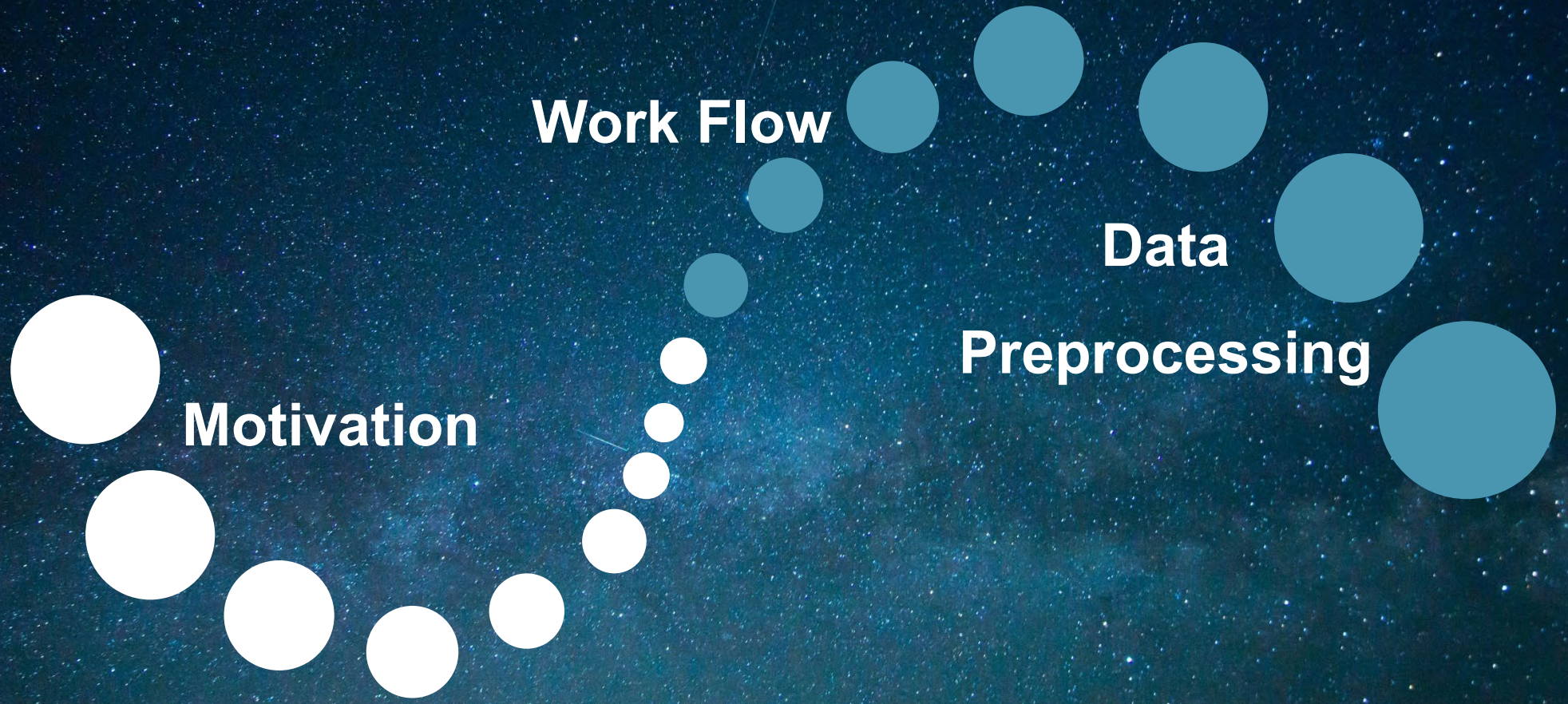
Machine Learning Project Presentation --Sloan Digital Sky Survey

SIQI SHEN
NAIQING CAI
HAO PAN

CONTENT


Introduction

Introduction



Introduction — Motivation

Our data is from Sloan Digital Sky, which is an official website releasing professional observations and dataset regarding sky and astronomy.

Astronomical features and classification hold a wealth of knowledge about the universe and its origin.

Task: Classify objects as stars, galaxies, and quasars.

