

•CONTENTS•

03

AI 머신러닝으로 배우는 데이터 과학, 똑똑한 컴퓨팅

- ❖ GDP와 관련 요인들의 상관관계에 예측 57
제주과학고등학교 고준호, 이우성, 임현성
- ❖ 머신러닝을 이용한 농구선수 포지션 추천 시스템 개발 68
제주제일고등학교 김서현, 양성현, 전승재
- ❖ 머신러닝을 활용한 제주 관광객 증감 요인 분석 및 예측 72
제주중앙여자고등학교 김영민, 양소연, 정지원
- ❖ 머신러닝을 이용한 영화 수익 예측 88
대기고등학교 강제호, 이창원, 황세호



GDP와 관련 요인들의 상관관계 예측

제주과학고등학교 1학년

고준호·이우성·임현성

I. 서 론

1. 탐구 동기

최근 뉴스를 보면 매일 경제에 관한 소식이 주를 이루고 있다. 경제라는 것은 그때의 상황마다 크게 변동할 수 있기 때문에 그만큼 정보가 많은 것이다. 그렇기에 최근 일본의 수출규제와 같은 문제는 논란과 기사가 많아지고 있다. 그리고 우리는 자료를 찾아보던 중 무역 갈등에 의해 경제 성장이 낮아지는 추세를 보인다는 기사를 보게 되었다. 또한, 머지않아 작년 환경오염에 관한 말이 많았을 때 즈음에 쓰인 기사 중엔 경제성장과 함께 환경 파괴가 진행된다는 말이 있다. 그래서 우리는 경제성장이 무역 갈등으로 인한 물가 변화 또는 환경오염 외에도 다른 어떤 요인이 경제성장에 영향을 미치는지 알아보고, 우리가 배운 머신러닝으로 경제성장과 요인들의 상관관계를 알 수 있도록 한 후, 더 나아가 미래 요인들의 변화로 인한 경제성장 정도를 예측할 수 있는 프로그램을 짜보기로 생각하였다.

2. 탐구 목적

인구수, 실업률, 국토의 면적, 보건비 지출량, CO₂ 배출량을 요인으로 하여 GDP와의 관계를 알아내어 요인의 변화에 따른 GDP의 변화를 유추할 수 있다.

3. 탐구 목표

- 가. 어떠한 요인이 GDP에 영향을 주는지 알아본다.
- 나. GDP 이외의 요인이 주어졌을 때 GDP를 유추할 수 있다.
- 다. 요인의 변화를 통해 미래의 GDP를 대략적으로 알 수 있다.

II. 탐구 방법

1. 탐구 재료 및 기구

- 가. 준비물 : 인터넷, 캐글
- 나. 기구 : 컴퓨터

2. 탐구 과정

가. GDP와의 관련 요인을 분석한다.

- 1) 통계청에 접속한다.
- 2) 선정된 데이터들을 통합하고, 각각의 통계자료에서 공통으로 조사된 나라들만으로 통합데이터를 구성한다.
- 3) 통합데이터의 각 나라에 해당하는 2015년 GDP를 추가한다.
- 4) 프로그램을 제작한다.
- 5) 최종데이터를 프로그램에 올린 뒤, 프로그램을 실행시킨다.
- 6) Heatmap의 결과를 통하여 관련성이 높은 요인들을 뽑아낸다.

나. 국가별 물가지수, 보건지출비, CO2 배출량의 미래예측 값을 구한다.

- 1) 연도별 물가지수, 보건지출비, CO2 배출량의 2008~2019까지의 데이터를 다운받는다.
- 2) 데이터를 정리한다.
- 3) 2008~2019 보건지출비 자료와 forecast 함수를 통하여 2029에서의 보건지출비를 예측한다.
- 4) CO2 배출량, 물가지수에 대해서도 forecast 함수를 통한 예측작업을 반복한다.

다. 여러 요인을 통한 국가들의 연도별 GDP 예측

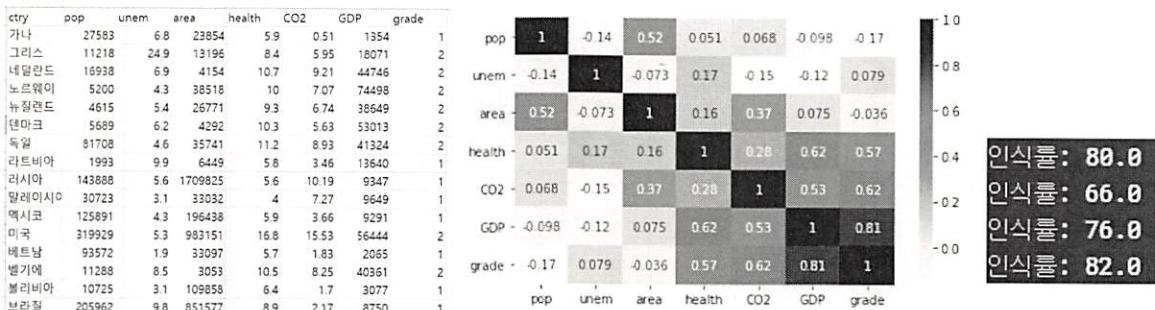
- 1) Test 데이터와 국가별 CO2, health, price, GDP, grade(4등급)를 정리한다.
- 2) 탐구 과정_2에서의 예측 자료를 dataset 으로 하고 프로그램을 작성한다.
- 3) 1)에서의 test 데이터를 모두 인공지능에 학습시킴. (결정 트리 Decision Tree 알고리즘을 이용함)
- 4) 각 국가, 연도별로 GDP등급을 예측한다.
- 5) 예측한 값들을 국가별로 2차원 리스트로 만든 후 리스트를 이용해서 꺾은선 그래프로 나타낸다.

III. 탐구결과 및 고찰

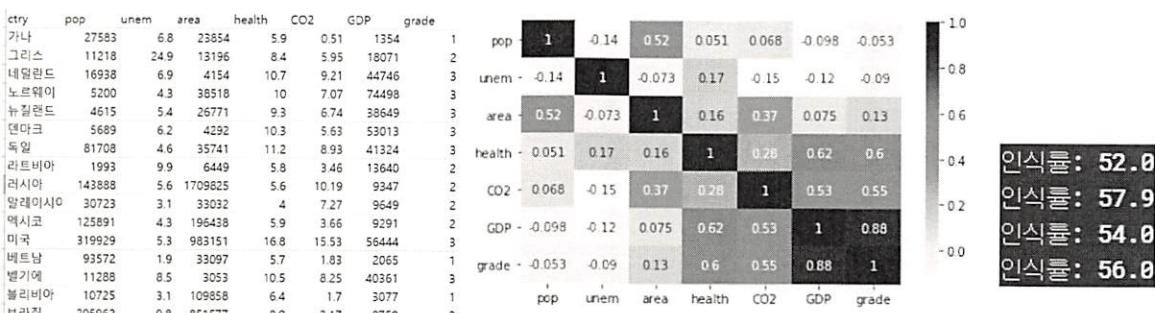
1. <탐구 가>의 결과

가. 등급의 수에 따른 히트맵, 인식률(써포트 벡터 머신, 논리 회귀, 결정 트리, 근접 이웃)

1) 등급을 두 개로 나누었을 때

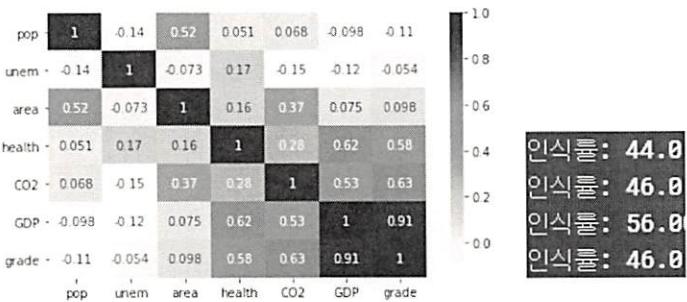


2) 등급을 세 개로 나누었을 때



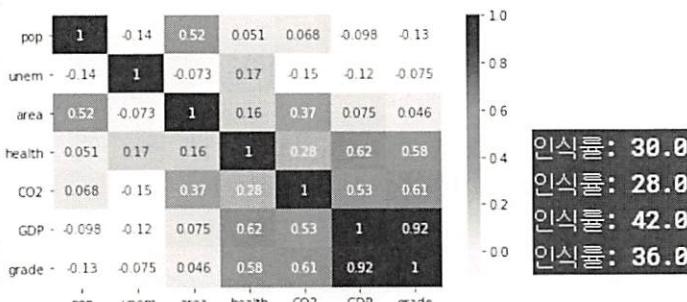
3) 등급을 네 개로 나누었을 때

ctry	pop	unem	area	health	CO2	GDP	grade
기나	27583	6.8	23854	5.9	0.51	1354	1
그리스	11218	24.9	13196	8.4	5.95	18071	4
네덜란드	16938	6.9	4154	10.7	9.21	44746	5
노르웨이	5200	4.3	38518	10	7.07	74498	4
뉴질랜드	4615	5.4	26771	9.3	6.74	38649	3
덴마크	5689	6.2	4292	10.3	5.63	53013	4
독일	81708	4.6	35741	11.2	8.93	41324	4
라트비아	1993	9.9	6449	5.8	3.46	13640	2
러시아	143888	5.6	1709825	5.6	10.19	9347	2
말레이시아	30723	3.1	33032	4	7.27	9649	2
멕시코	125891	4.3	196438	5.9	3.66	9291	2
미국	319929	5.3	983151	16.8	15.53	56444	4
베트남	93572	1.9	33097	5.7	1.83	2065	1
벨기에	11288	8.5	3053	10.5	8.25	40361	4
볼리비아	10725	3.1	109858	6.4	1.7	3077	1
브라질	205962	9.8	851577	8.9	2.17	8750	2



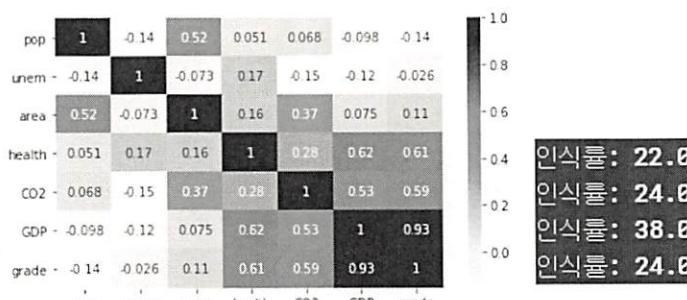
4) 등급을 다섯 개로 나누었을 때

ctry	pop	unem	area	health	CO2	GDP	grade
기나	27583	6.8	23854	5.9	0.51	1354	1
그리스	11218	24.9	13196	8.4	5.95	18071	3
네덜란드	16938	6.9	4154	10.7	9.21	44746	5
노르웨이	5200	4.3	38518	10	7.07	74498	5
뉴질랜드	4615	5.4	26771	9.3	6.74	38649	4
덴마크	5689	6.2	4292	10.3	5.63	53013	5
독일	81708	4.6	35741	11.2	8.93	41324	4
라트비아	1993	9.9	6449	5.8	3.46	13640	3
러시아	143888	5.6	1709825	5.6	10.19	9347	2
말레이시아	30723	3.1	33032	4	7.27	9649	2
멕시코	125891	4.3	196438	5.9	3.66	9291	2
미국	319929	5.3	983151	16.8	15.53	56444	4
베트남	93572	1.9	33097	5.7	1.83	2065	1
벨기에	11288	8.5	3053	10.5	8.25	40361	4
볼리비아	10725	3.1	109858	6.4	1.7	3077	1
브라질	205962	9.8	851577	8.9	2.17	8750	2

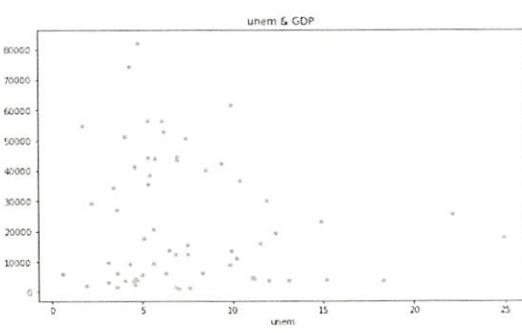
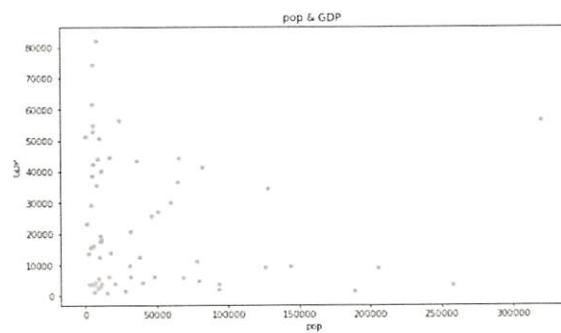


