|  |  |  |  |
| --- | --- | --- | --- |
| **2021 혁신성장 빅데이터 분석 프로젝트 요약서** | | | |
| **팀 명** | A조 | **팀 장** | 김승건 |
| **팀 원** | 김승건 오지현 정제원 | | |
| **주 제 명** | 빅데이터 프로젝트\_요약서-대전A반\_A조\_파이썬활용\_날씨데이터활용(2010~2020) | | |

|  |  |  |
| --- | --- | --- |
| 구분 | 프로젝트 상세내용 | |
| 프로젝트명 | 빅데이터 프로젝트\_요약서-대전A반\_A조\_파이썬활용\_날씨예측모델(2010~2020)  각 해 6월5일부터 8월5일까지 | |
| 구현 목표 | 기상청에서 제공한 2010~2020년 양력 기준 6/5 ~ 8/5 전국 기후 데이터를 이용하여 머신러닝을 통해 기후예측 모델을 제작하고, 실제 데이터를 활용하여 예측, 분석(정확도)을 하여본다.  기상청 - 관측되는 기상 정보와 예전의 데이터를 바탕으로 일기현상을 예측함.  머신러닝 - 훈련 데이터(Training Data)를 통해 학습된 알려진 속성을 기반으로 예측에 초점을 두고 있다 *(wikipedia)*  그렇다면, 충분한 관련된 데이터가 있다면  기상청에서 제공하는 데이터를 활용하여 모델링을 통해 직접 예측해볼 수 있지 않을까?  분석 순서   1. API 불러오기 2. CSV파일 생성 3. 데이터 전 처리 4. Lasso / Ridge를 통한 Feature Auto selection 5. Scoring 후 분석 6. 임의로 설정한 Feature를 통해 SVM RandomForest를 적용 7. Scoring 후 분석 8. 기상청 Feature Selection Set 적용 9. Scoring 후 분석 10. RandomForest, 최적값(depth) 산출 11. 교차검증(CrossValidation) 12. 최적값 모델로 예측 13. 혼돈행렬을 이용한 평가 | |
| 구현 내용 | 1. **Feature Selection 후 각 모델 Score** | 2**. Rasso/Ridge 자동 Feature 선택** |
|  |  |
| 3. **SVM/RandomForest Score** | 4. **RandomForest 최적값 찾기** |
|  |  |
|  | 5. 교차 검증 |  |

|  |  |  |
| --- | --- | --- |
| 각 팀원 역할 |  |  |
|  |  |
| 김승건  - 데이터 취득, 정제 및 코딩  - 방향성 확인 테스크, PPT, 및 보고서 생성  - 데이터 백업 및 확인  - 발표  오지현  - 데이터 취득 및 정제 및 코딩  - 코딩 Main  - 데이터 백업  - 발표  정제원  - 데이터 취득, 정제 및 코딩  - 방향성 확인, 테스크, PPT, 및 보고서 생성  - 데이터 백업 | |

|  |  |
| --- | --- |
| **상세설명 세부항목** | |
| **1. 주제 선정/개요**  1-1) 선정 배경  1-2) 예상점  기간 내 기후 데이터와 일기예보, 실제 날씨와의 상관관계를 알아본다.  1-3) 데이터 선정이유  Open API를 활용하여 실시간 데이터 및 일별 데이터를 확보 할 수 있음  실시간데이터를 활용하기에는 시간 및 리소스가 부족함  일별 데이터를 활용하여 데이터를 CSV 파일로 정제하여 데이터 안정성을 높여서 모델링에 초점을 둔다.  2010~2021 각 해의 06.05~08.05까지  여름 장마철을 중점.  실제 예보에는 흐림, 안개 와 같은 강수량과 상관이 없는 예보가 있음 이를 전부 라벨(y)로 분류할 수 없기때문에 맑음 / 비 두 가지로 분류함  **2. 분석 및 기능 설계/개발**  2-1) 데이터 수집 및 전처리  기상청 데이터 수집  Feature 65개 예) 기온 습도 풍속 풍향 등  2-2) 모델링을 통한 ML  SVM LASSO RIDGE RANDOM\_FOREST  2-3) 각 모델링의 결과값 및 정확도 Scoring  2-4) 모델링 결과값 시각화 | **3. 연구방향**  실제 기상 상황을 카테고리화 하는 것에 대한 모호성.  예) 강수량은 없으나 습도가 높을 경우의 결로, 안개 등을 어떻게 카테고리화 할 것 인가.  습도와 기온 등 적은 데이터로 일기예보를 하는 것에 대한 어려움을 어떻게 해결 할 것 인가.  **4. 결론**   1. 기상청의 예보는 “맑음”, “구름 많음”, “흐림” 등 소분류가 존재하기에 더 많은 변수에 영향을 받음. 2. 기상예보는 데이터를 기반으로 하기에, 센서 유무, 데이터 전송 등의 물리적, 기술적 한계가 존재함. 3. 제작한 모델링은 맑음/비옴 두 가지로만 나누어 위와 같은 복잡성을 배제함. 4. 제작한 모델로도 90% 이상의 정확도를 보여줌. (여름 한정) 5. 다른 계절의 경우 Feature(기온, 습도 등)가 매우 다른 양상을 가지기 때문에 학습 데이터양(범위)를 늘려야함. 6. 실제로 기상이 예보에 실패한다기 보다는, 자연현상의 특성상 급격하게 변할 수 있다. (한국 기상청 정확도 85%)   **5. 프로젝트 시사점**   1. 시간별 데이터까지 활용해서 예측 모델을 제작하였다면, 더 높은 정확도를 가질 수 있었을 것이다. 2. 1월부터 12월까지의 모든 데이터를 활용하였다면 다른 계절의 날씨 예측 해 볼 수 있었을 것이다. 3. Feature에 대한 이해도 부족. 4. 실제로는 기상청의 예보가 실사용자가 체감하는 것과는 달리 매우 높은 정확도를 보여준다. |