Appendix A: Effect of Data Inseparability on Classification Accuracy

(a) Separable data leading to a well-fitted model   (b) Inseparable data leading to a poorly fitted model
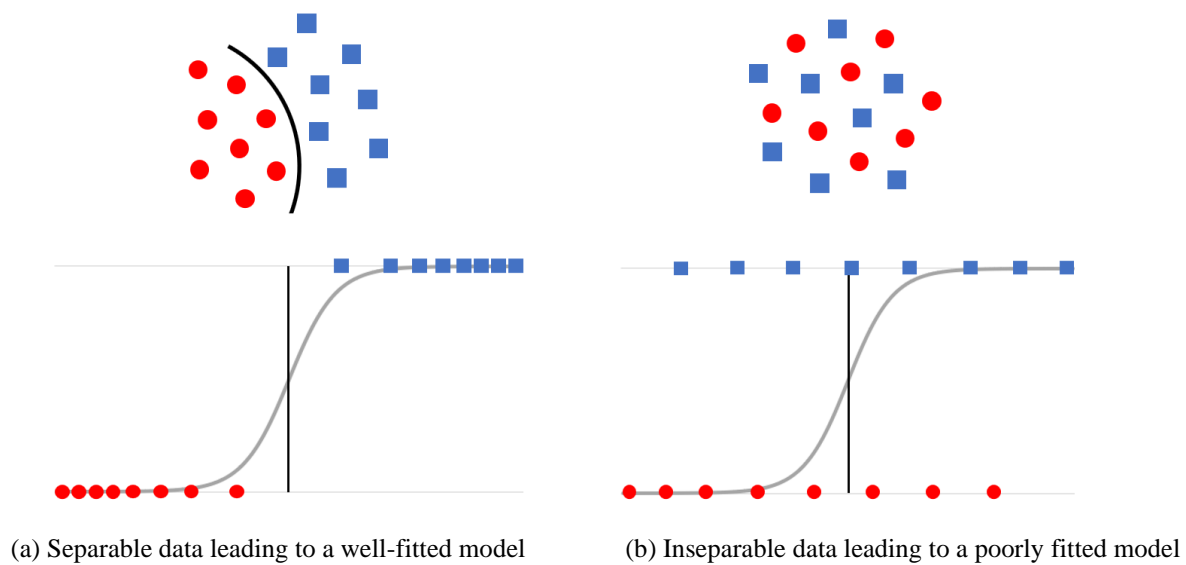
Figure A.1: Data inseparability and its effect on logistic regression performance

