Code Homework 1
Student: Nguyen Van Hoang
Student No.: 21025029

Exercise 2:
* Method: **Random Forest**
* Parameters used for training **RandomForestClassifier**:

- **criterion="gini"** for faster training time. Calculating Gini Index is less computationally expensive than calculating Entropy because it uses **logarithms**. The obtained results using the **"entropy"** criterion are slightly better, but it is not worth the time invested for training when using this criterion.
- **n_jobs=-1** (default: **None**) to use all processors of our physical machine, which helps improve training time => This will have great impact if we have a large dataset.
- **max_depth=5** => Setting a specific **max_depth** of each tree helps fighting with overfitting, instead of using default config that leads to full tree. Here I try to set it to 5 (corresponding with 5 features) and observe that it produces a good result, the bigger **max_depth** I set, the lower accuracy I get.

```
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 6 concurrent workers.
[Parallel(n_jobs=-1)]: Done  38 tasks      | elapsed:    0.2s
[Parallel(n_jobs=-1)]: Done 100 out of 100 | elapsed:    0.5s finished
[Parallel(n_jobs=6)]: Using backend ThreadingBackend with 6 concurrent workers.
[Parallel(n_jobs=6)]: Done  38 tasks      | elapsed:    0.0s
Accuracy:  0.879
F1 score:  0.7363834422657952
[Parallel(n_jobs=6)]: Done 100 out of 100 | elapsed:    0.0s finished
```

*With **n_jobs=-1**, we are using all processors and the training time is improved.*

```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    1.0s finished
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    0.0s finished
Accuracy:  0.879
F1 score:  0.7363834422657952
```

*With default **n_jobs**, we use only one processor to train the model and the elapsed time is higher.*

***Comparison:**

Classifier	Default RandomForestClassifier	RandomForestClassifier(**criterion**="gini", **max_depth**=5, **n_jobs**=-1, **verbose**=1)
Accuracy	0.838	0.879
F1 score	0.6553191489361702	0.7363834422657952