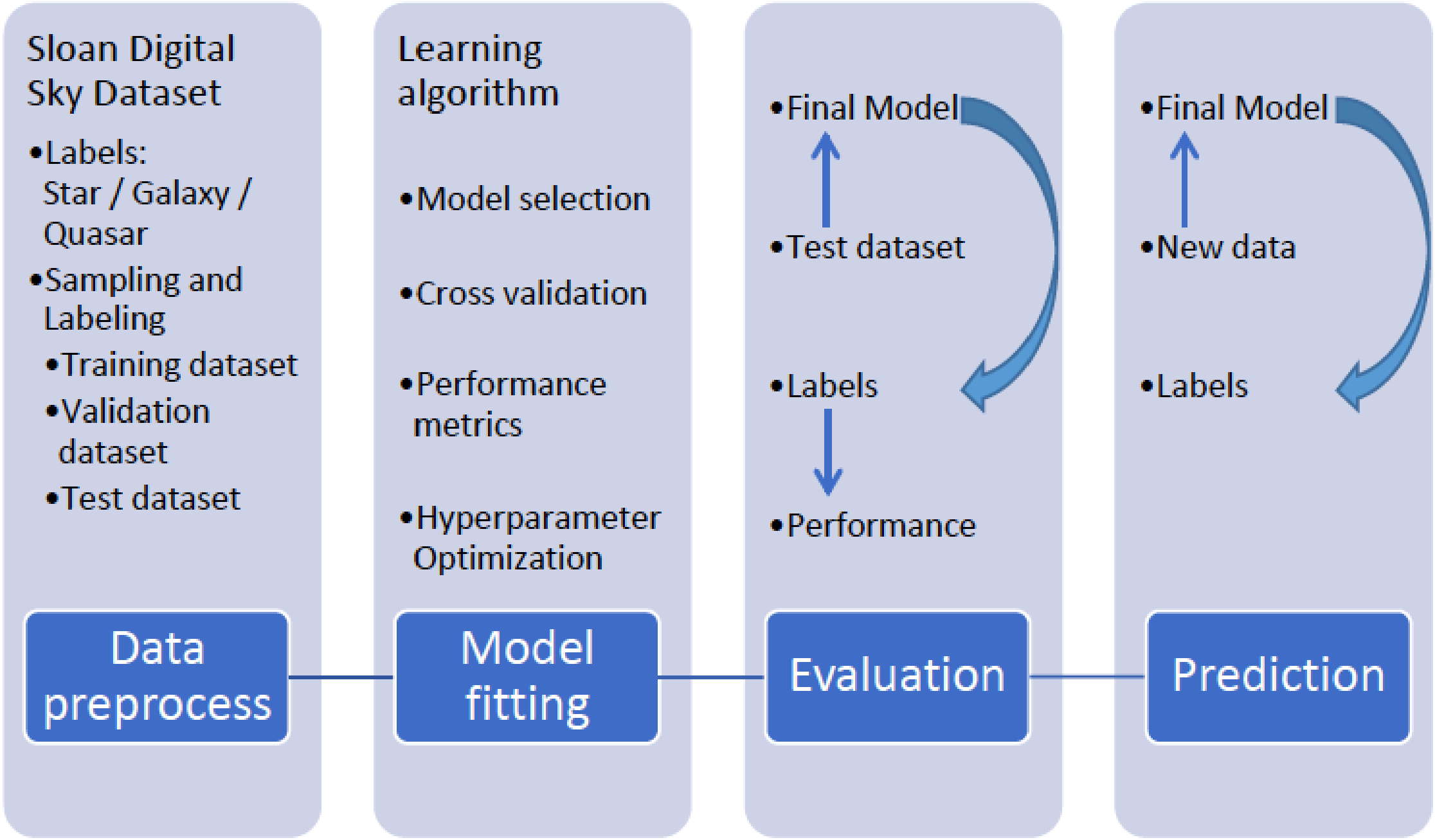
Introduction—Work Flow

Data Preprocessing 01

Model Fitting 02

Evaluation 03

Prediction 04


Sloan Digital Sky Dataset
•Labels: Star / Galaxy / Quasar
•Sampling and Labeling
•Training dataset
•Validation dataset
•Test dataset
Learning algorithm
•Model selection
•Cross validation
•Performance metrics
•Hyperparameter Optimization
•Final Model
•Test dataset
•Labels
•Performance
•Final Model
•New data
•Labels
Data preprocess
Model fitting
Evaluation
Prediction

Introduction—Data Preprocessing

"objid" and "specobjid" are nothing but some numbers given by human

Have extremely high-redshift value objects: quasars.

objid	ra	dec	u	g	r	i	z	run	rerun	camcol	field	specobjid	class	redshift	plate	mjd	fiberid
1.24E+18	183.5313	0.089693	19.47406	17.0424	15.94699	15.50342	15.22531	752	301	4	267	3.72E+18	STAR	-8.96E-06	3306	54922	491
1.24E+18	183.5984	0.135285	18.6628	17.21449	16.67637	16.48922	16.3915	752	301	4	267	3.64E+17	STAR	-5.49E-05	323	51615	541
1.24E+18	183.6802	0.126185	19.38298	18.19169	17.47428	17.08732	16.80125	752	301	4	268	3.23E+17	GALAXY	0.123111	287	52023	513
1.24E+18	183.8705	0.049911	17.76536	16.60272	16.16116	15.98233	15.90438	752	301	4	269	3.72E+18	STAR	-0.00011	3306	54922	510
1.24E+18	183.8833	0.102557	17.55025	16.26342	16.43869	16.55492	16.61326	752	301	4	269	3.72E+18	STAR	0.00059	3306	54922	512
1.24E+18	183.8472	0.173694	19.43133	18.46779	18.16451	18.01475	18.04155	752	301	4	269	3.65E+17	STAR	0.000315	324	51666	594

The original dataset is a csv document with 10000 rows and 18 columns. One of the columns is the label and the others are features.

