D (토론 1 - 6 조, 토론 2 - 5 조)

토론 1 - 6 조 5. 77 3 월 31 일

문제 정의 분석 보고서

说话 :

[데이터 설명]

- 반도체 제조 공정 데이터인 SECOM 데이터를 사용.

- 측정 시간, 공정 센서 데이터, 양품/불량품 여부에 관한 1567개의 object를 가짐.
  -> 데이터의 개수가 많지 않으므로 sample을 함부로 삭제하지 않아야 함.

- 591개의 feature을 가지는 다변량 데이터이고 측정 시간을 제외한 독립변수명 공개하지 않음.
  -> 분류에 악영향을 주거나 무의미한 feature을 제거 통해 차원을 감소시켜 overfitting 방지해야 함.

- Time 변수를 제외한 나머지 독립변수들은 수치형 변수이며, 종속변수는 양품(1) 및 불량품(-1)으로 구성.

- 이 데이터를 통해 가장 높은 분류 성능을 가지는 모델을 구축하고, 각 모델에서 제공하는 도구를 이용하여 중요한 신호를 파악.

3

데이터에 대한 설명이 상세합니다.

对数据的说明详细。

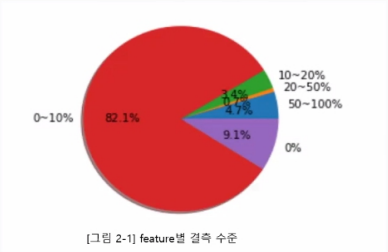
문제 정의 분석 보고서

1.결측치

说话 : 전창범 / Jeon, Chang B...

해당 데이터의 feature에서 많은 결측치가 있음을 발견.
결측이 없는 feature의 비율은 9%로 결측치가 있는 열을 모두 제거할 경우 데이터의 손실이 큼.
즉, 결측치 수준에 따른 결측치 처리 방법을 다르게 결정.

대체 : 결측치를 유의미한 관측치로 대체
삭제 : 결측치의 비율이 매우 높은 feature을 제거



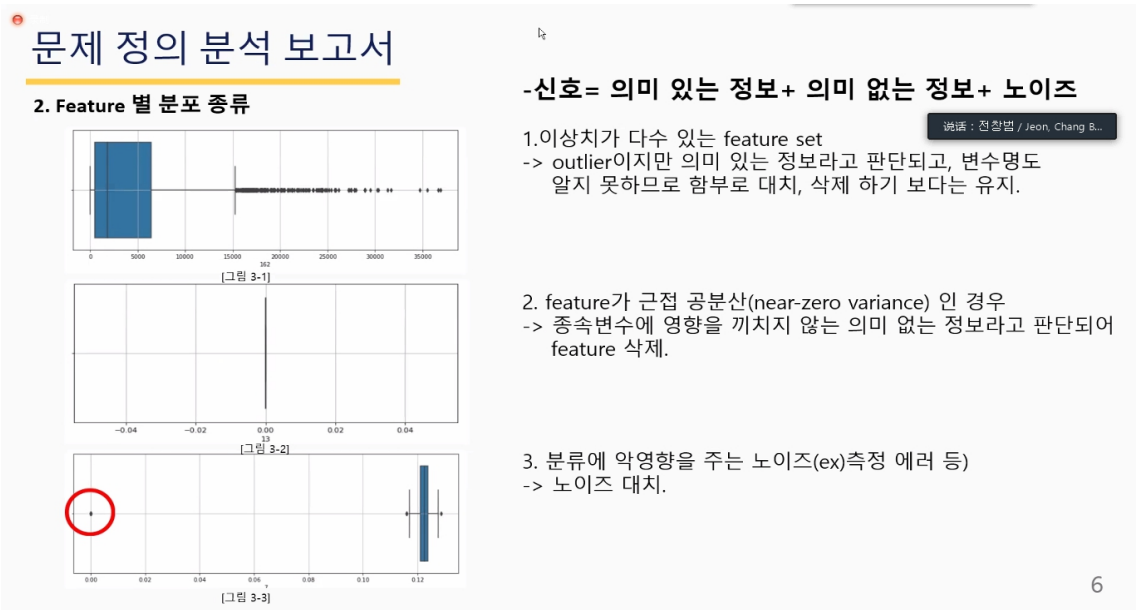
[그림 2-1] feature별 결측 수준

[그림 2-2] feature별 결측치 그래프

5

데이터 결손값 구분이 상세합니다.