Appendix B: ORN Data Inseparability by DVH Features

The following figures show that the two classes of ORN-negative and ORN-positive patients in our data set are inseparable by D30% and pre-radiation dental extraction (PDE) status. Given only two variables, this could be shown on a single plot. However, for better visualization and providing more insight on the structure of the data, we illustrate the data inseparability in smaller groups; that is, for each pair of DVH cluster and PDE status separately. Evidently, if the whole data set is separable, it remains separable when divided in smaller groups. Also, for each patient, we have plotted D30% against the standard deviation (SD) of the delivered dose, for visualization purposes only, so all data points are not located on a horizontal line.
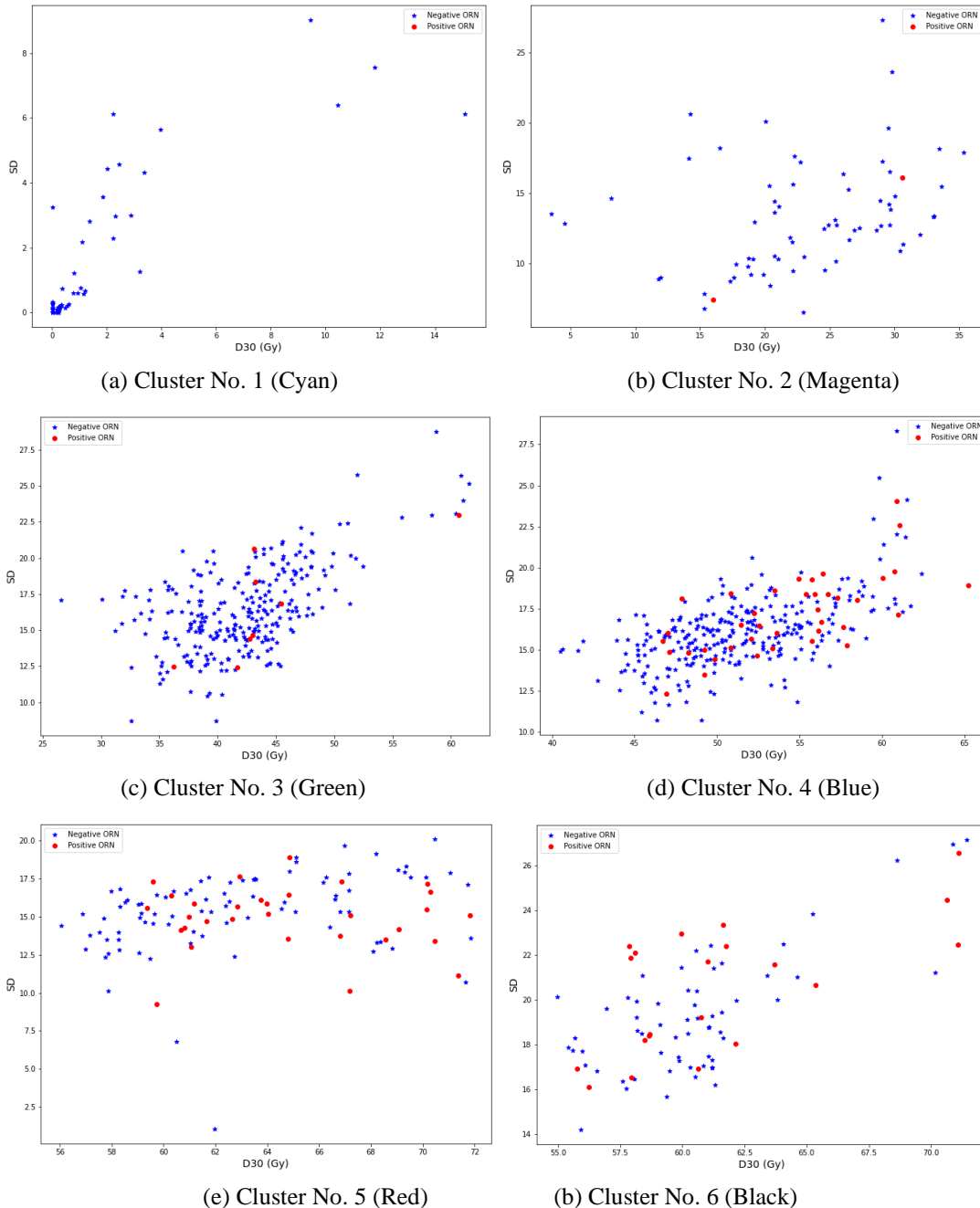

(a) Cluster No. 1 (Cyan)

(b) Cluster No. 2 (Magenta)

(c) Cluster No. 3 (Green)

(d) Cluster No. 4 (Blue)

(e) Cluster No. 5 (Red)

(b) Cluster No. 6 (Black)

Figure B.1: Data inseparability within the clusters for PDE = 0

(a) Cluster No. 1 (Cyan)

(b) Cluster No. 2 (Magenta)

(c) Cluster No. 3 (Green)

(d) Cluster No. 4 (Blue)