Dealing with missing data ✗

Getting categorical data into shape ✗

Over-sampling—SMOTE ✓

Feature selection ✓

Diving data into training set, and testing set

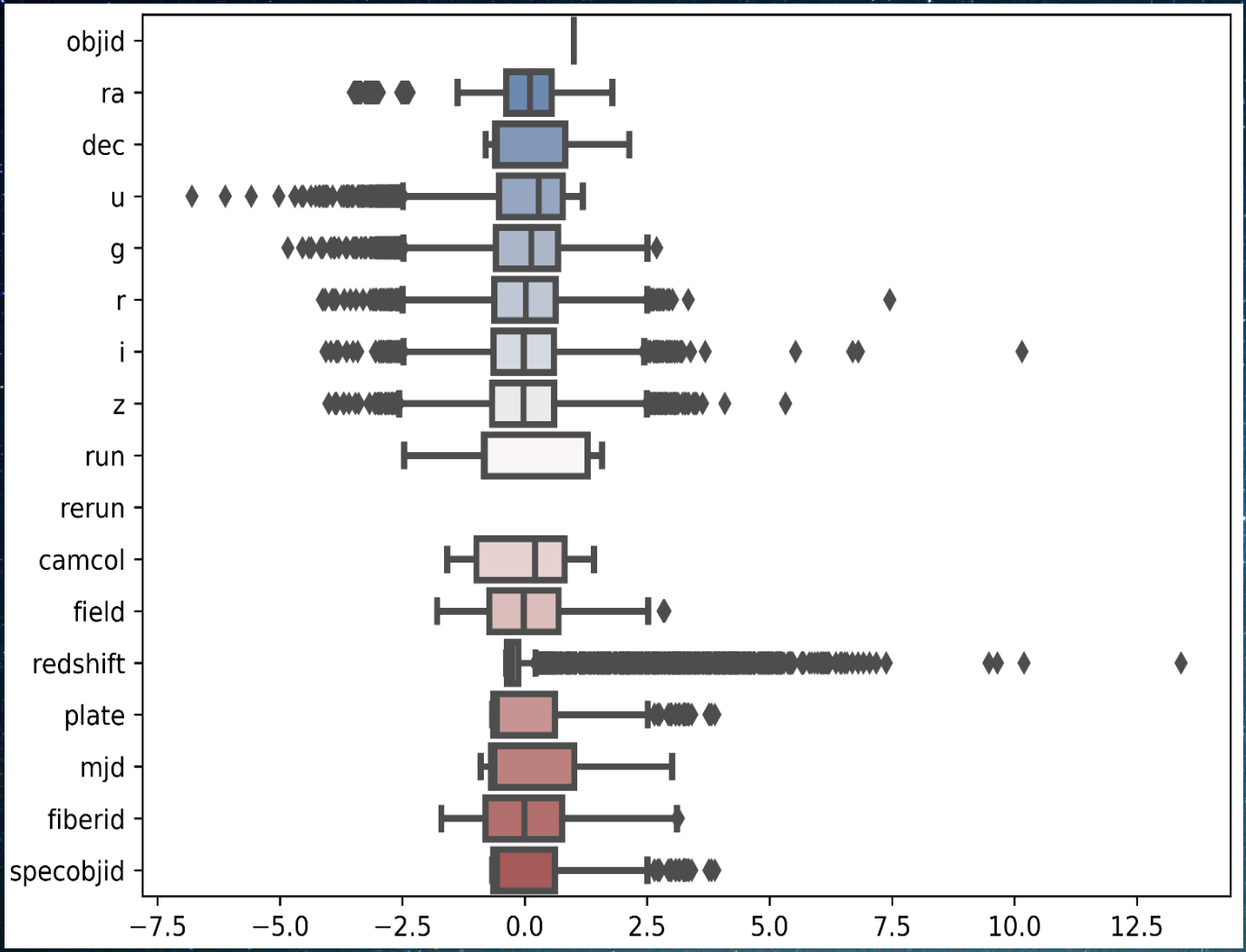
Introduction—Data Preprocessing

Synthetic Minority Oversampling Technique (SMOTE)

Imbalanced—oversampling (generate new samples in by interpolation)

	before	after
GALAXY	4998	4998
STAR	4152	4998
QSO	850	4998

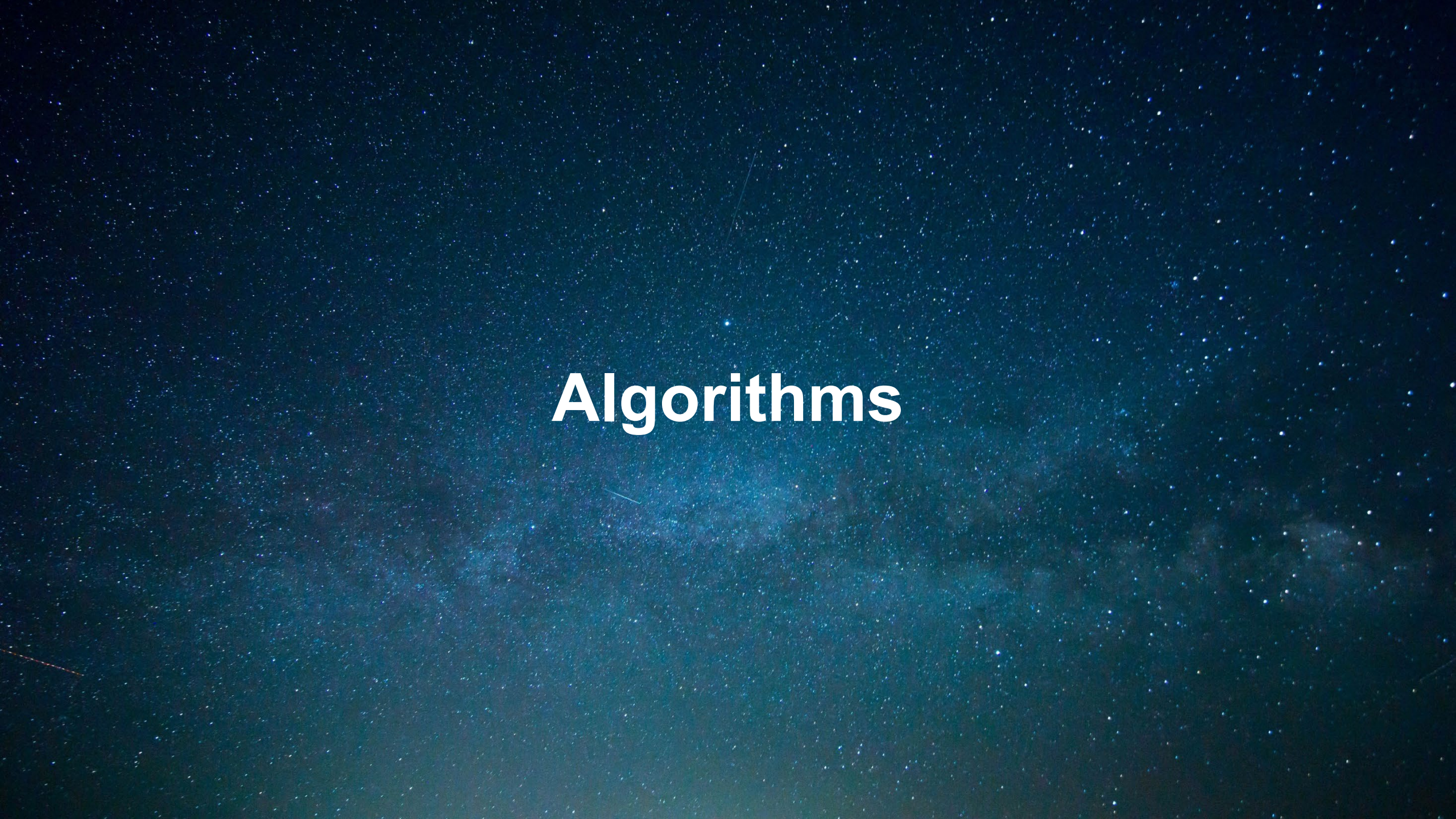
Introduction—Data Preprocessing



15 features
Thus we need to select features.

Boxplots of each features with their standardized values.

"rerun" is missing as its original value is just a constant.


