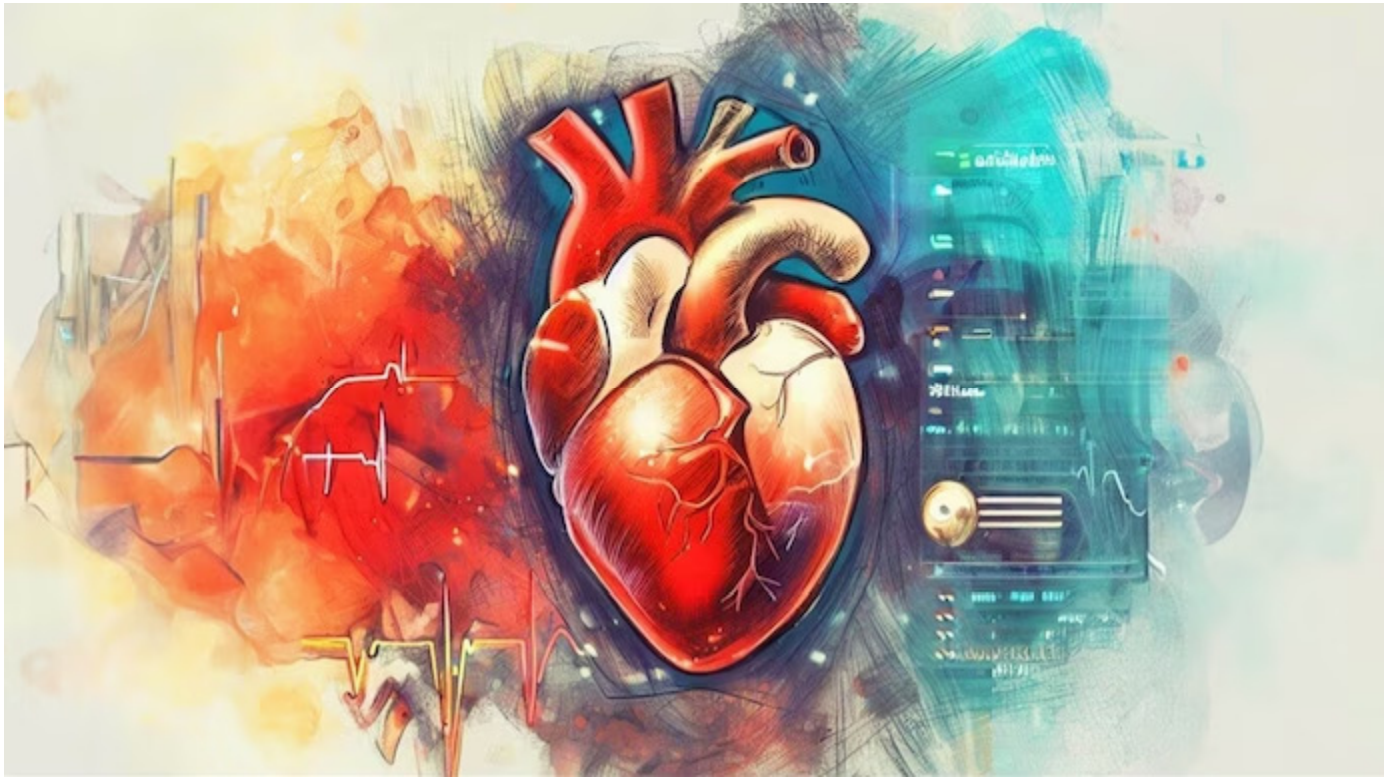


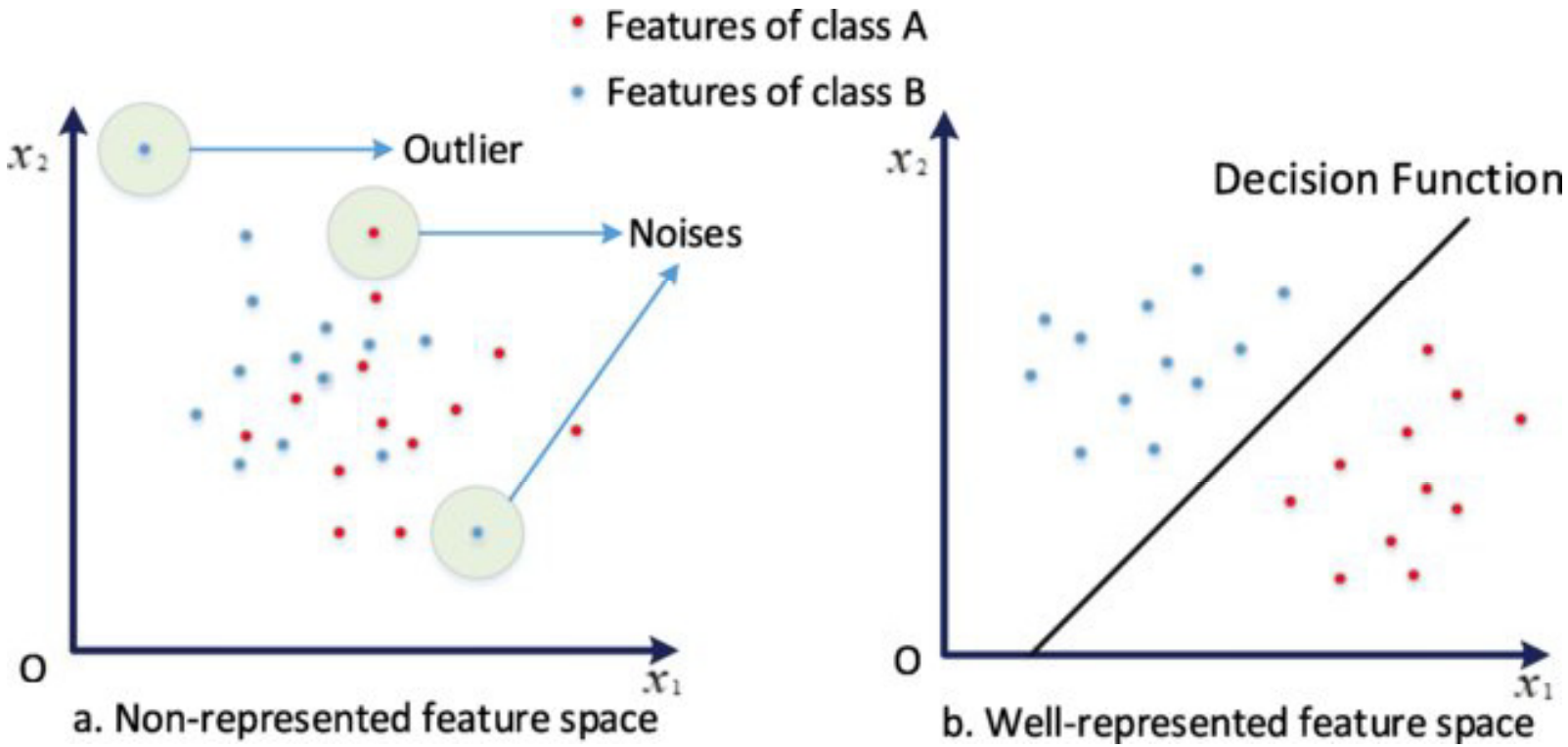
تحلیل داده‌های طرح پیشگیری از بیماری عروق کرونر زودرس توسط روش‌های یادگیری ماشین

دکتر حمید سعادت فر

تابستان ۱۴۰۳

Data Preparation

❑Feature Space: Choosing Informative Features



Some Examples

❑Feature Space: Choosing Informative Features

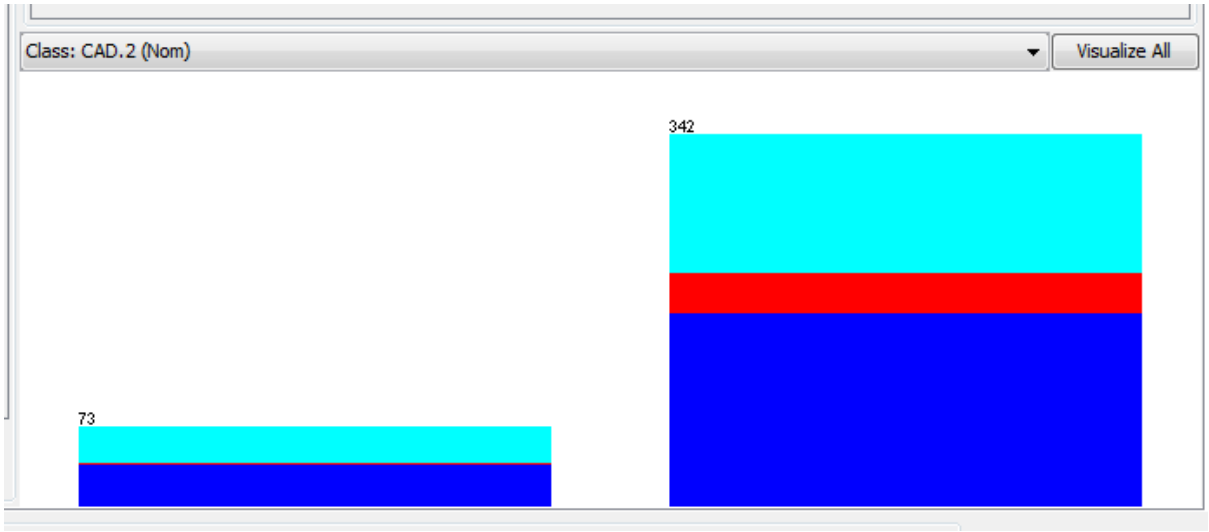
آزمایشات کلی					
HDL	AZ1		SGOT (AST)	AZ8	
LDL	AZ2		SGPT (ALT)	AZ9	
Chol	AZ3		APO A	AZ10	
TG	AZ4		APO B	AZ11	
hs-CRP	AZ5		HbA1C	AZ12	
IL1	AZ6		FBS	AZ12a	
IL6	AZ7				

به عنوان مثال، میانگین کلسترول برای افراد بیمار از افراد سالم کمتر می باشد.

Some Examples

❑Feature Space: Choosing Informative Features

حال که به سوالات فوق پاسخ دادید، با دقت فکر کنید و پاسخ سوالات زیر را به یاد بیاورید:		
آیا در **طول زندگی خود**، غالبا سطح فعالیت فیزیکی شما در همین حدود (که در سوالات بالا پاسخ دادید) بوده است؟ ۱) بله ۲) خیر	P17	



۸۲.۴٪ از افراد به سوال فوق جواب خیر داده‌اند!!!

Some Examples

❑Feature Space: Choosing Informative Features

Ethnicity	با توجه به قومیت های موجود، یکی از اعداد را بر اساس گفته بیمار، انتخاب نمایید؟				
	۱) نمی دانم ۲) فارس ۳) ترک آذربایجانی ۴) گیلک ۵) کرد ۶) عرب ۷) لر (به جز بختیاری) ۸) بلوچ ۹) ترکمن ۱۰) ترک قشقایی ۱۱) بختیاری ۱۲) تالشی ۱۳) ارمنی ۱۴) گرجی ۱۵) آشوری ۱۶) سایر (به دقت ثبت گردد)				
	* در صورتی که محل تولد را نمی دانید، کد ۱ را وارد نمایید.				
قومیت فرد	ET1		محل تولد فرد	ET8	
قومیت پدر	ET2		محل تولد پدر	ET9	
قومیت مادر	ET3		محل تولد مادر	ET10	
قومیت پدربزرگ پدری	ET4		محل تولد پدربزرگ پدری	ET11	
قومیت مادربزرگ پدری	ET5		محل تولد مادربزرگ پدری	ET12	
قومیت پدربزرگ مادری	ET6		محل تولد پدربزرگ مادری	ET13	
قومیت مادربزرگ مادری	ET7		محل تولد مادربزرگ مادری	ET14	

قومیت غالب فارس بوده و بررسی تاثیر قومیت امکان‌پذیر نمی باشد

Features with low variance: These features often carry little information and may even introduce noise into the model's predictions.

Some Examples

❑Feature Space: Choosing Informative Features

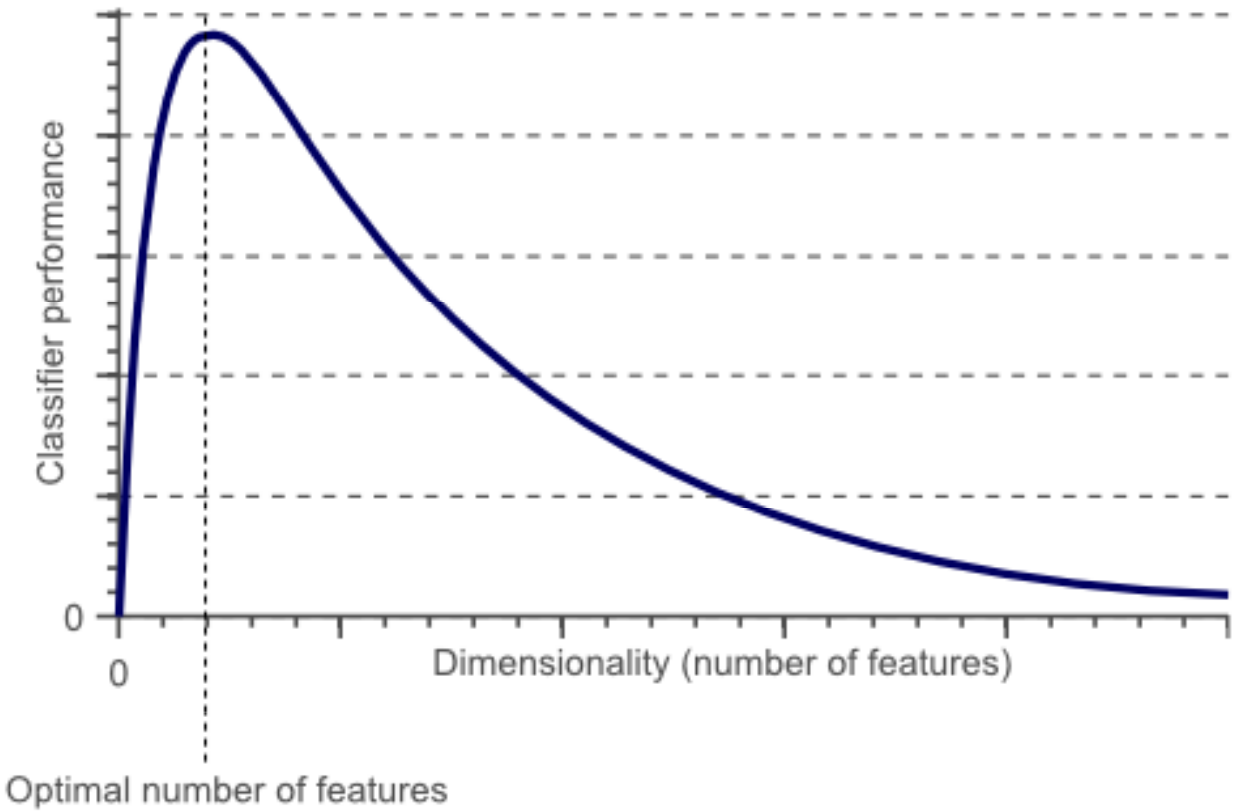
۹۶.۱۴٪ از افراد سابقه آسم، ۹۵.۶۶٪ حساسیت غذایی و ۹۳٪ حساسیت دارویی نداشته‌اند.