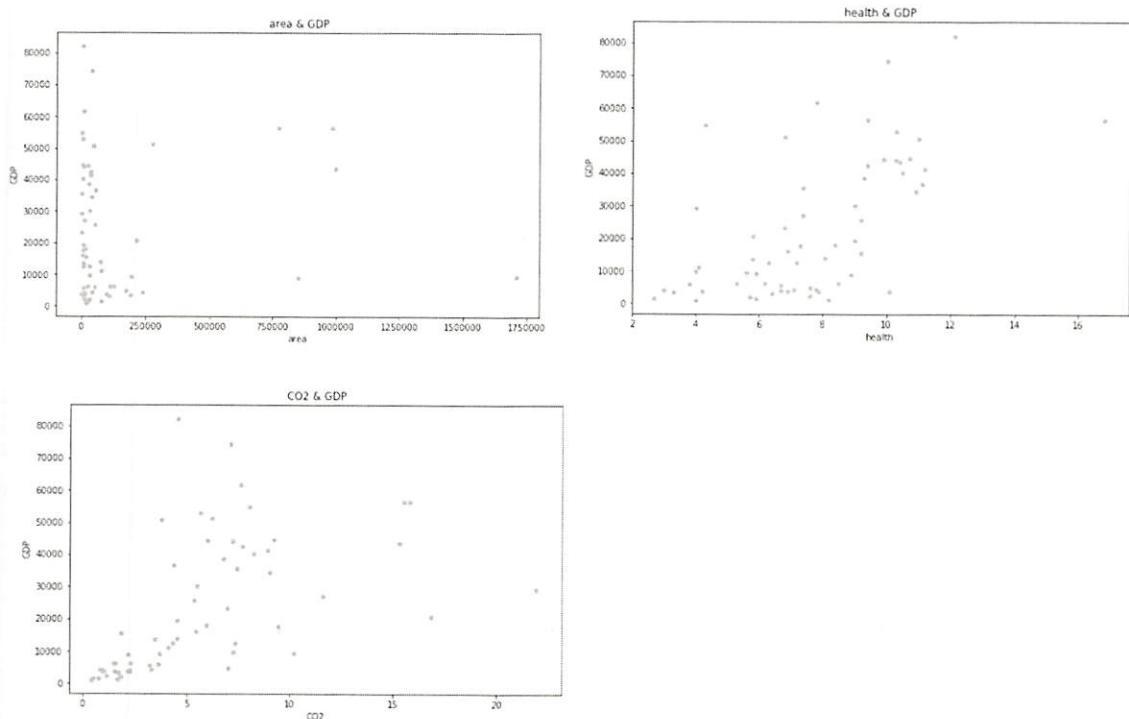
5) 등급을 여섯 개로 나누었을 때

ctry	pop	unem	area	health	CO2	GDP	grade
기나	27583	6.8	23854	5.9	0.51	1354	1
그리스	11218	24.9	13196	8.4	5.95	18071	4
네덜란드	16938	6.9	4154	10.7	9.21	44746	5
노르웨이	5200	4.3	38518	10	7.07	74498	5
뉴질랜드	4615	5.4	26771	9.3	6.74	38649	4
덴마크	5689	6.2	4292	10.3	5.63	53013	5
독일	81708	4.6	35741	11.2	8.93	41324	4
라트비아	1993	9.9	6449	5.8	3.46	13640	3
러시아	143888	5.6	1709825	5.6	10.19	9347	2
말레이시아	30723	3.1	33032	4	7.27	9649	2
멕시코	125891	4.3	196438	5.9	3.66	9291	2
미국	319929	5.3	983151	16.8	15.53	56444	4
베트남	93572	1.9	33097	5.7	1.83	2065	1
벨기에	11288	8.5	3053	10.5	8.25	40361	4
볼리비아	10725	3.1	109858	6.4	1.7	3077	1
브라질	205962	9.8	851577	8.9	2.17	8750	3



나. 각 요인에 대한 GDP의 점그래프





2. <탐구 나>의 결과

가. 2009~2015년도 CO2 배출량과 보건지출비

<CO2 배출량>

	2008	2009	2010	2011	2012	2013	2014	2015
가나	0.34	0.39	0.43	0.43	0.5	0.52	0.49	0.51
그리스	8.52	8.12	7.5	7.4	6.98	6.28	6.04	5.95
네덜란드	9.99	9.65	10.23	9.47	9.32	9.27	8.81	9.21
노르웨이	7.4	7.39	7.67	7.32	7.08	6.9	6.89	7.07
뉴질랜드	7.83	7.05	6.96	6.75	7.07	6.94	6.89	6.74
덴마크	8.85	8.49	8.51	7.54	6.59	6.88	6.1	5.63
독일	9.6	8.95	9.45	9.11	9.26	9.47	8.93	8.93
리트비아	3.64	3.35	3.86	3.56	3.44	3.43	3.37	3.46
러시아	10.88	10.09	10.7	11.22	10.83	10.55	10.34	10.19
레바논	3.84	4.61	4.2	4.03	4.27	3.9	3.99	3.88
루마니아	4.44	3.8	3.69	4.02	3.92	3.45	3.43	3.51
말레이시아	7	6.14	6.75	6.7	6.65	7.09	7.37	7.27
멕시코	3.9	3.77	3.85	3.94	3.92	3.8	3.63	3.66
미국	18.1	16.66	17.26	16.69	16	16.11	16.19	15.53
방글라데시	0.28	0.29	0.33	0.35	0.37	0.38	0.4	0.44
베트남	1.19	1.3	1.45	1.43	1.4	1.45	1.58	1.83
벨기에	9.68	8.96	9.52	8.45	8.35	8.41	7.83	8.25
볼리비아	1.22	1.27	1.38	1.48	1.63	1.63	1.73	1.7

<보건지출비>

국가별	2008	2009	2010	2011	2012	2013	2014	2015
가나	6.4	6.5	6.5	6.1	5.5	5.8	5.5	5.9
그리스	9.4	9.5	9.6	9.1	8.8	8.3	7.9	8.4
네덜란드	9.5	10.2	10.4	10.5	10.9	10.9	10.9	10.7
노르웨이	8	9.1	8.9	8.8	8.8	8.9	9.3	10
뉴질랜드	9.1	9.7	9.7	9.6	9.7	9.4	9.4	9.3
덴마크	9.5	10.7	10.4	10.2	10.3	10.2	10.3	10.3
독일	10.2	11.2	11	10.7	10.8	11	11.1	11.2
라트비아	8.5	8.9	8.6	7.9	7.6	7.5	7.8	5.8
러시아	5.2	5.9	5.3	5.1	5.3	5.5	5.7	5.6
레바논	7.9	7.1	7.5	8.1	8	7.5	7.4	7.4
말레이시아	3.1	3.4	3.3	3.5	3.6	3.7	3.9	4
멕시코	5.7	6.2	6	5.8	5.9	6	5.7	5.9
미국	15.3	16.3	16.4	16.4	16.4	16.3	16.5	16.8
방글라데시	2.5	2.6	2.7	2.8	2.7	2.7	2.7	2.6
베트남	5	5.2	5.9	5.8	6.5	6.4	5.8	5.7
벨기에	9.3	10.1	9.9	10.1	10.2	10.4	10.4	10.5
볼리비아	4.5	5.1	5.1	5.1	5.5	5.8	6.4	
브라질	8	8.4	8	7.8	7.8	8	8.4	8.9
사우디아라비아	2.9	4.1	3.5	3.6	3.9	4.4	5.1	5.8

3. <탐구 다>의 결과

가. 국가별로 CO₂, health, price, GDP를 정리한 후 GDP를 4등급으로 나눈다.

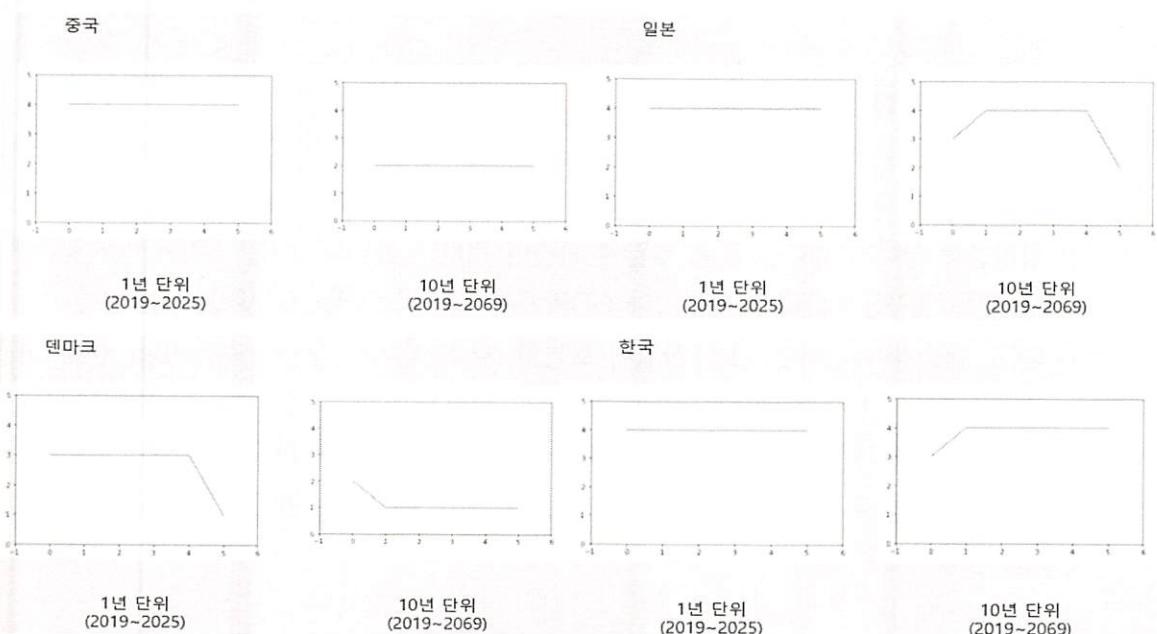
ctry	co2	health	trade	price	gdp	grade
가나	0.51	5.9	67.39	176	1354	1
그리스	5.95	8.4	39.38	100.8	18071	3
네덜란드	9.21	10.7	115.76	109.2	44746	4
노르웨이	7.07	10	46.61	108.6	74498	4
뉴질랜드	6.74	9.3	39.92	107.9	38649	4
덴마크	5.63	10.3	59.67	107.1	53013	4
독일	8.93	11.2	70.43	106.9	41324	4
라트비아	3.46	5.8	93.7	107.5	13640	3
러시아	10.19	5.6	40.46	151.5	9347	2
레바논	3.88	7.4	42.51	115	8452	2
루마니아	3.51	7.9	73.32	114.2	8978	2
말레이시아	7.27	4	126.56	112.8	9649	3
멕시코	3.66	5.9	68.02	119.4	9291	2
미국	15.53	16.8	21.08	108.7	56444	4
방글라데시	0.44	2.6	35.58	144.5	1210	1
베트남	1.83	5.7	170.19	145.8	2065	1

나. 2차원 리스트로 만드는 과정

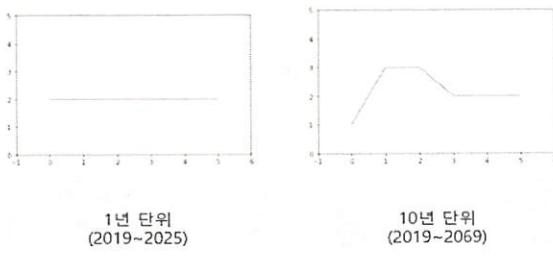
```
가나
[[5.025, 0.6325, 216.625], [3.691666667, 8.874166667, 347.5416667], [2.358333333, 1.115833332999999, 478.4583333000005], [1.825, 1.3575, 689.375], [-8.3883
3333, 1.599166667, 748.2916657000001], [-1.641666667, 1.848833330000002, 871.288333299998]]]
```

```
[[1, 1, 1, 1, 1, 1], [3, 1, 1, 1, 1, 1], [4, 4, 4, 4, 4, 1],
[4, 4, 4, 4, 4], [4, 4, 2, 3, 3, 1], [4, 1, 1, 1, 1, 1],
```

다. 여러 요인을 통한 국가들의 연도별 GDP 예측



인도



〈만든 프로그램은 <https://www.kaggle.com/wnsgh3678/ourstudy-10yr>

4. 고찰

- 가. forecast 함수에 대한 정확한 이해가 필요하다. 그 함수의 미래예측 정확성을 조절할 수 있는지, 예측하는 방법에 관한 깊은 탐구가 필요할 것이다.
- 나. 시간의 제약과 제한적인 나라들에 대해 조사한 값으로 인해 더 많은 국가에 대한 데이터 값을 마련하지 못하였다.

