PORTFOLIO 개요서

**목 차**

1. 연구보고서
2. 데이터 시각화
3. 머신러닝
4. 시계열 분석
5. 데이터분석프로젝트
6. 4회 데이터해커톤 분석보고서
7. 문화관광부 공모전
8. 환경융합공모전

1. 연구보고서

1) 데이터시각화

|  |  |
| --- | --- |
| 프로젝트 명 | 데이터시각화 연구보고서 |
| 프로젝트 목적 | 시각화 패키지별 연구 및 비교를 통한 활용법 학습 |
| 참여인원/역할 | 1명 |
| 기간 | 21.04.16 ~ 21.04.24 |
| 활용언어/TOOL | R / R\_studio |
| 주요 데이터 | 패키지 : ggplot2, plotly, esquisse, lattice, RGL, ggmap |
| 주요 내용 | 1. R에서 사용되는 주요 시각화 패키지를 활용 2. 각 패키지별 기능 비교와 성능 확인 3. 각 그래프가 가지는 특성들과 적합한 데이터 분별 |

2) 머신러닝 모델링

|  |  |
| --- | --- |
| 프로젝트 명 | 머신러닝 모델링 연구보고서 |
| 프로젝트 목적 | 머신러닝 모델 평가 및 성능 비교 |
| 참여인원/역할 | 3명 / PM 모델 선정 및 보고서 작성 |
| 기간 | 21.04.01 ~ 21.04.15 |
| 활용언어/TOOL | R / R\_studio |
| 주요 데이터 | Wisconsin Diagnostic Breast Cancer  Boston Housing Price |
| 주요 내용 | 1. 위스콘신 유방암 데이터를 활용하여 머신러닝 모델 평가 및 성능 비교 (Classification)  * Naïve Bayes, Logistic, Decision Tree, Neural network, SVM, Random Forest, xgBoost  1. 보스턴 집값 데이터를 활용하여 모델 평가 및 성능 비교 (Regression)  * Decision Tree, Neural network, Random Forest  1. 클러스터링 모델들의 특징 분석  * K-means, K-medoids, Gausian MM, DBSCAN 등 |

3) 시계열 분석

|  |  |
| --- | --- |
| 프로젝트 명 | 시계열 분석 연구보고서 |
| 프로젝트 목적 | 시계열 분석 이해 및 예측 활용 |
| 참여인원/역할 | 1명 |
| 기간 | 21.03 |
| 활용언어/TOOL | R / R\_studio |
| 주요 데이터 | World Bank - population  Kospi 지수 (10년) |
| 주요 내용 | 1. 1960 ~ 2019 년 대한민국 인구 데이터를 시계열로 분석하여 향후 인구 추세를 예측 2. 과거 10년간 일별 KOSPI 지수(종가 기준)를 시계열 분석 후 향후 지수 예측 |

2. 데이터분석프로젝트

1) 데이터해커톤 분석보고서

|  |  |
| --- | --- |
| 프로젝트 명 | 우리는 모두 한때 아이였다. |
| 프로젝트 목적 | 공공 빅데이터 분석 프로젝트 |
| 참여인원/역할 | 4명 / PM (기획, 보고서작성, 발표) |
| 기간 | 21.08 |
| 활용언어/TOOL | R, Python / R\_studio, Jupyter\_lab, QGIS |
| 주요 데이터 | 지방자치단체 기본통계 – 인구주택총조사  E-지방지표 – 통계청  사회보장통계 – 한국사회보장정보원  아동학대관련 데이터 – 아동권리보장원 |
| 주요 내용 | 1. 아동학대에 영향을 미칠 것으로 추정되는 외부요인들을 가정 2. 정리된 외부요인들을 대표할 수 있는 데이터들 수집 3. 수집된 외부요인들을 조사 및 점검 과정을 통해 독립변수로 선정 4. 아동학대 신고접수 건을 바탕으로 아동학대 발생률 산출 5. 산출된 발생률을 종속 변수로 설정하여 분석진행  * 회귀분석, 머신러닝 분류모델  1. 분석 결과를 도출한 뒤 유의미한 변수 도출과 지역적 특성 파악 2. 도출된 유의미한 변수들을 활용하여 아동학대 위험 지역 예측 선별 |

2) 문화관광부 공모전

|  |  |
| --- | --- |
| 프로젝트 명 | 문화 관광 업종의 특징에 따른 코로나-19 영향 분석 |
| 프로젝트 목적 | 문화관광부 데이터 분석 공모전 참여 |
| 참여인원/역할 | 1명 |
| 기간 | 21.06 |
| 활용언어/TOOL | R, Python / R\_studio, Jupyter\_lab |
| 주요 데이터 | 내국인 카드 지출 데이터 – 신한카드  설문조사 – 업종별 특징에 관련된 설문  SNS 토픽 크롤링 – 인스타그램 웹크롤링  코로나 확진자 데이터 – 질병관리본부 |
| 주요 내용 | 1. 설문 데이터를 활용하여 업종별 특징을 특정화 2. 카드 지출 내역과 SNS 웹크롤링 데이터를 활용, 결합하여 변수설정 3. 월별 코로나 심각도와 카드 사용건 수 증감률을 계산하여 변수 정리 4. 회귀와 랜덤포레스트를 활용하여 카드 사용 건 수에 영향을 주는 변수 도출 및 분석 5. LSTM(시계열)을 활용하여 확진자 수에 따른 업종 전망 예측 |

3) 환경융합 공모전

|  |  |
| --- | --- |
| 프로젝트 명 | 도로 재비산먼지와 버스정류장 데이터를 활용한 쉘터형 버스정류장 입지선정 |
| 프로젝트 목적 | 환경부 데이터분석 공모전 참여 |
| 참여인원/역할 | 1명 |
| 기간 | 21.09 |
| 활용언어/TOOL | R / R\_studio, QGIS |
| 주요 데이터 | 도로 재비산먼지 측정정보 (2017 ~ 2019) – 한국환경공단  서울시 버스정류장별 승하차 인원 – 서울 열린데이터 광장  서울시 버스정류장 위치 데이터 – 서울 열린데이터 광장  서울시 도로(도로명) SHP 데이터 - 국토교통부 |
| 주요 내용 | 1. 서울시 버스정류장별 승하차 데이터를 정제한 후 EDA와 통계적 분석을 통해 선정된 기준에 맞는 버스정류장 선별 2. 버스정류장 위경도 위치 데이터를 위에서 선별된 버스정류장 항목에 맞게 정비 3. 서울시 도로 재비산먼지 데이터를 1년 평균으로 정제 4. 위 데이터를 활용해 서울시 도로 재비산먼지 위험 지도를 만들고 그 해당 지역내 버스 정류장 탐색 5. 최종 선정된 버스정류장을 활용하여 정책 제언 |