00 오리엔테이션

스터디 개요

- 스터디 목적 : 캐글을 통한 머신러닝/딥러닝 실전 능력 향상
- 스터디 시간 : 일요일 오전 10 부터 2시간
- 스터디 장소 : 양재 토즈
- 스터디 운영 계획
  - 3월 : 머신러닝
  - 4월 : 머신러닝, 딥러닝
  - 5월 : 딥러닝 및 마무리
- 진행 방법
  - 한명 씩 돌아가며 교재의 내용 요약 정리 발표
  - 각자, 본인의 폴더를 만들고, 교재 및 캐글내의 자료를 참고하여 자신만의 풀이를 만들어 오기
  - 기본 스킬셋 획득이 제 1 목표 + 알파

01 왜 캐글인가?

Why 캐글?

- 데이터 과학 및 머신러닝 경진대회를 주최하는 온라인 커뮤니티
- 2022년 6월 기준 1천만명의 사용자
- 알고리즘 만으로는 해결하기 힘든 문제들 → ‘데이터'가 핵심
- 데이터 과학 및 머신러닝 역량을 키우기에 최적의 조건
  - 다양한 종류의 경진 대회, 코드와 아이디어의 자유로운 공유, 메달 시스템
- 취업시 우대

캐글 구성요소

- 경진대회, 데이터셋, 노트북, 토론, 강좌로 구성
- 경진대회
  - 기업은 데이터와 돈을 주고 대회 개최를 캐글에 의뢰,
  - 참여자는 기업이 요구하는 지표를 기준으로 모델을 제출
  - 상위의 성적의 참여자는 상금을 받고, 코드를 제공
  - Getting Started (입문용) → Playground(초보자를 위한) → Featured (일반) 의 수준의 난이도
- 데이터셋 : 누구나 공유 가능, 추천을 받으면 메달 획득 가능
- 코드 : 작성한 코드를 공유, 추천을 받으면 메달 획득 가능
- 토론 : 대회 관련 문의 가능, 아이디어 공유
- 강좌 : 기초, 중급, 고급 강좌 제공

캐글러 등급

- Novice, Contributor, Expert, Master, Grandmaster
- 경진대회, 데이터셋, 노트북, 토론마다 등급 부여, 등급을 올리려면 메달을 따야
- 메달 : 금메달, 은메달, 동메달

