AIDS STUDY

May 19, 2023

Ms. AL QOH, Oumaima
Mr. ARRIETA, Francisco
Ms. CAMENISCH, Lucia
Ms. GIANSANTE, Manuela
Ms. SCHMIDT, Emily
Ms. VALERA, Camille

Analytics Consulting – UNIGE – Spring 2023

Agenda

- Case Background
- Objective
- Results
- Conclusion
- Recommendations

CASE BACKGROUND: AIDS

- Cell count measurements are conducted to monitor patients affected by HIV/AIDS or who have other diseases such as cancer or hepatitis

- Most common diagnostics are CD4 and CD8 cell counts

- RNA counts (viral load) are also important among individuals diagnosed with HIV

CASE BACKGROUND: AIDS

Why are these cell counts important and what information do they convey?

- CD4: main indicator of HIV disease stage and progression
  - Cell counts below 200 cells/mm^3 indicate AIDS
- CD8: supporting indicator of HIV disease progression and immune function
- RNA Viral Load: measure of the amount of HIV in the bloodstream

CASE BACKGROUND: AIDS

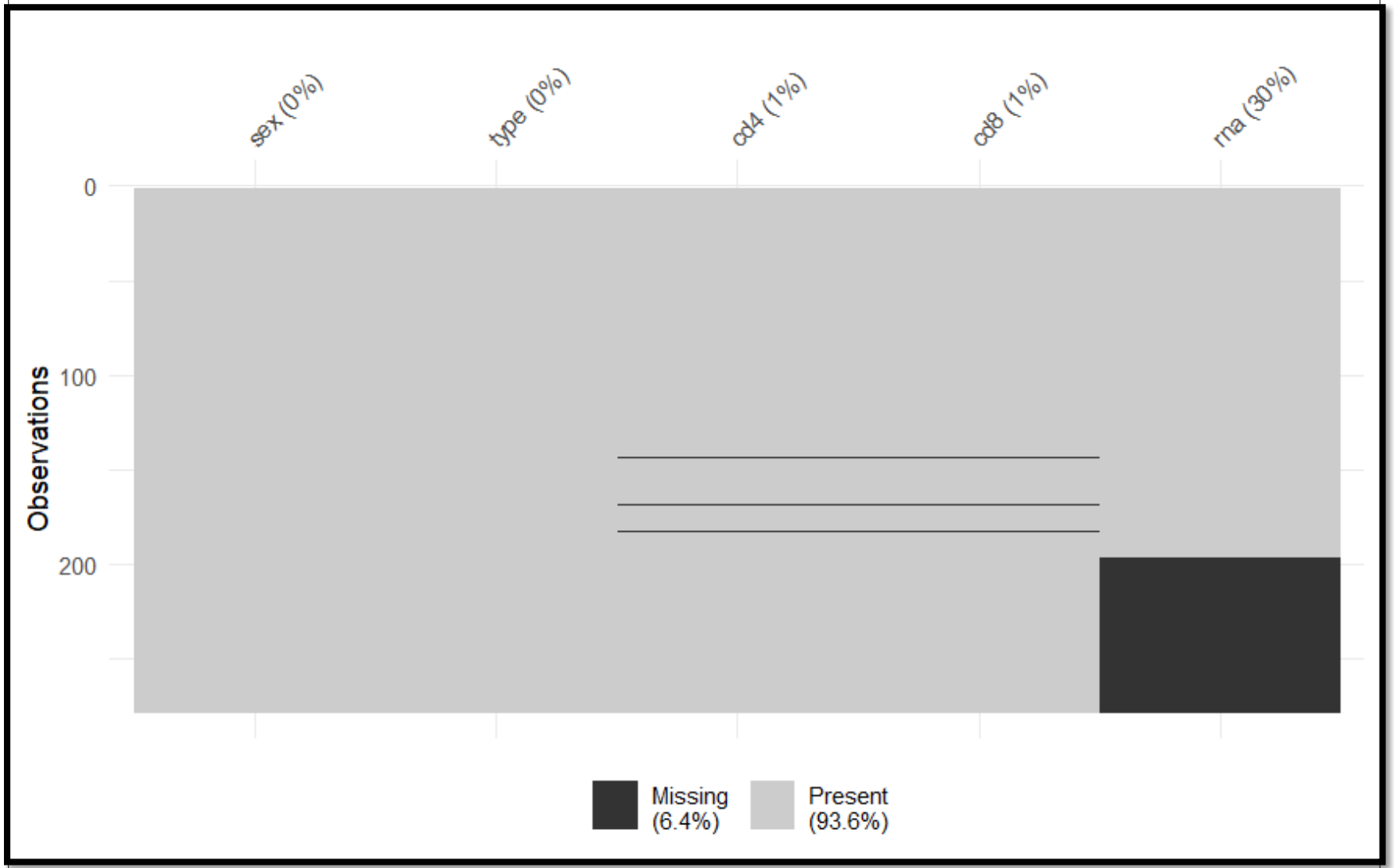
- A study was conducted measuring CD4, CD8, and RNA viral load among two groups of couples –
  - **Discordinant (DP):** one is HIV-positive
  - **Concordinant (CP):** both are HIV-positive
- One partner from each couple was included in the study
  - In the DP group, only the infected partner's cell counts were measured
- Drug users and non-monogamous couples were excluded

