Final report

網路成癮程度預測

Group5(廖芷萱、詹雅鈞、李姿慧、謝沛恩、李敏榕)

2024-12-29

Table of contents

1 資料來源

本研究所使用之資料來源為 Kaggle 競賽提供的 Healthy Brain Network (HBN) 資料集。該資料集為一臨床樣本，包含 3960 名年齡介於 5 至 22 歲的青少年，他們均接受過臨床及研究篩檢。資料集中包含以下兩類元素被納入分析範疇：(1) 體能活動資料，包括腕戴式加速度計記錄、體能評估及問卷調查數據；(2) 網路使用行為資料。

2 目標與動機

本研究旨在基於兒童和青少年的體能活動、身體測量、心理健康及網路行為等特徵，建立成癮嚴重程度 (sii) 的模型來預測參與者的網路成癮嚴重程度，為家庭及教育機構提供有針對性的建議，幫助減少過度使用網路的負面影響。

3 敘述統計

train

82 Variables **3960 Observations**

---

id

n	missing	distinct
3960	0	3960

```
lowest : 00008ff9 000fd460 00105258 00115b9f 0016bb22, highest: ff8a2de4 ffa9794a ffcd4dbd ffed1dd5 ffef538e
```

Basic_Demos.Enroll_Season

n	missing	distinct
3960	0	4

Value	Fall	Spring	Summer	Winter
Frequency	866	1127	970	997
Proportion	0.219	0.285	0.245	0.252

Basic_Demos.Age

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
3960	0	18	0.992	10.43	10	3.983	6	6	8	10	13	16	17

Value	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Frequency	112	369	436	490	467	420	334	291	236	200	178	151	114	74
Proportion	0.028	0.093	0.110	0.124	0.118	0.106	0.084	0.073	0.060	0.051	0.045	0.038	0.029	0.019

Value	19	20	21	22
Frequency	27	24	29	8
Proportion	0.007	0.006	0.007	0.002

```
For the frequency table, variable is rounded to the nearest 0
```

Basic_Demos.Sex

n	missing	distinct	Info	Sum	Mean
3960	0	2	0.701	1476	0.3727

CGAS.Season

n	missing	distinct
2555	1405	4

Value	Fall	Spring	Summer	Winter
Frequency	635	697	656	567
Proportion	0.249	0.273	0.257	0.222

CGAS.CGAS_Score

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2421	1539	59	0.995	65.45	65	14.11	45	50	59	65	75	80	85

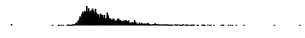
lowest : 25 30 31 33 35, highest: 91 92 93 95 999

Physical.Season

n	missing	distinct
3310	650	4

Value	Fall	Spring	Summer	Winter
Frequency	786	929	791	804
Proportion	0.237	0.281	0.239	0.243

Physical.BMI



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
3022	938	2658	1	19.33	18.64	5.205	14.13	14.70	15.87	17.94	21.57	25.69	29.31

lowest : 0 8.52244 9.69377 9.95917 10.2817, highest: 45.306 46.1029 47.6038 53.9184 59.132

Physical.Height

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
3027	933	306	1	55.95	55.75	8.542	45.00	46.50	50.00	55.00	62.00	66.04	68.75

lowest : 33 36 37.5 39 39.5, highest: 76 77 77.5 78 78.5

Physical.Weight

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
3076	884	783	1	89.04	84.5	47.43	42.0	47.1	57.2	77.0	113.8	148.5	173.4

lowest : 0 31.8 32.8 33 33.2 , highest: 298.8 299.6 302.4 306.4 315

Physical.Waist_Circumference

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
898	3062	44	0.996	27.28	26.5	5.956	21	21	23	26	30	35	38

lowest : 18 19 20 21 21.5, highest: 45.5 46 48 49 50

Physical.Diastolic_BP

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2954	1006	102	0.999	69.65	68.5	14.24	52	56	61	68	76	86	94

lowest : 0 11 14 22 28, highest: 135 136 145 146 179

Physical.HeartRate

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2967	993	88	0.999	81.6	81.5	15.32	60.3	64.6	72.0	81.0	90.5	99.0	105.0

```
lowest :  27  33  36  45  46, highest: 130 132 133 134 138
```

Physical.Systolic_BP

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2954	1006	129	0.999	117	115.5	18.04	95	100	107	114	125	138	149

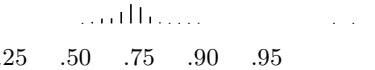
```
lowest :   0  49  57  60  62, highest: 193 194 197 198 203
```

Fitness_Endurance.Season

n	missing	distinct
1308	2652	4

Value	Fall	Spring	Summer	Winter
Frequency	332	385	253	338
Proportion	0.254	0.294	0.193	0.258

Fitness_Endurance.Max_Stage



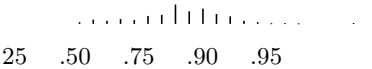
n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
743	3217	15	0.956	4.989	5	1.89	2	3	4	5	6	7	7

Value	0	1	2	3	4	5	6	7	8	9	10	11	12	26
Frequency	1	17	53	57	125	213	179	74	12	3	2	1	4	1
Proportion	0.001	0.023	0.071	0.077	0.168	0.287	0.241	0.100	0.016	0.004	0.003	0.001	0.005	0.001

Value	28
Frequency	1
Proportion	0.001

```
For the frequency table, variable is rounded to the nearest 0
```

Fitness_Endurance.Time_Mins



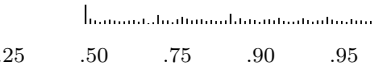
n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
740	3220	18	0.987	7.37	7.5	3.494	1	3	6	7	9	11	12

Value	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	12	30	21	34	27	55	63	135	88	106	69	52	22	12
Proportion	0.016	0.041	0.028	0.046	0.036	0.074	0.085	0.182	0.119	0.143	0.093	0.070	0.030	0.016

Value	14	15	16	20
Frequency	3	4	2	5
Proportion	0.004	0.005	0.003	0.007

```
For the frequency table, variable is rounded to the nearest 0
```

Fitness_Endurance.Time_Sec



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
740	3220	60	0.999	27.58	27.5	20.44	0.00	2.00	12.75	28.00	43.00	52.00	56.00

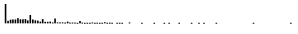
```
lowest :  0  1  2  3  4, highest: 55 56 57 58 59
```

FGC.Season

n	missing	distinct
3346	614	4

Value	Fall	Spring	Summer	Winter
Frequency	763	993	844	746
Proportion	0.228	0.297	0.252	0.223

FGC.FGC_CU



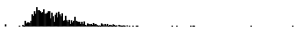
n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2322	1638	59	0.992	11.26	9.5	11.82	0.00	0.00	3.00	9.00	15.75	26.00	34.00

```
lowest :   0   1   2   3   4, highest:  78  80  85 100 115
```

FGC.FGC_CU_Zone

n	missing	distinct	Info	Sum	Mean
2282	1678	2	0.748	1087	0.4763

FGC.FGC_GSND



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1074	2886	341	1	22.42	21.05	10.84	11.10	12.50	15.10	20.05	26.60	35.50	42.70

```
lowest : 0     6.1   7.5   7.6   7.8  , highest: 80.4  81.2  81.8  106.4 124
```

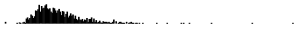
FGC.FGC_GSND_Zone

n	missing	distinct	Info	Mean	pMedian	Gmd
1062	2898	3	0.763	1.83	2	0.6163

```
Value          1     2     3
Frequency    305   633   124
Proportion 0.287 0.596 0.117
```

```
For the frequency table, variable is rounded to the nearest 0
```

FGC.FGC_GSD



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1074	2886	350	1	23.52	22.15	11.25	11.10	13.10	16.20	21.20	28.17	37.20	43.87

```
lowest : 0     5.1   6.2   6.3   6.5  , highest: 76.8  79.2  88.8  106   123.8
```

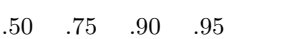
FGC.FGC_GSD_Zone

n	missing	distinct	Info	Mean	pMedian	Gmd
1063	2897	3	0.749	1.904	2	0.6117

```
Value          1     2     3
Frequency    255   655   153
Proportion 0.240 0.616 0.144
```

```
For the frequency table, variable is rounded to the nearest 0
```

FGC.FGC_PU



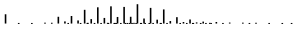
n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2310	1650	44	0.955	5.58	4.5	7.179	0	0	0	3	9	15	20

```
lowest :  0  1  2  3  4, highest: 41 47 49 50 51
```

FGC.FGC_PU_Zone

n	missing	distinct	Info	Sum	Mean
2271	1689	2	0.664	750	0.3303

FGC.FGC_SRL



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2305	1655	71	0.996	8.695	8.75	3.777	3.0	5.0	7.0	9.0	11.0	12.5	14.0

```
lowest : 0    1    2    3    3.5 , highest: 18.5 19   20   21   21.7
```

FGC.FGC_SRL_Zone

n	missing	distinct	Info	Sum	Mean
2267	1693	2	0.708	1403	0.6189

FGC.FGC_SRR

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2307	1653	73	0.996	8.806	9	3.784	4	5	7	9	11	13	14

```
lowest : 0    1    2    3    3.5 , highest: 18.5 19   19.5 20   21
```

FGC.FGC_SRR_Zone

n	missing	distinct	Info	Sum	Mean
2269	1691	2	0.707	1407	0.6201

FGC.FGC_TL

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2324	1636	43	0.986	9.253	9.5	3.325	4	5	7	10	12	12	13

```
lowest : 0   1   1.5 2   2.5, highest: 18  19  20  21  22
```

FGC.FGC_TL_Zone

n	missing	distinct	Info	Sum	Mean
2285	1675	2	0.505	1795	0.7856

BIA.Season

n	missing	distinct
2145	1815	4

```
Value        Fall Spring Summer Winter
Frequency     567    513    669    396
Proportion  0.264  0.239  0.312  0.185
```

BIA.BIA_Activity_Level_num

n	missing	distinct	Info	Mean	pMedian	Gmd
1991	1969	5	0.918	2.651	2.5	1.125

```
Value          1     2     3     4     5
Frequency    266   637   698   305    85
Proportion 0.134 0.320 0.351 0.153 0.043
```

For the frequency table, variable is rounded to the nearest 0

BIA.BIA_BMC

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1986	1	6.72	4.204	7.004	2.054	2.368	2.967	3.923	5.461	7.374	8.964

```
lowest : -7.78961 -6.40154 -5.02683 -4.86597 -4.16832, highest: 22.4353  22.9845  29.463   401.002  4115.36
```

BIA.BIA_BMI

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1803	1	19.37	18.65	5.145	14.16	14.71	15.91	17.97	21.46	25.57	29.16

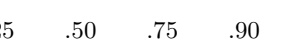
```
lowest : 0.0482667 10.6766   11.434    11.4685   11.6771
highest: 44.8404   45.311    46.1079   48.3754   53.9243
```

BIA.BIA_BMR

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1965	1	1237	1154	366.8	908.1	937.1	1004.7
.50	.75	.90	.95						
1115.4	1310.4	1541.3	1684.3						

lowest : 813.397 825.733 830.308 849.805 854.838, highest: 3600.18 3806.7 3987.68 11540.8 83152.2

BIA.BIA_DEE



n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1980	1	2065	1920	740.2	1332	1429	1606	1864	2218	2705	3083

lowest : 1073.45 1079.4 1111.29 1123.68 1124.63, highest: 6467.97 6779.05 7994.08 17311.2 124728

BIA.BIA_ECW

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1986	1	20.83	17.87	15.05	6.896	8.295	11.110
.50	.75	.90	.95						
15.928	25.162	33.384	38.585						

lowest : 1.78945 2.12534 2.20229 2.60779 2.71056, highest: 104.347 115.069 115.285 350.849 3233