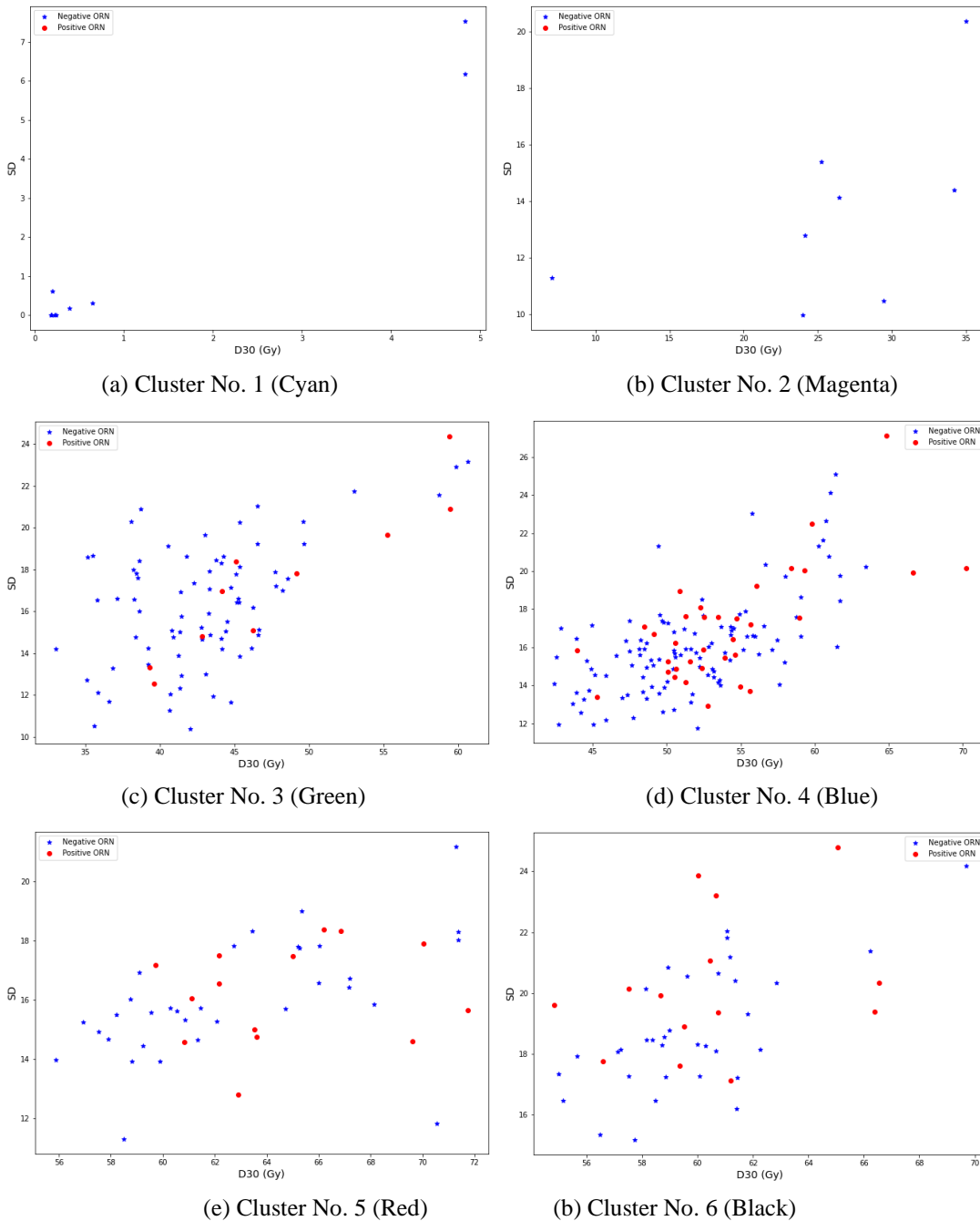(e) Cluster No. 5 (Red)

(b) Cluster No. 6 (Black)

