

*Business Analytics*

# 어떤 성적을 내야 **우승 가능성이 높아질까?**

*Final Term Project*

201501267 우창윤  
201501240 전영조  
201501220 손정습



## 서론

- 야구 란?
- 연구배경과 목적
- 데이터 수집과 계절데이터 추가

## 본론

- 데이터 확인
- 분석모델 설정
- 데이터 분석 과정
- Decision Tree
- Confusion Matrix
- Random Forest
- Improved Accuracy

## 결론

- 결과 및 제시사안
- Q&A

서론 - 소개

## 주제 선택 동기

# 야구?

BASEBALL

오심 벤치 클리어링

홈런

볼넷

SK 와이번스

롯데 자이언츠

스트라이크

우승팀은?

가을야구

KBO

2020 KBO 리그

키움 히어로즈

삼진아웃

심판

야구 펜

볼

스포츠 토포

홈런왕

불펜

덕 아웃

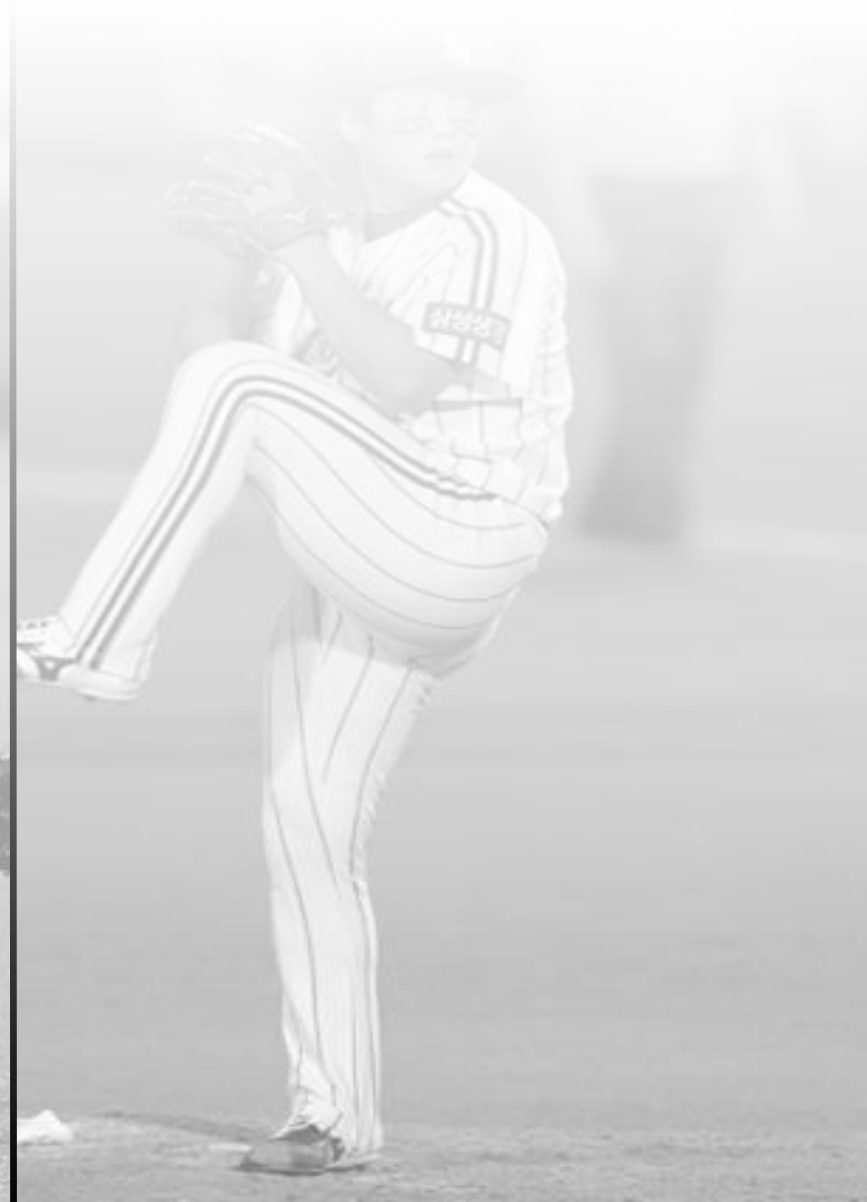
통위총(통일정책위원회)이 2014년 12월 15일 제1차 회의를 개최했다. 이 자리에서 통위총은 '통일정책'을 논의하고, '통일정책'을 수립하는 데 필요한 사항을 논의했다. 이 자리에서 통위총은 '통일정책'을 수립하는 데 필요한 사항을 논의했다. 이 자리에서 통위총은 '통일정책'을 수립하는 데 필요한 사항을 논의했다.



# 야구에는 이런 포지션이 있다!

## 1. 타격

득점을 위한 공격



야구에는 이런 포지션이 있다!



## 2. 수비

침착한 수비



# 야구에는 이런 포지션이 있다!



## 3. 투구

실점을 막기 위한 투구

## 0. 연구의 배경과 목적

### • 연구 배경

- 각 포지션 별로 데이터를 분석하기 때문에 **많은 데이터가 존재**한다.
- 야구는 여러 스포츠 중 데이터가 가장 **잘 정돈 되어 있는 스포츠**이다.
- 데이터를 분석하여 보다 **쉽게 예측**을 할 수 있다.

### 줄거리

게임의 역사를 바꾼 감동의 리그가 시작된다!

메이저리그 만년 최하위에 그나마 실력 있는 선수들은 다른 구단에 뺏기기 일수인 ‘오클랜드 애슬레틱스’. 돈 없고 실력 없는 오합지졸 구단이란 오명을 벗어 던지고 싶은 단장 ‘빌리 빈(브래드 피트)’은 경제학을 전공한 ‘피터’를 영입, 기존의 선수 선발 방식과는 전혀 다른 파격적인 ‘머니볼’ 이론을 따라 새로운 도전을 시작한다. 그는 경기 데이터에만 의존해 사생활 문란, 잦은 부상, 최고령 등의 이유로 다른 구단에서 외면 받던 선수들을 팀에 합류시키고, 모두가 미친 짓이라며 그를 비난한다.

과연 빌리와 애슬레틱스 팀은 ‘머니볼’의 기적을 이룰 수 있을까?



머니볼 (Moneyball, 2011)

네티즌 ★★★★★ 8.37 (2,418) | 기자·평론가 ★★★★★ 8.13 (6) 평점주기▶

드라마 | 2011.11.17. 개봉 | 133분 | 미국 | 12세 관람가

감독 배넷 밀러

관객수 641,323명

수상정보 17회 크리틱스 초이스 시상식(각색상), 46회 전미 비평가 협회상(남우주연상), 24회 시카고 비평가 협회상(각색상)▼

내용 게임의 역사를 바꾼 감동의 리그가 시작된다! 메이저리그 만년 최... 더보기

부가정보 공식사이트

- 데이터분석만으로 야구경기에서 충분히 좋은 성적을 거둘 수 있다.



## 0. 연구의 배경과 목적



왜 매년 야구는 우승팀만 예상을 할까?

