REPORT TITLE: Star Type Classification

This project aims to build a machine learning model that can **accurately classify stars into their respective types** (Brown Dwarf, Red Dwarf, White Dwarf, Main Sequence, Supergiant, or Hypergiant) based on measurable attributes.

- (Red Dwarf = 0, Brown Dwarf = 1, White Dwarf = 2, Main Sequence = 3, Supergiant = 4, or Hypergiant = 5)

- *Data loading and importing Python libraries*
- *EDA findings and visualizations*
- *Physical significance of Visualizations*
- *Training of several models on the training dataset*
- *Comparison of all three models*

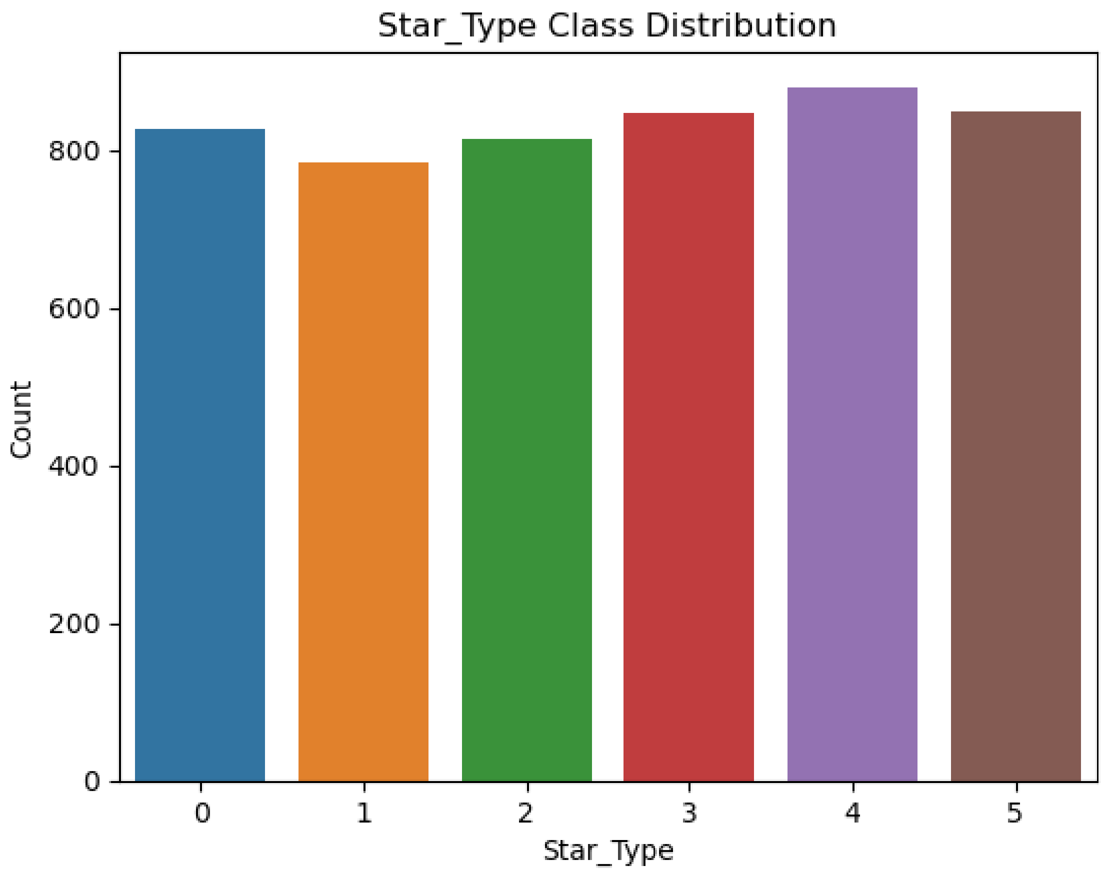
Aryan

21324004

BSMS Physics

Key Findings from EDA:

<u>Class Distribution in Star Type:</u> shows that class imbalance is minimum in the data set, or each class has the same data points as every other class.



Star Type

4 17.60%

5 17.00%

3 16.92%

0 16.52%

2 16.28%

1 15.68%

<u>Number of unique Spectral_Class values in df: 8</u>

Unique Spectral_Class values: ['A', 'D', 'G', 'F', 'K', 'M', 'O ', 'B', nan]

Spectral_Class value counts (percentage):