IV. 결 론

1. <탐구 가>의 결론

가. heatmap에서 알 수 있는 것

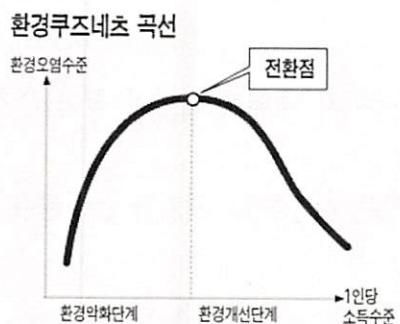
- 1) 2~6개로 등급을 나눈 heatmap 들에서 정신건강 상태와 CO2 배출량이 GDP와 관련이 있다는 것을 알 수 있다.
 - 가) 정신건강 상태가 높다는 것은 생활에 불만이 적다는 것이다. 이는 직장에서의 능률 향상에 좋은 영향을 끼치고, 그에 따라 GDP 또한 높아지는 것으로 생각할 수 있다.
 - 나) CO2 배출량이 많다는 것은 산업이 발달할 때 일어나는 당연한 현상이다. 물론 최근에 친환경 발전, 신재생 에너지 등이 늘어나는 추세지만 아직 화석연료가 차지하는 비중이 크다. 따라서 CO2 배출량이 많다는 것이 GDP가 높아지는 일이라고 생각할 수 있다.
- 2) 2~6개로 등급을 나눈 heatmap 들에서 인구수, 실업률, 국토 면적이 GDP와 관련이 거의 없다는 것을 알 수 있다.

나. 인식률 측정을 통한 인식 방법 선택

- 1) 2등급으로 나눈 경우는 균접 이웃 방식, 3등급으로 나눈 경우는 논리 회귀 방식, 4, 5, 6 등급으로 나눈 경우는 결정 트리 방식이 인식률이 가장 높았다. 이후 2번의 재측정 이후 가장 높은 횟수가 많았던 결정 트리 방식을 사용하기로 했다.
- 2) 등급을 나누는 수가 많아질수록 전체적인 인식률이 떨어진다.

다. 점그래프에서 알 수 있는 것

- 1) GDP와 관련도가 낮았던 인구수, 실업률, 국토 면적의 그래프는 아무런 사실을 알 수 없었다.
- 2) GDP와 관련도가 높았던 행복지수와 CO2의 배출량의 그래프는 형태가 보인다.
 - 가) 행복지수 & GDP 그래프는 행복지수가 높으면 GDP 또한 높아지는 비례 함수의 꼴을 나타내고 있다.
 - 나) CO2 배출량 & GDP 그래프는 환경 쿠즈네츠 곡선의 형태와 유사했다.



2. <탐구 나>의 결론

가. forecast 함수를 사용하여 요인들의 미래의 값을 구할 수 있었다.

3. <탐구 다>의 결론

가. <탐구 나>에서 구한 요인들의 값으로 국가별 미래의 GDP를 구한다.

- 1) 중국에 관한 결과 분석

가) 최근 뉴스들을 보면, 중국의 미래 경제성장 전망을 긍정적으로 보는 기사들이 있었다. 하지만, 대다수의 블로그와 기사들에서 분석한 중국의 경제 전망은 조금 달랐다. 2000

년대 초 중국이 세계 수출시장에서 점유율이 불과 3%에 불과했지만, 2010년도부터 최대 12% 상승률을 보였다. 하지만 이런 정부의 경제정책은 점점 한계점에 다다랐다. 그 래프를 통하여 중국의 경제성장률이 점점 감소하며 GDP 그래프가 수평선에 가까워짐을 알 수 있었다.

2) 일본에 관한 결과 분석

가) 전문가들은 일본의 경제 전망을 낙관적으로 본다. 2012년 12월 이후 장기간 동안 안정적인 경제 상황이 실제로 지속되고 있는 상황이다.

나) 일본 인구의 노동생산성이 증가하면서 경제적인 안정이 찾아올 것이라는 예측

3) 덴마크에 관한 결과 분석

가) 완만한 호황이 계속될 전망

나) 전문가들은 견고한 경기 회복세가 경제성장에 영향을 미칠 것이라는 예상

다) 덴마크의 경제성장률 긍정적

라) 2010년대부터 시작된 덴마크 정부의 구조적 정책은 균형적인 경제 성장세를 가져올 것이라는 예상

4) 한국에 관한 결과 분석

가) 경제전문가들은 은행사업, 해운사업 등 특정 분야에서 경제적 하락이 찾아올 것이라고 예상한다.

나) 한국은행에서 한국경제보고서를 산출해냈다. 물가는 상승하고, 경기가 둔화된다. 그러므로 경제성장률이 낮아질 것이다.

5) 인도에 관한 결과 분석

가) 인도 금융시장이 창출하는 부가가치는 지속해서 증가 전망이다. 인도는 젊고 풍부한 노동력, 시장경제와 민주주의 전통, 과학기술 분야 우수성, 영어 사용능력 등 다대한 중장기적 경제성장 잠재력을 보유하고 있음.

V. 참가소감

우리가 공통적으로 AI라는 주제에 관심이 끌렸던 것은 사람의 역할을 대신, 또는 편리하게 해주는 것을 스스로 할 수 있게 된다는 머신러닝에 있었다. 먼저, 우리는 처음에 머신러닝과 딥러닝의 차이를 몰랐다. 그리고 자료를 조사하던 중 그 차이를 알게 되었는데 그때 우리는 다시 한번 이 주제에 대해서 영감을 받았다. 우리가 배운 머신러닝이라는 것은 아무래도 학습이라는 것을 시켜줘야 한다. 스스로 생각할 수 있는 딥러닝이 좋다고 생각할 수 있었지만 우리는 그렇게 생각하지 않았다. 우리가 필요한 것만 정확히 알려주는 머신러닝에 더욱 흥미가 솟구쳤다. 그리고 우리의

첫 수업은 매우 인상에 남았다. 우리가 생각했던 수업은 오로지 인공지능을 스스로 생각할 수 있게 해주는 알고리즘과 그 알고리즘을 짜는 명령어 정도를 배우는 것으로 한정했기 때문이다. 한마디로, 우리가 머신러닝에 관해 아직도 잘 알고 있지 못했음을 알게 해주었다. 첫날 수업의 핵심은 아마 사람의 생각과 인공지능이란? 정도로 생각한다. 우리는 처음엔 잘 몰랐지만, 이 수업을 들은 후에 우리가 배운 것은 사람과 인공지능의 차이를 인식할 수 있게 되는 능력이었다. 그 덕에 본격적으로 배운 여러 명령어와 기능들을 한층 쉽게 이해할 수 있었다. 그 후 스스로 머신러닝을 실행 할 수 있도록 학습을 시키는 과정에서 복잡한 명령어와 과정이 다소 힘들기도 하였으나 납득이 갈 만한 결과가 나와 뿌듯하였다. 우리는 이번 공학 아카데미를 계기로 실력을 키워 더욱 발전할 수 있었다. 다음에도 이러한 기회가 있다면 본래의 지식을 갈고닦고, 새로운 지식을 받아들이기 위해 적극적으로 참여할 것이다.

머신러닝을 이용한 농구선수 포지션 추천 시스템 개발

제주제일고등학교 1학년

김서현·양성현·전승재

I. 서 론

1. 탐구 동기

먼저 “머신러닝을 이용해 어떤 데이터를 분석해볼까?” 하는 생각이 들었다. kaggle이라는 사이트에서 웹서핑을 하다가 농구선수 약 2만 명의 (키, 몸무게, 포지션, BMI 등) 데이터 셋을 발견하였다. 그때 우리는 농구선수들의 키, 몸무게는 포지션에 영향을 얼마나 주는지에 대해 의문이 생겼다. 우리는 키와 몸무게에 따른 포지션 데이터를 시각화해 보고, 인공지능에 학습시켜 신체조건에 따른 적합한 포지션을 추천하는 인공지능을 개발하기로 결정하였다.

2. 탐구 목적

주어진 Dataset을 시각화하여 농구선수의 신체 조건과 포지션의 상관관계를 파악한다.

주어진 Dataset을 인공지능에게 학습시켜 인공지능이 농구선수의 신체조건에 따른 최적의 포지션을 배정하는 프로그램을 만들고자 한다.

3. 탐구 목표

가. 데이터 정리 및 시각화

- 데이터를 Heatmap으로 시각화 하여 각 특징들의 상관관계를 확인한다. 그 뒤 가장 상관 관계가 높은 특징들을 선별한다.
- 데이터를 Graph으로 시각화 하여 각 특징에 따른 2만 명의 선수 데이터의 분포를 한눈에 볼 수 있도록 한다.

나. 머신러닝을 이용한 데이터 분류

- 약 2만 개의 농구선수 데이터를 키, 몸무게, 나이를 기준으로 분류하여 가장 잘 맞을 포지션을 추천해주고자 한다.

II. 탐구 방법

1. 탐구 재료 및 기구

가. 준비물 : 노트북

나. 기구 : 노트북, WIFI

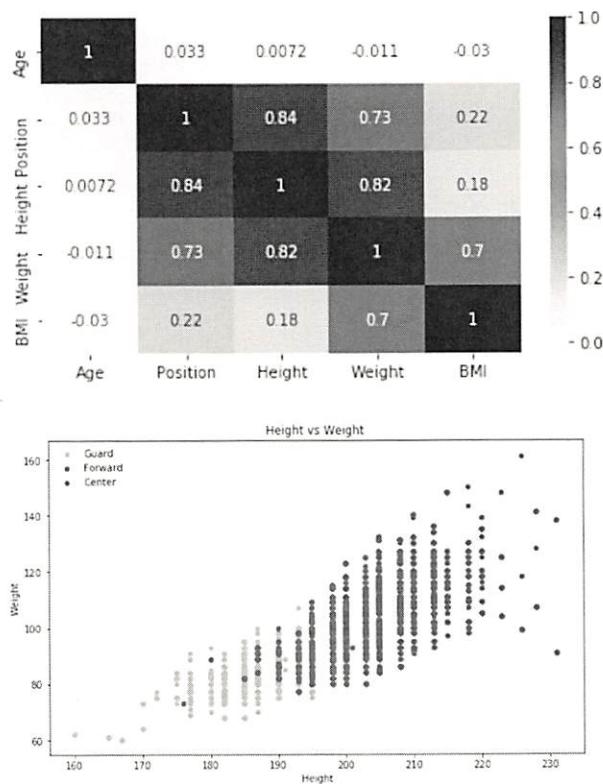
2. 탐구 과정

- 1) kaggle에서 .csv파일을 불러온 후 Excel 통해 데이터 자료를 정리한다.
- 2) python 모듈 pandas를 이용하여 분류한 데이터를 시각화 한다.
- 3) 그 데이터를 써포트 벡터 머신(SVM), 논리 회귀(Logistic Regression), 결정 트리 (Decision Tree), 근접 이웃(K-Nearest Neighbours) 알고리즘으로 분류하고 그중 가장 인식률이 높은 알고리즘을 선택한다.
- 4) 그 선택된 알고리즘을 이용하여 포지션과 상반관계가 큰 키, 나이, 몸무게를 입력하여 포지션 추천 결과를 얻는다.
- 5) 해당 유저가 입력한 값과 포지션 추천 값을 기존 데이터에 추가하여 사람들이 많이 이용 할수록 인식률 및 정확도를 높힌다.

