[박사학위논문 Proposal]

kNN 알고리즘을 활용한 선호직무 수강과목 추천 시스템

A Study on the Recommendation System for Preferred Job Courses Using the kNN Algorithm

2021년 9월 25일

지도교수 안 태 호 교수님

박사과정 전 정 우

숭실대학교 대학원 프로젝트경영학과

Contents

제1장 서론

1. 연구의 배경
2. 연구의 목적
3. 연구의 범위

1.1 연구의 배경 (1/3)

- 졸업하고 취직에 성공한 직장인 중 19.1%는 취업 후 1년 이내 이직 등의 사유로 일자리가 변경되었으며(고등교육기관 졸업자 일자리 이동통계, 2018), 이는 일자리(직무) 매칭이 잘못되었음을 단적으로 보여주는 사례임. 취업 포털 사람인이 941개 기업 대상으로 실시한 조사에서도, 71.2%의 기업이 구인기업-취업자 간 일자리 미스 매칭 경험이 있다고 응답함.
- 이렇듯 일자리 미스 매칭은 구인·구직자 모두에게 일어날 수 있는 사회적 문제로, 이로 인한 사회적 비용 발생도 상당할 것으로 보임.
- 이러한 일자리 미스 매칭 문제를 해소하기 위해서는 대학 교육 단계에서 부터 학생들에 적합한 진로 탐색을 원활히 할 수 있고, 그에 따른 일자리 및 직무 매칭이 이루어 질 수 있는 대학 교육 지원 시스템이 필요해 보임.
- 조사결과에 따르면 졸업 후 진로로 취업을 계획하고 있으나(70.3%), 자신에게 맞는 진로·일자리를 찾지 못해 졸업 후의 진로가 가장 큰 고민거리라고 답하는 대학생이 많고(57.9%), 무엇을 하고 싶은지 몰라서 진로를 결정하지 못하고 있다는 조사 결과(36.4%)를 보면, 여전히 상당수의 대학생들이 진로 탐색에 어려움 때문에 자신에 맞는 제대로 된 진로 설정을 하지 못하고 있는 실정임(대학 진로교육 현황조사, 2018 ; 한국직업능력개발원, 2019) .

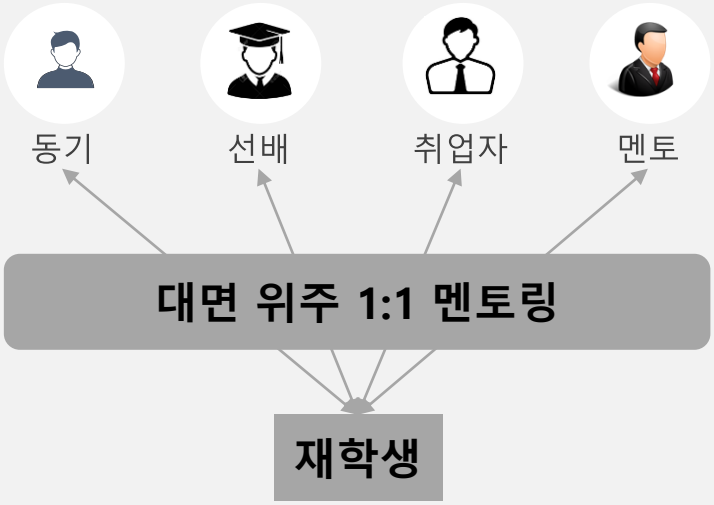
1.1 연구의 배경 (2/3)

- 대학생들의 직무 및 진로 탐색 시, 전공 및 교양 과목의 선택은 중요한 역할을 함.
- 그러나 산업 패러다임 변화에 따라 기업의 직무 역량은 고도화되고, 이에 따른 수강 과목의 분야와 수가 증가하고 있어 학생들이 어떠한 수강 과목을 선택해야 하고, 자신의 선호 직무에 맞는 수강 과목을 어떻게 선택해야 할지 어려움을 겪고 있음.
- 또한 최근 코로나 사태로 인한 비대면 사회로의 전환으로 대학 취업지원 담당자 및 교수, 선배 등과의 대면 상담이 어려워져 적합한 수강 과목 선택 및 이를 통한 진로 탐색의 어려움이 가중되고 있음. (코로나 학번의 빼앗긴 봄)
- 이에, 비대면 상황 하에서도 활용 가능하고, 산업 환경 변화에 대응할 수 있는 취업역량 강화 까지 가능한 선호 직무 중심의 수강과목 추천 시스템 및 이를 개발할 수 있는 연구가 필요함.
- 현재, 학업 만족도를 위한 교양과목 추천 중심으로 한 수강과목 추천 연구는 다수 존재 하지만, 선호 직무를 고려한 수강과목 추천과 관련된 선행 연구는 미흡함.

1.1 연구의 배경 (3/3)

"취업자 **직무정보** 및 **수강이력**을 기반으로 **선호직무별 수강과목 추천**"

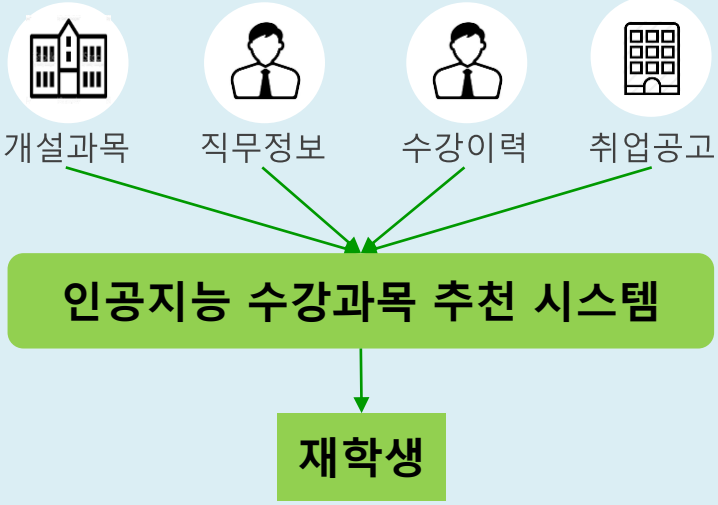
현행 : AS-IS (지인의 경험에 의존) 감성적 의사결정



문제점

- 제한적 멘토 섭외로 진로 탐색의 범용성 한계
- 사례 위주의 경험 정보 및 구두 전달로 정보 왜곡
- **단발적인 멘토링으로 정보의 파편화 현상 발생**

향후 : TO-BE (취업자의 수강이력 기준) 과학적 의사결정



기대효과

- 대면 멘토링 대비 비용 최소화. **경제적 이익**
- 군집화된 직무별 정보 취득. **타 직무정보 취득**
- 직무 환경변화에 따른 **신속한 재조정 가능**

1.2 연구의 목적

- 본 연구는 강의 만족도 향상을 위한 수강과목 추천이 아닌 취업자의 직무정보와 수강이력 정보를 활용하여 취업역량 강화를 위해 재학생의 선호직무에 대한 인공지능 기반의 추천시스템을 구축하고자 함
- 본 연구는 특정 대학의 수강이력 정보와 강의 만족도 정보를 중심으로 한 교양과목 추천 연구의 한계점을 극복하고, 대학생의 취업역량 강화를 위한 선호직무 중심의 수강과목 추천시스템을 구축하여 취업 포털에 베타 서비스를 게시하고자 함