Figure B.2: Data inseparability within the clusters for PDE = 1

Figures B.1 and B.2 show that, in all D30%-PDE categories, the class of ORN-negative patients is dominant, and the two classes are highly mixed (inseparable). As we have used SD (y axis) for better visualization, the data inseparability by D30% (for each PDE status) is implied by nonexistence of vertical separators of "Negative ORN" and "Positive ORN" points. However, the fact that these classes cannot be separated by other lines/curves implies that including SD as a feature is not expected to improve the performance of a logistic regression model. Finally, the existence of high (linear) correlations among Dx% features implies that replacing D30% with any other DVH feature (including the mean dose) will not improve the inseparability.
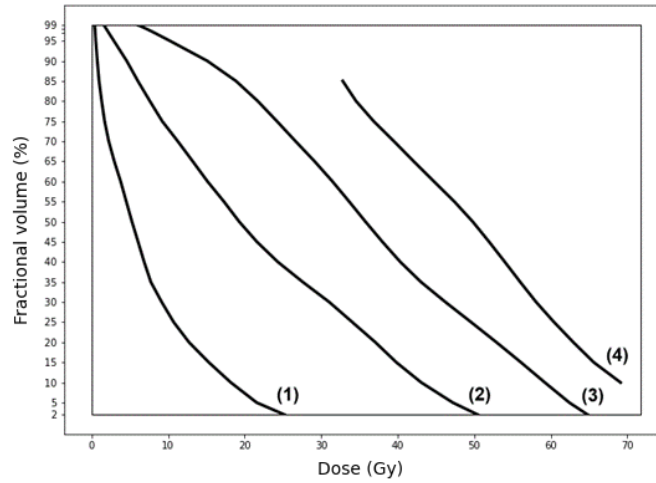
Appendix C: Coordinates of the Borderlines

Figure C.1: Borderline numbers for K = 5

Table C.1: Coordinates of the borderlines for K = 5 (Gy)

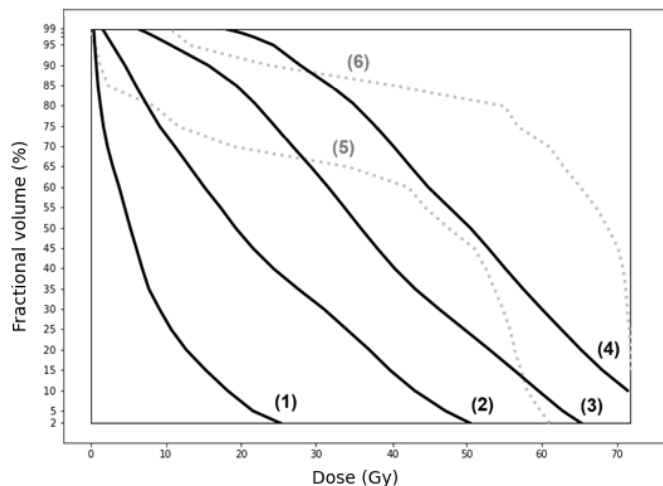| Line No.	(1)	(2)	(3)	(4)
D2%	25.19	50.46	64.83	-
D5%	21.56	47.19	62.43	-
D10%	18.18	43.08	59.24	69.14
D15%	15.31	39.83	56.12	65.64
D20%	12.72	37.08	52.91	62.97
D25%	10.72	34.07	49.60	60.45
D30%	9.15	31.06	46.24	58.10
D35%	7.73	27.59	43.04	56.02
D40%	6.85	24.32	40.34	54.00
D45%	6.06	21.60	38.00	51.95
D50%	5.26	19.28	35.79	49.81
D55%	4.52	17.29	33.66	47.41
D60%	3.79	15.15	31.48	44.75
D65%	2.95	13.21	29.14	42.10
D70%	2.23	11.26	26.65	39.51
D75%	1.69	9.22	24.21	36.86
D80%	1.30	7.60	21.71	34.56
D85%	0.96	6.04	18.86	32.82
D90%	0.73	4.58	15.11	-
D95%	0.53	2.90	10.22	-
D97%	0.46	2.22	8.20	-
D98%	0.43	1.89	7.14	-
D99%	0.38	1.57	5.96	-