گزینه	در طول زندگی خود حدودا چند بار علایم زیر را حس کردید؟					کد	گزینه
	۱	۲	۳	۴			
حساسیت فصلی	هرگز	یک تا پنج بار	شش تا ده بار	بیش از ده بار		Ar1	
حساسیت دارویی	هرگز	یک تا پنج بار	شش تا ده بار	بیش از ده بار		Ar3	
حساسیت غذایی	هرگز	یک تا پنج بار	شش تا ده بار	بیش از ده بار		Ar5	
کهیر پوستی	هرگز	یک تا پنج بار	شش تا ده بار	بیش از ده بار		Ar7	
آیا سابقه ابتلا به آسم را دارا می باشید؟		۱) بلی ↓	۲) خیر			Ar9	
اگر بلی، مدت ابتلا به آن چند سال بوده است؟						Ar11	
آیا سابقه ابتلا به اگزما را دارا می باشید؟		۱) بلی ↓	۲) خیر	۳) نمی دانم		Ar12	
اگر بلی، مدت ابتلا به آن چند سال بوده است؟						Ar14	

Data Preparation

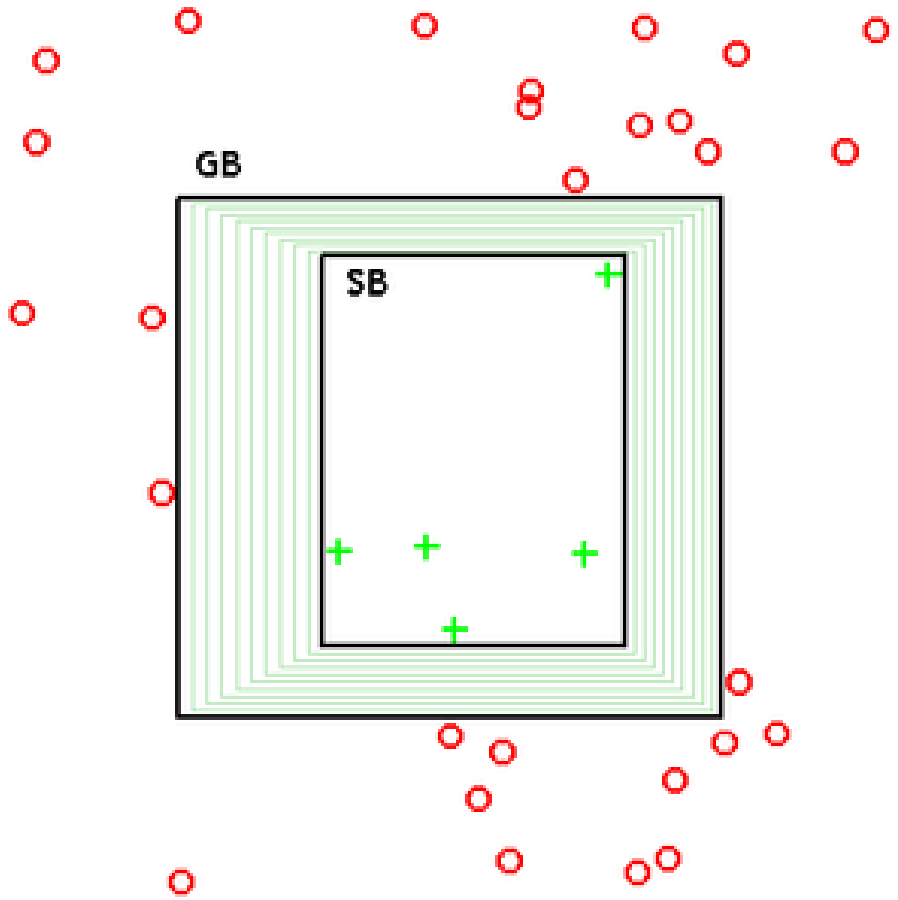
❑Feature Space: Curse of Dimensionality

بالغ بر ۲۰۰ فیلد
اطلاعاتی و تنها
۴۱۵ نمونه !!!



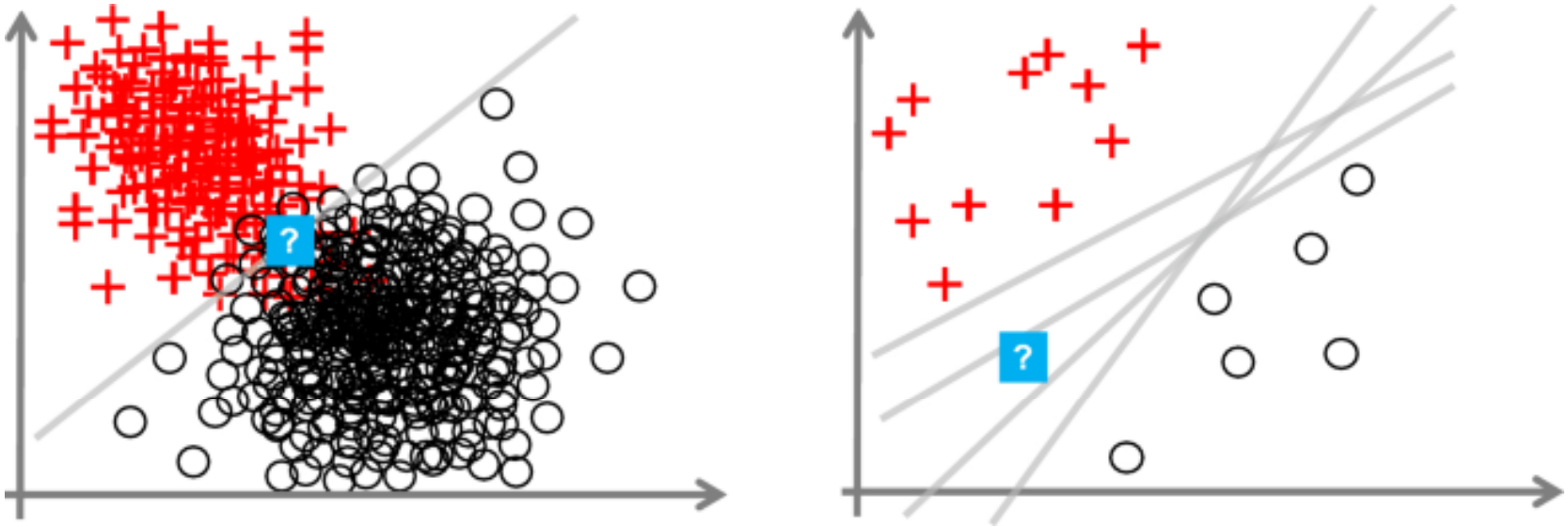
Data Preparation

- **Feature Space:** Curse of Dimensionality



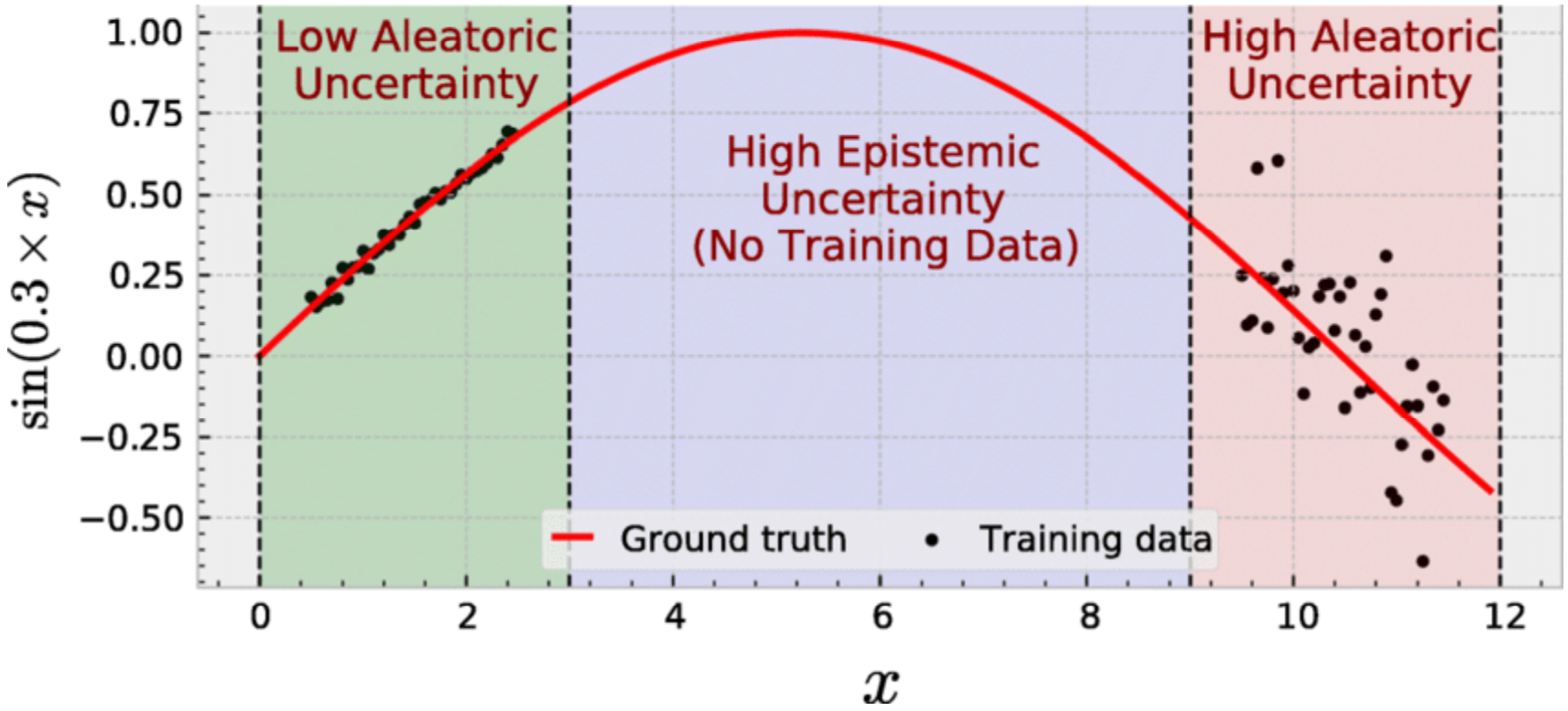
Data Preparation

- **Feature Space:** Curse of Dimensionality



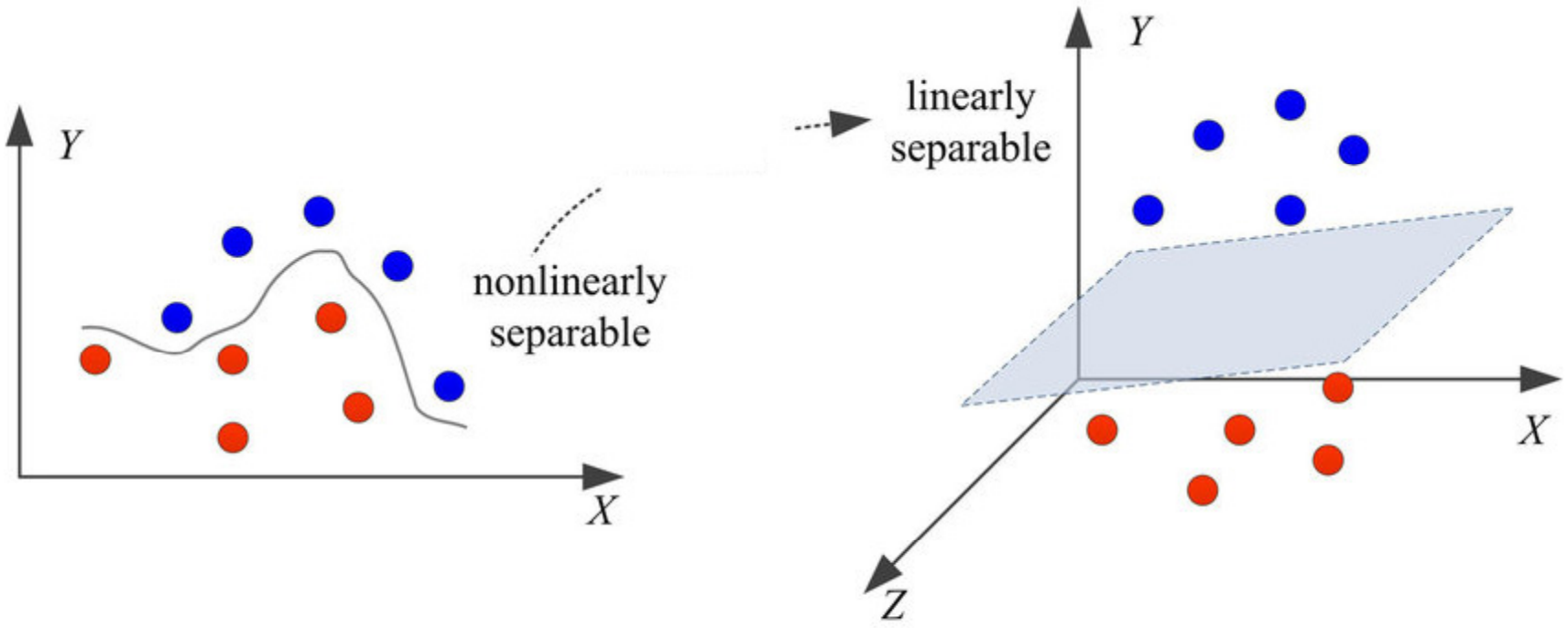
Data Preparation

❑Feature Space: Curse of Dimensionality



Data Preparation

- **Feature Space:** Considering all important features as possible



Data Preparation

❑Feature Space: Highly correlated features provide redundant information

		چربی خون (بسیار مهم: دقت شود که سوالات زیر، تا قبل از شروع علایم بیماری قلبی نظیر درد سینه یا احساس فشار روی قفسه سینه را مد نظر دارد و این نکته بایستی به درستی به بیمار تفهیم شود).
	HL1	**سابقه افزایش چربی خون:** ۱) بلی ۲) خیر
	HL1m	**مدت ابتلا به چربی خون:** ماه:
	HL1y	سال:
		سابقه مصرف دارو جهت درمان چربی خون:
	HL2	۱) بلی ۲) خیر ۳) علیرغم تجویز پزشک دارو مصرف نمی‌کردم
	HL2a	**اگر بلی، مدت مصرف دارو (سال):**
	HL3	**آیا به طور کلی دارو را مرتب مصرف می کردید؟** ۱) بلی ۲) خیر

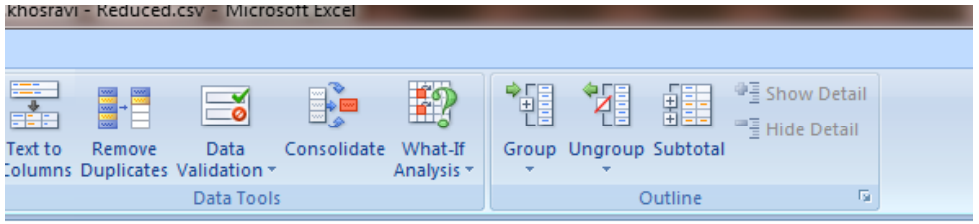
همبستگی بالای ۰.۹ که در موارد متعدد می‌توان شاهد بود

Data Preparation

❑Feature Space:

Handling Missing Values

در مواردی نرم‌افزارها یا توابع یادگیری ماشین مقادیر گم شده را (اگر مدیریت نشوند) صفر در نظر می‌گیرند که می‌تواند به تحلیل‌ها آسیب بزند

