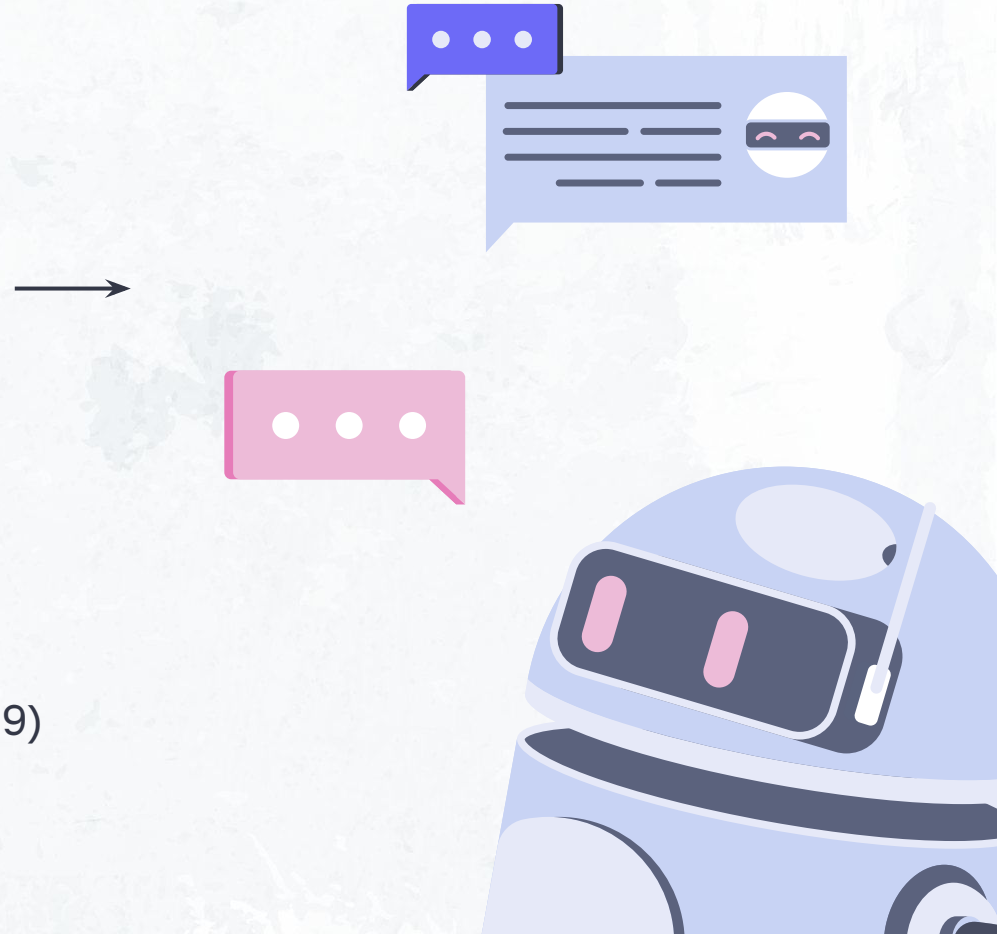
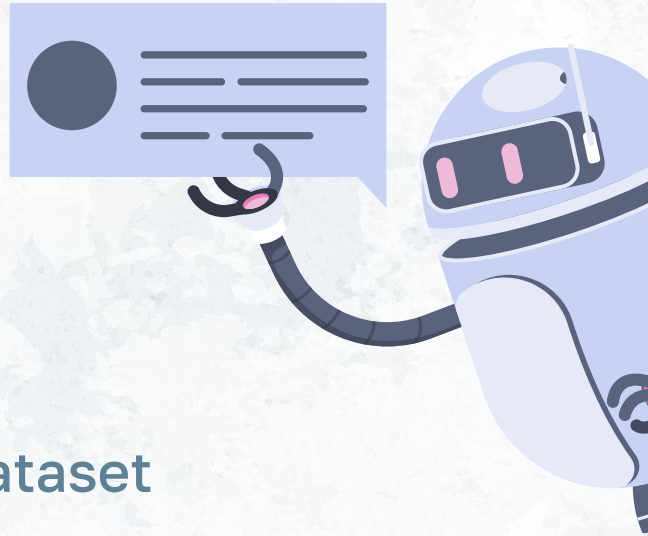
CS638 - Machine Learning (Case Study)



Nithish Kumar S (CB.SC.P2CSE24009)