Figure C.2: Borderline numbers for K = 6

Table C.2: Coordinates of the borderlines for K = 6 (Gy)

Line No.	(1)	(2)	(3)	(4)	(5)	(6)
D2%	25.19	50.44	65.23	80.98	61.04	73.73
D5%	21.56	47.18	62.78	75.94	59.83	72.81
D10%	18.18	43.07	59.51	71.41	58.04	72.39
D15%	15.31	39.83	56.33	68.08	57.24	72.01
D20%	12.72	37.08	53.08	65.25	56.39	71.81
D25%	10.72	34.07	49.74	62.63	55.80	71.64
D30%	9.15	31.06	46.37	60.04	54.83	71.43
D35%	7.73	27.59	43.17	57.54	53.71	71.18
D40%	6.85	24.32	40.46	55.16	52.56	70.84
D45%	6.06	21.6	38.12	52.90	51.04	70.15
D50%	5.26	19.28	35.91	50.51	47.57	68.81
D55%	4.52	17.29	33.78	47.73	44.54	67.25
D60%	3.79	15.14	31.62	44.93	42.28	65.17
D65%	2.95	13.19	29.33	42.60	34.40	62.97
D70%	2.23	11.22	26.91	40.33	19.13	61.01
D75%	1.69	9.18	24.60	37.93	11.67	56.82
D80%	1.30	7.57	22.21	35.27	8.46	55.04
D85%	0.96	6.02	19.42	31.87	2.32	40.46
D90%	0.73	4.56	15.66	27.91	1.24	23.82
D95%	0.53	2.90	10.67	24.33	0.64	13.17
D97%	0.46	2.22	8.62	21.71	0.02	11.89
D98%	0.43	1.89	7.55	20.09	0.01	11.22
D99%	0.38	1.57	6.34	18.03	0.01	10.50

Appendix D: Comparing Risk Indices for K = 5 and K = 6
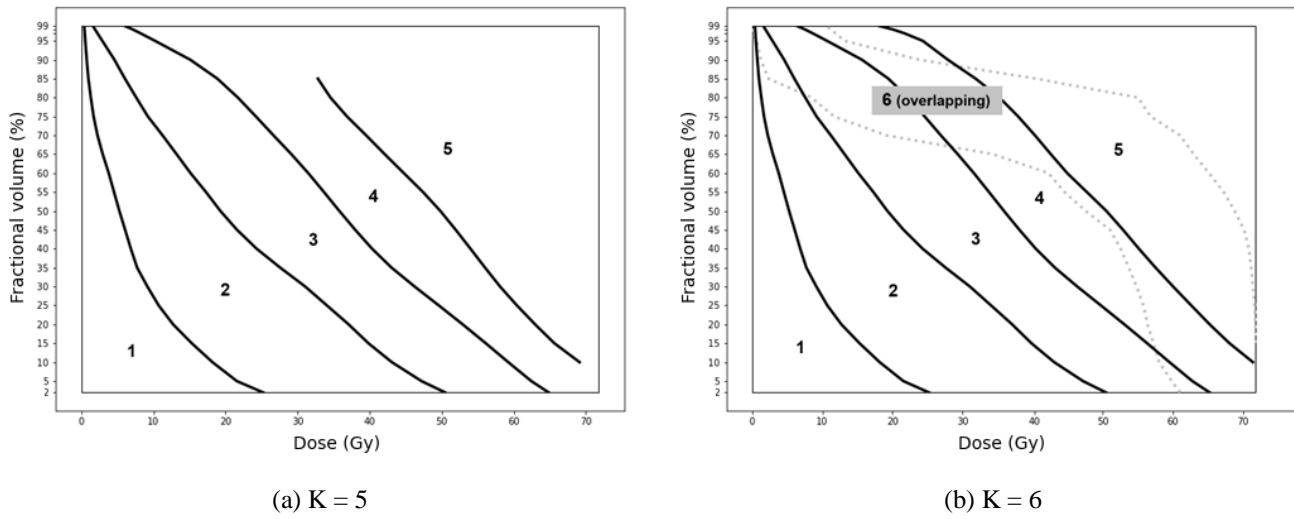

