# **미국의 의료비용 데이터를 이용한 비용 예측**

**요 약**

미국인들의 나이, 비만도, 흡연유무 등 다양한 데이터를 이용하여 의료비용을 예측할 수 있는 모델을 만든다.

예측한 모델은 선형회귀분석 평균제곱오차법을 사용했다. 이를 통해 한국인들의 데이터를 수집한다면, 우리나라에 알맞는 기준을 통해 의료비용을 예측하고 대비할 수 있다.

# **서론**

개인마다의 여러 조건(ex. 나이, 성별, bmi 등)에 따라 의료 비용이 달라지게 된다. 의료 보험 설계에서도 데이터를 통해 적용할 수 있게되며, 개개인의 조건에 맞춘 예측을 통해 비용적 측면에서 미리 인지하고 예방을 할수 있을 것이다. 미국의 데이터통계 자료를 통해 진행하였지만, 우리나라의 데이터 통계자료가 확보되면 우리나라에서도 적용할 수 있다. 미리 어느정도의 예측이 가능하다면 개인의 갑작스러운 경제적 부담도 줄일 수 있을 것이다.

# **관련연구**

우리 나라는 건강보험제도가 시행되고 있음에도 불구하고 이용자의 본인부담비율이 전체 진료비의 50% 를 상회할 정도로 매우 높은 수준이다. 이런 높은 본인부담 수준은 건강보험의 위험 분산 및 소득보호 기능을 저하시키며 질병 발생 시 진료비의 발생 수준에 따라 개인의 경제생활에 지대한 영향을 미칠 수 있다. 특히 높은 본인부담금은 개인의 경제적 능력에 따라, 체감하는 부담의 크기가 달라질 수 있다 예컨대 고소득층의 경우에 높은 본인부담은 다른 재화나 서비스에 대한 소비 감소로 인한 후생의 감소를 초래하지만 빈곤에 빠지게 하는 일은 드물다. 그러나 중산층이나 저소득층의 경우에 높은 본인부담은 소비 감소로 인한 후생의 감소외에도 빈곤에 빠지게 할 가능성이 높다. 건강보험의 보장성은 결국 예기치 못한 의료비 지출로부터 개인이나 가정을 얼마나 보호해줄 수 있느냐가 관건이다 . (참고문헌[1] 참조)

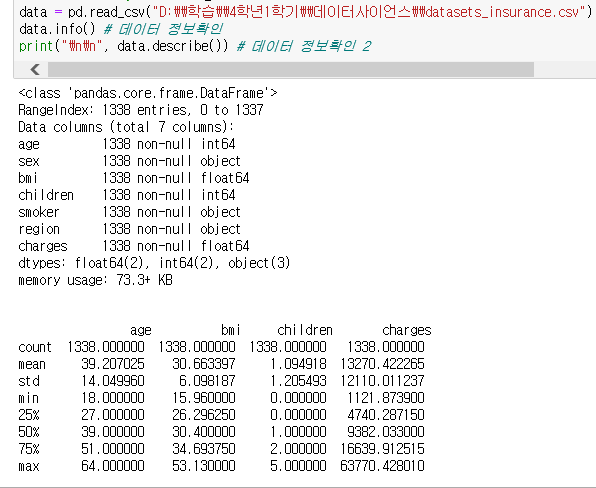
# **데이터**

데이터 타입으로는 나이(age: int), 성별(sex: string), 비만도(bmi: float), 부양가족 수(children: int), 흡연(smoker: Bool), 지역(region: string), 비용(charges: float)으로 구성되어 있다. bmi는 kg(weight)/m^2(height)을 통해 계산되었으며, 미국의 데이터를 인용하여 비용의 기준은 달러(\)이다. 부양가족 수(children)은 건강보험이 적용되는 어린이수/ 부양가족 수 를 나타내는 데이터이다.