  1. 취업자의 직무정보를 도출
  2. 취업자의 수강과목을 도출
  3. 1과 2를 매칭하여 학습 데이터셋 구성
  4. 재학생의 선호직무에 맞는 인공지능 수강과목 추천 시스템 구축
  5. 추천 시스템을 포털에 베타 서비스 개시

- 본 연구의 효과

실용적 가치		학술적 가치
선호직무에 맞는 개인화 된 수강과목 추천 미스매칭률 감소를 통한 사회적 비용 절감	⊕	강의 만족도 향상을 위한 수강과목 추천에서 취업역량 강화를 위한 선호직무 기반 추천 연구

1.3 연구의 범위

선행연구의 분석	● 추천 시스템에 적용된 인공지능 연구방법의 문헌조사
연구방법 및 연구절차 제시	● 학생의 선호직무, 취업자의 직무정보 및 수강이력 간의 관계를 바탕으로 연구방법 설정
데이터 수집 및 데이터 전처리	● 비식별화 및 가명처리 된 연구 목적의 수강과목 추천 원천 데이터 구성
Feature 도출 및 데이터셋 구성	● 원천데이터로부터 대표 Feature 도출을 통한 학습 데이터셋 구성
알고리즘 매칭률 실험	● 하이퍼 파라미터 조정을 통한 인공지능 알고리즘의 손실률 및 정확도 실험
알고리즘 재학습 및 서비스 구축	● 수강과목 추천에 최적화된 인공지능 알고리즘의 반복적 재학습 및 검증 ● 취업 포털과 증명발급 포털에 수강과목 추천 서비스 개시
결과분석 및 결론	● 연구결과 요약 및 분석을 통한 결론 도출 ● 연구의 시사점, 한계점 및 향후 연구과제 제시

제2장 이론적 배경

1. 추천 시스템
2. 협업 필터링(C.F)
3. 인공지능 알고리즘
4. 콜드 스타트 문제

2.1 추천 시스템 (1/2)

- (추천시스템의 등장) 아마존닷컴은 1997년 사용자 구매이력 데이터로 특정 서적을 구매했던 고객들이 선호하는 서적을 추천함으로써, 각 고객별로 맞춤화된 서적 목록을 추천하는 시스템을 제공하였음.
  아마존이 추천시스템의 태동을 알렸다면, 넷플릭스는 Netfilx Prize를 통해서 1등 100만 달러 상금으로 추천시스템의 저변을 세계적으로 확대하였다는 데 그 의의가 있음.

- (추천시스템의 종류) 개인화의 판단 주제에 따라 **설정형 추천**과 **학습형 추천**으로 구분됨.

- (설정형 추천) 사용자가 자신의 판단에 따라 사전 설정에 의해 추천
  - 구글의 이메일, RSS, 캘린더, 검색 등 사용자가 설정하는 기준에 따라 구성됨.
  - 사용자가 스스로 설정된 정보만 제공되는 형태로 추가적인 연구가 진행되지 못함.

- (학습형 추천) 사용자의 인구통계학적 정보를 물론 구매 여부 등을 분석하여 추천
  - 각 사용자가 어떠한 콘텐츠를 선호하는지 학습하여야 함.
  - 각 사용자의 선호도와 유사한 선호패턴을 갖는 집단인 Neighbor를 결정해야 함.
  - 사용자가 콘텐츠를 어느 상황에서 소비하는지 파악함으로써 적절한 추천 시점을 결정함.

  ※ [학습형 추천의 적용 시나리오]
  학기를 마치고 성적증명서를 발급하는 재학생에게 전공학과 또는 선호직무를 중심으로 수강과목 추천 안내서를 출력할 수 있도록 함

2.1 추천 시스템 (2/2)

- (추천시스템의 알고리즘) 협업적 필터링(collaborative filtering), 내용기반 필터링(content based filtering), 하이브리드 기법(hybrid method)으로 구분됨.

- 협업적 필터링(collaborative filtering) : 각 사용자별로 그 사용자의 가장 선호정보가 비슷한 사용자들을 이웃으로 찾은 다음, 이웃들이 선호하는 상품을 추천하는 것임.
  (Pearson correlation coefficient, constrained Pearson correlation coefficient, Jaccard coefficient, cosine vector).
  - (수동화) 2000년 중반까지 사용자가 접한 뉴스 또는 음악의 선호를 사용자가 직접 평가한 데이터 및 구매여부 데이터를 바탕으로 협업적 필터링에 대한 연구가 이루어짐.
  - (자동화) 2000년 후반부터는 전자상거래에서 사용자의 행동 및 탐색 패턴을 고려하여 사용자가 선호도를 직접 입력하지 않더라도, 사용자의 이용 패턴을 통해 추천된 선호도를 활용함으로써 자동화된 협업적 필더링 기법이 연구됨.

- 내용기반 필터링(content based filtering): 각 사용자 별 특정 콘텐츠 또는 상품의 속성을 학습하는 것을 기반으로 함. 상품의 내용과 사용자가 요구하는 정보간의 유사도를 계산하여 그 결과를 순위화하여 나타냄 (가중치 기법, 적합성 피드백, 확률검색 모형).
  ※ 모든 콘텐츠에 대해 등록자가 수작업으로 분류해야 하는 번거로움을 극복하기 위하여, 최근에는 텍스트마이닝을 활용하여 문서 내에 문장들로 부터 해당 문서의 내용을 정확히 표현하는 키워드를 추출하는 자동화된 내용기반 필터링 기법도 연구되고 있음.

2.2 인공지능 알고리즘 (1/4)

- 인공지능 알고리즘 비교

	랜덤포레스트	K-NN	가우시안 나이브베이즈
장점	분류, 회귀 사용 가능	분류, 회귀 사용	정규 분포 / 조건부 확률 계산
	결측치 다루기 쉬움	단순하고 효율적임	단순하고 빠르며 효과적임
	대용량 데이터 처리 좋음	훈련 단계가 빠름	노이즈,결측치있어도잘처리함
	오버피팅 회피, 정확도 향상	수치 분류작업 성능 좋음	많은 예제도 잘 처리함
	중요한 변수 선정 가능	-	예측 추정 확률 얻기 쉬움
단점	데이터 수↑,속도가 떨어짐	특징-클래스관계이해제한적	독립적 결함 가정에 의존
	결과 해석 어려움	데이터 ↑, 분류 느려짐	수치 데이터에 좋지 않음
	-	적절한 K 선택 중요	예측된 범주보다 덜 신뢰적
	-	추가적인 전처리 필요	-

2.2 인공지능 알고리즘 (2/4)

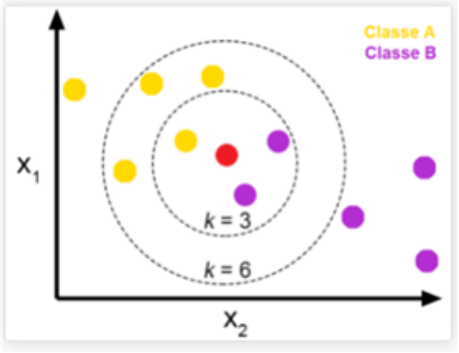
- KNN (K Nearest Neghbor)

  • 패턴 인식에서, KNN은 분류나 회귀에 사용되는 비모수 방식으로, 데이터가 주어지면 그 데이터의 주변을 살펴본 뒤 더 많은 데이터가 포함되어 있는 범주로 분류하는 방식임.
  거리 기반으로 분류하는 '클러스터링'과 유사한 개념이긴 하나, 기존 관측치의 Y값(Class)가 존재한다는 점에서 비지도학습에 해당하는 '클러스터링'과 차이가 있음.