Algorithms

Algorithms

01

KNN

02

Decision Tree

03

Random Forest & ExtraTree

04

ANN

Algorithms

KNN: Label a new object as the plurality label of its k nearest neighbors

Decision Tree: Fork Data into Subspaces

Random Forest: Decision Trees+Bagging+Random Feature Subset on trees

ExtraTrees: Random Features on each Node of Trees

ANN: Formed by Several Layers, Activation Functions and Parameters

Algorithms

Hyperparameters needed to be tuned:

- KNN: K
- Decision Tree: Max Depth
- Random Forest & ExtraTrees: Number of Trees
- ANN: Alpha & Number of Neurons in two different layers & Activation Function

Method:

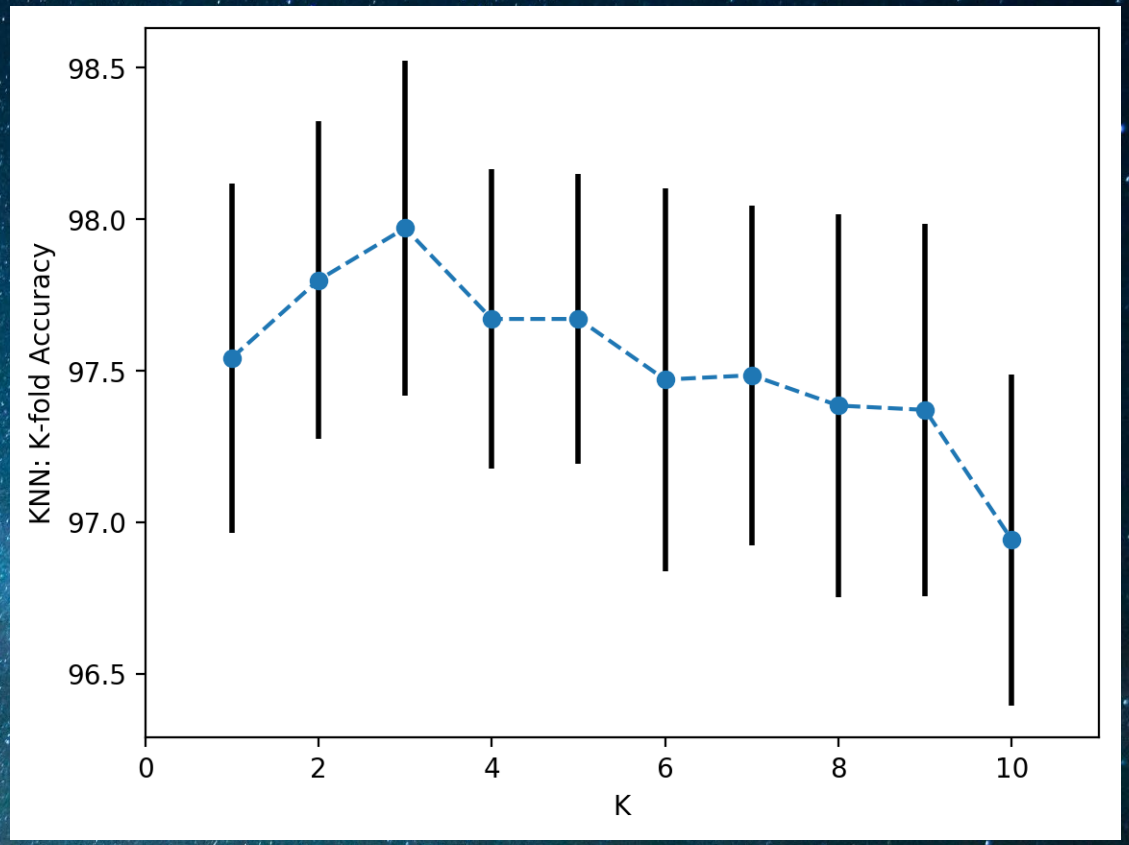
- Grid Search
- K-fold Validation on Each Choice of the Models

