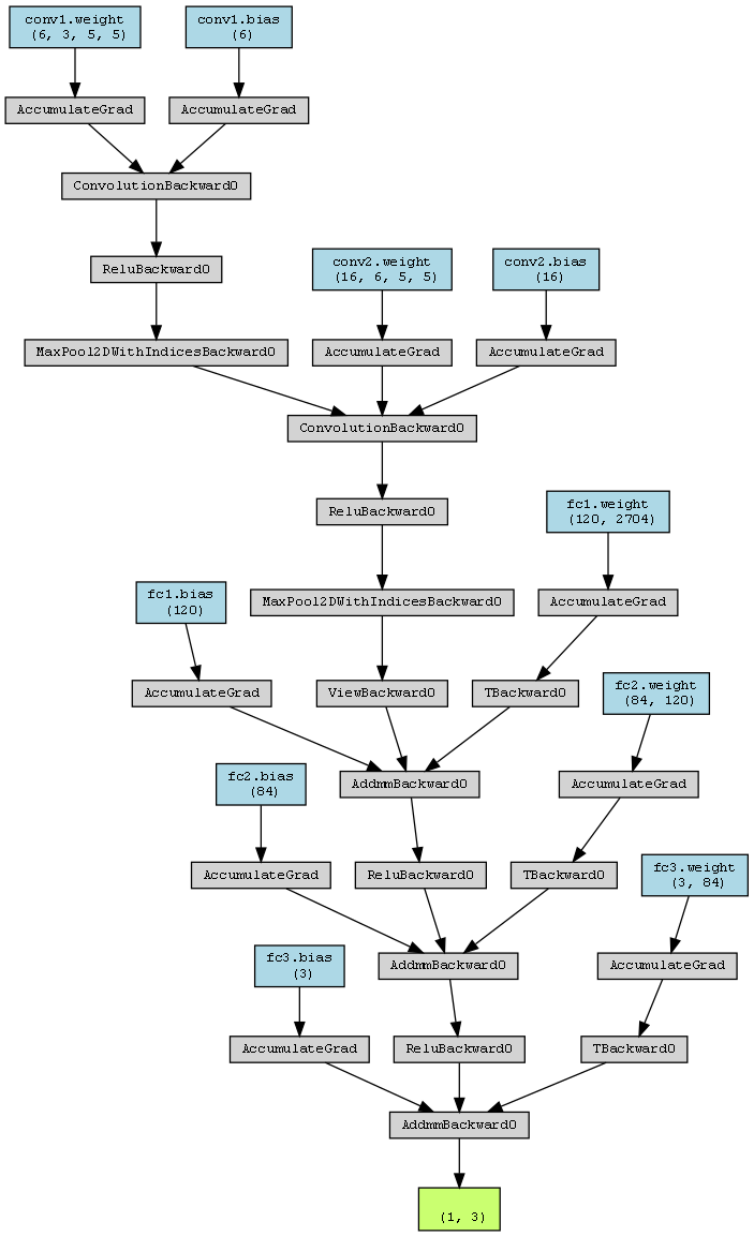
Section 3

Methodology

Meta Data

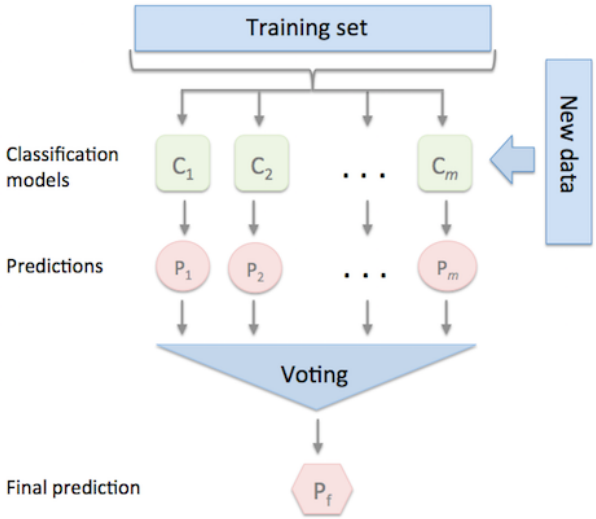
- **Explanatory Variables**: u, g, r, i, z, redshift
- **Response Variable**: class
  - GALAXY: 0
  - QSO: 1
  - STAR: 2
- **kNN**: k = 3
- **Decision Tree**:
  - Gini impurity
  - max_depth: 4
- **Logistic Regression**
  - C: 1
  - penalty: l2
  - $P(Y_i = k) = \frac{e^{\beta_k \cdot x_i}}{\sum_{j=1}^{3} e^{\beta_j \cdot x_i}}$, $i = 0, 1, 2$

