SECTION:

MATHEMATICAL SCIENCES(INCLUDING STATISTICS)

DETECTION OF BLACK HOLES USING A LOGISTIC REGRESSION MODEL AND SVC ALGORITHM

AUTHOR-
Swapnanil Basu (bswapnanil2004@gmail.com)
St. Xavier's College(Autonomous), Kolkata
30, Mother Teresa Sarani,
Kolkata - 700016,West Bengal, India

CO-AUTHOR-
Sneha Maheshwari (sneha.maheshwari.2004@gmail.com)
St. Xavier's College(Autonomous), Kolkata
30, Mother Teresa Sarani,
Kolkata - 700016,West Bengal, India

KEYWORDS

1. **Black Hole**
2. **Cosmic Body**
3. **Gravity**
4. **Singularity**
5. **Stellar Populations**
6. **Educational Endeavors**
7. **Practical Applications**
8. **Vizier Dataset**
9. **Response Variable**
10. **Covariates/Predictors**
11. **Magnitude**
12. **Luminosity**
13. **Radius**
14. **Mass**
15. **Logistic Regression**
16. **SVC Algorithm**
17. **Assumptions**
18. **Logit Function**
19. **Support Vector Classifier (SVC)**
20. **Kernel Functions**
21. **Non-Linear Patterns**
22. **Multi-Class Classification**
23. **Decision Boundaries**
24. **Interpretation**

DECLARATION OF ISCA MEMBERSHIP INTENT

We, the undersigned authors of Detection of black holes using a logistic regression model and SVC algorithm, hereby declare our intent to become members of the Indian Science Congress Association (ISCA) before the commencement of its 109th Session.

Author- Swapnanil Basu
Co-author- Sneha Maheshwari

ACKNOWLEDGEMENTS

We would like to express our heartfelt gratitude to Rev. Dr. Dominic Savio, SJ, the esteemed principal of St. Xavier's College, Kolkata, for his unwavering support and encouragement throughout our academic journey.

We extend our sincere thanks to Dr. Ayan Chandra for his invaluable guidance and mentorship. His expertise and dedication have been instrumental in shaping our research and academic pursuits. We are profoundly grateful for the knowledge and insights we have gained under his tutelage.

We also wish to acknowledge the entire faculty of the Department of Statistics at St. Xavier's College for their dedication to excellence in education. Their commitment to fostering a vibrant learning environment has enriched our academic experience.

To all those who have contributed to our growth and learning, your support has been instrumental in our academic achievements, and we are deeply appreciative of your kindness and guidance.

ABSTRACT

This research project involves the innovative application of some machine learning algorithms in detection of black holes. We work with a dataset, made with the help of VizieR catalog access tool and The Hipparcos and Tycho   Catalogues,comprising various features such as Absolute Magnitude of the star, Spectral Type of the star,etc. Calculating some other features which are important features for the detection of a black hole we then design and train machine learning models,such as Logistic Regression model and SVC models, to classify stars and detect black holes based on the above features. We rigorously validate the machine learning models,comparing them with each other based on the evaluation metrics such as precision,recall and F1 score.

This research on the intersection of astrophysics and machine learning helps to revolutionize the ability to monitor black holes around the cosmos. It not only provides an efficient approach to black hole detection but also has the potential to discover the unknown black holes.

LITERATURE REVIEW

Black hole detection is not only about identifying and studying these enigmatic objects but also to gain deeper insights about the evolution of the cosmos, the fundamental laws of physics and the mysteries around them. This project is rooted in the quest for knowledge and to make profound contributions to cosmology and our understanding of the fundamental laws that govern the cosmos. We use machine learning algorithms such as Logistic Regression and SVC for this task. Logistic Regression provides a flexible and interpretable approach for binary classification problems in astrophysics. Its ability to model the probability of an event based on the given data makes it a reliable tool for use. It can be used in Stellar classification, supernova classification, etc. SVC provides a powerful tool for binary classification and helps to deal with non-linearly separable data.It is used in projects like exoplanet detection.For this purpose,we use high quality datasets to work with.High quality datasets are not only about quantity but also about accuracy and completeness of the data.It serves as the foundation upon which these machine learning algorithms can make significant discoveries and contribute to the field of astrophysics.