BIA.BIA_FFM

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1985	1	74.02	65.17	39.07	38.98	42.08	49.28
.50	.75	.90	.95						
61.07	81.83	106.43	121.67						

lowest : 28.9004 30.2144 30.7017 32.7784 33.3145, highest: 325.73 347.727 367.004 1171.51 8799.08

BIA.BIA_FFMI

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1945	1	15.03	14.4	2.574	12.77	13.00	13.41	14.09	15.43	17.35	18.87

lowest : 7.86485 11.3229 11.4432 11.8963 12.1529, highest: 58.4569 61.4583 65.5384 82.4902 217.771

BIA.BIA_FMI

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1985	1	4.336	4.116	4.072	0.8423	1.4141	2.3069
.50	.75	.90	.95						
3.6986	5.9877	9.1522	11.0375						

lowest : -194.163 -66.378 -45.8722 -45.1576 -44.5091, highest: 19.901 20.4853 25.1517 27.6857 28.2515

BIA.BIA_Fat

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1988	1	16.86	19.27	31.73	2.981	5.041	8.602
.50	.75	.90	.95						
16.175	30.273	48.072	64.565						

lowest : -8745.08 -1044.51 -217.522 -198.528 -195.933, highest: 119.986 121.166 126.01 129.226 153.82

BIA.BIA_Frame_num

n	missing	distinct	Info	Mean	pMedian	Gmd
1991	1969	3	0.832	1.745	1.5	0.7126

Value	1	2	3
Frequency	779	940	272
Proportion	0.391	0.472	0.137

For the frequency table, variable is rounded to the nearest 0

BIA.BIA_ICW

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1980	1	33.17	29.97	13.76	20.59	21.75	24.46	28.86	35.48	46.23	52.48

lowest : 14.489 17.555 17.845 17.9973 18.0323, highest: 125.694 138.491 152.738 428.264 2457.91

BIA.BIA_LDM

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1984	1	20.02	17.43	11.53	9.384	10.429	12.983
.50	.75	.90	.95						
16.439	22.168	27.963	31.911						

lowest : 4.63581 5.47079 5.8564 5.88392 6.12639, highest: 94.1672 95.6882 98.9813 392.4 3108.17

BIA.BIA_LST

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25
1991	1969	1986	1	67.3	60.86	34.01	35.24	38.14	45.20
.50	.75	.90	.95						
57.00	77.11	100.82	115.44						

lowest : 23.6201 23.9473 24.7088 24.9603 25.2618, highest: 315.215 334.766 355.058 770.511 4683.71

BIA.BIA_SMM

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1984	1	34.39	29.56	20.68	15.89	17.38	21.14	27.42	38.18	52.92	60.90

lowest : 4.65573 11.3825 11.7991 12.1372 12.5903, highest: 215.413 223.449 254.611 823.028 3607.69

BIA.BIA_TBW

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1991	1969	1989	1	54	47.88	28.29	28.42	30.70	35.89	44.99	60.27	79.14	90.33

lowest : 20.5892 21.6173 21.7241 21.7776 22.0463, highest: 230.042 253.56 268.022 779.114 5690.91

PAQ_A.Season

n	missing	distinct
475	3485	4

Value	Fall	Spring	Summer	Winter
Frequency	98	123	117	137
Proportion	0.206	0.259	0.246	0.288

PAQ_A.PAQ_A_Total

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
475	3485	256	1	2.179	2.135	0.9586	1.077	1.140	1.490	2.010	2.780	3.398	3.670

lowest : 0.66 0.99 1 1.01 1.02, highest: 4.42 4.52 4.54 4.58 4.71

PAQ_C.Season

n	missing	distinct
1721	2239	4

Value	Fall	Spring	Summer	Winter
Frequency	354	506	391	470
Proportion	0.206	0.294	0.227	0.273

PAQ_C.PAQ_C_Total

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
1721	2239	404	1	2.59	2.575	0.8935	1.35	1.58	2.02	2.54	3.16	3.68	3.96

lowest : 0.58 0.77 0.88 0.96 0.99, highest: 4.63 4.66 4.74 4.75 4.79

PCIAT.Season

n	missing	distinct
2736	1224	4

Value	Fall	Spring	Summer	Winter
Frequency	667	762	659	648
Proportion	0.244	0.279	0.241	0.237

PCIAT.PCIAT_01

n	missing	distinct	Info	Mean	pMedian	Gmd
2733	1227	6	0.969	2.371	2.5	1.901

Value	0	1	2	3	4	5
Frequency	543	339	599	418	482	352
Proportion	0.199	0.124	0.219	0.153	0.176	0.129

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_02

n	missing	distinct	Info	Mean	pMedian	Gmd
2734	1226	6	0.965	2.178	2	1.923

Value	0	1	2	3	4	5
Frequency	682	337	572	408	426	309
Proportion	0.249	0.123	0.209	0.149	0.156	0.113

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_03

n	missing	distinct	Info	Mean	pMedian	Gmd
2731	1229	6	0.965	2.4	2.5	1.799

Value	0	1	2	3	4	5
Frequency	494	288	652	503	517	277
Proportion	0.181	0.105	0.239	0.184	0.189	0.101

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_04

n	missing	distinct	Info	Mean	pMedian	Gmd
2731	1229	6	0.825	0.8393	0.5	1.144

Value	0	1	2	3	4	5
Frequency	1473	699	293	106	111	49
Proportion	0.539	0.256	0.107	0.039	0.041	0.018

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_05

n	missing	distinct	Info	Mean	pMedian	Gmd
2729	1231	6	0.969	2.298	2.5	1.936

Value	0	1	2	3	4	5
Frequency	586	373	575	421	386	388
Proportion	0.215	0.137	0.211	0.154	0.141	0.142

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_06

n	missing	distinct	Info	Mean	pMedian	Gmd
2732	1228	6	0.89	1.064	1	1.281

Value	0	1	2	3	4	5
Frequency	1152	873	362	141	134	70
Proportion	0.422	0.320	0.133	0.052	0.049	0.026

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_07

n	missing	distinct	Info	Mean	pMedian	Gmd
2729	1231	6	0.704	0.5863	0.5	0.8957

Value	0	1	2	3	4	5
Frequency	1799	576	186	57	74	37
Proportion	0.659	0.211	0.068	0.021	0.027	0.014

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_08

n	missing	distinct	Info	Mean	pMedian	Gmd
2730	1230	6	0.918	1.247	1	1.417

Value	0	1	2	3	4	5
Frequency	1036	766	466	213	179	70
Proportion	0.379	0.281	0.171	0.078	0.066	0.026

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_09

n	missing	distinct	Info	Mean	pMedian	Gmd
2730	1230	6	0.891	1.063	1	1.266

Value	0	1	2	3	4	5
Frequency	1134	880	402	123	107	84
Proportion	0.415	0.322	0.147	0.045	0.039	0.031

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_10

n	missing	distinct	Info	Mean	pMedian	Gmd
2733	1227	6	0.926	1.305	1	1.423

Value	0	1	2	3	4	5
Frequency	977	719	583	199	191	64
Proportion	0.357	0.263	0.213	0.073	0.070	0.023

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_11

n	missing	distinct	Info	Mean	pMedian	Gmd
2734	1226	6	0.946	1.685	1.5	1.715

Value	0	1	2	3	4	5
Frequency	895	442	601	336	344	116
Proportion	0.327	0.162	0.220	0.123	0.126	0.042

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_12

n	missing	distinct	Info	Mean	pMedian	Gmd
2731	1229	6	0.51	0.2446	0	0.3952

Value	0	1	2	3	4	5
Frequency	2141	540	34	7	6	3
Proportion	0.784	0.198	0.012	0.003	0.002	0.001

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_13

n	missing	distinct	Info	Mean	pMedian	Gmd
2729	1231	6	0.926	1.34	1	1.502

Value	0	1	2	3	4	5
Frequency	999	711	508	220	185	106
Proportion	0.366	0.261	0.186	0.081	0.068	0.039

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_14

n	missing	distinct	Info	Mean	pMedian	Gmd
2732	1228	6	0.876	1.036	1	1.314

Value	0	1	2	3	4	5
Frequency	1285	706	369	175	125	72
Proportion	0.470	0.258	0.135	0.064	0.046	0.026

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_15

n	missing	distinct	Info	Mean	pMedian	Gmd
2730	1230	6	0.938	1.5	1.5	1.625

Value	0	1	2	3	4	5
Frequency	951	599	528	300	221	131
Proportion	0.348	0.219	0.193	0.110	0.081	0.048

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_16

n	missing	distinct	Info	Mean	pMedian	Gmd
2728	1232	6	0.934	1.452	1.5	1.601

Value	0	1	2	3	4	5
Frequency	937	714	507	203	210	157
Proportion	0.343	0.262	0.186	0.074	0.077	0.058

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_17

n	missing	distinct	Info	Mean	pMedian	Gmd
2725	1235	6	0.949	1.628	1.5	1.597

Value	0	1	2	3	4	5
Frequency	791	595	639	348	241	111
Proportion	0.290	0.218	0.234	0.128	0.088	0.041

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_18

n	missing	distinct	Info	Mean	pMedian	Gmd
2728	1232	6	0.947	1.614	1.5	1.668

Value	0	1	2	3	4	5
Frequency	824	661	585	250	219	189
Proportion	0.302	0.242	0.214	0.092	0.080	0.069

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_19

n	missing	distinct	Info	Mean	pMedian	Gmd
2730	1230	6	0.902	1.159	1	1.382

Value	0	1	2	3	4	5
Frequency	1119	797	410	157	159	88
Proportion	0.410	0.292	0.150	0.058	0.058	0.032

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_20

n	missing	distinct	Info	Mean	pMedian	Gmd
2733	1227	6	0.87	0.9437	0.5	1.171

Value	0	1	2	3	4	5
Frequency	1248	875	343	102	113	52
Proportion	0.457	0.320	0.126	0.037	0.041	0.019

For the frequency table, variable is rounded to the nearest 0

PCIAT.PCIAT_Total

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2736	1224	93	0.998	27.9	27	22.92	0	0	12	26	41	56	65

lowest : 0 1 2 3 4, highest: 89 90 91 92 93

SDS.Season

n	missing	distinct
2618	1342	4

Value	Fall	Spring	Summer	Winter
Frequency	619	712	635	652
Proportion	0.236	0.272	0.243	0.249

SDS.SDS_Total_Raw

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2609	1351	62	0.999	41.09	40	11.26	28	30	33	39	46	55	61

lowest : 17 24 25 26 27, highest: 82 84 85 93 96

SDS.SDS_Total_T

n	missing	distinct	Info	Mean	pMedian	Gmd	.05	.10	.25	.50	.75	.90	.95
2606	1354	49	0.999	57.76	56.5	14.43	41	43	47	55	64	76	84

```
lowest :  38  40  41  42  43, highest:  95  97  98  99 100
```

PreInt_EduHx.Season

n	missing	distinct
3540	420	4

Value	Fall	Spring	Summer	Winter
Frequency	828	985	821	906
Proportion	0.234	0.278	0.232	0.256

