Diabetic EDA

DATE 2023.07.01.Sat NO.

distplot → 연속형 값. 히스토그램.

distribution(분포) 확인은 중요해.

≫ 오직 numeric feature만 비교가능하다.
alphabet과 string은 비교X

피어슨 상관계수
: 각각의 feature의 분산들이 어떤식으로 되어있어서 어떤 상관관계가 있나?
여기서 중점의 모양을 보고 그것을 수치화해서 그들의 Correlation을 판단.

다중공선성 $y=X$ .. 똑같은 값을 가질때. 그러나 redundant해.
필요없다. 비슷한 정보를 가진것이 많으면 성능저하!!
e.g) 앙상블같은 부스팅 모델은 여러 트리를 합쳐서 만드는 모델.
이러한 모델의 가정은 다양한 (diverse)한 개별 tree를 만들어야
성능이 좋아진다. // EDA중. feature Engineering and data Cleaning.

모델 학습은 컴퓨터가 알아서 함. 사람이 해야할건 정제, EDA. idea,
중요한 문제! 데이터이해가 최우선.

모델이 학습하기 좋은형태로의 변환도 중요하다.

≫
나이같은 경우 연속형 값이다. 아는 문제가있다.
성별의 경우 남/여로 (Sex)로 나눌수있지만 Age로 나누면
좀.. 애매하다. 문제가있다. 나누기가.. 그래서 카테고리값으로..
바꿀 필요가 있다. range-hand로 풀자.
80까지 있으니까 5개 bins. 80-16 = 5.
bins_size = 16.

DATE . . NO.

연속적인 value와 연속적인 value를 비교하면 상관관계 유무를 좀 판단가능.
Survived같은 binary value는 범주형 변수에 어울림.

qcut은 continous_value를 나눠준다.
등분 sample 수로 맞춰준다. range안에서
cut으로 나누면 같은 간격으로 나눈다. (구간의 폭)
qcut은 같은 sample 수로.. qcut을쓰자. 깔끔수를 맞춰줌.

cut, qcut 둘다 확인하자.
» 레벨 인코딩을 활용해서 String → 숫자로.

data feature와 많은 factor의 관계도?