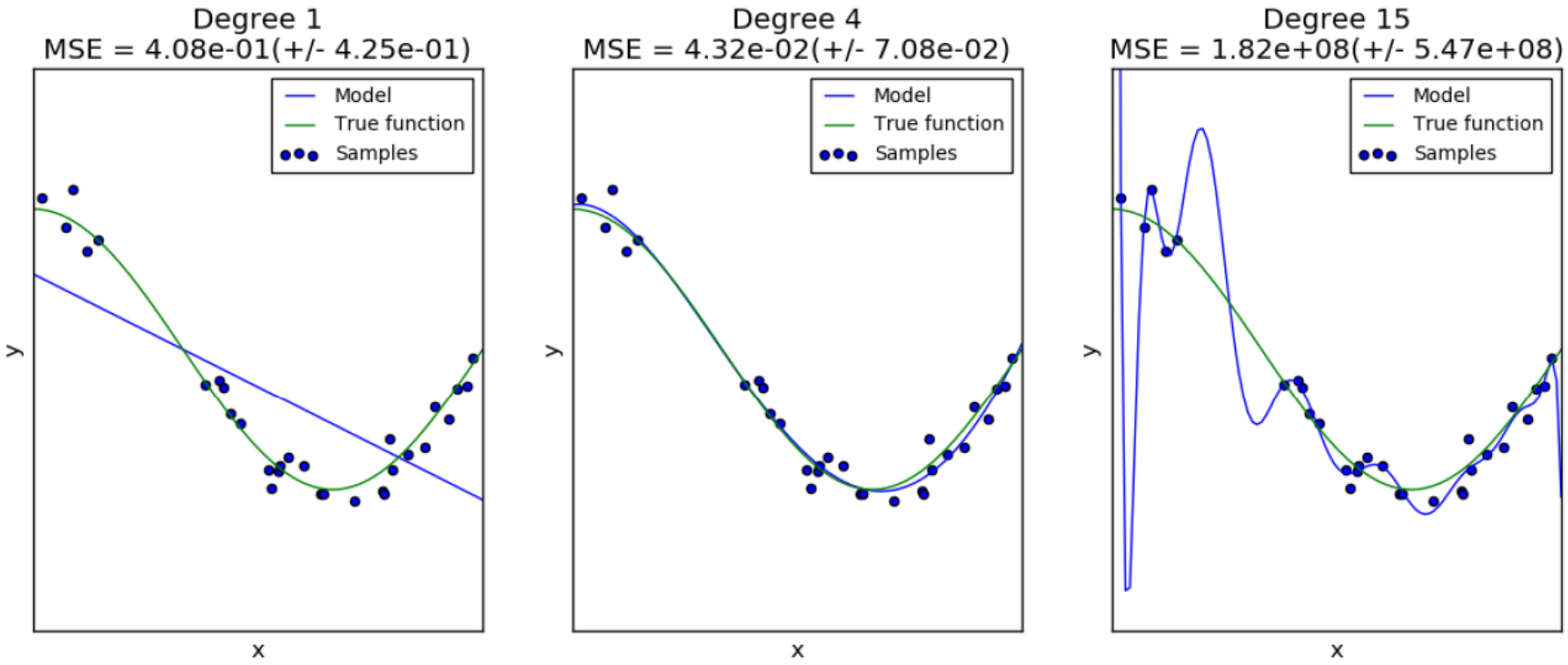


	DB	DC	DD	DE	DF	DG	DH	DI	DJ
	AR13	AR14	DP1	DP2	DP3	LS21H	LS21ha	LS22	LS23
2			1	1	1	1	22	15	
2			3	2	2	1	23	10	
2			4	4	1	1	23	5	
2			3	2	1	1	23	1	
2			3	2	1	1	22	10	
2			2	1	1	1	22	20	
2			1	1	1	2		20	
2			1	1	1	1	21	5	
1	1	4	3	2	1	1	23	5	
1	2	7	3	2	1	1	22	5	
2			2	2	1	1	23	60	
2			3	1	1	1	22	5	
2			3	2	1	1	23	10	
2			1	1	1	1	21	5	
2			3	4	3	1	23	50	
2			4	4	2	1	22	60	
2			3	3	1	1	22	10	
1	2	10	1	1	1	2		2	
2			4	2	1	1	23	60	
2			3	2	2	1	22	5	
2			3	2	1	2		5	
1	2	20	3	3	1	1	21	2	
2			2	3	1	2		10	
2			4	4	1	1	23	10	

High Generalization Ability

❑Important Point:

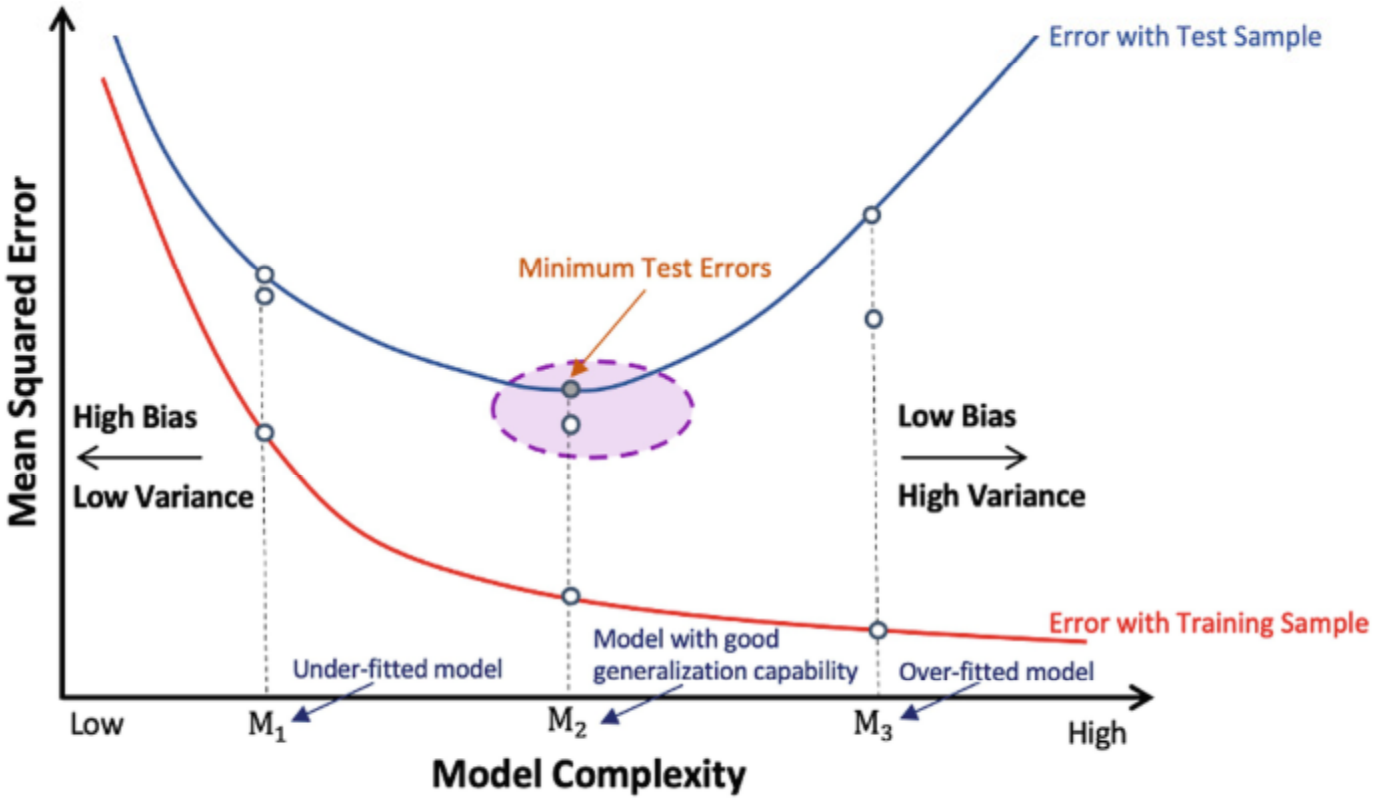
high accuracy on training data does not guarantee effectiveness in the real world



High Generalization Ability

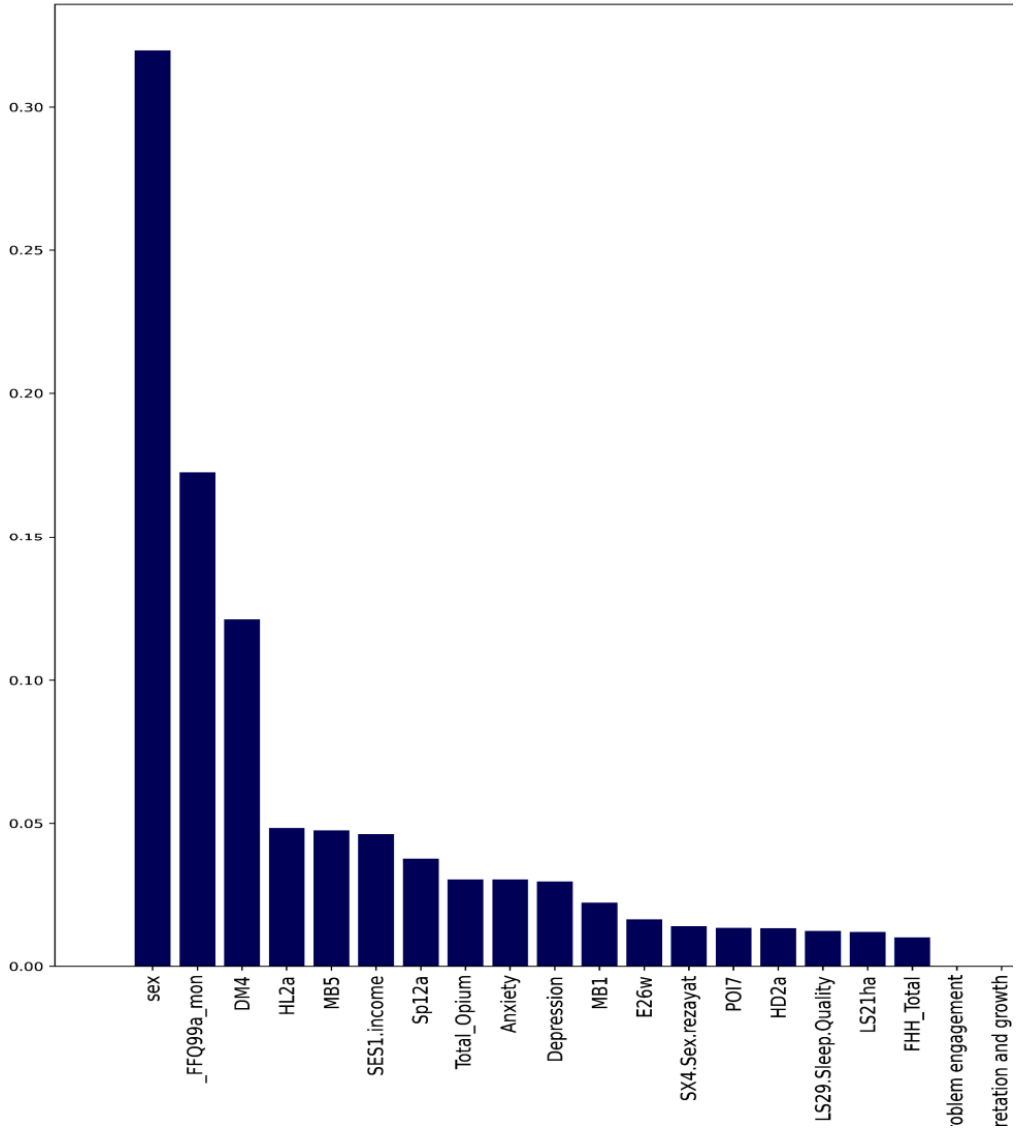
❑Important Point:

high accuracy on training data does not guarantee effectiveness in the real world