우승팀만의 규칙이 있지 않을까??

**그래서!**

**우승팀의 규칙을 생각해 보기로 했습니다!**

# 1. 데이터 수집과 계절 데이터 추가

<http://www.statiz.co.kr/stat.php?lr=5>

시즌기록실

시즌기록실

통산기록실

팀기록실

특별기록실

연도별 상수

WAR Special

팀기록실

종합

타격

투구

수비

2020

연도

시작

끝

팀:전체

포지션

정규

규정

상황

옵션

[자동 : 전체]

완료

삼성:두산 롯데:키움 KT:SK LG:한화 NC:KIA

기본확장가치클러치타석타구1타구2파워팀배팅1팀배팅2도루주루구종가치구종구사

리그 기록

순	이름	팀	정렬	G	타석	타수	득점	안타	2타	3타	홈런	루타	타점	도루	도실	볼넷	사구	고4	삼진	병살	희타	희비	비율					
			WAR*																				타율	출루	장타	OPS	wOBA	wRC+
1	리그	20	52.57	4573	14071	12425	1885	3378	630	49	361	5189	1787	209	89	1236	182	28	2491	287	96	132	.272	.343	.418	.761	.340	99.8

팀 기록

순	이름	팀	정렬	G	타석	타수	득점	안타	2타	3타	홈런	루타	타점	도루	도실	볼넷	사구	고4	삼진	병살	희타	희비	비율					
			WAR*																				타율	출루	장타	OPS	wOBA	wRC+
1	NC	20	10.39	480	1472	1285	247	391	76	6	55	644	239	22	9	134	31	4	256	26	10	12	.304	.380	.501	.882	.386	125.5
2	LG	20	7.90	472	1397	1236	212	359	68	7	38	555	204	26	9	116	17	5	227	28	10	18	.290	.355	.449	.804	.355	114.7
3	키움	20	7.31	467	1473	1279	216	341	67	7	43	551	200	22	5	156	17	3	279	26	6	15	.267	.350	.431	.781	.348	105.9
4	두산	20	7.23	450	1412	1243	210	367	70	4	35	550	198	12	9	122	18	3	216	39	8	21	.295	.361	.442	.804	.357	114.0
5	KT	20	7.14	472	1423	1279	207	375	67	10	41	585	194	18	10	108	12	2	256	18	11	13	.293	.351	.457	.808	.356	110.9
6	KIA	20	5.21	437	1434	1258	184	344	60	2	38	522	177	13	5	136	17	0	245	29	12	11	.273	.350	.415	.765	.343	99.7
7	롯데	20	3.91	428	1405	1245	168	331	65	2	26	478	155	26	7	125	17	4	234	28	6	12	.266	.338	.384	.722	.327	90.8
8	삼성	20	2.60	471	1381	1214	184	308	66	2	35	483	176	28	15	118	22	2	253	27	10	17	.254	.327	.398	.725	.324	85.8
9	SK	20	1.44	463	1368	1202	142	284	46	8	29	433	135	22	11	128	15	5	267	31	14	9	.236	.315	.360	.676	.304	77.6
10	한화	20	-0.56	433	1306	1184	115	278	45	1	21	388	109	20	9	93	16	0	258	35	9	4	.235	.298	.328	.626	.287	67.9

