Modelling – Dataset A (Dean)

Group K

2025-10-02

Table 1: First 10 rows of Dataset A (Health)

Dataset	SurveyYear	CharacteristicId	CharacteristicCategory	CharacteristicLabel	IndicatorId	IndicatorType	Value	DenominatorW
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_DOC	I	28.5	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_DOC	I	30.0	4122
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_DOC	I	27.3	2010
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NRS	I	66.6	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NRS	I	65.0	4122
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NRS	I	68.4	2010
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_TBA	I	0.1	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_OTH	I	0.6	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_OTH	I	0.7	4122
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_MIS	S	1.2	2871

Table 2: Filtered Health (I) – head(10)

Dataset	SurveyYear	CharacteristicId	CharacteristicCategory	CharacteristicLabel	IndicatorId	IndicatorType	Value	Denominator
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_DOC	I	28.5	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_DOC	I	30.0	4122
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_DOC	I	27.3	2010
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NRS	I	66.6	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NRS	I	65.0	4122
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NRS	I	68.4	2010
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_TBA	I	0.1	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_OTH	I	0.6	2871
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_OTH	I	0.7	4122
Access_To_Healthcare_01	1998	1000	Total	Total	RH_ANCP_W_NON	I	2.9	2871

Table 3: Class balance: HighValue (No/Yes)

Var1	Freq
No	82
Yes	82

Table 4: Train/Test split sizes

Split	Rows
Train	116
Test	48

pdf 2

Table 5: Logistic Regression – Overall Metrics (Test)

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.667	0.333	0.516	0.796	0.5	0.015	0.803

Table 6: Logistic Regression – Class Metrics (Test)

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
0.708	0.625	0.654	0.682	0.654	0.708	0.68	0.5	0.354	0.542	0.667

pdf 2 pdf 2

Table 7: Decision Tree – Overall Metrics (Test)

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.708	0.417	0.559	0.83	0.5	0.003	0.423

Table 8: Decision Tree – Class Metrics (Test)

Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision	Recall	F1	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
0.792	0.625	0.679	0.75	0.679	0.792	0.731	0.5	0.396	0.583	0.708

Table 9: Decision Tree – Top 10 Feature Importance

	Overall	Feature
IndicatorId	15.976	IndicatorId
DW_log	11.800	DW_log
SurveyYear	0.704	SurveyYear

Table 10: Model Performance Comparison (Test Set)

Model	Accuracy	Kappa	Sensitivity	Specificity	AUC
Logistic Regression	0.667	0.333	0.708	0.625	0.832
Decision Tree	0.708	0.417	0.792	0.625	0.786

Summary. We trained two models on Dataset A's health indicators (binary target: HighValue $\geq$ median(Value)). The table above shows test-set metrics. Based on AUC (primary) and Accuracy (secondary), the better model in this run is **Logistic Regression** (AUC = 0.832, Accuracy = 0.667). We also saved ROC curves, a tree plot, and CSV metrics in `../outputs/`.