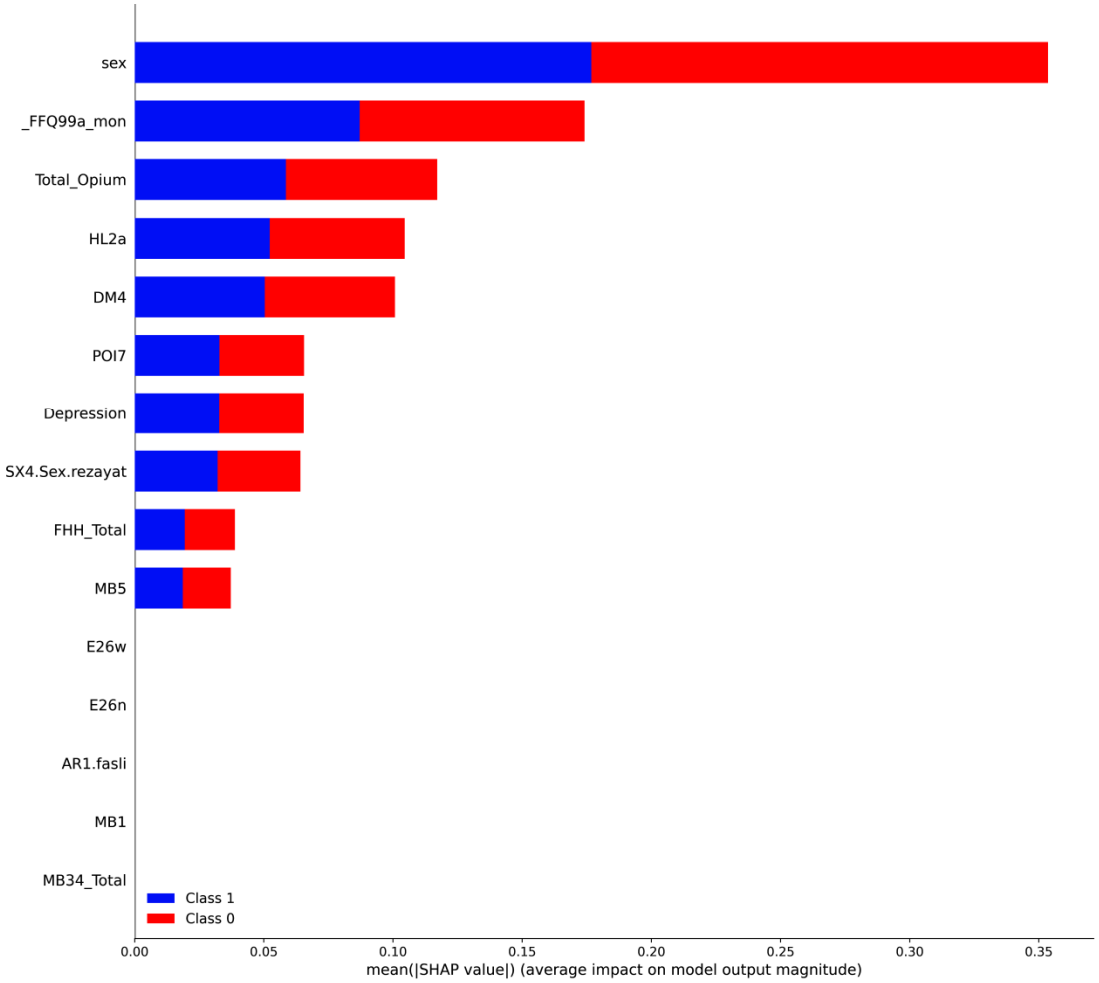
Importance of Features

- Decision Tree



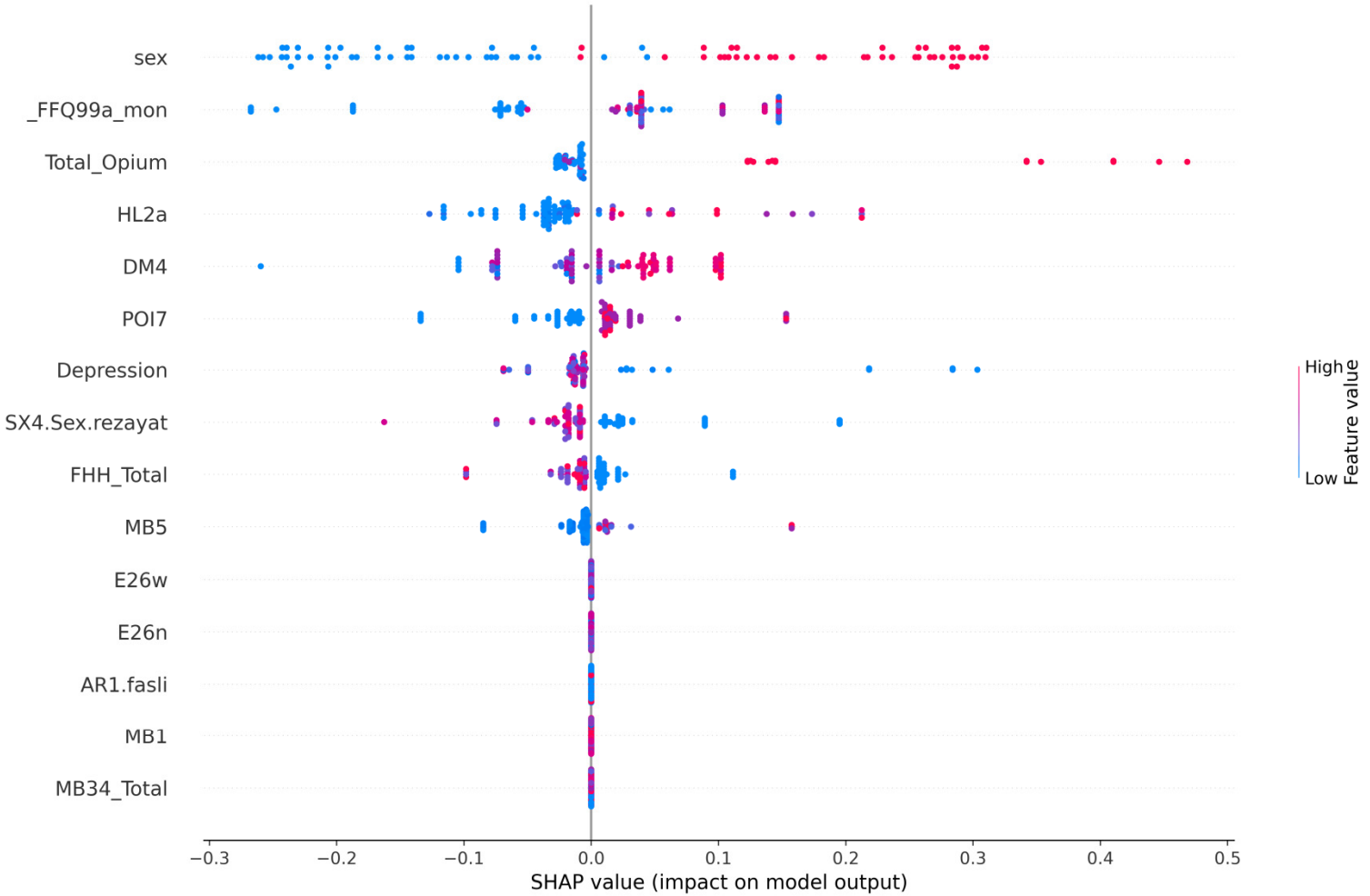
Importance of Features

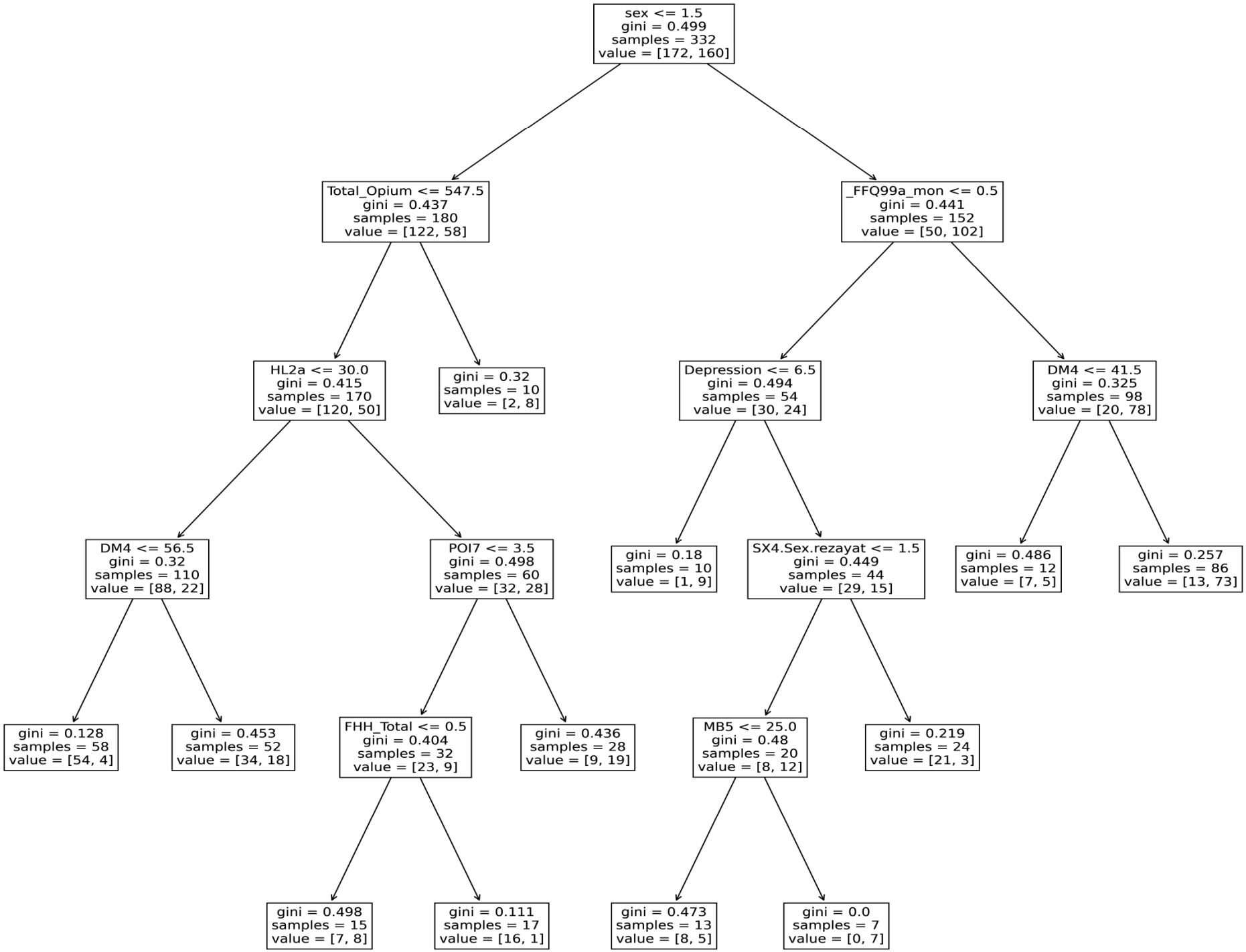
❑SHAP Analysis

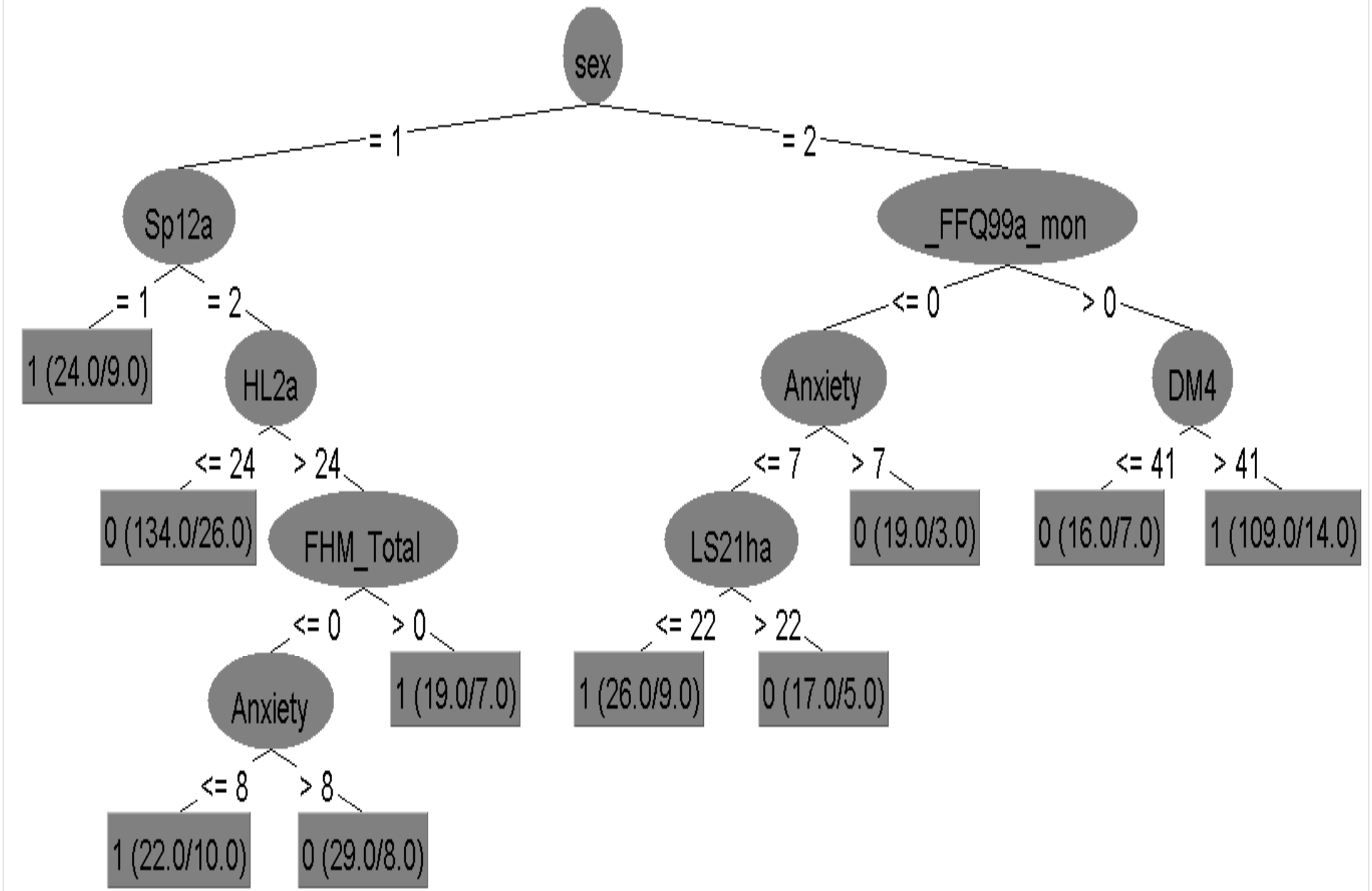


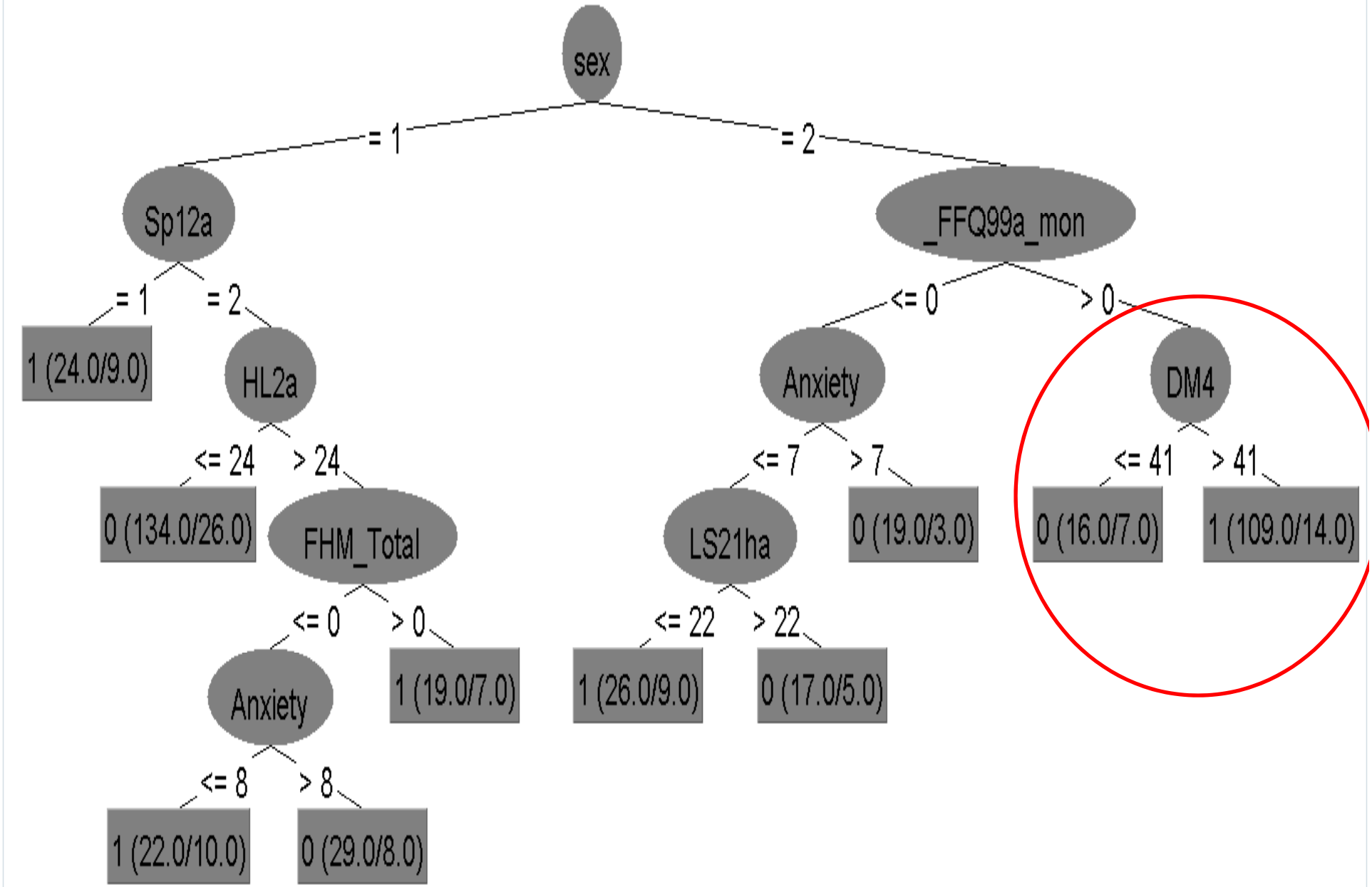
Importance of Features

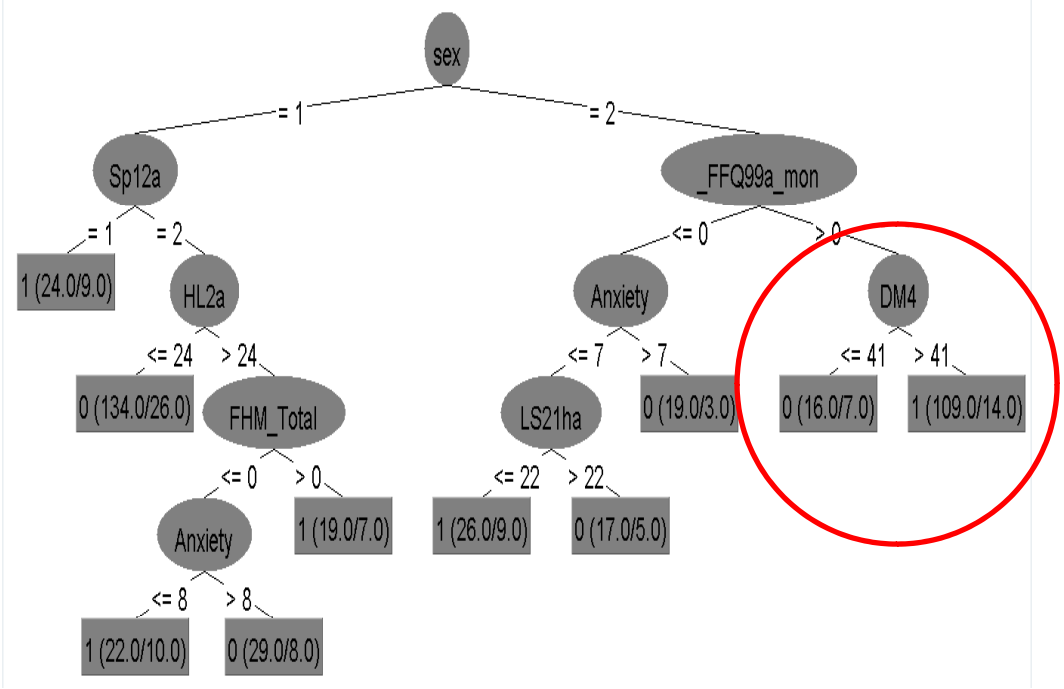
❑SHAP Analysis




sex <= 1.5
gini = 0.499
samples = 332
value = [172, 160]
Total_Opium <= 547.5
gini = 0.437
samples = 180
value = [122, 58]
_FFQ99a_mon <= 0.5
gini = 0.441
samples = 152
value = [50, 102]
HL2a <= 30.0
gini = 0.415
samples = 170
value = [120, 50]
gini = 0.32
samples = 10
value = [2, 8]
Depression <= 6.5
gini = 0.494
samples = 54
value = [30, 24]
DM4 <= 41.5
gini = 0.325
samples = 98
value = [20, 78]
DM4 <= 56.5
gini = 0.32
samples = 110
value = [88, 22]
POI7 <= 3.5
gini = 0.498
samples = 60
value = [32, 28]
gini = 0.18
samples = 10
value = [1, 9]
SX4.Sex.rezayat <= 1.5
gini = 0.449
samples = 44
value = [29, 15]
gini = 0.486
samples = 12
value = [7, 5]
gini = 0.257
samples = 86
value = [13, 73]
gini = 0.128
samples = 58
value = [54, 4]
gini = 0.453
samples = 52
value = [34, 18]
FHH_Total <= 0.5
gini = 0.404
samples = 32
value = [23, 9]
gini = 0.436
samples = 28
value = [9, 19]
MB5 <= 25.0
gini = 0.48
samples = 20
value = [8, 12]
gini = 0.219
samples = 24
value = [21, 3]
gini = 0.498
samples = 15
value = [7, 8]
gini = 0.111
samples = 17
value = [16, 1]
gini = 0.473
samples = 13
value = [8, 5]
gini = 0.0
samples = 7
value = [0, 7]

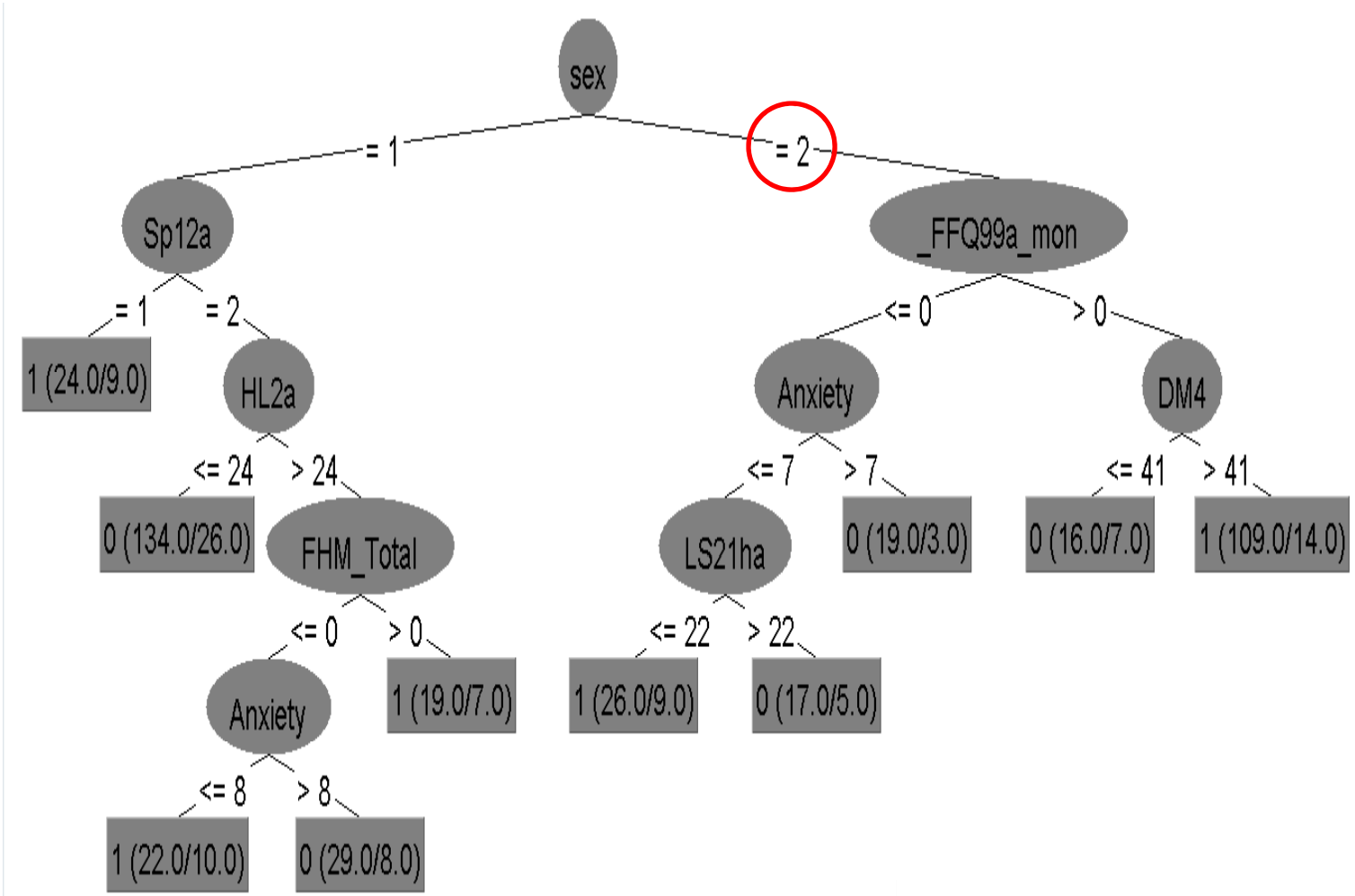

sex
= 1
= 2
Sp12a