Spectral Class

M 22.865294%

B 17.179539%

D 16.588547%

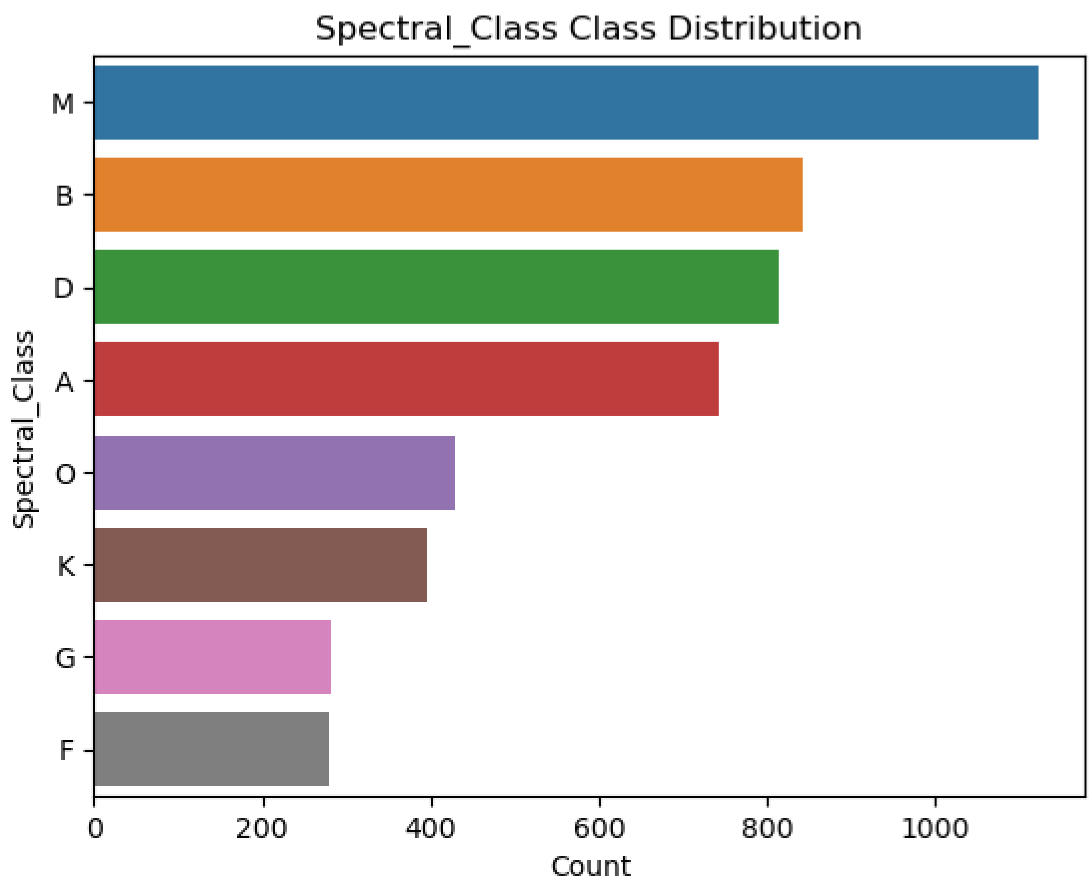
A 15.121255%

O 8.742613%

K 8.049725%

G 5.746892%

F 5.706134%



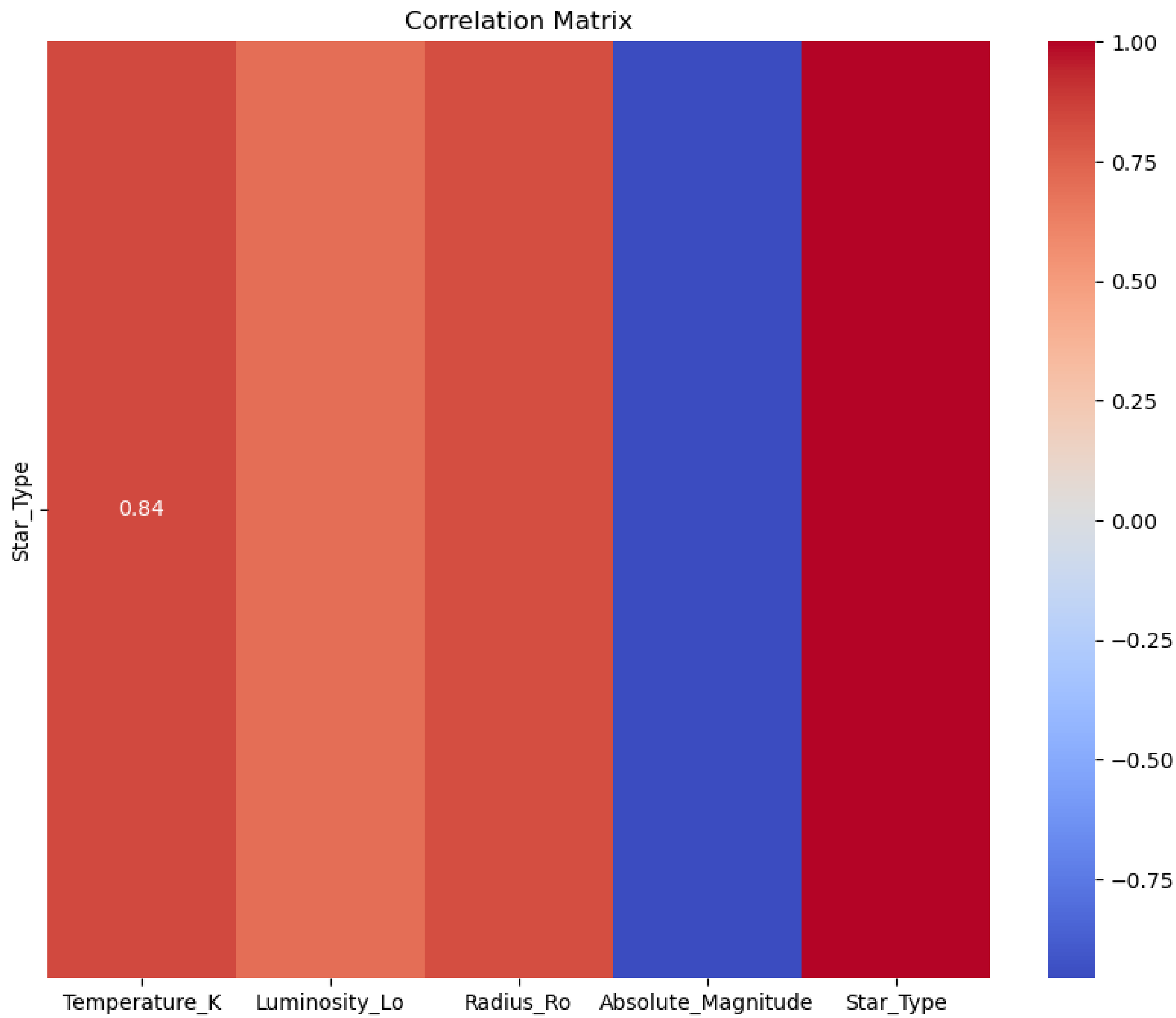
Correlation of Star_Type with other features:

Temperature_K 0.835004

Luminosity_Lo 0.697074

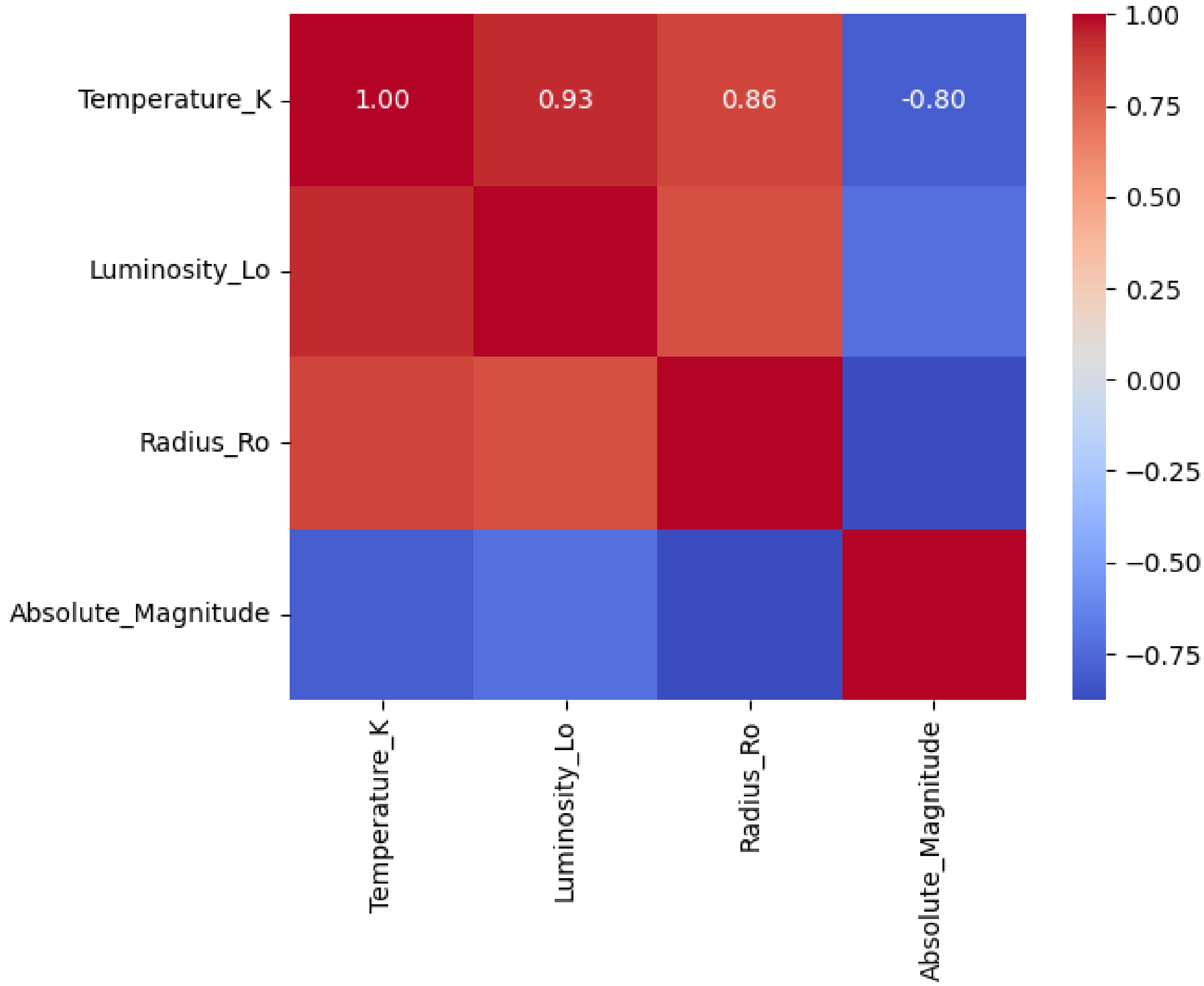
Radius_Ro 0.821730

Absolute_Magnitude -0.956318

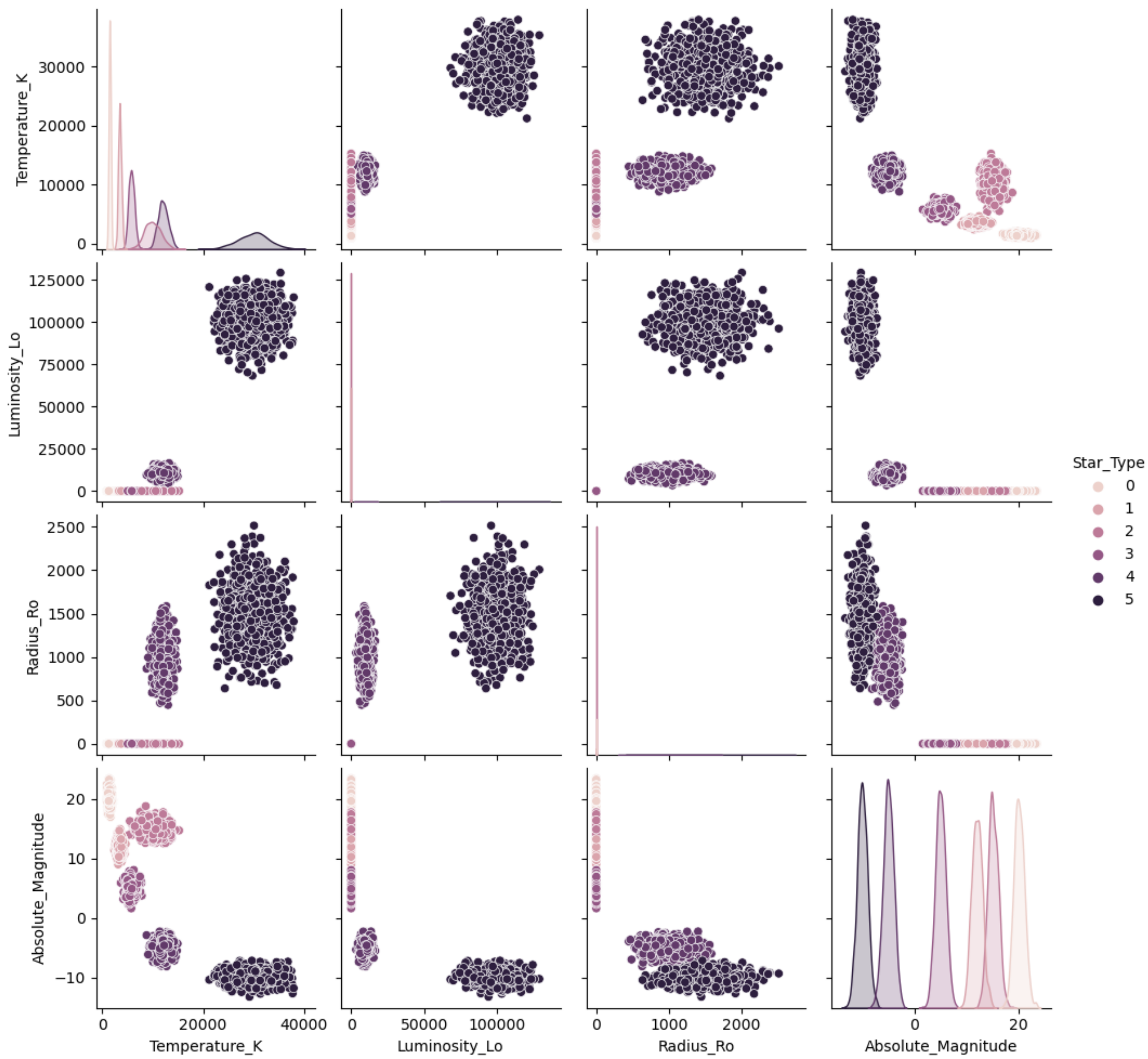


This show how Star_Type is related with all other numerical variables in a heatmap view.

The below heat map displays correlations among all other variables.