  • 수식 :
  일반적으로 점과 점 사이의 거리를 구하는 방법(유클리드 거리)

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$



  • 장·단점:
  단순하고 효율적임. / 훈련 단계가 빠름. / 수치 기반 데이터 분류 작업에서 성능이 우수함.
  모델을 생성하지 않아 특징과 클래스간 관계를 이해하는데 제한적임.
  데이터가 많아지면 분류가 느려짐. / 적절한 K의 선택이 중요함.
  명목 특징 및 누락 데이터를 위한 추가 처리가 필요함.

2.3 인공지능 알고리즘 (3/4)

- 랜덤포레스트(RandomForest)

• 기계 학습에서의 랜덤포레스트(RandomForest : 이하 랜덤포레스트)는 분류, 회귀 분석 등에 사용되는 앙상블 학습 방법의 일종으로, 훈련 과정에서 구성한 다수의 결정 트리로 부터 분류 또는 회귀 분석을 출력함으로써 동작함.

• 수식 :
랜덤포레스트의 모든 트리들은 독립적으로 훈련 단계를 거침.
테스트 단계에서 데이터 포인트 v는 모든 트리에 동시로 입력되어 종단 노드에 도달하게 됨.
이러한 테스트 단계는 병렬적으로 진행될 수 있으며 높은 계산 효율성을 얻을 수 있음.
랜덤포레스트의 예측 결과는 모든 트리의 예측 결과들의 평균을 얻음.
다른 방법으로는 트리들의 결과들을 곱하는 방법이 있음.

$$p(c|\mathbf{v}) = \frac{1}{T}\sum_{t=1}^{T} p_t(c|\mathbf{v}) \qquad p(c|\mathbf{v}) = \frac{1}{Z}\prod_{t=1}^{T} p_t(c|\mathbf{v}),$$

• 장·단점:
분류(Classification) 및 회귀(Regression) 문제에 모두 사용. / 결측치를 다루기 쉬움.
분류(Classification) 모델에서 상대적으로 중요한 변수를 선정 및 Ranking이 가능함.
대용량 데이터 처리에 효과적임. /
모델의 노이즈를 심화시키는 오버피팅 문제를 회피하여, 모델 정확도를 향상 시킴.
데이터 수가 많아지면 속도가 떨어짐. / 결과 해석의 어려움.

2.3 인공지능 알고리즘 (4/4)

- 나이브베이즈 (Naïve Bayes)

• 확률 기반 머신러닝 분류 알고리즘으로 데이터를 나이브(단순)하게 독립적인 사건으로 가정하고, 이 독립 사건들을 베이즈 이론에 대입시켜 가장 높은 확률의 레이블로 분류를 실행함.
가우시안 나이브베이즈 : 설명변수가 연속형 변수일 때 사용함. 적은 데이터로도 효율적인 성능을 낼 수 있음. 특징들의 값들이 정규 분포되어 있다는 가정 하에 조건부 확률을 계산함.

• 수식 :
트레이닝 데이터가 연속적인 속성 x를 포함하는 것으로 가정하면, 먼저 클래스에 따라 데이터를 나눈 뒤에, 각 클래스에서 x의 평균과 분산을 계산함. 클래스 c와 연관된 x 값의 평균을 μ_c라 하고, 분산을 σ_c^2 라고 하면, 주어진 클래스의 값들의 확률 분포가 M과 S로 매개변수화되어 정규분포식을 통해 계산될 수 있음.

$$p(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(v-\mu_c)^2}{2\sigma_c^2}}$$

• 장·단점:
단순하고 빠르며 매우 효과적임. / 노이즈와 결측 데이터가 있어도 잘 수행함. / 예측에 대한 추정된 확률을 얻기 쉬움. / 예측에 대한 추정된 확률을 얻기 쉬움.
모든 속성은 동등하게 중요하고 독립적이라는 알려진 결함 가정에 의존함.

※ 베이즈 이론 : 어떤 사건 B가 일어났을 때 사건 A가 일어날 확률 / 어떤 사건 A가 일어났을 때 사건 B가 일어날 확률 / 어떤 사건 A가 일어날 확률

2.4 콜드 스타트 문제

- 협업 필터링을 통하여 추천할 때, 가장 큰 문제는 콜드 스타트 문제임 .
  협업 필터링의 사용자 기반 알고리즘은 대상 사용자의 데이터를 기반으로 점수를 계산하고 최근접 이웃법을 통하여 이웃을 탐색함.

- 추천시스템의 초기 상태에서 대상 사용자의 데이터가 없을 때 콜드 스타트 문제가 발생함.

- 데이터가 많지 않을 경우에는 각 사용자의 선호 정보가 유사한 이웃을 찾기 어려워 유용하지 못할 뿐만 아니라, 사용자 수가 많을 경우 연산속도가 매우 느려지는 한계가 있음.

- 본 연구에서는 이 문제를 해결하기 위하여 취업 포털의 5년간의 직무정보 및 수강이력정보를 바탕으로 추천해 줌으로써 콜드 스타트 문제를 해결하고자 함.

  ※ (차원감소 기법) 일반적인 연구 방법에는 콜드 스타트 문제를 해결하는 방법
  사용자가 평가하지 않은 아이템의 초기 선호도 값을 평균값 등으로 모두 채운 다음 추정 작업을 진행하므로 모든 아이템에 대해 선호도를 예측할 수 있음.

제3장 연구 방법

1. 연구 절차
2. 연구 데이터
3. 학습 데이터셋
4. 알고리즘 실험
5. 시스템 구성도

3.1 연구 절차

- 다음과 같은 단계로 연구 진행

- 1. 원천 데이터 수집 및 전처리
  - 개설과목 : 국내 100개 대학의 개설과목 수집 (최대 230개 대학)
  - 직무정보 : 취업 포털의 이력서에 기록된 경력정보 수집 (스카우트 포털)
  - 수강과목 : 이력서에 첨부된 성적증명서의 과목명 또는 발급된 증명서의 과목명 수집
  - 선호직무 : 취업 포털의 직무명을 기준으로 한 채용 공고명 군집화 처리
  - 전처리 : 개인정보 비식별화 및 가명 처리