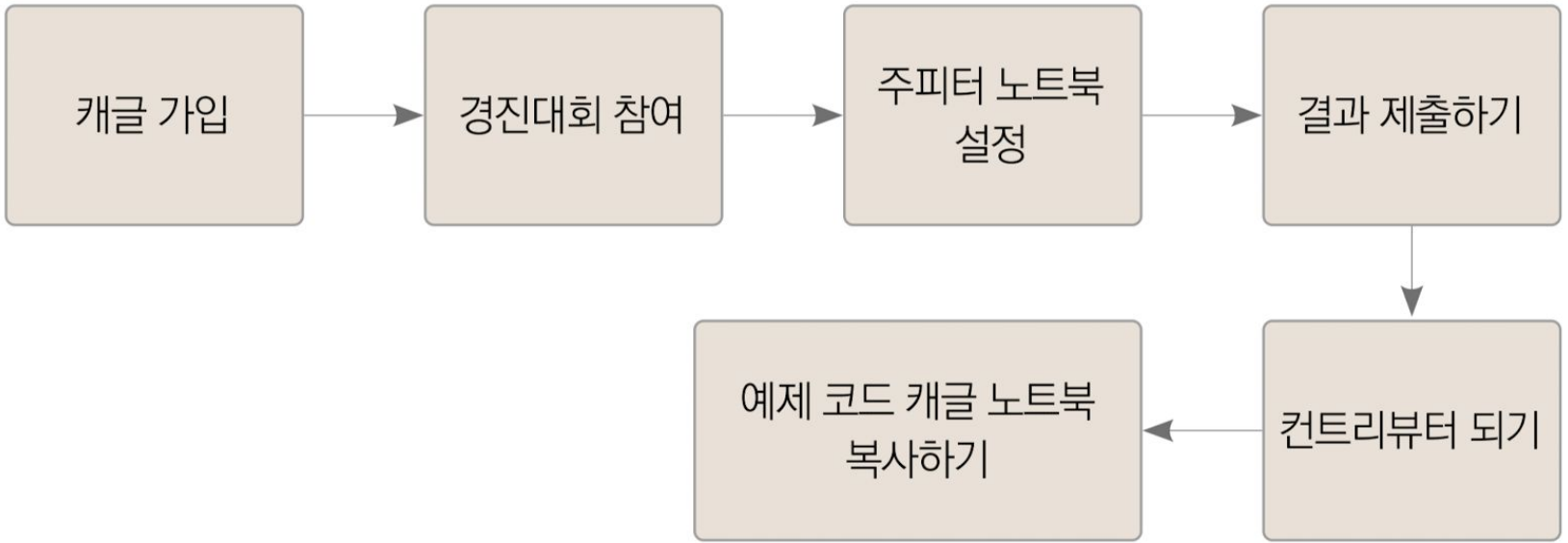
▼ 경진대회 메달 조건

	0~99팀	100~249팀	250~999팀	1000팀 이상
동메달	상위 40%	상위 40%	상위 100위	상위 10%
은메달	상위 20%	상위 20%	상위 50위	상위 5%
금메달	상위 10%	상위 10위	상위 10위 + 0.2%[1]	상위 10위 + 0.2%

- Contributor : 간단한 조건만 만족하면 가능
- Expert : 경진대회 (동메달 2개), 데이터셋 (동메달 3개), 노트북 (동메달 5개), 토론 (50개)
- Master : 경진대회 (금 1, 은2), 데이터셋 (금1, 은4), 노트북 (은10), 토론 (은 50 포함 200개)
- Grandmaster : 경진대회(솔로 금 1, 금 5개), 데이터셋(금5, 은5), 노트북 (금15) 토론 (금50개 포함 500개)