III. 탐구결과 및 고찰

1. 탐구 결과

- 가. 데이터를 Heatmap을 이용하여 분석한 결과 선수의 포지션의 상관관계는 그 선수의 키와 몸무게에서 가장 높게 나왔다.
- 나. 데이터에 따르면 키가 크고 몸무게가 크면 클수록 가드 -> 포워드 -> 센터 순으로 포지션이 적절하다고 나왔다.
- 다. 4개의 알고리즘 중 주로 써포트 벡터 머신(SVM)이 인식률이 가장 높게 나왔는데, 그때의 인식률이 약 80% 대 였다.



인식률 계산 중입니다...

SVM 인식률: 80.93242535358827

LR 인식률: 79.51807228915662

TREE 인식률: 78.7323205866946

KNC 인식률: 77.26558407543216

써포트 벡터 머신(SVM) 알고리즘을 사용합니다.

IV. 결론

Heatmap을 통해 포지션과 가장 상관관계가 큰 특징은 키와 몸무게라는 것을 알 수 있었다. 데이터 측면에서 보면 키가 작고 몸무게가 작으면 주로 가드, 키가 크고 몸무게가 많이 나가면

센터, 그 중간 값은 포워드라는 결과가 도출되었다.

알고리즘의 인식률은 상황에 따라 달라지므로 여러 알고리즘들 중 하나의 알고리즘을 선택하여 사용하는 것이 좋았다. 인식률이 달라지는 원인은 학습데이터와 테스트 데이터를 나누는 함수인 `train_test_split` 함수가 Dataset을 무작위로 세팅해 나눈 것 때문이라고 예측하였다.

상관관계가 낮은 특징들을 포함하여 인공지능에 학습시키면 인식률의 변화는 크지 않다는 것을 알았다.

V. 참가소감

이번 공학아카데미에 참가하면서 데이터 과학과 인공지능은 매우 밀접한 관계가 있다는 것을 알게 되었고, 매우 적은 부분이지만 간접적으로 나마 인공지능 개발을 경험한 것이 매우 의미 있었다. 웹사이트에서 데이터를 찾고 정리하고 머신러닝을 이용하여 데이터를 분류, 예측하는 방법을 터득 할 수 있었다. 터득한 내용을 바탕으로 활용할 수 있는 기회가 되었다.

머신러닝을 활용한 제주 관광객 증감 요인 분석 및 예측

제주중앙여자고등학교 2학년

김영민·양소연·정지원

I. 서 론

1. 탐구 동기

최근 한국공항공사 제주지역본부는 여름철 관광 성수기(7월 25일~8월 11, 18일)에 항공기 운항 편수는 9003편으로, 전년 동기 대비 2.0% 증가할 것으로 전망하고 있다. 또한, 현재 제주도에는 성수기에 하루 평균 8만 9천여명이 오고 가는 등 제주도로 여행을 오는 사람들이 증가하고 있다는 것을 알 수 있다. 2017년 ‘효리네 민박’이 방영할 당시 효리네 민박 효과로 제주 관광객이 급증했다.

이처럼 제주도 관광객 수는 해마다 증가하고 있다. 제주도 관광객 수가 증가하면 수익이 늘어나 경제 성장률이 증가할 것이다. 그러나 만약 제주도 관광객 수가 급증한다는 사실을 예측하지 못한 경우에는 어떻게 될까? 제주도 관광객 수가 급증한다면, 그리고 관광 산업 분야에서 이를 예측하지 못했다면, 교통 산업 분야에서 관광객 수에 맞춰 교통편을 조절하지 못해 성수기에 관광객이 제주도로 이동할 교통편이 부족해지는 문제점이 발생할 수 있다.

그렇다면, 반대로 제주도 관광객 수가 급감한다는 사실을 예측하지 못한 경우에는 어떤 상황이 발생할까? 이 같은 경우에는 관광 수입이 감소하고 물량이 남아 경제성장률이 마이너스를 기록하는 상황이 발생할 것이다. 따라서 본 탐구는 위에서 언급한 문제점을 해결하기 위해 제주도 관광객 수 증감에 영향을 미치는 요인들을 파악하고, 머신러닝 기술을 활용하여 월별로 들어오는 제주도 관광객의 수를 예측하기 위한 정확도가 높은 프로그램을 만드는 것을 목표로 한다. 더 나아가, 이 프로그램을 관광이나 유통 등의 산업 분야에 적용시켜 예측한 제주도 관광객 수에 따라 수요를 조절하고 예측할 수 있도록 할 것이다.

2. 탐구 목적

제주도 관광객 수에 영향을 미치는 요인들에 대해 파악하고 상관관계가 가장 큰 요인으로 년도마다 제주도로 들어오는 관광객의 수를 예측하여 관광이나 유통 등의 사업분야에 적용시켜 예측 수요에 따라 조절한다.

3. 탐구 목표

- 가. 국내 4개 지역의 관광객 증감 요인에 대해 파악하고 시각화하여 분석한다.
- 나. 분석하여 선정한 증감 요인을 토대로 제주도 관광객 수를 예측 프로그램을 만든다.
- 다. 개발한 프로그램을 관광 및 유통 등 산업 분야에 적용할 수 있는 방안을 모색한다.
- 라. 전 세계에서 들어오는 제주의 외국인 관광객 수를 예측한다.

II. 탐구 방법

1. 개발환경

가. 사용 언어

python(파이썬): 컴퓨터 언어의 일종으로 간결하고 생산성 높은 프로그래밍 언어이다. 외부에 풍부한 라이브러리가 있어 다양한 용도로 확장하기 좋다. 실제로 파이썬은 웹 개발뿐만 아니라 데이터 분석, 머신러닝, 그래픽, 학술 연구 등 여러 분야에서 활용되고 있다. 생산성이 높은 것도 큰 장점이다. 속도가 느리다는 평가도 있으며, 모바일 앱 개발 환경에서 사용하기 힘들다. 또한 컴파일 시 타입 검사가 이뤄지지 않아 개발자가 실수할 여지가 조금 더 많다거나 멀티코어를 활용하기 쉽지 않다는 지적도 있다.

나. kaggle 프로그램 참고

1) 자료 시각화

```
plt.figure(figsize=(10,6))
sns.plotting_context('notebook',font_scale=1.2)
cols = ['sqft_lot','sqft_above','sqft_living', 'bedrooms','grade','price']
```

```

g = sns.pairplot(train_df[cols], hue='bedrooms', size=2)
g.set(xticklabels=[ ])

plt.figure(figsize=(10,6))
sns.plotting_context('notebook', font_scale=1.2)
cols = ['sqft_lot', 'sqft_above', 'sqft_living', 'bedrooms', 'grade', 'price']
g = sns.pairplot(train_df[cols], hue='bedrooms', size=2)
g.set(xticklabels=[ ])

```

2) 데이터 학습 과정

```

gildong = LinearRegression()
gildong.fit(train_df_part1[features], train_df_part1['price'])
score = gildong.score(train_df_part2[features], train_df_part2['price'])
print(format(score, '.3f'))

```

```

gildong = LinearRegression()
gildong.fit(train_df_part1[features], train_df_part1['price'])
score = gildong.score(train_df_part2[features], train_df_part2['price'])
print(format(score, '.3f'))

```

```

gildong = LinearRegression()
gildong.fit(train_df_part1[features], train_df_part1['price'])
score = gildong.score(train_df_part2[features], train_df_part2['price'])
print(format(score, '.3f'))

```

2. 탐구 과정

가. 제주도 관광객 증감 요인 설정

- 1) 날씨
- 2) TV 프로그램
- 3) 날짜/시기
- 4) 미디어

나. 각 요인별 데이터 수집 및 정리

1) 날씨

- 지역별 기온과 강수량

기상청 홈페이지에서 2017년도 월별 평균 강수량, 기온 조사 후 데이터를 정리한다.

2) TV 프로그램

- 지역 관련 예능 프로그램 시청률

지역을 배경으로 한 특정 TV프로그램에서 4개 지역(제주, 강릉, 부산, 여수)을 선정하여 각 지역별 방영 당시 시청률 조사

3) 날짜/시기

- 날짜 범위 설정

2017년 1월~2018년 12월 최근 2년간 요인별로 월별 데이터를 수집하여 날짜/시기에 대해 예측할 배경 정보로 사용한다.

4) 미디어

- 네이버 검색어

네이버 검색어 중 그 지역의 관광과 관련된 단어 일일 검색 통계자료를 조사하여 월별 지역 관광 관련 평균 검색 현황 파악

년월	1	2	3	4 water
17-Jan	49	55.6	50.7	19.6 43.725
17-Feb	81.8	86.9	39.1	39.6 61.85
17-Mar	36.3	60.4	48.2	34.2 44.775
17-Apr	222.5	106.7	51.4	85.1 116.425
17-May	59.2	66.4	38.5	31.1 48.8
17-Jun	247.4	141.2	60.7	114.8 141.025
17-Jul	51.8	427.5	35.2	23.1 134.4
17-Aug	231.8	437.3	159.5	279.7 277.075
17-Sep	102.8	98.3	83.4	58.3 85.7
17-Oct	231.4	389.1	162	152.4 233.725
17-Nov	6.7	16.5	19.2	6.7 12.275
17-Dec	13.3	31.8	25.4	16.7 21.8
18-Jan	63.7	102.4	58	47.7 67.95
18-Feb	10.5	125.7	86.6	75 74.45
18-Mar	151.6	179.5	118	101.2 137.575
18-Apr	312	234.1	112.5	175.6 208.55
18-May	356.1	151.8	98.8	256.7 215.85
18-Jun	304.3	377.2	211.1	186.7 269.825
18-Jul	23.2	41.8	48.7	20.8 33.625
18-Aug	159.6	125	376.5	127.2 197.075
18-Sep	473.8	521.8	187.7	136.6 329.975
18-Oct	210.2	184.2	347	121.2 215.65
18-Nov	55.6	45.9	28.6	42.1 43.05
18-Dec	45.6	57.3	96	55 63.475

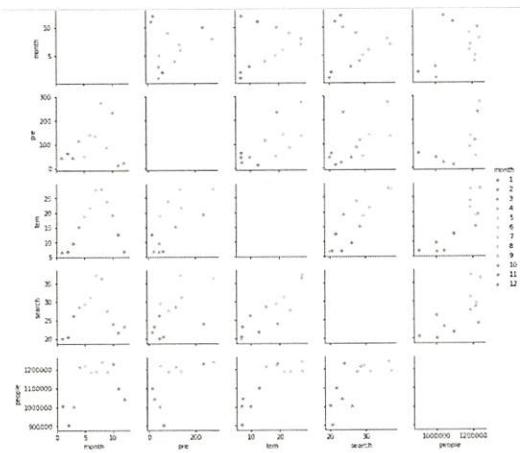
[그림 1. 지역별 강수량 데이터]

년월	지점	평균기온(℃)	최저기온(℃)	최고기온(℃)
Jan-17	108	-1.8	-12.6	11.4
Feb-17	108	-0.2	-9.3	12.3
Mar-17	108	6.3	-5.3	18.9
Apr-17	108	13.9	3.1	27.8
May-17	108	19.5	10	30.3
Jun-17	108	23.3	14.5	34.1
Jul-17	108	26.9	21.9	35.4
Aug-17	108	25.9	16.1	35.3
Sep-17	108	22.1	11.2	31.4
Oct-17	108	16.4	2.5	29.4
Nov-17	108	5.6	-6.6	18.4
Dec-17	108	-1.9	-12.3	8.7
Jan-18	108	-4	-17.8	8.7
Feb-18	108	-1.6	-13.4	10.4
Mar-18	108	8.1	-6.7	22.1
Apr-18	108	13	0.1	26.3
May-18	108	18.2	6.9	29.6
Jun-18	108	23.1	16.3	32.9
Jul-18	108	27.8	17.7	38.3
Aug-18	108	28.8	20.2	39.6
Sep-18	108	21.5	10.8	30.9
Oct-18	108	13.1	0.7	25.6
Nov-18	108	7.8	-3.1	19.4
Dec-18	108	-0.6	-14.4	13.5