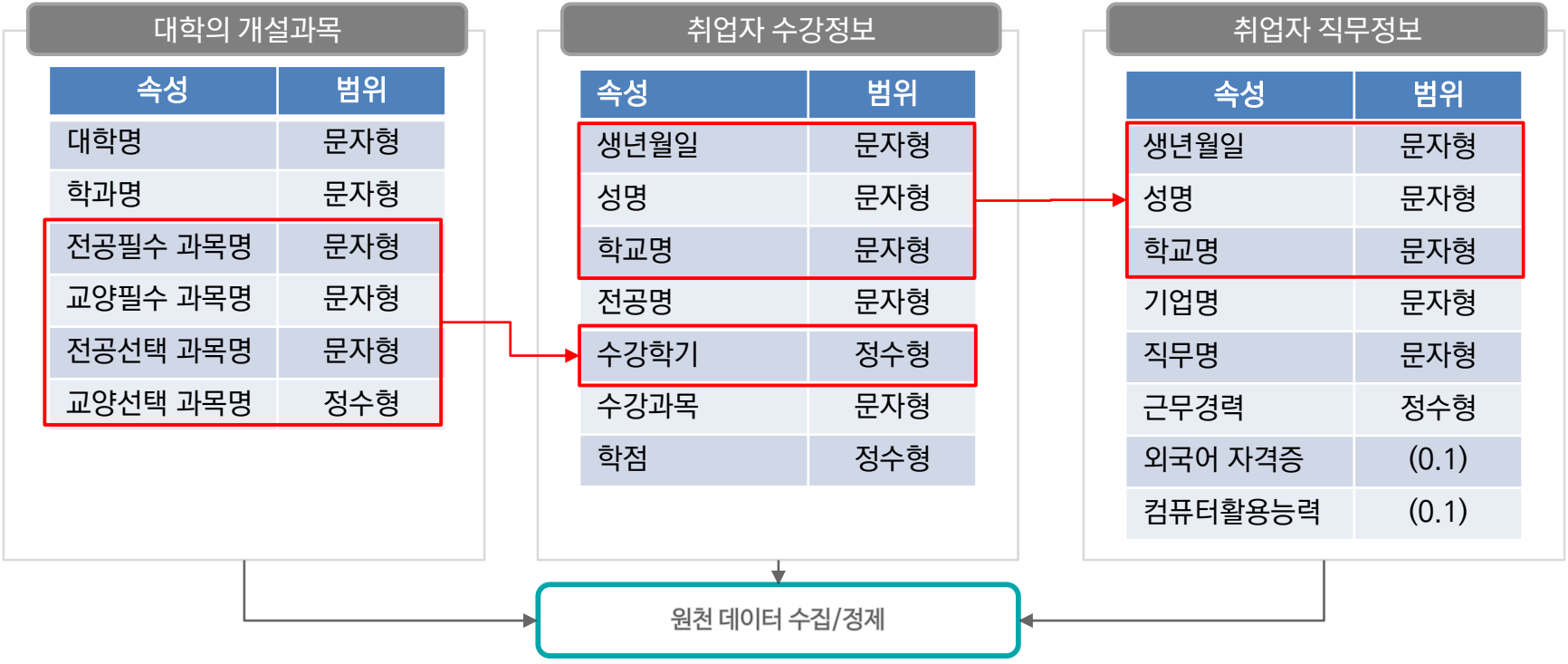
  2. Feature Engineering을 통한 인공지능 학습 데이터셋 구성
  3. kNN 알고리즘을 적용한 학습의 손실율과 정확도 산출 및 추천 시스템 구축
  4. 취업 포털(스카우트)과 대학 증명서발급 포털(웹민원센터)에 서비스 개시

- 연구 방법 절차도

3.2 연구 데이터

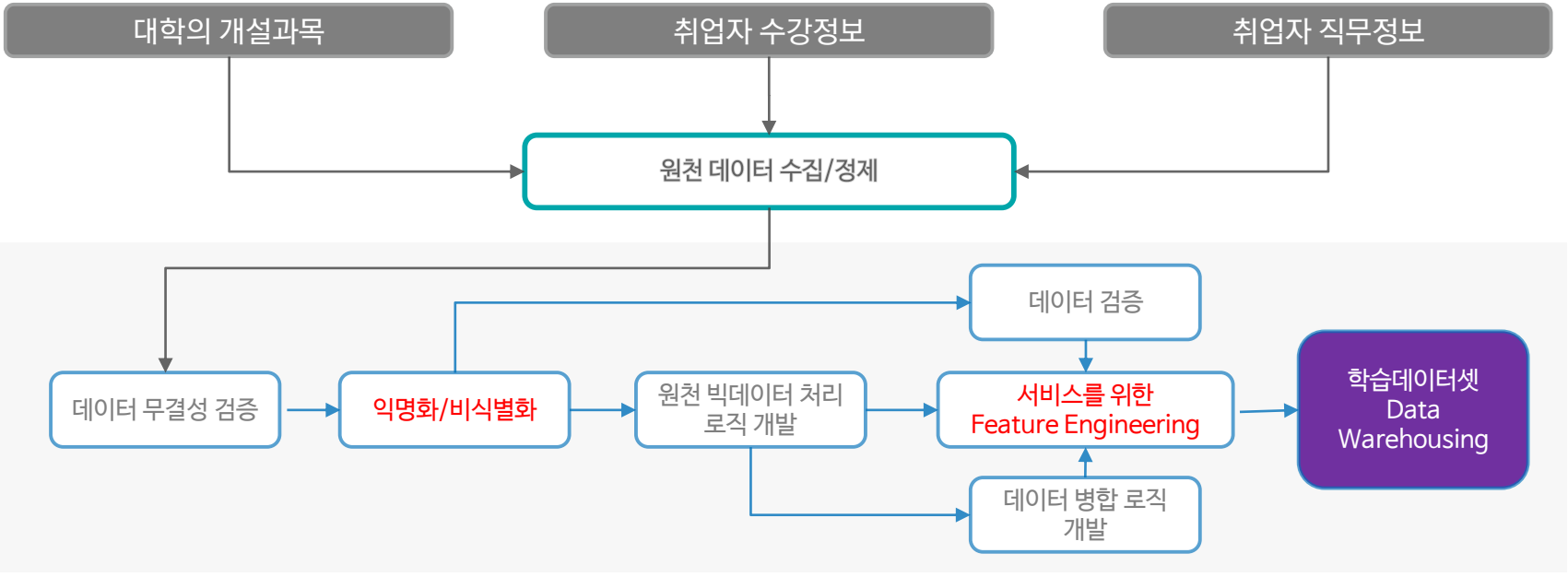
- 연구 데이터의 정의
  범용성 확보를 위해 100개 대학의 개설과목과 2015년도에서 2020년 까지 취업자가 수강한 과목명을 취업자의 직무정보로 군집화한 학습 데이터셋 구성

- 제증명서 인터넷 발급 사이트인 웹민원센터(www.webminwon.com)의 대학 개설과목 정보와 취업 포털 사이트인 스카우트(www.scout.co.kr)의 직무정보를 원천 데이터로 수집.정제하고 수강과목 학습 데이터셋을 구성하고자 함 (골드 스타트 문제 해결 방안)



3.3 학습 데이터셋

- 학습 데이터셋 구성 방법
  원천 데이터 수집 및 정제 후 익명화와 비식별화 과정을 통해 연구 목적의 학습 데이터를 구성하고 데이터의 특이점 도출(Feature Engineering)를 통해 최종 학습 데이터셋을 완성함.