_FFQ99a_mon
= 1
= 2
<= 0
> 0
1 (24.0/9.0)
HL2a
Anxiety
DM4
<= 24
> 24
<= 7
> 7
<= 41
> 41
0 (134.0/26.0)
FHM_Total
LS21ha
0 (19.0/3.0)
0 (16.0/7.0)
1 (109.0/14.0)
<= 0
> 0
<= 22
> 22
Anxiety
1 (19.0/7.0)
1 (26.0/9.0)
0 (17.0/5.0)
<= 8
> 8
1 (22.0/10.0)
0 (29.0/8.0)


sex
= 1
= 2
Sp12a
_FFQ99a_mon
= 1
= 2
<= 0
> 0
1 (24.0/9.0)
HL2a
Anxiety
DM4
<= 24
> 24
<= 7
> 7
<= 41
> 41
0 (134.0/26.0)
FHM_Total
LS21ha
0 (19.0/3.0)
0 (16.0/7.0)
1 (109.0/14.0)
<= 0
> 0
<= 22
> 22
Anxiety
1 (19.0/7.0)
1 (26.0/9.0)
0 (17.0/5.0)
<= 8
> 8
1 (22.0/10.0)
0 (29.0/8.0)

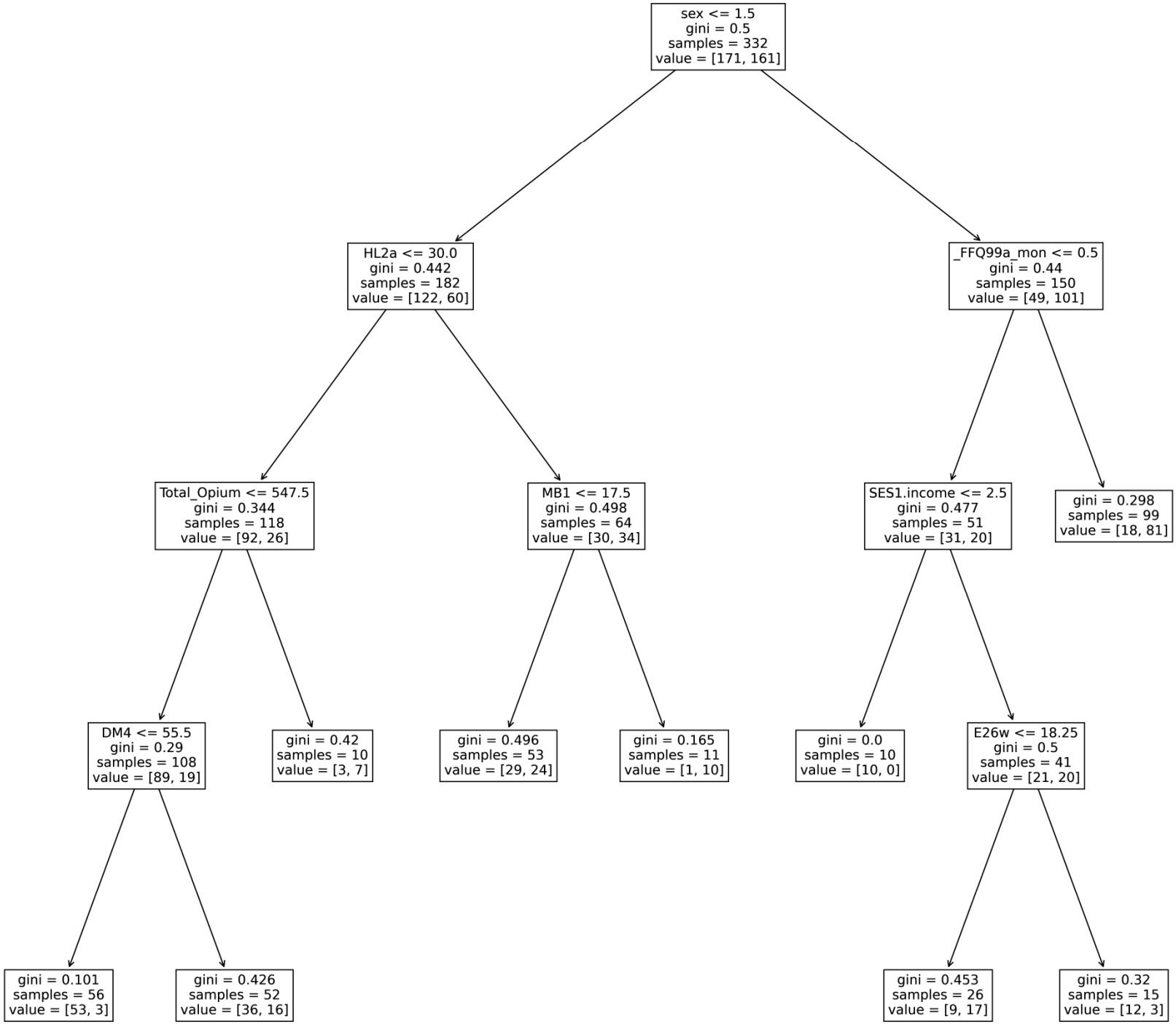


?

	ttest	p_value
DM4	-1.327923	1.849371e-01
Problem engagement	-1.097542	2.730439e-01
itive re-interpretation and growth	-1.880270	6.077483e-02
Anxiety	3.455493	6.061906e-04
Depression	2.135237	3.332944e-02
E26w	-4.201707	3.248494e-05
E26n	-5.297011	1.918409e-07
MB1	-2.023201	4.369613e-02
MB5	0.709169	4.786196e-01
MB34_Total	-4.802218	2.198318e-06
SES221	-0.904293	3.663675e-01
LS21ha	-1.684921	9.275917e-02
AR1.fasli	2.957579	3.278753e-03
LS29.Sleep.Quality	-0.633098	5.270202e-01
Total_Opium	-4.091137	5.162854e-05
Sp18a	-4.127003	4.447430e-05
HL2a	-0.870769	3.843864e-01
HD2a	-1.712770	8.750539e-02
HH2a	0.156492	8.757217e-01
FHL_Total	-0.001056	9.991579e-01
FHD_Total	0.536827	5.916764e-01
FHH_Total	0.668681	5.040729e-01
FHS_Total	0.898206	3.695988e-01
FHM_Total	-2.998210	2.879817e-03
_FFQ99a_mon	-2.023890	4.362480e-02
_FFQ100_mon	-1.704206	8.909479e-02
@_FFQ103	-2.089698	3.725624e-02
@_FFQ104	-2.119288	3.466229e-02
@_FFQ105	-1.464332	1.438640e-01
BMI	1.173332	2.413388e-01

