해석 핵심 : 분기할 때 사용된 변수, 분기 기준

1. 이 트리가 말해주는 한 줄 요약

"82승 이상 팀은 기본적으로 **실점(ra)**이 적은 팀이고, **장타력(SLG)**과 **타자 볼넷(total_bb)**으로 승률이 갈린다."

실제로 분기에서 쓰인 변수 : ra(시즌 총 실점), slg(장타율), total_bb(타자 볼넷 수)

pitch_so(투수 삼진수), pitch_bb(투수 볼넷 수), sb(도루성공률)는 아예 트리에 사용되지 않음
→ 상대적으로 영향이 작다고 판단된 것.

2. 루트 노드 해석 : RA ≥ 750 이냐 아니냐

맨 위 split : ra >= 750

- RA ≥ 750 (실점이 많은 팀) → 왼쪽 서브 트리 → 기본값이 Loss (미진출/82승 미만)
- RA < 750 (실점이 적은 팀) → 오른쪽 서브 트리 → 기본값이 Win (82승 이상)

👉 실점 750개가 "강팀/약팀을 가르는 1차 기준선"
투수·수비력이 일정 수준 이하로 떨어지면(=실점이 많아지면), 공격이 아무리 좋아도 82승 어려움

3. 왼쪽 서브 트리 : "투수도 안 좋으면 공격이 살 길이다"

3-1. RA ≥ 750 & SLG < 0.42 → 거의 다 Loss

왼쪽 제일 큰 잎:

- 조건 : ra >= 750 그리고 slg < 0.42
- 결과 : 거의 전부 Loss (푸른색 잎, Win 거의 없음)

➡️ 투수력 나쁘고 장타력까지 평균 이하인 팀은 82승 거의 불가능
= "투수 안 좋은 팀(실점이 많음)이 살아남으려면 최소 SLG 0.42는 찍어야 한다"

3-2. RA ≥ 750 & SLG ≥ 0.42 → 일부만 살아남는 공격형 팀

같은 RA≥750 그룹에서도, 장타력이 높은 팀은 한 번 더 갈려:

- 조건 대략 : ra >= 750, slg >= 0.42
- 그 안에서 ra >= 842 이면 Loss, 그보다 낮으면 Win 쪽으로 분류

➡️ 실점이 많아도(=투수 약점),
장타력(SLG ≥ 0.42) + RA가 최악(842 이상)까진 아니면 82승을 할 수 있는 팀으로 본다는 뜻.

인사이트

- 아주 안 좋은 투수진(842실점 수준) → 공격이 좋아도 한계
- 그보다 약간 덜 나쁜 수준이면, 화력으로 커버한 강타선 팀이 실제로 존재했다는 걸 반영.

4. 오른쪽 서브 트리 : "투수 좋은 팀들 안에서의 미세한 차이"

루트에서 RA < 750인 팀들은 기본적으로 Win(82승 이상 가능성이 더 큼) 으로 시작해서
slg와 ra, total_bb로 다시 나뉘어.

4-1. SLG < 0.4 인 팀

- 조건 : ra < 750 & slg < 0.4
- 여기서 다시 RA로 갈려 : RA ≥ 616 → Loss 잎 / RA < 616 → Win 잎

➡️ 공격이 약한 팀(장타율 0.4 미만)은 진짜 말도 안 되게 실점을 줄여야(616 미만) 82승이 가능
= "수비/투수 절대강팀만이 빈약한 타선을 커버할 수 있다."

4-2. SLG ≥ 0.4 인 팀

- 조건 : ra < 750 & slg ≥ 0.4
- 다시 ra >= 713 으로 갈리는데, RA(실점)가 조금 높은 편이어도(713 이상) 여전히 Win
- RA가 더 낮고(ra <713) 그 안에서도 : slg < 0.42 이면서 total_bb < 499 → Loss
  slg < 0.42 이면서 total_bb ≥ 499 → Win
  slg ≥ 0.42 → 대부분 Win

➡️ 여기서 나오는 메시지는 :
1. 투수력이 아주 완벽하지 않아도(750 > RA >= 713) 장타력(SLG ≥ 0.4) & 볼넷(total_bb ≥ 499)이 받쳐주면 82승 이상 충분히 가능.
2. 반대로, 장타력은 어느 정도 있는데(SLG 약 0.4 근처) 볼넷이 적으면(total_bb < 499)
실제로는 82승까지 못 가는 팀이 많았다는 걸 반영.

5. 이 트리에서 보이는 "강팀 타입" 정리

1. **투수 절대강팀형** > 실점 억제로 버티는 팀
   - ra(실점) < 616
   - SLG(장타율, 타자 능력)는 다소 낮아도 괜찮음 (slg < 0.4도 가능)

2. **균형형 강팀** > 공격과 수비 모두 준수한 팀
   - ra < 750 + slg ≥ 0.4
   - 볼넷도 어느 정도 많음 (total_bb ≥ 499)

3. **공격 폭발형 (투수 약점 커버) > 실점이 많음에도 장타와 득점력으로 우승하는 팀**
   - ra ≥ 750 (하지만 실점이 최악 수준까지는 아닌 팀)
   - slg ≥ 0.42

- 투수도 평범/나쁘고(ra ≥ 750) 장타도 약한(slg < 0.42) 팀
- 투수는 괜찮은 편이지만(ra < 750) 장타력도 낮고(SLG < 0.4) 볼넷도 많지 않은 팀

은 대부분 82승 미만(Loss) 으로 떨어짐

6. 추가로 말할 수 있는 인사이트

1. 투수 지표(pitch_so, pitch_bb) 대신 순수 실점 RA만 살아남았다
   - 삼진/볼넷 같은 세부 투구지표보다 **결과 지표(실점 합계)**가 승리 예측에 더 직접적으로 연결된다는 해석 가능.
2. 도루(SB)는 트리에 전혀 쓰이지 않았다
   - 이 모델, 이 기간(2000–2015) MLB에서는 팀의 시즌 승수(82승 기준)에 거의 설명력을 주지 못했다는 함의.
3. SLG 임계값 0.40, 0.42가 계속 등장
   - "평균 이상 장타력 여부"가 강팀·약팀을 가르는 핵심 기준으로 잡힘.
4. 출루(볼넷)과 장타의 조합이 중요
   - SLG가 애매할 때 타자 볼넷수(total_bb)가 다시 한 번 Win/Loss를 가른다.
   - "장타+볼넷 = 강팀", "장타만 있거나 볼넷만 있으면 애매"라는 식의 해석 가능.