Images



- Structure:
  - 2 layers of convolution and 1 maxpooling
  - 3 layers of full connecting
- Output:
  - $\vec{y} = (y_1, y_2, y_3)$
  - $y_{pred} = argmax_i\{\vec{y}\}$
  - Probability through softmax $P(y = j \mid \mathbf{z}) = \frac{e^{z_j}}{\sum_{k=1}^{3} e^{z_k}}$
- Training:
  SGD with different momentum, Adam, 10 epoch, batch size 64,lr 0.001

Voting Classifier



- Soft Voting:
  - Models $\{C_1, \cdots C_n\}$
  - For a given inputs, C_i has a predict probability $P_i(y_j|x)$
  - The probabilities for voting classifier $P(y_j|x) = \frac{1}{m}\sum_{i=1}^{m} P_i(y_j|x)$
  - The prediction $p(x) = \arg\max_{y_j} P(y_j|x)$
- Weighted Hard Voting: For a given inputs, C_i has a predict $y^i|x : y^i_{k=j} = 1, y^i_{k\neq j} = 0$ $y_{pred} = \sum_i w_i \cdot y^i|x$, here we use accuracy of each model as their weight, the predict class is $argmax_j\ y_{pred}$
- Construction:

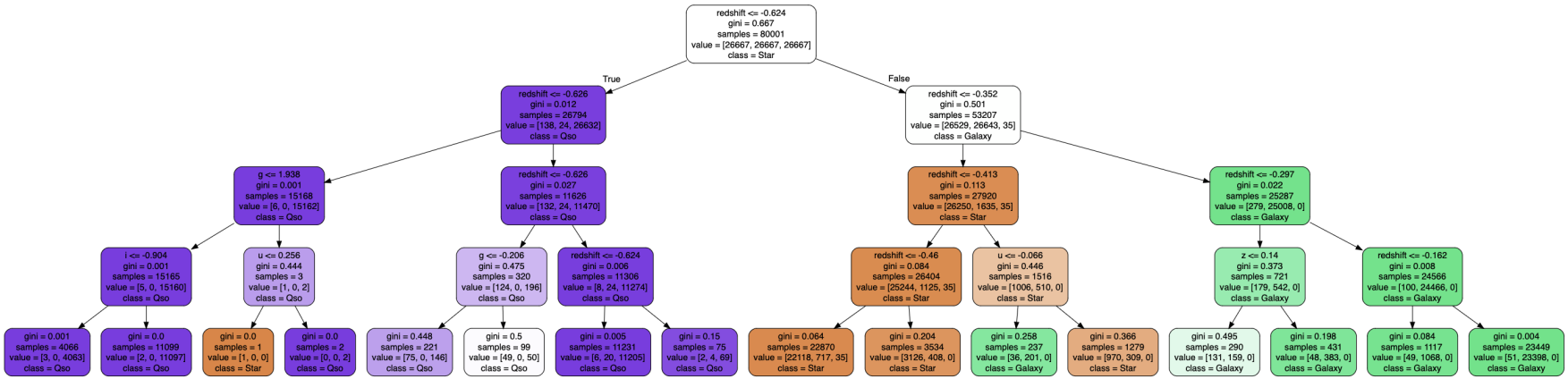
  The candidate models are KNN, Logistic Regression, Decision Tree, CNN for celestial objects image and CNN for spectrum image.

Section 4

Results

Metadata

- Decision Tree:



- Logistic Regression:

Table 2: Coefficients of Logistic Regression

	Intercept	u	g	r	i	z	redshift
Galaxy	15.09801	1.110865	-1.698055	-0.1525521	0.6145228	-0.0238246	23.35724
Qso	16.80773	-2.883481	5.212935	0.7959545	-1.2216091	-2.1410609	32.50714
Star	-31.90574	1.772616	-3.514880	-0.6434025	0.6070863	2.1648855	-55.86438

Metadata

- The accuracy of KNN: 96.80%, Tree: 97.44%, Logistic Regression: 97.00%
- Soft voting: 97.75%, Hard voting: 97.55%