Table of contents

01 →



Linear Regression

Dataset used : Laptop Price Prediction Dataset

Dataset Description

The dataset contains a diverse set of attributes, providing a holistic view of laptops from various manufacturers, models, and technical configurations. Each data entry comprises essential features that significantly influence laptop pricing, including:

Input Features : Company, TypeName,Inches, screen resolution, Cpu, Ram, Memory, GPU, OpSys, Weight

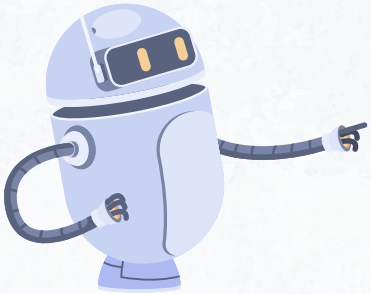
Output Feature : Price

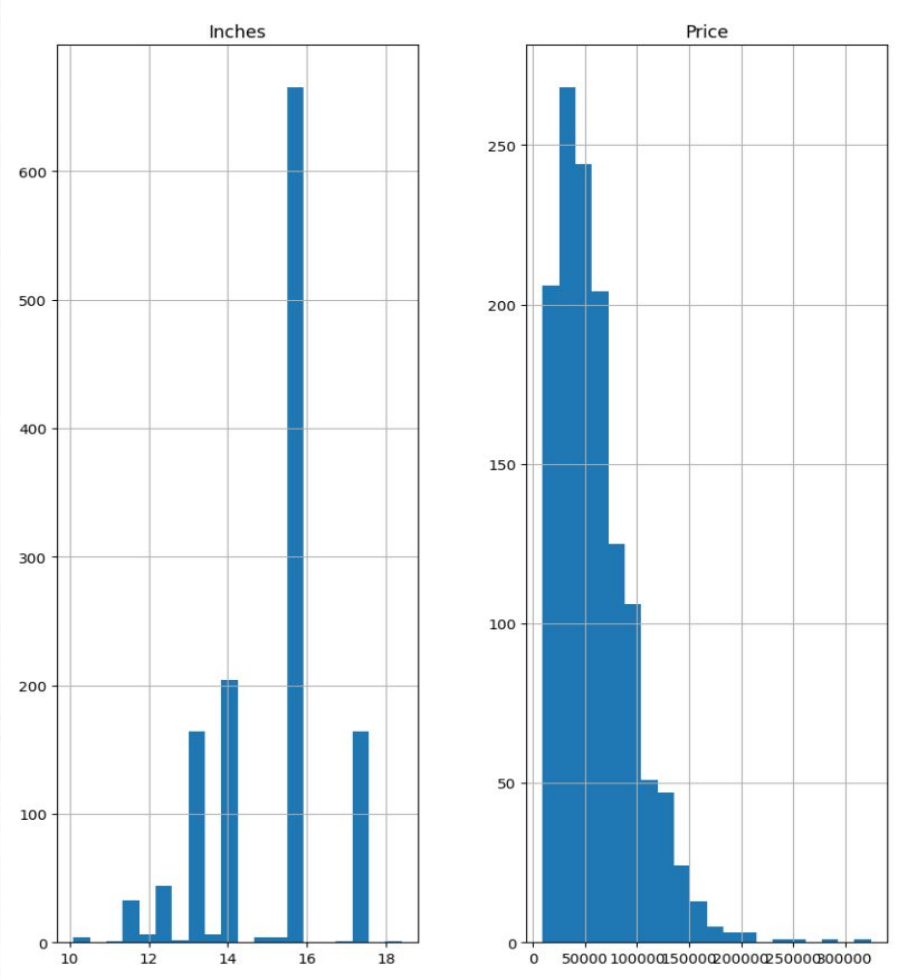
Dataset Description

Loading 5 data points from the dataset....