Algorithms — KNN – Model Fitting

Grid search with cross validation

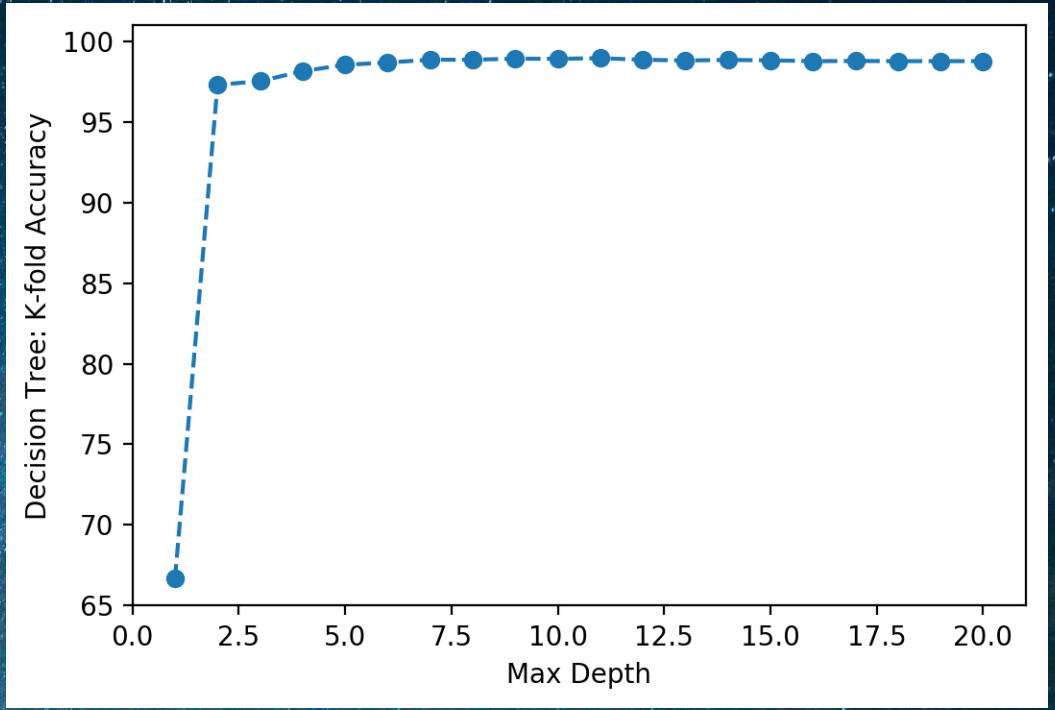
Set k(hyperparameter) from 1 to 10

3 is the best for validation

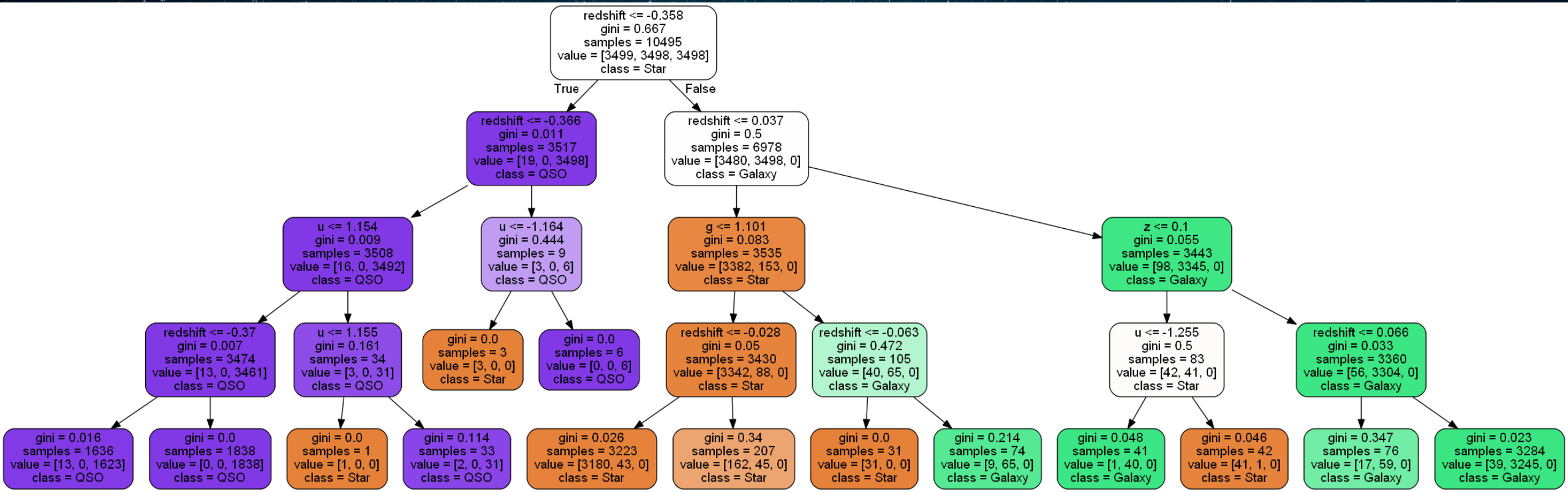


Algorithms —Decision Tree