OBJECTIVE

- Determine if CD4, CD8, and RNA viral load measurements are able to provide distinction between couples classified as Discordinant (DP) and Concordinant (CP).

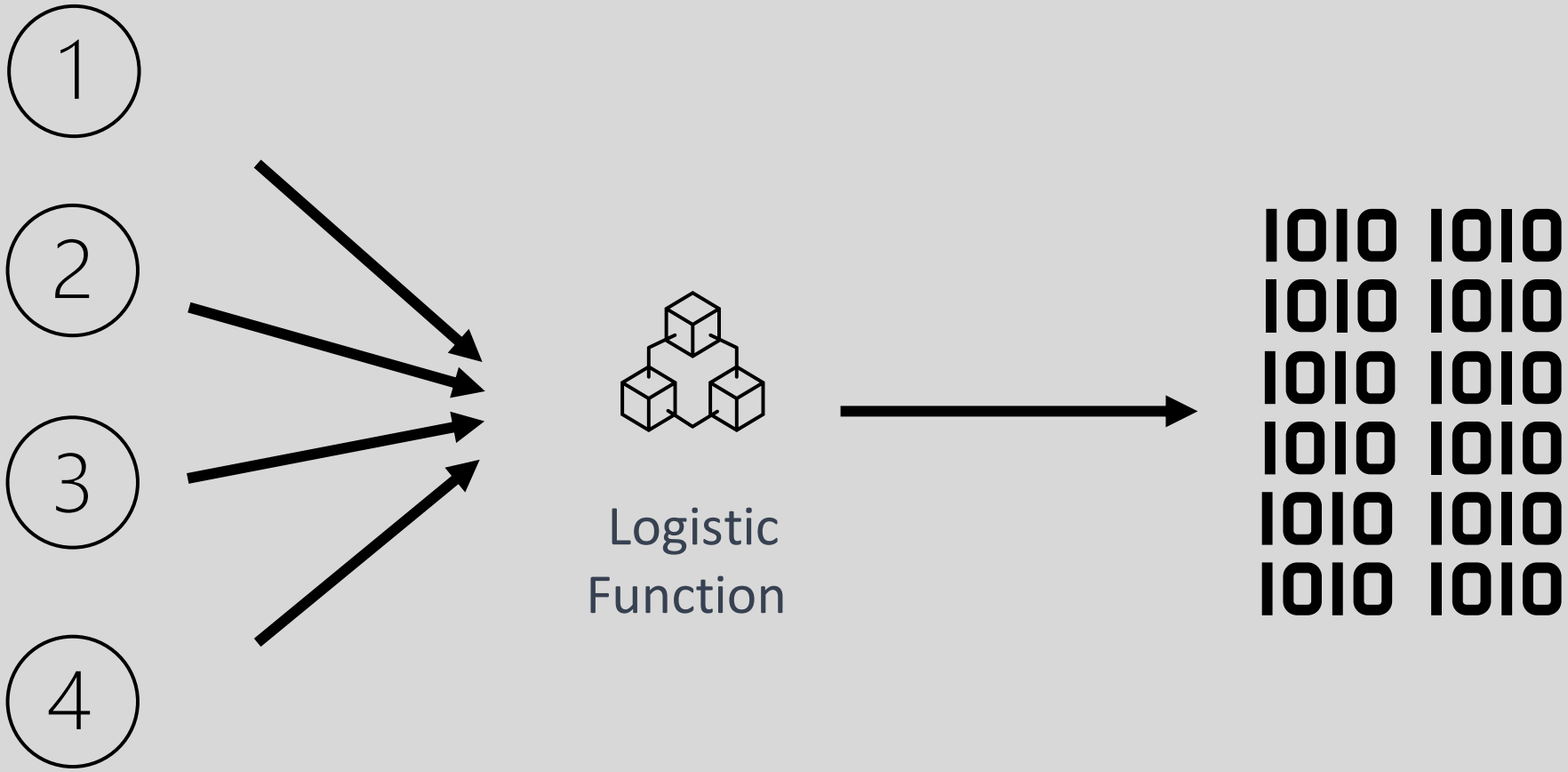
Missing Data

