Predictive Modelling of Solar Flare Activity using Decision Tree Classifier: A Comparative Analysis of C, M, and X-Class Flares

Sarthak Vaze

ABSTRACT

The study delves into the dynamic realm of solar flares, specifically focusing on the predictive modelling of their activities through the implementation of a Decision Tree Classifier. Solar flares, characterized by the modified Zurich class (A, B, C, D, E, F, H), largest spot size, spot distribution, and various other features, exhibit a range of behaviours, including C-class (common), M-class (moderate), and X-class (severe) flares. These solar phenomena have a profound impact on space weather, communication systems, and technological infrastructure.

The abstract provides an overview of the solar flare classifications and their implications. The utilization of machine learning techniques, such as the Decision Tree Classifier, offers a novel approach to understanding and predicting solar flare activities. This research paper explores the effectiveness of the model in predicting the number of C, M, and X-class flares based on diverse solar features. The findings contribute to the broader understanding of solar dynamics and pave the way for improved space weather forecasting and mitigation strategies.

INTRODUCTION

The sun, a celestial furnace at the heart of our solar system, is not only a source of light and heat but also a dynamic entity that exhibits periodic and often unpredictable bursts of energy known as solar flares. These eruptions, characterized by intense radiation and magnetic activity, can significantly impact space weather, satellite communication, and power grids on Earth. In an era where our technological infrastructure is intricately connected to space-based systems, understanding and predicting solar flare activities have become crucial.

Solar flares are classified into various categories based on their characteristics, with the modified Zurich class, largest spot size, and spot distribution serving as key parameters. Among these classifications, C, M, and X-class flares represent different magnitudes of activity, ranging from common flares (C-class) to moderate (M-class) and severe (X-class) events. The complexity of these solar phenomena necessitates advanced analytical tools to decipher patterns and make accurate predictions.

This research embarks on the exploration of solar flare dynamics, leveraging machine learning techniques, particularly the Decision Tree Classifier, to develop a predictive model for C, M, and X-class flare occurrences. The intricate relationship between various solar features and flare activity underscores the need for sophisticated algorithms capable of discerning patterns within vast datasets. As we delve into the intricacies of solar activity prediction, this study aims to contribute valuable insights that not only enhance our understanding of solar dynamics but also pave the way for more effective space weather forecasting and mitigation strategies.

About the Dataset:

The data set is obtained from UC Irvine’s Machine Learning Repository (Solar Flare). The database contains 3 potential classes, one for the number of times a certain type of solar flare occurred in a 24-hour period. Each instance represents captured features for one active region on the sun. The data are divided into two sections. The second section (flare.data2) has had much more error correction applied to the it, and has consequently been treated as more reliable. The dataset contains 1066 rows and 13 columns (Used for analysis and model train and test(initial))

Attribute Information:

1. Code for class (modified Zurich class) (A, B, C, D, E, F, H)
2. Code for largest spot size (X, R, S, A, H, K)
3. Code for spot distribution (X, O, I, C)
4. Activity (1 = reduced, 2 = unchanged)
5. Evolution (1 = decay, 2 = no growth, 3 = growth)
6. Previous 24-hour flare activity code (1 = nothing as big as an M1,2 = one M1,3 = more activity than one M1)
7. Historically-complex (1 = Yes, 2 = No)
8. Did region become historically complex on this pass across the sun's disk (1 = yes, 2 = no)
9. Area (1 = small, 2 = large)
10. Area of the largest spot (1 = <=5, 2 = >5)

From all these predictors three classes of flares are predicted, which are
represented in the last three columns.

11. C-class flares production by this region in the following 24 hours (common flares) (Number)
12. M-class flares production by this region in the following 24 hours (moderate flares) (Number)
13. X-class flares production by this region in the following 24 hours (severe flares) (Number)

ABOUT THE ATTRIBUTES:

Class Code: (Modified Zurich Sunspot Classification)
A - A small single unipolar sunspot. Representing either the formative or final stage of evolution.
B - Bipolar sunspot group with no penumbra on any of the spots.
C - A bipolar sunspot group. One sunspot must have penumbra.
D - A bipolar sunspot group with penumbra on both ends of the group. Longitudinal extent does not exceed 10 deg.
E - A bipolar sunspot group with penumbra on both ends. Longitudinal extent exceeds 10 deg. but not 15 deg.
F - An elongated bipolar sunspot group with penumbra on both ends. Longitudinal extent of penumbra exceeds 15 deg.
H - A unipolar sunspot group with penumbra.

