Modelling – Dataset A (Dean)

Group K

2025-10-02

## Rows: 633 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (5): Dataset, CharacteristicCategory, CharacteristicLabel, IndicatorId, ...  
## dbl (5): SurveyYear, CharacteristicId, Value, DenominatorWeighted, Denominat...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

First few rows of Dataset A (Health)

| Dataset | SurveyYear | CharacteristicId | CharacteristicCategory | CharacteristicLabel | IndicatorId | IndicatorType | Value | DenominatorWeighted | DenominatorUnweighted |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_DOC | I | 28.5 | 2871 | 2903 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_DOC | I | 30.0 | 4122 | 4148 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_DOC | I | 27.3 | 2010 | 2041 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_NRS | I | 66.6 | 2871 | 2903 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_NRS | I | 65.0 | 4122 | 4148 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_NRS | I | 68.4 | 2010 | 2041 |

Filtered Health (I) – head()

| Dataset | SurveyYear | CharacteristicId | CharacteristicCategory | CharacteristicLabel | IndicatorId | IndicatorType | Value | DenominatorWeighted | DenominatorUnweighted |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_DOC | I | 28.5 | 2871 | 2903 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_DOC | I | 30.0 | 4122 | 4148 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_DOC | I | 27.3 | 2010 | 2041 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_NRS | I | 66.6 | 2871 | 2903 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_NRS | I | 65.0 | 4122 | 4148 |
| Access\_To\_Healthcare\_01 | 1998 | 1000 | Total | Total | RH\_ANCP\_W\_NRS | I | 68.4 | 2010 | 2041 |

set.seed(42)  
idx <- caret::createDataPartition(model\_df$HighValue, p = 0.7, list = FALSE)  
train\_df <- model\_df[idx, ] %>% mutate(  
 SurveyYear = forcats::fct\_drop(SurveyYear),  
 IndicatorId = forcats::fct\_drop(IndicatorId)  
)  
test\_df <- model\_df[-idx, ] %>% mutate(  
 SurveyYear = factor(SurveyYear, levels = levels(train\_df$SurveyYear)),  
 IndicatorId = factor(IndicatorId, levels = levels(train\_df$IndicatorId))  
)

## Setting direction: controls < cases

## png   
## 2

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 |

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 |

## Setting direction: controls < cases

## png   
## 2

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 |

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 |

Decision Tree – Top 10 Feature Importance

|  | Overall | Feature |
| --- | --- | --- |
| IndicatorId | 15.976471 | IndicatorId |
| DW\_log | 11.799693 | DW\_log |
| SurveyYear | 0.704405 | SurveyYear |

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 ≥ median(Value)). Both achieved reasonable discrimination. The comparison 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, tree plot, metrics, and variable importance in ../outputs/ .