**세미프로젝트 기획안**

**과목명: 실전 빅데이터 분석 프로젝트**

**기획안 작성일자 : 2022-05-06**

|  |  |
| --- | --- |
| **팀 명** | 3조 : 파이썬 어렵조 |
| **프로젝트 조직**  (구성원 및 역할) | 팀장 : 박영우 : 데이터 수집 및 전처리, 모델링, 발표 팀원 : 김연우 : 데이터 수집 및 전처리, 모델링, 발표  팀원 : 문석찬 : 데이터 수집 및 전처리, 모델링, 발표자료 제작  팀원 : 김나은 : 데이터 수집 및 전처리, 모델링, 발표자료 제작  팀원 : 오효석 : 데이터 수집 및 전처리, 모델링 |
| **프로젝트 주제 및 개요**  해결하고자 하는 문제  최종 산출물의 청사진 | 주제 : 아파트 실거래가 예측 모델 만들기  청사진 : 기존에 아파트 실거래가를 설명하는 기준금리 등의 거시경제 데이터, 공원 위치 등의 변수에 지하철역 위치와 도심과의 거리 변수를 추가하여 아파트 실거래가를 예측하는 새로운 모델 만들기 |
| **프로젝트 수행 방향**  수행 방법/도구 | **1. 데이터 수집 및 전처리**  1-1. 소상공인 데이터와 K-Means Clustering을 이용해 스타벅스 가맹점의 위치와 도심의 위치 도출  1-2. 지하철역의 위도 및 경도 데이터 받아오기  1-3. 국토교통부 아파트 실거래가 데이터를 이용해 각 아파트의 기본적인 feature 및 실거래가, 위경도 정보 등 도출  1-4. 한국은행의 기준금리 데이터, 정부 부동산 정책 관련 변수 등 거시경제 관련 데이터 받아오기  1-5. 각 구별 범죄율, 서울시 공원 데이터 등 집값에 연관이 있을 것으로 추정되는 기타 변수의 데이터를 받아오기  1-6. 해당 아파트와 가장 가까운 도심 사이의 하버사인(Haversine) 거리 도출  1-7. 해당 아파트와 가장 가까운 지하철역 사이의 하버사인(Haversine) 거리 도출  **2. 모델 구축 및 평가**  2-1. 아파트 실거래가 데이터를 Test&Train 데이터로 분할  2-2. Train 데이터와 여러 머신러닝 모델들을 이용해서 가장 성능이 좋은 ML 모델을 찾기  2-3. 구축한 모델 및 Test 데이터를 이용하여 실거래가 예측 및 평가 |
| **프로젝트 추진 일정** | 05/03~05/05 : 데이터 수집  05/06~05/14 : 데이터 전처리  05/15~05/17 : 모델 구축 및 기능 확인  05/18~05/19 : 발표준비 및 발표 |