3. 타격, 투구 등 설정

1. 조사 하고자 하는 연도 설정

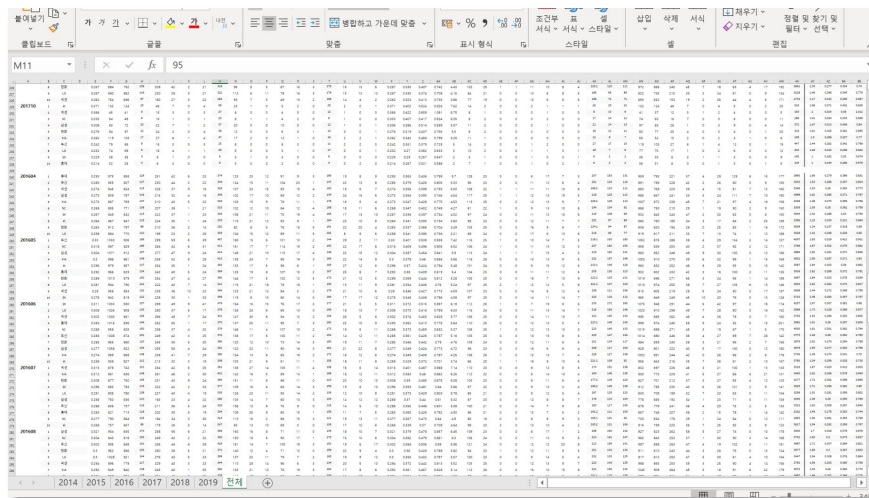
2. 데이터 내보내기

# 1. 데이터 수집과 계절 데이터 추가

가 가 가		가		내년		병합하고 가운데 맞춤		%		조건부 표 셀		삽입 삭제 서식		[새우기] 지우기		정렬 및 찾기 및 필터 선택	
클립보드		글꼴		맞춤		표시 형식		스타일		셀		편집					
M11		95															
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z AA AB AC AD AE AF AG AH AI AJ AK AL AM AN AO AP AQ AR AS AT AU AV AW AX AY AZ BA BB		2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2030 2031 2032 2033 2034 2035 2036 2037 2038 2039 2040 2041 2042 2043 2044 2045 2046 2047 2048 2049 2050 2051 2052 2053 2054 2055 2056 2057 2058 2059 2060 2061 2062 2063 2064 2065 2066 2067 2068 2069 2070 2071 2072 2073 2074 2075 2076 2077 2078 2079 2080 2081 2082 2083 2084 2085 2086 2087 2088 2089 2090 2091 2092 2093 2094 2095 2096 2097 2098 2099 2100 2101 2102 2103 2104 2105 2106 2107 2108 2109 2110 2111 2112 2113 2114 2115 2116 2117 2118 2119 2120 2121 2122 2123 2124 2125 2126 2127 2128 2129 2130 2131 2132 2133 2134 2135 2136 2137 2138 2139 2140 2141 2142 2143 2144 2145 2146 2147 2148 2149 2150 2151 2152 2153 2154 2155 2156 2157 2158 2159 2160 2161 2162 2163 2164 2165 2166 2167 2168 2169 2170 2171 2172 2173 2174 2175 2176 2177 2178 2179 2180 2181 2182 2183 2184 2185 2186 2187 2188 2189 2190 2191 2192 2193 2194 2195 2196 2197 2198 2199 2200 2201 2202 2203 2204 2205 2206 2207 2208 2209 2210 2211 2212 2213 2214 2215 2216 2217 2218 2219 2220 2221 2222 2223 2224 2225 2226 2227 2228 2229 2230 2231 2232 2233 2234 2235 2236 2237 2238 2239 2240 2241 2242 2243 2244 2245 2246 2247 2248 2249 2250 2251 2252 2253 2254 2255 2256 2257 2258 2259 2260 2261 2262 2263 2264 2265 2266 2267 2268 2269 2270 2271 2272 2273 2274 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285 2286 2287 2288 2289 2290 2291 2292 2293 2294 2295 2296 2297 2298 2299 2300 2301 2302 2303 2304 2305 2306 2307 2308 2309 2310 2311 2312 2313 2314 2315 2316 2317 2318 2319 2320 2321 2322 2323 2324 2325 2326 2327 2328 2329 2330 2331 2332 2333 2334 2335 2336 2337 2338 2339 2340 2341 2342 2343 2344 2345 2346 2347 2348 2349 2350 2351 2352 2353 2354 2355 2356 2357 2358 2359 2360 2361 2362 2363 2364 2365 2366 2367 2368 2369 2370 2371 2372 2373 2374 2375 2376 2377 2378 2379 2380 2381 2382 2383 2384 2385 2386 2387 2388 2389 2390 2391 2392 2393 2394 2395 2396 2397 2398 2399 2400 2401 2402 2403 2404 2405 2406 2407 2408 2409 2410 2411 2412 2413 2414 2415 2416 2417 2418 2419 2420 2421 2422 2423 2424 2425 2426 2427 2428 2429 2430 2431 2432 2433 2434 2435 2436 2437 2438 2439 2440 2441 2442 2443 2444 2445 2446 2447 2448 2449 2450 2451 2452 2453 2454 2455 2456 2457 2458 2459 2460 2461 2462 2463 2464 2465 2466 2467 2468 2469 2470 2471 2472 2473 2474 2475 2476 2477 2478 2479 2480 2481 2482 2483 2484 2485 2486 2487 2488 2489 2490 2491 2492 2493 2494 2495 2496 2497 2498 2499 2500 2501 2502 2503 2504 2505 2506 2507 2508 2509 2510 2511 2512 2513 2514 2515 2516 2517 2518 2519 2520 2521 2522 2523 2524 2525 2526 2527 2528 2529 2530 2531 2532 2533 2534 2535 2536 2537 2538 2539 2540 2541 2542 2543 2544 2545 2546 2547 2548 2549 2550 2551 2552 2553 2554 2555 2556 2557 2558 2559 2560 2561 2562 2563 2564 2565 2566 2567 2568 2569 2570 2571 2572 2573 2574 2575 2576 2577 2578 2579 2580 2581 2582 2583 2584 2585 2586 2587 2588 2589 2590 2591 2592 2593 2594 2595 2596 2597 2598 2599 2600 2601 2602 2603 2604 2605 2606 2607 2608 2609 2610 2611 2612 2613 2614 2615 2616 2617 2618 2619 2620 2621 2622 2623 2624 2625 2626 2627 2628 2629 2630 2631 2632 2633 2634 2635 2636 2637 2638 2639 2640 2641 2642 2643 2644 2645 2646 2647 2648 2649 2650 2651 2652 2653 2654 2655 2656 2657 2658 2659 2660 2661 2662 2663 2664 2665 2666 2667 2668 2669 2670 2671 2672 2673 2674 2675 2676 2677 2678 2679 2680 2681 2682 2683 2684 2685 2686 2687 2688 2689 2690 2691 2692 2693 2694 2695 2696 2697 2698 2699 2700 2701 2702 2703 2704 2705 2706 2707 2708 2709 2710 2711 2712 2713 2714 2715 2716 2717 2718 2719 2720 2721 2722 2723 2724 2725 2726 2727 2728 2729 2730 2731 2732 2733 2734 2735 2736 2737 2738 2739 2740 2741 2742 2743 2744 2745 2746 2747 2748 2749 2750 2751 2752 2753 2754 2755 2756 2757 2758 2759 2760 2761 2762 2763 2764 2765 2766 2767 2768 2769 2770 2771 2772 2773 2774 2775 2776 2777 2778 2779 2780 2781 2782 2783 2784 2785 2786 2787 2788 2789 2790 2791 2792 2793 2794 2795 2796 2797 2798 2799 2800 2801 2802 2803 2804 2805 2806 2807 2808 2809 2810 2811 2812 2813 2814 2815 2816 2817 2818 2819 2820 2821 2822 2823 2824 2825 2826 2827 2828 2829 2830 2831 2832 2833 2834 2835 2836 2837 2838 2839 2840 2841 2842 2843 2844 2845 2846 2847 2848 2849 2850 2851 2852 2853 2854 2855 2856 2857 2858 2859 2860 2861 2862 2863 2864 2865 2866 2867 2868 2869 2870 2871 2872 2873 2874 2875 2876 2877 2878 2879 2880 2881 2882 2883 2884 2885 2886 2887 2888 2889 2890 2891 2892 2893 2894 2895 2896 2897 2898 2899 2900 2901 2902 2903 2904 2905 2906 2907 2908 2909 2910 2911 2912 2913 2914 2915 2916 2917 2918 2919 2920 2921 2922 2923 2924 2925 2926 2927 2928 2929 2930 2931 2932 2933 2934 2935 2936 2937 2938 2939 2940 2941 2942 2943 2944 2945 2946 2947 2948 2949 2950 2951 2952 2953 2954 2955 2956 2957 2958 2959 2960 2961 2962 2963 2964 2965 2966 2967 2968 2969 2970 2971 2972 2973 2974 2975 2976 2977 2978 2979 2980 2981 2982 2983 2984 2985 2986 2987 2988 2989 2990 2991 2992 2993 2994 2995 2996 2997 2998 2999 3000 3001 3002 3003 3004 3005 3006 3007 3008 3009 3010 3011 3012 3013 3014 3015 3016 3017 3018 3019 3020 3021 3022 3023 3024 3025 3026 3027 3028 3029 3030 3031 3032 3033 3034 3035 3036 3037 3038 3039 3040 3041 3042 3043 3044 3045 3046 3047 3048 3049 3050 3051 3052 3053 3054 3055 3056 3057 3058 3059 3060 3061 3062 3063 3064 3065 3066 3067 3068 3069 3070 3071 3072 3073 3074 3075 3076 3077 3078 3079 3080 3081 3082 3083 3084 3085 3086 3087 3088 3089 3090 3091 3092 3093 3094 3095 3096 3097 3098 3099 3100 3101 3102 3103 3104 3105 3106 3107 3108 3109 3110 3111 3112 3113 3114 3115 3116 3117 3118 3119 3120 3121 3122 3123 3124 3125 3126 3127 3128 3129 3130 3131 3132 3133 3134 3135 3136 3137 3138 3139 3140 3141 3142 3143 3144 3145 3146 3147 3148 3149 3150 3151 3152 3153 3154 3155 3156 3157 3158 3159 3160 3161 3162 3163 3164 3165 3166 3167 3168 3169 3170 3171 3172 3173 3174 3175 3176 3177 3178 3179 3180 3181 