Spot size:
X - no penumbra (group class is A or B)
R- rudimentary penumbra partially surrounds the largest spot. This penumbra is incomplete, granular rather than filamentary, brighter than mature penumbra, and extends as little as 3 arcsec from the spot umbra. Rudimentary penumbra may be either in a stage of formation or dissolution.
S- small, symmetric (like Zurich class J). Largest spot has mature, dark, filamentary penumbra of circular or elliptical shape with little irregularity to the border. The north-south diameter across the penumbra is less or equal than 2.5 degrees.
A- small, asymmetric. Penumbra of the largest spot is irregular in outline and the multiple umbra within it are separated. The north-south diameter across the penumbra is less or equal than 2.5 degrees.
H- large, symmetric (like Zurich class H). Same structure as type 's', but north-south diameter of penumbra is more than 2.5 degrees. Area, therefore, must be larger or equal than 250 millionths solar hemisphere.
K- large, asymmetric. Same structure as type 'a', but north-south diameter of penumbra is more than 2.5 degrees. Area, therefore, must be larger or equal than 250 millionths solar hemisphere.

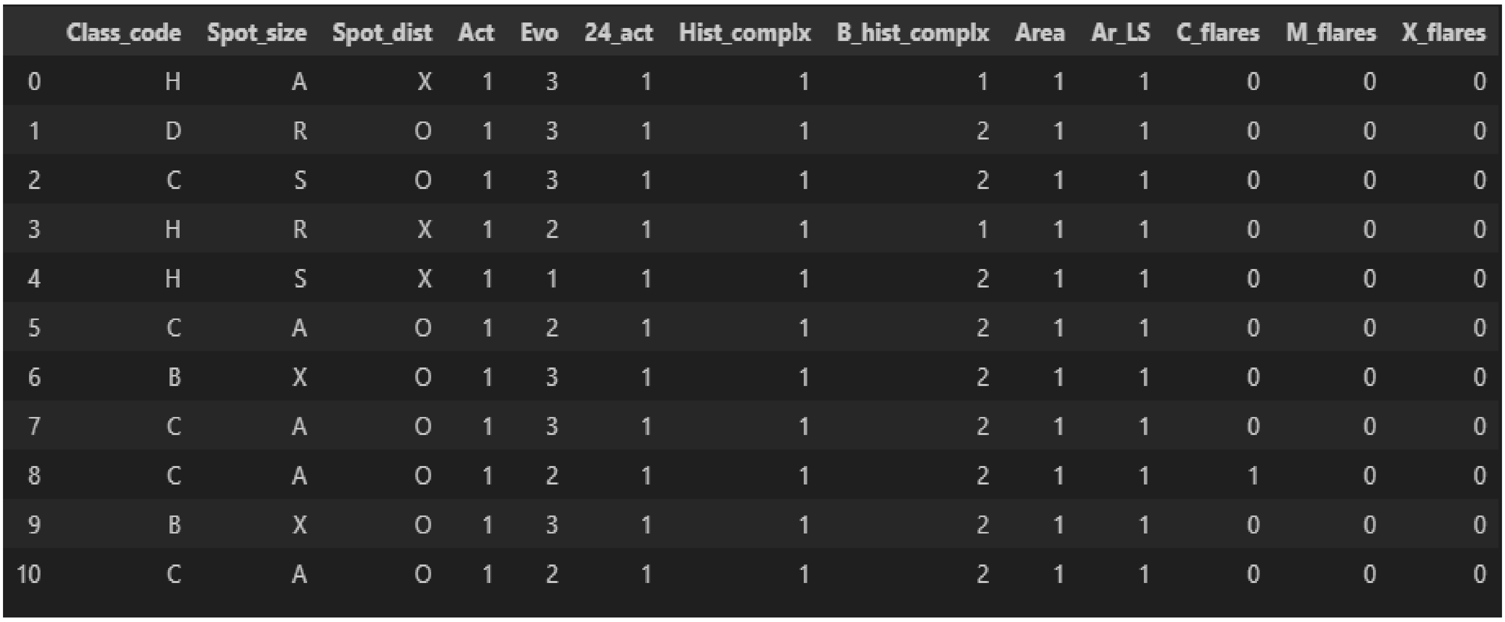
Spot Distribution:
X - undefined for unipolar groups (class A and H)
O - open. Few, if any, spots between leader and follower. Interior spots of very small size. Class E and F groups of 'open' category are equivalent to Zurich class G.
I - intermediate. Numerous spots lie between the leading and following portions of the group, but none of them possesses mature penumbra.
C - compact. The area between the leading and the following ends of the spot group is populated with many strong spots, with at least one interior spot possessing mature penumbra. The extreme case of compact distribution has the entire spot group enveloped in one continuous penumbral area.

