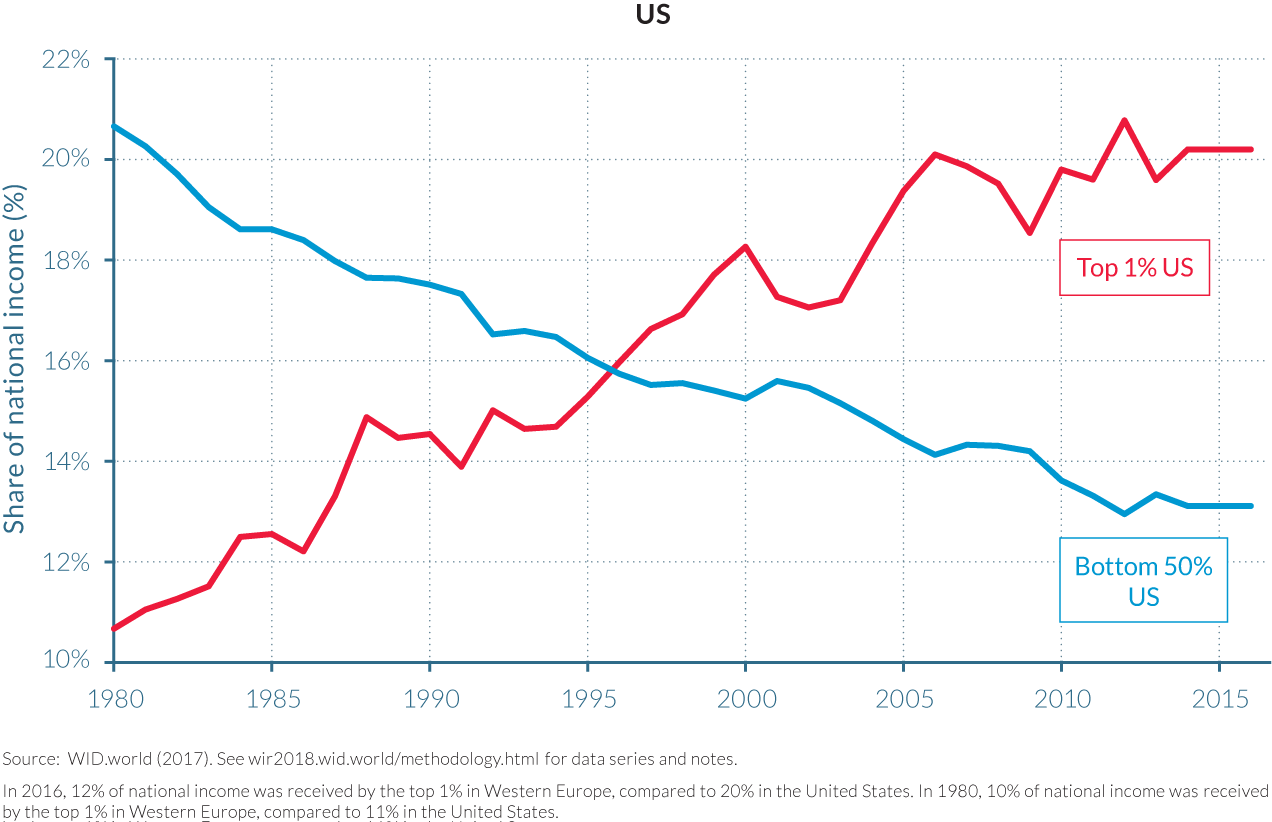
**Term Project Proposal (Group #06)**

김세준, 박창훈, 최현호, 황지수

1. 문제 설정

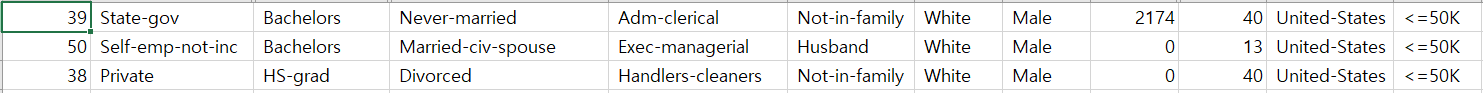


21세기로 접어들면서 세계 각국은 극심한 경제적 불평등으로 몸살을 앓고 있습니다. 중산층이 몰락하고 사회계층의 양극화가 일어나는 오늘날, 불평등을 낳는 원인에 대한 연구의 필요성은 나날이 증대되고 있습니다. 이를 위해 저희 팀에서는 사람들의 개인소득과 관련된 사회조사자료를 활용해, 사람들이 경제적 상류층이 되는데 어떠한 요소가 핵심적인 영향을 미치는지, 특정한 조건의 사람이 상류층에 속할 확률은 어느정도인지를 알아보고자 합니다.

1. 활용할 데이터셋

모델을 학습시키는데 사용할 데이터셋은 1994 US Adult Census Data에서 추출한 미국인 48842명의 연봉 데이터입니다. *(출처:* [*https://www.kaggle.com/johnolafenwa/us-census-data#adult-training.csv*](https://www.kaggle.com/johnolafenwa/us-census-data#adult-training.csv)*)* 각 row의 feature는 연령, 성별, 인종, 교육 수준, 직장 유형, 결혼 상태 등이며, 연봉 5만달러를 기준선으로 label되어있습니다. OECD기준 중산층의 정의가 중위소득의 50~150%의 수입을 올리는 자라는 것을 감안하면, 연봉 5만 달러는 1994년에 상류층을 판가름하는 기준으로 적절하다 볼 수 있습니다. 저희 팀의 목표는 데이터에서 주어진 Label을 바탕으로 supervised learning을 시행해, 특정 정보가 주어졌을 때 연봉 5만달러 이상일지를 classify 하는 모델을 만들고, 어떠한 요소가 얼마나 큰 영향을 미치는지 알아보는 것입니다. 