3182 3183 3184 3185 3186 3187 3188 3189 3190 3191 3192 3193 3194 3195 3196 3197 3198 3199 3200 3201 3202 3203 3204 3205 3206 3207 3208 3209 3210 3211 3212 3213 3214 3215 3216 3217 3218 3219 3220 3221 3222 3223 3224 3225 3226 3227 3228 3229 3230 3231 3232 3233 3234 3235 3236 3237 3238 3239 3240 3241 3242 3243 3244 3245 3246 3247 3248 3249 3250 3251 3252 3253 3254 3255 3256 3257 3258 3259 3260 3261 3262 3263 3264 3265 3266 3267 3268 3269 3270 3271 3272 3273 3274 3275 3276 3277 3278 3279 3280 3281 3282 3283 3284 3285 3286 3287 3288 3289 3290 3291 3292 3293 3294 3295 3296 3297 3298 3299 3300 3301 3302 3303 3304 3305 3306 3307 3308 3309 3310 3311 3312 3313 3314 3315 3316 3317 3318 3319 3320 3321 3322 3323 3324 3325 3326 3327 3328 3329 3330 3331 3332 3333 3334 3335 3336 3337 3338 3339 3340 3341 3342 3343 3344 3345 3346 3347 3348 3349 3350 3351 3352 3353 3354 3355 3356 3357 3358 3359 3360 3361 3362 3363 3364 3365 3366 3367 3368 3369 3370 3371 3372 3373 3374 3375 3376 3377 3378 3379 3380 3381 3382 3383 3384 3385 3386 3387 3388 3389 3390 3391 3392 3393 3394 3395 3396 3397 3398 3399 3400 3401 3402 3403 3404 3405 3406 3407 3408 3409 3410 3411 3412 3413 3414 3415 3416 3417 3418 3419 3420 3421 3422 3423 3424 3425 3426 3427 3428 3429 3430 3431 3432 3433 3434 3435 3436 3437 3438 3439 3440 3441 3442 3443 3444 3445 3446 3447 3448 3449 3450 3451 3452 3453 3454 3455 3456 3457 3458 3459 3460 3461 3462 3463 3464 3465 3466 3467 3468 3469 3470 3471 3472 3473 3474 3475 3476 3477 3478 3479 3480 3481 3482 3483 3484 3485 3486 3487 3488 3489 3490 3491 3492 3493 3494 3495 3496 3497 3498 3499 3500 3501 3502 3503 3504 3505 3506 3507 3508 3509 3510 3511 3512 3513 3514 3515 3516 3517 3518 3519 3520 3521 3522 3523 3524 3525 3526 3527 3528 3529 3530 3531 3532 3533 3534 3535 3536 3537 3538 3539 3540 3541 3542 3543 3544 3545 3546 3547 3548 3549 3550 3551 3552 3553 3554 3555 3556 3557 3558 3559 3560 3561 3562 3563 3564 3565 3566 3567 3568 3569 3570 3571 3572 3573 3574 3575 3576 3577 3578 3579 3580 3581 3582 3583 3584 3585 3586 3587 3588 3589 3590 3591 3592 3593 3594 3595 3596 3597 3598 3599 3600 3601 3602 3603 3604 3605 3606 3607 3608 3609 3610 3611 3612 3613 3614 3615 3616 3617 3618 3619 3620 3621 3622 3623 3624 3625 3626 3627 3628 3629 3630 3631 3632 3633 3634 3635 3636 3637 3638 3639 3640 3641 3642 3643 3644 3645 3646 3647 3648 3649 3650 3651 3652 3653 3654 3655 3656 3657 3658 3659 3660 3661 3662 3663 3664 3665 3666 3667 3668 3669 3670 3671 3672 3673 3674 3675 3676 3677 3678 3679 3680 3681 3682 3683 3684 3685 3686 3687 3688 3689 3690 3691 3692 3693 3694 3695 3696 3697 3698 3699 3700 3701 3702 3703 3704 3705 3706 3707 3708 3709 3710 3711 3712 3713 3714 3715 3716 3717 3718 3719 3720 3721 3722 3723 3724 3725 3726 3727 3728 3729 3730 3731 3732 3733 3734 3735 3736 3737 3738 3739 3740 3741 3742 3743 3744 3745 3746 3747 3748 3749 3750 3751 3752 3753 3754 3755 3756 3757 3758 3759 3760 3761 3762 3763 3764 3765 3766 3767 3768 3769 3770 3771 3772 3773 3774 3775 3776 3777 3778 3779 3780 3781 3782 3783 3784 3785 3786 3787 3788 3789 3790 3791 3792 3793 3794 3795 3796 3797 3798 3799 3800 3801 3802 3803 3804 3805 3806 3807 3808 3809 3810 3811 3812 3813 3814 3815 3816 3817 3818 3819 3820 3821 3822 3823 3824 3825 3826 3827 3828 3829 3830 3831 3832 3833 3834 3835 3836 3837 3838 3839 3840 3841 3842 3843 3844 3845 3846 3847 3848 3849 3850 3851 3852 3853 3854 3855 3856 3857 3858 3859 3860 3861 3862 3863 3864 3865 3866 3867 3868 3869 3870 3871 3872 3873 3874 3875 3876 3877 3878 3879 3880 3881 3882 3883 3884 3885 3886 3887 3888 3889 3890 3891 3892 3893 3894 3895 3896 3897 3898 3899 3900 3901 3902 3903 3904 3905 3906 3907 3908 3909 3910 3911 3912 3913 3914 3915 3916 3917 3918 3919 3920 3921 3922 3923 3924 3925 3926 3927 3928 3929 3930 3931 3932 3933 3934 3935 3936 3937 3938 3939 3940 3941 3942 3943 3944 3945 3946 3947 3948 3949 3950 3951 3952 3953 3954 3955 3956 3957 3958 3959 3960 3961 3962 3963 3964 3965 3966 3967 3968 3969 3970 3971 3972 3973 3974 3975 3976 3977 3978 3979 3980 3981 3982 3983 3984 3985 3986 3987 3988 3989 3990 3991 3992 3993 3994 3995 3996 3997 3998 3999 4000 4001 4002 4003 4004 4005 4006 4007 4008 4009 4010 4011 4012 4013 4014 4015 4016 4017 4018 4019 4020 4021 4022 4023 4024 4025 4026 4027 4028 4029 4030 4031 4032 4033 4034 4035 4036 4037 4038 4039 4040 4041 4042 4043 4044 4045 4046 4047 4048 4049 4050 4051 4052 4053 4054 4055 4056 4057 4058 4059 4060 4061 4062 4063 4064 4065 4066 4067 4068 4069 4070 4071 4072 4073 4074 4075 4076 4077 4078 4079 4080 4081 4082 4083 4084 4085 4086 4087 4088 4089 4090 4091 4092 4093 4094 4095 4096 4097 4098 4099 4100 4101 4102 4103 4104 4105 4106 4107 4108 4109 4110 4111 4112 4113 4114 4115 4116 4117 4118 4119 4120 4121 4122 4123 4124 4125 4126 4127 4128 4129 4130 4131 4132 4133 4134 4135 4136 4137 4138 4139 4140 4141 4142 4143 4144 4145 4146 4147 4148 4149 4150 4151 4152 4153 4154 4155 4156 4157 4158 4159 4160 4161 4162 4163 4164 4165 4166 4167 4168 4169 4170 4171 4172 4173 4174 4175 4176 4177 4178 4179 4180 4181 4182 4183 4184 4185 4186 4187 4188 4189 4190 4191 4192 4193 4194 4195 4196 4197 4198 4199 4200 4201 4202 4203 4204 4205 4206 4207 4208 4209 4210 4211 4212 4213 4214 4215 4216 4217 4218 4219 4220 4221 4222 4223 4224 4225 4226 4227 4228 4229 4230 4231 4232 4233 4234 4235 4236 4237 4238 4239 4240 4241 4242 4243 4244 4245 4246 4247 4248 4249 4250 4251 4252 4253 4254 4255 4256 4257 4258 4259 4260 4261 4262 4263 4264 4265 4266 4267 4268 4269 4270 4271 4272 4273 4274 4275 4276 4277 4278 4279 4280 4281 4282 4283 4284 4285 4286 4287 4288 4289 4290 4291 4292 4293 4294 4295 4296 4297 4298 4299 4300 4301 4302 4303 4304 4305 4306 4307 4308 4309 4310 4311 4312 4313 4314 4315 4316 4317 4318 4319 4320 4321 4322 4323 4324 4325 4326 4327 4328 4329 4330 4331 4332 4333 4334 4335 4336 4337 4338 4339 4340 4341 4342 4343 4344 4345 4346 4347 4348 4349 4350 4351 4352 4353 4354 4355 4356 4357 4358 4359 4360 4361 4362 4363 4364 4365 4366 4367 4368 4369 4370 4371 4372 4373 4374 4375 4376 4377 4378 4379 4380 4381 4382 4383 4384 4385 4386 4387 4388 4389 4390 4391 4392 4393 4394 4395 4396 4397 4398 4399 4400 4401 4402 4403 4404 4405 4406 4407 4408 4409 4410 4411 4412 4413 4414 4415 4416 4417 4418 4419 4420															