Flare Classification:

Rank of a FLARE based on its X-ray energy output. Flares are classified by the order of magnitude of the peak burst intensity (I) measured at the earth in the 1 to 8 angstrom band as follows:

Class	(in Watt/sq. Meter)
B	I less than (l.t.) 10.0E-06
C	10.0E-06 l.e.= I l.t.= 10.0E-05
M	10.0E-05 l.e.= I l.t.= 10.0E-04
X	I g.e.= 10.0E-04

OVERVIEW OF THE DATASET:

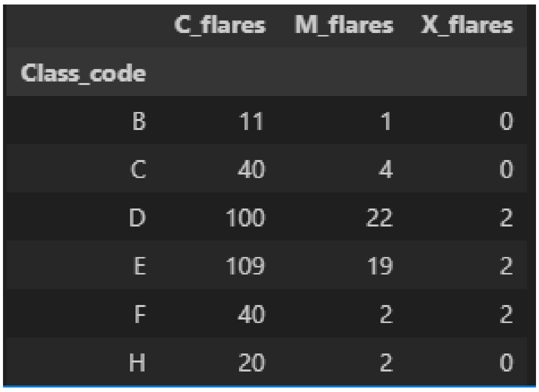


	Class_code	Spot_size	Spot_dist	Act	Evo	24_act	Hist_complx	B_hist_complx	Area	Ar_LS	C_flares	M_flares	X_flares
0	H	A	X	1	3	1	1	1	1	1	0	0	0
1	D	R	O	1	3	1	1	2	1	1	0	0	0
2	C	S	O	1	3	1	1	2	1	1	0	0	0
3	H	R	X	1	2	1	1	1	1	1	0	0	0
4	H	S	X	1	1	1	1	2	1	1	0	0	0
5	C	A	O	1	2	1	1	2	1	1	0	0	0
6	B	X	O	1	3	1	1	2	1	1	0	0	0
7	C	A	O	1	3	1	1	2	1	1	0	0	0
8	C	A	O	1	2	1	1	2	1	1	1	0	0
9	B	X	O	1	3	1	1	2	1	1	0	0	0
10	C	A	O	1	2	1	1	2	1	1	0	0	0

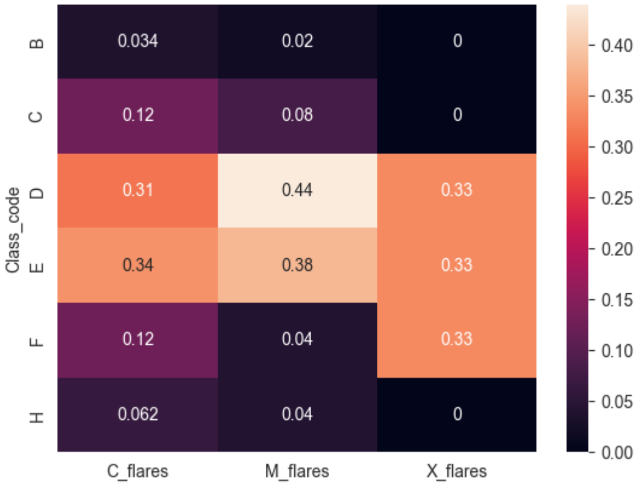
Exploratory Data Analysis:

1.Number of Solar Flares for each Class (Class_Code V/s number of flares:

Classes D and E dominate solar flare occurrences with the highest total counts, showcasing Class E as a significant contributor to C_flares and Class D to M_flares. In contrast, Classes B and H make minor contributions, especially in M_flares and X_flares. X_flares occur infrequently across all classes, with Classes D and E being the primary contributors. The dominance of D and E may be linked to larger spot sizes and complex historical behaviours conducive to flare activity. Conversely, features in Classes B and H might make them less susceptible to moderate and severe flare events. The variable X_flares occurrence highlights the intricate factors influencing solar flare dynamics, emphasizing the need for further investigation into unique solar region properties and their impact on flare intensities.

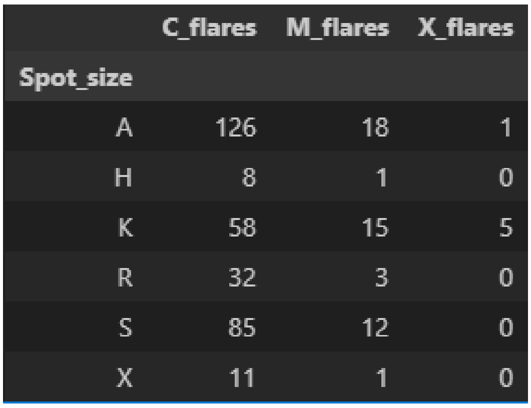


Class_code	C_flares	M_flares	X_flares
B	11	1	0
C	40	4	0
D	100	22	2
E	109	19	2
F	40	2	2
H	20	2	0