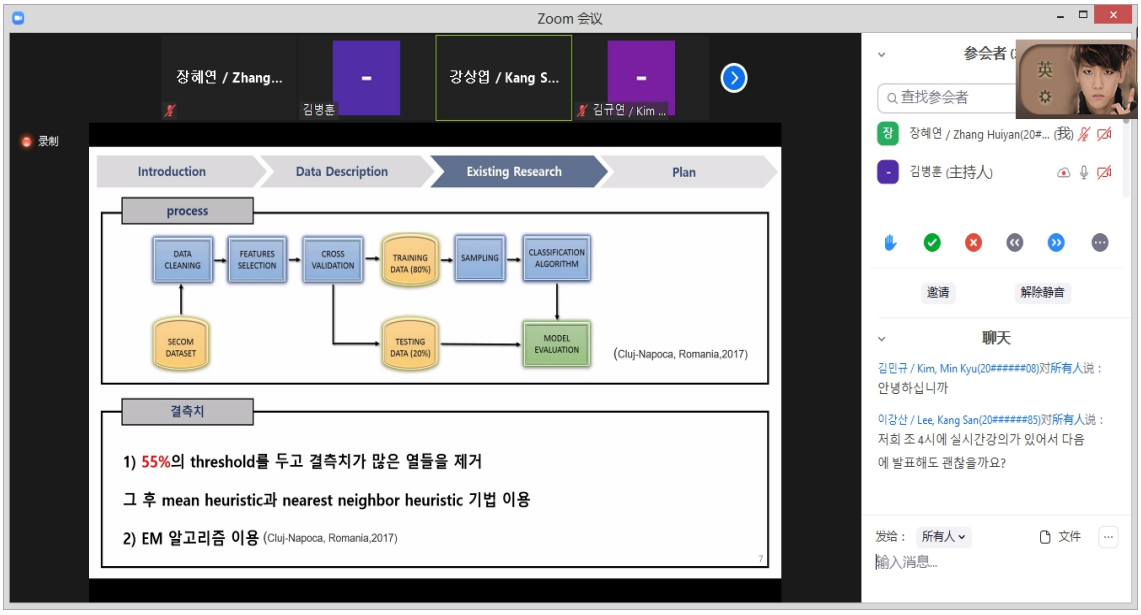
数据缺失值划分细。



구체적으로　삭제된　데이터를　예시합니다.