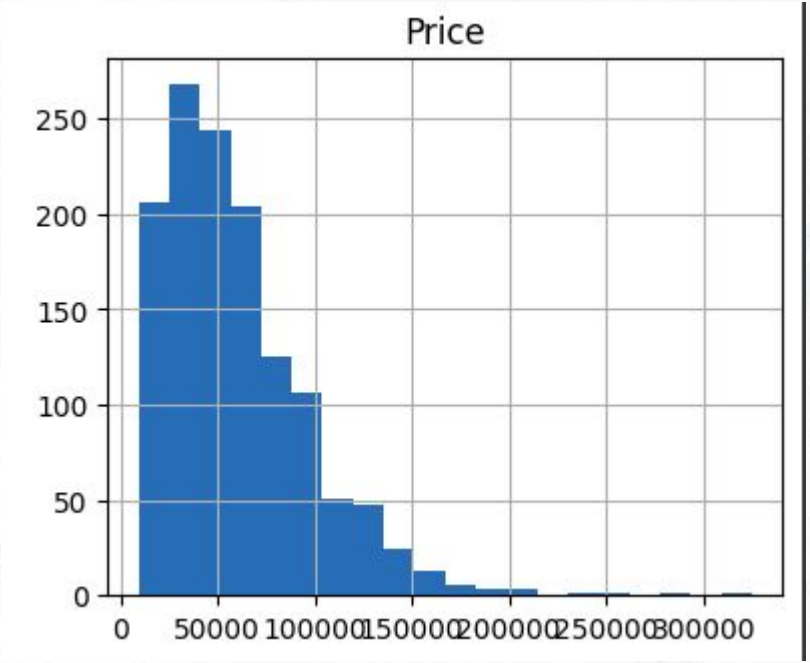
	Unnamed: 0	Company	TypeName	Inches	ScreenResolution	Cpu	Ram	Memory	Gpu	OpSys	Weight	Pric
0	0	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	128GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	71378.683
1	1	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	128GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	47895.523
2	2	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i5 7200U 2.5GHz	8GB	256GB SSD	Intel HD Graphics 620	No OS	1.86kg	30636.000
3	3	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.7GHz	16GB	512GB SSD	AMD Radeon Pro 455	macOS	1.83kg	135195.336
4	4	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 3.1GHz	8GB	256GB SSD	Intel Iris Plus Graphics 650	macOS	1.37kg	96095.808

