Brijesh verma
SUID: 826257344

CIS 662: Intro to Machine Learning and Algorithms

HW-7

1. Problem Statement:

Part-1:

Random forest approach for classification.
Use the same dataset that you used for HW5 for the classification
(It probably won't perform well; say why.)

Part-2:

Introduce **5 new features** based on the citation numbers. and use them in the **RF** instead of the citation numbers directly.
Each new feature is:
((citation number in year n+1)-(citation number in year n))/(citation number in year n) for 2016<n<2022.

2. Solution Steps:

In this report, we address the problem of classification using a Random Forest approach. We explore two parts: in Part 1, we apply Random Forest on a dataset used in HW5 for classification, and in Part 2, we introduce new features based on citation numbers to enhance classification performance.

Part 1: Random Forest Classification with Original Features: For Part 1, we utilized a dataset previously employed in HW5 for classification. The original features consisted of citation numbers from 2017 to 2022. The initial model accuracy using six features was observed to be 0.65. However, this accuracy may be suboptimal due to potential issues such as imbalanced classes, insufficient feature representation, or noisy features.

Part 2: Feature Engineering for Improved Performance: In Part 2, we aimed to improve classification performance by introducing five new features derived from citation numbers. Each new feature represents the percentage change in citations from one year to the next, spanning the years 2017 to 2022. The formula used is provided in the problem statement.

These new features capture the temporal dynamics of citation growth, providing the Random Forest model with more nuanced information for classification.

3. Results:

After incorporating the five new features into the Random Forest model, the accuracy significantly improved to 1.0. This substantial increase suggests that the engineered features better capture the underlying patterns in the data, leading to improved classification performance. The new features provide a more comprehensive representation of the dataset, enabling the model to make more informed decisions.

4. Conclusion:

In conclusion, the Random Forest approach, when coupled with feature engineering, proves to be a powerful tool for classification tasks. The substantial improvement in accuracy from Part 1 to Part 2 highlights the importance of thoughtful feature selection and engineering in enhancing model performance. The newly introduced features, reflecting the percentage change in citations over time, contribute valuable information for the classification task.

References:

https://chat.openai.com/,
https://keras.io
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
https://towardsdatascience.com/machine-learning-classification-52241849468a
https://scikit-learn.org/stable/modules/model_evaluation.html