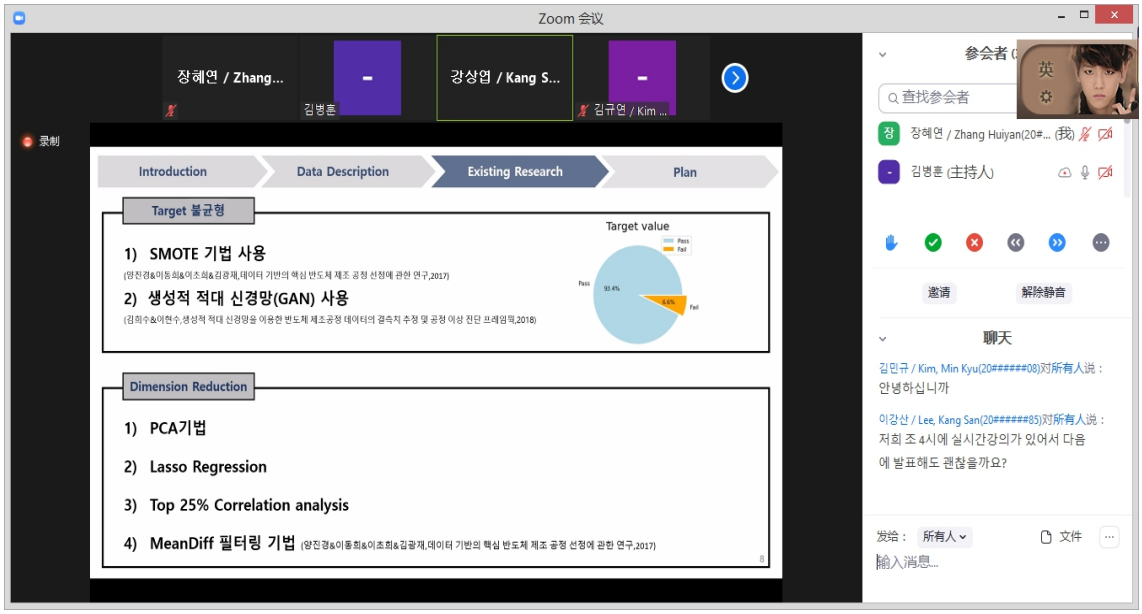
例举具体删除的数据。

토론 2 - 5 조 5. 47　5 월　4 일



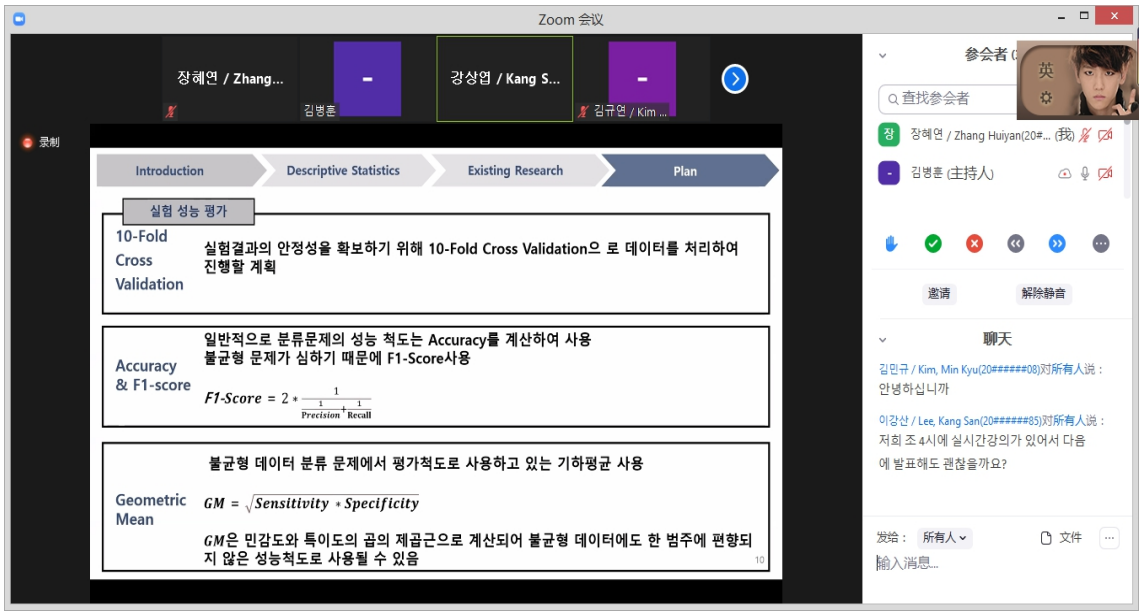
이용한　기술과　알고리즘을　썼습니다.

写出了利用的技术和算法。



절차가 상세하게 설명되어 있습니다.

步骤写的详细。



계획 중 실험성능을 평가합니다.

在计划中评价实验性能。