Feature Distribution

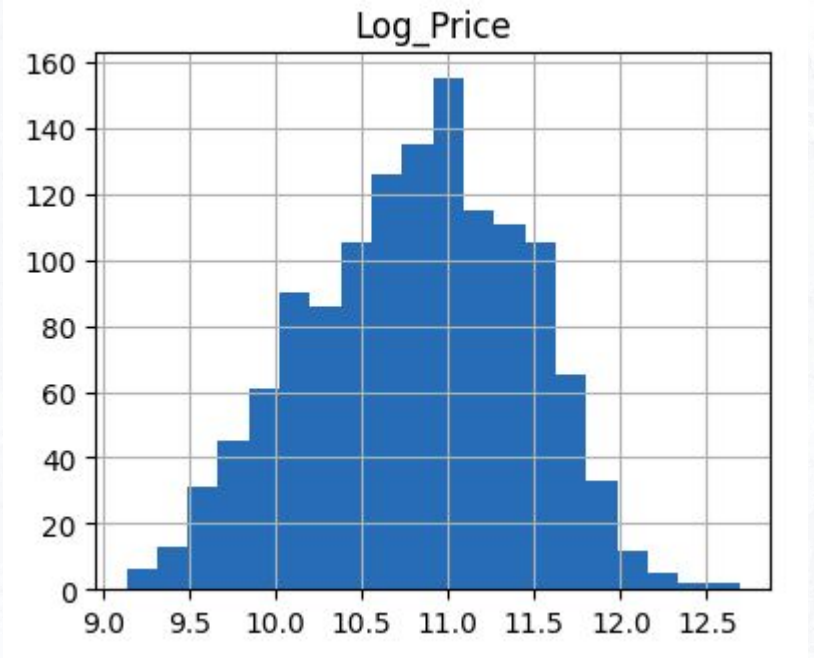




Feature Distribution after Log Transformation

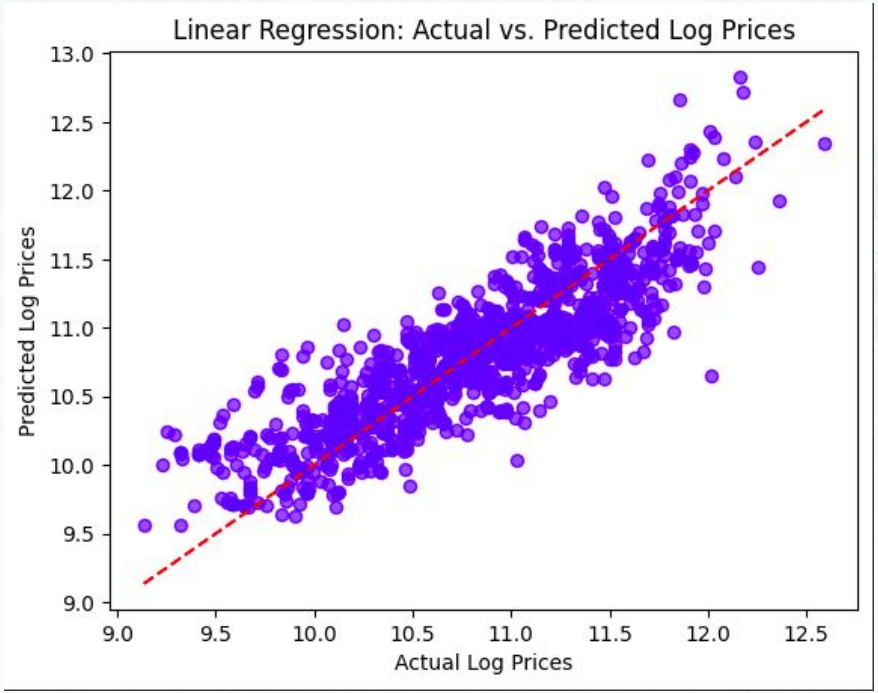


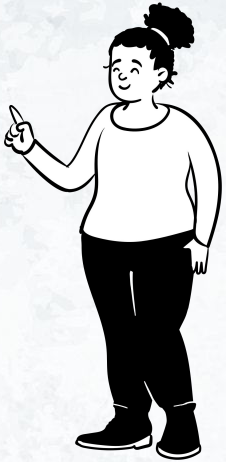
Before



After

Scattering Plot for Testing Data





Model Evaluation and Inference

Mean Squared Error : 0.109913126845355

R-squared : 0.6992435799146

02 →

Lasso and Ridge Regression

Dataset used : Laptop Price Prediction Dataset

Lasso and Ridge Regression

Lasso

Mean Squared Error : 0.11975677382913097

R-squared : 0.6912944333167976

Ridge

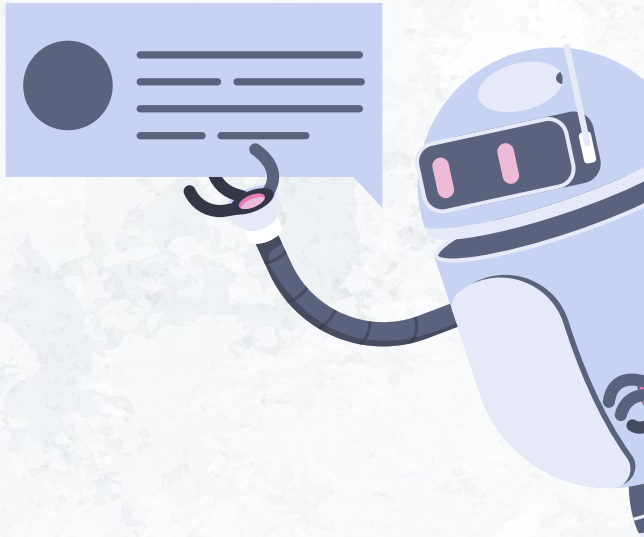
Mean Squared Error : 0.12228311851107884