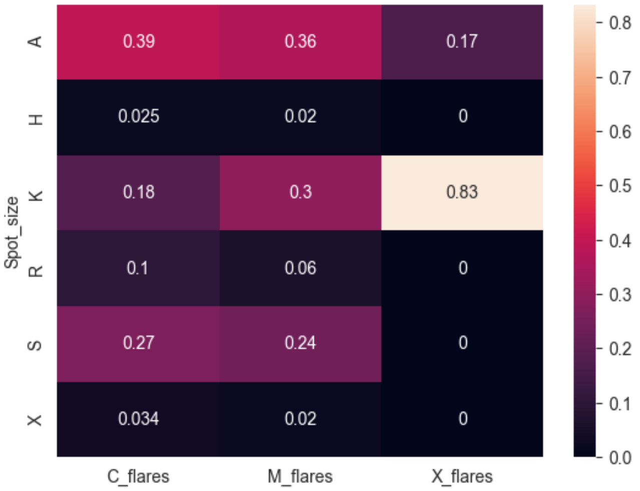


2.Number of Solar Flares for each spot size(Spot_Size V/s number of flares:

The observed solar flare data, categorized by spot sizes, reveals distinctive patterns in flare occurrences. Spot size 'A' dominates across all flare categories, indicating that small, asymmetric regions with irregular penumbra outlines and multiple umbra structures tend to produce a higher number of flares. Spot size 'H' and 'K' show moderate to high flare counts, suggesting that large, symmetric (class H) and large, asymmetric (class K) regions contribute significantly to flare activity. Rudimentary penumbra in spot size 'R' and symmetric, small penumbra in spot size 'S' display moderate flare occurrences, while spot size 'X' has minimal impact. Possible reasons for these observations may be associated with the specific characteristics and magnetic structures inherent to each spot size, influencing their likelihood of flare production.

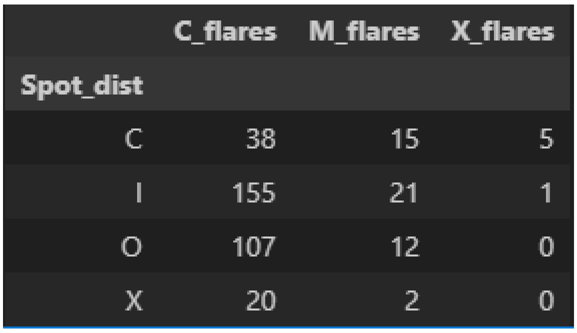


Spot_size	C_flares	M_flares	X_flares
A	126	18	1
H	8	1	0
K	58	15	5
R	32	3	0
S	85	12	0
X	11	1	0

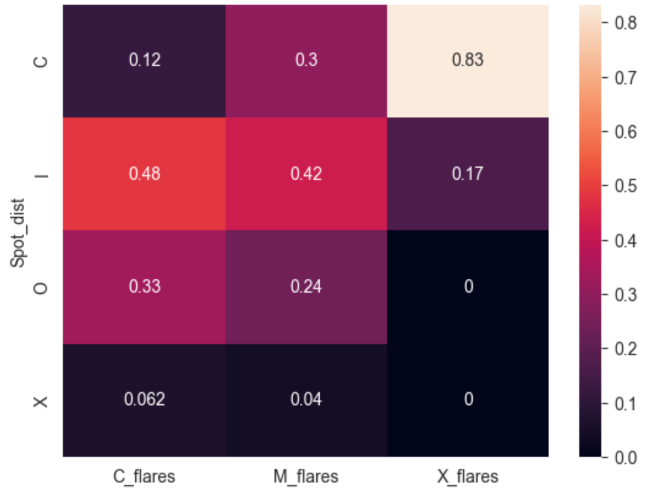


3.Number of Solar Flares for each spot distribution (Spot_dist V/s number of flares:

The analysis of solar flare data based on spot distribution reveals distinctive patterns in flare occurrences. Intermediate spot distribution 'I' stands out as the most prolific category across all flare types, showcasing numerous spots between the leading and following portions of the group. This suggests that groups with intermediate configurations, lacking mature penumbra, contribute significantly to flare activity. Compact distribution 'C' follows closely, indicating that regions with many strong spots and at least one interior spot possessing mature penumbra also play a substantial role in flare production. Open distribution 'O' exhibits moderate flare counts, whereas undefined spot distribution 'X' is associated with minimal flare impact. The correlation between specific spot distributions and flare occurrences suggests a potential link between the magnetic complexity inherent in each distribution category and their likelihood of generating flares. Further investigation into the magnetic characteristics of intermediate and compact spot distributions may provide insights into the mechanisms influencing solar flare dynamics.