이를 위해 decision tree, random forest 등의 여러 기법을 적용한 모델을 각각 만들어 학습시키고, 가장 정확한 결과를 보여주는 모델을 선택하려 합니다.

<샘플 데이터>



왼쪽부터 나이, workclass, 최종 학위, 결혼 여부, 직업 특성, 가족 구성, 인종, 성별, 수입, 주 당 근로시간, 연봉

1. 예상되는 결과
2. 미국에서 시행된 사회조사 자료에 따르면 교육수준에 따라 평균 수입이 크게 달라진다고 합니다. 이에 따라 데이터셋에 존재하는 여러가지 feature 중 교육수준이 상류층이 되는데 핵심적인 영향을 끼칠 거라 예상합니다.
3. 능력주의를 표방하는 미국사회의 특성상 시간당 임금의 차이도 크기에, 절대적인 노동시간과 상류층 여부에는 유의미한 상관관계가 없을 것으로 예상합니다.
4. 다만 자료 특성상 feature와 instance의 갯수가 상당히 많아, 다른 요소에 비해 어떤 요소가 더 결정적인 영향을 미칠지 직관적으로 가설을 세우기 어렵습니다. 때문에 모델을 학습시킨 결과를 분석하며 feature의 상대적 중요도를 알아보려 합니다.
5. 팀원별 역할

우선 첫 발표를 앞두고 팀원 모두가 모여 연구주제와 방법을 의논하였습니다. 프레젠테이션과 proposal을 위한 자료 수집은 김세준, proposal 문서 작성은 박창훈, 프레젠테이션 파일 작성은 황지수, 발표자는 최현호 학우가 분담하였습니다. 앞으로 실제로 모델을 만들고 학습시키기 위해 요구되는 머신러닝에 대한 지식, 라이브러리와 툴의 사용법 등은 팀원 모두가 제대로 숙지하고 있어야 하기에 분담하기보다는 팀원 모두가 지식을 공유하며 함께 갖추어나가려 합니다. 최종 발표와 보고서 작성에 대해서는 차후에 역할을 배분할 계획입니다.