R-squared : 0.698684188792644

03 →

Logistic Regression

Dataset used : Mushroom Dataset



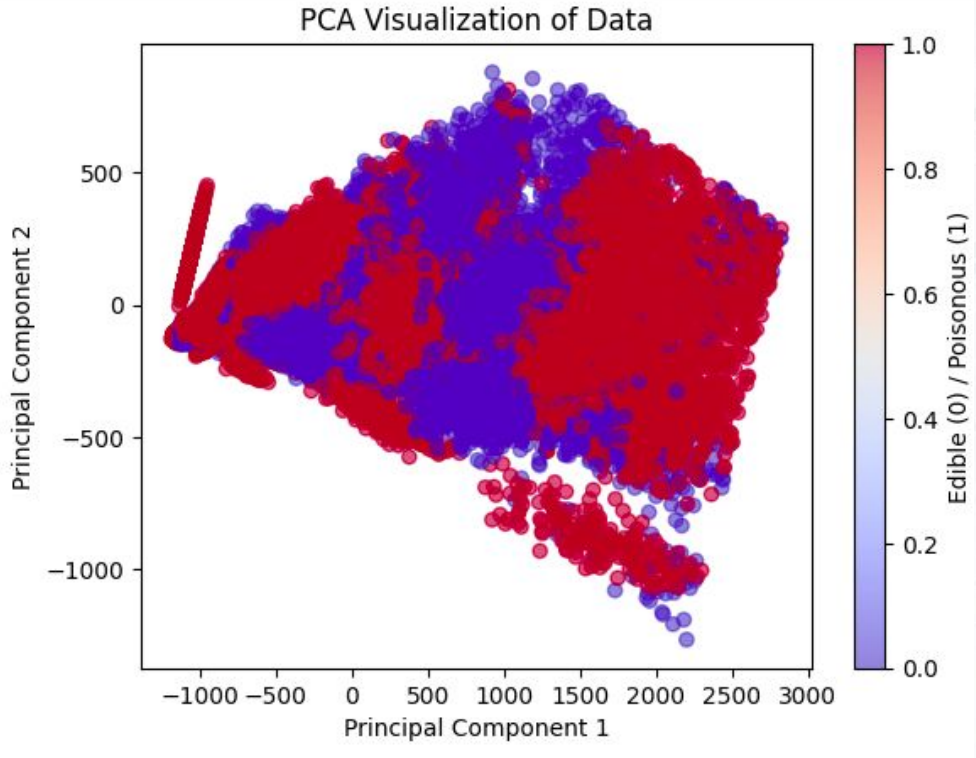
Dataset Description

This dataset is a cleaned version of the original Mushroom Dataset for Binary Classification Available at UCI Library. This dataset was
cleaned using various techniques such as Modal imputation, one-hot encoding, z-score normalization, and feature selection.

Input Features : Cap Diameter, Cap Shape, Gill Attachment, Gill Color, Stem Height, Stem Width, Stem Color, Season

Output Feature : Target Class

Is the Dataset Linear or NonLinear ?



Model Evaluation and Inference

```
Accuracy: 0.6394
Classification Report:
              precision    recall  f1-score   support

           0       0.58      0.72      0.64      4909
           1       0.71      0.58      0.64      5898

    accuracy                           0.64     10807
   macro avg       0.65      0.65      0.64     10807
weighted avg       0.65      0.64      0.64     10807

Confusion Matrix:
 [[3512 1397]
 [2500 3398]]
```

