**슬라이드 1: 소개** "오늘 제가 여러분에게 소개할 주제는 사이킷런의 model selection입니다. 이것은 데이터 과학에서 매우 중요한 단계로, 우리의 모델이 실제 세계 데이터에 얼마나 잘 작동하는지를 평가하는 데 도움이 됩니다. 오늘 우리는 특히 학습/테스트 데이터 세트 분리, 교차 검증, 그리고 하이퍼파라미터 최적화에 초점을 맞출 것입니다."

**슬라이드 2: Model Selection 소개** "Model Selection은 학습 데이터와 테스트 데이터를 분리하고, 모델을 평가하기 위해 교차 검증을 사용하며, 최적의 모델 성능을 달성하기 위해 하이퍼파라미터를 조정하는 과정을 포함합니다. 이를 통해 우리는 모델의 일반화 능력을 평가하고 최적화할 수 있습니다."

**슬라이드 3: Train\_test\_split** "train\_test\_split 함수는 데이터를 학습 세트와 테스트 세트로 분할하는 과정입니다. 이것은 모델의 성능을 검증하기 위해 필수적이며, 이를 통해 모델이 본 적 없는 데이터에 대해 얼마나 잘 작동하는지를 평가할 수 있습니다. 주요 매개변수로는 특성과 레이블 데이터, 테스트 세트의 크기, 데이터 세트의 무작위 분할을 제어하는 random\_state, 그리고 데이터를 섞을지 여부를 결정하는 shuffle이 있습니다."

**슬라이드 4: 교차 검증** "교차 검증은 단일 검증 데이터 세트를 사용하는 대신 여러 데이터 세트에 대한 학습과 평가를 반복함으로써 모델의 성능을 더 잘 평가하는 방법입니다. 이 접근 방식은 모델이 특정한 데이터 세트에 과적합되는 것을 방지하고, 모델의 일반화 능력을 더 정확하게 평가할 수 있도록 합니다."

**슬라이드 5: KFold** "KFold 교차 검증은 전체 데이터를 K개의 세트로 나누고, 각 세트에 대해 학습과 평가를 반복합니다. 이 방법은 모든 데이터가 최소 한 번은 검증 세트로 사용될 수 있도록 보장합니다. KFold의 주요 단점은 데이터의 분포가 균등하지 않을 수 있다는 것입니다, 이를 해결하기 위해 Stratified KFold를 사용할 수 있습니다."

**슬라이드 6: Stratified KFold** "Stratified KFold는 각 분류 그룹 내 데이터의 비율을 유지하면서 데이터를 분할합니다. 이는 특히 불균형 데이터 세트에서 중요하며, 더 신뢰할 수 있는 검증 점수를 제공합니다."

**슬라이드 7: Cross\_val\_score** "cross\_val\_score 함수는 교차 검증을 더 쉽게 구현할 수 있도록 해줍니다. 이 함수는 데이터를 폴드로 나누고, 학습과 평가를 자동으로 반복하여, 각 반복에서의 성능 점수를 반환합니다."