# **방법**

확보된 데이터 확인 : data.info()

확보된 데이터 확인2: data.describe()

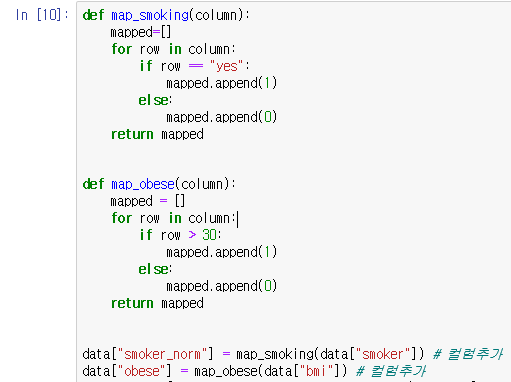


데이터 전처리

- smoker : yes(1), no(0) 으로 변환

- bmi : 30이상(1 = 비만) , 30이하(0 = 정상)

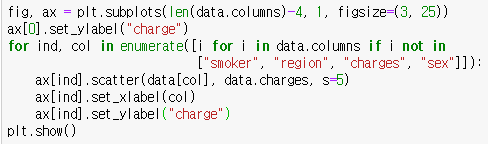
새로운 컬럼에 추가하였다.

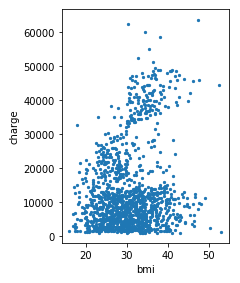
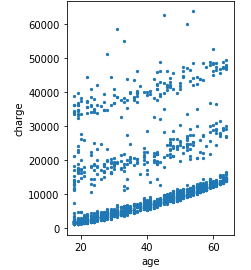


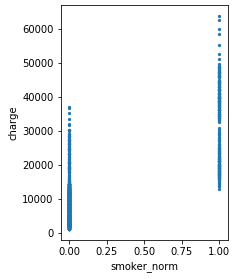
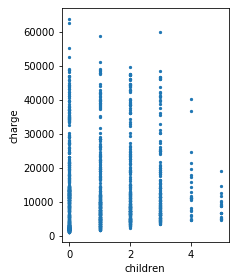
위 data.info()에서 확인된 정보들 중에서 dtypes 이 숫자가 아닌 object로 이루어진 데이터들의 컬럼들을 수집.

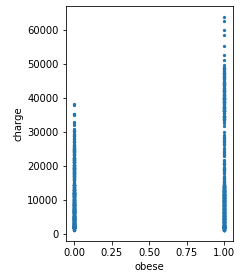


각 컬럼들과 비용을 scatter그래프로 시각화

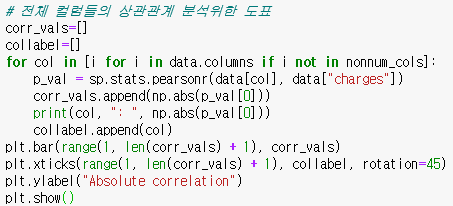




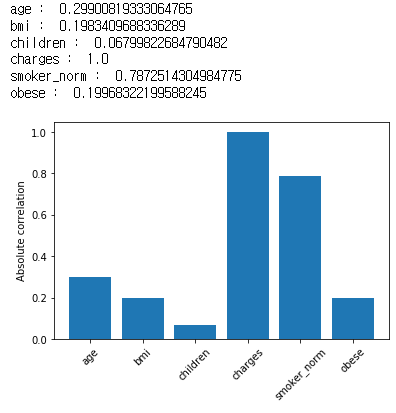




pearsonr 함수를 사용하여 비용과 각 컬럼들을 1:1로 매칭하여 상관계수를 도출.



그 결과를 시각화



이를 통하여 흡연률(smoker\_norm), 비만도(obese), 나이(age) 이렇게 세가지 컬럼이 비용과 상관관계가 가장 높게 나타났다.

