HW7

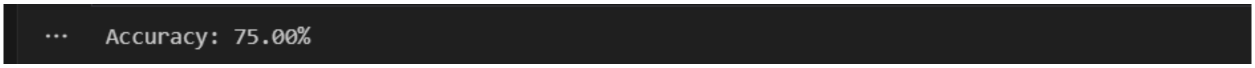
Part 1:

In Part 1, we are working with the original dataset and applying a Random Forest classifier after normalizing the citation data and converting categories to numerical values. The accuracy obtained in Part 1 is 75%.

This accuracy is less than the accuracy we got with the logistic regression approach.

The primary features used for training the model are the normalized citation counts for each year (2017 to 2022). While normalization is applied, there is no additional feature engineering to capture more complex relationships in the data. The model might struggle if the underlying patterns are not well-represented by the chosen features.

Also, the dataset is small, the model may not be able to learn the patterns effectively.

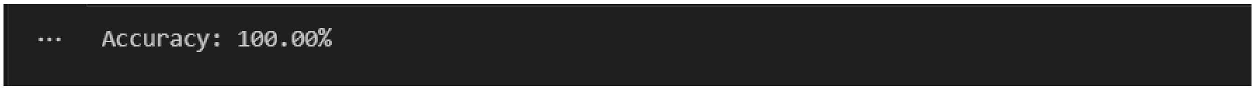


```
...   Accuracy: 75.00%
```

Part 2:

In Part 2, we introduce new features related to citation changes over the years and exclude raw citation numbers. The accuracy increases to 100%, which could be attributed to the following reasons:

1. The new features, calculated as citation changes over the years, may capture more informative patterns in the data, leading to a better representation of the underlying relationships.
2. By excluding raw citation numbers and focusing on changes, the model may become less sensitive to absolute citation counts and more sensitive to trends and patterns in citation behavior.
3. Raw citation counts can be noisy and influenced by various external factors.
4. Citation_change features allow the model to capture non-linear patterns in citation growth or decline. Linear models might struggle to capture complex relationships, and the Random Forest classifier is well-suited for handling non-linear relationships.



```
...   Accuracy: 100.00%
```

References:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
https://www.codementor.io/@agarrahul01/multiclass-classification-using-random-forest-on-scikit-learn-library-hkk4lwawu