Some of the previous studies done on this topic, includes black hole attack detection using K-nearest neighbors algorithm.It states that KNN algorithm is used for clustering and fuzzy inference is used to calculate trust.The paper concludes that the proposed method can be combined with other algorithms such as SVC,neural network to obtain better results.Another work includes detection and interception of black hole attack with justification using anomaly based intrusion detection system in MANETs, stating that the research has done extensive investigation on machine learning models such as decision trees,SVM,KNN (Read: Black Hole Attack Detection Using K-Nearest Neighbor Algorithm and Reputation Calculation in Mobile Ad Hoc Networks) and neural networks. They have concluded that SVM works comparatively better than the other algorithms since it has an elevated accuracy rate.There are also limitations for black hole detection using machine learning

algorithms since the data available for this projects are relatively low and imbalance between positive and negative classes affect the performance of the model.Moreover astronomical data can be noisy due to factors such as atmospheric conditions,and background cosmic radiation.But, despite these challenges,machine learning offers significant promise for advancing black hole detection and astrophysical research.

INTRODUCTION

A black hole is a cosmic body of extremely intense gravity from which nothing, not even light, can escape. A black hole can be formed by the death of a massive star. When such a star has exhausted the internal thermonuclear fuels in its core at the end of its life, the core becomes unstable and gravitationally collapses inward upon itself, and the star's outer layers are blown away. The crushing weight of constituent matter falling in from all sides compresses the dying star to a point of zero volume and infinite density called the singularity.

Why Do We Pursue Black Hole Detection?
The endeavor to detect black holes serves multiple significant purposes, spanning from deepening our comprehension of the fundamental principles governing the cosmos to unraveling the enigmas concealed within the universe. Below are some primary motivations underpinning these undertakings:

Exploring Stellar Populations:
The identification and study of black holes in various regions of the Milky Way and distant galaxies aid astronomers in deciphering the distribution and characteristics of stellar populations. This, in turn, contributes to a more comprehensive grasp of the cosmic tapestry.

Inspiring Educational Endeavors:

Initiatives centered around black hole detection and research serve as a wellspring of inspiration for students and the broader public, fostering engagement in the realms of science and astronomy. These projects offer valuable opportunities for educational outreach, nurturing an enduring interest in STEM disciplines.

Practical Applications:
While the primary objective of black hole research is to advance scientific knowledge, it also holds the potential to yield unforeseen practical applications and cutting-edge technologies that can benefit society. Over time, numerous innovations stemming from astronomy and astrophysics research have found valuable applications in diverse fields such as medicine, telecommunications, and materials science.

In essence, black hole detection projects play a pivotal role in the pursuit of expanded cosmic understanding, the validation of fundamental physics theories, the cultivation of future generations of scientists, and the prospect of technological and pragmatic advancements. They stand as an indispensable cornerstone of contemporary astrophysical research.

Dataset in hand:

The dataset used for the purpose of detection of Black holes is the Vizier dataset.
This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France (DOI: 10.26093/cds/vizier). The original description of the VizieR service was published in A&AS 143, 23
The Hipparcos and Tycho Catalogues (ESA 1997)

For this data, the response variable is considered as “Black Hole” and the covariates/predictors are “Vmag”, “Plx”, “e_Plx”, “B-V”, “SpType”, “Amag”, “GiantorDwarf”, “TargetClass(renamed as GiantorDwarf later)”, “Temperature”, “Luminosity”, “Radius” and “Mass”.