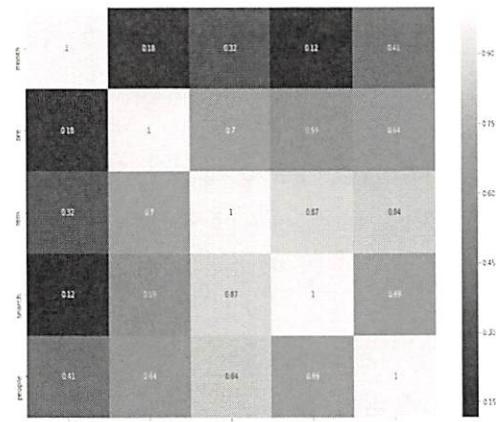
[그림 2. 지역별 기온 데이터]

다. 데이터를 기반으로 한 CSV 파일 변환

라. 데이터 시각화 및 분석

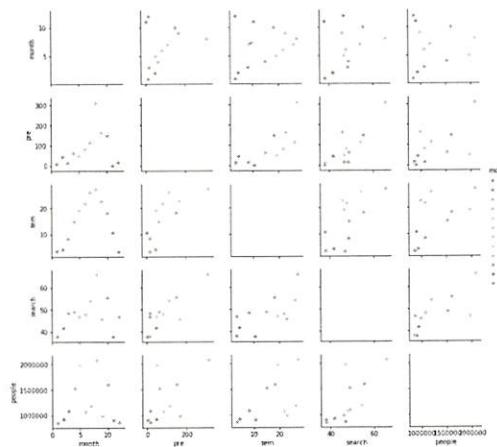


[그림 3. pairplot 제주 관광객 증감
각 요인별 분석]

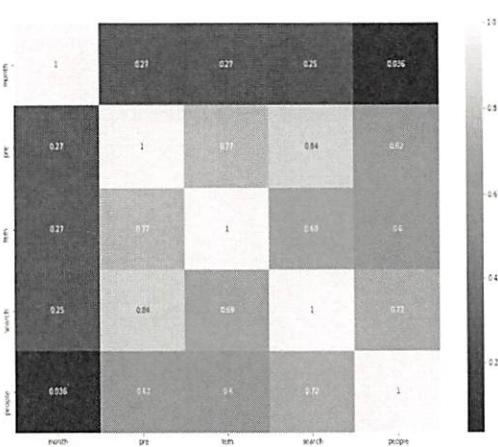


[그림 4. heatmap 제주 관광객 증감
각 요인별 분석]

제주도의 시기별 관광객 증감은 4월~8월, 즉 봄과 여름에 비교적 관광객 수가 많았고 겨울에는 관광객 수가 가장 적었다. 강수량에 따른 관광객 수는 강수량이 많은 시기인 7~8월에 관광객 수가 많았고, 강수량이 적은 시기인 12월~3월에 관광객 수가 적었다. 기온이 온난한 5월부터 9월 관광객 수가 많은 것을 알 수 있었고 10월~4월 사이 비교적 관광객 수가 적었다. 연관 검색어 중 제주 관광 관련 검색을 가장 많이 한 시기는 5~8월이고 5~10월 사이 관광객 수가 많은 것을 볼 수 있다. 위의 두 가지 결과 자료를 통해 분석, 예측해 본 결과 제주도의 관광객 증감은 시기와 가장 연관이 있으며 시기별 특징인 기온, 강수 등의 요인에도 영향을 미친다.

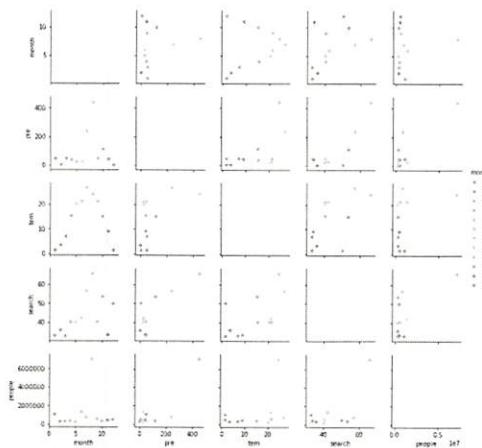


[그림 5. pairplot 여수 관광객 증감
각 요인별 분석]

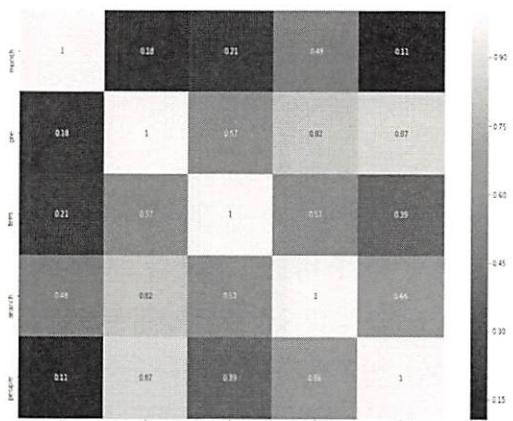


[그림 6. heatmap 여수 관광객 증감
각 요인별 분석]

여수의 시기별 관광객 증감은 4월~5월, 7월, 10월에 비교적 관광객 수가 많았고 12월 ~2월에는 관광객 수가 비교적 적었다. 강수량에 따른 관광객 수는 강수량이 비교적 적은 4월과 강수량이 가장 많은 8월에 관광객 수가 많았고, 그 외에는 강수량이 비교적 적었다. 기온이 온난한 5월부터 10월 관광객 수가 많은 것을 알 수 있었고 11월~4월 사이 비교적 관광객 수가 적었다. 연관 검색어 중 여수 관광 관련 검색을 가장 많이 한 시기는 5~8월이고 5~10월 사이 관광객 수가 많은 것을 볼 수 있다. 위의 두 가지 결과 자료를 통해 분석, 예측해 본 결과 여수의 관광객 증감은 기온이 온난한 시기에 관광객이 가장 많은 것을 보아 기온이 관광객 증감에 영향을 미친다.



[그림 7. pairplot 강릉 관광객 증감 각 요인별 분석]

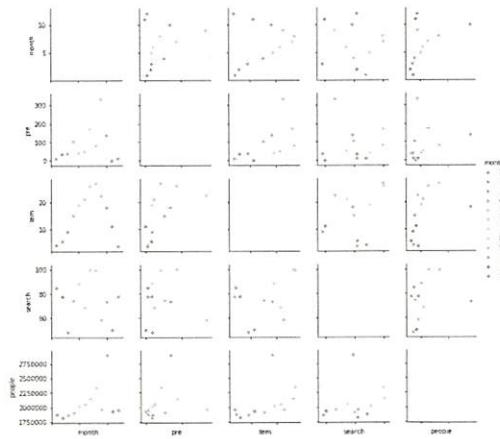


[그림 8. heatmap 강릉 관광객 증감 각 요인별 분석]

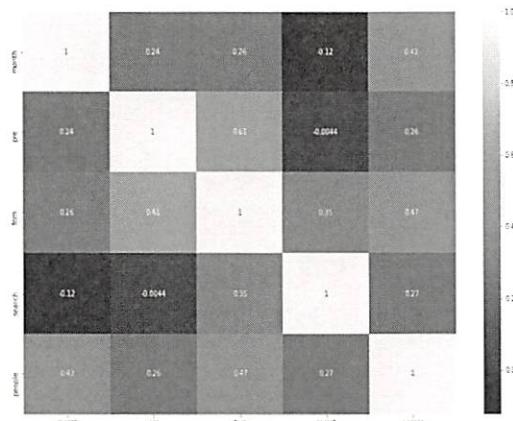
강릉의 시기별 관광객 증감은 1월과 6월~8월에 비교적 관광객 수가 많았고 2월~4월에는 관광객 수가 비교적 적었다. 강수량에 따른 관광객 수는 강수량이 많은 시기인 8월에 월등하게 관광객 수가 많았고, 강수량이 적은 시기인 그 외의 시기에는 관광객 수가 적었다. 기온이 높은 8월 관광객 수가 많은 것을 알 수 있었고 10월~4월 사이 비교적 관광객 수가 적었다. 연관 검색어 중 강릉 관광 관련 검색을 가장 많이 한 시기는 8월이고 6월~8월 사이 관광객 수가 많은 것을 볼 수 있다. 위의 두 가지 결과 자료를 통해 분석, 예측해 본 결과 강릉의 관광객 증감은 8월만 모든 결과에서 월등히 관광객 수가 많았으므로 시기와 가장 연관이 있으며 시기별 특징인 기온, 강수 등의 요인에도 영향을 미친다.

03

AI 러닝머신으로 배우는 데이터 과학, 똑똑한 컴퓨팅



[그림 9. pairplot 부산 관광객 증감
각 요인별 분석]



[그림 10. heatmap 강릉 관광객 증감
각 요인별 분석]

부산의 시기별 관광객 증감은 4월~9월까지 증가하였고 9월에 가장 관광객이 많았다. 10월~2월에는 관광객 수가 비교적 적었다. 강수량에 따른 관광객 수는 8월이 강수량이 가장 많았지만 관광객은 비교적 적고 9월에 강수량은 비교적 적지만 관광객 수는 가장 많다. 기온에 따른 관광객 수는 기온이 비교적 높은 5월~8월에는 관광객 수가 적었고 온난한 9월 관광객 수가 많다. 연관 검색어 중 부산 관광 관련 검색을 많이 한 시기는 4월~8월이지만 관광객 수가 비교적 적고 9월에 관광객 수가 많은 것을 볼 수 있다. 위의 두 가지 결과 자료를 통해 분석, 예측해 본 결과 부산의 관광객 증감은 9월만 모든 결과에서 월등히 관광객 수가 많았으므로 시기와 가장 연관이 있으며 시기별 특징인 기온, 강수 등의 요인에도 영향을 미친다.

	A	B	C	D	E	F	
1	month	pre	tem	search	people	제주	
2	1	43.725	6.65	20.03813	1,005,438		
3	2	61.85	6.775	20.51777	905,821		
4	3	44.775	9.65	26.17955	1,000,871		
5	4	116.425	15.15	28.55828	1,214,588		
6	5	48.8	18.75	29.44585	1,219,337		
7	6	141.025	21.65	31.19541	1,187,388		
8	7	134.4	27.9	37.23628	1,191,311		
9	8	277.075	28	36.4325	1,240,389		
10	9	85.7	23.75	27.59208	1,186,048		
11	10	233.725	19.175	23.95147	1,229,679		
12	11	12.275	12.7	21.72643	1,097,987		
13	12	21.8	6.875	23.27598	1,043,775		

[그림 11. 제주의 각 요인별 데이터]

	A	B	C	D	E	F	
1	month	pre	tem	search	people	여수	
2	1	11.5	3.4	38.07479	861419		
3	2	45.1	4.3	41.76578	928678		
4	3	14.7	8.3	48.49696	1083662		
5	4	62.8	14.9	48.99174	1525028		
6	5	49.5	19.2	46.72445	1982849		
7	6	81.6	21.7	47.77465	1071816		
8	7	113.5	26.1	54.21441	1189305		
9	8	311.6	27.2	65.89268	2078157		
10	9	163.9	22.7	45.75033	985762		
11	10	148.3	18.1	55.42521	1596797		
12	11	0.6	10.6	37.79871	906505		
13	12	16.5	3.3	46.85782	873536		

[그림 12. 여수의 각 요인별 데이터]

	A	B	C	D	E	F	G
1	month	pre	tem	search	people	area	강릉
2	1	48.5	1.5	33.10718	1066421	2	
3	2	3.5	3.5	35.8266	313497	2	
4	3	48.5	7.1	33.1018	298183	2	
5	4	39.8	15.4	40.31872	375468	2	
6	5	25	20.3	40.14198	273691	2	
7	6	27.2	21.2	42.1986	1335580	2	
8	7	238.1	26.8	56.86373	815772	2	
9	8	444.1	24.3	65.87023	7094478	2	
10	9	45.7	21.3	40.64336	571185	2	
11	10	114.5	15.2	53.60353	345983	2	
12	11	41.5	9	33.67733	437550	2	
13	12	5.5	1.5	50.24307	491984	2	

[그림 13. 강릉의 각 요인별 데이터]

	A	B	C	D	E	F
1	month	pre	tem	search	people	부산
2	1	12	4.1	84.98442	1,878,400	
3	2	33.8	5.5	77.62397	1,828,607	
4	3	35.7	9	48.02646	1,867,876	
5	4	105.1	15	74.56469	1,908,158	
6	5	39.2	19	88.39547	2,018,860	
7	6	49.8	21.3	68.71529	2,046,082	
8	7	172.1	26.1	100	2,142,403	
9	8	82.5	27	99.60446	2,338,653	
10	9	335	22.6	58.49384	1,963,002	
11	10	138.3	18.1	73.59118	2,897,177	
12	11	0.3	11.2	49.90786	1,925,853	
13	12	10.6	3.6	77.46704	1,954,268	

[그림 14. 부산의 각 요인별 데이터]

마. 분석 자료를 통한 예측 요인 선정

각 지역별 요인과 데이터를 분석해보았을 때 2016년도부터 2019년도까지의 제주도 데이터를 통해 키워드와 관광객수와의 상관관계는 없었다. 제주도의 관광객 수에 기온이고, 강수량, 시기순으로 영향을 미친다.