PreInt_EduHx.computerinternet_hoursday

n	missing	distinct	Info	Mean	pMedian	Gmd
3301	659	4	0.87	1.061	1	1.177

Value	0	1	2	3
Frequency	1524	413	1004	360
Proportion	0.462	0.125	0.304	0.109

For the frequency table, variable is rounded to the nearest 0

---

sii

n	missing	distinct	Info	Mean	pMedian	Gmd
2736	1224	4	0.781	0.5804	0.5	0.767

Value	0	1	2	3
Frequency	1594	730	378	34
Proportion	0.583	0.267	0.138	0.012

For the frequency table, variable is rounded to the nearest 0

3.1 資料描述

本研究使用之訓練資料集包含 3960 筆樣本，共計 82 個變數。其中包含 59 個解釋變數，主要分為以下類別：

1. 參與者基本資料 (Demographics)
2. 兒童全球評估量表 (Children's Global Assessment Scale)
3. 身體量測 (Physical Measures)
4. 健體測驗生命指標及跑步機測試 (FitnessGram Vitals and Treadmill)
5. 兒童版健體測驗 (FitnessGram Child)
6. 生物電阻抗分析 (Bio-electric Impedance Analysis)
7. 身體活動問卷青少年版 (Physical Activity Questionnaire (Adolescents))
8. 身體活動問卷兒童版 (Physical Activity Questionnaire (Children))
9. 兒童睡眠障礙量表 (Sleep Disturbance Scale)
10. 網絡使用時間 (Internet Use)

研究之反應變數為網絡成癮嚴重程度 (Severity Impairment Index, SII)，其定義基於父母評估孩子網路成癮程度的問卷 (Parent-Child Internet Addiction Test, PCIAT)，並以問卷總分

(PCIAT_Total) 量化嚴重程度。該指數依據總分範圍將樣本分為四個層級：0 = 無 (None)、1 = 輕度 (Mild)、2 = 中度 (Moderate)、3 = 重度 (Severe)。

此外，資料集中有 22 個變數為問卷中的題目分數，分別對應 PCIAT 問卷的各項評估指標。

4 前處理

4.1 檢視資料中反應變數與解釋變數的缺失值情況

4.1.1 反應變數分析

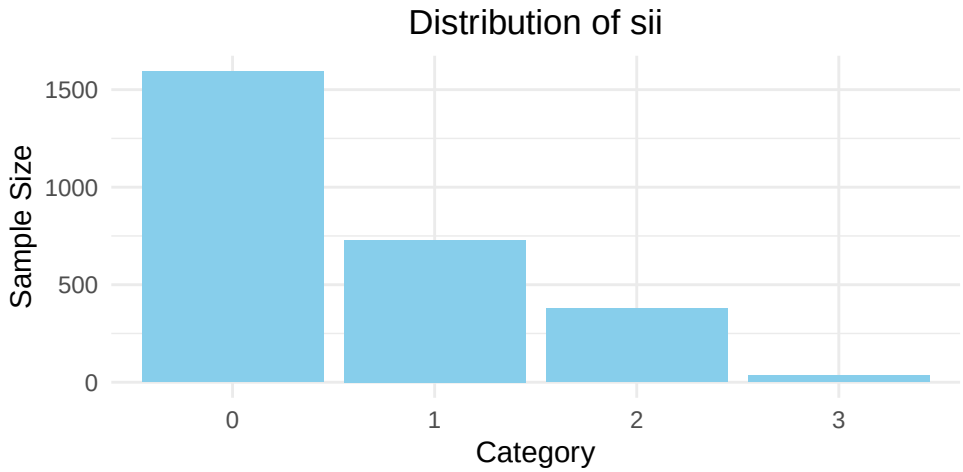


Figure 1: Distribution of sii

如圖 1 所示，反應變數網絡成癮嚴重程度 (Severity Impairment Index, SII) 的分佈顯示出類別不平衡的現象。

在所有樣本中，類別 0 (無，None) 占據了最大比例，共有 1594 個樣本；其次是類別 1 (輕度，Mild)，擁有 730 個樣本；類別 2 (中度，Moderate) 則有 378 個樣本；而類別 3 (重度，Severe) 僅有 34 個樣本。

此類別不平衡可能會對後續分析或模型訓練過程中的結果產生偏差，因此，為減少類別不平衡對模型的影響，我們將類別 2 (中度，Moderate) 與類別 3 (重度，Severe) 合併為類別 2 (中重度，Moderate-Severe)。

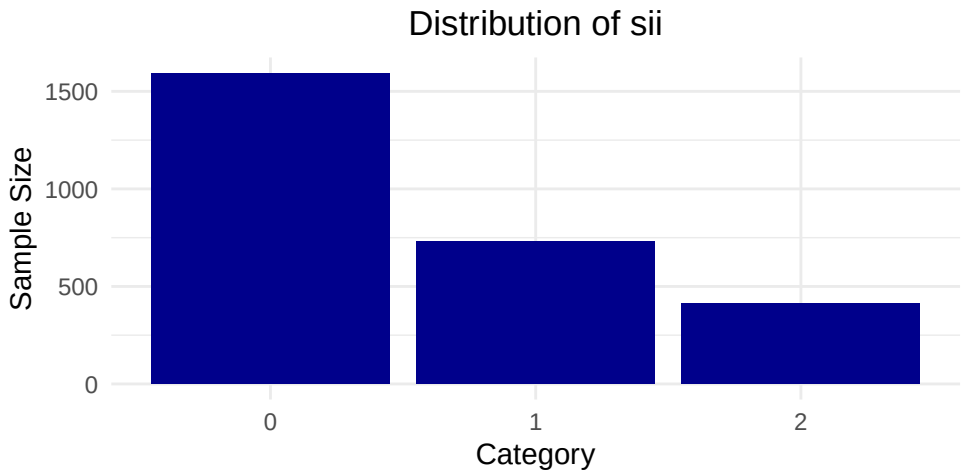


Figure 2: After combination the distribution of sii

此處的調整 (如圖 2) 有助於提升模型的穩定性並減少少數類別樣本數量對結果的過度影響。

4.1.2 檢視遺失值

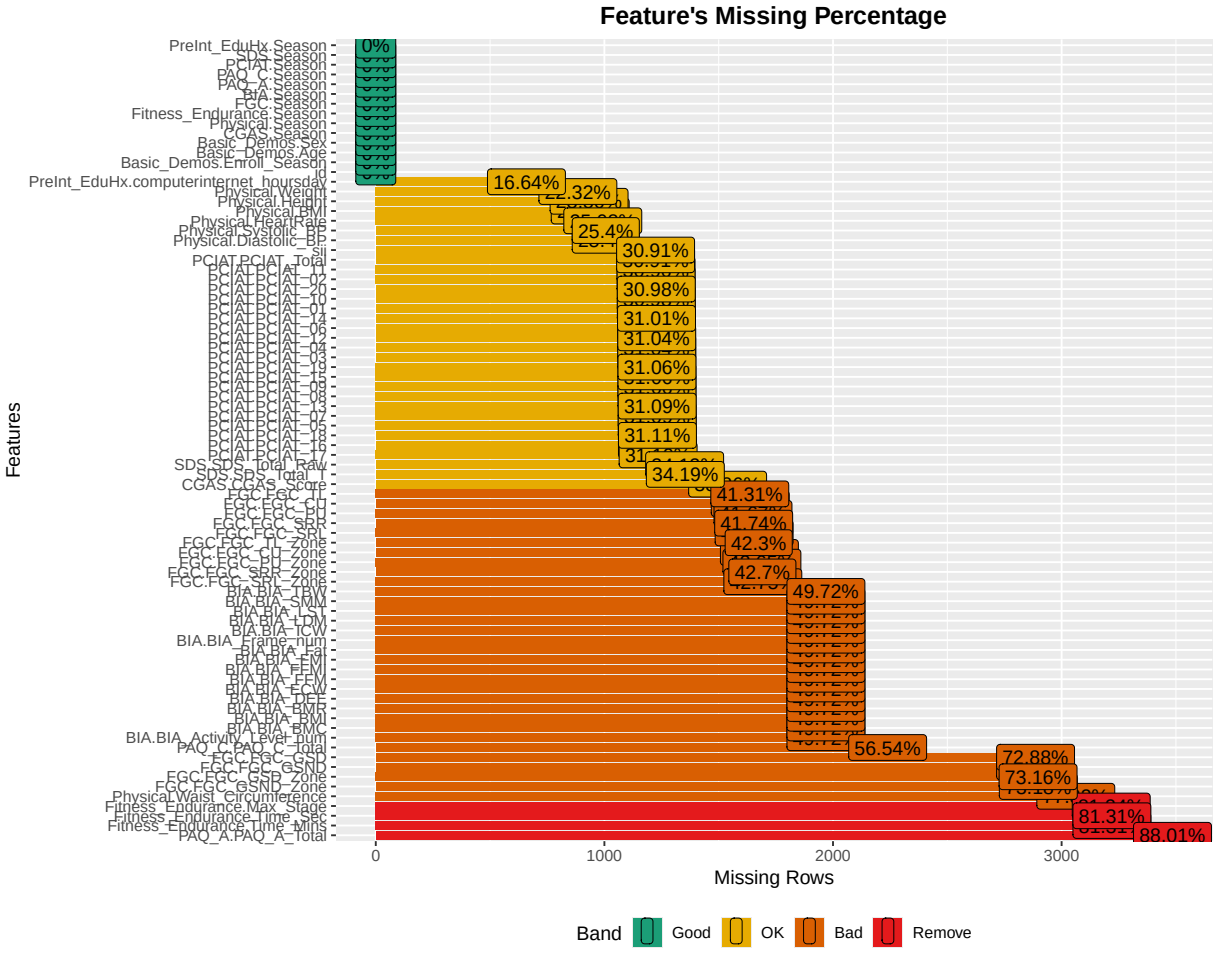


Figure 3: Missing value percentage of original data

本研究對數據集中的特徵進行了遺失值分析，結果如圖 3 所示。資料集中存在多個具有較高比例遺失值的變數，因此，為提高分析準確性，本研究將根據變數的含義及其與其他變數的相關性進行變數選擇。