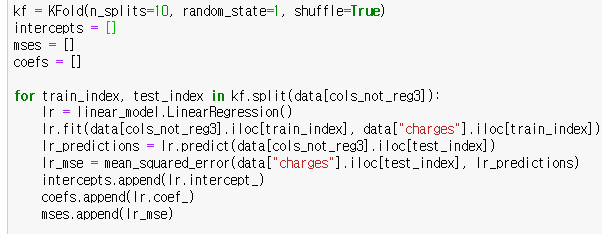
보다 편하게 처리하기 위해 앞서말한 컬럼들을 cols\_not\_reg3 리스트에 모아 주었다.



shuffle=True 로 주어 전체 데이터에서 학습 데이터와 테스트 데이터를 랜덤으로 나누어 주고, 학습과 동시에 테스트도 같이 실행되는 KFold 교차검증법을 10번 사용하였다. (참고문헌[2] 참조)

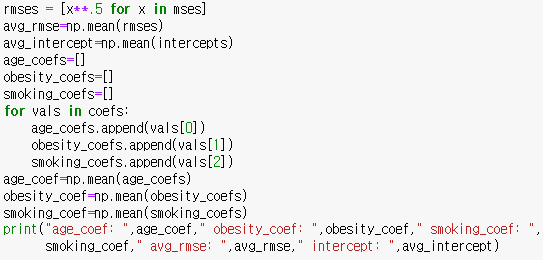
sklearn의 선형회귀분석을 사용하였고, 평균제곱오차법(mse:mean\_squared\_error)을 사용.

이를 통하여, lr.coef\_(학습을 통해 얻은 가중치), lr.intercept\_(학습을 통해 얻은 절편), lr.\_mse(학습을 통해 얻은 오차) 를 파악



lr.\_mse로 파악한 값들을 토대로 rmse를 구하고, 평균을 구함(성능 분석) - '5.실험'에서 사용.

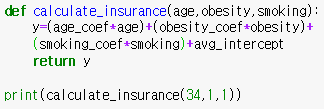
앞서 찾아낸 나이, 비만유무, 흡연유무에 대한 coefs(가중치)들을 분류해주고 평균을 도출.



최종적으로 각 컬럼에 대한 가중치와 입력된 데이터를 토대로 아래의 완성 된 Formular 를 통하여 의료 비용을 예측해줌.

predict = (해당 컬럼 가중치(coef) \* 입력된 데이터)

+ ... + 학습을 통해 얻은 절편(intercept)



# **실험**

예측 정확도 평가를 위해 rmse를 사용

(참고문헌[3] 참조)

KFold 의 n\_splits를 5~20 까지 조절하며 실험하여

rmse 값을 도출 해 냈고, 이를 통하여 n\_splits값이

18일때 rmse가 확연히 낮다는 것을 확인.

(예측에 사용 된 입력 데이터 : age=34, obesity=1, smoking=1)

KFold(교차검증방법) n\_splits 값을 18로 설정하여 값을 도출 해 냈다.



# **결론**

위 자료는 미국인들의 데이터를 토대로 만들어진 모델인데, 한국인들의 데이터를 수집하여 모델을 구성한다면 우리나라에 대한 의료비용을 예측할 수 있을 것 이다.

다양한 컬럼(예를들어 암, 치매 등 각종 질병 등) 들을 추가하여 새롭게 상관관계를 도출해내 다양한 질병에 대한 의료비용을 예측할 수 있음. 그로인해 금전적으로 혹은 심리적으로 대비를 할 수 있다.

추가적으로, 기준을 비용이 아닌 각종 질병에 대한 발생율과 같은 다른 기준으로 설정해서, 해당 기준에 대한 예측을 기대할 수 있다.

# **참고문헌**

1. 보건의료 비용 지출의 형평성, 이태진(한림대학교 보건대학원), 양봉민, 권순만, 오주환, 이수형(서울대학교 보건대학원), 보건경제연구 제9권 제2호, 25~34, 2003.
2. <https://velog.io/@skyepodium/K-Fold-%EA%B5%90%EC%B0%A8%EA%B2%80%EC%A6%9D>
3. <https://otexts.com/fppkr/accuracy.html>