7. 간단 요약

"2000–2015년 MLB 팀 데이터를 기반으로 82승 이상 여부를 의사결정나무로 예측한 결과,
첫 번째로 중요한 변수는 시즌 실점 RA였고, RA 750점을 기준으로 강팀과 약팀이 1차적으로 구분되었다. **그 이후에는** 팀의 장타력(SLG)과 볼넷 수(total_bb)가 추가로 승률 차이를 설명했다.
도루나 세부 투구지표(pitch_so, pitch_bb)는 최종 트리에서 사용되지 않아 장기적인 시즌 승수에는 상대적으로 영향이 적은 것으로 나타났다."

📌 1. Confusion Matrix 완전 해석

Confusion Matrix

	실제 Loss	실제 Win
예측 Loss	56	5
예측 Win	15	67

🔍 해석 포인트 1 : Loss 예측은 매우 정확함

- Loss 중 56/71을 맞춤 → 정확도 0.788
- Loss를 Win으로 잘못 예측한 경우는 15건

➡️ "82승 못 넘는 팀"을 잘 골라낸다.

🔍 해석 포인트 2 : Win 예측도 상당히 좋은 편

- Win 중 67/72를 맞춤 → 정확도 0.930
- Win을 Loss로 잘못 예측한 경우는 5건

➡️ "강팀(82승 이상)"까지 예측 성능이 매우 뛰어나다.

🔍 해석 포인트 3: 전체 패턴

- Loss보다 Win 쪽 Precision이 더 높음
  → 예측 Win 중 91.8%가 진짜 Win
- MLB 강팀의 특징을 잘 잡아냈다는 뜻

🔍 McNemar's Test P-value = 0.044

- 이 값이 0.05보다 작으면
  → 양쪽 클래스의 오차가 의미 있게 다르다는 뜻
- 즉, 모델이 Loss/Win을 "동등하게" 잘 예측하는 것은 아님

결론 : Loss보다 Win 쪽이 더 잘 맞는 모델이다. 즉, 강팀(82승 이상)을 설명하는 능력이 더 강함.

📌 2. 모델 정확도 지표 해석

✔ Accuracy = 0.8601

→ 전체 예측 중 **86.01%**가 맞았다는 뜻

→ 의사결정나무 모델로는 매우 우수한 성능

✔ 95% CI = (0.7923, 0.9124)

→ 신뢰구간도 좁고 0.8 이상 유지

→ 안정적으로 높은 정확도

✔ Kappa = 0.72

- 0.7 이상이면 Substantial Agreement(높은 일치도)
- 클래스 불균형을 고려한 정확도이므로 신뢰도가 높다.

✔ Sensitivity (Recall for Loss) = 0.7887

- Loss를 Loss로 맞춘 비율
- Loss 팀은 잘 잡아내지만 Win보다는 다소 낮음

✔ Specificity (Recall for Win) = 0.9306

- Win을 Win으로 맞춘 비율
- Win 클래스의 재현율이 매우 높음

➡️ 강팀(82승 이상) 예측 성능이 매우 뛰어남.

✔ PPV (Precision for Loss) = 0.9180 (예측 Loss 중 91.8%가 실제 Loss)

→ Loss를 "함부로" 찍지 않음

✔ NPV (Precision for Win) = 0.8171 (예측 Win 중 81%가 진짜 Win)

→ 잘못 예측한 Win 비율이 낮음

📌 Balanced Accuracy = 0.8596

민감도/특이도의 평균 = 0.86

→ 양쪽 클래스 모두 균형적으로 잘 맞춤.

🔥 **인사이트 정리**

1) 실점(RA)이 압도적 1위 (73.65)
   - 팀 승률을 가장 많이 설명하는 변수
   - 실점 억제력이 강팀/약팀을 1차적으로 나눔

2) 장타율(SLG)이 매우 중요한 공격 변수 (56.45) > **장타력은 득점/승리에 가장 직접적으로 연결됨**
   - RA 다음으로 영향력
   - 평균 0.40~0.42가 강팀/약팀을 구분하는 기준으로 작동
   - MLB의 실제 승리 패턴과 일치

3) 투수 탈삼진(pitch_so)이 3순위 (28.7)
   - 직접 트리에 포함되지는 않았지만 RA와 상관이 높아서 간접적 영향이 컸을 가능성 O

4) 투수 볼넷(pitch_bb) 또한 중요 (27.3)
   - 이 역시 RA에 영향을 주므로 중요도 상승

5) 타자 볼넷(total_bb)은 중간 중요도 (14.4)
   - RA·SLG보다는 영향이 낮지만 트리에서는 Win/Loss를 결정짓는 최종 분기 중 하나로 사용됨
   - 출루력 높은 팀이 더 높은 승률을 보인다는 MLB 이론과 일치.

6) 도루(SB)는 최하위 (5.13)
   - 시즌 승률(82승 기준)과는 거의 무관
   - MLB에서 오래전부터 알려진 사실 그대로 반영됨