바. 제주 관광객 수 증감 예측 프로그램 구현

1) 변수 설정

```
In[2]:  
# CSV 파일 읽어오기  
gildong = pd.read_csv("../input/movement/jejun.csv")  
jiwon = pd.read_csv("../input/movement/yeosun.csv")  
soyeon = pd.read_csv("../input/movement/gangneungn.csv")  
yeongmin = pd.read_csv("../input/movement/busann.csv")
```

[그림 15. 예측 프로그램 코딩-변수 설정]

2) 모듈 선언

```
In[1]: import numpy as np # 수학 연산 수학을 위한 모듈
import pandas as pd # 데이터 처리를 위한 모듈
import seaborn as sns # 데이터 시각화 모듈
import matplotlib.pyplot as plt # 데이터 시각화 모듈

import os
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import explained_variance_score
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import svm
from sklearn import metrics
from sklearn.tree import DecisionTreeClassifier

# 어떤 파일이 있는지 표시하기
from subprocess import check_output
print(check_output(["ls", ".../input"]).decode("utf8"))

movement
testdata
testfile
```

[그림 16. 예측 프로그램 코딩-모듈 선언]

3) Pairplot 제작

```
In[6]: plt.figure(figsize=(10,6))
sns.set_context('notebook', font_scale=1.2)
cols = ['month', 'pre', 'tem', 'search', 'people']
g = sns.pairplot(gildong[cols], hue='month', size=2)
g = sns.pairplot(jiwon[cols], hue='month', size=2)
g = sns.pairplot(soyeon[cols], hue='month', size=2)
g = sns.pairplot(yeongmin[cols], hue='month', size=2)
g.set(xticklabels=[])
```

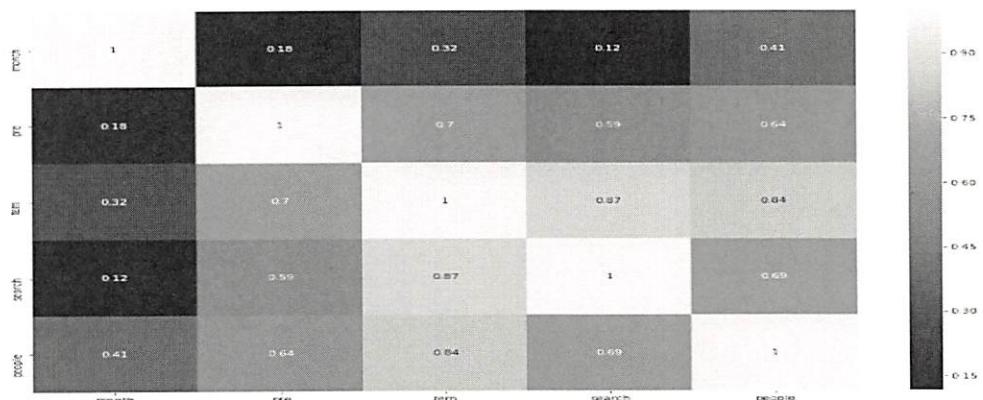
[그림 17. 예측 프로그램 코딩-Pairplot 제작]

4) Heatmap 제작

```
In[7]: plt.figure(figsize=(15,10))
columns = ['month', 'pre', 'tem', 'search', 'people']
sns.heatmap(gildong[columns].corr(), annot=True)

#CHECK THE PPT SLIDE
```

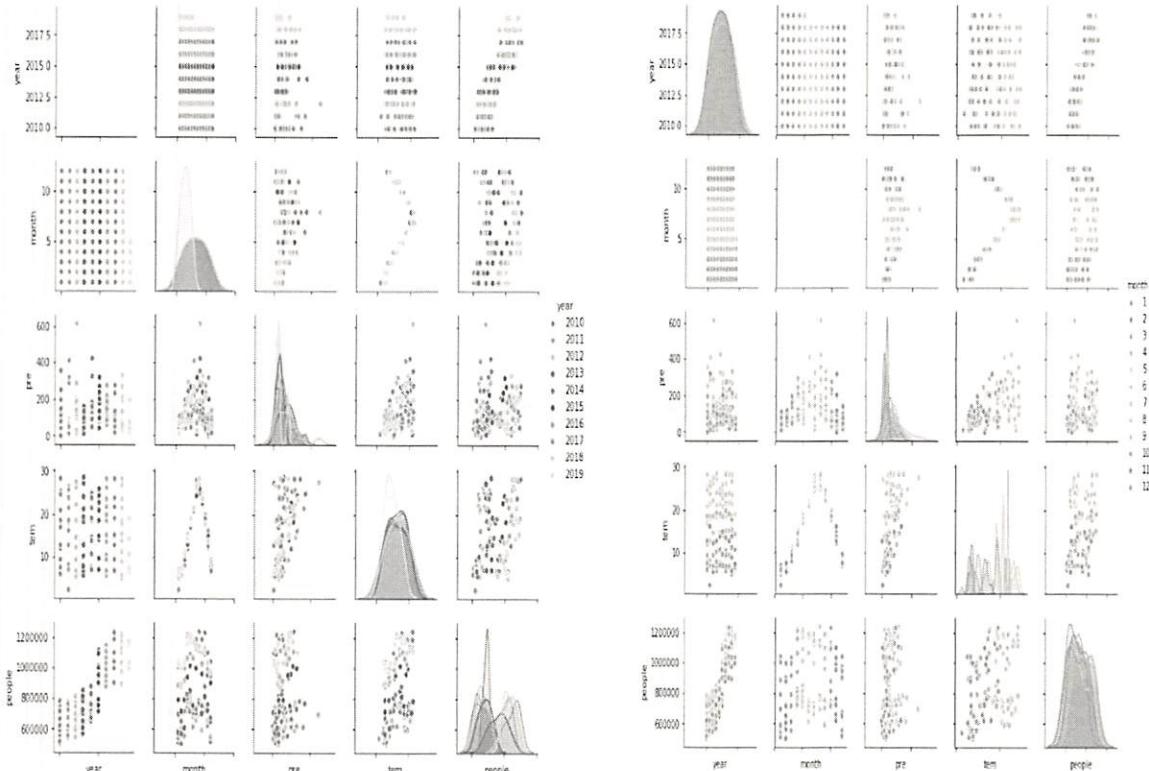
```
Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x7f695cf5400>
```



[그림 18. 예측 프로그램 코딩-Heatmap 제작]

III. 탐구결과 및 고찰

1. 탐구 결과



제주도 관광객 증감 요인을 파악하기 위해 지역별로 월별 관광객 수와 기온, 강수량, 연관 검색량, 날짜와의 상관관계를 파악하였다. 이를 통해 관광객 수에 영향을 미치는 요인은 기온, 강수량, 연관 검색량이라는 것을 알게 되었고 2016년부터 2019년 5월까지 제주도 관광객 수와 위 요인과의 상관관계를 알아보았다. 이는 지역별로 조사한 월별 관광객 수와 여러 요인들과의 상관관계 탐구에서 자료가 조금밖에 없었던 점을 보완하고자 탐구를 진행했다. 2016년부터 2019년 5월까지 제주도 관광객 수와 기온, 강수량, 연관 검색량과의 상관관계 탐구를 통해 제주도 관광객 수 증감 요인으로 기온, 강수량, 날짜라고 결론을 지었다.

1) SVM(서포트 벡터 머신) 알고리즘 이용하여 예측하기

```
In[43]: baby1 = svm.SVC()
baby1.fit(train_X, train_y)
prediction = baby1.predict(test_X)

plt.plot(prediction)
plt.plot(test_y.values.tolist())

Out[43]: <matplotlib.lines.Line2D at 0x7f0624fb4e48>
```

```
▶ score = baby1.score(train_df_part2[features], train_df_part2['people'])
print(score)
```

```
0.5555555555555556
```

- 예측 프로그램

```
baby1 = svm.SVC()
baby1.fit(train_X, train_y)
prediction = baby1.predict(test_X)
```

- 예측 내용 그래프 표현

```
plt.plot(prediction)
plt.plot(test_y.values.tolist())
```

- 정확도 도출

```
score = baby1.score(train_df_part2[features], train_df_part2['people'])
print(score)
```

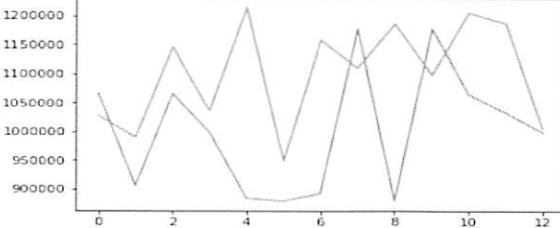
2) 근접 이웃 알고리즘(K-Nearest Neighbours) 이용하여 예측하기

```

▶ baby4 = KNeighborsClassifier(n_neighbors=3)
baby4.fit(train_X, train_y)
prediction = baby4.predict(test_X)

plt.plot(prediction)
plt.plot(test_y.values.tolist())

Out[49]:
[<matplotlib.lines.Line2D at 0x7f06207b8588>]



```

- 예측 프로그램

```

baby2 = LogisticRegression()
baby2.fit(train_X, train_y)
prediction = baby2.predict(test_X)

```

- 예측 내용 그래프 표현

```

plt.plot(prediction)
plt.plot(test_y.values.tolist())

```

- 정확도 도출

```

score = baby2.score(train_df_part2[features], train_df_part2['people'])
print(score)

```

3) 논리회귀 알고리즘 이용하여 예측하기

```
In[45]: baby2 = LogisticRegression()
baby2.fit(train_X, train_y)
prediction = baby2.predict(test_X)

plt.plot(prediction)
plt.plot(test_y.values.tolist())

Out[45]: [<matplotlib.lines.Line2D at 0x7f062086de48>]
```

```
score = baby2.score(train_df_part2[features], train_df_part2['people'])
print score
```

```
0.5555555555555556
```

- 예측 프로그램

```
baby3 = DecisionTreeClassifier()
baby3.fit(train_X, train_y)
prediction = baby3.predict(test_X)
```

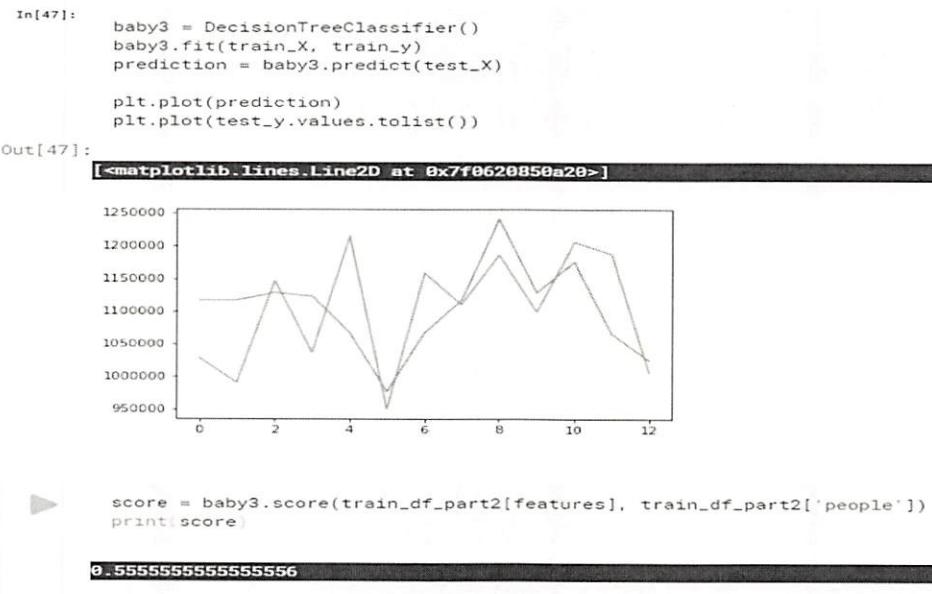
- 예측 내용 그래프 표현

```
plt.plot(prediction)
plt.plot(test_y.values.tolist())
```

- 정확도 도출

```
score = baby3.score(train_df_part2[features], train_df_part2['people'])
print(score)
```

4) 결정트리 알고리즘 이용하여 예측하기



- 예측 프로그램

```
baby4 = KNeighborsClassifier(n_neighbors=3)  
baby4.fit(train_X,train_y)  
prediction = baby4.predict(test_X)
```

- 예측 내용 그래프 표현

```
plt.plot(prediction)  
plt.plot(test_y.values.tolist())
```

- 정확도 도출

```
score = baby4.score(train_df_part2[features], train_df_part2['people'])  
print(score)
```

두 번의 탐구를 통해 분석해 본 제주도 관광객 수 증감 요인을 토대로 ‘제주도 관광객 수 예측 프로그램’을 만들기를 진행했다. 2010년부터 2019년까지 제주도 관광객 수와 기온, 강수량, 날짜의 데이터 값을 토대로 SVM (Support Vector Machine) 알고리즘과 논리회귀 알고리즘, 결정 트리 알고리즘, 근접이웃 알고리즘. 총 4개의 알고리즘을 활용해 머신러닝을 한 결과, SVM 알고리즘, 논리회귀 알고리즘, 결정트리 알고리즘이 가장 높은 정확도를 보였다.