## 4. 엑셀로 데이터 정리

## 1. 데이터 수집과 계절 데이터 추가



The image shows a screenshot of a Microsoft Excel spreadsheet. The spreadsheet is very wide, with many columns visible. The rows are organized by year, with 2014 at the top and 2019 at the bottom. Each year has 12 rows, one for each month. The data appears to be numerical values, possibly representing sales or other metrics. The spreadsheet is titled 'M11' in the top left corner. The status bar at the bottom indicates that the data is from 2014 to 2019.

2014년 ~ 2019년 데이터의 양 多

월별데이터

월별 데이터는 Factor가 너무 많아서

계절 데이터로 취합

계절별데이터



## 2. 데이터 확인

- 수집된 데이터가 제대로 읽혔는지 확인

	계절	타율	득점	안타	X2 타	X3 타	홈런	루타	타점	도루	도실	볼넷	사구	고4	삼진 허용
1	봄	0.297	61	82	22	0	8	128	58	6	3	42	6	0	66
2	봄	0.282	32	82	8	0	7	111	31	8	2	30	3	2	73
3	봄	0.268	49	75	10	2	14	131	46	5	1	39	6	1	67
4	봄	0.258	40	70	6	1	15	123	40	2	3	25	6	2	57
5	봄	0.252	41	68	18	1	4	100	40	6	1	26	2	0	60
6	봄	0.246	36	62	8	1	6	90	36	3	1	37	3	0	48
7	봄	0.237	30	66	14	0	4	92	28	3	3	32	3	1	64
8	봄	0.234	34	62	15	1	4	91	33	3	3	35	4	0	62
9	봄	0.217	35	56	13	1	8	95	34	8	2	23	5	0	61
10	봄	0.210	28	57	13	0	5	85	25	6	1	31	4	0	63
11	봄	0.309	153	267	53	7	17	385	146	22	7	102	9	4	163
12	봄	0.298	124	235	56	2	23	364	115	10	8	62	17	1	156
13	봄	0.284	144	230	46	9	20	354	133	15	8	109	14	4	150
14	봄	0.277	112	211	34	4	15	298	104	19	8	81	10	3	137
15	봄	0.268	99	204	27	6	14	285	89	4	4	75	12	5	166
16	봄	0.267	103	208	43	5	14	303	95	13	6	79	8	4	167
17	봄	0.262	96	202	45	5	13	296	92	20	4	77	9	2	168
18	봄	0.252	88	176	35	1	14	255	85	15	4	69	9	4	174
19	봄	0.246	87	200	27	2	16	279	80	14	6	68	8	2	173
20	봄	0.245	95	196	30	3	22	298	91	15	3	87	14	4	177
21	봄	0.296	131	274	42	7	22	396	127	17	8	81	15	3	161
22	봄	0.283	128	252	43	3	16	349	122	14	8	91	13	5	176

- Data Class 확인

```
> str(data)
'data.frame': 455 obs. of 47 variables:
 $ 계절 : Factor w/ 3 levels "가을","봄","여름": 2 2 2 2 2 2 2 2 2 2 ...
 $ 타율 : num 0.297 0.282 0.268 0.258 0.252 0.246 0.237 0.234 0.217 0.21 ...
 $ 득점 : int 61 32 49 40 41 36 30 34 35 28 ...
 $ 안타 : int 82 82 75 70 68 62 66 62 56 57 ...
 $ X2타 : int 22 8 10 6 18 8 14 15 13 13 ...
 $ X3타 : int 0 0 2 1 1 1 0 1 1 0 ...
 $ 홈런 : int 8 7 14 15 4 6 4 4 8 5 ...
 $ 루타 : int 128 111 131 123 100 90 92 91 95 85 ...
 $ 타점 : int 58 31 46 40 40 36 28 33 34 25 ...
 $ 도루 : int 6 8 5 2 6 3 3 3 8 6 ...
 $ 도실 : int 3 2 1 3 1 1 3 3 2 1 ...
 $ 볼넷 : int 42 30 39 25 26 37 32 35 23 31 ...
 $ 사구 : int 6 3 6 6 2 3 3 4 5 4
```