4.1.3 解釋變數相關係數矩陣

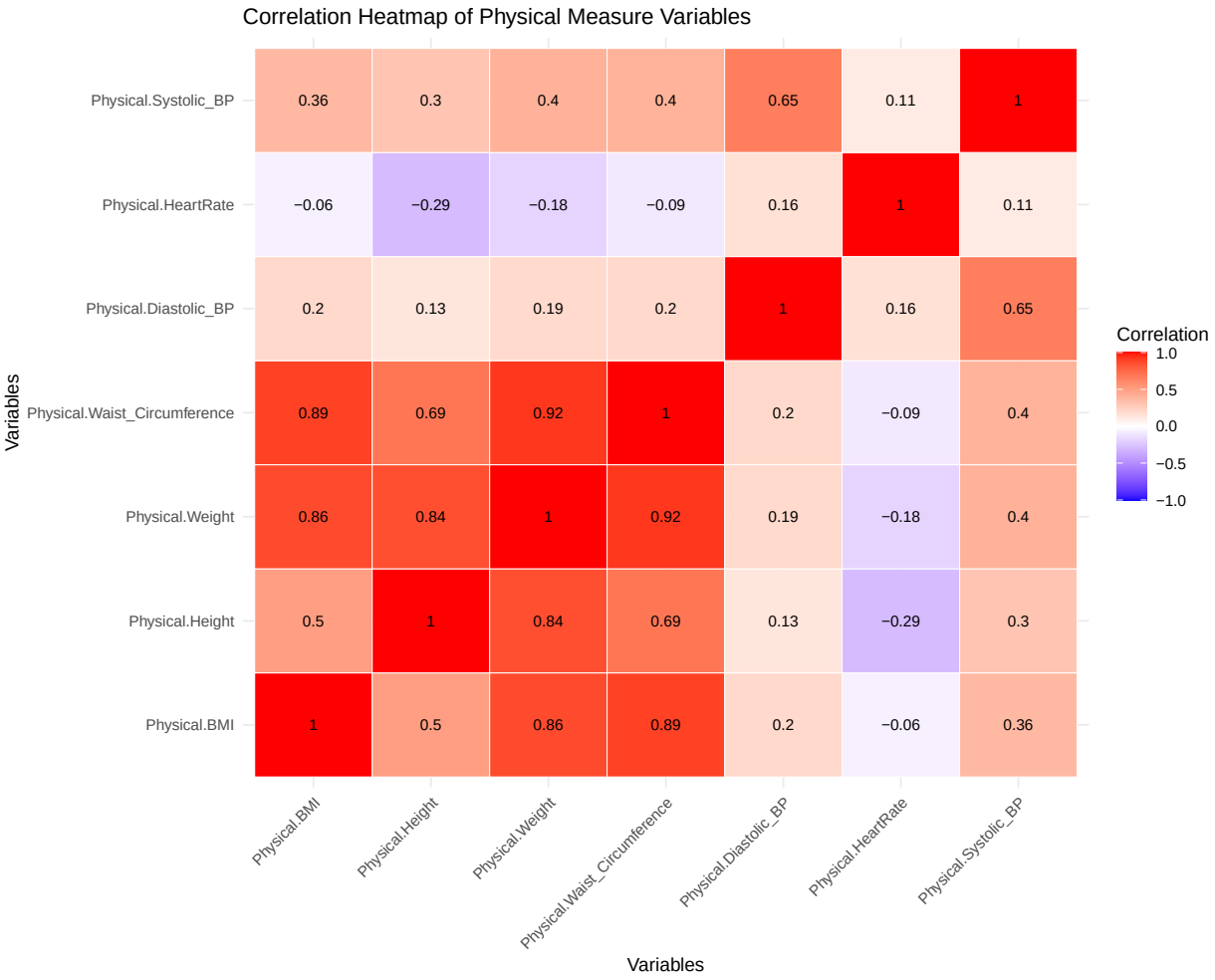


Figure 4: Correlation Heatmap of Physical Measure Variables

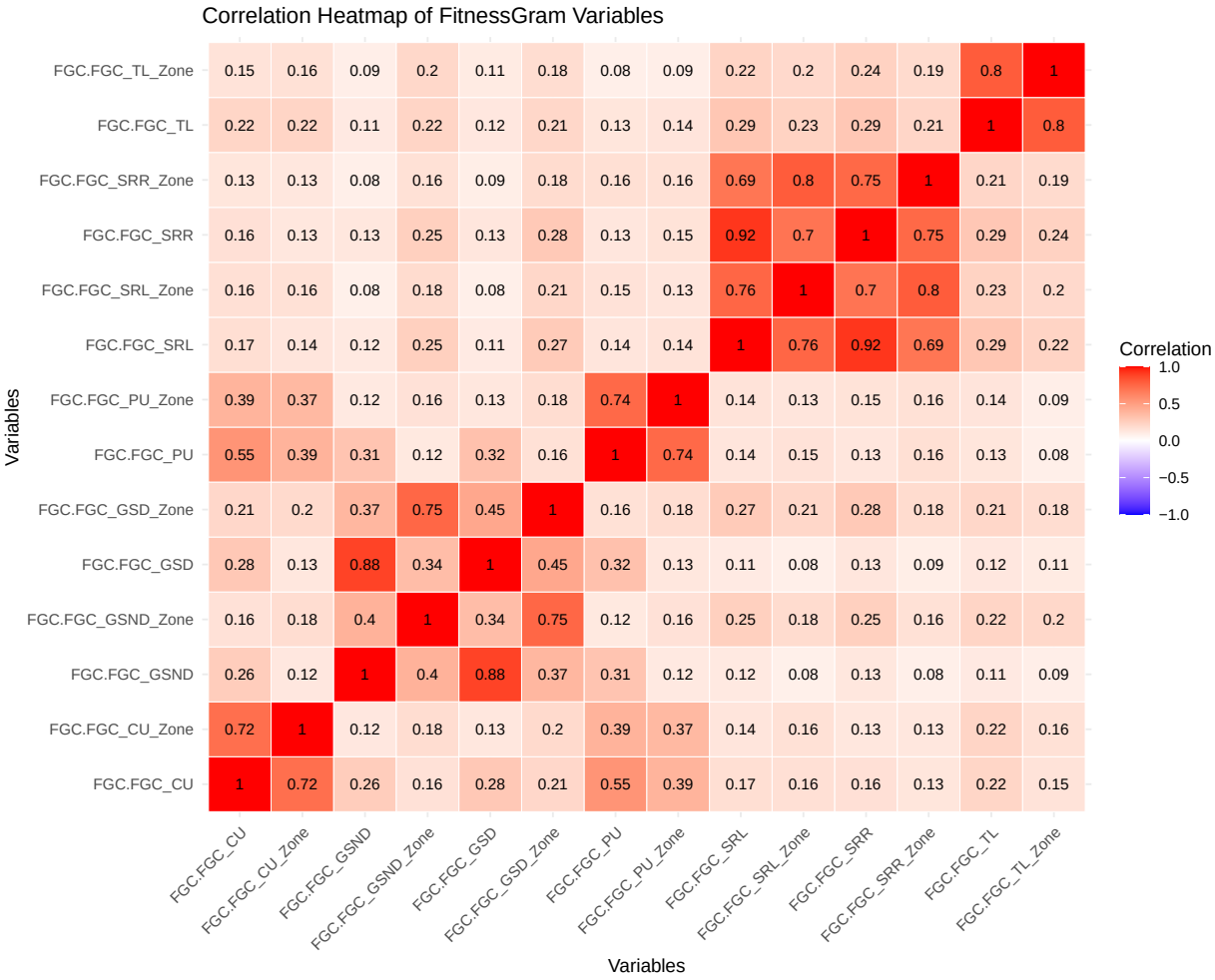


Figure 5: Correlation Heatmap of FitnessGram Variables

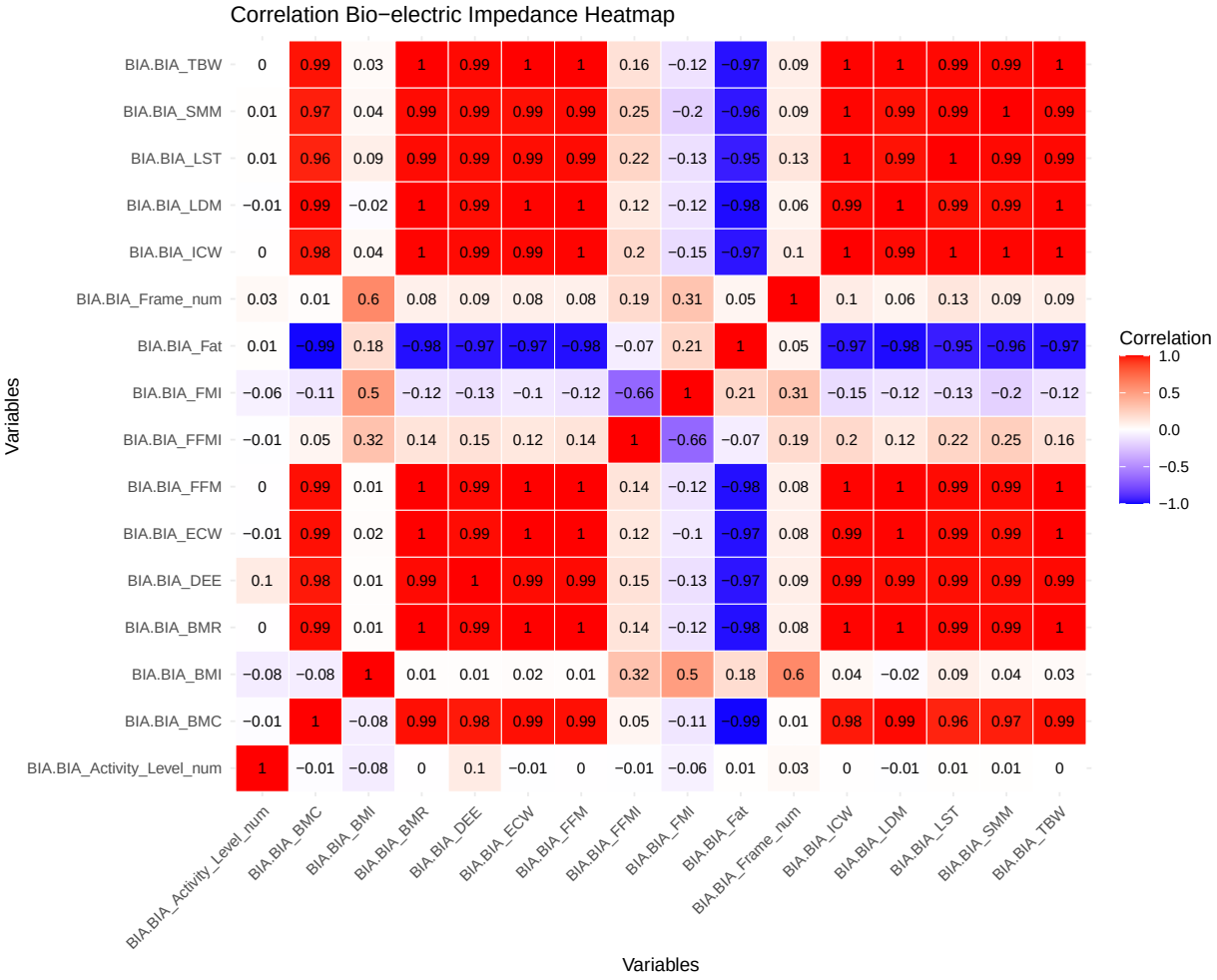


Figure 6: Correlation Heatmap of Bio-electric Impedance Analysis

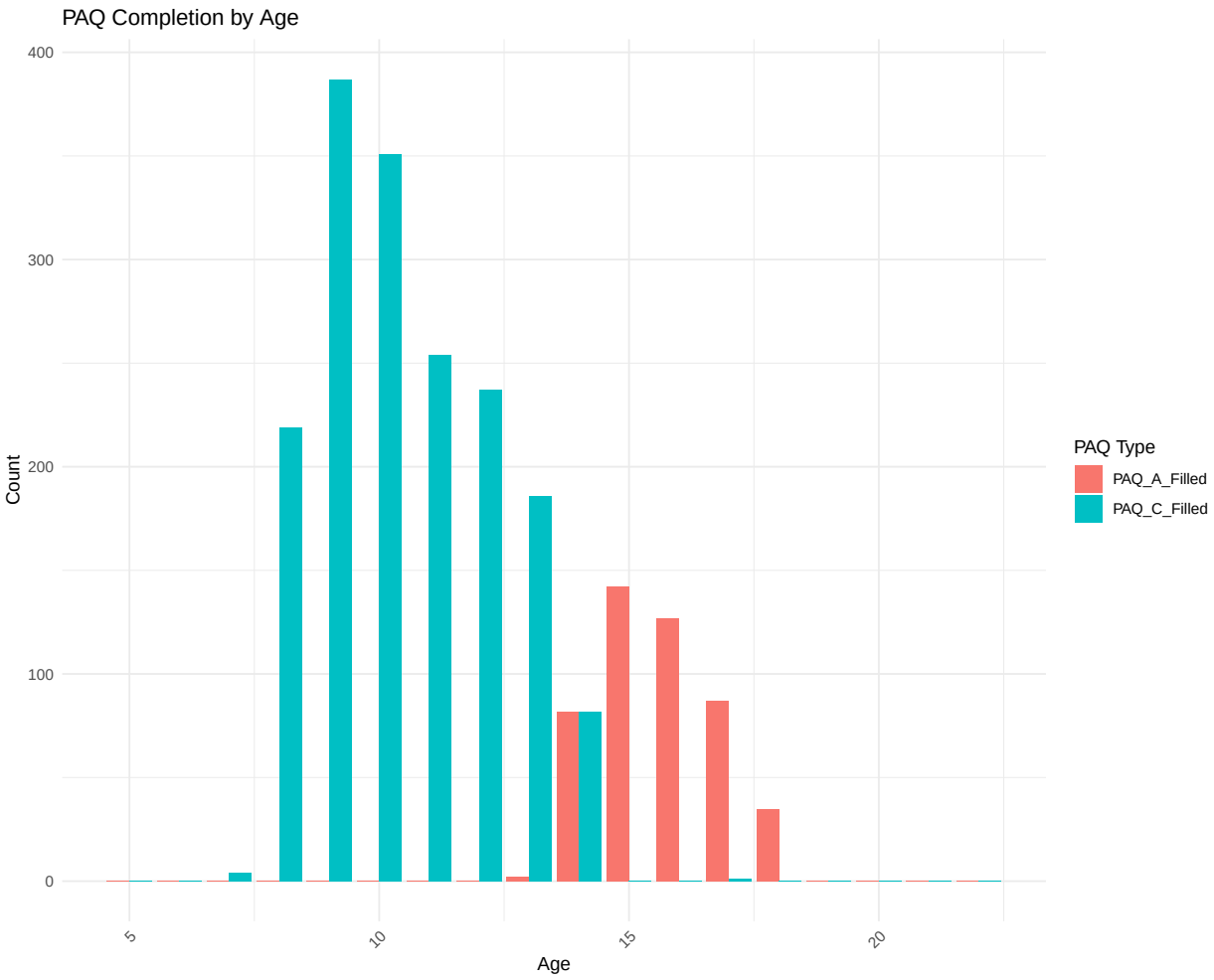


Figure 7: PAQ Completion by Age

本研究對資料中的解釋變數進行了篩選與處理，具體過程如下：

首先，我們發現資料中的解釋變數分類均記錄了數據收集時的季節。由於這些變數對研究目標的重要性較低，因此將其刪除。隨後，在身體量測 (Physical Measures) 分類中，如圖 4 可見，身體質量指數 (BMI)、身高 (Height)、體重 (Weight) 及腰圍 (Waist circumference) 之間呈現高度相關 (相關性 > 0.8)。為避免共線性問題並基於變數的重要性考量，刪除了身高 (Height)、體重 (Weight) 及腰圍 (Waist circumference)。

在健體測驗生命指標及跑步機測試 (FitnessGram Vitals and Treadmill) 分類中，如圖 3 可見，跑步機的速度或傾斜度的最高階段 (Maximum Stage Reached)、完成時間的分鐘 (Time Mins) 及完成時間的秒數 (Time Sec) 變數的遺失值比例均超過 80%。考量到兒童版健體測驗 (FitnessGram Child) 已能充分反映兒童體能數據，刪除了上述三個變數。此外，兒童版健體測驗中的變數分為體能測試的實際得分 (Total) 及根據性別、年齡和體重計算的健康標準 (Zone, 1=Weak, 2=Normal, 3=Strong)。最終僅保留健康標準 (Zone) 資料。此外，圖 5 中可了解到該分類中的坐姿體前屈左側測試 (Sit & Reach Left) 及坐姿體前屈右側測試 (Sit & Reach Right) 因缺失值比例達 41% 至 42%，且兩者之間高度相關 (相關性 > 0.7)，為減少缺失值的影響，刪除了坐姿體前屈左側測試 (Sit & Reach Left)。

在生物電阻抗分析 (Bio-electric Impedance Analysis) 分類中，BMI 變數與身體量測 (Physical Measures) 分類中的 BMI 重複，且其遺失值比例更高 (49.72%)，因此刪除。此外，圖 6 中可得知此分類中的骨礦物質含量 (BMC)、基礎代謝率 (BMR)、每日能量消耗 (DEE)、細胞外水分 (ECW)、去脂體重 (FFM)、細胞內水分 (ICW)、瘦體乾重 (LDM)、瘦軟組織 (LST)、骨骼肌質量 (SMM) 及總身體水分 (TBW) 之間存在極高度相關性 (相關性 > 0.9)。基於變數的重要性，最終僅保留骨骼肌質量 (SMM)。

在身體活動問卷青少年版 (Physical Activity Questionnaire (Adolescents)) 及身體活動問卷兒童版 (Physical Activity Questionnaire (Children)) 中，青少年版適用於 14-19 歲的青少年，而兒童版適用於 8-14 歲的兒童。如圖 7 可見，兩者數據幾乎互斥，我們將其合併為一個變數；若數據同時來自兩個測驗，則取平均值，因兩者的評分方式一致。此外，資料集本身包含來自 5 至 22 歲青少年的數據，對於不在上述測驗涵蓋年齡層內的樣本，後續將進行缺失值插補。