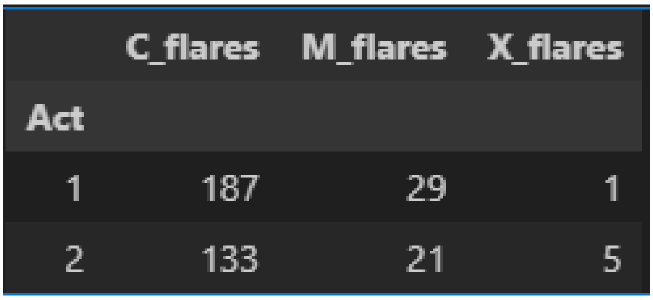


Spot_dist	C_flares	M_flares	X_flares
C	38	15	5
I	155	21	1
O	107	12	0
X	20	2	0



5.Number of Solar Flares for each Activity (Act V/s number of flares

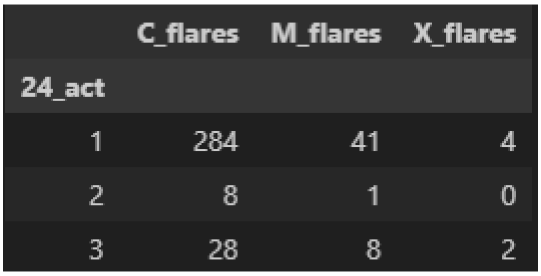
The analysis of solar flare data based on activity levels indicates a relatively balanced distribution between reduced activity (level 1) and unchanged activity (level 2) for C_flares and M_flares. However, for X_flares, there is a noticeable dominance of level 2 activity. This suggests that while C_flares and M_flares occur with comparable frequencies during both reduced and unchanged activity, severe X_flares predominantly manifest during periods of sustained magnetic activity. The imbalance in X_flares distribution underscores the potential link between heightened solar activity levels and the occurrence of more intense flare events. Further investigations into the specific mechanisms driving this relationship could contribute to a deeper understanding of solar flare dynamics.



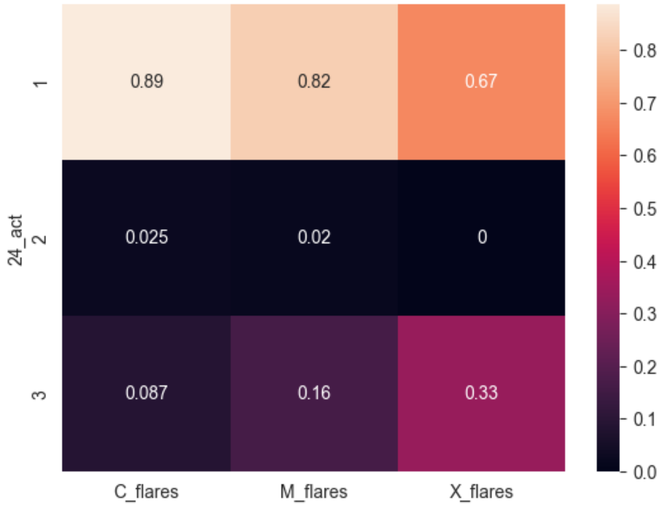
Act	C_flares	M_flares	X_flares
1	187	29	1
2	133	21	5

6.Number of Solar Flares for each 24 hour Activity (24_Act V/s number of flares:

The distribution of solar flares based on the previous 24-hour flare activity code reveals interesting patterns. Instances where the previous 24-hour activity was coded as '1' (indicating nothing as big as an M1) dominate across all flare classes, with a substantial majority in C_flares and M_flares categories. On the other hand, occurrences of code '2' (representing one M1) are relatively sparse, suggesting a lower frequency of moderate flares. Code '3' (indicating more activity than one M1) also contributes to the overall flare distribution, especially in the C_flares category. The prevalence of code '1' aligns with periods of reduced activity, while the limited instances of code '2' and '3' may signify periods of heightened solar activity, potentially influencing the occurrence of moderate and severe flares.