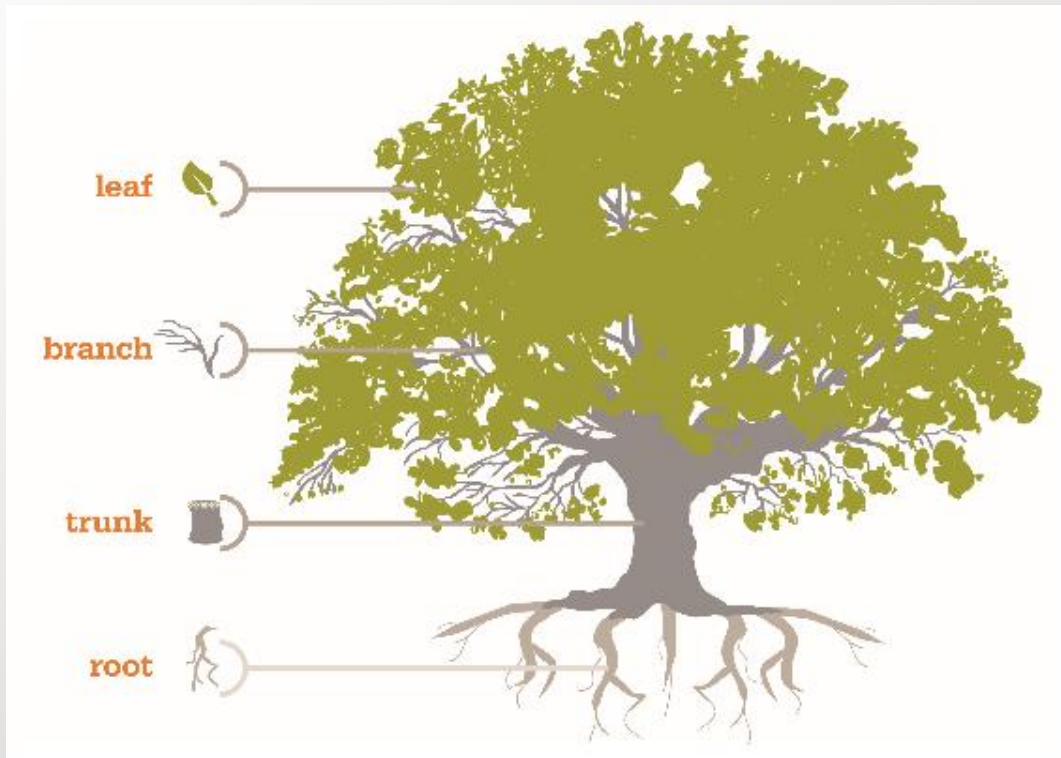
Nominal Data → Factor

```
$ 사구어부 : int 3 1 2 3 6 4 9 3 6 3 ...
$ 삼진 : int 73 67 57 53 54 59 63 69 61 65 ...
$ 투구 : int 1309 1239 1249 1169 1142 1161 1282 1215 1217 1083 ...
$ WHIP : num 1.68 1.69 1.4 1.52 1.35 1.32 1.56 1.23 1.32 1 ...
$ 허용타율 : num 0.277 0.299 0.251 0.254 0.256 0.242 0.261 0.211 0.246 0.205 ...
$ 출루허용율 : num 0.365 0.373 0.327 0.362 0.345 0.322 0.37 0.301 0.325 0.266 ...
$ 허용OPS : num 0.833 0.85 0.741 0.77 0.682 0.735 0.747 0.594 0.662 0.587 ...
$ 우승여부 : Factor w/ 2 levels "비우승","우승": 1 1 1 1 1 2 1 1 1 1 ...
```

Y Value → Factor

### 3. 분석 모델 선정

- Analytics Model : Decision Tree



의사결정나무 또는 나무 모형은 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석 방법

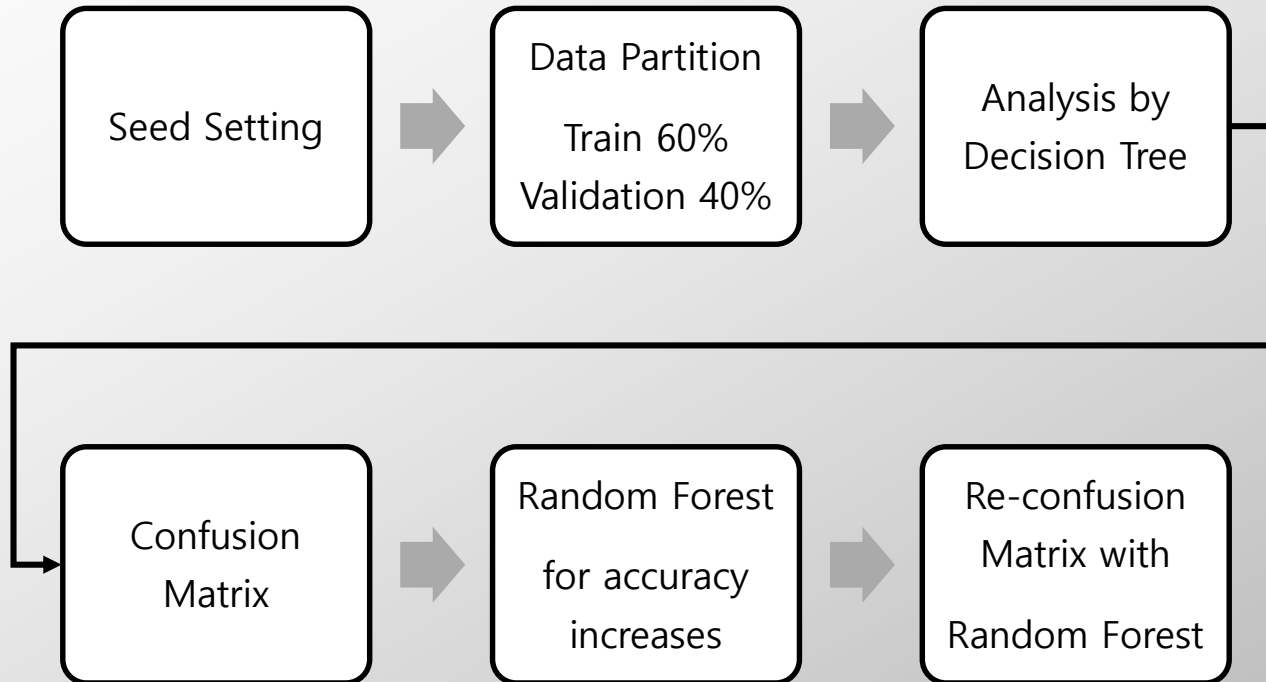
- The reason why we select the decision tree as model

Decision Tree Characteristics	
Strength	
해석의 용이성	나무 구조로 표현되어 사용자가 모형을 이해하기 쉬움
교호작용효과의 해석	두 개 이상의 변수가 결합하여 목표 변수에 어떻게 영향을 주는지 알기 쉬움
비모수적 모형	선형성, 정규성, 등분산성 등의 가정이 필요하지 않음

야구 데이터를 통해 우승 규칙을 찾기 위해 **변수의 별 다른 가정 없이** 다양한 변수를 통해 쉽게 규칙을 찾고 생성된 규칙들을 **가시적으로 해석**하기에 Decision Tree가 용이

### 3. 데이터 분석 과정

- Framework



- 분석 과정 설명

1. 원활한 분석을 위해 Seed를 고정 후
2. 데이터에서 60%으로 학습용 40%을 검증용으로 분할
3. 'rpart'(R library) 를 이용해 나무 생성
4. 'caret'(R library) 를 이용해 Accuracy 확인
5. 모델 개선을 위해 Random Forest 생성
6. 개선된 Accuracy 확인