The variables are described as:

Vmag: Visual Apparent Magnitude of the Star
Plx: Distance Between the Star and the Earth
e_Plx: Standard Error of Plx
B-V: B-V color index
SpType: Spectral type
Amag: Absolute Magnitude of the Star
Dwarf (0) and Giant (1): Star size

Formulae used:

Some of the other predictors were calculated using the following formulae:

1. **Temperature(in kelvins):**
   $T = 4600K(\frac{1}{0.92(B-V)+1.7} + \frac{1}{0.92(B-V)+0.62})$
   where $(B - V)$**: B-V color index.(A hot star has a B-V color index close to 0 or negative, while a cool star has a B-V color index close to 2.0.)**

2. **Magnitude and luminosity are related with the formula:**
   $M = - 2.5 * log_{10}(L/L_0)$
   where M **is the absolute magnitude of the star,** L **is its luminosity,** $L_0 = 3.0128 * 10^{28}W$ **is the zero-point luminosity.**

3. **Calculating the radius of the star in meters:**

   $r = \sqrt{\frac{L}{4\pi\sigma T^4}}$

4. **And finally, to get mass of stars the following relationship was used:**

   $(\frac{L}{L_{Sun}}) = (\frac{M}{M_{Sun}})^4$

PROPOSED METHODOLOGIES

The two proposed methodologies used to detect black holes from the dataset are:

1. **Logistic Regression**
2. **SVC Algorithm**

What is Logistic Regression?

It is a predictive algorithm using independent variables to predict the dependent variable, just like Linear Regression, but with a difference that the dependent variable should be a categorical variable.

Independent variables can be numeric or categorical variables, but the dependent variable will always be categorical

Logistic regression is a statistical model that uses Logistic function to model the conditional probability.

For binary regression, we calculate the conditional probability of the dependent variable Y, given independent variable X

It can be written as P(Y=1|X) or P(Y=0|X)

Logistic Regression can be used for binary classification or multi-class classification.

Assumptions for Logistic Regression:

1. **No outliers in the data. An outlier can be identified by analyzing the independent variables**
2. **No correlation (multicollinearity) between the independent variables.**

Logistic Regression function:

Logistic regression uses logit function, also referred to as log-odds; it is the logarithm of odds. The odds ratio is the ratio of odds of an event A in the presence of event B and the odds of event A in the absence of event B.

$$\ln\left(\frac{P}{1-P}\right) = \theta_1 + \theta_2 x + \mathrm{e}$$

$$\frac{P}{1-P} = \mathrm{e}^{\theta_1+\theta_2 x+\mathrm{e}}$$

$$P = \frac{1}{1+\mathrm{e}^{-}}(\theta_1 + \theta_2 x)$$

$$\sigma(z) = \frac{1}{1+\mathrm{e}^{-z}} \quad \text{Where} \quad z = \theta^T x$$

$$\theta^T \mathbf{x} = \sum_{i=1}^{m} \theta_i x_i = \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m$$

logit or logistic function

where,

- **P is the probability that event Y occurs. P(Y=1)**
- **P/(1-P) is the odds ratio**
- **θ is a parameters of length m**

Logit function estimates probabilities between 0 and 1, and hence logistic regression is a non-linear transformation that looks like S-function.

Why is Logistic Regression used in detecting black holes?

We know that Logistic Regression is a statistical technique, used primarily for binary classifications and in some cases , multi-class classification as well. We are implementing Logistic Regression since our dependent variable is categorical (1,if the given star is a probable black hole and 0 if the given star is not a probable black hole) . Moreover we have more than one independent variable (such as Temperature,Mass,Radius,etc.) which are continuous in nature. Additionally , Logistic Regression is computationally efficient . Hence,it can be used in situations dealing with large datasets, such as ours.

What is a Support Vector Classifier (SVC)?

An SVM classifier, or support vector machine classifier, is a type of machine learning algorithm that can be used to analyze and classify data. A support vector machine is a supervised machine learning algorithm that can be used for both classification and regression tasks. The Support vector machine classifier works by finding the hyperplane that maximizes the margin between the two classes. The Support vector machine algorithm is also known as a max-margin classifier. Support vector machine is a powerful tool for machine learning and has been widely used in many tasks such as hand-written digit recognition, facial expression recognition, and text classification. Support vector machine has many advantages over other machine learning algorithms, such as robustness to noise and the ability to handle large datasets.