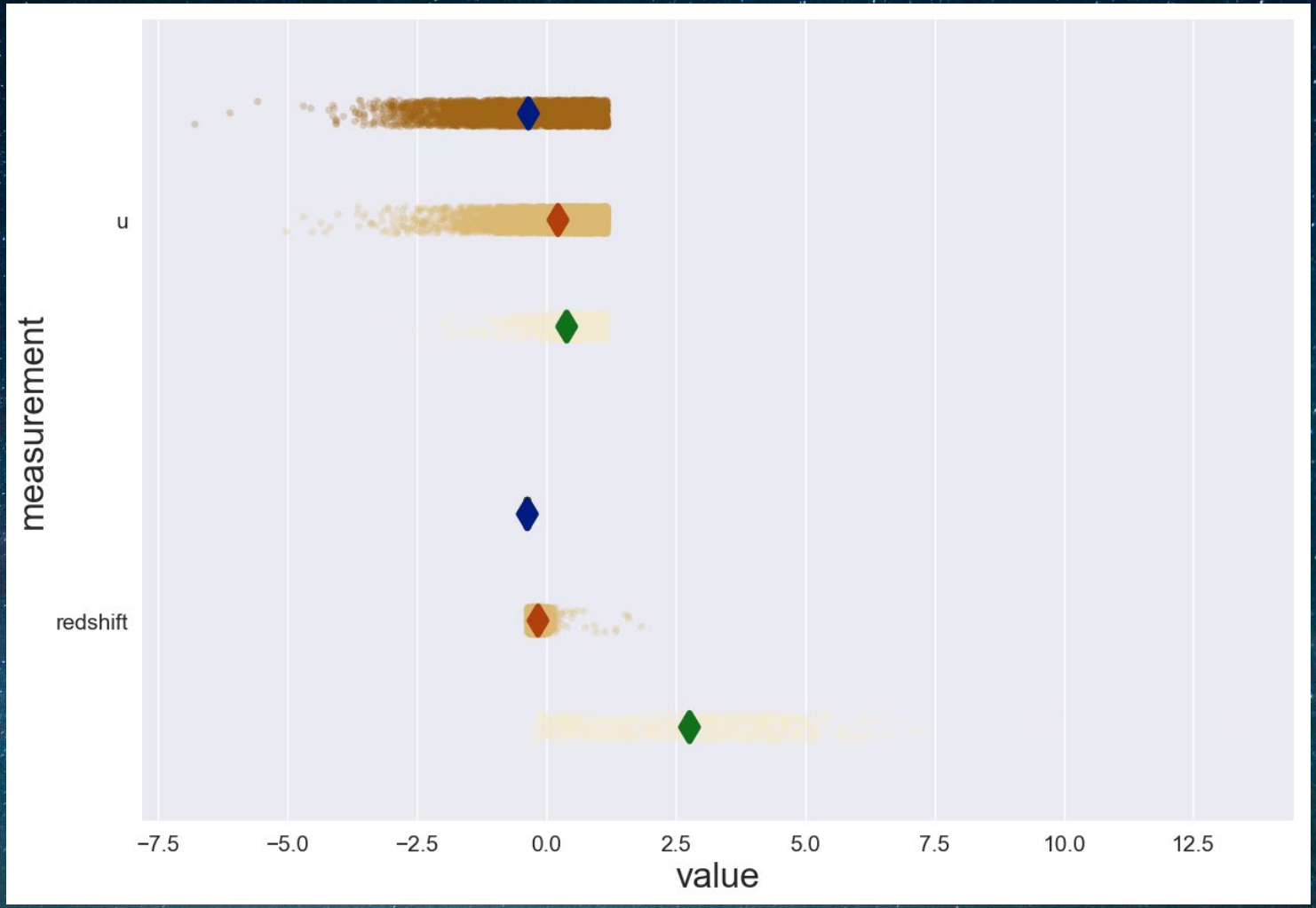
Max Depth=4



Algorithms —Decision Tree

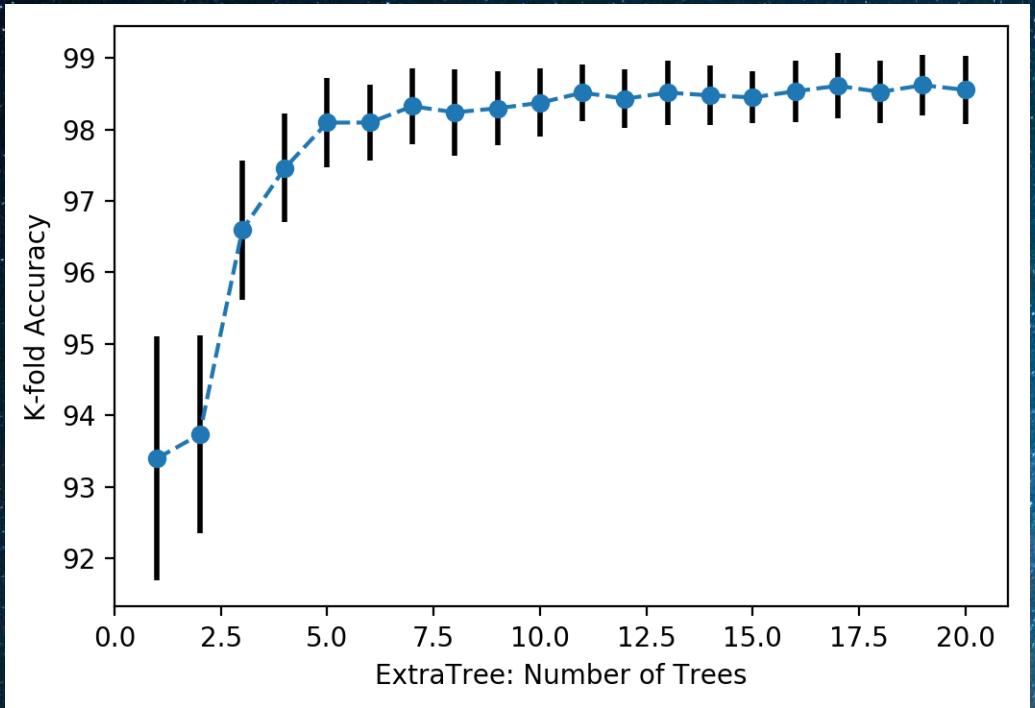


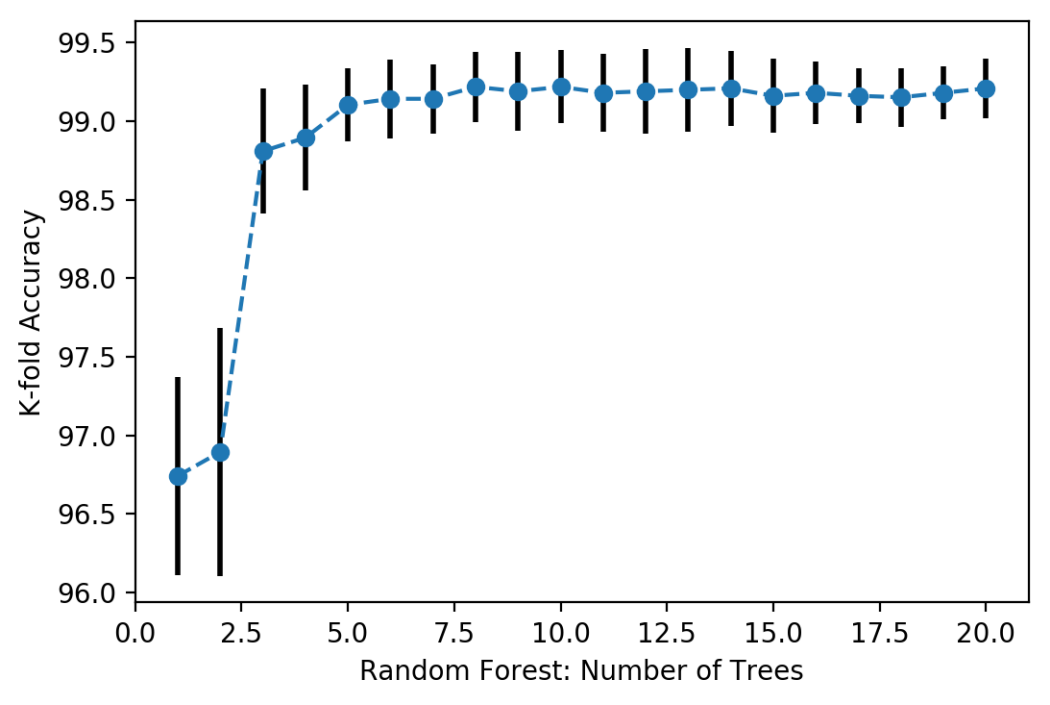
Algorithms —Decision Tree



Algorithms —Random Forest & Extra Tree

max_features="log2", criterion="gini"