在兒童睡眠障礙量表 (Sleep Disturbance Scale) 中，變數分為原始分數 (Raw Score) 及標準化分數 (Total T-Score)。基於標準化分數的解釋性更強，僅保留 Total T-Score。最後，對於網絡使用時間 (Internet Use) 分類，未發現異常，故保留所有變數。

上述處理步驟有效簡化了資料結構，減少了冗餘與噪音數據，從而提升了分析的準確性與科學性。

本研究最終選取 22 個解釋變數包括以下幾個分類及其具代表性的指標：

1. 參與者基本資料 (Basic Demographics)

- 年齡 (Age)
- 性別 (Sex)

2. 兒童全球評估量表 (Children's Global Assessment Scale, CGAS)

- CGAS 總分 (CGAS_Score)

3. 身體量測 (Physical Measures)

- 身體質量指數 (BMI)

- 舒張壓 (Diastolic_BP)
- 心率 (HeartRate)
- 收縮壓 (Systolic_BP)

3. 兒童版健體測驗 (FitnessGram Zones)

- 上肢力量 (FGC_CU_Zone)
- 通用肌耐力 (FGC_GSND_Zone)
- 全身肌耐力 (FGC_GSD_Zone)
- 上肢推舉力量 (FGC_PU_Zone)
- 坐姿體前屈右側 (FGC_SRR_Zone)
- 身體總力量 (FGC_TL_Zone)

4. 生物電阻抗分析 (Bio-electric Impedance Analysis, BIA)

- 活動水平 (BIA_Activity_Level_num)
- 去脂體質量指數 (FFMI, Fat-Free Mass Index)
- 脂肪質量指數 (FMI, Fat Mass Index)
- 體脂肪百分比 (Fat)
- 體型 (Frame_num)
- 骨骼肌質量 (SMM, Skeletal Muscle Mass)

5. 身體活動問卷 (Physical Activity Questionnaire)

- 合併後的總分 (PAQ_Total_Combined)

6. 兒童睡眠障礙量表 (Sleep Disturbance Scale, SDS)

- 標準化總分 (SDS_Total_T)

7. 網絡使用時間 (Internet Use)

- 每日使用電腦與網絡的平均時數 (PreInt_EduHx.computerinternet_hoursday)

此外，本研究選取了 1 個反應變數，即網絡成癮嚴重程度 (Severity Impairment Index, SII)。

4.1.4 最終解釋變數遺失值分析

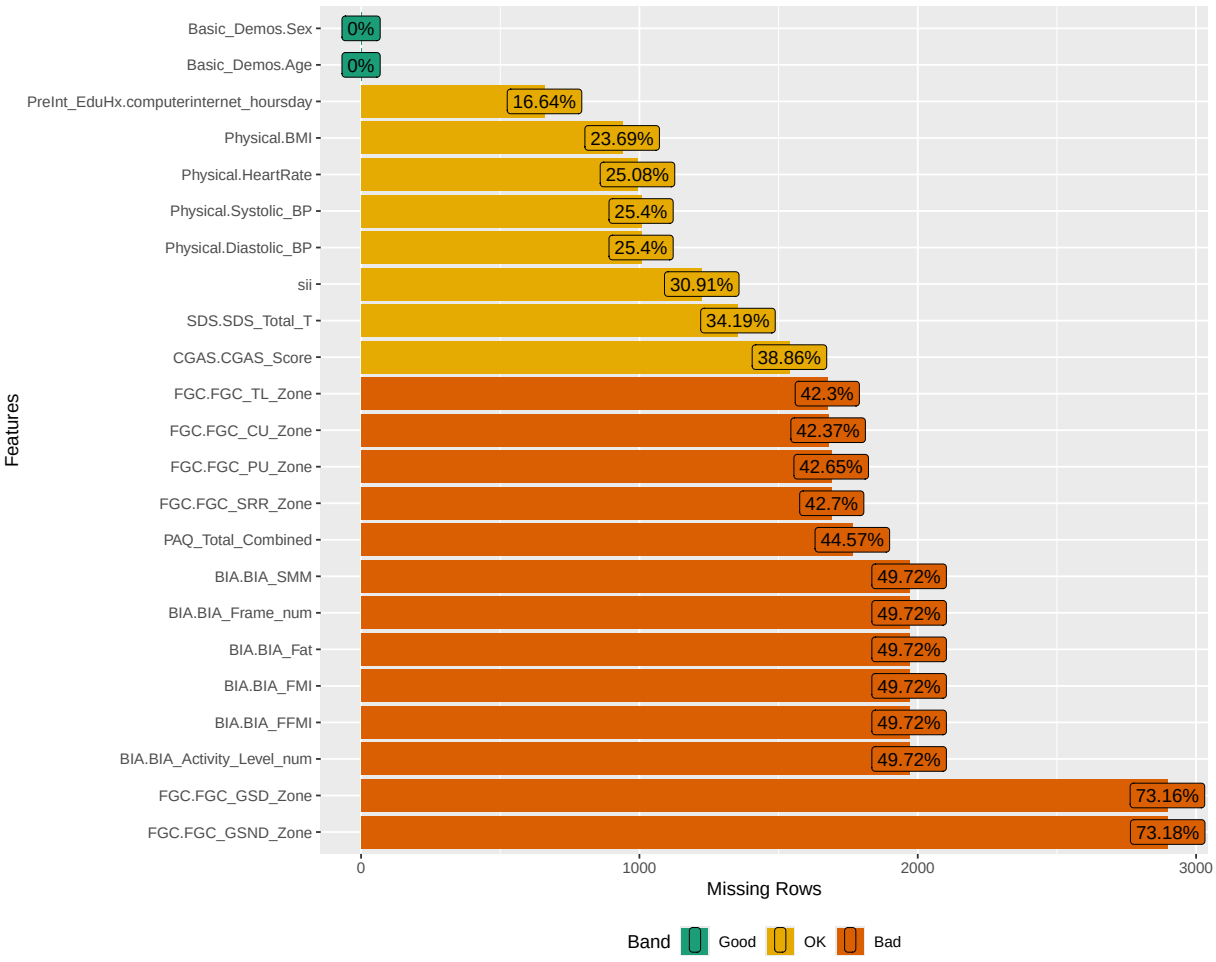


Figure 8: Missing value percentage of reduced data

在圖 8 中，我們觀察到以下兩個變數：

FGC.FGC_GSND_Zone：基於參試者的年齡和性別，非優勢手握力的測試結果被分類為不同的「健康適能區間」。FGC.FGC_GSD_Zone：基於參試者的年齡和性別，優勢手握力的測試結果被分類為不同的「健康適能區間」。由於以上兩個變數的缺失值比例均超過 70%，考慮到過多的遺失值可能影響數據的有效性與分析結果的穩健性，故決定將其刪除。

處理遺失值

在遺失值處理過程中，我們首先對反應變數**網絡成癮嚴重程度 (Severity Impairment Index, SII)** 進行處理。為確保模型預測結果的準確性與可靠性，將所有含有該變數缺失值的資料刪除，從而避免遺失值對分析結果的影響。