3.4 알고리즘 실험

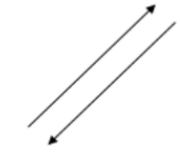
수강이력 데이터 표준화

- 대학별로 상의한 개설과목을 군집화를 통한 표준화 작업

대학명 | 학과명 | ··· | 개설과목

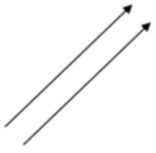
수강이력 데이터 정형화

- 코사인 유사도(Consine Similarity)를 통한 데이터 정형화



코사인 유사도 : -1

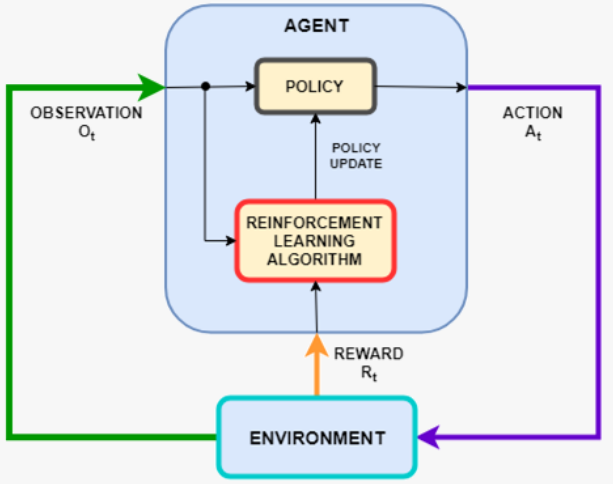
코사인 유사도 : 0

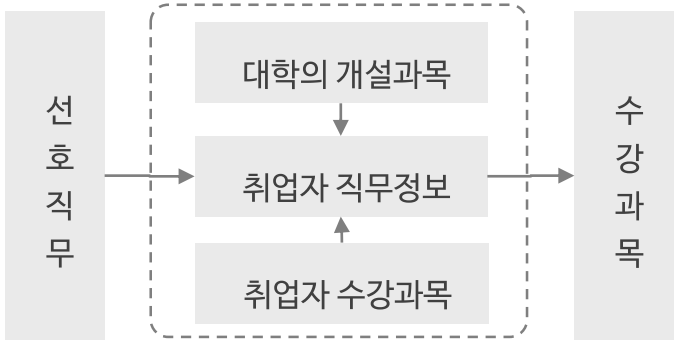


코사인 유사도 : 1

$$similarity = cos(\Theta) = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

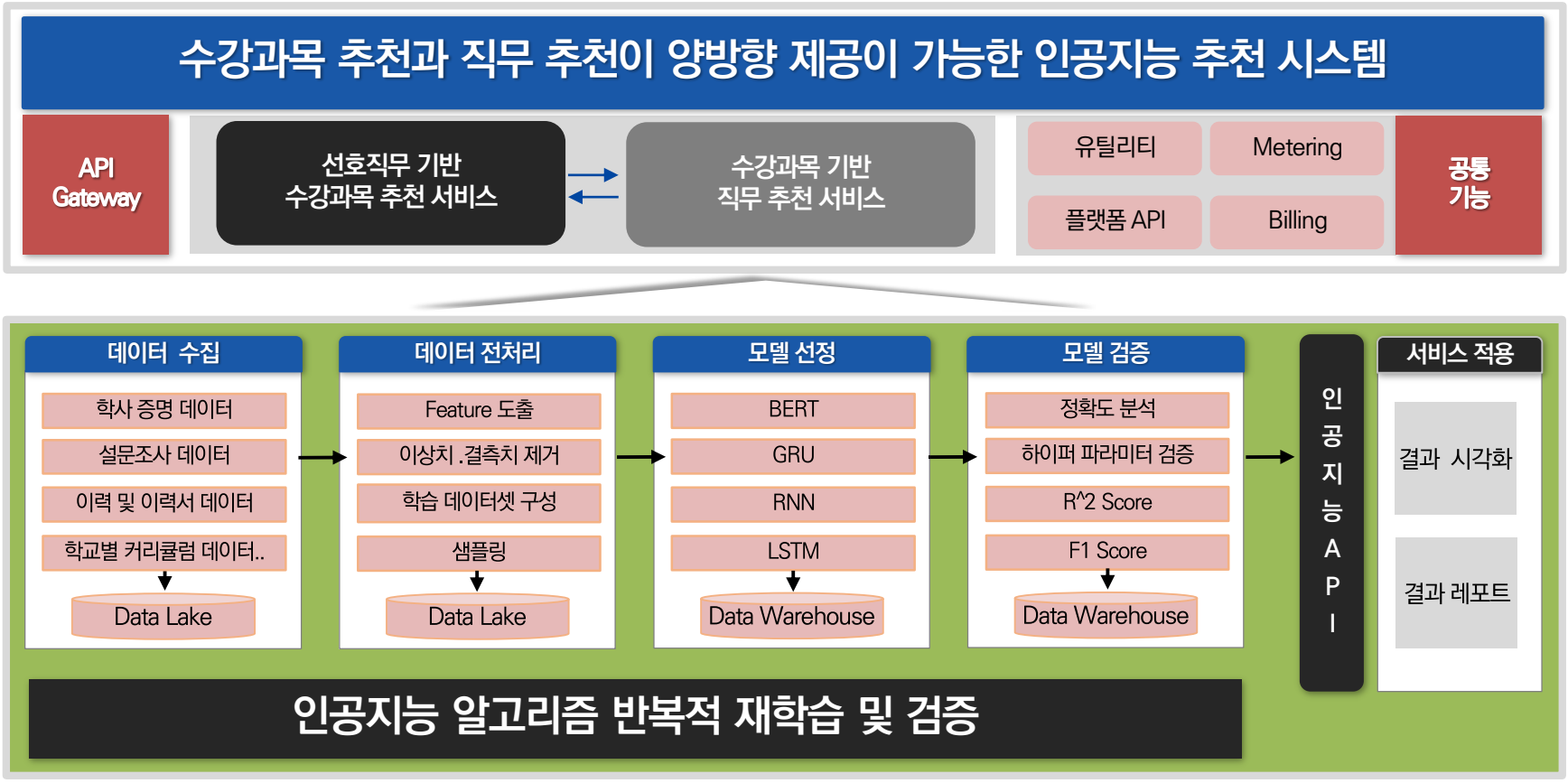
kNN 알고리즘의 반복적 재학습 및 검증





3.5 시스템 구성도

- 시스템 확장성을 고려한 API Gatway 방식의 시스템 구성
  추천 시스템 베터 서비스 이후 성능 검증을 완료 하고, 추천 시스템의 적용을 요청하는 다수의 대학을 확장 설치를 고려한 API 기술 기반의 시스템 구성