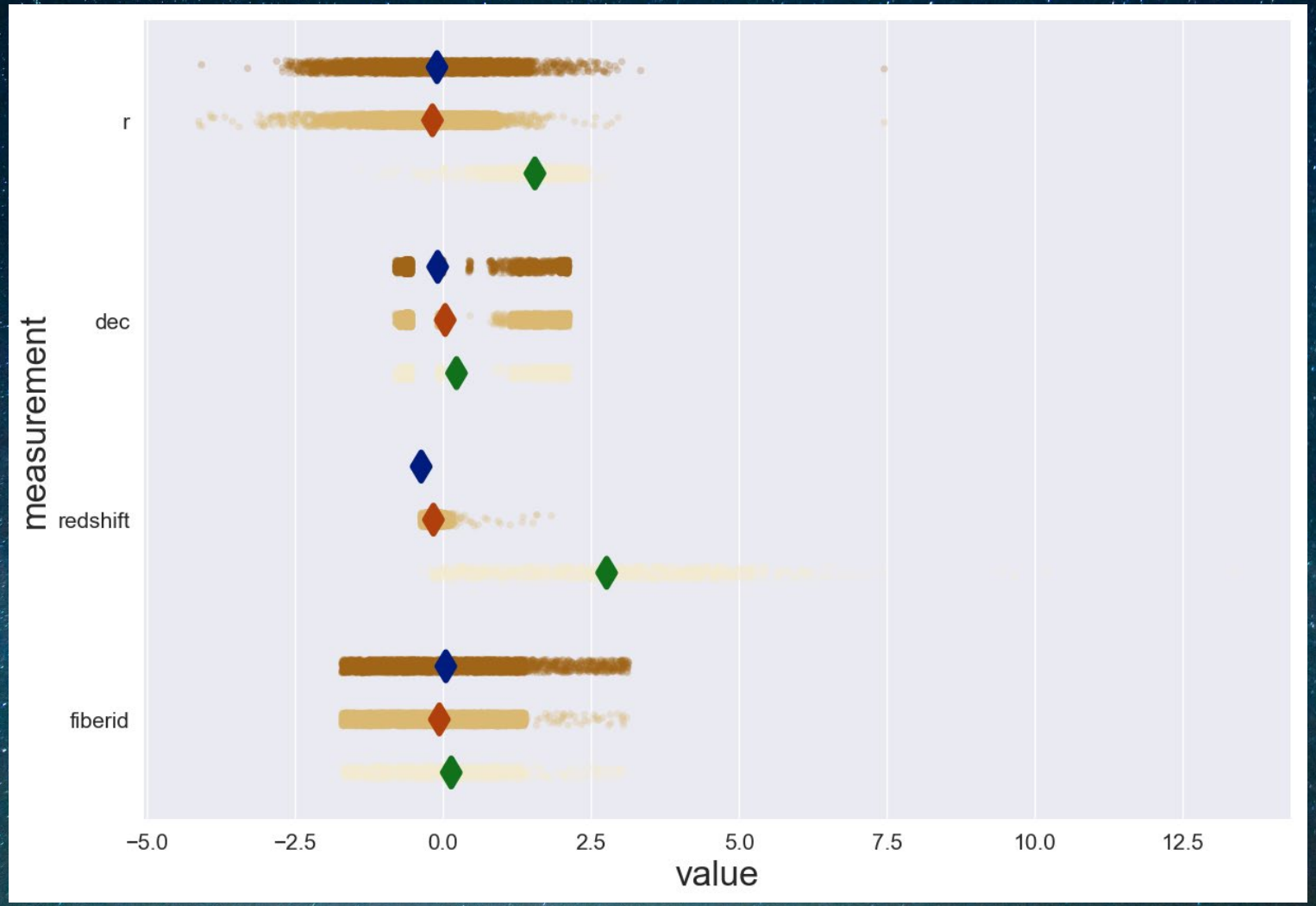


Number of Trees=10

First four feature importance: redshift fiberid dec r

Algorithms —Random Forest & Extra Tree

First four feature importance: redshift fiberid dec r



Algorithms —ANN

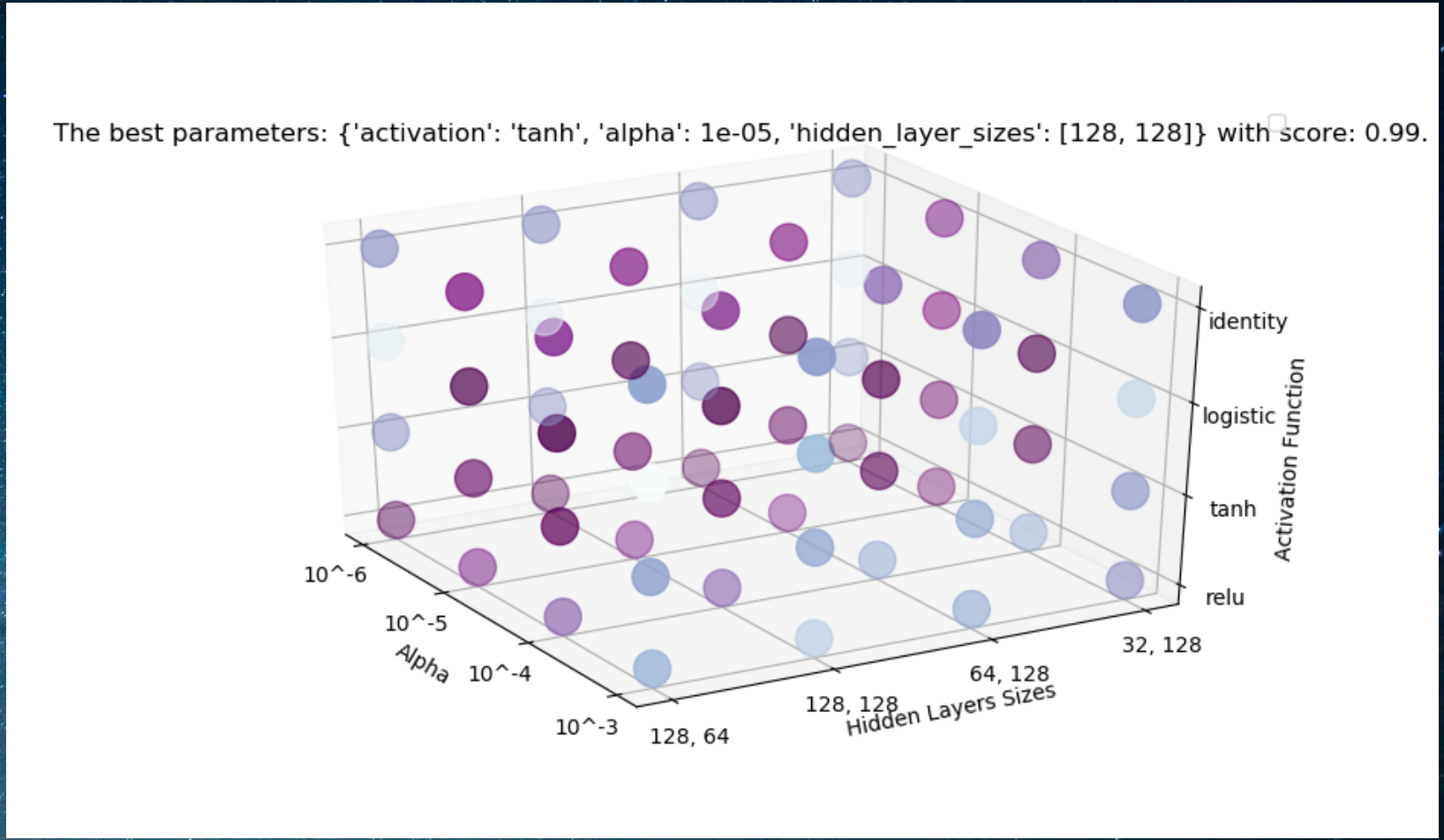
epsilon=1e-08
learning_rate_init=0.001
max_iter=2000
Solver="adam"
beta_1= 0.9
beta_2= 0.999

Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t.