SVM can be used to solve nonlinear problems by using kernel functions. For example, the popular RBF (radial basis function) kernel can be used to map data points into a higher dimensional space so that they become linearly separable. Once the data points are mapped, SVM will find the optimal hyperplane in this new space that can separate the data points into two classes.

Why is SVC used in detecting black holes?

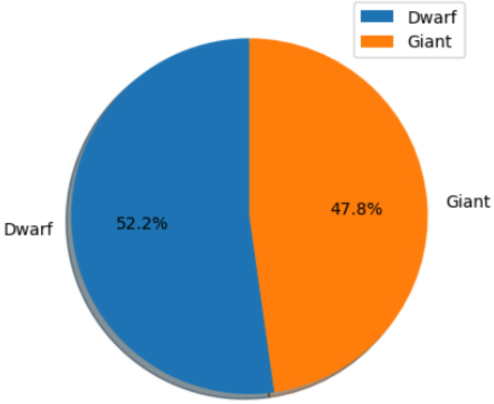
Since we have astronomical data in hand , it contains complex patterns and relationships which can be captured by SVC as SVC can capture both linear and non-linear patterns. SVC with the kernel functions, helps in the modeling of non-linear decision boundaries .
SVC can be extended to multi-class classification,helping to identify the different types of black holes.
SVC also offers transparent decision boundaries, helping in the interpretation of results more easily.

The conditions for a star to be a probable black hole are :

1. **If the star is a Hot type and giant (Like in the Hertzsprung–Russell diagram) [in our data hot = 1].**
2. **If the B-V of the star is below 1 which means its color is more on the blue side**
3. **If the star is giant [in our data giant=1]**
4. **If it's a main sequence star.**
5. **If the mass of the star is 20 times bigger than our estimated sun in this data.**

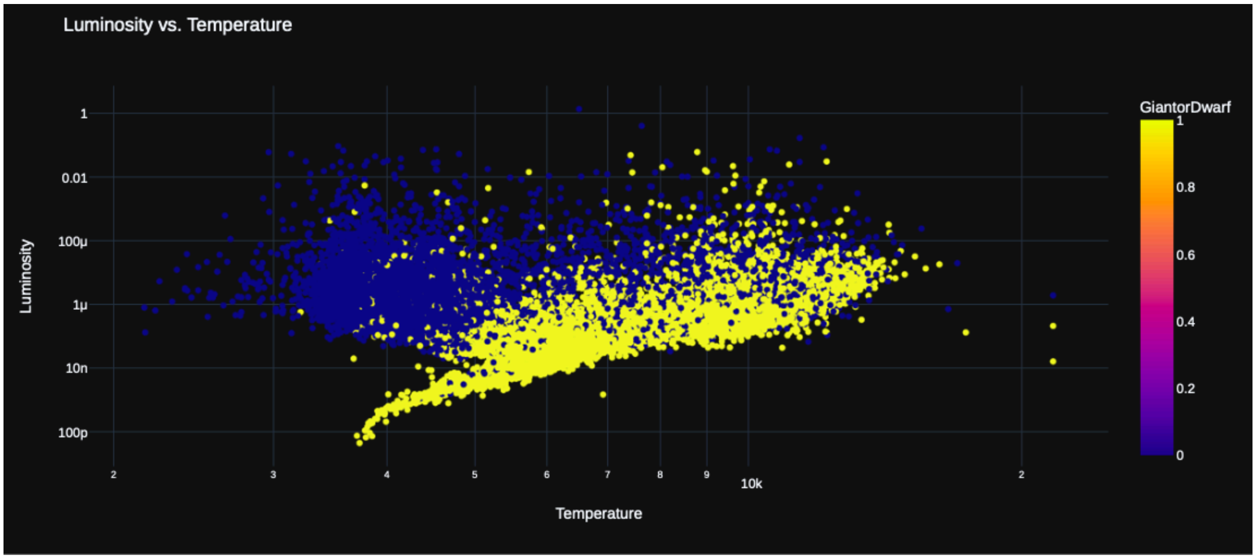
OBSERVATIONS

The types of stars in the dataset are:



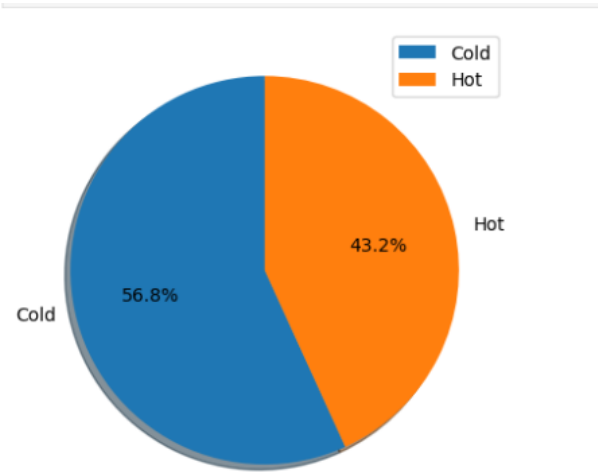
From the above pie-chart we can see that based on our data, we have a percentage of 52.2% dwarf stars and a percentage of 47.8% hot stars.

Luminosity vs Surface Temperature Hertzsprung-Russell Diagram for Main Sequence Stars is given as:



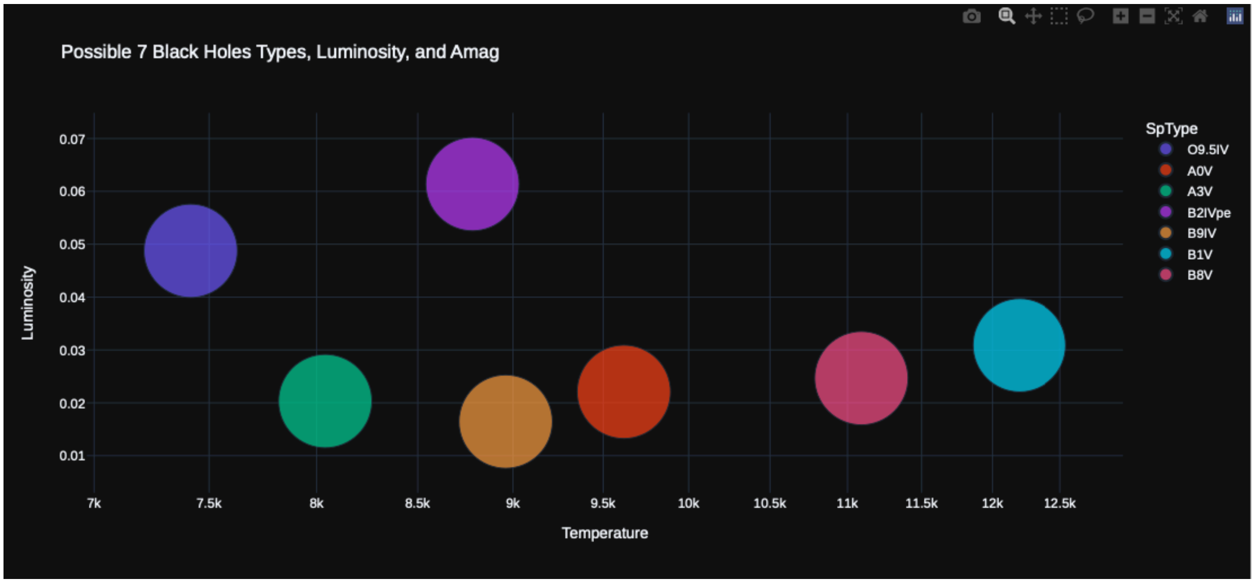
Our dataset exhibits a pattern reminiscent of the diagram, with supergiants and hypergiants positioned prominently at the upper end. In contrast, typical stars follow a life path within the main sequence. Dwarfs, on the other hand, predominantly populate the lower portion of the graph, characterized by lower temperatures and luminosity.