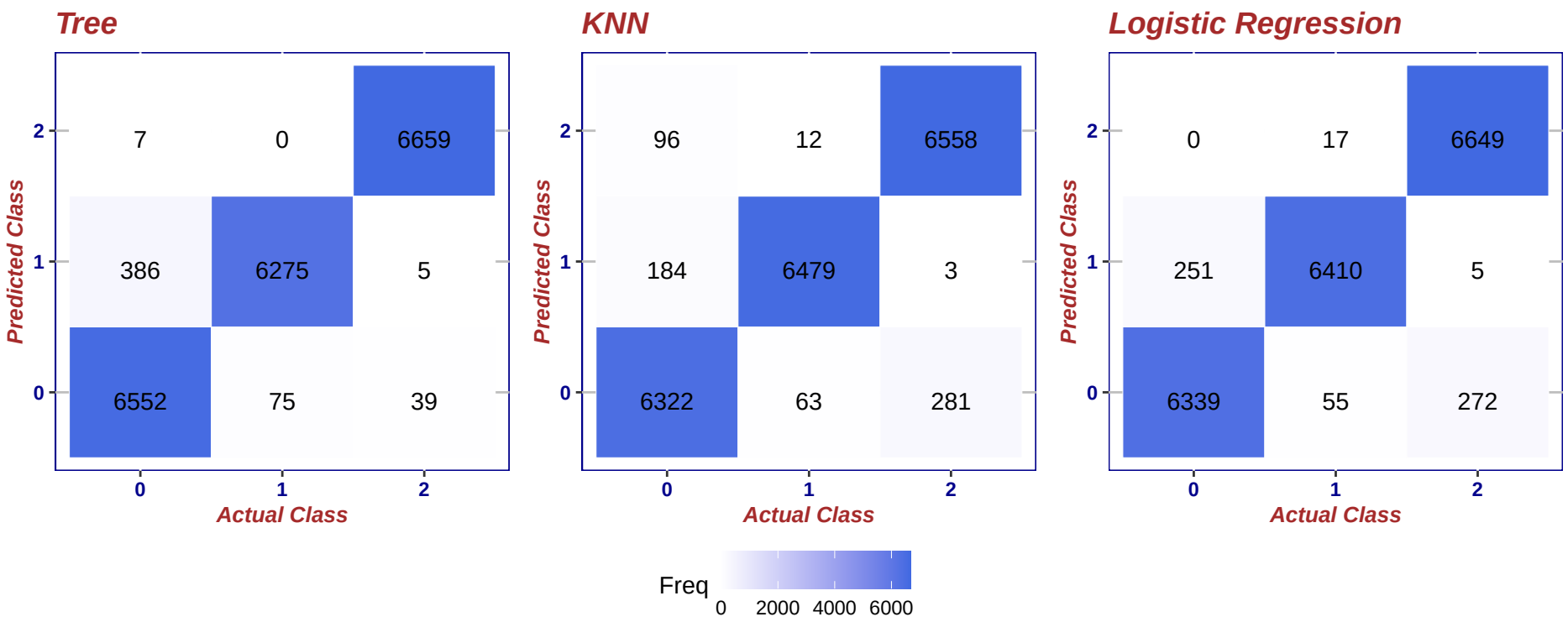


Figure 8: Confusion Matrices of Metadata Models

Images

- The accuracy of CNN of images 91.73%, CNN of spectrum is 88.91%
- Soft voting: 97.71%, Hard voting: 91.74%