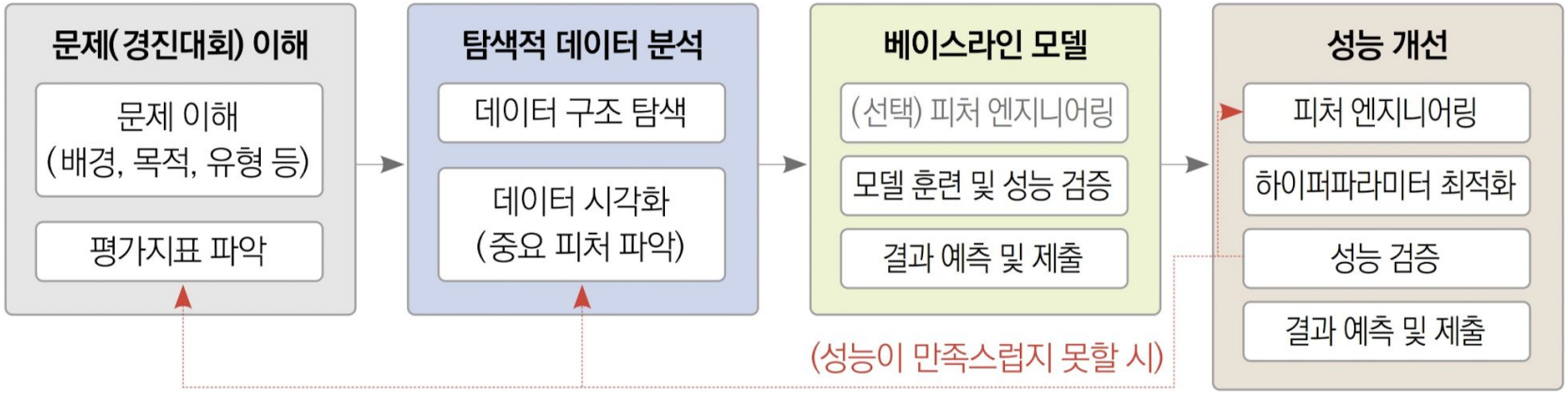
02 캐글 정복 첫걸음

요약



03 문제해결 프로세스 및 체크리스트

머신러닝 문제해결 프로세스

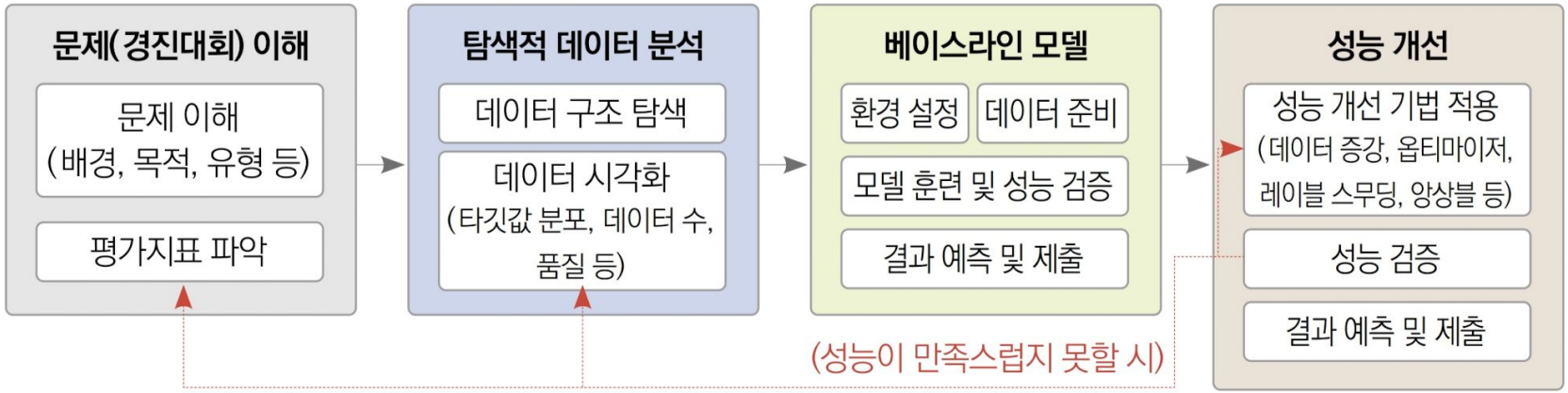


정형 데이터 (엑셀, csv) 위주

머신러닝 문제해결 체크리스트

- 문제 이해
  - 제목, 미션, 문제유형, 평가지표
- 데이터 둘러보기
  - 파일별 용도 파악, 데이터양(레코드수, 피쳐수, 전체 용량) 피쳐이해 (이름, 의미, 타입, 결측값, 고윳값 실제값, 데이터 종류 등), 훈련 데이터와 테스트 데이터 차이, 타깃값
- 데이터 시각화
  - 시각화를 위한 피쳐 엔지니어링, 수치형 데이터 시각화, 범주형 데이터 시각화, 데이터 관계, 피쳐 파악, 이상치 파악
- 베이스라인 모델
  - 준비 : 데이터 불러오기, 피쳐 엔지니어링, 평가지표 계산 함수 준비
  - 모델훈련 : 모델 생성, 훈련
  - 성능 검증 : 예측, 평가
  - 예측 및 결과 제출 : 최종예측, 제풀 파일 생성, 제출
- 성능 개선
  - 피쳐 엔지니어링 : 이상치 제거, 결측값 처리, 데이터 인코딩, 타입 변경, 파생 피쳐 생성, 시차 피쳐 생성, 스케일링, 데이터 다운 캐스팅, 조합 생성, 필요 없는 피쳐 제거
  - 훈련 : 하이퍼파라미터 의미 파악 및 선별, 값 번위 설정, 최적화 기법 선택 및 적용
  - 성능 검증 : 예측,평가
  - 예측 및 제출 : 최종 예측, 제출 파일 생성 및 제출

딥러닝 문제해결 프로세스



비정형 데이터 (이미지, 음성, 텍스트 등) 위주

딥러닝 문제해결 체크리스트

- 문제 이해
  - 제목, 미션, 문제 유형, 평가지표
- eda
  - 데이터 둘러보기 : 파일별 용도 파악, 데이터양, 피쳐 이해, 훈련 데이터와 테스트 데이터 차이, 타깃값
  - 데이터 시각화: 타깃값 분포, 분류별 이미지 출력, 이미지 형태 확인, 불량 이미지 포함 여부
- 베이스라인 모델
  - 환경 설정 : 시드값 고정, gpu 설정
  - 데이터 준비 : 훈련/검증 데이터 분리, 데이터겟 클레스 정의, 데이터 증강, 테이터셋 생성, 데이터 로더 생성
  - 모델훈련 : 모델 생성, 훈련
  - 성능 검증 : 예측, 평가
  - 예측 및 결과 제출 : 최종 예측, 제출 파일 생성, 제출
- 성능 개선
  - 데이터 증강 : 변환기 목록
  - 모델 개선 : 사전 학습된 모델 물색, 선정 모델 목록
  - 훈련 단계 최적화 : 손실 함수, 홉티마이저, 스케줄러, 에폭수
  - 예측 단계 최적화 : 테스트 단계 데이터 증강, 레이블 스무딩
  - 성능 검증 : 예측, 성능 평가
  - 예측 및 결과 제출 : 최종 예측, 제출 파일 생성, 제출