제4장 결론

1. 한계점 및 개선점
2. 향후 연구 계획

한계점 및 개선점

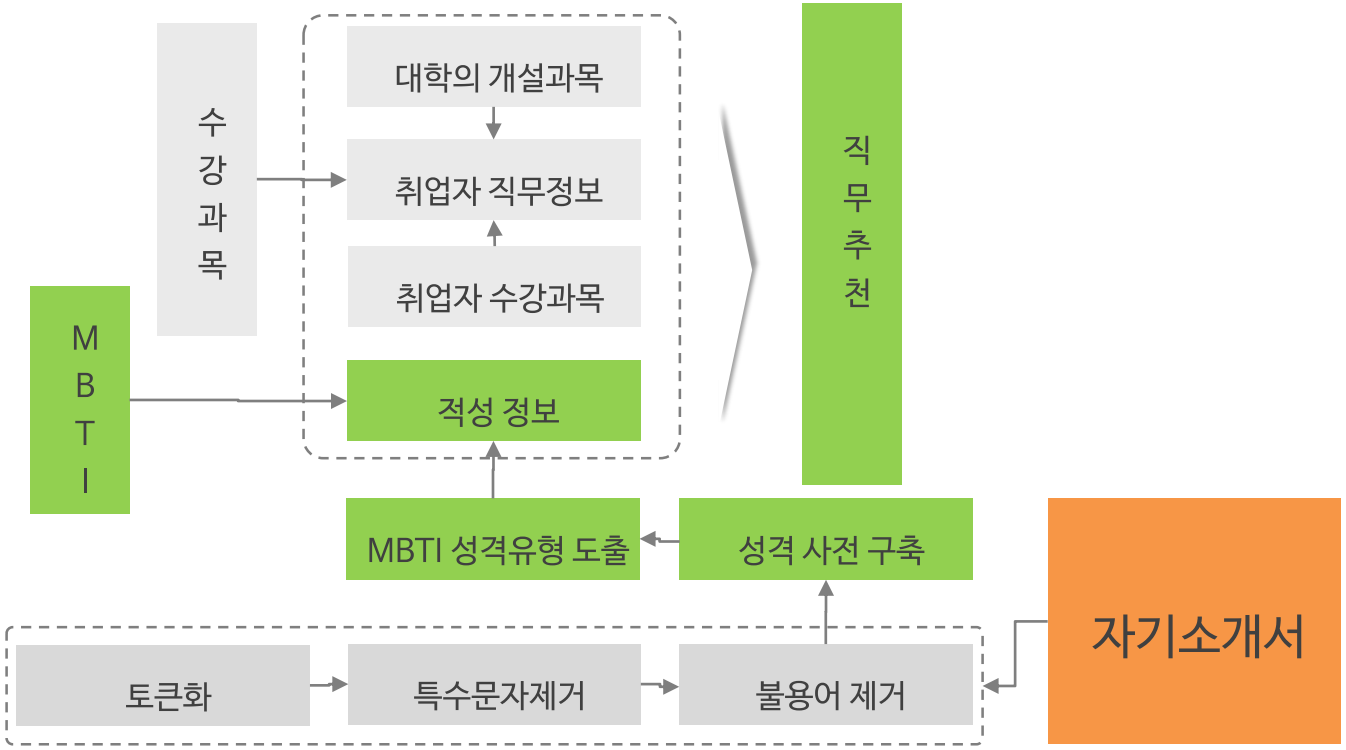
- 프로토타입 선행 연구 결과
  - 본 연구에 앞서 프로토 타입 선행 연구를 진행함(**학생 1675명의 수강과목 및 이력서**)
  - 선행 연구의 머신러닝 RamdomForest 의 한계를 인지하고 본 연구에서는 딥러닝 K Nearest Neghbor 를 중심으로 연구하고자 함.

한계점	**01**	**다양한 범주값에 비해 적은 데이터 수로, 여러 모델을 시도하기에는 어려움이 있었음**
	02	Auto ML 을 사용해 자연어 처리를 시도했으나 다양한 범주의 값으로 인해 제대로 분류되지 않았음
	03	1987년도부터 수집된 데이터, 특정 직업에 편중된 데이터로 현 시대의 트렌드를 반영하지 못한 아쉬움이 있었음
	04	정확도, F-1 스코어를 높이기 위해 학생들이 많이 들은 과목을 위주로 추천해주는 모델을 만들어 다양한 과목 추천의 어려움이 있었음

개선점	**01**	**프로젝트 진행초반 논문 등과 같은 선행연구에 관한 정보를 통해 도움을 얻을 수 있도록 함**
	02	학교별 모델을 만들어 정확도가 높은 학교별 맞춤 서비스 필요
	03	학생들의 강의평가와 같은 여러 정보를 추가해 수강신청시 다방면으로 도움을 받을 수 있도록 발전
	04	여러 서비스가 융합된 플랫폼을 개발하여 활용

향후 연구 계획

- 본 연구를 기반으로 재학생이 이수 예정인 수강과목을 입력 시 직무 추천이 가능한 심화 연구가 가능할 것으로 예상됨.
- 이를 위하여 취업자의 자기 소개서를 기반으로 적성 정보를 구성하고, 재학생이 MBTI 정보를 추가로 입력하므로서 완성도 높은 직무 추천이 가능한 심화 연구를 진행하고자 함.



제5장 참고문헌

참고문헌

[국내 문헌]

이태명, 김재경. (2020). BERT와 LSTM모형에 기반한 기업 평판 지수 도출 방법 및 적용. 한국지능정보시스템학회 학술대회논문집, 153-155.

이청용, 이병현, 이흠철, 김재경. (2021). CNN 기반 리뷰 유용성 점수 예측을 통한 개인화 추천 서비스 성능 향상에 관한 연구. 한국지능정보시스템학회 학술대회논문집, 15-15.

유재준, 유준영, 조재춘. (2020) kNN알고리즘 기반의 교양과목 추천 모델 연구. 한국컴퓨터교육학회 학술발표대회논문집, 24(1), 107-109.

장예화, 이병현, 정재호, 김재경. (2021). MBTI 성격유형을 반영한 심층 신경망 기반 직무 추천 서비스. 인터넷전자상거래연구, 21(4), 99-113.

김근호, 정종인, 김창석, 강신천, 김의정. (2020). MBTI와 성적을 활용한 진로 추천 시스템의 연구. 한국정보통신학회 종합학술대회 논문집, 24(1), 49-52.

김용수. (2012). 개인화 서비스를 위한 추천 시스템의 연구동향. Ie 매거진, 19(1), 37-42.

이재규, 박희성. (2021). 네트워크 분석을 활용한 딥러닝 기반 전공과목 추천 시스템. 한국지능정보시스템학회 학술대회논문집, 86-86.

김영현, 이인환, 권준희. (2020). 도보 여행객을 위한 여행 추천 시스템의 설계 및 구현. Proceedings of KIIT Conference, 214-215.

참고문헌

김영재, 원준연, 정진석, 김선호, 윤용운. (2016). 패턴 분석 알고리즘을 이용한 수업 추천. 한국정보과학회 학술발표논문집, 1659-1661.

손기락, 김소현. (2007). 협동적 필터링을 이용한 K-최근접 이웃 수강 과목 추천 시스템. 한국정보교육학회, 11(3), 281-288.

박희망, 김태성, 황일용, 정세빈, 윤현주. (2020). 협업 필터링과 내용기반 추천을 활용한 여행경로 추천 시스템. Proceedings of KIIT Conference, 452-454

김두형, 신우석, 한기웅, 이진숙, 문기범, 이수강, 한수연, 권혜정, 한성원. (2020). 협업필터링을 활용한 대학 교양과목 추천 시스템. 대한산업공학회 추계학술대회 논문집, 2551-2556.

[해외 문헌]

Wenzhong Liang, Guangquan Lu, Xiaoyu Ji, Jian Li, Dingrong Yuan. (2014). Difference Factor' KNN Collaborative Filtering Recommendation Algorithm. Lecture Notes in Computer Science, 175-184.