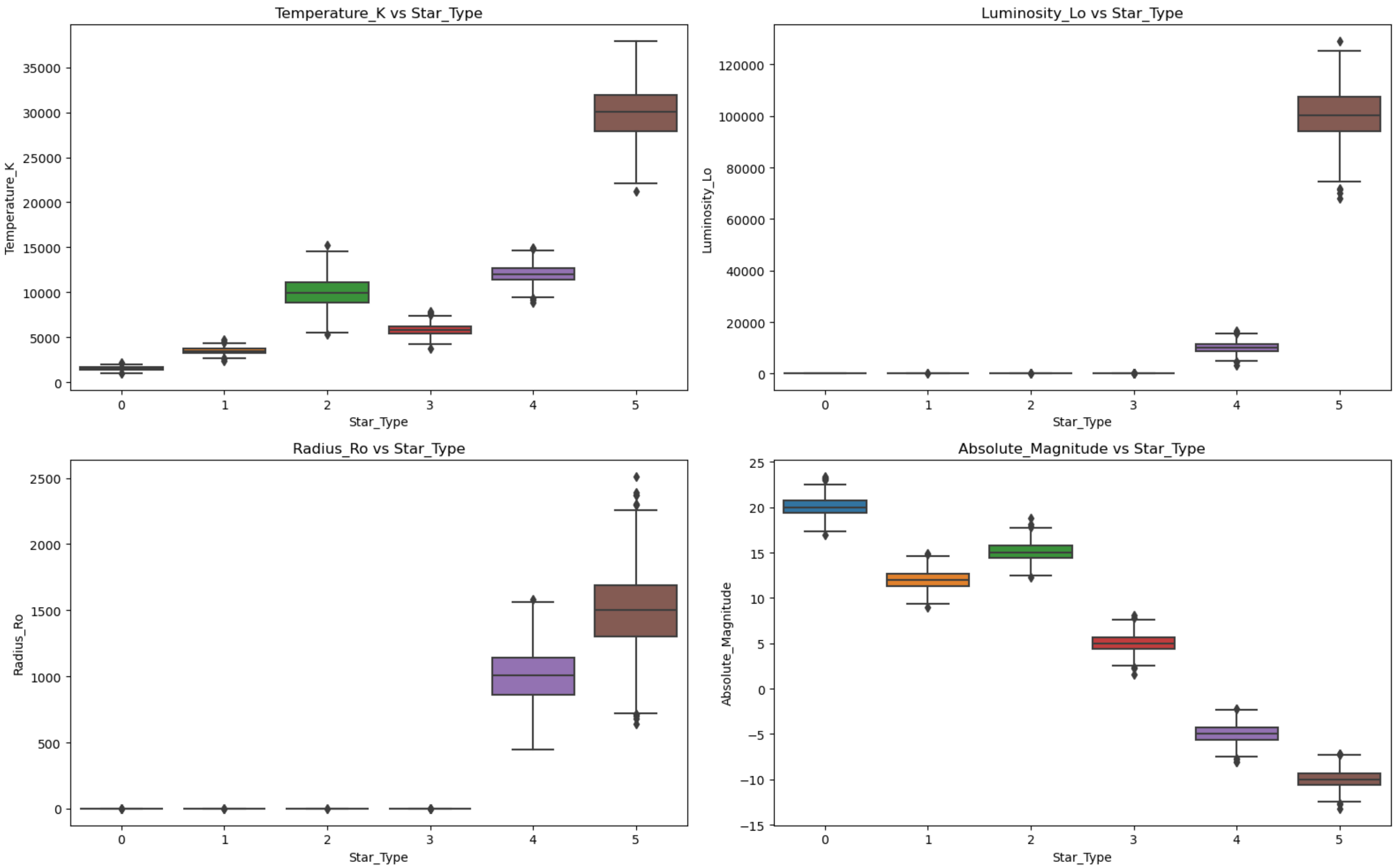

Temperature_K
Luminosity_Lo
Radius_Ro
Absolute_Magnitude
1.00
0.93
0.86
-0.80
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Temperature_K
Luminosity_Lo
Radius_Ro
Absolute_Magnitude

Outliers Visualisation using scatter plot:



outliers in the above data is minimum, hence only scaling will do the job.

Different Variables vs Star type:



Value ranges for Temperature_K by Star_Type:

Star_Type 0: min=945.43, Q1=1378.82, median=1508.67, Q3=1654.49, max=2117.46

Star_Type 1: min=2391.01, Q1=3270.78, median=3483.13, Q3=3704.44, max=4670.84

Star_Type 2: min=5325.02, Q1=8833.92, median=9915.24, Q3=11113.55, max=15243.82

Star_Type 3: min=3713.10, Q1=5382.32, median=5784.50, Q3=6200.75, max=7851.17

Star_Type 4: min=8804.38, Q1=11398.26, median=12016.64, Q3=12721.79, max=14938.19

Star_Type 5: min=21211.65, Q1=27896.67, median=30111.47, Q3=31946.15, max=37911.34

Value ranges for Luminosity_Lo by Star_Type:

Star_Type 0: min=0.00, Q1=0.00, median=0.00, Q3=0.00, max=0.00

Star_Type 1: min=0.01, Q1=0.03, median=0.04, Q3=0.05, max=0.07

Star_Type 2: min=-0.01, Q1=0.01, median=0.01, Q3=0.01, max=0.03

Star_Type 3: min=-0.78, Q1=0.70, median=0.98, Q3=1.35, max=2.93

Star_Type 4: min=3340.99, Q1=8742.54, median=10129.07, Q3=11443.07, max=16459.66

Star_Type 5: min=68121.88, Q1=94106.16, median=100295.94, Q3=107460.09, max=129029.42

Value ranges for Radius_Ro by Star_Type:

Star_Type 0: min=0.04, Q1=0.09, median=0.10, Q3=0.12, max=0.16

Star_Type 1: min=0.35, Q1=0.63, median=0.70, Q3=0.78, max=1.02

Star_Type 2: min=-0.00, Q1=0.01, median=0.01, Q3=0.01, max=0.02

Star_Type 3: min=-0.16, Q1=0.80, median=0.99, Q3=1.21, max=2.01

Star_Type 4: min=448.08, Q1=858.27, median=1009.92, Q3=1142.26, max=1585.45

Star_Type 5: min=638.56, Q1=1301.53, median=1504.54, Q3=1688.56, max=2513.22

Value ranges for Absolute_Magnitude by Star_Type:

Star_Type 0: min=16.95, Q1=19.36, median=20.02, Q3=20.72, max=23.41

Star_Type 1: min=8.99, Q1=11.28, median=11.97, Q3=12.67, max=14.94

Star_Type 2: min=12.32, Q1=14.40, median=14.99, Q3=15.75, max=18.78

Star_Type 3: min=1.61, Q1=4.37, median=5.00, Q3=5.67, max=8.06

Star_Type 4: min=-8.12, Q1=-5.61, median=-4.98, Q3=-4.30, max=-2.25

Star_Type 5: min=-13.23, Q1=-10.61, median=-9.99, Q3=-9.30, max=-7.16

Star Type	Temperature (K)	Luminosity (L⊙)	Radius (R⊙)	Absolute Magnitude	Key Features
Red Dwarf (0)	945 – 2,117	~0	0.037 – 0.158	16.95 – 23.41	Coolest, smallest, faintest, extremely long-lived
Brown Dwarf (1)	2,391 – 4,671	0.010 – 0.071	0.351 – 1.024	8.99 – 14.94	Substellar, not true stars, bridge between planets and stars
White Dwarf (2)	5,325 – 15,244	-0.005 – 0.026	~0.002 – 0.018	12.32 – 18.78	Hot, dense stellar remnants, small (Earth-sized), faint
Main Sequence (3)	3,713 – 7,851	-0.784 – 2.926	-0.157 – 2.012	1.61 – 8.06	Stable, hydrogen-burning phase, includes the Sun
Supergiant (4)	8,804 – 14,938	3,341 – 16,460	448 – 1,585	-8.12 – -2.26	Massive, very luminous, short-lived, often end as supernovae
Hypergiant (5)	21,212 – 37,911	68,122 – 129,029	639 – 2,513	-13.23 – -7.16	Extremely massive, hottest, brightest, rare, unstable, significant mass loss

Salient Features of Each Star Type

Red Dwarf (0)

Temperature: Very cool, ranging from approximately 945 K to 2,117 K.

Luminosity: Extremely faint, with luminosities close to zero.

Radius: Small, between 0.037 and 0.158 times the Sun's radius.

Absolute Magnitude: Very dim, with values from 16.95 to 23.41, making them some of the faintest stars.

Summary: Red dwarfs are the smallest, coolest, and longest-lived stars, often too faint to be seen with the naked eye.

Brown Dwarf (1)

Temperature: Cool, between 2,391 K and 4,671 K.

Luminosity: Very low, from 0.010 to 0.071 times the Sun's luminosity.

Radius: Small, between 0.351 and 1.024 solar radii.

Absolute Magnitude: Dim, with values from 8.99 to 14.94.

Summary: Brown dwarfs are substellar objects, not massive enough to sustain hydrogen fusion, bridging the gap between the largest planets and the smallest stars.

White Dwarf (2)

Temperature: Hot, from 5,325 K to 15,244 K.

Luminosity: Very low, can even be negative, indicating they are much dimmer than the Sun.

Radius: Extremely small, from -0.002 to 0.018 solar radii (the negative value is likely a data artifact; white dwarfs are about Earth-sized).

Absolute Magnitude: Moderately dim, from 12.32 to 18.78.

Summary: White dwarfs are dense stellar remnants, the final evolutionary state for most stars, with high temperatures but very low luminosity due to their small size.

Main Sequence (3)

Temperature: Moderate, from 3,713 K to 7,851 K.

