

# 딥러닝 기반 핵심 산업별 빅데이터 분석

## <머신러닝&딥러닝 파일럿 프로젝트>

주 제	Costa Rican Household Poverty Level Prediction	링 크	<a href="https://www.kaggle.com/c/costa-rican-household-poverty-prediction">https://www.kaggle.com/c/costa-rican-household-poverty-prediction</a>
팀 명	원픽	일 자	2018년 11월 29일
팀 장	문정연<jymoon1115@gmail.com>	팀 원	공정배, 이진수

### 1. 과제 개요

적절한 복지 예산 집행을 위해서는 소득 자격 확인 알고리즘이 매우 중요하다. 라틴 아메리카에서 주로 사용되는 PMT(Proxy Means Testing) 방법은 주택의 벽의 재질, 소유하고 있는 가전제품 종류 등의 가시적인 평가지표를 활용해 모델을 만들어 적용한다. 그러나 여전히 모델의 정확성을 높일 필요가 있다. 이에 따라 기존 평가 지표 외의 다양한 코스타리카의 가계 특성에 대한 데이터 세트를 기반으로 보다 적합한 모델을 찾아내고자 한다.

텐서플로우를 활용하여 TRAINING 데이터를 학습하여 TEST 데이터의 경제 수준을 1-4(1 =극빈 빈곤 / 2 = 중등빈곤 / 3 = 취약 가구/ 4 = 비취약 가구)의 4가지 단계로 구분하여 나타내고자 한다. 이전에 배운 텐서플로우의 다중분류기법을 적절히 활용할 수 있는 예제임은 물론 사회적으로 꾸준히 논의가 이루어지고 있는 데이터를 분석한다는 것에서 의의가 있다.

### 2. 데이터 설명

Idhogar은 세대 식별 고유 값으로 같은 행끼리는 같은 가구원임을 파악할 수 있다. 같은 세대임에도 불구하고 TEST 최종 값이 다르게 나타나는 경우가 있으며 이러한 오류를 줄이기 위해서 세대주만을 뽑아 분석을 진행하고자 한다.

크게 유의미할 것으로 보이지 않는 데이터 열은 피쳐로 선택하지 않으며, 동일한 내용을 담고 있는 열은 각각의 피쳐값을 척도로 파악해 통일하여 하나의 피쳐로 만들어낼 예정이다. 각 피쳐 간의 격차를 줄이기 위해 normalize를 시행할 예정이다.

### 3. 과제 수행 내역

### 4. 결과 보고

감사합니다