## 4. Decision Tree

### # seed설정

- 예측결과 재현성을 위해 seed의 값을 고정

### # train/validation 데이터분할

- data의 수가 많지 않으므로 6:4의 비율로 데이터 분할

### # 데이터 파티션 확인

Nrow(df) #455개

Nrow(train.df) #272개

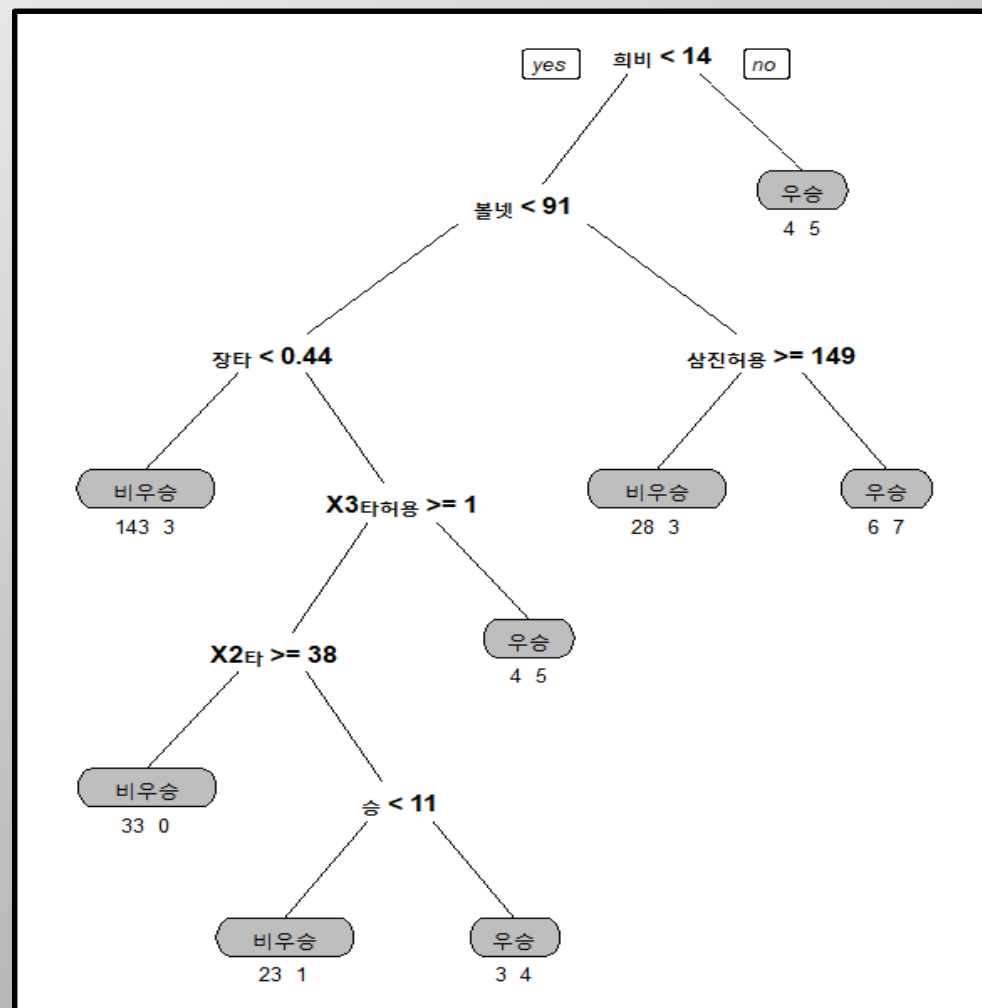
Nrow(valid.df) #183개

### # 의사결정나무

- Train 데이터의 10%정도인 20으로 minbucket 설정  
→ (20개의 오류가 발생할 시 pruning)
- 우승에 영향력 있는 주원인만 고려하기 위해 maxdepth=7 설정

# Main Variables → 희비, 볼넷, 장타, 삼진허용, 3루타, 2루타, 승

### • Tree Plot





## 5. Confusion Matrix

- Confusion Matrix

	Reference	
Prediction	비우승	우승
비우승	227	7
우승	17	21

*Train data confusion matrix*

Accuracy : 0.9118  
 95% CI : (0.8716, 0.9426)  
 No Information Rate : 0.8971  
 P-Value [Acc > NIR] : 0.24677  
  
 Kappa : 0.5875

McNemar's Test P-Value : 0.06619

	Reference	
Prediction	비우승	우승
비우승	138	13
우승	26	6

Accuracy : 0.7869  
 95% CI : (0.7204, 0.8438)

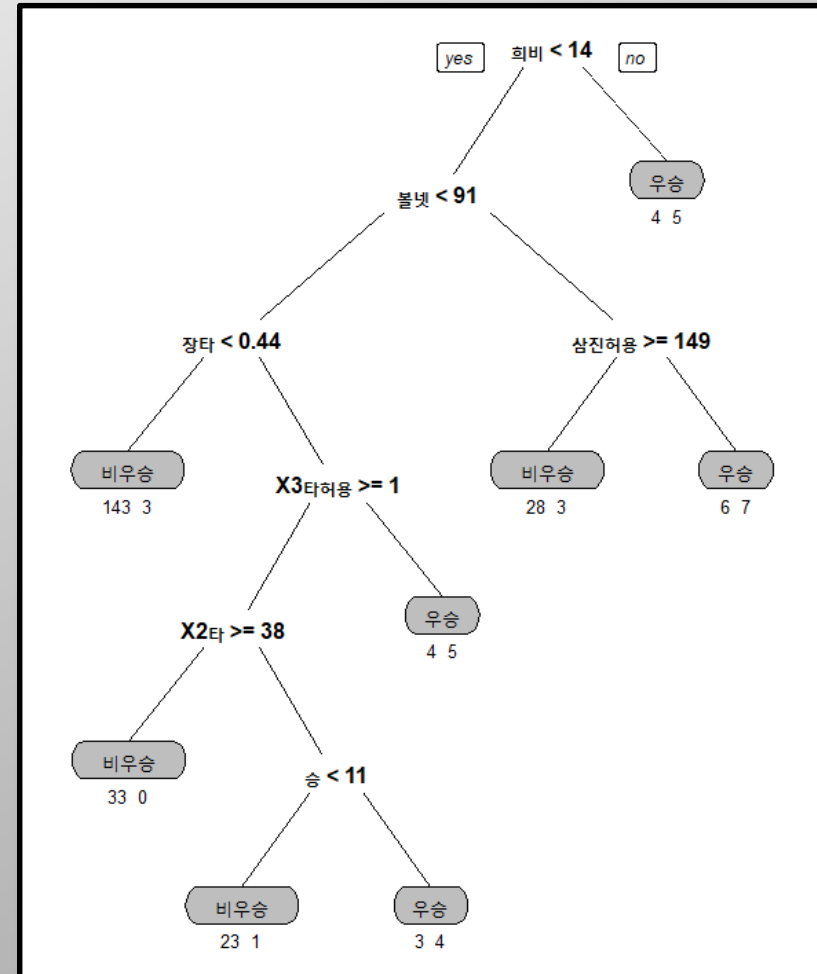
No Information Rate : 0.8962  
 P-Value [Acc > NIR] : 1.00000