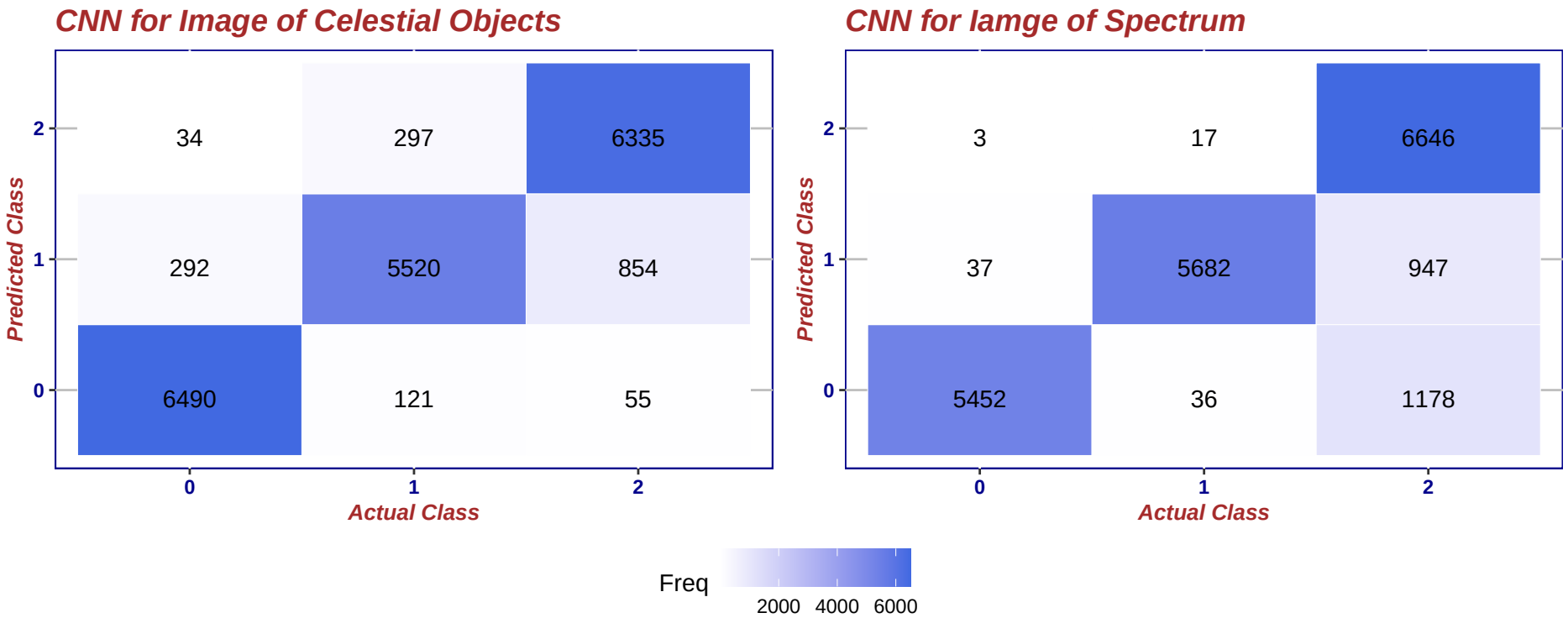


Figure 9: Confusion Matrices of CNN Models

Voting Classifier

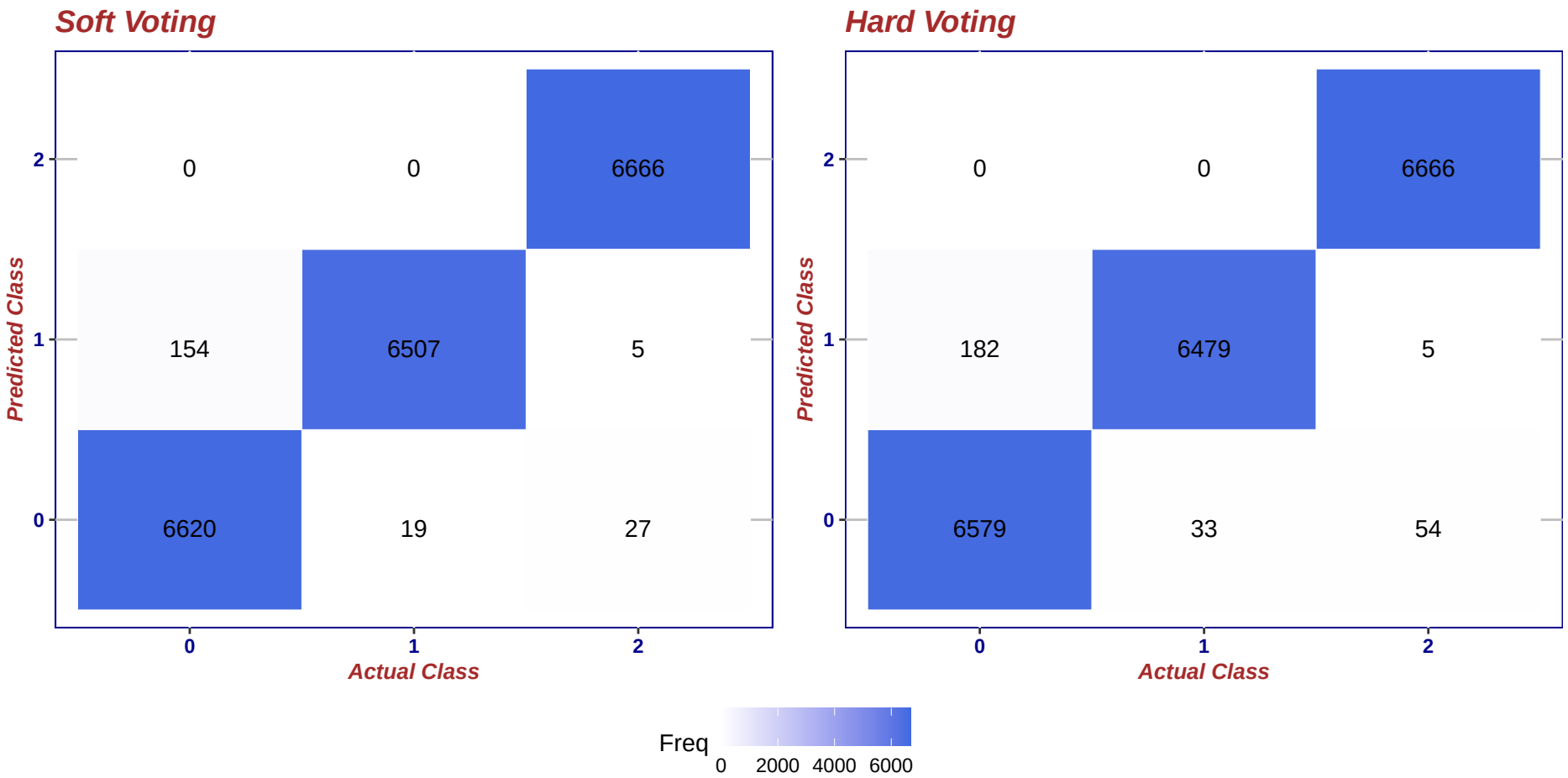


Figure 10: Confusion Matrices of Voting Classifier

Overall

Table 3: Evaluation of Models

Data		M	M	M	IC	IS	M+IC+IS	M+IC+IS
Model		kNN	DT	LR	CNN	CNN	SVC	HVC
Accuracy		0.968	0.9744	0.97	0.9173	0.8891	0.9897	0.9863
Precision	Galaxy	0.9576	0.9434	0.9619	0.9522	0.9927	0.9773	0.9731
	Qso	0.9886	0.9882	0.9889	0.9296	0.9908	0.9971	0.9949
	Star	0.9585	0.9934	0.96	0.8745	0.7577	0.9952	0.9912
Recall	Galaxy	0.9484	0.9829	0.9509	0.9736	0.8179	0.9931	0.9869
	Qso	0.9719	0.9413	0.9616	0.8281	0.8524	0.9761	0.9719
	Star	0.9838	0.9989	0.9974	0.9503	0.997	1	1
F1	Galaxy	0.953	0.9628	0.9564	0.9628	0.8969	0.9851	0.98
	Qso	0.9802	0.9642	0.9751	0.8759	0.9164	0.9865	0.9833
