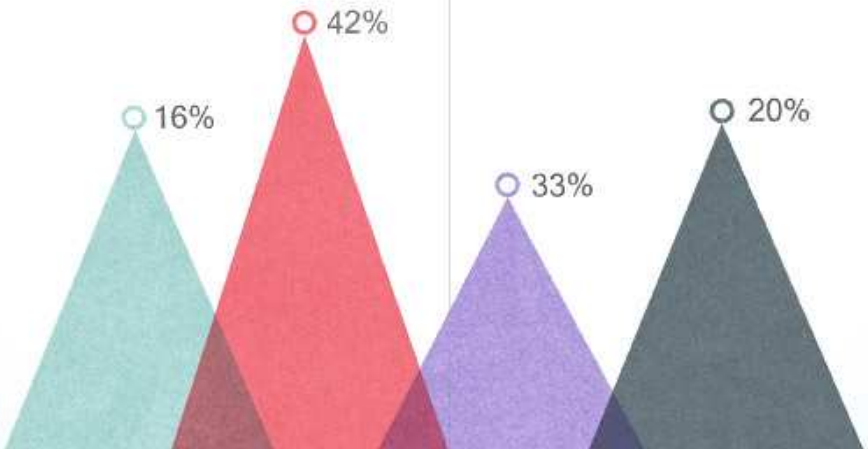
와이즈인컴퍼니
Wisein Company

과제

[데이터 사이언스 과정]
프로젝트 과제



와이즈인컴퍼니 WiseIN Company

공지 및 프로젝트 소개

1. 프로젝트 과제는 **데이터사이언스 자격증 발급과 관련된 중요한 종합과제**입니다.

2. 본 과제는 실제 와이즈인컴퍼니에서 수행한 분석 데이터를 기반으로 합니다. 따라서 **학습 목적 이외에 데이터의 외부 유출을 엄격히 금합니다.**

3. 모듈2 머신러닝의 종합과제는 특허 전략 예측 데이터로서 와이즈인컴퍼니가 진행했던 특허분석 사업 데이터입니다.

4. 정답이 정해져 있는 것이 아니기때문에 세부 과제별로 최선의 방법을 모색하고, 학습자의 판단에 따라 분석결과를 정리해야 합니다. 평가 역시 정답이 아닌 분석 로직(logic)과 분석 효율성을 중심으로 하며, 학습 이외의 내용이 과제로 제출될 수 있습니다. 학습 이외의 과제는 다양한 정보(구글링 등)를 통합하여 해결하시기 바랍니다.

5. 본 과제 제출마감은 **2019년 7월 31일(수요일)**까지입니다. 과제물은 wise@wiseinc.co.kr로 제출 바랍니다.

- 제출 과제물: 1) 결과물 정리 ppt 2) python 코드문서 3) 분석에 적용된 데이터파일

6. 본 과제의 풀이는 화상강의가 아닌 별도의 영상강의를 추후 제공할 예정입니다 (홈페이지 강좌 업로드)

