SIT384 Cyber security analytics

High Distinction Task 6.3HD: Comparing Classification Models for a Dataset

Task description:

You are asked to apply different classification techniques to a given Spambase dataset, to potentially enhance the accuracy of the learnt models via selecting better parameters, and/or pre-processing etc., to compare the results, and to summarize your findings in a report.

The Spambase Data Set can be retrieved from https://archive.ics.uci.edu/ml/datasets/Spambase or be directly downloaded from task resources.

The classification algorithms to apply are:

1. Decision Trees
2. Random Forest
3. Support Vector Machines

You must compare and interpret the results of using different approaches for the dataset. Other requirements are:

- Classification models that achieve higher accuracies will get more points.
- In your report after comparing the experimental results, write a paragraph or two trying to explain/speculate why, in your opinion one classification algorithm outperformed the others.
- Include a brief discussion in your report, how you have selected the parameters of particular data mining algorithms.
- Finally, at the end of your report provide a 1-2 paragraphs summary that summarizes the most important findings of this task

The github website of the prescribed textbook has quite some useful supplemental material (**code examples**, IPython notebooks, etc.), available at https://github.com/amueller/introduction_to_ml_with_python, especially chapter 2.

Submission:

Submit the following files to OnTrack:

1. Your program source code (e.g. task6_3.py)
2. A screen shot of your program running
3. Your comparison and discussion report

Check the following things before submitting:

1. Add proper comments to your code