In order to detect black holes, we are mainly interested in hot stars so we plot a pie chart to see the percentage of hot stars in the dataset.



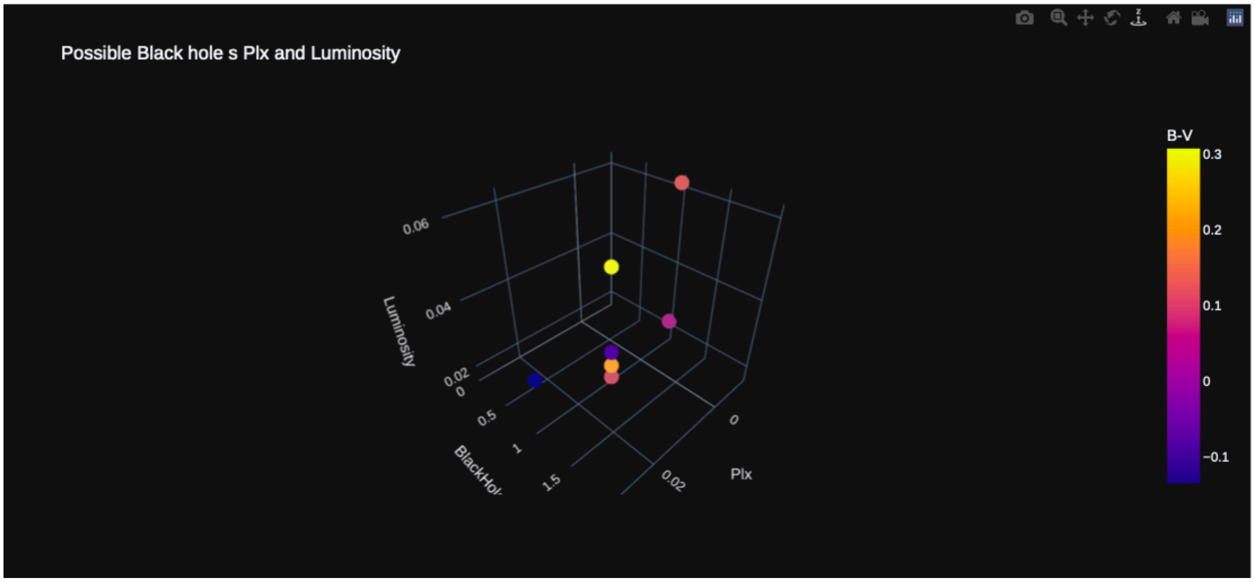
We can deduce that 56.8% of the stars are Cold stars whereas 43.2% stars are the Hot stars.

Finally we obtained,



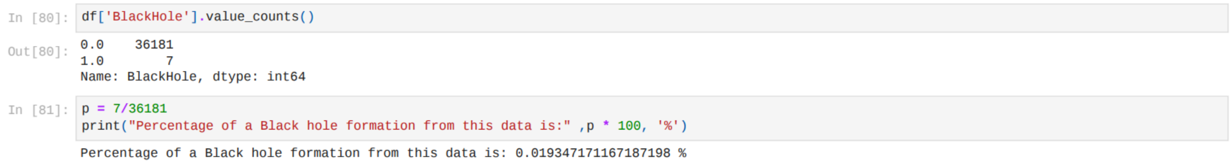
In the above figure,we get to see the Temperate and Luminosity of the probable 7 black hole types.

For example the black hole type with SpType B2IVpe has a temperature ranging between 8.5k K(Kelvin) - 9k K(Kelvin) and Relative Luminosity around 0.06.
Similarly for the SpType A3V,the Temperature ranges around 8k K(Kelvin) and Relative Luminosity around 0.02.