D. A. Adeniyi, Z. Wei, Y. Yang. (2017). Personalised news filtering and recommendation system using Chi-square statistics-based K-nearest neighbour (χ2SB-KNN) model. ENTERPRISE INFORMATION SYSTEMS, 1283-1316.

V. Subramaniyaswamy. (2017). Adaptive KNN based Recommender System through Mining of User Preferences. WIRELESS PERSONAL COMMUNICATIONS, 2229-2247.

감사합니다.

질/의/응/답

Appendix

프로토타입 선행 연구 결과

- 데이터 소개
- 데이터 전처리
- 최종 데이터셋

- 분석 방법
- Model1 입력변수 및 목표 변수 설정
- Model1 분석 결과
- Model2 분석 결과

- 프로토타입 서비스
- 한계점 및 개선점

데이터 소개 – 학생 1675명의 수강과목 및 이력서

Appendix

학사 데이터

순번	성별	학교명	단과명	주전공	복수전공	부전공	입학일	졸업일
A	여	OO대학교		사회체육학과			19930311	20010227
A	남	OO대학교	자연과학대학	수학과			35187	38571
A		OO여자대학교	공과대학	자원공학과	X	X	19940302	19970221
A	여	OO대학교		정보통계학과			33039	37162
A0005	여	OO여자대학교	전자정보공과대학	컴퓨터공학	X	X	37319	40939
A0006	남	OO대학교		디자인과	X	X	19980301	20020219
A0007		OO대학교		컴퓨터공학과			39820	43029
A0008	남	OO대학교		아동학과			20020303	20080203
A0009	남	OO대학교		경영학	회계학전공	X	39210	45938
A0010	여	OO대학교		전자과				

성적 데이터

순번	년도	학기	이수구분	교과목명	성적	학점
A	1996	1	교양	채플	P	P
A	1998	2	전공	운동보건학	B	2
A	1997	2	전공	스포츠사회학	A-	3
A	1998	1	교양	세계정치의이해	A0	3
A0005	1998	1	전선	씨름	B+	1
A0006	1995	1	교선6	레크리에이션	A	2
A0007	1994	2	전선	사회체육강독	A+	3
A0008	1992	2	전공	스포츠의학	C	3
A0009	1995	2	전공	선형대수	B	2

이력서 데이터

순번	근무경력
A	-2004년 11월~ 2005월 현재: OO공단 OO지도원 근무 / 일반서무 담당 - 사업계획서 작성
A	
A0004	2002년2월17일~2004년4월 6일 (주)OO기업 e-biz사업부 개발팀 근무
A0005	
A0006	
A0007	
A0008	거래처 관리 및 신규 거래처 개척 |
A0009	디자인 담당업무를 진행했습니다. 의류 샘플 선정부터 디자인에 관해 모든 업무 처리
A0010	OO기업 마케팅팀 8개월 근무 | 경력은 많지 않지만 성실하게 근무 가능합니다!

데이터 전처리 (1/2)

순 번

- 중복되는 순번 제거
- A0001~A2162 (1675명)

성 적

- 학교별로 표기의 차이
- A+부터F까지 9~0으로 분류
- 오류 값 (예:AB+, /, S, AD) 처리
- (raw데이터에서 범주 개수:87->10개)

단 과 명

- 학교별로 다양한 단과명이 존재
- 유사한 단과명의 경우, 하나로 분류
- 결측치는 주전공을 보고 분류 (raw데이터에서 범주 개수:139->15개)

학 점

- 수업의 방식에 따라 학점의 차이
- 실습 등을 제외한 1, 2, 3학점만 사용
- (raw데이터에서 범주 개수:60->3개)

이 수 구 분

- 학교별로 표기의 차이
- '전필', '전선', '교양' 3가지로 구분
- 표기 오류의 경우 직접 과목을 보며 분류
- (raw데이터에서 범주 개수:331->3개)

전체학기

- 년도와 학기 컬럼을 이용해 학기를 순서에 따라 1~8학기로 구분
- 9학기 이상은 데이터에서 제외

데이터 전처리 (2/2)

교 과 목 명

- 2만개이상의교과목명을100개로 라벨링작업
- raw데이터에서범주개수:23953

한국사특강 세계속의한국역사 한국역사속의권력과통치
한국사개설
한국고대사
understangofkoreanhistory
한국전통사상론
한국역사의이해
한국역사탐구
인물로본한국역사 한국사의역사인식
…

한국역사

교 과 목 명

- 경력데이터와전공확인후 '직무' 도출작업
- NCS기반중분류까지라벨링(약40종)

순번	경력사항
A0011	OO기업 의류디자이너 | 스포츠 아웃도어 의류디자인

코드	NCS 직무 상세
01	관리직(임원·부서장)
02	경영·행정·사무직
024	광고·조사·상품기획·행사기획 전문가
03	금융·보험직
…	
41	예술·디자인·방송직
42	스포츠·레크리에이션직
…	
88	인쇄·목재·공예 및기타 설치·정비·생산직
89	제조 단순직
90	농림어업직

최종 데이터셋

초기 데이터셋

	순번	이수구분	교과목명	성적	학점	semester	단과명	직무중분류
0	A0001	교양	철학	9	2	6	체육대학	26.0
1	A0001	전선	스포츠일반	7	3	5	체육대학	26.0
2	A0001	전선	스포츠일반	8	3	5	체육대학	26.0
3	A0001	전선	스포츠기타	9	2	5	체육대학	26.0
4	A0001	전선	스포츠일반	7	2	5	체육대학	26.0
...	...	...	...	...	...	...	...	...
50656	A2162	교양	동양학	9	3	4	생활과학대학	24.0
50657	A2162	교양	예술학	8	3	4	생활과학대학	24.0
50658	A2162	전선	패션기타	8	2	4	생활과학대학	24.0
50659	A2162	교양	영어회화	7	2	5	생활과학대학	24.0
50660	A2162	전선	패션기타	6	3	2	생활과학대학	24.0

A0001

A2162

최종 데이터셋

	순번	단과명	직무중분류	1_1교과목명	1_1성적	1_1이수구분	1_1학점	1_2교과목명	1_2성적	1_2이수구분	...
0	A0001	체육대학	26	스포츠일반	7	전선	3	구기스포츠	7	전선	...
1	A0004	공과대학	13	산업일반	7	교양	2	생활학	8	교양	...
2	A0005	공과대학	61	물리학	7	교양	2	화학	7	교양	...
3	A0006	공과대학	13	구기스포츠	6	교양	1	철학	7	전선	...
4	A0009	공과대학	61	컴퓨터소프트웨어	5	전선	1	영어회화	5	교양	...
...	...	...	...	...	...	...	...	...	...	...	...
1059	A2157	상경대학	26	경영학	8	전선	3	커뮤니케이션	8	교양	...
1060	A2159	자연과학대학	29	영어회화	4	교양	3	국문학	6	교양	...
1061	A2160	사회과학대학	24	회계학	6	전선	2	통계학	8	전선	...
1062	A2161	사회과학대학	2	한국역사학	2	교양	3	경제학	5	전선	...
1063	A2162	생활과학대학	24	국문학	6	교양	3	환경학	8	교양	...