2. 고찰

초기 요인 설정을 하였을 때 블로그 포스트, SNS, 축제와 행사 같은 여러 요인들을 고려했는데 데이터 수집을 하는 과정에서 요인별 데이터가 설정한 시기마다 데이터가 없는 경우가 있어서 요인들의 개수가 점점 줄었는데 다량의 데이터를 수집할 수 있는 요인을 더 설정했다면 정확도가 더 높아졌을 것으로 예상된다. 또한, 데이터 수집의 기간이 2010년~2019년이었는데 더 긴 기간을 설정하고 많은 데이터를 수집하였다면 정확성을 높일 수 있었을 것이다.

IV. 결론

우리는 두 차례에 걸쳐 제주도 관광객 증감 요인이 기온, 강수량, 날짜라는 것을 파악했고, 이 증감 요인을 토대로 제주도 관광객 수 예측 프로그램을 만들었다. 프로그램을 만들 때, 4개의 알고리즘(SVM 알고리즘, 논리회귀 알고리즘, 결정트리 알고리즘, 근접이웃 알고리즘)을 활용해 각각의 정확도를 알아본 결과 SVM 알고리즘과 논리회귀 알고리즘, 결정트리 알고리즘이 우리가 만든 제주도 관광객 수 예측 프로그램에 적합하다는 결론을 내렸다.

앞서 서론부분에서 언급한 바와 같이 우리가 프로그램을 만드는 이유는 관광 및 유통 등의 산업 분야에 이 프로그램을 접목시켜 제주도 관광객 수에 따라 수요를 예측하고, 경제성장률이 감소하는 것을 예방할 수 있을 것이다.

V. 참가소감

IT분야에 관심을 갖게 된 계기가 영화 ‘아이언 맨’에서 AI 서비스를 보고 ‘자비스와 같은 인공지능을 나도 만들고 싶다.’라고 생각해서인 만큼, 나는 중학교 때부터 IT, 특히 인공지능 분야에 관심이 많았다. 그래서 나는 공학 아카데미 프로그램 중에서 ‘AI 머신러닝으로 배우는 데이터 과학, 똑똑한 컴퓨팅’ 프로그램에 신청했다. 이 프로그램을 통해 단기간에 머신러닝에 대해 배우고, 이를 활용한 프로그램을 만들며 이제까지 어렵다고만 생각했던 인공지능을 만드는 것에 대해 수학적 지식(행렬, 선형대수학 등)이 그다지 높지 않은 나도 할 수 있다는 자신감을 얻게 되었다. 이 프로그램이 끝나서도 공학 아카데미를 통해 얻은 지식을 활용한 프로그램을 스스로 만들어보고 싶다는 생각을 하게 되었다. 인공지능과 머신러닝, 데이터 과학에 대해 배울 기회를 주신 제주 특별자치도교육청 분들과 제주대학교 교수님들과 멘토 선생님분들께 감사의 말씀을 드리고 싶다. (김영민)

고등학교 1학년 때 머신러닝에 대해 배울 때는 이론적으로 머신러닝의 종류와 빅데이터를 기반으로 학습을 시킨다는 내용을 배웠었다. 하지만 이러한 머신러닝은 전문가들만 사용이 가능한 어려운 내용인줄 알았다. 또한 파이썬을 배울 때도 파이썬을 사용하여 인공지능을 만든다는 내용을 들었는데 지금의 나와는 거리가 먼 내용이라고 생각했었다. 하지만 이번 2019 고등학생 공학아카데미에 참석하여 수업을 들으면서 파이썬이 어떻게 머신러닝을 만들 때 사용이 되는지 직접 코딩을 하면서 배우니 더 쉽게 받아 들어졌다. 또한 제주 관광객을 통계낸 자료를 2010년부터 2019년까지 총 9년동안의 통계자료를 사용하여 예측프로그램을 만들었는데 이 자료도 적다고 하시는 걸 듣고 정말 빅데이터가 머신러닝에 왜 꼭 필요한지 알 수가 있었다. 이번 19년 7월 22일부터 7월 25일 4일간 수업을 듣고 제주 관광객 예측프로그램을 만들면서 머신러닝과 파이썬의 많은 부분을 추가적으로 알게 되었다. (양소연)

인공지능의 주요 기술인 머신러닝을 직접 프로그래밍으로 실습해보며 배우고 주제 선정을 통해 프로그램을 구축하는 과정이 흥미로웠다. 머신러닝은 높은 수준의 기술이기 때문에 지금까지 공부해왔던 언어 내에서 구현할 수 없다고 생각하였는데 파이썬 언어로 구현해낼 수 있었다. 또한, 탐구 주제 특성상 kaggle에 데이터 자료가 없어서 직접 자료를 찾아와야 했는데 2010년~2019년까지 여러 자료를 도출해내는 과정에서 각 요인에 따라서 다른 자료를 참고하여 데이터를 뽑아와야 했기 때문에 많은 시간을 필요로 했다. 하지만 직접 수집한 데이터를 기반으로 머신러닝 프로그램을 구축하는 것이 흥미로웠다. 프로그램이 스스로 데이터를 학습하고 실행을 거듭할수록 정확성이 높아지는 수치를 바로 확인할 수 있어서 빅데이터를 활용하였다면 더 높은 정확성 갖춘 프로그램을 구축할 수 있었을 것 같다. 이번 2019 고등학생 공학아카데미에 참가하게 되어 인공지능과 머신러닝에 관련된 오픈소스를 참고할 수 있는 프로그램이나 직접 코딩하여 실행해볼 수 있는 사이트를 알게 되었고 이후, 이 탐구에서 조금 더 나아가 빅데이터를 활용한 예측, 분류 프로그램을 제작해보고 싶다.(정지원)

머신러닝을 이용한 영화 수익 예측

대기고등학교 2학년
강제호·이창원·황세호

I. 서 론

1. 탐구 동기

2019 고등학생 공학 아카데미 활동을 위해 많은 사람들이 알고 관심을 가질만한 주제를 찾던 도중 영화라는 주제를 선택하게 되었다. 그리고 인터넷 검색을 하던 도중 최근 개봉된 영화가 어마무시한 수익을 냈다는 기사를 보고 수익을 내기 위해 영화를 제작하기 전 수익을 먼저 예측하는 과정이 필요할 것이라는 생각을 가지게 되었다. 영화의 수익을 예측하기 위해 영화의 예산, 장르, 제작 나라 등의 요인으로 수익을 예측해보는 활동을 하면 흥미로울 것 같다는 생각을 가지게 되었으며 이 활동을 진행하기 위해 kaggle을 이용한 영화 수익 예측에 관하여 탐구하게 되었다.

2. 탐구 목적

우리 팀원들은 모두 컴퓨터공학과 진학을 목표로 하고 있다. 그러나 학교에서 학생의 역할을 수행하면서 컴퓨터에 대해 혼자 공부를 따로 하기에는 힘든 감이 없지 않아 있었다. 그렇기에 우리 팀은 이번 공학 아카데미를 통해 여러 가지 요인들이 영화 수익에 어떤 영향을 미치고 그에 따라 수익을 높이기 위해서 어떻게 해야 하는지 알아볼 것이다. 또한 이 활동을 통해 컴퓨터에 대한 지식을 쌓을 것이다.

3. 탐구 목표

- 가. 영화 수익에 영향을 미치는 요인들을 알아보고 영화 수익을 예측한다.
- 나. 다양한 데이터셋을 올바르게 사용하여 예측값의 정확도를 높인다.

II. 탐구 방법

1. 탐구 재료 및 기구

- 가. 준비물 : 노트북

2. 탐구 과정

가. 데이터 불러오기

- 1) 요인 : 예산, 장르, 제작사 나라, 출연진 수, 영화 상영시간
- 2) 모델링을 위해 K-fold, MiniMaxScaler, Cross_val_Score 알고리즘을 사용했다.
- 3) train.csv와 test.csv를 읽어온다.

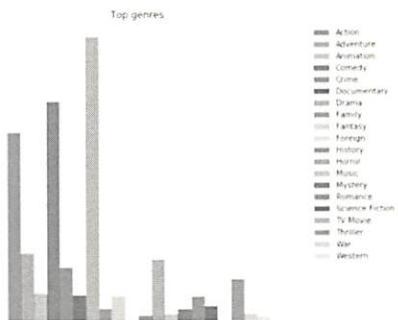
나. 시각화 하기

- 1) 장르에 따른 수익률

먼저 장르별로 어떤 데이터가 더 많은지 시각화로 확인한다. 그러기 위해선 먼저 genres라는 디렉터리를 새로 추가해 이름 별로 추가시켜 준다.

```
genres = {}
for i in data['genres']:
    if(not(pd.isnull(i))):
        if (eval(i)[0]['name']) not in genres:
            genres[eval(i)[0]['name']] = 1
        else:
            genres[eval(i)[0]['name']] += 1

df = pd.DataFrame([genres])
df.index = ['top genres']
df.plot(kind = 'bar', stacked = False, figsize = (12, 7))
plt.title('Top genres')
plt.axis('off')
plt.show()
```



장르별로 개수가 정확히 몇 개인지 알아보기 위해 OrderedDict를 이용해봅시다.

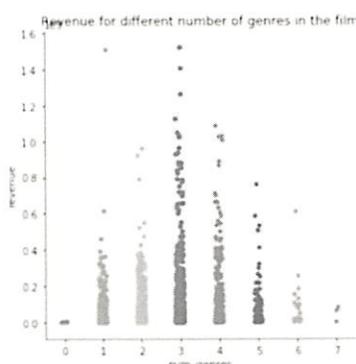
```
genres = OrderedDict(genres)
OrderedDict(sorted(genres.items(), key=lambda t: t[1]))
```

```
OrderedDict([('TV Movie', 1),
             ('Foreign', 2),
             ('Western', 13),
             ('History', 16),
             ('War', 28),
             ('Music', 29),
             ('Mystery', 33),
             ('Family', 36),
             ('Science Fiction', 41),
             ('Romance', 67),
             ('Fantasy', 68),
             ('Documentary', 71),
             ('Animation', 76),
             ('Thriller', 116),
             ('Crime', 147),
             ('Horror', 178),
             ('Adventure', 187),
             ('Action', 526),
             ('Comedy', 564),
             ('Drama', 785)])
```

train 데이터의 영화 속 다양한 장르의 수익에 대해 확인해 봅시다.

```
#adding number of genres for each movie
genres_count=[]
for i in data['genres']:
    if(not(pd.isnull(i))):
        genres_count.append(len(eval(i)))
    else:
        genres_count.append(0)
data['num_genres']=genres_count

#Genres v/s revenue
sns.catplot(x='num_genres', y='revenue', data=data);
plt.title('Revenue for different number of genres in the film');
```



이후를 위해 test 데이터의 영화 속 다양한 장르의 수와 수와의 상관관계를 확인할 코드를 만듭시다.

```
genres_count_test = []
for i in test['genres']:
    if(not(pd.isnull(i))):
        genres_count_test.append(len(eval(i)))
    else:
        genres_count_test.append(0)
test['num_genres'] = genres_count_test
```

마지막으로 genres 타이틀을 사용했으니 똑같이 제거해줍시다.