The above figure is a 3D plot of the probable black holes.We can visualize some of the important features of the probable black holes. For instance ,We can see one observation with the B-V color index of 0.3 and Relative Luminosity around 0.04.Observations on the other points can be made as well.

CONCLUSION

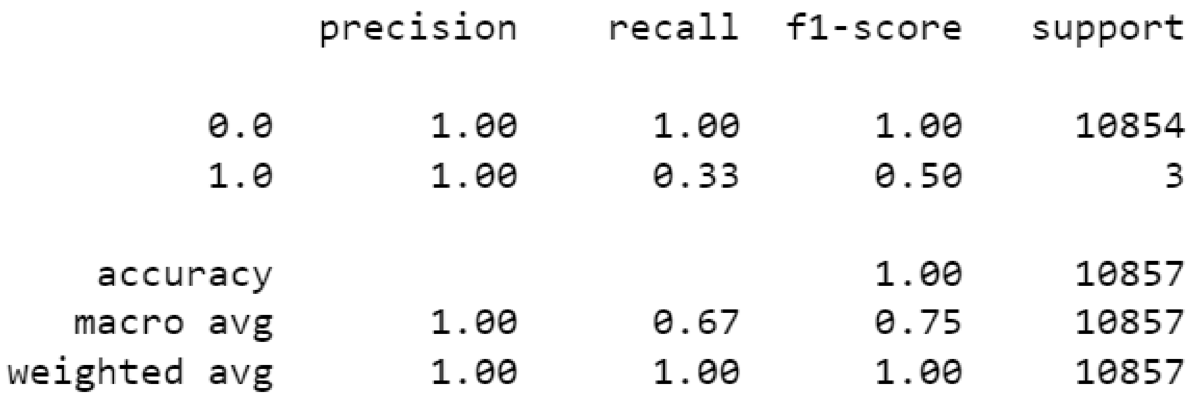


```
In [80]: df['BlackHole'].value_counts()

Out[80]: 0.0    36181
         1.0        7
         Name: BlackHole, dtype: int64

In [81]: p = 7/36181
         print("Percentage of a Black hole formation from this data is:" ,p * 100, '%')

         Percentage of a Black hole formation from this data is: 0.019347171167187198 %
```

From the above detection , we come to the conclusion that there are a total of 7 stars out of 36188 stars , which satisfy all the conditions to be a probable black hole.
Hence ,we observe that the percentage of a black hole formation from our dataset is 0.019347171%.

We know that while interpreting a classification report, we should focus on precision , recall and f1 score for assessing the model's performance for each class.(Class 1 and Class 0)

For our Logistic Regression model , we have the following classification report :



```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     10854
         1.0       1.00      0.33      0.50         3

    accuracy                           1.00     10857
   macro avg       1.00      0.67      0.75     10857
weighted avg       1.00      1.00      1.00     10857
```

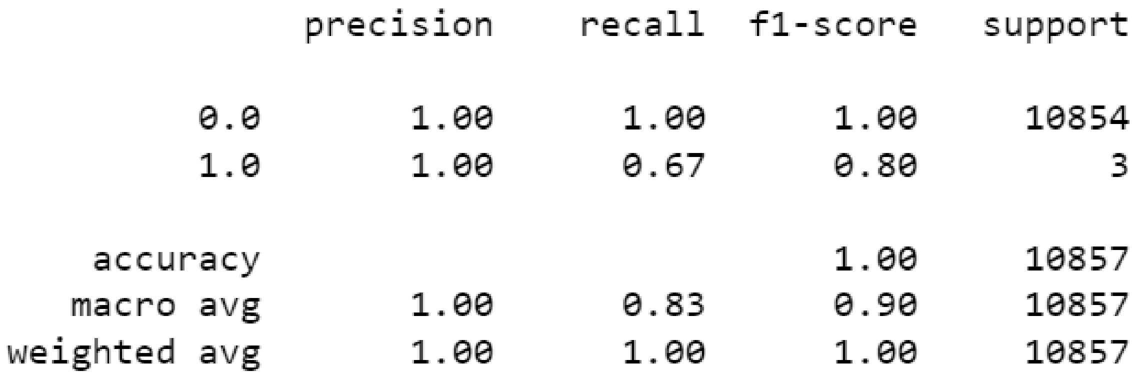
We can see that for both Class 1 and Class 0 , the precision is 1.00 which implies that 100% of the instances predicted as Class 1 and Class 0 were correct.

