HW7 Report
Gina Roh

1. 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.)

The training score is perfect at 1.0, but the test score is significantly lower at 0.35. I've tried out various combinations of maximum depth and bootstrapping iterations to prevent overfitting, but it seems that the test scores consistently stay low.

Random forest models tend to be limited by the data they were originally trained on, and their performance can be poor when applied to datasets that fall outside the scope of their initial training data[1]. In this case, it appears that the random forest model may not be well-suited for the test dataset, which lies outside the range of the original training data. Additionally, the training dataset is widely distributed and sparse, contributing to the low performance.

One potential solution to this problem might be using pre-processed data. In task2, we utilize the ratio of changes in citation numbers compared to the previous year. This transformation can help make the dataset more compact and tightly distributed, leading to a substantial enhancement in the model's performance. We can observe significantly higher scores for both the training set and the test set, with both achieving a perfect score of 1.0.

```
-----part1-----
    univ_rank first_initial last_initial  cit_2017  cit_2018  cit_2019  cit_2020  cit_2021  cit_2022  category  predict
0          51             I            P        38       102       159       245       277       381         2        1
1          51             S            M       153       333       510       749       963      1048         1        2
2          51             A            B      5524      8950     12526     14204     16734     17508         1        0
3          51             W            H       161       183       206       215       179       262         2        0
4          51             F            N        70        96        88       133       157       156         0        2
5          51             M            I       238       386       641       602      1025      1249         2        2
6          51             R            F        41       115       210       312       473       554         2        1
7          51             S            J        54        72       113       139       144       141         0        1
8          51             S            Z       135        92       160       184       238       332         2        1
9          51             J            Z      1678      2066      2635      3253      4319      4125         0        0
10         52             E            G       151       147       156       152       169       167         0        0
11         52             M            C        85       121       202       264       376       383         0        1
12         52             W            E      1375      1264      1038       998       947       784         0        0
13         52             A            D       183       286       356       395       449       490         1        2
14         52             R            C        89       128       103       109       108       103         0        2
15         52             V            C        19        22        52       116       172       188         1        2
16         52             T            B       503       463       584       722       945       893         0        2
17         52             W            A        47        82        98       128       178       346         2        2
18         52             K            A       139       125        84        80        74        47         0        0
19         52             S            H       205       201       220       210       202       187         0        0
Feature_importance:  [0.15877299 0.15615642 0.16106573 0.14920135 0.14487029 0.22993322]
Training score:  1.0
Test score:  0.35
```

[1] Refer to https://blog.frankfurt-school.de/wp-content/uploads/2018/10/Neural-Networks-vs-Random-Forests.pdf

2. Part-2:

Introduce 5 new features based on the cita7on numbers. and use them in the RF instead of the citation numbers directly.

Each new feature is: ((citation number in year n+1)-(cita7on number in year n))/(cita7on number in year n) for 2016<n<2022

```
-----part2-----
    univ_rank first_initial last_initial  change18  change19  change20  change21  change22  category  predict
0          51             I            P  1.684211  0.558824  0.540881  0.130612  0.375451         2        2
1          51             S            M  1.176471  0.531532  0.468627  0.285714  0.088266         1        1
2          51             A            B  0.620203  0.399553  0.133961  0.178119  0.046253         1        1
3          51             W            H  0.136646  0.125683  0.043689 -0.167442  0.463687         2        2
4          51             F            N  0.371429 -0.083333  0.511364  0.180451 -0.006369         0        0
5          51             M            I  0.621849  0.660622 -0.060842  0.702658  0.218537         2        2
6          51             R            F  1.804878  0.826087  0.485714  0.516026  0.171247         2        2
7          51             S            J  0.333333  0.569444  0.230088  0.035971 -0.020833         0        0
8          51             S            Z -0.318519  0.739130  0.150000  0.293478  0.394958         2        2
9          51             J            Z  0.231228  0.275411  0.234535  0.327698 -0.044918         0        0
10         52             E            G -0.026490  0.061224 -0.025641  0.111842 -0.011834         0        0
11         52             M            C  0.423529  0.669421  0.306931  0.424242  0.018617         0        0
12         52             W            E -0.080727 -0.178797 -0.038536 -0.051102 -0.172122         0        0
13         52             A            D  0.562842  0.244755  0.109551  0.136709  0.091314         1        1
14         52             R            C  0.438202 -0.195312  0.058252 -0.009174 -0.046296         0        0
15         52             V            C  0.157895  1.363636  1.230769  0.482759  0.093023         1        1
16         52             T            B -0.079523  0.261339  0.236301  0.308864 -0.055026         0        0
17         52             W            A  0.744681  0.195122  0.306122  0.390625  0.943820         2        2
18         52             K            A -0.100719 -0.328000 -0.047619 -0.075000 -0.364865         0        0
19         52             S            H -0.019512  0.094527 -0.045455 -0.038095 -0.074257         0        0
Feature importance:  [0.09792896 0.05374935 0.05545764 0.10595865 0.6869054 ]
Training score:  1.0
Test score:  1.0
```