Luminosity: Ranges from slightly less than the Sun to almost three times as luminous.

Radius: From slightly smaller to about twice the Sun's radius.

Absolute Magnitude: Bright, between 1.61 and 8.06.

Summary: Main sequence stars, including the Sun, are in the stable, hydrogen-burning phase of their lives.

Supergiant (4)

Temperature: Hot, from 8,804 K to 14,938 K.

Luminosity: Extremely bright, from 3,341 to 16,460 times the Sun's luminosity.

Radius: Very large, from 448 to 1,585 solar radii.

Absolute Magnitude: Exceptionally bright, from -8.12 to -2.26.

Summary: Supergiants are massive, luminous stars nearing the end of their lives, often leading to spectacular supernovae.

Hypergiant (5)

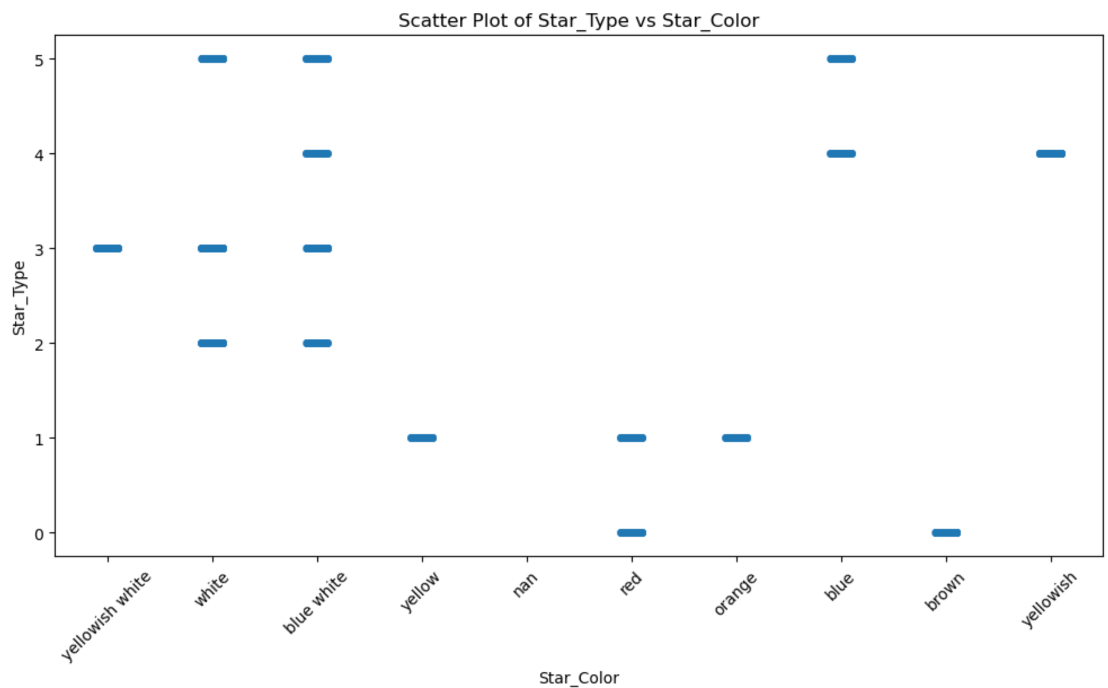
Temperature: Extremely hot, from 21,212 K to 37,911 K.

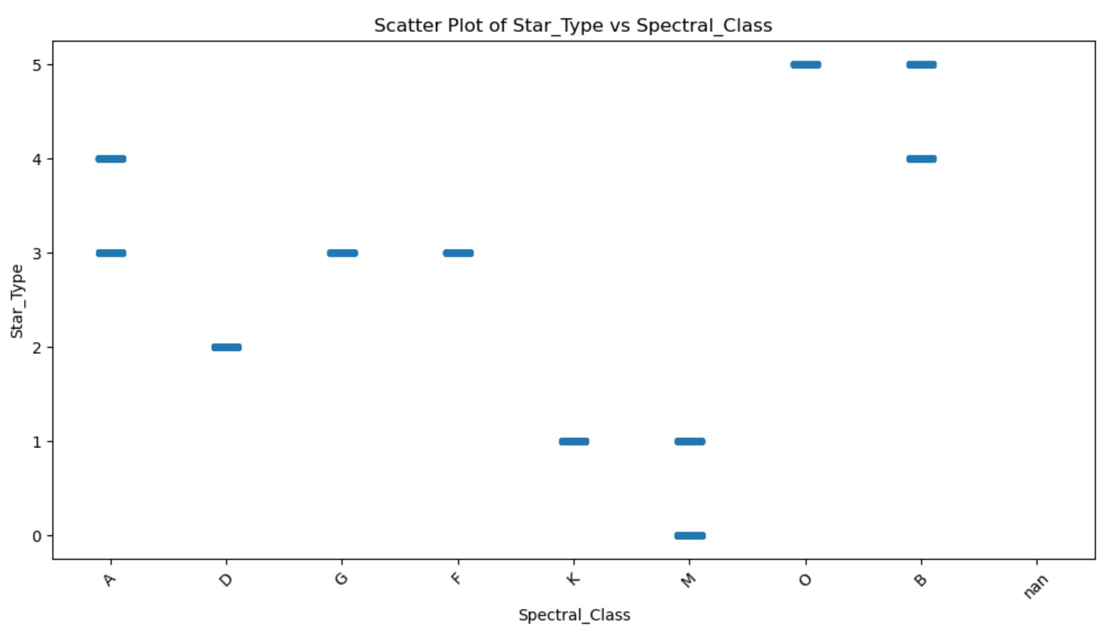
Luminosity: Among the brightest, from 68,122 to 129,029 times the Sun's luminosity.

Radius: Enormous, from 639 to 2,513 solar radii.

Absolute Magnitude: Incredibly bright, from -13.23 to -7.16.

Summary: Hypergiants are rare, extremely massive stars with immense luminosity and size, prone to instability and significant mass loss.