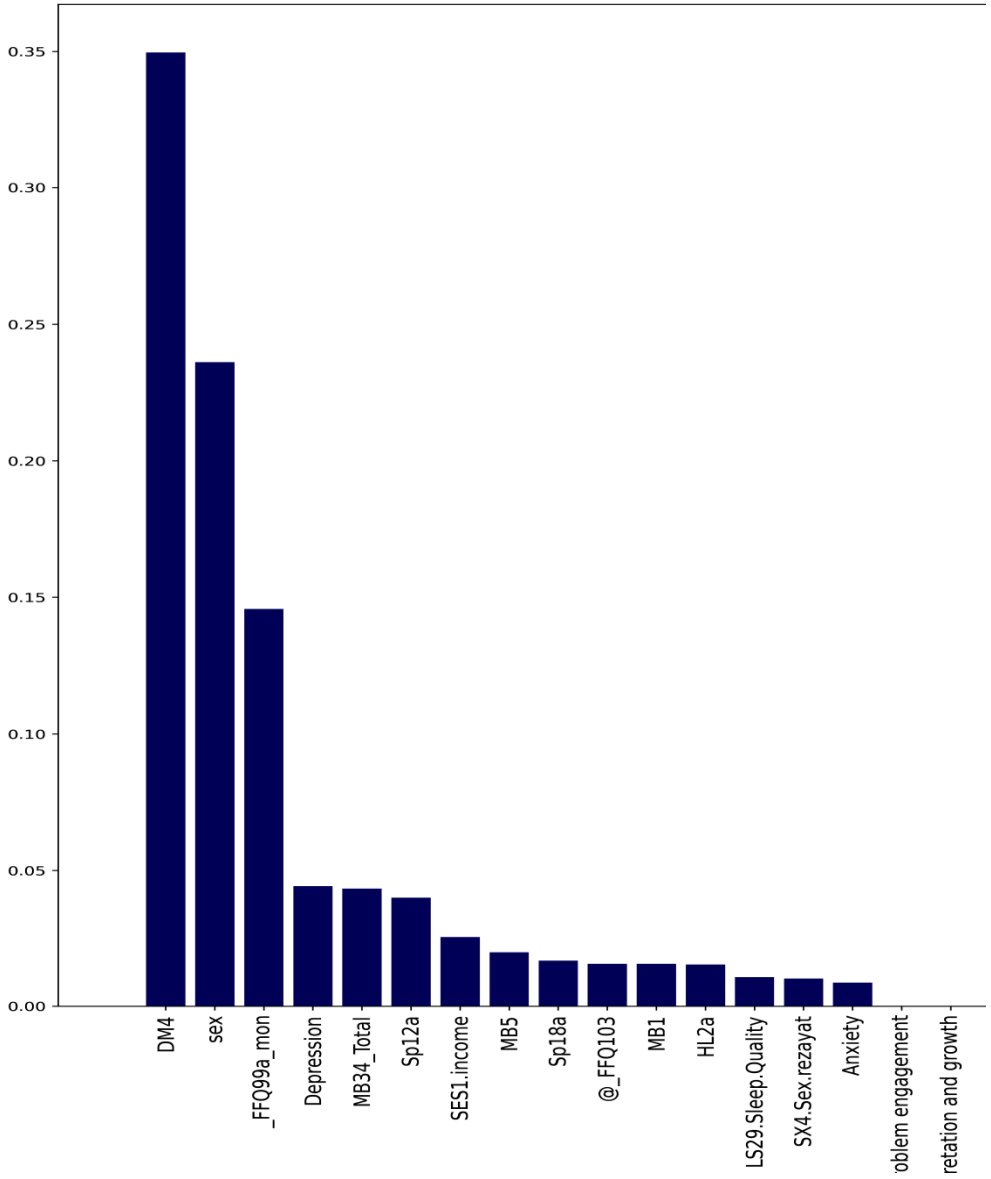


	ttest	p_value
DM4	-3.624673	0.000422
MB5	4.317094	0.000032
_FFQ99a_mon	3.177573	0.001878


sex <= 1.5
gini = 0.5
samples = 332
value = [171, 161]
HL2a <= 30.0
gini = 0.442
samples = 182
value = [122, 60]
_FFQ99a_mon <= 0.5
gini = 0.44
samples = 150
value = [49, 101]
Total_Opium <= 547.5
gini = 0.344
samples = 118
value = [92, 26]
MB1 <= 17.5
gini = 0.498
samples = 64
value = [30, 34]
SES1.income <= 2.5
gini = 0.477
samples = 51
value = [31, 20]
gini = 0.298
samples = 99
value = [18, 81]
DM4 <= 55.5
gini = 0.29
samples = 108
value = [89, 19]
gini = 0.42
samples = 10
value = [3, 7]
gini = 0.496
samples = 53
value = [29, 24]
gini = 0.165
samples = 11
value = [1, 10]
gini = 0.0
samples = 10
value = [10, 0]
E26w <= 18.25
gini = 0.5
samples = 41
value = [21, 20]
gini = 0.101
samples = 56
value = [53, 3]
gini = 0.426
samples = 52
value = [36, 16]
gini = 0.453
samples = 26
value = [9, 17]
gini = 0.32
samples = 15
value = [12, 3]