	Star	0.971	0.9962	0.9784	0.9109	0.861	0.9976	0.9956

Note:

M: Metadata. IC: Image of Celestial Objects. IS: Image of Spectrum.

Section 5

Conclusions

Conclusions

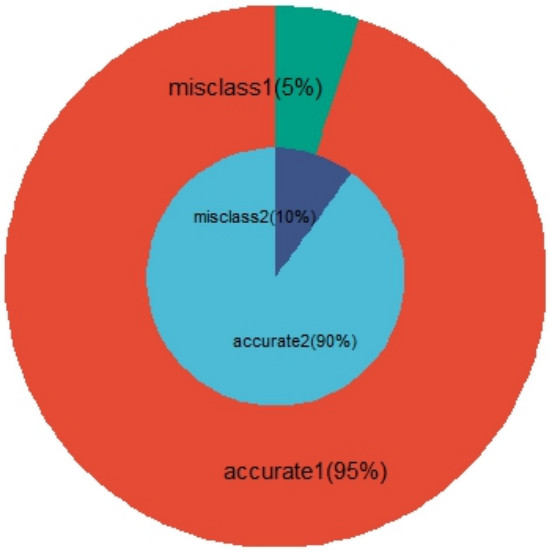


Figure 11: Voting

- For all the combination of classifiers, Soft Voting performs better than Hard Voting
- For every data set, Voting method performs equal or better than single model

Section 6

Future Work

Future Work

- Include more models
- Use less data
- Add noise to the data
- Consider more voting methods

References I

[1] Jialin Gao, Jianyu Chen, Jiaqi Wei, Bin Jiang, and A-Li Luo. Deep multimodal networks for m-type star classification with paired spectrum and photometric image. *Publications of the Astronomical Society of the Pacific*, 135:044503, 05 2023.