5 插補缺失值

其次，對於缺失值比例低於 50% 的變數，我們採用了多重插補方法 (Multiple Imputation by Chained Equations, MICE) 進行處理。該方法通過利用其他變數的信息進行迭代插補，生成合理的插補值，從而減少缺失值對分析結果的影響，提升數據的完整性與模型的準確性。

在實際操作中，我們設置了多重插補的迭代次數 (iteration) 為 50 次，並生成了 5 個插補後的資料集。隨後，我們基於這 5 個資料集分別構建並運行模型進行預測，最終選擇預測表現最佳的資料集作為後續分析的基礎。

6 模型訓練

6.1 Ordinal Logistic Regression

有序邏輯斯迴歸用來處理反應變數為順序類別變數的資料，通常採用累積邏輯模型，其核心為一個類別的累積機率建模。

$$\text{logit}(P(Y \leq j)) = \log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \alpha_j - \mathbf{x}^\top \beta, \quad j = 1, 2, \ldots, J-1$$

適用條件：

- 反應變數為順序類別變數
- 解釋變數可以有連續和類別變數
- 變數之間獨立、無多重共線性
- 平行線假設 (Parallel Lines Assumption)：表示各個反應變數會服從平行的線性模型，即迴歸係數會一致，但截距項不同。

Table 1: Comparison between Logistic Regression and Ordinal Logistic Regression

特性	Logistic Regression	Ordinal Logistic Regression
資料型態	適用於二元類別資料，例如：0/1、是/否	適用於有序類別資料，例如：低/中/高
類別數量	二元類別	多個有序類別
類別順序考量	忽略類別之間的順序關係	考慮類別之間的順序

特性	Logistic Regression	Ordinal Logistic Regression
模型假設	預測的 log-odds 為線性函數的形式	假設平行線和反應變數為順序變數
解釋重點	解釋單一類別相對於另一類別的機率 (Odds Ratio)	解釋類別累積機率，或在不同閾值間的隱變數變化
模型輸出	每個觀測值歸屬於某一類別的機率	預測每個觀測值落在某一類的累積機率
適用情境	適用於二元分類問題，例如：是否患病 (是/否)	適用於有序類別問題，例如：滿意度 (不滿意/滿意/非常滿意)
效能表現	快速、適合處理大量二元分類問題	計算較複雜，適合處理類別數較多且有序的問題
實現方式	R 套件 `glm` 或 Python 套件 `statsmodels`	R 套件 `MASS::polr` 或 Python 套件 `statsmodels` 的 `OrderedModel`

在進行預測前，先利用 `with()` 和 `pool()` 進行多重插捕資料集的參數估計值合併。其中可以利用 fmi 觀察插捕的效果好壞，利用 `summary()` 觀察模型的解釋性和參數推論。

```
Class: mipo    m = 5
                                        term m      estimate          ubar
1                          Basic_Demos.Age 5  0.1602533060 6.267122e-04
2                         Basic_Demos.Sex1 5 -0.5543477195 8.167040e-03
3                          CGAS.CGAS_Score 5 -0.0036881542 1.298485e-05
4                             Physical.BMI 5 -0.0033800827 8.241345e-04
5                    Physical.Diastolic_BP 5  0.0001165481 1.619068e-05
6                       Physical.HeartRate 5  0.0084260777 9.935645e-06
7                     Physical.Systolic_BP 5  0.0006333018 1.144750e-05
8                          FGC.FGC_CU_Zone1 5  0.1653638835 8.350661e-03
9                          FGC.FGC_PU_Zone1 5  0.2202421920 8.811074e-03
10                        FGC.FGC_SRR_Zone1 5 -0.0220600859 8.087529e-03
11                         FGC.FGC_TL_Zone1 5  0.1869224742 1.099611e-02
12             BIA.BIA_Activity_Level_num2 5 -0.1936294483 1.884532e-02
13             BIA.BIA_Activity_Level_num3 5 -0.1614330980 1.932263e-02
14             BIA.BIA_Activity_Level_num4 5 -0.2132379970 2.793241e-02
15             BIA.BIA_Activity_Level_num5 5 -0.0816720297 5.128097e-02
16                             BIA.BIA_FFMI 5 -0.0395349444 1.118655e-03
17                              BIA.BIA_FMI 5 -0.0294525538 7.815243e-04
18                              BIA.BIA_Fat 5  0.0057620036 9.497788e-06
19                        BIA.BIA_Frame_num2 5  0.3221581155 9.633119e-03
20                        BIA.BIA_Frame_num3 5 -0.0367332671 2.909160e-02
21                              BIA.BIA_SMM 5  0.0112518546 5.069249e-05
22                        PAQ_Total_Combined 5  0.0791133899 3.253829e-03
23                           SDS.SDS_Total_T 5  0.0370497022 1.016557e-05
24 PreInt_EduHx.computerinternet_hoursday1 5  0.8429431647 1.567123e-02
25 PreInt_EduHx.computerinternet_hoursday2 5  0.8210578169 1.009400e-02
26 PreInt_EduHx.computerinternet_hoursday3 5  1.3698647317 2.213017e-02
27                                       0|1 5  5.1789363566 4.526080e-01
28                                       1|2 5  6.9145438997 4.610294e-01
```

	b	t	dfcom	df	riv	lambda	fmi
1	6.406964e-05	7.035958e-04	2708	294.11798	0.12267763	0.10927236	0.11526814
2	7.129018e-05	8.252588e-03	2708	2498.22326	0.01047481	0.01036623	0.01115755
3	1.620464e-06	1.492941e-05	2708	214.30899	0.14975575	0.13025005	0.13825478
4	8.299468e-04	1.820071e-03	2708	13.21491	1.20846317	0.54719643	0.60304672
5	1.052269e-06	1.745340e-05	2708	585.84449	0.07799075	0.07234825	0.07549900
6	1.929826e-06	1.225144e-05	2708	106.51884	0.23307905	0.18902199	0.20383182
7	6.571027e-07	1.223602e-05	2708	697.72884	0.06888169	0.06444277	0.06711301
8	1.028956e-03	9.585408e-03	2708	218.69699	0.14786222	0.12881530	0.13667454
9	6.591756e-03	1.672118e-02	2708	17.65302	0.89774603	0.47305910	0.52408708
10	3.328117e-03	1.208127e-02	2708	35.87875	0.49381463	0.33057290	0.36500955
11	8.696851e-03	2.143233e-02	2708	16.66737	0.94908287	0.48693818	0.53911210
12	7.445951e-03	2.778047e-02	2708	37.86894	0.47413038	0.32163395	0.35483109
13	6.029701e-03	2.655827e-02	2708	52.45390	0.37446458	0.27244396	0.29868400
14	1.508421e-03	2.974252e-02	2708	757.88743	0.06480306	0.06085920	0.06332774
15	2.548304e-02	8.186062e-02	2708	28.18790	0.59631560	0.37355746	0.41372961
16	1.250489e-03	2.619242e-03	2708	12.05961	1.34142024	0.57290879	0.62962886
17	9.179685e-04	1.883086e-03	2708	11.56873	1.40950468	0.58497694	0.64195145
18	2.342045e-06	1.230824e-05	2708	73.99957	0.29590610	0.22833915	0.24838240
19	8.313506e-03	1.960933e-02	2708	15.27687	1.03561557	0.50874811	0.56250477
20	9.751716e-03	4.079365e-02	2708	47.41509	0.40224877	0.28685978	0.31515052
21	1.185438e-05	6.491774e-05	2708	80.14459	0.28061852	0.21912733	0.23791082
22	2.008882e-03	5.664488e-03	2708	21.77622	0.74086826	0.42557399	0.47194312
23	1.345831e-06	1.178057e-05	2708	195.05847	0.15886925	0.13708988	0.14580357
24	1.823643e-04	1.589007e-02	2708	2368.96127	0.01396426	0.01377194	0.01460352
25	1.115915e-04	1.022791e-02	2708	2396.32689	0.01326628	0.01309259	0.01391524
26	8.613603e-04	2.316380e-02	2708	1130.44593	0.04670694	0.04462274	0.04630854
27	1.518665e-01	6.348478e-01	2708	47.35005	0.40264390	0.28706067	0.31537998
28	1.535156e-01	6.452481e-01	2708	47.85865	0.39958132	0.28550061	0.31359806

先觀察插捕後的效果。FMI (Fraction of Missing Information) 為衡量因資料缺失而導致的不確定性，它表示每個估計量中的總變異有多少來自於遺失值的插補過程，值介於 0 到 1 之間，越小越好。從 fit 有序邏輯斯迴歸模型結果可以看出有些變數的 fmi 還是有點偏高，像是 BMI(Physical.BMI)、上肢推舉力量 (FGC.FGC_PU_Zone)、身體總力量 (FGC.FGC_TL_Zone)、去脂體質量指數 (BIA.BIA_FFMI)、脂肪質量指數 (BIA.BIA_FMI)、體型 (BIA.BIA_Frame_num2)，大於 0.5。

	term	estimate	std.error
1	Basic_Demos.Age	0.1602533060	0.026525380
2	Basic_Demos.Sex1	-0.5543477195	0.090843756
3	CGAS.CGAS_Score	-0.0036881542	0.003863859
4	Physical.BMI	-0.0033800827	0.042662285
5	Physical.Diastolic_BP	0.0001165481	0.004177727
6	Physical.HeartRate	0.0084260777	0.003500205
7	Physical.Systolic_BP	0.0006333018	0.003498003
8	FGC.FGC_CU_Zone1	0.1653638835	0.097905098
9	FGC.FGC_PU_Zone1	0.2202421920	0.129310405
10	FGC.FGC_SRR_Zone1	-0.0220600859	0.109914824

```
11                          FGC.FGC_TL_Zone1  0.1869224742 0.146397854
12              BIA.BIA_Activity_Level_num2 -0.1936294483 0.166674729
13              BIA.BIA_Activity_Level_num3 -0.1614330980 0.162967096
14              BIA.BIA_Activity_Level_num4 -0.2132379970 0.172460190
15              BIA.BIA_Activity_Level_num5 -0.0816720297 0.286112946
16                             BIA.BIA_FFMI -0.0395349444 0.051178533
17                              BIA.BIA_FMI -0.0294525538 0.043394544
18                              BIA.BIA_Fat  0.0057620036 0.003508310
19                      BIA.BIA_Frame_num2  0.3221581155 0.140033304
20                      BIA.BIA_Frame_num3 -0.0367332671 0.201974392
21                             BIA.BIA_SMM  0.0112518546 0.008057155
22                      PAQ_Total_Combined  0.0791133899 0.075262792
23                         SDS.SDS_Total_T  0.0370497022 0.003432283
24 PreInt_EduHx.computerinternet_hoursday1  0.8429431647 0.126055829
25 PreInt_EduHx.computerinternet_hoursday2  0.8210578169 0.101133114
26 PreInt_EduHx.computerinternet_hoursday3  1.3698647317 0.152196590
27                                      0|1  5.1789363566 0.796773383
28                                      1|2  6.9145438997 0.803273373
     statistic          df      p.value
1   6.04150836  294.11798 4.607187e-09
2  -6.10221049 2498.22326 1.208999e-09
3  -0.95452596  214.30899 3.408929e-01
4  -0.07922882   13.21491 9.380378e-01
5   0.02789749  585.84449 9.777534e-01
6   2.40730970  106.51884 1.779028e-02
7   0.18104667  697.72884 8.563835e-01
8   1.68902220  218.69699 9.264035e-02
9   1.70320549   17.65302 1.060700e-01
10 -0.20070164   35.87875 8.420653e-01
11  1.27681157   16.66737 2.191714e-01
12 -1.16172050   37.86894 2.526191e-01
13 -0.99058707   52.45390 3.264349e-01
14 -1.23644765  757.88743 2.166752e-01
15 -0.28545381   28.18790 7.773836e-01
16 -0.77249077   12.05961 4.546974e-01
17 -0.67871559   11.56873 5.106676e-01
18  1.64238709   73.99957 1.047530e-01
19  2.30058213   15.27687 3.589922e-02
20 -0.18187091   47.41509 8.564593e-01
21  1.39650475   80.14459 1.664182e-01
22  1.05116204   21.77622 3.047099e-01
23 10.79447659  195.05847 1.344211e-21
24  6.68706217 2368.96127 2.830631e-11
25  8.11858532 2396.32689 7.460501e-16
26  9.00062698 1130.44593 9.284755e-19
27  6.49988625   47.35005 4.553349e-08
28  8.60795855   47.85865 2.762803e-11
```