(a) K = 5   (b) K = 6

Figure D.1: Dose-volume partitioning with K = 5 and K = 6

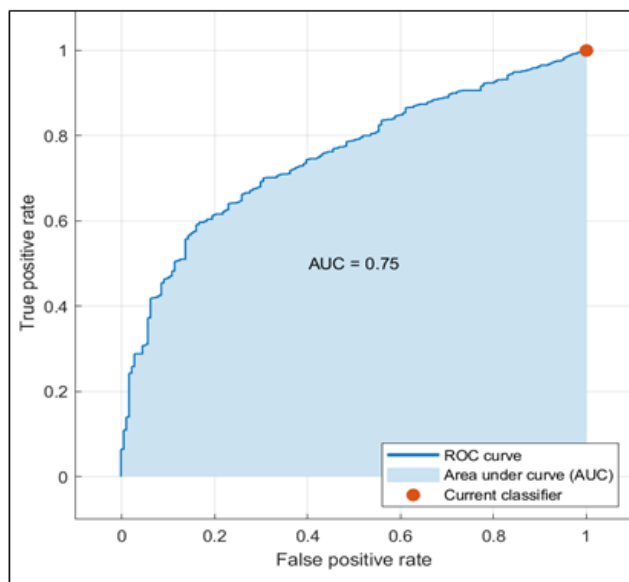Table D.1: Risk indices for K = 5 and K = 6 and their differences

Region	PDE = 0			PDE = 1		
No.	K = 5	K = 6	Δr	K = 5	K = 6	Δr
1	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
2	2.9 %	2.9 %	0.0 %	0.0 %	0.0 %	0.0 %
3	3.0 %	2.9 %	0.1 %	10.8 %	11.6 %	0.8 %
4	12.8 %	12.3 %	0.5 %	24.3 %	23.6 %	0.7 %
5	25.1 %	26.3 %	0.9 %	28.9 %	29.8 %	0.6 %
6	—	25.6 %		—	29.2 %	

PDE = Pre-radiation dental extraction status. Δr = Difference between risk indices for K = 5 and K = 6.

Appendix E: Logistic Regression vs. Naive Bayes
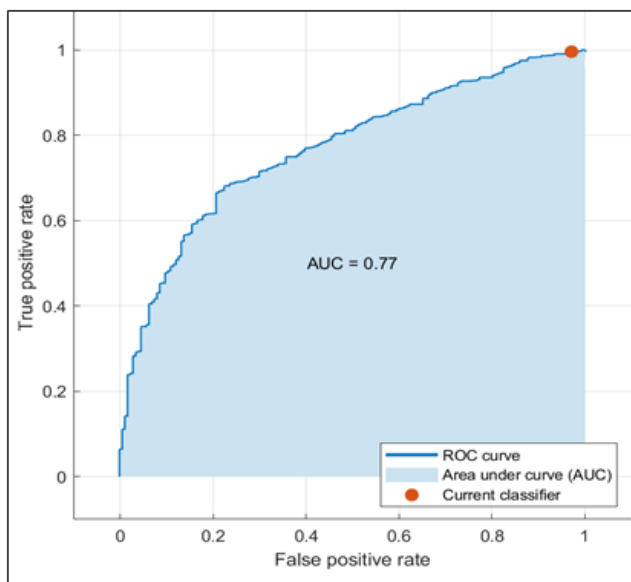
(a) Logistic Regression   (b) Naive Bayes

Figure E.1: Comparison of the performance of logistic regression and naive Bayes on the data (The features include D30% and PDE status for both models.)