SEOULTECH
Seoul National University of Science & Technology

Assigment 1

20102122 정효안

Code	ITM 526		Title	Business Analytics		
Type	Assignment	Questions	1	Weighting	5%	

1. Build classification models → (1), (2), (3), (5): please refer to the code

(1) Load the dataset (10pts)

(2) Split the dataset (Training / Validation) (10pts)

(3) Train the models while changing some hyperparameters (30pts)
- Use two different learning algorithms you know.

(4) Describe the meaning of the hyperparameters you adjusted.(10pts)
- min_samples_leaf: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches.

- max_depth: The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Too large max_depth can cause overfitting.

- C: Inverse of regularization strength; must be a positive float. Like in support vector machines, smaller values specify stronger regularization.

(5) Performance tables for training set / validation set according to hyperparameter settings (30pts)
- Performance metric: accuracy

	min_samples_leaf	training accuracy	validation accuracy
0	1	1.000000	0.937063
1	2	0.990610	0.937063
2	5	0.969484	0.930070
3	7	0.964789	0.881119
4	10	0.957746	0.923077
5	20	0.929577	0.888112

	max_depth	training accuracy	validation accuracy
0	2	0.955399	0.895105
1	3	0.967136	0.888112
2	4	0.978873	0.895105
3	5	0.988263	0.944056
4	6	0.995305	0.937063
5	7	0.997653	0.951049
6	8	1.000000	0.937063

	C	training accuracy	validation accuracy
0	0.01	0.748826	0.769231
1	0.10	0.938967	0.937063
2	1.00	0.967136	0.979021
3	10.00	0.981221	0.979021
4	100.00	0.988263	0.965035
5	1000.00	0.990610	0.965035
6	10000.00	0.997653	0.937063

(6) Find the best hyperparameters for each learning algorithms (10pts)

→Learning algorithms used (- best hyperparameter)

1. Decision Tree
   - min_samples_leaf: 2
   - max_depth: 7
2. Logistic Regression
   - C: 10.0

   →This is because the validation accuracy is the highest in these cases. (The ultimate goal of training is to get a high score on test set.)