Similarly we can see that the recall for class 0 is 1.00 indicating that the model correctly identified 100% of the actual class 0 instances, whereas, for class 1 the recall is 0.33 , indicating that only 33% of the

actual class 1 instances were identified correctly implying that the model is not effectively capturing all instances of the positive class.

The F1 score is the Harmonic mean between the precision and recall and hence provides a balance between the above metrics.
The F1 score for class 1 is 0.50 indicating the balance.The F1 score for class 0 is 100%.
Thus we observe that our Logistic Regression model does not work well on our dataset for detecting black holes since it has a low recall and low F1 score.It may be because of the model complexity and hence we choose another model to work with.

We choose the SVC model for the task.The classification report is given as :



```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00     10854
         1.0       1.00      0.67      0.80         3

    accuracy                           1.00     10857
   macro avg       1.00      0.83      0.90     10857
weighted avg       1.00      1.00      1.00     10857
```

Here we can see that, same as Logistic regression it has a precision of 1.00 for both the classes. But unlike Logistic Regression, it gives a better recall and F1 score to interpret the model. The recall of Class 1 is 0.67 which means that 67% of the actual class 1 instances were identified correctly , implying that the model is moderately effective in identifying positive instances. Finally class 1 has a F1 score 0.8 showing that overall , the model is performing well and has achieved a good balance between making accurate positive predictions and capturing most actual positive instances.

Hence we can compare and see that the SVC model works much better than the Logistic Regression model.

Hence, in the light of the data we can conclude that out of 36188 observations/stars , there are 17286 main sequence stars. And based on our analysis,among those 17286 main sequence stars, there are 7 probable black holes which can be formed. It is implied that, roughly among 2469 main sequence stars, there is a probable black hole that can be detected.

Roughly one out of every 1000 that forms , is massive enough to form a black hole. Most of these are invisible to us and only a dozen can be identified. The nearest one is some 1600 light years from Earth.

REFERENCES

Introduction: https://www.britannica.com/science/black-hole

Logistic Regression:
https://towardsdatascience.com/quick-and-easy-explanation-of-logistics-regression-709df5cc3f1e

SVC:
https://vitalflux.com/svm-classifier-scikit-learn-code-examples/#:~:text=SVC%2C%20or%20Support%20Vector%20Classifier,the%20data%20into%20two%20classes

Radius of a star:
https://cas.sdss.org/dr4/en/proj/advanced/hr/radius1.asp

Star Mass:
https://www.researchgate.net/figure/Mass-radius-relation-for-Kepler-red-giants-with-RGB-stars-in-blue-and-clump-stars-in_fig6_257882348

Blackbody Radiation and Quantization of Energy for Luminosity and B-V:
https://web.njit.edu/~gary/321/Lecture2.html

Some definitions:

Main sequence star: A main sequence star is any star that has a hot, dense core which fuses hydrogen into helium to produce energy. https://study.com/academy/lesson/main-sequence-star-definition-facts-quiz.html

Hertzsprung-Russell Diagram: It plots the temperature of stars against their luminosity (the theoretical HR diagram), or the color of stars (or spectral type) against their absolute magnitude (the observational HR diagram, also known as a color-magnitude diagram). Depending on its initial mass, every star goes through specific evolutionary stages dictated by its internal structure and how it produces energy. https://astronomy.swin.edu.au/cosmos/h/hertzsprung-russell+diagram