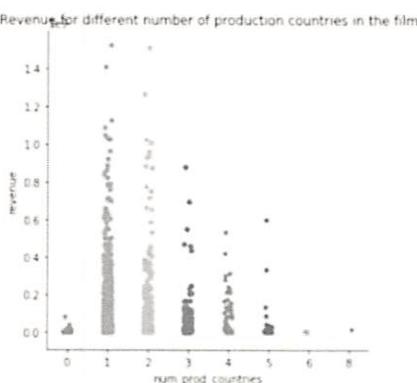
```
#Dropping genres
data.drop(['genres'], axis=1, inplace = True)
test.drop(['genres'], axis=1, inplace = True)
```

2) 제작사 나라의 다양성에 따른 수익률

여러 나라의 제작사에 따라 수익률에 변화가 있는지 확인해봅시다.

```
#production_countries
#Adding production_countries count for data
prod_coun_count = []
for i in data['production_countries']:
    if(not(pd.isnull(i))):
        prod_coun_count.append(len(eval(i)))
    else:
        prod_coun_count.append(0)
data['num_prod_countries'] = prod_coun_count

#Number of prod countries vs revenue
sns.catplot(x='num_prod_countries', y='revenue', data=data);
plt.title('Revenue for different number of production countries in the film');
```



제작사 나라에 대한 데이터도 제거해줍니다.

```
#Dropping production_countries
data.drop(['production_countries'], axis=1, inplace = True)
test.drop(['production_countries'], axis=1, inplace = True)
```

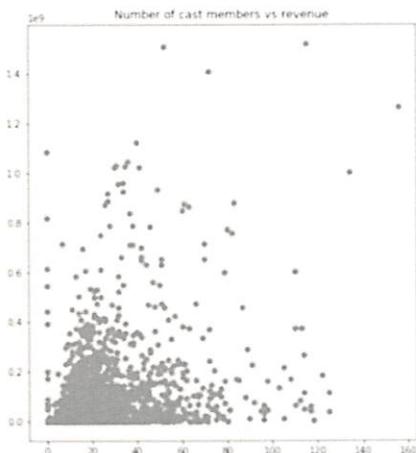
3) 출연진 수에 따른 수익률

```

total_cast=[]
for i in data['cast']:
    if(not(pd.isnull(i))):
        total_cast.append(len(eval(i)))
    else:
        total_cast.append(0)
data['cast_count']=total_cast

plt.figure(figsize=(16, 8))
plt.subplot(1, 2, 1)
plt.scatter(data['cast_count'], data['revenue'])
plt.title('Number of cast members vs revenue')

```



cast 타이틀 제거

```

#Dropping cast
data= data.drop(['cast'],axis=1)
test= test.drop(['cast'],axis=1)

```

4) 예산, 인기도, 런타임과 수익성 히트맵

```

#check correlation between variables
col = ['revenue', 'budget', 'popularity', 'runtime']

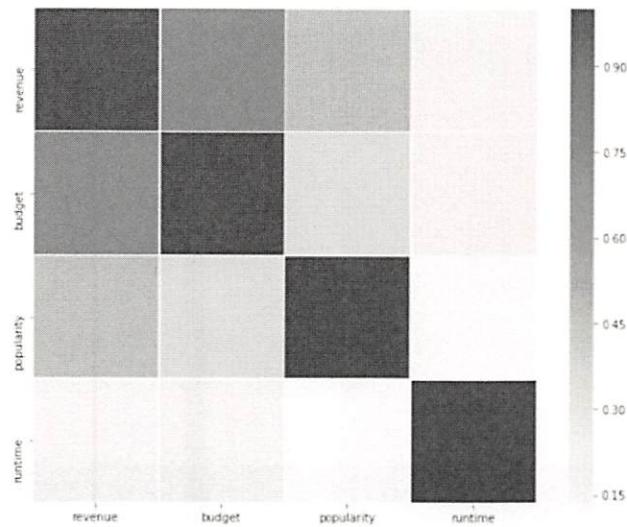
plt.subplots(figsize=(10, 8))

corr = data[col].corr()

sns.heatmap(corr, xticklabels=col, yticklabels=col, linewidths=.5, cmap="Reds")

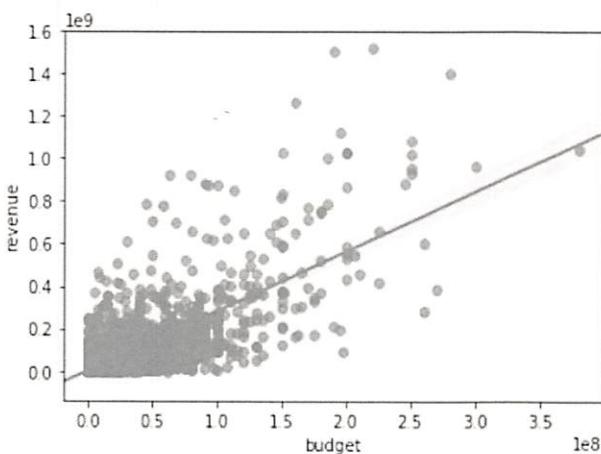
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f7aa6a82298>
```



예산과 수익성이 크게 비례하고 있다는 것을 히트맵을 통해 확인할 수 있듯이 그래프를 그려 그 관계를 알아본다.

```
#budget and revenue are highly correlated
sns.regplot(x="budget", y="revenue", data=data)
```



4. Modeling

- 1) y값을 예측값 revenue로 정해서 진행해봅시다. cols 변수에는 revenue와 id를 포함하지 않는 열들로만 구성되어 있습니다. 이때 cols에 포함된 열들의 값들만 X로 지정해줍니다. np.log1p(y)는 $\log(1 + y)$ 를 의미합니다.

```
y= data['revenue'].values
cols = [col for col in data.columns if col not in ['revenue', 'id']]
X= data[cols].values
y = np.log1p(y)
```

2) Linear Regression

```
from sklearn.linear_model import LinearRegression
clf = LinearRegression()
scores = cross_val_score(clf, X, y, scoring="neg_mean_squared_error", cv=10)
rmse_scores = np.sqrt(-scores)
print(rmse_scores.mean())
```

2.4832232893952897

3) Random Forest Regression

```
from sklearn.ensemble import RandomForestRegressor
regr = RandomForestRegressor(max_depth=10, min_samples_split=5, random_state=0,n_estimators=100)
scores = cross_val_score(regr, X, y, scoring="neg_mean_squared_error", cv=10)
rmse_scores = np.sqrt(-scores)
print(rmse_scores.mean())
```

2.2267428632215664

› 둘 중 Linear Regression의 예측 점수가 높기 때문에 Testing에서는 선형 회귀법을 사용할 것입니다.

4. Testing

```
cols = [col for col in test.columns if col not in ['id']]
X_test= test[cols].values

regr.fit(X,y)
y_pred = regr.predict(X_test)

y_pred=np.expml(y_pred)
pd.DataFrame({'id': test.id, 'revenue': y_pred}).to_csv('submission_RF.csv', index=False)
```

III. 탐구결과 및 고찰

1. <탐구 1>의 결과

가. 드라마, 코미디, 액션 순으로 인기가 많고 한 영화에 장르가 3개인 경우 수익률이 가장 높다.
나.

- 1) 하나 이상의 나라에서 제작사가 협찬해줄 때 수익률이 증가했으며 제작사의 수가 늘어날 수록 수익률은 감소하는 것을 알 수 있다.
 - 2) 협찬해준 제작사의 수가 0개, 6개 이상 일때 수익률이 적다.
- 다. 대체로 출연진 수가 적을수록 수익률도 적게 나타났다.
라.
- 1) 수익과 예산의 색이 가장 강렬한 것으로 보아 수익성에는 예산과 크게 비례함을 알 수 있다.
 - 2) 인기와 영화 상영 시간과는 거의 상관관계가 없다는 것도 알 수 있다.
- 마. 표어의 유무만으로도 수익성에 영향을 미칠 수 있다.

IV. 결 론

이번 활동에서 머신러닝을 이용하여 영화 개봉 전 수익을 예측하는 기법을 제시하였고 여러 가지 변수와 수익과의 상관관계를 통해 수익을 높이기 위해 어떻게 해야 할지 확인하였다.

머신러닝을 활용하는 이번 활동에서 정확도를 높이기 위해 학습을 많이 시키는 것이 좋고 개봉 전 영화의 수익을 정확히 판단하는 것은 제한적이다. 하지만 좀 더 다양한 데이터로 각 분석 기법에 맞는 변수 들을 활용하면 보다 높은 정확도와 높은 예측모델을 구현할 수 있을 것으로 기대된다.

V. 참가소감

이번 기회를 통해 머신러닝을 직접 체험해 보며 배우게 되어 정말 유익하고 재미있었다. 처음에는 설명만 듣고 하려니 어렵고 마음대로 되지 않아 힘들기도 하였다. 그러나 직접 코드를 만져보고 결과물을 보고 예측의 정확도를 점점 높여가며 재미를 느낄 수 있었고 컴퓨터를 배울 때 이론만이 아닌 직접 체험해 보며 배워야 한다는 말의 의미를 알 수 있었고 의미 있는 경험을 한 것 같다.

학 생 탐 구 활 동 평 가

지도교수	컴퓨터공학과 변 영 철
탐구주제	AI 머신러닝으로 배우는 데이터 과학, 똑똑한 컴퓨팅
탐구활동 평 가	<ul style="list-style-type: none"> ◦ 각각 3명으로 구성된 5개의 팀이 참가하여 각자 하고자 하는 내용에 대하여 탐구하고 발표함. ◦ 데이터 과학관련 기본 지식 습득을 위하여 인공지능의 기본, 파이썬을 이용한 데이터 사이언스, 데이터 엔지니어링을 위한 튜토리얼으로 진행함. ◦ 고등학생이어서 프로그래밍에 대한 지식이 충분치 않음에도 불구하고 의외로 각자 재미있는 주제를 선정하고 4일간 열심히 탐구 수행 ◦ 특히 탐구관련하여 수집하는 데이터가 상당한 고민에서 나온 것이어서 높이 평가함.
보고서 평 가	<ul style="list-style-type: none"> ◦ 고등학교 탐구보고서를 많이 작성해 보았는지 서론과 본론, 결론에 대한 일목요연한 정리가 눈에 띤다. ◦ 특히 사용하고자 하는 데이터를 모으는 과정과 이를 이용하여 전처리하는 과정, 그리고 최종적으로 이를 머신러닝 알고리즘을 이용하여 예측과 분류를 수행하는 방법이 체계적임. ◦ 보고서 작성에 있어서 3명의 팀원 간 역할 분담이 잘 이루어지고 있음을 긍정적으로 평가함. ◦ 영화평점예측, 농구선수 포지션 추천, 제주관광객 수 예측, 경제성장 요인 분석, 대륙별 종교분포 및 미래예측 등 재미있는 주제에 대한 보고서가 제출됨.
발표평가	<ul style="list-style-type: none"> ◦ 팀별 3인이 모두 나와서 발표 전체 내용에 대해 각자 맡은 부분을 역할 분담하여 발표를 수행함. ◦ 경제성장 요인분석과 관련하여 상당히 많은 내용에 대하여 데이터를 찾고 이를 기반으로 발표가 이루어짐. ◦ 영화평점의 경우 실제 데이터에 기반하여 예측이 이루어짐. ◦ 종교예측은 캐글에 있는 실제 데이터를 활용하여 탐구가 이루어짐. ◦ 제주 관광객 수 예측은 월별 방문자 수 관광공사 데이터와 해당 월의 제주관련 키워드 검색 정보를 조합하여 머신러닝을 수행하였음. ◦ 전체적으로 실제 데이터를 활용하여 데이터 분석도 하고 실제 머신러닝 알고리즘도 적용하는 등 우수한 결과를 냈다.