Jingxia Zhu (Jingxia6)

IE598 MLF F18

Module 7 Homework (Random Forest)

Using the Wine dataset, described in Raschka chapter 4 and 10 fold cross validation;

**Part 1: Random forest estimators**

Fit a random forest model, try several different values for N\_estimators, report in-sample accuracies.

|  |  |
| --- | --- |
| N\_estimators | In-sample accuracy |
| 1 | ﻿0.90 (+/- 0.13) |
| 2 | ﻿0.86 (+/- 0.20) |
| 5 | ﻿0.96 (+/- 0.10) |
| 10 | ﻿0.97 (+/- 0.07) |
| 15 | ﻿0.98 (+/- 0.06) |
| 25 | ﻿0.97 (+/- 0.07) |
| 50 | ﻿0.97 (+/- 0.07) |
| 100 | ﻿0.97 (+/- 0.07) |
| 250 | ﻿0.97 (+/- 0.07) |
| 500 | ﻿0.97 (+/- 0.07) |
| 1000 | ﻿0.97 (+/- 0.07) |

**Part 2: Random forest feature importance**

Display the individual feature importance of your best model in Part 1 above using the code presented in Chapter 4 on page 136. {importances=forest.feature\_importances\_ }

﻿ ﻿ 1) Color intensity 0.199674

2) Flavanoids 0.194217

3) Alcohol 0.160306

4) Proline 0.127067

5) OD280/OD315 of diluted wines 0.091433

6) Hue 0.051251

7) Total phenols 0.042253

8) Alcalinity of ash 0.033179

9) Nonflavanoid phenols 0.030194

10) Malic acid 0.023970

11) Magnesium 0.021691

12) Proanthocyanins 0.015880

13) Ash 0.008885



**Part 3: Conclusions**

Write a short paragraph summarizing your findings. What is the relationship between n\_estimators, in-sample CV accuracy and computation time? What is the optimal number of estimators for your forest? Which features contribute the most importance in your model according to scikit-learn function? What is feature importance and how is it calculated? (If you are not sure, refer to the Scikit-Learn.org documentation.)

Overall, in-sample CV accuracy increases when the n estimator goes up. However, in terms of running time, the higher the n\_estimator is, the longer the time is spent on computing. The optimal number of estimators for my forest is 15, considering both accuracy and computational complexity. Color intensity contributes the most importance in my model. Feature importance measures the relevant importance among different features. It is calculated as the mean decrease impurity over all trees in the random forest of ensemble.

**Part 4: Appendix**

Link to github repo: https://github.com/jzhuuhzj/IE598\_F18\_HW7.git