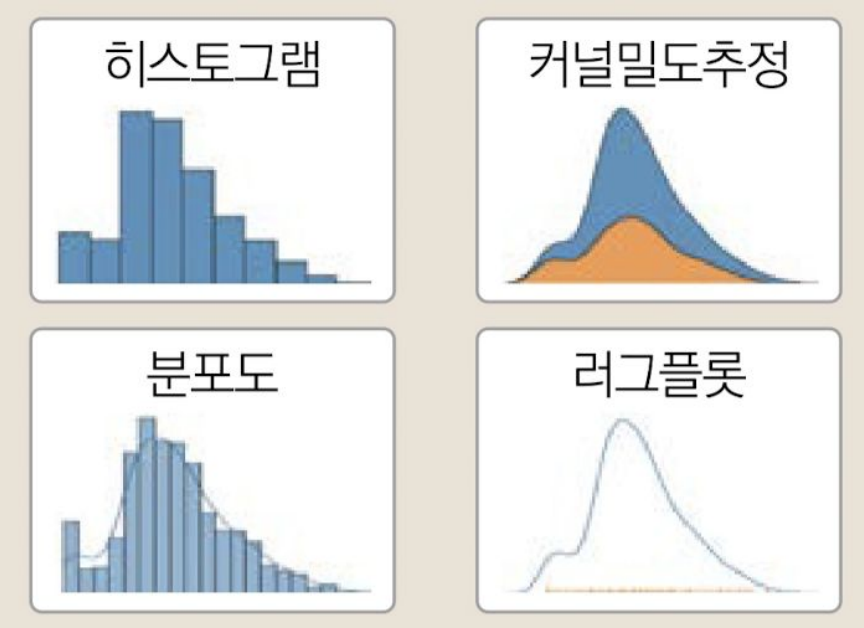
04 데이터를 한눈에

데이터 종류

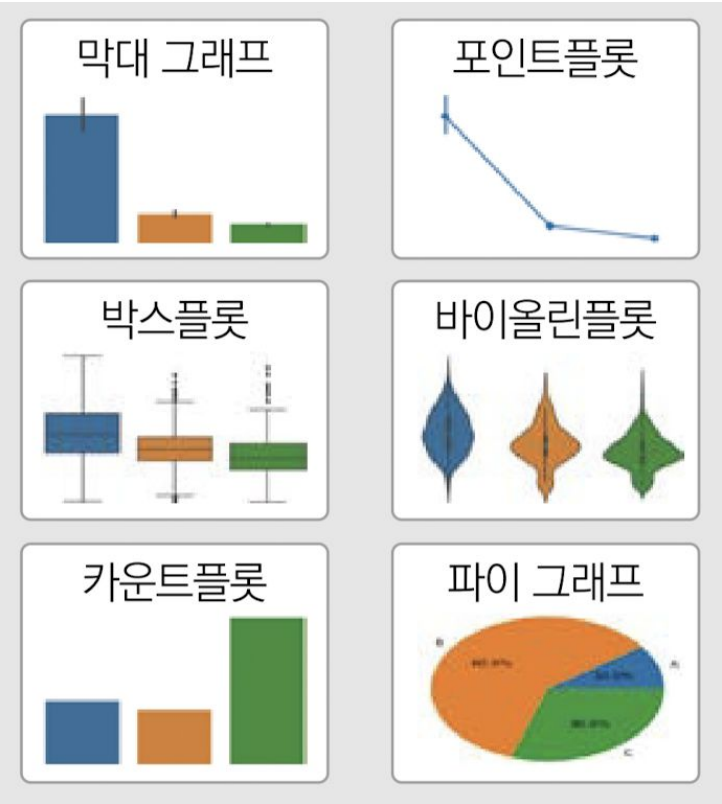
▼ 데이터 종류

대분류	소분류	예시
수치형 데이터 (사칙 연산이 가능한 데이터)	연속형 데이터	키, 몸무게, 수입
	이산형 데이터	과일 개수, 책의 페이지 수
범주형 데이터 (범주로 나누어지는 데이터)	순서형 데이터	학점, 순위(랭킹)
	명목형 데이터	성별, 음식 종류, 우편 번호

수치형 데이터 시각화



범주형 데이터 시각화



데이터 관계 시각화