04 →

KNN

Dataset used : Stellar Classification Dataset - SDSS17

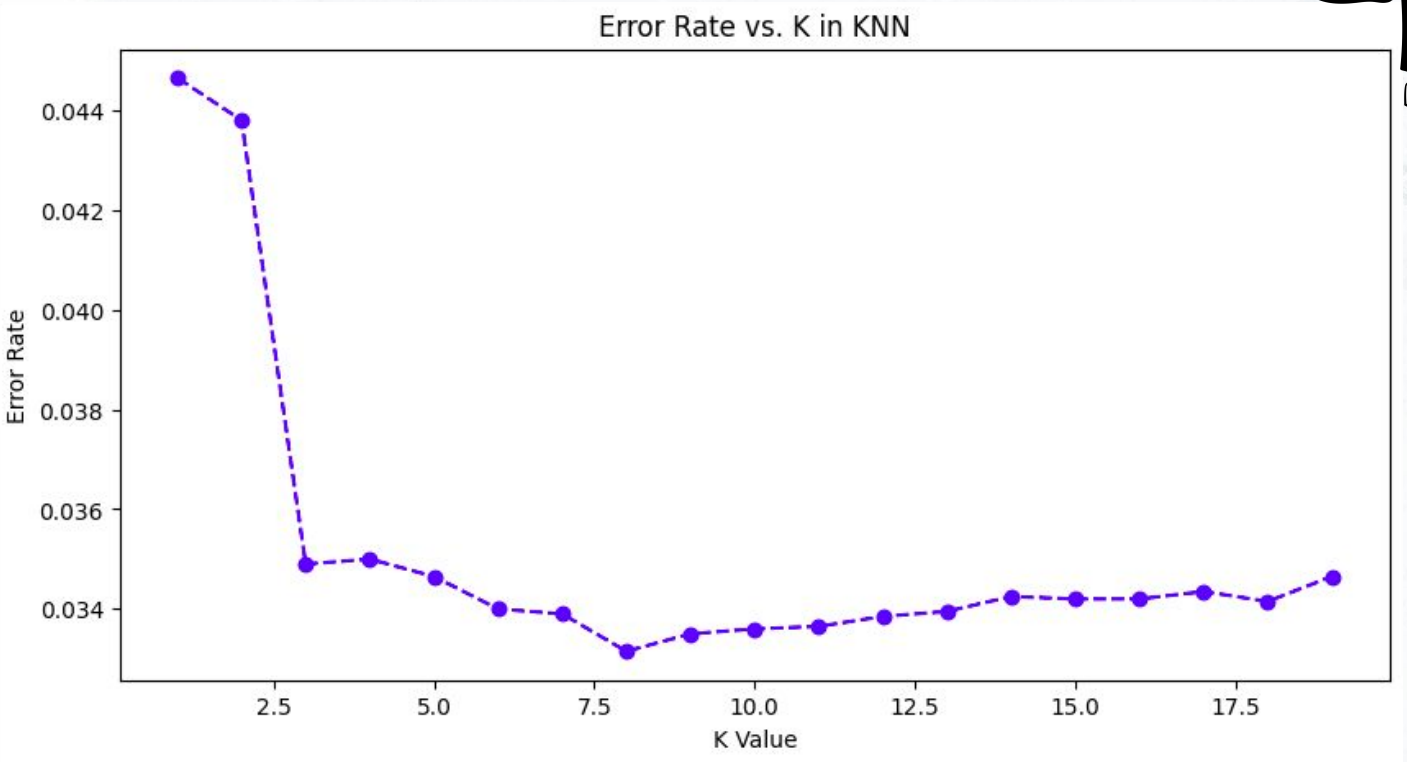
Dataset Description

Sloan Digital Sky Survey DR17
The data consists of 100,000 observations of space taken by the SDSS (Sloan Digital Sky Survey). Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.

Input Features : obj_ID, alpha, delta, u, g, r, i, z, run_ID, rereun_ID, cam_col, field_ID, spec_obj_ID, redshift, plate, MJD, fiber_ID

Output Feature : Class (Star or Galaxy or Quasar)

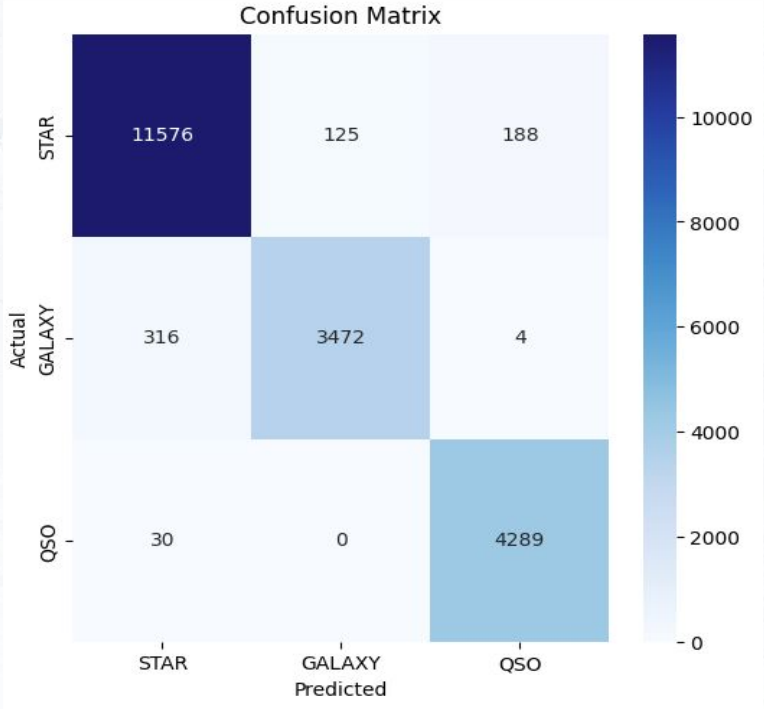
Choosing the best k-value



Model Evaluation and Inference

Accuracy: 96.69%
Classification Report:

	precision	recall	f1-score	support
GALAXY	0.97	0.97	0.97	11889
QSO	0.97	0.92	0.94	3792
STAR	0.96	0.99	0.97	4319
accuracy			0.97	20000
macro avg	0.96	0.96	0.96	20000
weighted avg	0.97	0.97	0.97	20000



Key Takeaways

Thank You!