Kappa : 0.1207

*Validation data confusion matrix*

McNemar's Test P-Value : 0.05466

- Tree Plot



## 6. Random Forest

- Parameter Set

# **mtry=20** (복원 추출 선택변수 개수)

- 수집한 기존 데이터의 변수가 많으므로 총 변수의 수의 약 50%인 20개로 Bagging

# **ntree=1000** (ntree는 의사결정 나무의 개수를 의미)

# **nodesize=7** (나무의 깊이를 설정하는 인자)

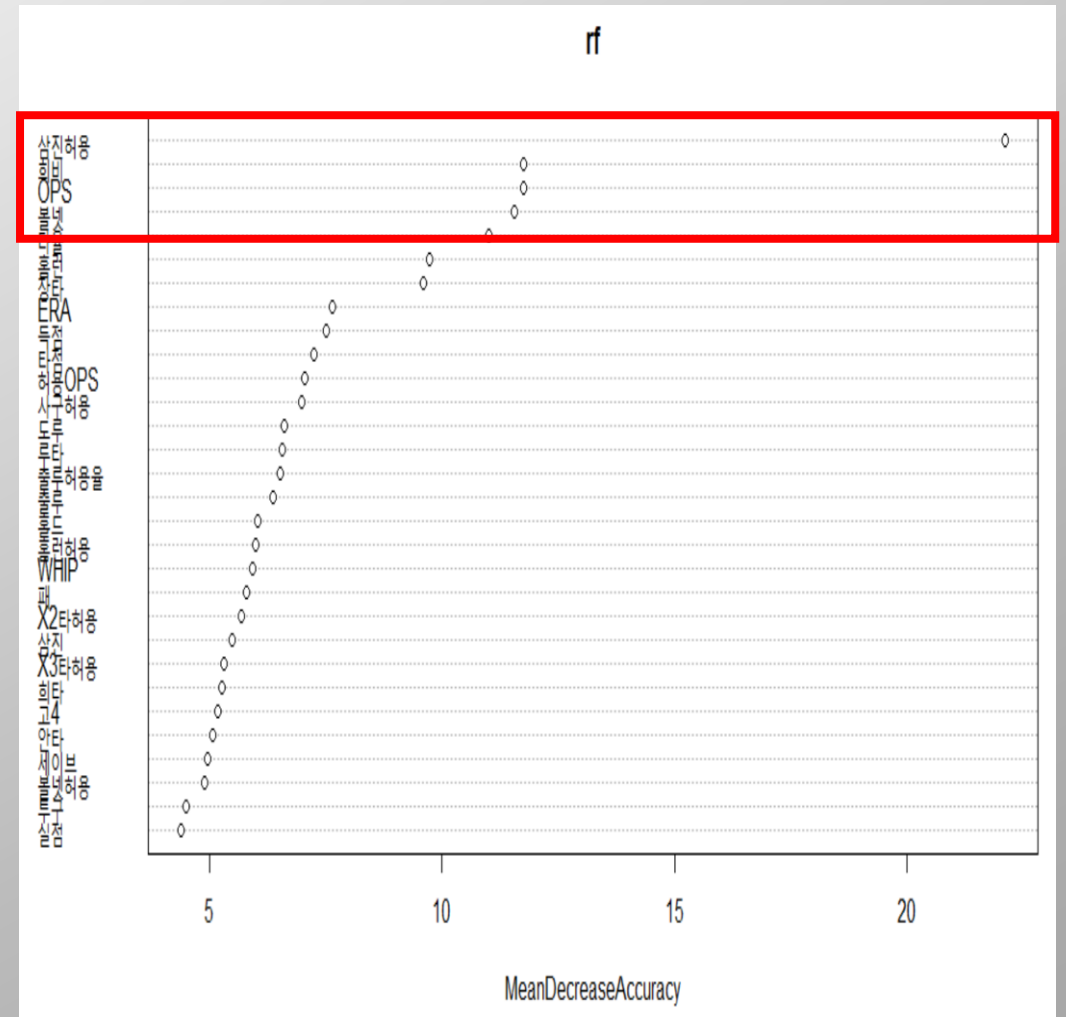
- 의사결정나무의 maxdepth와 동일하게 7로 설정

- Result

# Main Variables → 삼진허용, 희비, OPS, 볼넷, 타율

\* 기준 : Mean Decrease Accuracy(MDA) : 특정한 변수의 값을 다른 값으로 대체하였을 때 정확도가 감소하는 정도

- Mean Decrease Accuracy plot



## 7. Improved Accuracy

- Mean Decrease Accuracy-Mean Decrease Gini

	비우승	우승	MeanDecreaseAccuracy	MeanDecreaseGini
삼진허용	21.8538565	-2.43019258	21.3575852	3.16248659
OPS	13.0713241	-3.82868047	12.6873022	1.23835847
희비	8.7854331	8.58345516	11.4149500	4.20092431
장타	10.5366556	-1.71429935	10.4048742	1.21319640
홈런	8.9867485	-1.13463556	8.8619184	1.89742616
득점	8.5813111	-2.42747734	8.6026486	1.18353245
ERA	8.8513501	-4.56692813	8.3679811	0.8552305
볼넷	7.7864470	2.93452975	8.3275197	1.99646113
타점	8.0865279	-1.90052333	8.0462649	0.86169212
사구허용	8.4663976	-5.84355947	7.7631591	1.35289042

- Mean Decrease Gini(MDG) : 특정 변수가 모델에 적용 될 때 불순도를 제거 능력 수치
- Main value → 삼진허용, OPS, 희비, 장타, 홈런

- Random forest Accuracy

	Reference	
Prediction	비우승	우승
비우승	138	13
우승	26	6

Accuracy : 0.7869  
95% CI : (0.7174, 0.8564)

	Reference	
Prediction	비우승	우승
비우승	160	19
우승	4	0

Accuracy : 0.8743  
95% CI : (0.8174, 0.9186)

No Information Rate : 0.8962  
P-Value [Acc > NIR] : 0.861400

Kappa : -0.0375

Mcnemar's Test P-Value : 0.003509

**Improved Accuracy : 0.0874**

## 8. 결과 및 제시사항

- Analysis Result

- Random Forest Accuracy : 0.8743  
→ 충분히 신뢰 가능한 모델 생성
- Main Variables in Decision Tree  
→ **희비**, **볼넷**, **장타**, **삼진허용**, 3루타, 2루타, 승
- Main Variables in Random Forest  
→ **삼진허용**, **희비**, OPS, 타율, **장타**, 홈런, **볼넷**
- DT와 Rf의 공통 Main Variables  
→ **희비**, **삼진허용**, **장타**, **볼넷**



- Suggested Strategy for WIN

우승 전략 1. **희비** (희생플라이)

→ 스코어링 포지션(특히 3루 주자가 있는 상황)에서는  
**안타 대신 희생플라이와 같이 팀득점을 가져오는 '팀배팅'을 해야 함**

우승 전략 2. 삼진허용 (타자가 삼진을 당하는 경우)

→ 최대한 투수가 던진 공에 contact(공에 맞추어 보려는 타격)을  
많이 하여 **삼진을 당하지 않으려 노력**해야 함

우승 전략 3. **장타** (2루타 이상)

→ 장타는 주자가 있는 경우에는 주자를 불러들일 수 있고 주자가  
없는 상황에서도 스코어링 포지션을 만들 수 있으므로 **장타능력이  
있는 선수를 배치** 혹은 **영입할 필요**

우승 전략 4. 볼넷

→ 적극적인 스윙보다는 **공을 많이 보고 신중하게 타격**해야 함



Q&A

Q&A

경청 해주셔서 감사합니다.