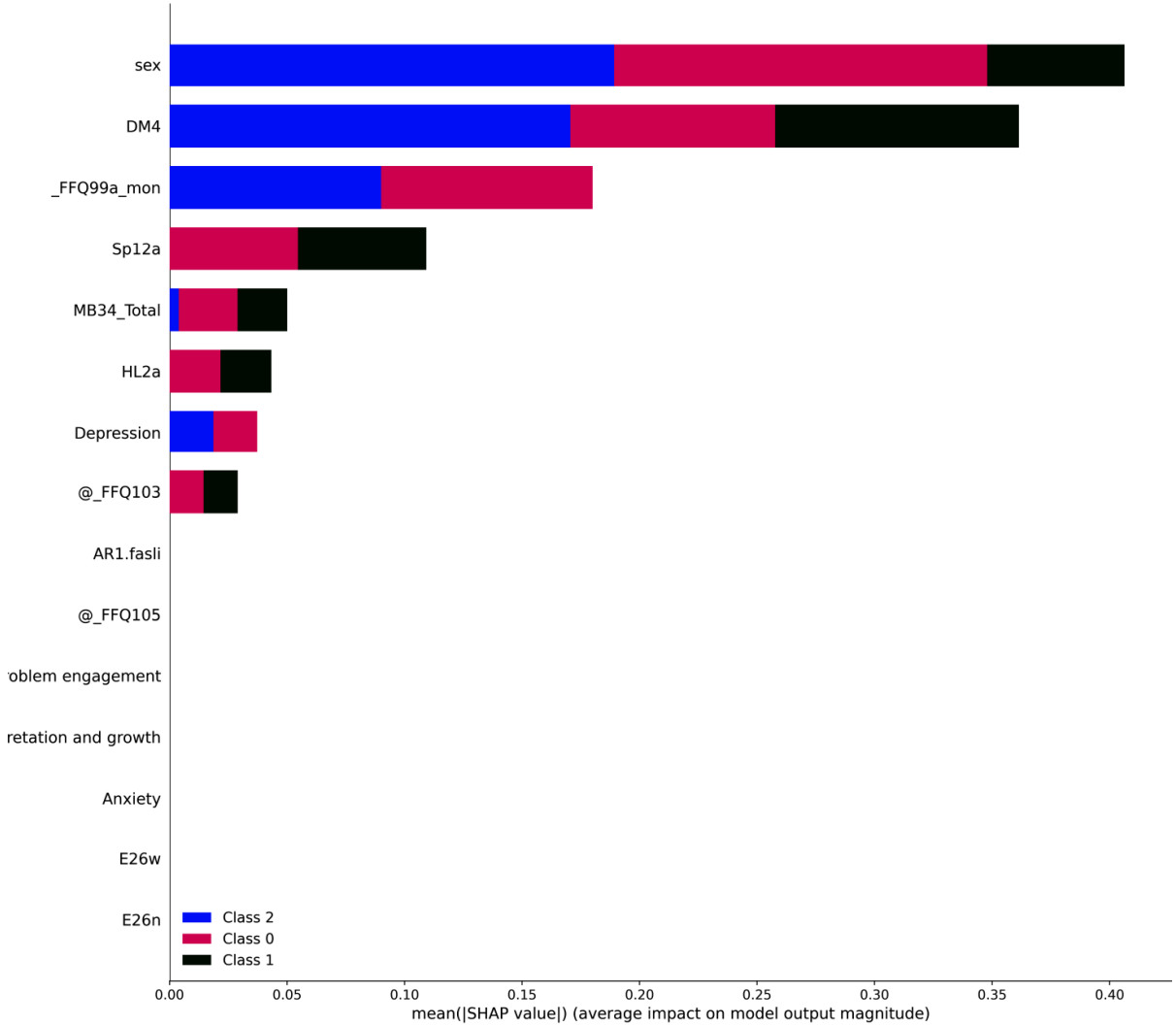
Importance of Features

- Decision Tree
- 3 Classes

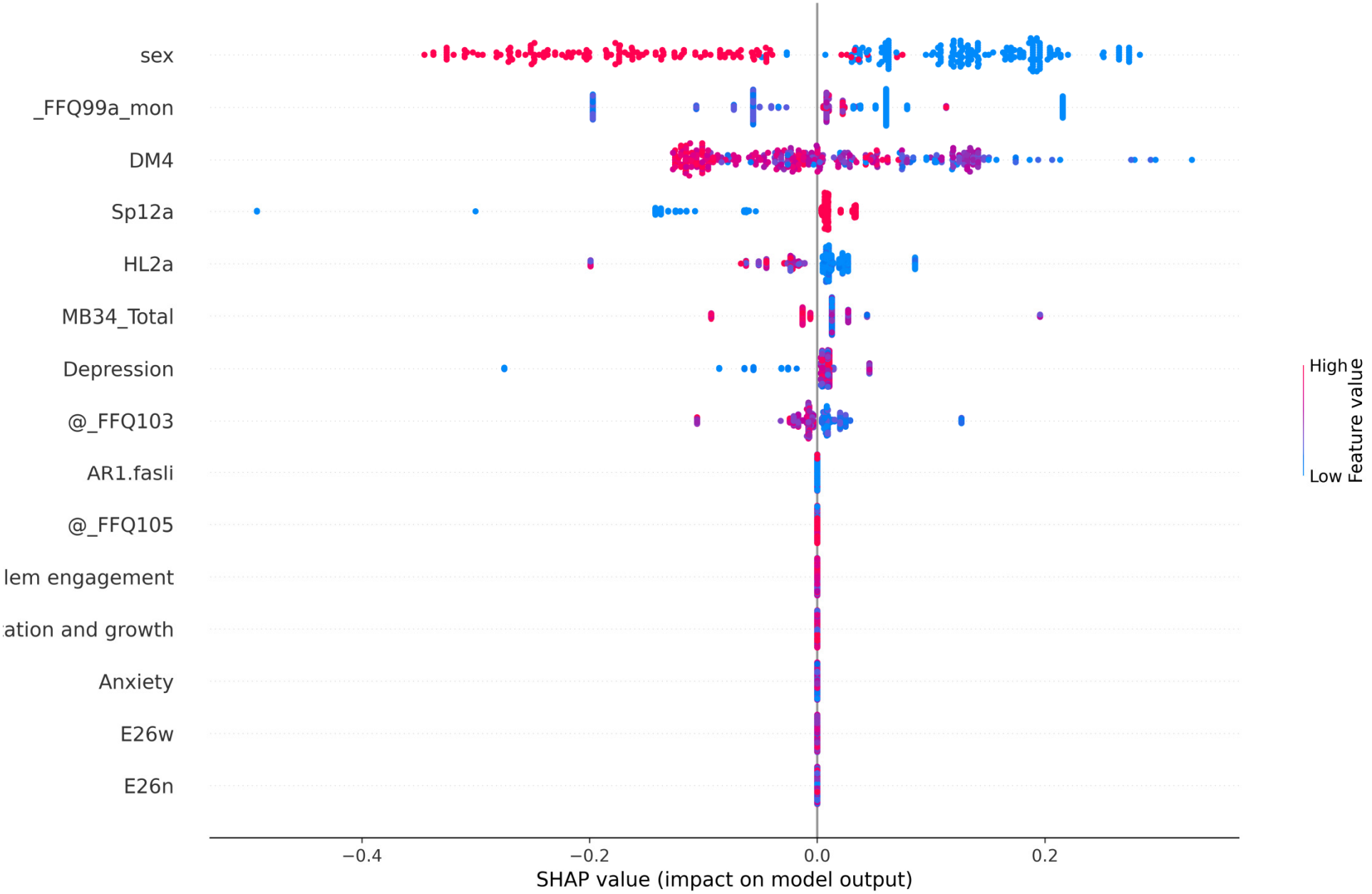


Importance of Features

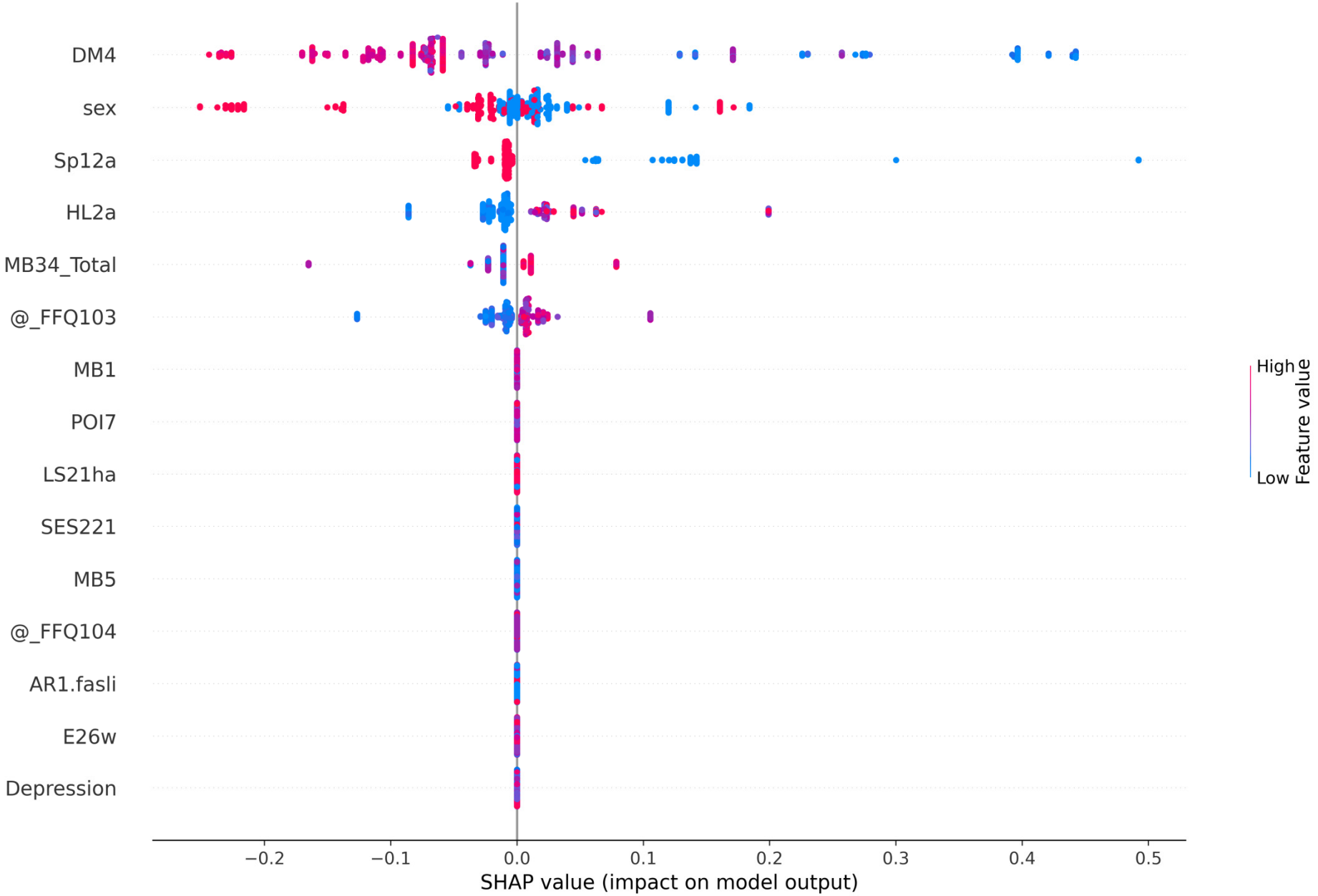
- SHAP Analysis
- 3 Classes



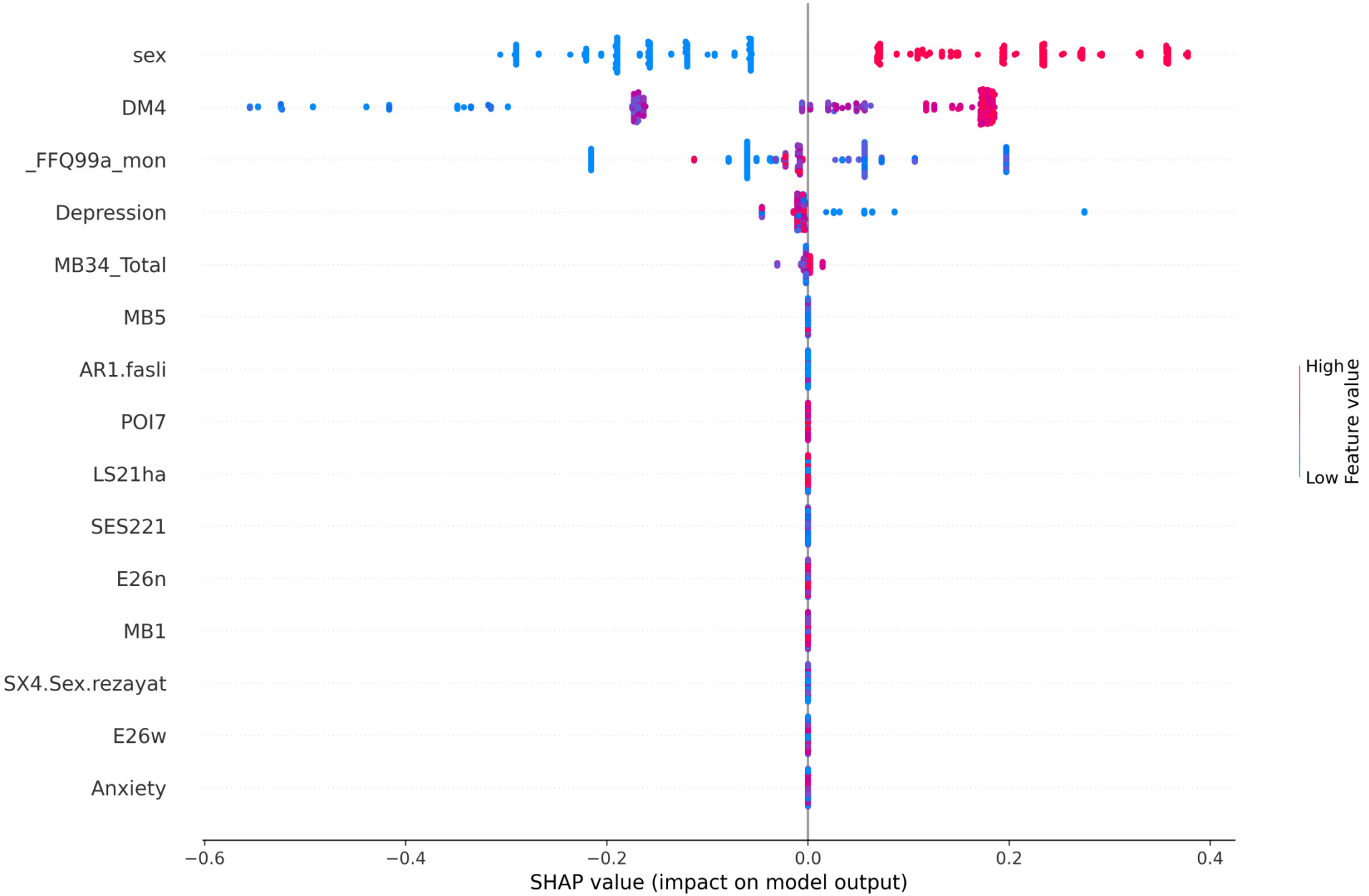
❑Class 0

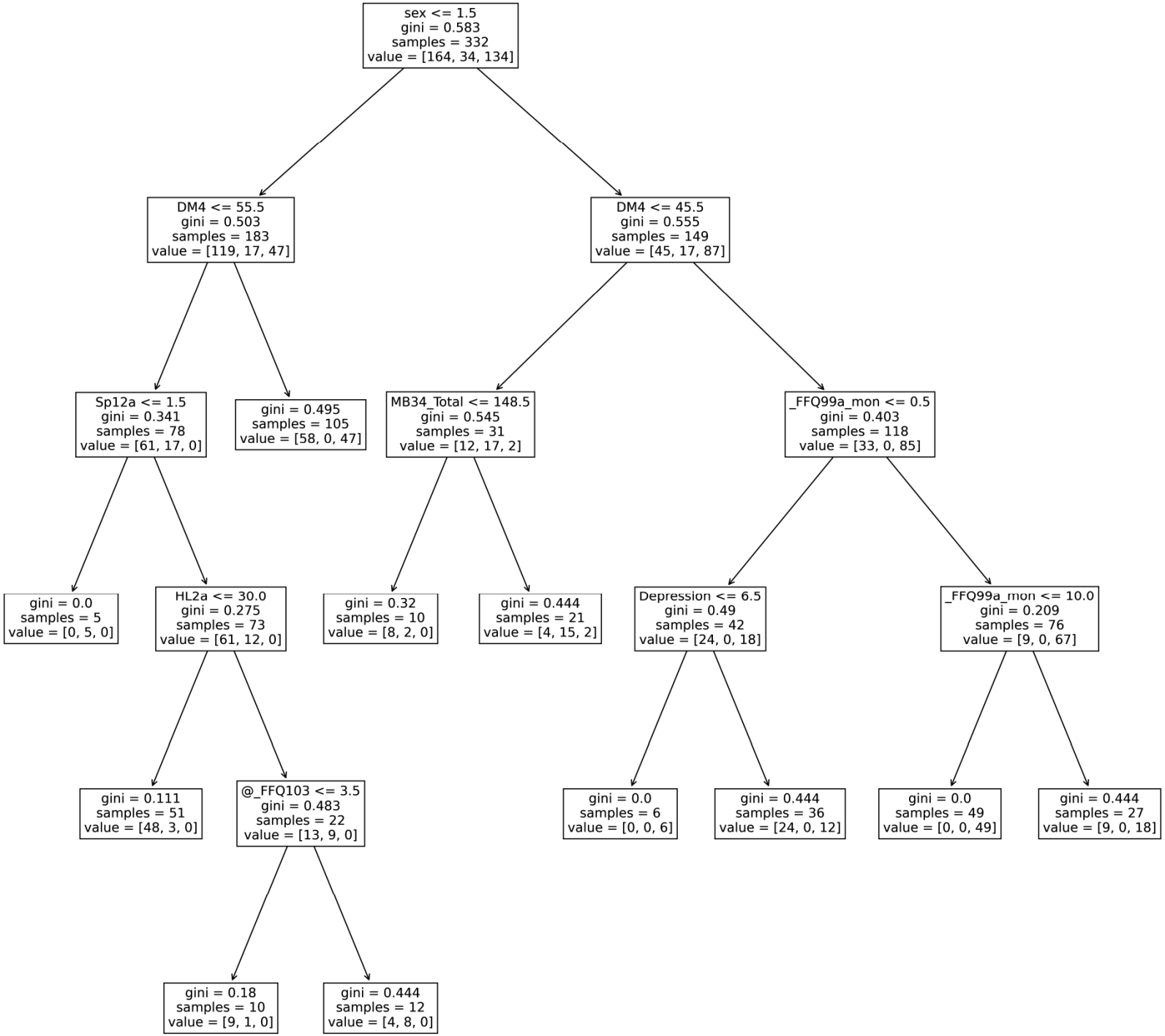

sex
_FFQ99a_mon
DM4
Sp12a
HL2a
MB34_Total
Depression
@_FFQ103
AR1.fasli
@_FFQ105
lem engagement
ation and growth
Anxiety
E26w
E26n
−0.4
−0.2
0.0
0.2
SHAP value (impact on model output)
High
Low
Feature value

❑Class 1


DM4
sex
Sp12a
HL2a
MB34_Total
@_FFQ103
MB1
POI7
LS21ha
SES221
MB5
@_FFQ104
AR1.fasli
E26w
Depression
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
SHAP value (impact on model output)
High
Low
Feature value

❑Class 2


sex
DM4
_FFQ99a_mon
Depression
MB34_Total
MB5
AR1.fasli
POI7
LS21ha
SES221
E26n
MB1
SX4.Sex.rezayat
E26w
Anxiety
−0.6
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Low
Feature value

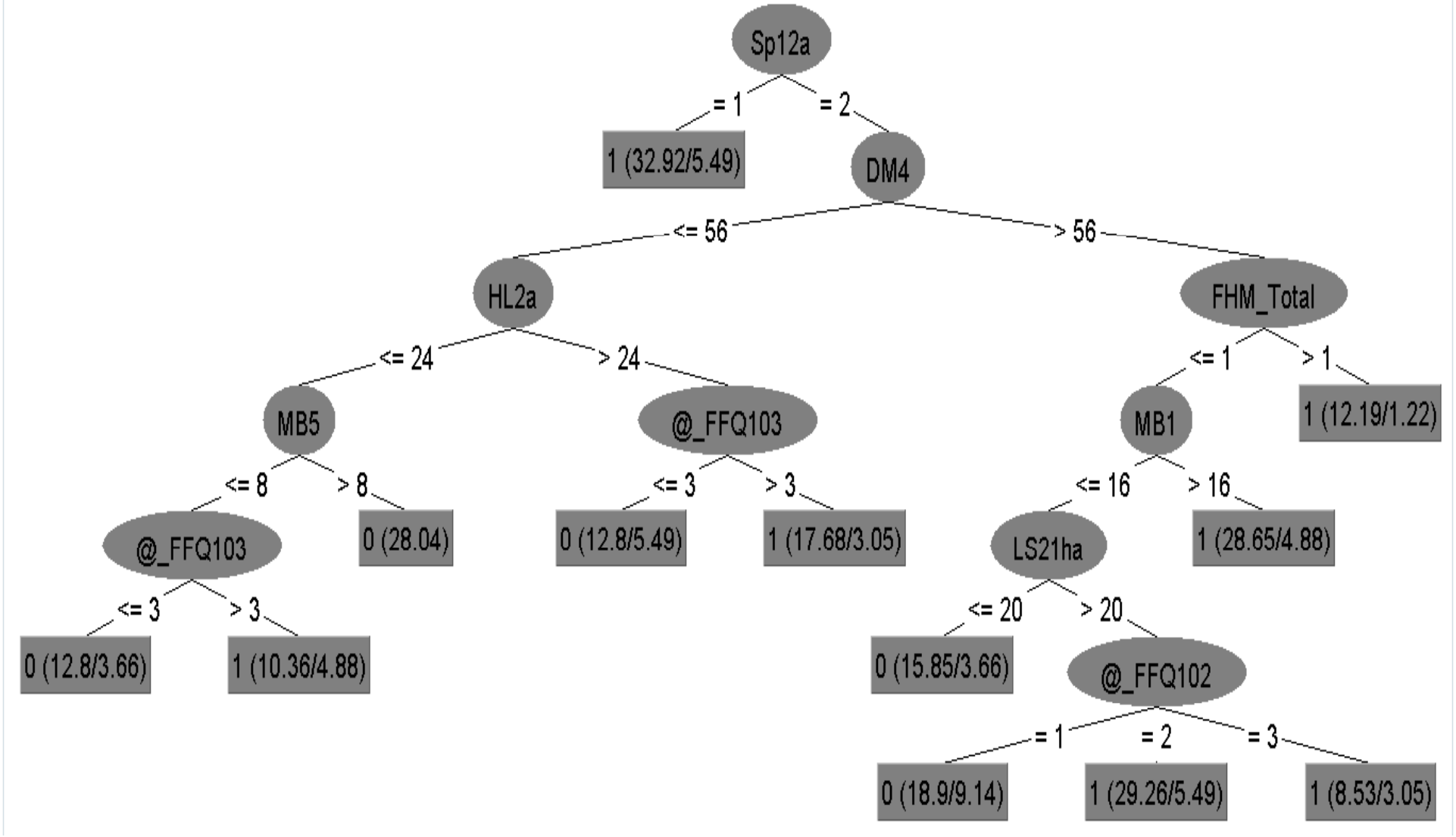

sex <= 1.5
gini = 0.583
samples = 332
value = [164, 34, 134]
DM4 <= 55.5
gini = 0.503
samples = 183
value = [119, 17, 47]
DM4 <= 45.5
gini = 0.555
samples = 149
value = [45, 17, 87]
Sp12a <= 1.5
gini = 0.341
samples = 78
value = [61, 17, 0]
gini = 0.495
samples = 105
value = [58, 0, 47]
MB34_Total <= 148.5
gini = 0.545
samples = 31
value = [12, 17, 2]
_FFQ99a_mon <= 0.5
gini = 0.403
samples = 118
value = [33, 0, 85]
gini = 0.0
samples = 5
value = [0, 5, 0]
HL2a <= 30.0
gini = 0.275
samples = 73
value = [61, 12, 0]
gini = 0.32
samples = 10
value = [8, 2, 0]
gini = 0.444
samples = 21
value = [4, 15, 2]
Depression <= 6.5
gini = 0.49
samples = 42
value = [24, 0, 18]
_FFQ99a_mon <= 10.0
gini = 0.209
samples = 76
value = [9, 0, 67]
gini = 0.111
samples = 51
value = [48, 3, 0]
@_FFQ103 <= 3.5
gini = 0.483
samples = 22
value = [13, 9, 0]
gini = 0.0
samples = 6
value = [0, 0, 6]
gini = 0.444
samples = 36
value = [24, 0, 12]
gini = 0.0
samples = 49
value = [0, 0, 49]
gini = 0.444
samples = 27
value = [9, 0, 18]
gini = 0.18
samples = 10
value = [9, 1, 0]
gini = 0.444
samples = 12
value = [4, 8, 0]