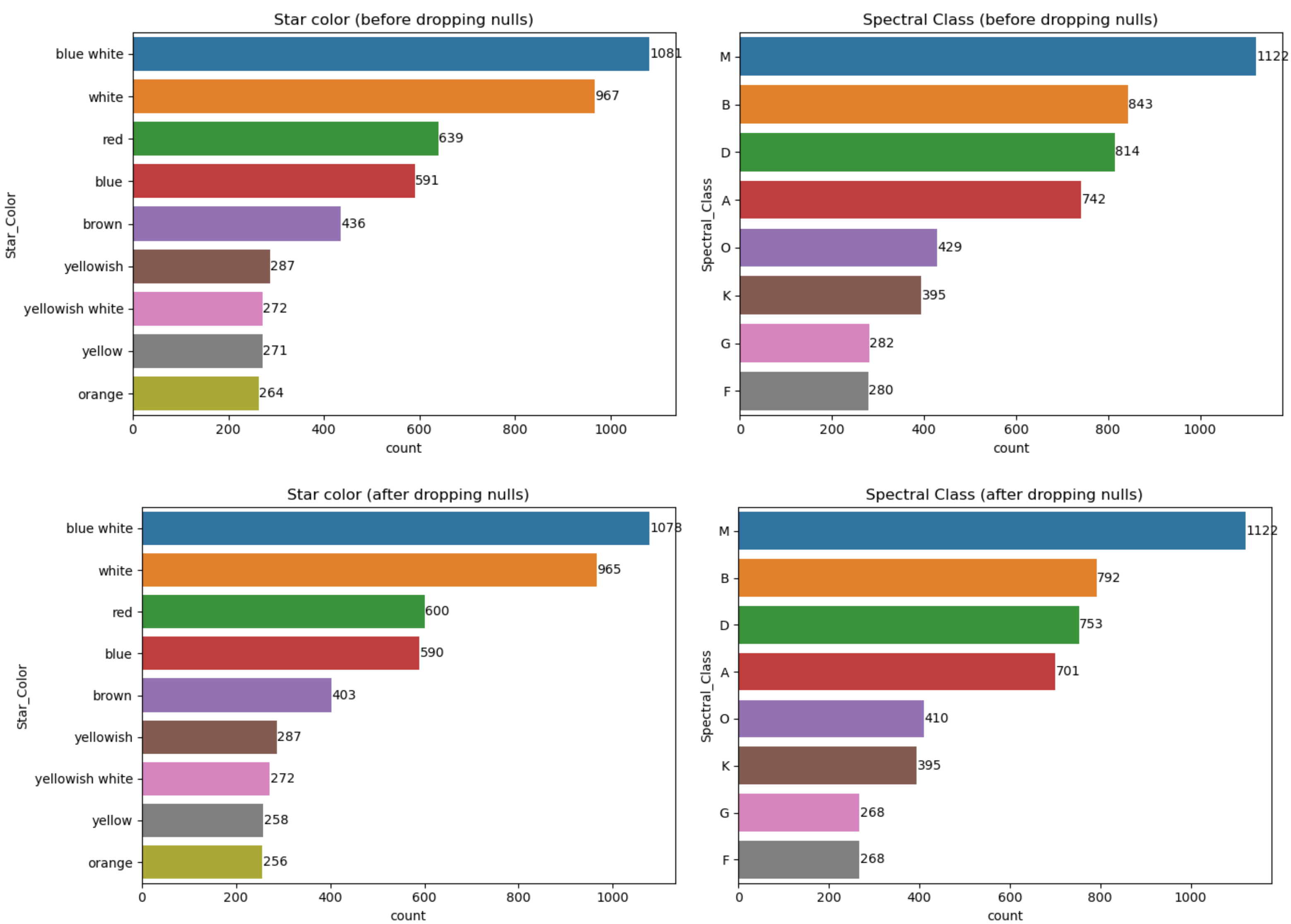
Handling Missing data:

Temperature_K	0.14%
Luminosity_Lo	0.00%
Radius_Ro	0.00%
Absolute_Magnitude	0.00%
Star_Color	3.84%
Spectral_Class	1.86%
Star_Type	0.00%

```
cols=df.columns
len(df[cols].dropna())/len(df)
```

The above code will show us that even after dropping the missing value columns the remaing data left will be 94.18%.

As the missing values in each column is <5%, to preserve accuracy , the columns are simply dropped and remaining execution is done on the remaining 94.18% data.

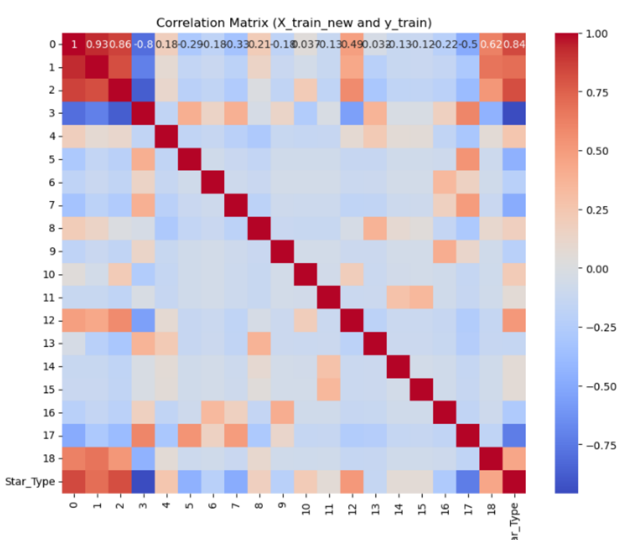


Even after deleting the missing values, the distribution remains identical, which indicates that the missing values were purely at random.

Handling the Categorical values using one-hot encoding

```
cat_cols = ['Star_Color', 'Spectral_Class']
trf2= ColumnTransformer(transformers=[
    ('scale', StandardScaler(),[col for col in X_train.columns if col not in
cat_cols]),
    ('encode',OneHotEncoder(sparse=False,drop='first',handle_unknown='ignore'),ca
t_cols),
],remainder='passthrough')
```

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0.235165	-0.250169	0.755518	-0.931446	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.160372	-0.478352	-0.652989	0.777245	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
0.219380	-0.345748	1.015685	-1.031322	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
-0.407752	-0.478324	-0.650718	-0.054619	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0.121996	-0.223995	0.778990	-1.073657	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
-0.696069	-0.478351	-0.652073	0.500990	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
-0.438222	-0.478341	-0.652295	-0.204527	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
2.334846	2.448255	1.942139	-1.662307	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0
-0.655244	-0.478351	-0.651783	0.444950	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
-0.407194	-0.478320	-0.651911	-0.134782	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

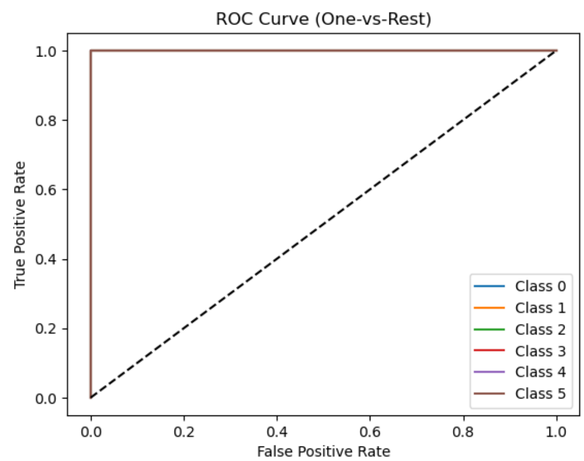


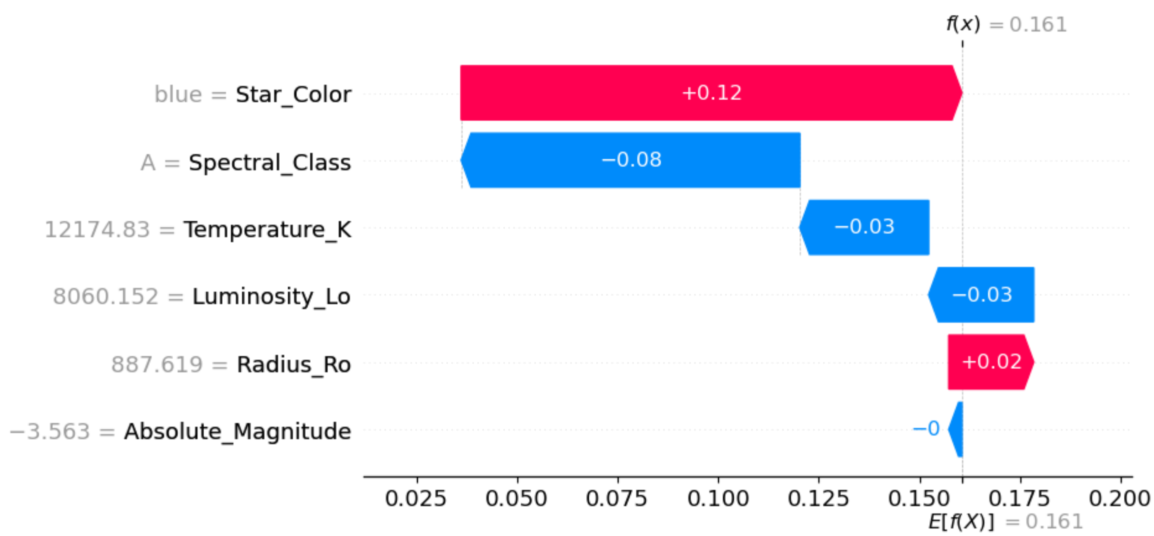
: Heatmap shows correlation between columns after encoding to ensure variable dependencies on each other

Training different models and explaining them using plots.