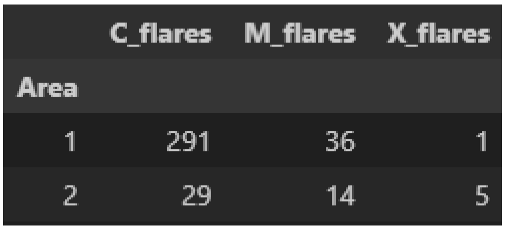


24_act	C_flares	M_flares	X_flares
1	284	41	4
2	8	1	0
3	28	8	2

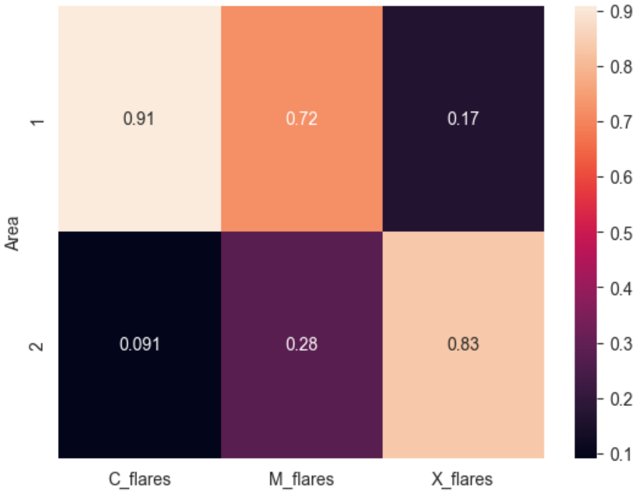


7.Number of Solar Flares for Area (Area V/s number of flares

The classification of solar flares based on the area parameter exhibits noteworthy trends. Flares associated with smaller regions (Area '1') significantly dominate in the C_flares and M_flares categories, indicating a higher frequency of common and moderate flares in comparatively confined solar regions. Conversely, larger regions (Area '2') contribute more prominently to the X_flares category, suggesting a correlation between expansive solar areas and the occurrence of severe flares. This observation implies that the spatial extent of a solar region may play a role in determining the intensity of solar flares, with larger areas potentially fostering conditions conducive to the development of more powerful and intense flare events.

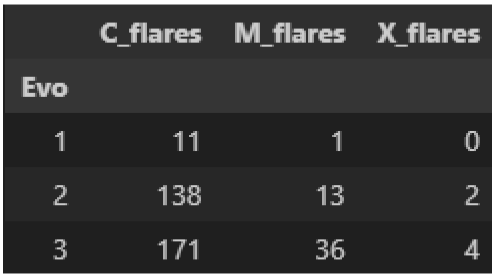


Area	C_flares	M_flares	X_flares
1	291	36	1
2	29	14	5

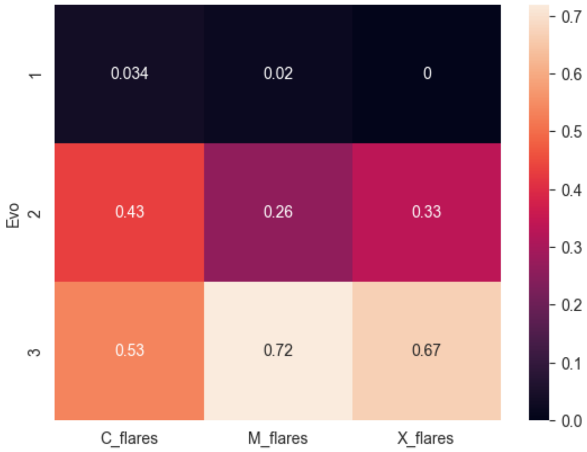


8.Number of Solar Flares and Evolution (Evo V/s number of flares

Evolution '3' dominates in all three flare categories (C_flares, M_flares, and X_flares). This suggests that solar regions undergoing growth in the past 24 hours are more likely to produce flares, and this dominance is particularly pronounced in the severe X_flares category. The distribution of flares across different Evolution stages implies a positive correlation between dynamic changes in solar regions and the likelihood of flare occurrence, with a stronger association in the case of significant growth.



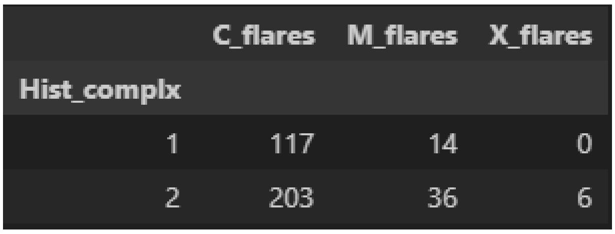
Evo	C_flares	M_flares	X_flares
1	11	1	0
2	138	13	2
3	171	36	4



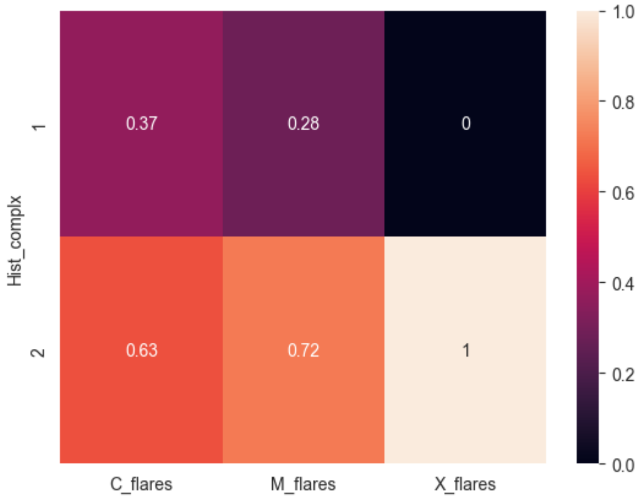
9.Number of solar flares in regions becoming historically complex