Features' Average Comparison

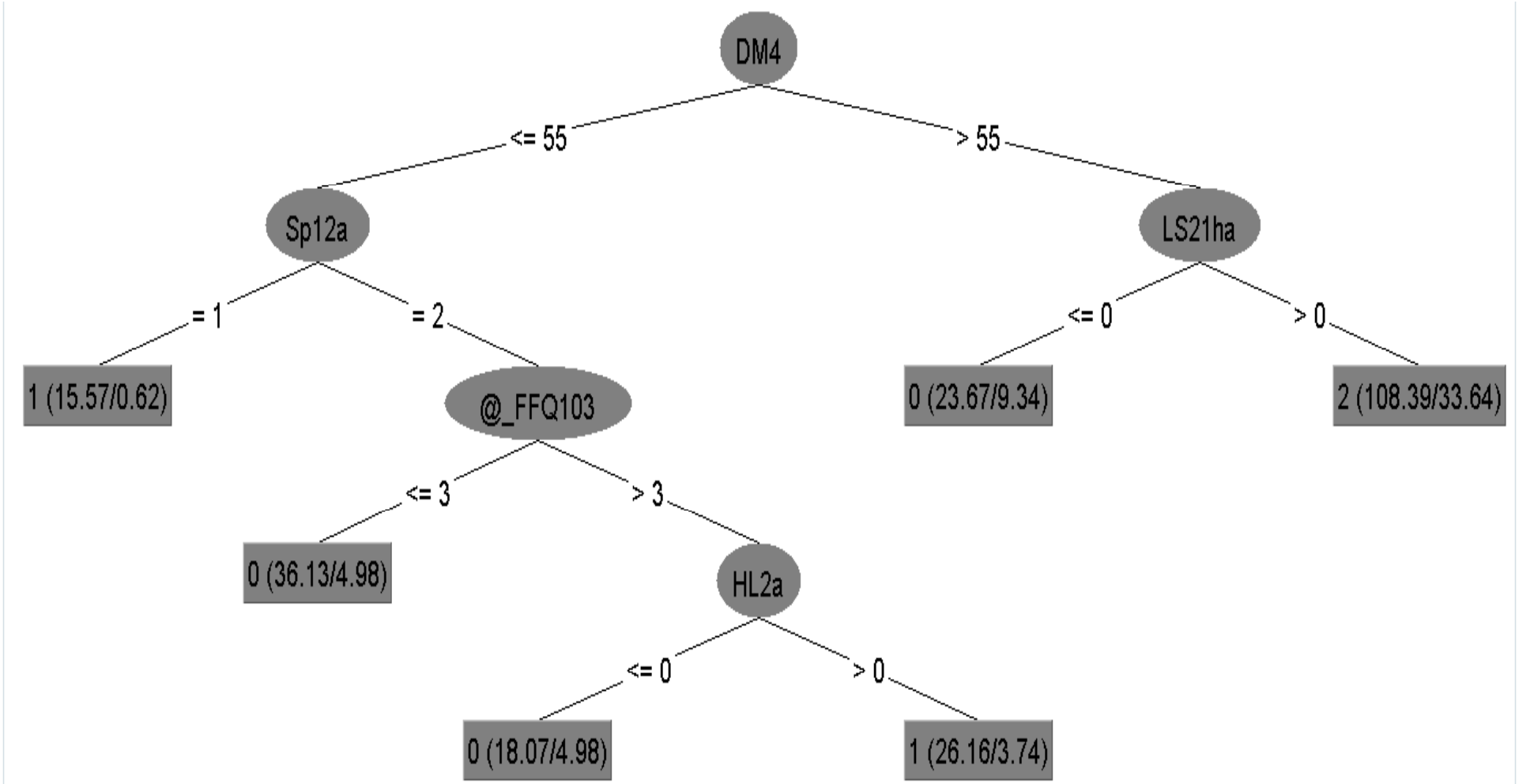
Feature Name	Women				Men		
	CAD 0 (60)	CAD 1 (20)	CAD 2 (107)		CAD 0 (155)	CAD 1 (19)	CAD 2 (54)
DM4	49.1	41.5	54.19		55.36	49.89	61.98
Problem engagement	9.17	9.6	9.44		8.94	9.42	8.63
Positive re-interpretation and growth	6.73	6.45	6.95		6.49	6.68	6.70
Anxiety	7.08	7.4	6.24		8.97	9.32	8.20
Depression	10.68	10.15	10.06		11.89	12.53	11.76
E26w	18.18	17.83	17.99		16.54	16.74	16.74
E26n	37.91	38.68	38.64		34.60	34.82	34.71
MB1	12.27	14.7	13.53		12.30	11.05	12.93
MB5	44.17	70.75	24.40		23.94	17.16	14.76
MB34_Total	143.73	178.5	158.83		75.98	72.68	67.37
AR1.fasli	0.42	0.25	0.26		0.64	0.21	0.30

Sp12a	1.67	1.65	1.56		1.94	1.68	1.83
Total_Opium	219	190.3	252.96		34.33	211.37	109.11
Sp18a	5.5	6.2	8.04		1.37	3.05	2.56
HL2a	12.6	12	19.57		29.19	57.47	42.89
HD2a	7	3.6	24.11		25.39	58.11	38.44
HH2a	12	6	22.99		38.79	36	52
FHL_Total	0.67	0.75	0.82		0.76	0.68	0.57
FHD_Total	0.42	0.7	0.59		0.66	0.63	0.35
FHH_Total	0.85	0.85	0.96		1.12	1	1
FHS_Total	0.17	0.3	0.23		0.28	0.11	0.15
FHM_Total	0.43	0.9	0.64		0.37	0.47	0.52
_FFQ99a_mon	6.17	8.4	7.90		5.81	5.37	7.19
_FFQ100_mon	1.28	3.25	1.78		0.55	0.11	0.39
@_FFQ102	1.9	2	1.75		1.70	1.63	1.78
@_FFQ103	4.67	5	4.66		3.19	4.21	2.85
@_FFQ104	2.87	3	2.98		2.83	2.84	2.98
@_FFQ105	5.45	4.65	6.21		5.79	6.11	6.13
SES1.income	3.73	4.55	4.23		3.24	2.74	2.83
BMI	25.88	25.89	26.79		28.20	27.65	27.38

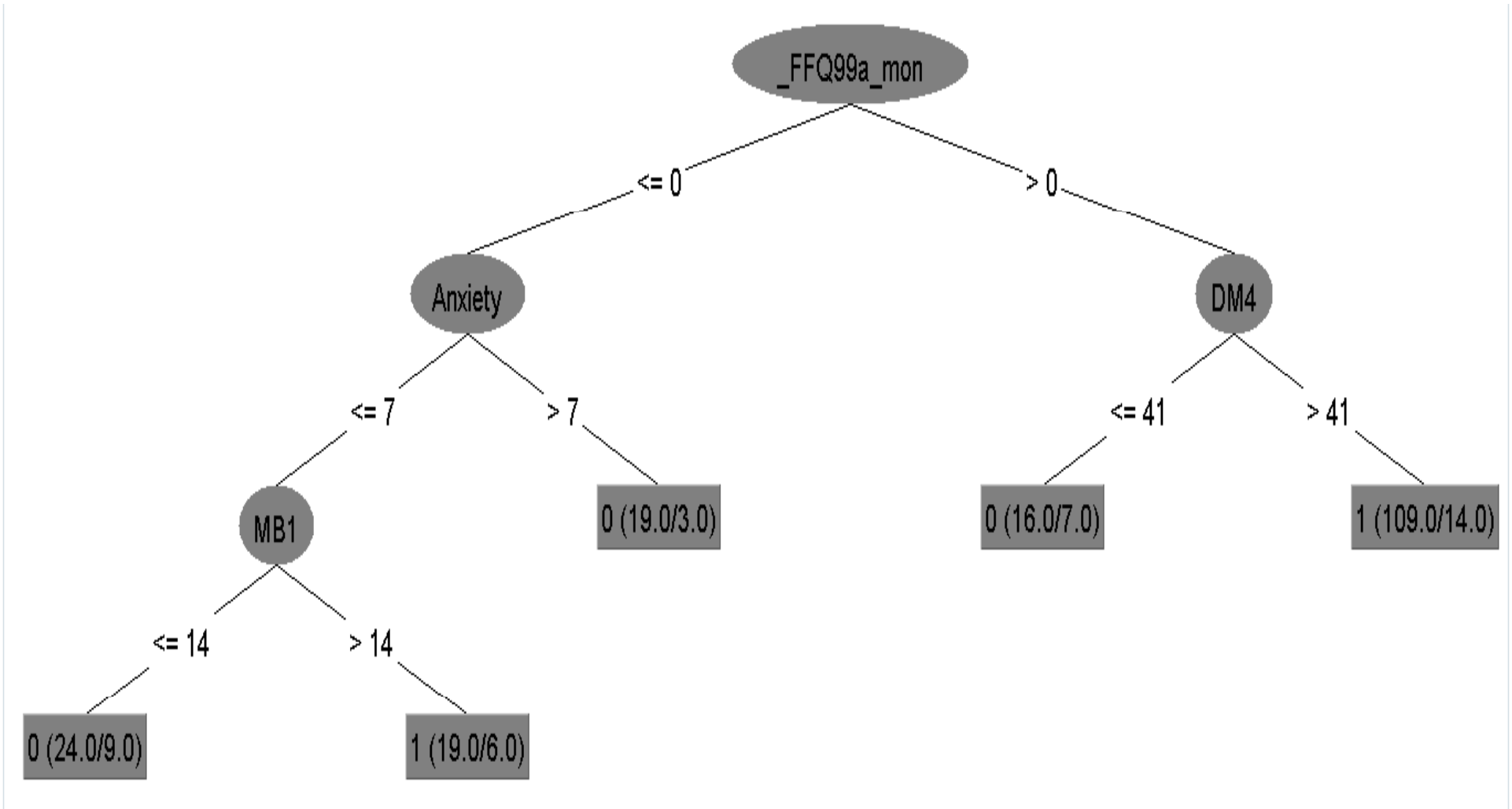
Men's Classification (2 C)



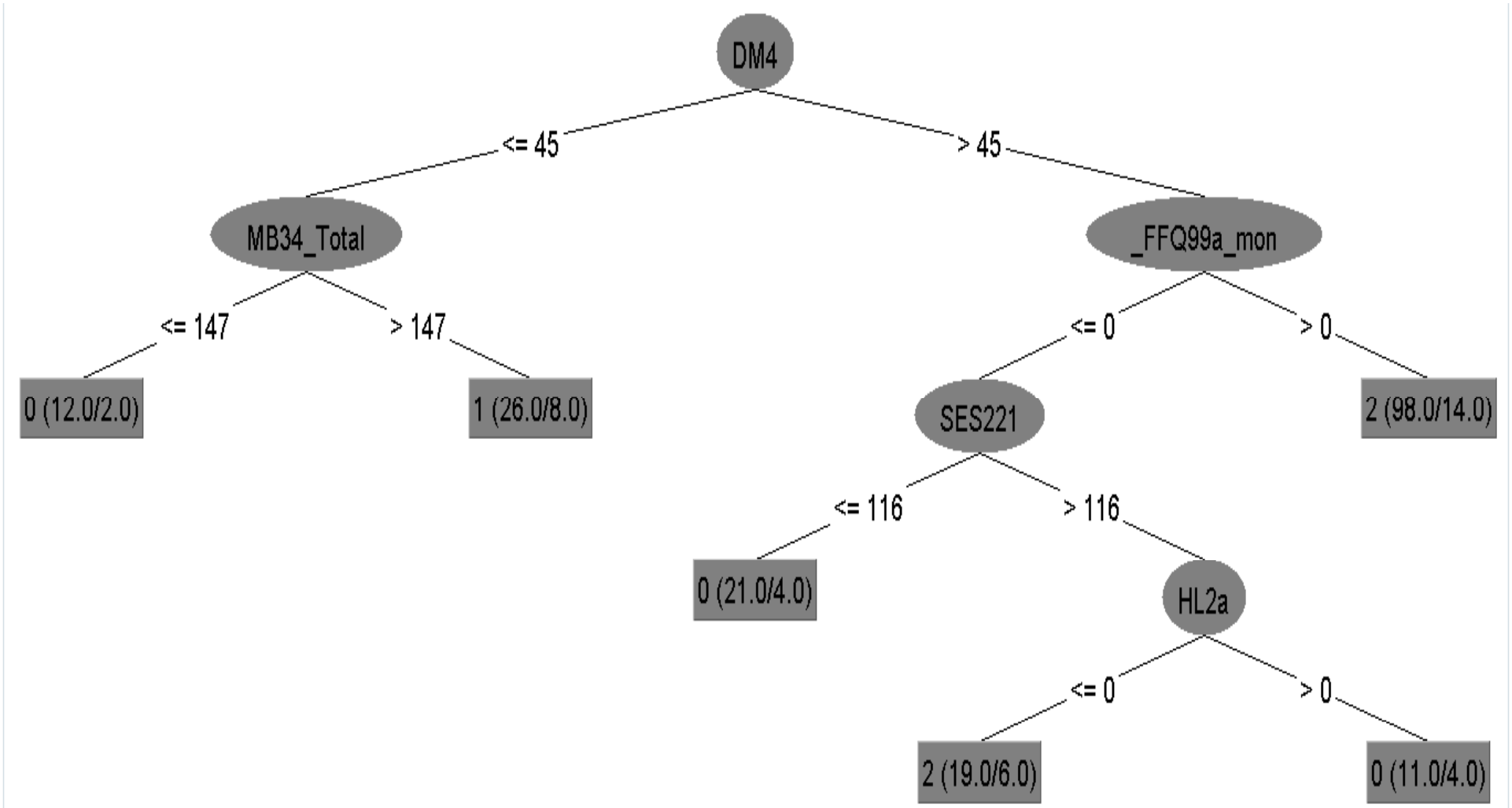
Men's Classification (3 C)

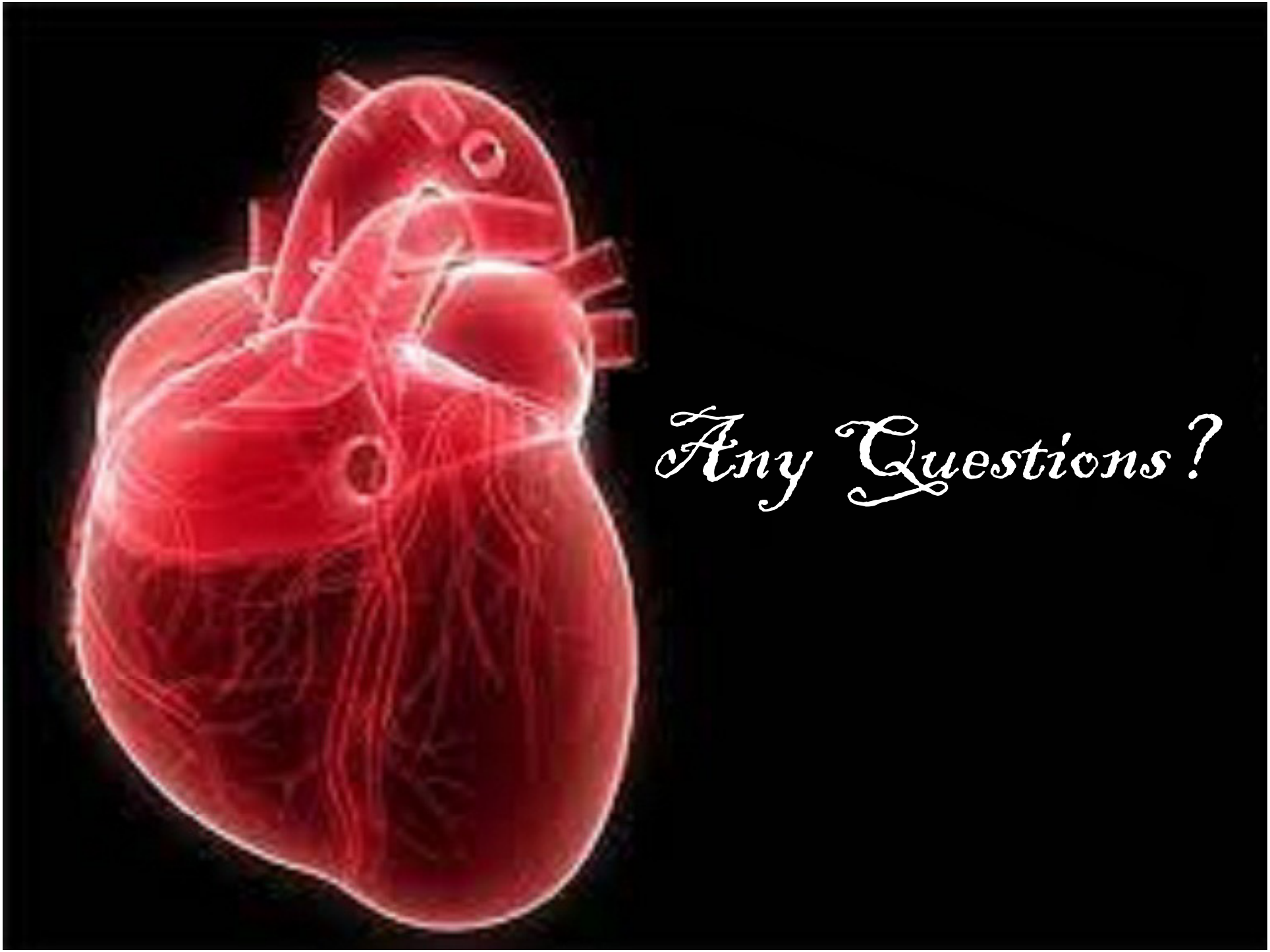


Women's Classification (2 C)



Women's Classification (3 C)




Any Questions?