분석 방법

입력변수 설정	목표변수 설정	ML 모델 설정	파라미터 설정	모델 데스트
학생의 희망직무 전공 (단과명) 1~2학년 수강과목 이수구분 성적 학점	Model1 (메인) 3학년 1,2학기 4학년 1,2학기 추천 수강과목 Model2 (서브) 3~4학년 과목별 수강여부	[분류모델] Decision Tree Catboost RandomForest LightGBM	Pipeline MInMaxScaler Gridsearch Label Encoding SMOTE (Oversamplin Depth Learning rate Iterations	Accuracy F1score ROC AUC

Model1 입력변수 및 목표 변수 설정

입력변수 설정	목표변수 설정	목표변수 설정 기준
단과명, 직무코드, 교과목명, 이수구분, 성적, 학점	2학년 2학기 까지 데이터 → 학생별 3학년 1학기 1개 과목	목표변수로 삼을 교과목 선정 기준 1. 성적이 우수한 과목 2. 학점 비중이 높은 과목 - 1순위 A+ 3학점 - 2순위 A+ 2학점 - 3순위 A 3학점 - 4순위 A 2학점 . . .
	3학년 1학기 까지 데이터 → 학생별 3학년 2학기 1개 과목	
	3학년 2학기 까지 데이터 → 학생별 4학년 1학기 1개 과목	
	4학년 1학기 까지 데이터 → 학생별 4학년 2학기 1개 과목	

Model1 분석 결과

모델 선정 배경

- 상대적으로 데이터가 많은 3학년 1학기의 경우 빠르고 성능이 좋은 Light GBM 선정
- 학기가 높아질수록 듣는 수강과목이 줄어들어 Row의 수가 감소. 과적합을 방지하기 위하여 Random Forest 선정

성능 향상 작업

- Label Encoding
- Pipeline
- MinMax Scalar
- SMOTE 기법 사용
- GridSearch, Cross Validation

<추천대상학기>	Row 개수	Best Model	Train Data		Test Data	
			ACC	F1 SCORE	ACC	F1 SCORE
3학년1학기	277개	Light GBM	79%	78%	46%	42%
3학년2학기	256개	Random Forest	89%	90%	54%	47%
4학년1학기	222개	Random Forest	93%	93%	43%	48%
4학년2학기	166개	Random Forest	98%	98%	47%	45%

Model2 분석 결과 (서브 모델)

01 목표변수 설정
- 각 교과목을 목표변수로 설정 〉 과목별로 모델 설정 및 학습
- 목표변수 : 3~4학년에 해당 과목을 수강한 경우 1 , 아닌 경우 0
- 교과목선정 기준 : 200명이상이 수강한 과목을 선정

02 Over sampling
- SMOTE 기법 활용

03 Catboost 학습 모델 사용
- 이진분류 모델에서 가장 우수한 성능을 내어 채택

04 학습모델 정확도 평가
f1score
ROC - AUC

교과목명	3학년 이후 수강한 학생	임계값	Train data	Testdata	
			f1macroavg	f1macroavg	ROC-AUC
경영학	365	0.5	0.92	0.60	0.6375
경제학	302	0.5	0.90	0.64	0.6929
미술학	294	0.5	0.84	0.58	0.6336
컴퓨터일반	269	0.5	0.92	0.72	0.7840
국제사	265	0.4	0.87	0.52	0.5158
스포츠일반	256	0.5	0.88	0.48	0.5308
국문학	253	0.5	0.83	0.57	0.6446
영어회화	243	0.5	0.94	0.56	0.6038
미디어	239	0.5	0.87	0.57	0.5963
가정학	238	0.4	0.89	0.50	0.5107
교육학	226	0.5	0.83	0.46	0.5305
사회학	226	0.5	0.85	0.57	0.6637
비즈니스	215	0.5	0.87	0.68	0.7165
철학	214	0.5	0.93	0.54	0.5942
컴퓨터소프트웨어	208	0.3	0.95	0.59	0.6728

프로토타입 서비스 도출 – Front UI

프로토타입 서비스 도출 – Report

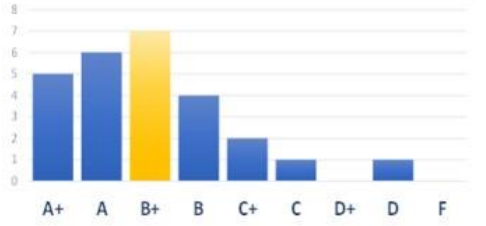
이 름 라이언

학 번 201925127

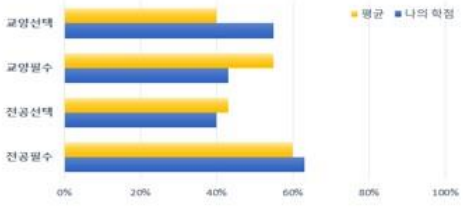
학 과 통계학과

희망직무 데이터분석가

나의 성적분포



분야별 이수한 학점



AI 가 추천하는 다음학기 수강과목

	1학년 수강과목		2학년 수강과목		3학년 1학기 추천 수강과목		3학년 2학기
	1-1	1-2	2-1	2-2			
전공	응용 수학1 기초통계학1 통계학원론 미분과 적분1	응용 수학2 기초통계학2 정보통계학 선형대수학	기초확률론 보험학원론 탐색적 자료분석	행렬대수학 미시경제학 재무관리 전산실습	전공 필수: 회귀분석, 통계적 품질혁식론 전공선택: 다변량자료 분석, 시계열분석		전공 필수: 비모수통계학 전공선택: 통계적 인공지능
교양	실용컴퓨터1 CSP 진로탐색 실용영어1	실용컴퓨터2 경영과 경제	현재중국론 프랑스 예술과 광기	아메리칸재즈 테니스 기초	교양필수: 경영학원론, 정치학개론 교양선택: 빅데이터 마케팅, 비즈니스 커뮤니케이션		교양필수: 금융통계학 교양선택: AI와 미래산업

수강한 과목의 희망직무적합도

good! 73%

희망직무 채용공고

CACAO BANK
[프로덕트] 데이터 분석가 모집

한계점 및 개선점

한계점	**01**	**다양한 범주값에 비해 적은 데이터 수로, 여러 모델을 시도하기에는 어려움이 있었음**
	02	Auto ML 을 사용해 자연어 처리를 시도했으나 다양한 범주의 값으로 인해 제대로 분류되지 않았음
	03	1987년도부터 수집된 데이터, 특정 직업에 편중된 데이터로 현 시대의 트렌드를 반영하지 못한 아쉬움이 있었음
	04	정확도, F-1 스코어를 높이기 위해 학생들이 많이 들은 과목을 위주로 추천해주는 모델을 만들어 다양한 과목 추천의 어려움이 있었음

개선점	**01**	**프로젝트 진행초반 논문 등과 같은 선행연구에 관한 정보를 통해 도움을 얻을 수 있도록 함**
	02	학교별 모델을 만들어 정확도가 높은 학교별 맞춤 서비스 필요
	03	학생들의 강의평가와 같은 여러 정보를 추가해 수강신청시 다방면으로 도움을 받을 수 있도록 발전
	04	여러 서비스가 융합된 플랫폼을 개발하여 활용