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

Algorithms —ANN

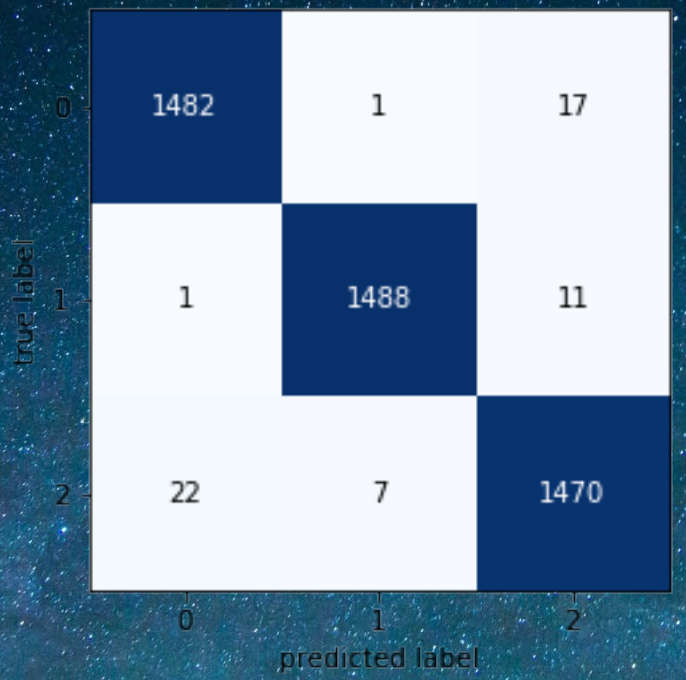
epsilon=1e-08
learning_rate_init=0.001
max_iter=2000
Solver=“adam”
beta_1= 0.9
beta_2= 0.999



Results and Conclusions

Result & Conclusion

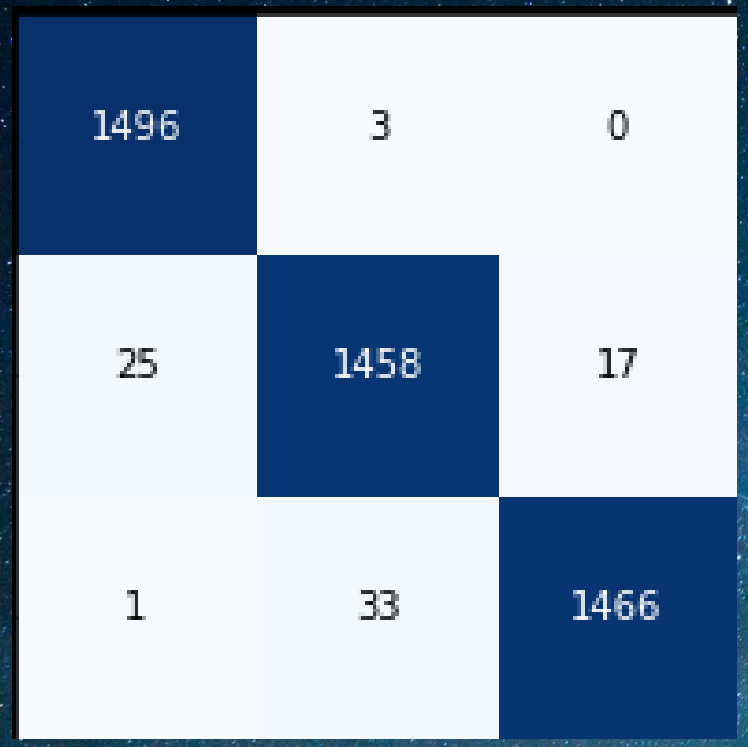
Decision tree



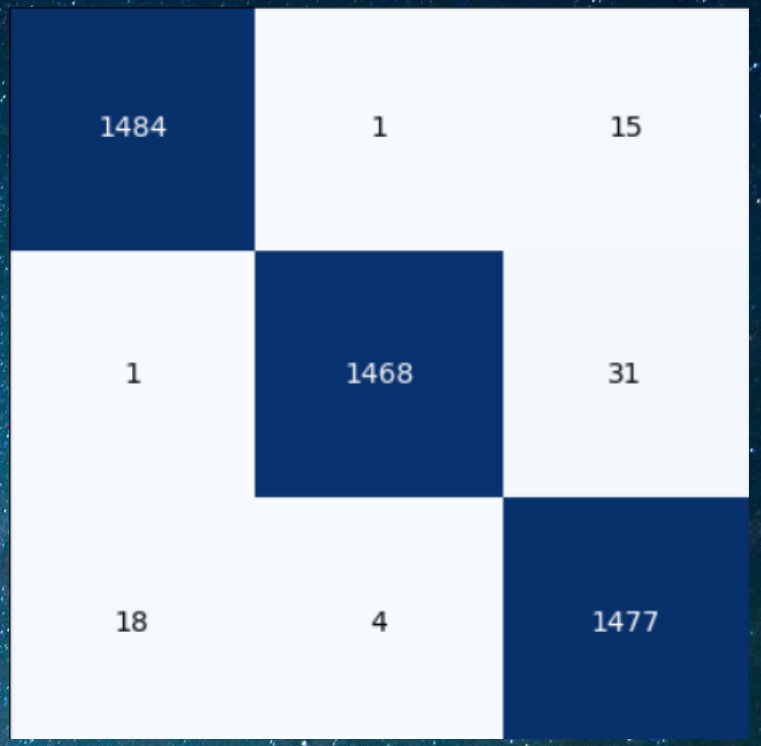
ExtraTree

Random Forest

Result & Conclusion



KNN



ANN

Result & Conclusion

Algorithms	Hyperparameters	Accuracy on training dataset	Accuracy on test dataset	Most important Features	Time cost	Choice
KNN	K = 3	98.4%	98.2%	dec, run, camcol and redshift	> 1 hour	
Decision Tree	Max depth = 4	98.8%	98.2%	/	5 seconds	
Extra Tree	Number of Trees = 10	100%	98.7%	redshift fiberid dec r	10 seconds	
Random Forest	Number of Trees = 10	99.9%	99.4%	redshift fiberid dec r	19 seconds	✓
ANN	alpha=10^{-5}, hidden layer size=[128,128]	99.0%	99%	/	12.7 minutes	

Future Research


THANK YOU