1 과제

1. 데이터구조 및 과제1

머신러닝과제_특허 데이터.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	출원번호	대분류	피인용횟수	전체 피인용 평균	해당분야 피인용수 평균	특허 영향력	전체 대상특허 건수	해당분야 특허건수	특허 집중도	해당분야 특허건수	해당분야 최근3년 출원특허	최근 특허 활동력	패밀리국가수	전체 평균 패밀리국가수	해당분야 평균 패밀리국가수	시장 확보력	소분류별 특허건수	기술이전 특허건수	해당분야 기술이전 특허건수	사업화 성공률	사분면
714	1020130078499	전기/전자	0	0.55	0.00	0.00	733	7	0.01	7	5	0.71	2	1.72	2.71	1.58	7	0	0	0.00	4사분면
715	1020130015090	전기/전자	0	0.55	0.00	0.00	733	7	0.01	7	5	0.71	4	1.72	2.71	1.58	7	0	0	0.00	4사분면
716	1020120140570	전기/전자	0	0.55	0.00	0.00	733	7	0.01	7	5	0.71	1	1.72	2.71	1.58	7	0	0	0.00	4사분면
717	1020090060748	전기/전자	0	0.55	0.00	0.00	733	7	0.01	7	5	0.71	3	1.72	2.71	1.58	7	0	0	0.00	4사분면
718	1020110021039	화학	0	0.55	0.33	0.60	733	6	0.01	6	0	0.00	1	1.72	2.00	1.16	6	0	4	0.67	4사분면
719	1020100038385	화학	1	0.55	0.33	0.60	733	6	0.01	6	0	0.00	4	1.72	2.00	1.16	6	1	4	0.67	4사분면
720	1020100001923	화학	1	0.55	0.33	0.60	733	6	0.01	6	0	0.00	1	1.72	2.00	1.16	6	1	4	0.67	4사분면
721	1020110019456	화학	0	0.55	0.33	0.60	733	6	0.01	6	0	0.00	1	1.72	2.00	1.16	6	1	4	0.67	4사분면
722	1020110023975	화학	0	0.55	0.33	0.60	733	6	0.01	6	0	0.00	1	1.72	2.00	1.16	6	0	4	0.67	4사분면
723	1020100001447	화학	0	0.55	0.33	0.60	733	6	0.01	6	0	0.00	4	1.72	2.00	1.16	6	1	4	0.67	4사분면
724	1020120117320	화학	0	0.55	0.60	1.08	733	5	0.01	5	1	0.20	4	1.72	3.20	1.86	5	0	0	0.00	4사분면
725	1020110126080	화학	0	0.55	0.60	1.08	733	5	0.01	5	1	0.20	1	1.72	3.20	1.86	5	0	0	0.00	4사분면
726	1020110123217	화학	0	0.55	0.60	1.08	733	5	0.01	5	1	0.20	1	1.72	3.20	1.86	5	0	0	0.00	4사분면
727	1020100114007	화학	2	0.55	0.60	1.08	733	5	0.01	5	1	0.20	5	1.72	3.20	1.86	5	0	0	0.00	4사분면
728	1020100114008	화학	1	0.55	0.60	1.08	733	5	0.01	5	1	0.20	5	1.72	3.20	1.86	5	0	0	0.00	4사분면
729	1020120044074	화학	1	0.55	2.17	3.91	733	6	0.01	6	1	0.17	1	1.72	1.33	0.78	6	0	1	0.17	4사분면
730	1020080132341	화학	1	0.55	2.17	3.91	733	6	0.01	6	1	0.17	1	1.72	1.33	0.78	6	0	1	0.17	4사분면
731	1020097009556	화학	0	0.55	2.17	3.91	733	6	0.01	6	1	0.17	3	1.72	1.33	0.78	6	1	1	0.17	4사분면
732	1020060126557	화학	7	0.55	2.17	3.91	733	6	0.01	6	1	0.17	1	1.72	1.33	0.78	6	0	1	0.17	4사분면
733	1020070026163	화학	0	0.55	2.17	3.91	733	6	0.01	6	1	0.17	1	1.72	1.33	0.78	6	0	1	0.17	4사분면
734	1020060019595	화학	4	0.55	2.17	3.91	733	6	0.01	6	1	0.17	1	1.72	1.33	0.78	6	0	1	0.17	4사분면
735																					

- 모 기관에서 출원 및 등록한 745건이다.
- 입력변수: 피인용건수 ~ 사업화성공률
- 분류범주: 사분면 (1~4)

1 과제

2. 과제

- 학습데이터와 테스트데이터를 7:3의 비율로 나누고, 정규화 과정을 진행한 후 과업을 진행하시오
- 먼저 KNN, 로짓, SVM, 의사결정나무분석, 랜덤포테스트 등 개별 알고리즘으로 예측력이 가장 높은 하이퍼파라미터를 결정하시오
- 이후 투표기반 앙상블을 적용하여 예측분류율을 높이십시오. hard. & soft voting을 각각 비교해서 가장 높은 모델로 결정하시오

1 과제

2. 과제

과제1(배점 15점)

KNN 모델 적용 최종 결과 (예측분류율 및 최적의 하이퍼 파라미터)

결과물 제시

1 과제

2. 과제

과제2(배점 15점)

로지스틱모델 적용 최종 결과 (예측분류율 및 최적의 하이퍼 파라미터)

결과물 제시

1 과제

2. 과제

과제3(배점 15점)

의사결정나무 모델 적용 최종 결과 (예측분류율 및 최적의 하이퍼 파라미터)

결과물 제시

1 과제

2. 과제

과제4(배점 15점)

랜덤포레스트 모델 적용 최종 결과 (예측분류율 및 최적의 하이퍼 파라미터)

결과물 제시

1 과제

2. 과제

과제5(배점 40점)

앙상블 모델 적용 최종 결과 (예측분류율 및 최적의 하이퍼 파라미터)

결과물 제시

수고하셨습니다!

와이즈인컴퍼니