模型結果以迴歸係數 (estimate)、標準誤差 (std.error)、統計量 (statistic)、P-value 呈現，可以得知年齡 (Basic_Demos.Age)、性別 (basic_demos.sex)、心率 (Physical.HeartRate)、體型 (BIA.BIA_Frame_num2)、兒童睡眠障礙量表標準化總分 (SDS.SDS_Total_T)、每日使用電腦與網絡的平均時數 (PreInt_EduHx.computerinternet_hoursday)，這些變數在顯著水準為 0.05 下，為此模型的顯著變數，表示這些變數可能對網絡成癮嚴重程度 (sii) 有影響。而 $0|1: p < 0.05$ 、$1|2: p < 0.05$ ，表示不同網絡成癮嚴重程度 (sii) 類別之間的分界點顯著，能有效區分不同類別。

Table 2: Result of Ordinal Logistic Regression

	x
Accuracy	0.6208791
Quadratic Weighted Kappa	0.3571736

利用多重插捕五個資料集進行建模預測，把驗證集放入模型做測試，可以得到平均準確率為 62.08%(表 2)，還算可以接受，但也沒有達到不錯的表現。平均 QWK 值為 0.357，表示模型對有序分類的預測有效果，但一致性不高。

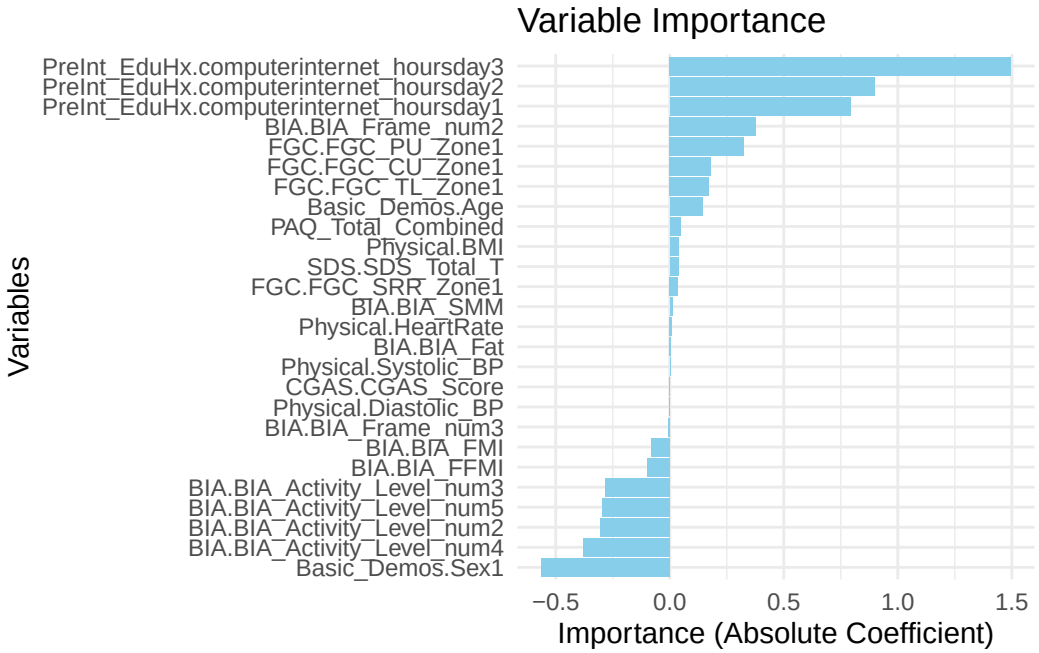


Figure 9: Importance variable of Ordinal Logistics

從圖 9 觀察 Ordinal Logistic Regerssion 模型中的特徵重要性，特徵重要性前幾名為每日使用電腦與網路平均時數 (PreInt_EduHx.computerinternet_hoursday)、性別 (Basic_Demos.Sex)、活動水平 (BIA_Activity_Level_num)，表示在 Ordinal Logistic Regerssion 模型中這些變數是重要的。其中，性別和活動水平是負相關，表示男性的網路成癮程度可能高於女性，而活動水平越高，網路成癮程度傾向於降低。

6.2 Ordinal Forest

Ordinal Forest 是一種隨機森林演算法的變體，專門用來處理有序類別變數的預測問題。將數據中的類別 (例如：低、中、高) 視為具有順序的數值，並在建模時利用這種順序資訊來提高預測準確率。

工作原理:

- 建立分數集：首先，會建立多個分數集。對於每個分數集，會先從 Uniform(0,1) 分佈中隨機抽取 J-1 個值 (其中 J 是類別的數量)。這些值會被排序，並定義為區間邊界。然後，將每個類別值替換為其對應區間中點的逆標準常態分佈函數 $(\Phi{-}1)$。這個過程會產生一個連續變數，用於訓練回歸森林。

- 生成回歸森林：使用新建立的連續變數作為反映變數，並使用原始的解釋變數，會建立一個回歸森林。

- 評估森林效能：根據其袋外 (OOB) 預測效能，使用效能函數 g 來評估每個森林的效能。先將預測的連續變數轉換回原始的類別值，然後再評估預測的準確性。效能函數 g 可以根據不同的需求進行選擇，例如：希望每個類別的預測準確度相同，或者希望正確分類的樣本數量最多。

- 選擇最佳森林和建立優化的分數集：選擇具有最高效能分數的預先定義數量的森林。然後，通過平均這些選定森林中的分數集來計算優化的分數集。

- 訓練最終的回歸森林：使用優化的分數集和原始的解釋變數來訓練最終的回歸森林。這個最終的森林用於預測新的觀察結果。

序數森林通過嘗試許多不同的分數集，並選擇在預測原始序數反應變數方面表現最佳的分數集，以迭代的方式找到最佳的連續表示法。

Table 3: Comparison between Random Forest and Ordinal Forest

特性	普通隨機森林 (Random Forest)	有序森林 (Ordinal Forest)
資料型態	適用於類別型 (分類) 或數值型 (迴歸) 資料	專為處理有序類別資料設計，例如“低”、“中”、“高”
類別順序考量	忽略類別之間的順序關係	考慮類別之間的順序關係，避免預測結果與真實值相差過遠
分裂準則	以資訊增益 (Information Gain) 或基尼係數 (Gini Index) 為基準	使用順序敏感的分裂準則，優化有序類別的預測
適用情境	適用於所有類別型問題，例如是否患病 (是/否)	適用於有序類別問題，例如風險分級 (低/中/高)
模型輸出	類別標籤或數值預測	類別標籤，並確保輸出順序的合理性
誤差懲罰	預測錯誤時，無法區分「小錯誤」與「大錯誤」	預測錯誤時，較大懲罰遠離真實值的錯誤
特徵重要性	提供變數重要性評估，例如基於分裂次數	提供有序資料的變數重要性評估
對類別不平衡的處理	支援權重調整或重新取樣 (Resampling)	同樣支援權重調整或重新取樣，並針對小樣本類別提供改進
效能表現	快速、靈活，適合大規模資料	效能較高，但計算量稍多於普通隨機森林
實現方式	R 套件 `randomForest` 或 Python 套件 `scikit-learn`	R 套件 `ordfor`

參數設置

- classweights 是用來為每個類別賦予不同的權重，當類別不平衡時，為了減少多數類別對模型的影響，可以給少數類別賦予更高的權重。

- perffunction 是設定模型的性能評估方式，用於選擇最佳特徵組合。

設定 perffunction 為"proportional" 指的是在 Ordinal Forest 演算法中，使用 gclprop 效能函數來評估每個森林的效能。

gclprop 函數會根據每個類別的樣本數量比例來設定權重，優先考慮較大類別的預測準確度。

也就是說，模型會更重視將較多樣本的類別預測正確，而較小類別的預測準確度則可能較低。

以下是 gclprop 函數的公式：

$$gclprop(y, \hat{y}) = \sum_{j=1}^{J} \left(\frac{\#\{y_i = j : i \in \{1, \dots, n\}\}}{n} \cdot Yind(y, \hat{y}, j) \right)$$

其中：

- $\#\{y_i = j : i \in \{1, \dots, n\}\}$ 表示屬於類別 j 的樣本數量。
- n 表示總樣本數量。
- $Yind(y, \hat{y}, j)$ 表示類別 j 的 Youden 指數，用於衡量類別 j 的預測準確度。

使用 gclprop 函數時，模型會傾向於將樣本預測到樣本數量較多的類別，以最大化整體的預測準確度。然而，這也意味著較小類別的預測準確度可能會受到影響。

Table 4: Result of Ordinal Forest

	x
Accuracy	0.6051282
Quadratic Weighted Kappa	0.3208794

整體效能 (表 4)：

- 準確率 (Accuracy)：60.51% 表示模型對目標變數的預測中有超過一半是正確的，但這可能不足以滿足高準確性的需求，特別是如果應用場景需要非常準確的分類。
- Quadratic Weighted Kappa (QWK)：QWK 值為 0.329，表示模型對有序分類的預測有一定效果，但一致性並不高。(QWK=0 表示隨機預測)。QWK 特別適用於有序類別資料，它會根據類別之間的相對距離進行評估。此低分數表示模型在區分不同類別時未能充分考慮類別順序。

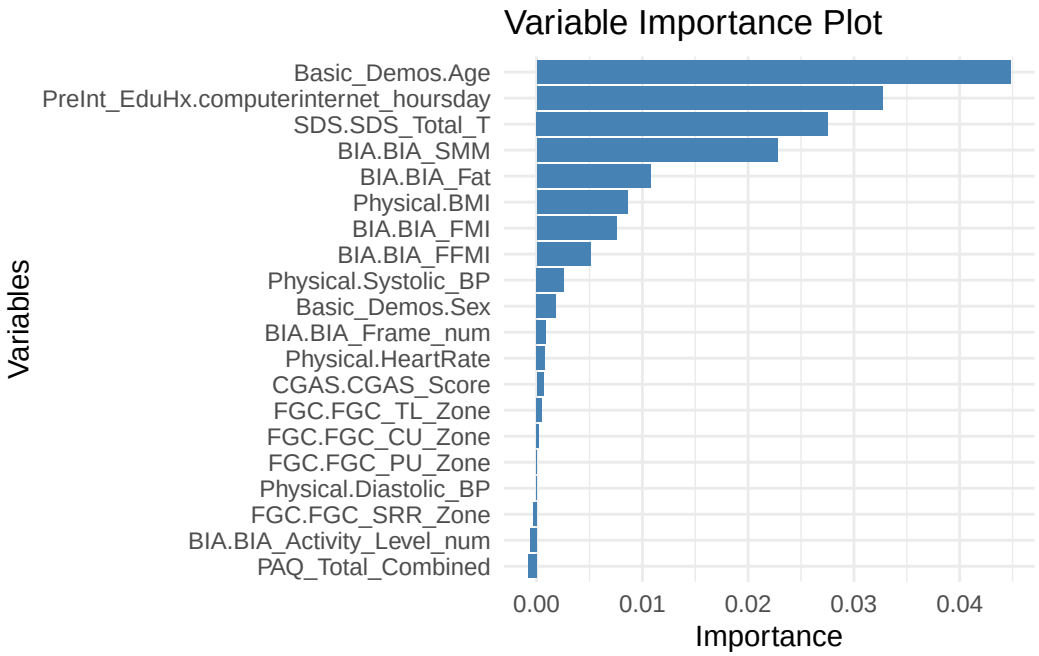


Figure 10: Importance variable of Ordinal Forest

在 Ordinal Forest 模型中，特徵重要性通過評估每個特徵對於模型預測的貢獻度來進行排名。根據該模型的分析 (圖 10)，特徵重要性排名前三的變數為：年齡 (Basic_Demos.Age)、每日使用電腦與網絡的平均時數 (PreInt_EduHx.computerinternet_hoursday) 以及兒童睡眠障礙量表標準化總分 (SDS.SDS_Total_T)。這表明，在此分析中，年齡對預測結果的影響最大，其次是每日使用電腦與網絡的時間，最後是兒童睡眠障礙量表的總分，這些變數在模型的預測中扮演著關鍵角色。