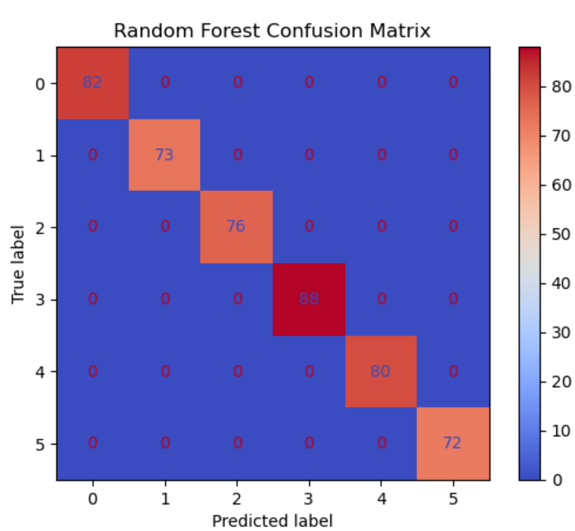
Random Forest Pipeline Metrics:

Precision (macro): 1.0
Accuracy: 1.0
Recall (macro): 1.0
F1 Score (macro): 1.0
F2 Score (macro): 1.0
Confusion Matrix:
[[82 0 0 0 0 0]
[0 73 0 0 0 0]
[0 0 76 0 0 0]
[0 0 0 88 0 0]
[0 0 0 0 80 0]
[0 0 0 0 0 72]]
AUC-ROC (macro, OVR): 1.0





A SHAP waterfall plot is a visualization designed to explain the prediction of a machine learning model for a single data point by showing how each feature contributes to the final model output.

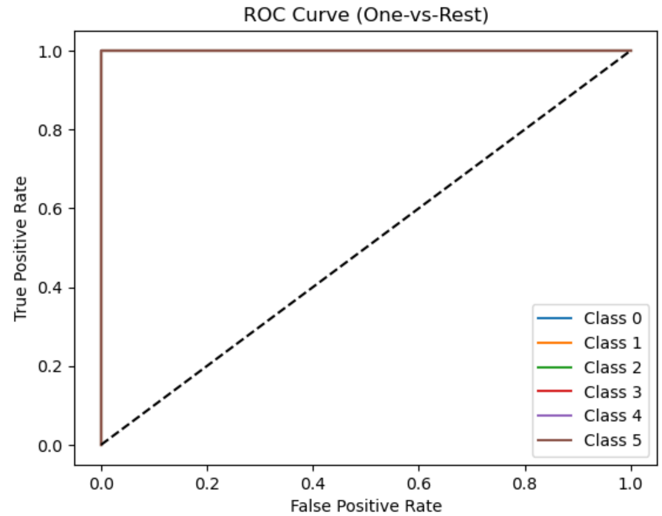


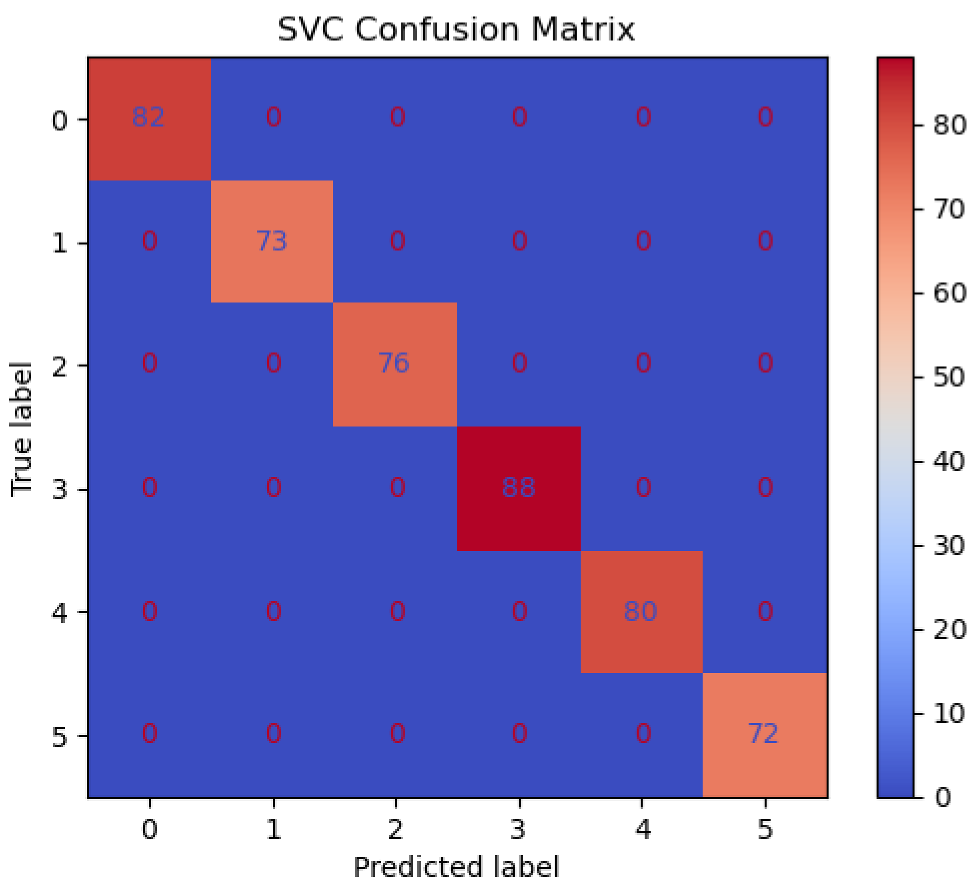
SVC Pipeline Metrics:

SVC Pipeline Metrics:
Precision (macro): 1.0
Accuracy: 1.0
Recall (macro): 1.0
F1 Score (macro): 1.0
F2 Score (macro): 1.0
Confusion Matrix:

```
[[82  0  0  0  0  0]
 [ 0 73  0  0  0  0]
 [ 0  0 76  0  0  0]
 [ 0  0  0 88  0  0]
 [ 0  0  0  0 80  0]
 [ 0  0  0  0  0 72]]
```

AUC-ROC (macro, OVR): 1.0





K-Nearest Neighbors Pipeline Metrics:

KNN Pipeline Metrics:

Accuracy: 1.0

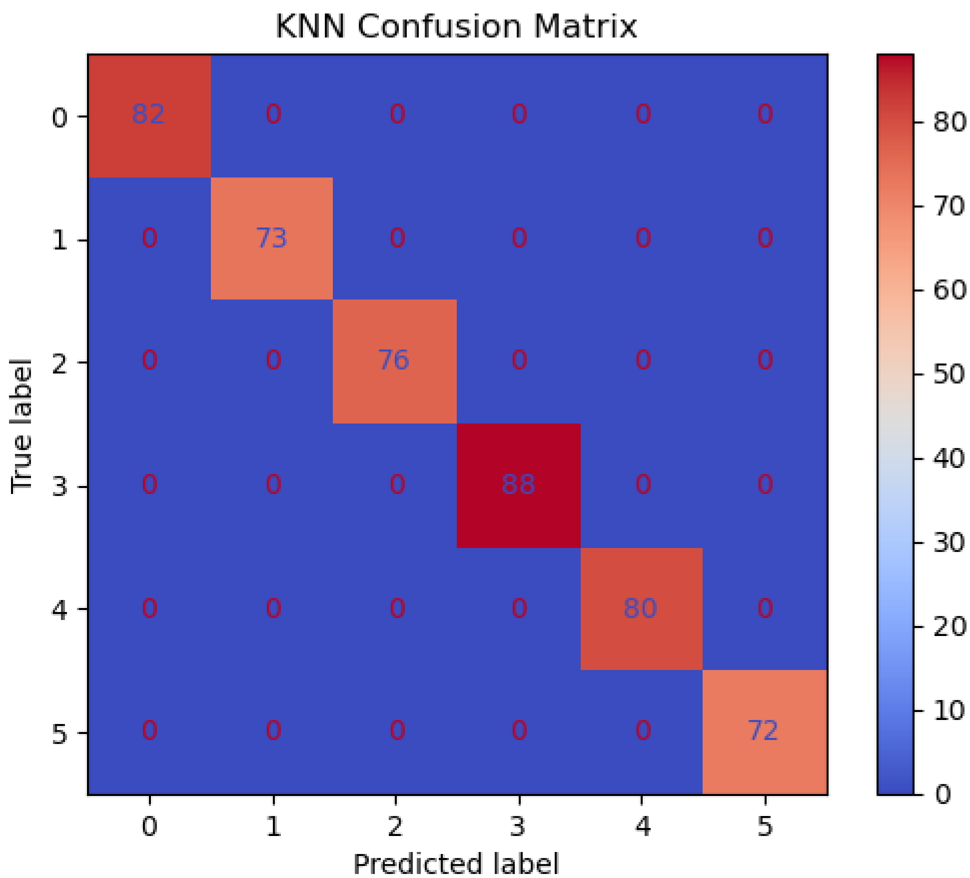
Recall (macro): 1.0

F1 Score (macro): 1.0

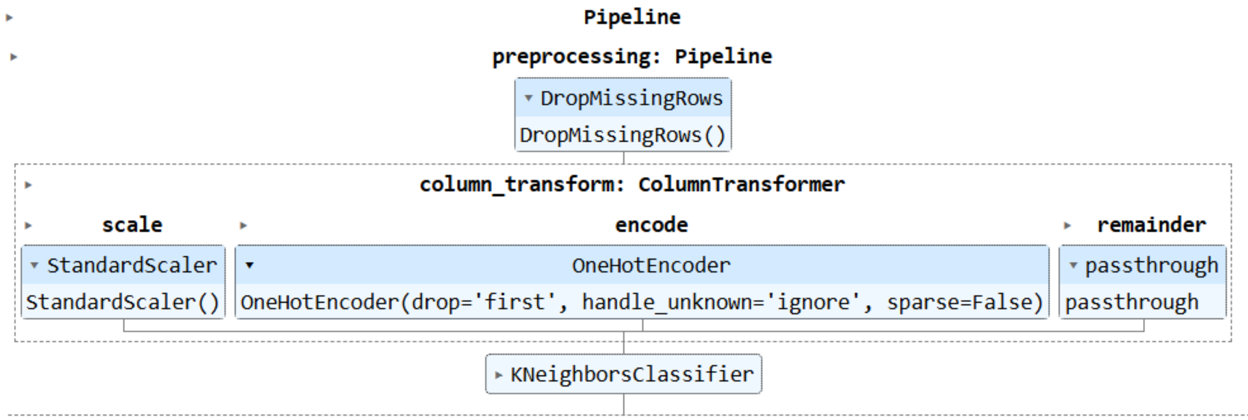
F2 Score (macro): 1.0

Confusion Matrix:

```
[[82  0  0  0  0  0]
 [ 0 73  0  0  0  0]
 [ 0  0 76  0  0  0]
 [ 0  0  0 88  0  0]
 [ 0  0  0  0 80  0]
 [ 0  0  0  0  0 72]]
```



Deploying Pipelines



USING THE DEPLOYED PIPELINES:

```python
# Remove 'S.No.' column from validate_df to match training features
validate_features = validate_df.drop(columns=['S.No.'])

# Normalize Star_Color in the same way as training data
validate_features['Star_Color'] =
validate_features['Star_Color'].str.lower().str.replace('-', ' ').str.strip()

# Preprocess the validation features using the preprocessing pipeline
validate_processed = preprocessing_pipeline.transform(validate_features)

# Predict Star_Type using the trained Random Forest pipeline
validate_predictions = rf_pipeline.predict(validate_features)

print(validate_predictions)
```

How Can These Models Help Physicists?

Capability	Benefit to Physicists
Speed & Scale	Processes millions of objects in hours, not years.
Accuracy	Near-perfect classification for common stars; robust for rare types.
Physical Insights	Derives fundamental properties (temperature, radius) directly from imaging.
Bias Mitigation	Physics-based synthetic data improves reliability in underrepresented classes.
New Discovery Pathways	Identifies anomalies and novel phenomena missed by traditional methods.

1. Automated Classification at Scale

- **Handles massive datasets** from sky surveys (e.g., Kepler, *Gaia*, SDSS) far faster than manual methods.
- **Reduces human bias and labor**, freeing physicists for higher-level analysis.

2. High-Precision Classification

- Achieves **>99% accuracy** for stars/galaxies and **>94% for quasars** using photometric data.
- **Light-curve analysis** (e.g., variable stars) reaches **99% accuracy** with models like Swin Transformers.

3. Physical Parameter Estimation

- Predicts **stellar properties** (e.g., temperature, luminosity) directly from broad-band photometry, with errors **<200 K** for temperature regression.
- Enables **data-driven discovery** of rare objects (e.g., hypergiants) by identifying outliers.

4. Mitigating Data Challenges

- **Self-regulating models** counter class imbalance and biases by generating synthetic data grounded in physics (e.g., using *Gaia* parameters).
- **Handles sparse/missing data** common in astronomical datasets.

5. Novel Applications

- **Single-band classification**: Identifies spectral types from diffraction patterns in images (e.g., *Hubble* or *Euclid*), achieving **half-spectral-class precision**.
- **Unsupervised discovery**: Detects new stellar classes without pre-defined labels (e.g., contact binaries).

6. Accelerating Research Workflows

- **Rapid candidate screening**: Prioritizes promising targets for follow-up spectroscopy or observation.
- **Democratizes analysis**: Tools like **PySSED** or **StarWhisper LightCurve** make advanced classification accessible.