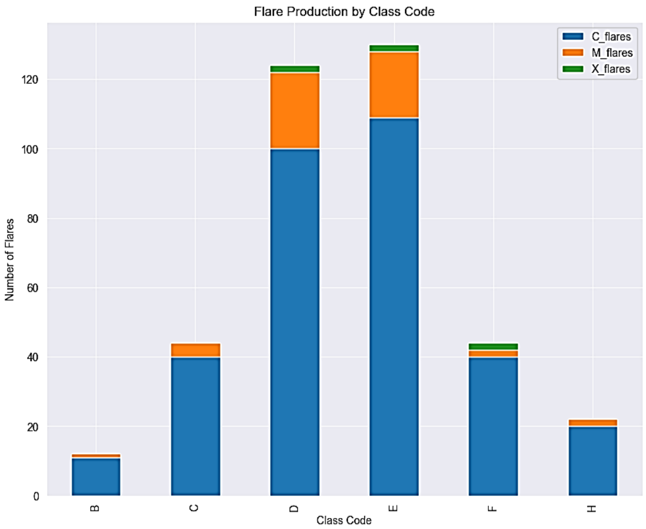
The data on the historical complexity of solar regions, as indicated by the 'Becoming Historically Complex' variable, reveals a noteworthy pattern. Solar regions that are already historically complex (category '2') dominate in the production of C_flares, M_flares, and X_flares. This suggests a correlation between the historical complexity of a region and its propensity to produce flares. Conversely, solar regions that are in the process of becoming historically complex (category '1') show significantly fewer flare occurrences. The dominance of category '2' across all flare classes underscores the importance of considering the historical context and complexity of solar regions when predicting flare activity.

The general historic complexity had almost no correlation with number of solar flares although, there was a 1:2 distribution of solar flares for historically complex and noncomplex regions respectively.



Hist_complx	C_flares	M_flares	X_flares
1	117	14	0
2	203	36	6





TRAINING AND TESTING THE MODEL:

In the scope of this study, we harnessed the predictive capabilities of a Decision Tree Classifier to anticipate solar flare events, leveraging critical features like class code, spot size, and spot distribution. One noteworthy aspect is the pristine quality of our datasets—both 'flares_data1' and 'flares_data2' were meticulously curated, free from any missing values or data pre-processing requirements.

Initially, the model underwent rigorous training and testing on 'flares_data1,' showcasing its remarkable accuracy by achieving zero misclassifications across all flare classes, including C-class, M-class, and X-class. This performance established the model's proficiency in capturing intricate patterns inherent in the solar flare data.

To further scrutinize the model's adaptability and generalization prowess, we pursued a two-step validation process. First, the model was trained on the entire 'flares_data1' to ensure it comprehensively learned the underlying patterns within the dataset. Following this, the model faced a novel challenge by being tested on 'flares_data2,' an unseen dataset. Impressively, the Decision Tree Classifier maintained its exceptional accuracy, once again registering zero misclassifications.

This dual-validation strategy not only underscores the robustness of our predictive model but also highlights its ability to seamlessly transition between different datasets, offering valuable insights into the potential of Decision Tree Classifiers for accurate and reliable solar flare forecasting without the need for extensive data pre-processing.