6.3 CatBoost

CatBoost 是一種基於梯度提升 (Gradient Boosting) 的機器學習方法，專為處理分類特徵而設計，此方法提出了有序目標編碼 (Ordered Target Encoding) 的來避免資料洩漏 (data leakage)，並有效提升模型的表現。

6.3.1 Ordered Target Encoding

Ordered Target Encoding 首先根據樣本的順序對數據進行排序，確保每個樣本的編碼只會參考之前的樣本，而不會使用未來樣本的目標變數資訊。

對於每個類別特徵的每個樣本，此方法計算它與之前所有相同類別值的目標變數的加權平均值，為了避免在少數樣本的情況下出現極端的編碼值，過程引入了平滑參數 (a)，以減少少數樣本對編碼值的影響，從而避免過度擬合。

計算公式為：

$$\hat{x}_k^i = \frac{\sum_{j=1}^{i-1} \mathbb{I}(x_{\sigma_j,k} = x_{\sigma_i,k}) \cdot y_{\sigma_j} + a \cdot p}{\sum_{j=1}^{i-1} \mathbb{I}(x_{\sigma_j,k} = x_{\sigma_i,k}) + a}$$

其中，$x_{\sigma_j,k}$ 表示第 j 個樣本在類別特徵 k 上的取值，$x_{\sigma_i,k}$ 表示第 i 個樣本在類別特徵 k 上的取值，$\mathbb{I}(\cdot)$ 是指示函數，當條件成立時其值為 1，否則為 0，y_{σ_j} 是第 j 個樣本的目標變數值，a 是平滑參數，p 是全局目標變數均值，即所有樣本目標變數的平均值。

將原本的類別特徵換成計算出的編碼值，使模型能夠使用數值特徵進行訓練。這樣的編碼方法減少了資訊洩漏，並提升模型的穩定性和泛化能力。

Table 5: Comparison between Ordered Target Encoding and Other Encoding Methods

方法	優點	缺點	適用情況
Ordered Target Encoding	防止資訊洩漏 (Target Leakage)，提高泛化能力	計算較為複雜；需要大量數據支持	類別特徵較多、數據量大
Target Encoding	簡單、高效率	容易出現資訊洩漏)、需處理極端值與樣本不均的問題	類別變數較少
One-Hot Encoding	易於理解、編碼時無需計算	類別數量過多時，會導致特徵維度爆炸，增加計算量	類別數量少，特徵較簡單

6.3.2 CatBoost 模型建構

為了增強穩健性，CatBoost 會多次進行隨機排列並進行計算，且會以貪婪方式選擇特徵組合，通過結合有用的特徵來擴展特徵空間，提升模型的預測能力。CatBoost 使用 Oblivious Tree 預測，通過梯度提升方法最小化損失函數，以達到最佳預測效果。其中 Oblivious Tree 同一層的所有節點使用相同的分割條件，這種結構有助於減少過擬合風險，並提高模型的穩定性。

解決資料不平衡問題，避免多數類別影響資料:

本研究通過計算每個類別的標記次數 (class_counts)，能夠識別資料集中的類別分佈，特別是少數類別和多數類別之間的比例。透過反向比例計算，CatBoost 為每個類別分配權重 (class_weights)，這樣少數類別的權重會較高，使模型在訓練過程中能夠更加關注少數類別，減少多數類別的影響。

Table 6: Result of Catboost

Evaluation	Average
Accuracy	0.5744
Quadratic Weighted Kappa	0.2645

根據比較五個模型的結果 (表 6)，平均準確率為 57.44%，準確率較低，平均 Quadratic Weighted Kappa (QWK) 值為 0.2645，顯示模型在處理有序分類時的一致性較差，意味著模型未能充分考慮類別間的順序關係，預測效果有限。這些結果表明，模型可能需要通過調整學習率、迭代次數與樹的深度等參數，以提升模型的準確性和穩定性。

```
Best imputation dataset: 4

Confusion Matrix on Test Set:

   test_predictions
      0   1   2
  0 213  74  31
  1  47  56  43
  2  10  23  49
```

從表現最佳的模型之混淆矩陣來看，模型在類別 0 和類別 2 之間的誤分類較為嚴重，尤其是類別 0 常被誤分為類別 2，類別 1 則常被誤分為類別 2。這表示類別 1 和類別 2 之間的區別較為模糊，可能是特徵重疊所致。

特徵重要性 (基於 CatBoost 模型在構建過程中使用該特徵進行分裂的頻率及貢獻)

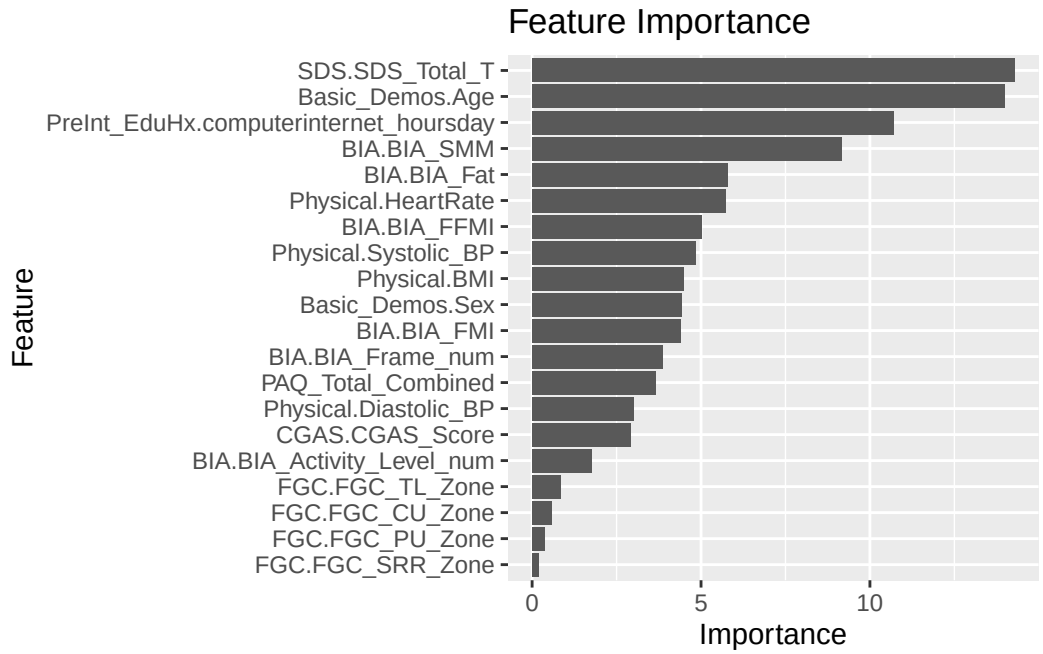


Figure 11: Importance variable of CatBoost

在 CatBoost 模型中，特徵重要性通過計算特徵在建樹過程中分裂節點的頻率和貢獻來評估，根據 CatBoost 模型的分析，(圖 11) 特徵重要性排名前三的變數為：兒童睡眠障礙量表標準化總分 (SDS.SDS_Total_T) 、年齡 (Basic_Demos.Age) 以及每日使用電腦與網絡的平均時數 (PreInt_EduHx.computerinternet_hoursday)。這些特徵不僅在 CatBoost 模型中顯示出高貢獻，也在 Ordinal Forest 是重要變數，也在 Ordinal Logistic Regression 模型中被證明是顯著的變數，顯示它們對預測目標變數具有重要的影響力。因此，雖然未能達到很好的預測效果，此結果暗示這些變數對目標變數的影響可能較大，因此可以視為評估網路成癮的參考變數。

7 結論

Table 7: Result of All Model

Model	Kappa	Accuracy
Ordinal Logistic Regression	0.3571736	0.6208791
Ordinal Forest	0.3208794	0.6051282
CatBoost	0.2645000	0.5744000

根據模型評估結果 (表 7)，Ordinal Logistic Regression 的準確率為 0.6209，Kappa 值為 0.3572，顯示其在預測準確性上表現較好，Ordinal Forest 的準確率為 0.6051，Kappa 值為 0.3209，表現略遜色於 Ordinal Logistic Regression，但仍能提供相對穩定的預測結果。CatBoost 準確率 (0.5744)，Kappa 值 (0.2645) 都不高，表現較其他兩個模型弱。總體來看，Ordinal Logistic Regression 在此次序行類別資料中表現最好，是處理此分類問題的最佳選擇。

綜合三個模型，可以推測每日使用電腦與網絡的平均時數（PreInt_EduHx.computerinternet_hoursday）、年齡（Basic_Demos.Age）、兒童睡眠障礙量表標準化總分（SDS.SDS_Total_T）、性別（basic_demos.sex）對目標變數的影響可能較大，可以視為評估網路成癮的參考。

8 工作分配

Table 8: Work Assignment Table

負責人	工作項目
廖芷萱	資料描述、資料前處理、口頭報告、書面報告製作
詹雅鈞	資料前處理、CatBoost、書面報告製作
李姿慧	資料前處理、Ordinal Forest、書面報告製作
謝沛恩	Ordinal Logistic Regression、書面報告製作
李敏榕	資料描述、簡報製作

9 參考資料

[1]Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of Big Data, 7(1), 94. https://doi.org/10.1186/s40537-020-00369-8

[2]J. K. Sayyad, K. Attarde and N. Saadouli(2024), "Optimizing e-Commerce Supply Chains With Categorical Boosting: A Predictive Modeling Framework," in IEEE Access, vol. 12, pp. 134549-134567, 2024, doi: 10.1109/ACCESS.2024.3447756

[3]https://www.w3computing.com/articles/using-catboost-for-categorical-feature-handling-in-machine-learning/

[4]https://www.youtube.com/watch?v=KXOTSkPL2X4

[5]Hornung, R. (2017). Ordinal forests. Journal of Machine Learning Research, 18(159), 1–25.

[6]Institute for Digital Research and Education. (n.d.). Ordinal logistic regression in R. UCLA: Statistical Consulting Group. https://stats.oarc.ucla.edu/r/dae/ordinal-logistic-regression/

[7]Cheng Hua, Dr. Youn-Jeng Choi, Qingzhou Shi. (2021). Binary logistic regression. In Advanced regression techniques. Retrieved from Binary Logistic Regression (Bookdown)

[8]Shawn. (2024). 順序羅吉斯回歸 (Ordinal Logistic Regression)：介紹與解讀. Medium. Retrieved from Ordinal Logistic Regression 簡介與解讀