1. 데이터가 아래와 같이 split 되었을 때 information gain을 계산하시오
2. AUC가 0.5일 경우 classifier의 예측능력이 전혀 없다고 볼 수 있다 (참/거짓)
3. Data Science는 고객들의 행위를 분석하기 위해서 데이터로부터 숨겨진 패턴을 찾아내는데 주로 사용된다 (참/거짓)
4. 분류와 회귀분석은 둘 다 supervised data mining tasks이다 (참/거짓)
5. Logistic regression은 연속적인 변수의 값을 예측하는 classification 문제에 사용된다
6. total expected profit을 계산하시오
7. Data Science는 저장매체에 저장된 빅데이터를 access하고 처리하기 위해서 ㅇ 기술을 이용한다
8. 어떤 event가 일어날 확률이 0.5이면 odds 값은 1이다
9. SVM은 ㅇ 함수를 이용하여 margin을 벗어나는 잘 못 분류된 instance에 대해서 margin으로부터 거리에 비례해서 penalty를 부여한다
10. K-NN모델은 training이 매우 빠르므로 실시작 application 에 적합하다 (참/거짓)
11. Entropy의 최소값은 항상 0이다 (참/거짓)
12. 회귀분석 문제를 있는대로 고르시오
13. TF는 corpus에서 어떤 용어 또는 단어가 얼마나 자주 나타나는지 빈도수를 나타낸다 (참/거짓)
14. 어떤 classifier가 완벽하게 분류한다고 가정했을 때 고객 리스트들 중 50%만 타겟한다고 할 때 이 classifier의 lift값은?
15. Jaccard distance값을 구하시오
16. CRISP-DM의 여러 단계들 중에서 data mining 기법을 데이터에 적용하는 첫 단계는?
17. Naive Bayes classifier를 이용하여 spam으로 분류될 확률을 구하시오
18. Big data의 4가지 특징