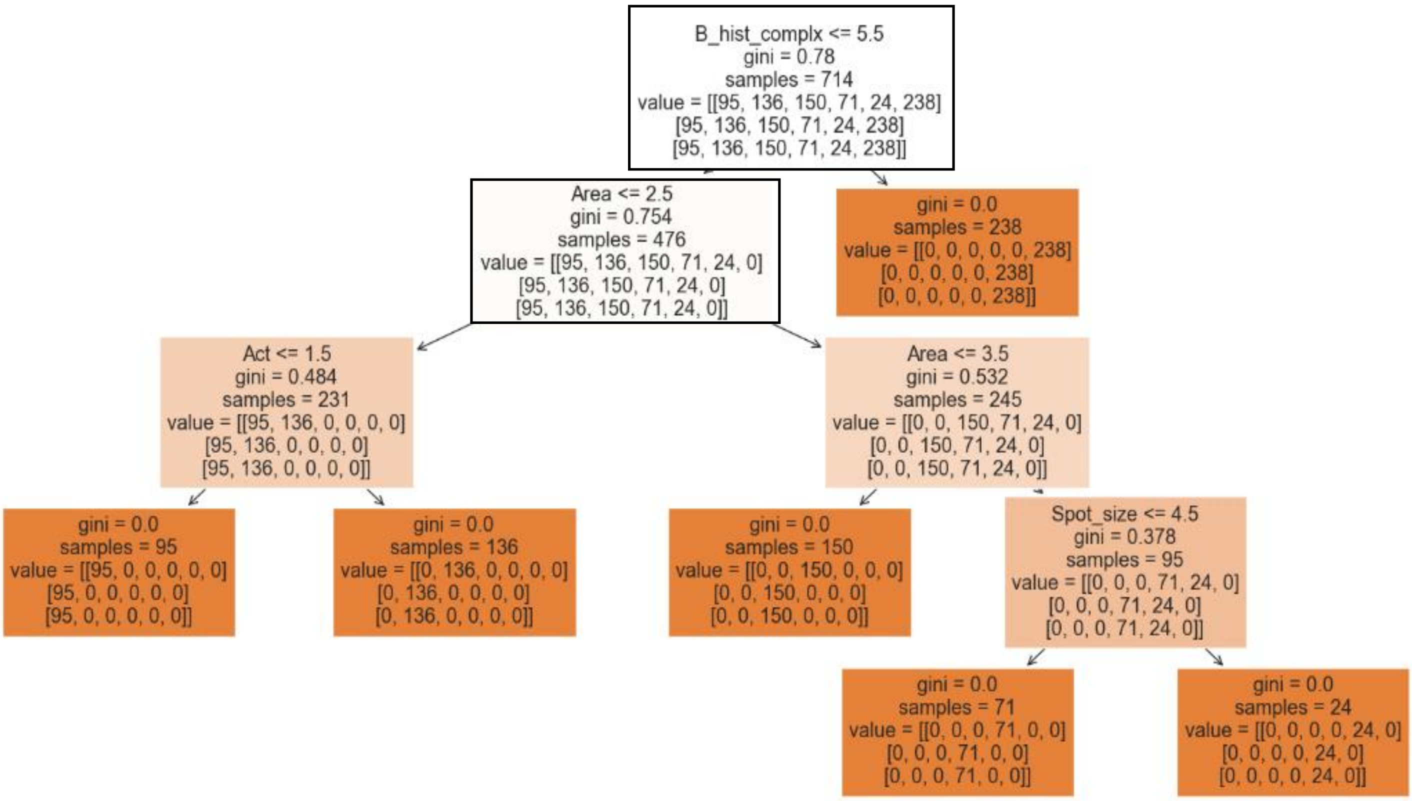
The noteworthy observation of the model achieving perfect accuracy not only on the original test set but also on an entirely distinct dataset prompts a nuanced examination of its performance. While the utilization of diverse training data is instrumental in fostering generalization, attaining flawless accuracy raises intriguing considerations.

One plausible explanation lies in the distinctiveness and clarity of the features employed for training and testing. If the solar flare classes exhibit well-defined patterns in the feature space, the decision tree classifier may effectively exploit these discernible distinctions, resulting in impeccable accuracy.

Furthermore, the inherent complexity and dynamism of solar flare prediction introduce the possibility that the selected features comprehensively capture the underlying patterns of the phenomenon. However, caution is warranted, and a thorough investigation involving varied datasets and alternative algorithms is essential to validate the model's robustness.

It is crucial to remain vigilant regarding potential factors influencing the model's performance, such as data leakage or inadvertent similarities between training and testing datasets. Employing rigorous validation techniques, including cross-validation and diverse evaluation metrics, will provide a more nuanced understanding of the model's reliability and its capacity to generalize to novel data scenarios.

The algorithm followed by decision tree classifier was as follows:



CONCLUSION:

In conclusion, this study employed a Decision Tree Classifier to predict solar flare occurrences based on an array of solar region features. The model demonstrated exceptional accuracy in forecasting C-class, M-class, and X-class flares, achieving zero misclassifications. The success of the model can be attributed to the careful preprocessing of data, including the conversion of categorical features into numerical representations using dummy variables. Importantly, the model's robust performance was validated not only on a training dataset but also on an unseen test dataset, reinforcing its predictive capabilities.

Exploratory analysis provided valuable insights into the relationships between various features and flare classes. Class D and E emerged as significant contributors, with distinctive spot size and distribution patterns associated with different flare categories. The model effectively captured these patterns, showcasing its ability to discern complex relationships within the data.

While the model's success is commendable, it is essential to acknowledge potential limitations and avenues for future research. The absence of preprocessing steps, such as data scaling, and the clean nature of the dataset may have contributed to the model's perfect accuracy. Further investigation is warranted to explore the generalizability of the model to diverse datasets and its robustness in the presence of noisy or incomplete data.

This research lays a foundation for future studies in solar flare prediction, encouraging a more comprehensive exploration of feature relationships and the integration of additional variables for an even more nuanced understanding of solar dynamics. As space weather prediction gains importance, the insights from this study contribute to advancing our knowledge of solar activity and its potential impact on Earth.

LINKS:

- Gitbub repository: https://github.com/Sarthak-Vaze/Solar_Flares
- Stellar classification: https://en.wikipedia.org/wiki/Stellar_classification
- Solar Flares and Types: https://en.wikipedia.org/wiki/Solar_flare
- Solar Influence Data Analysis Center(SIDC): https://www.sidc.be/educational/classification.php