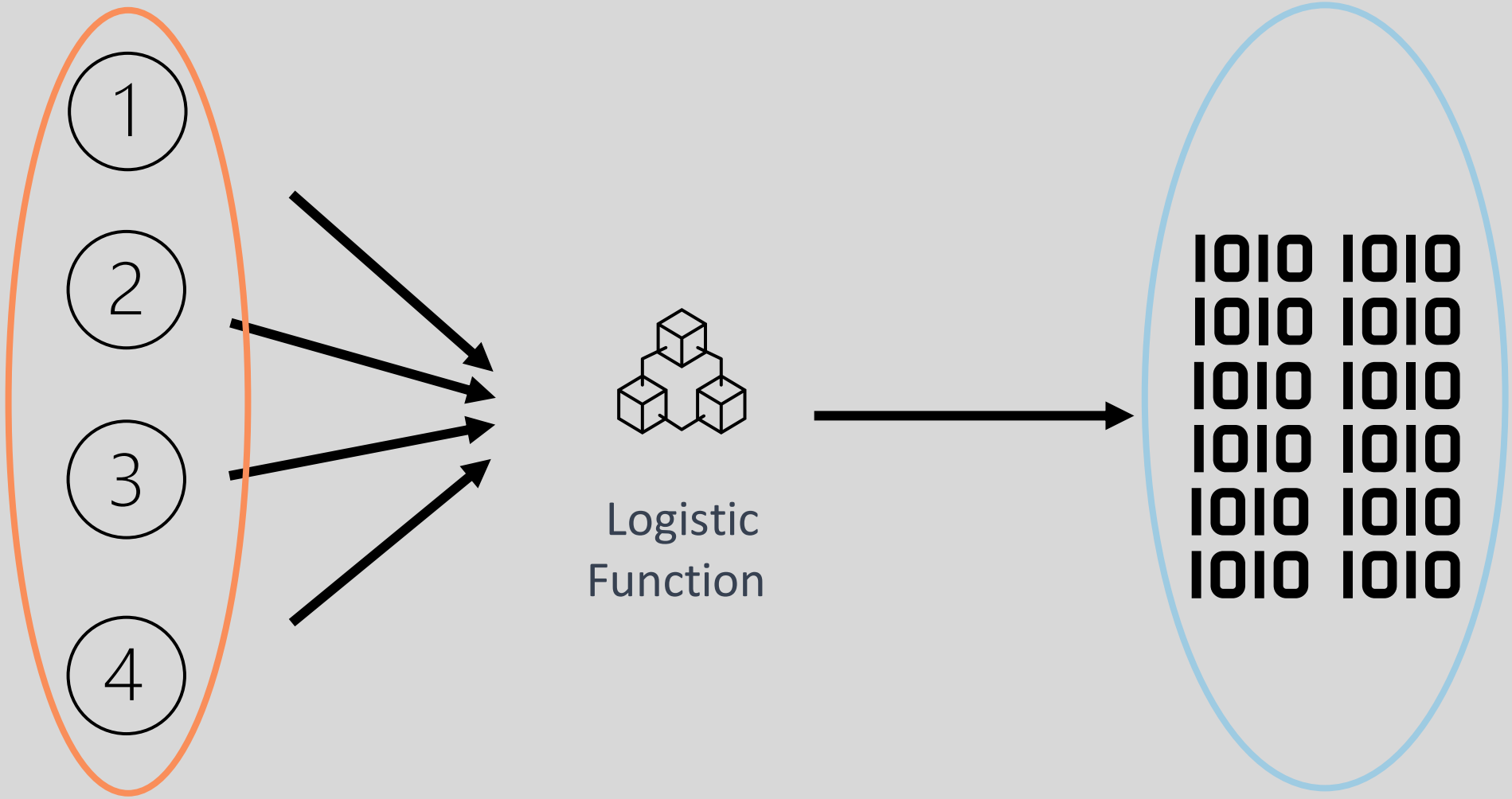
- The provided data presents a 6.4% portion of missing records:
- Respectively:
  - CD4 1%
  - CD8 1%
  - RNA 30%
- Discarding missing data points means deleting whole records.
- The **information loss** is too great!

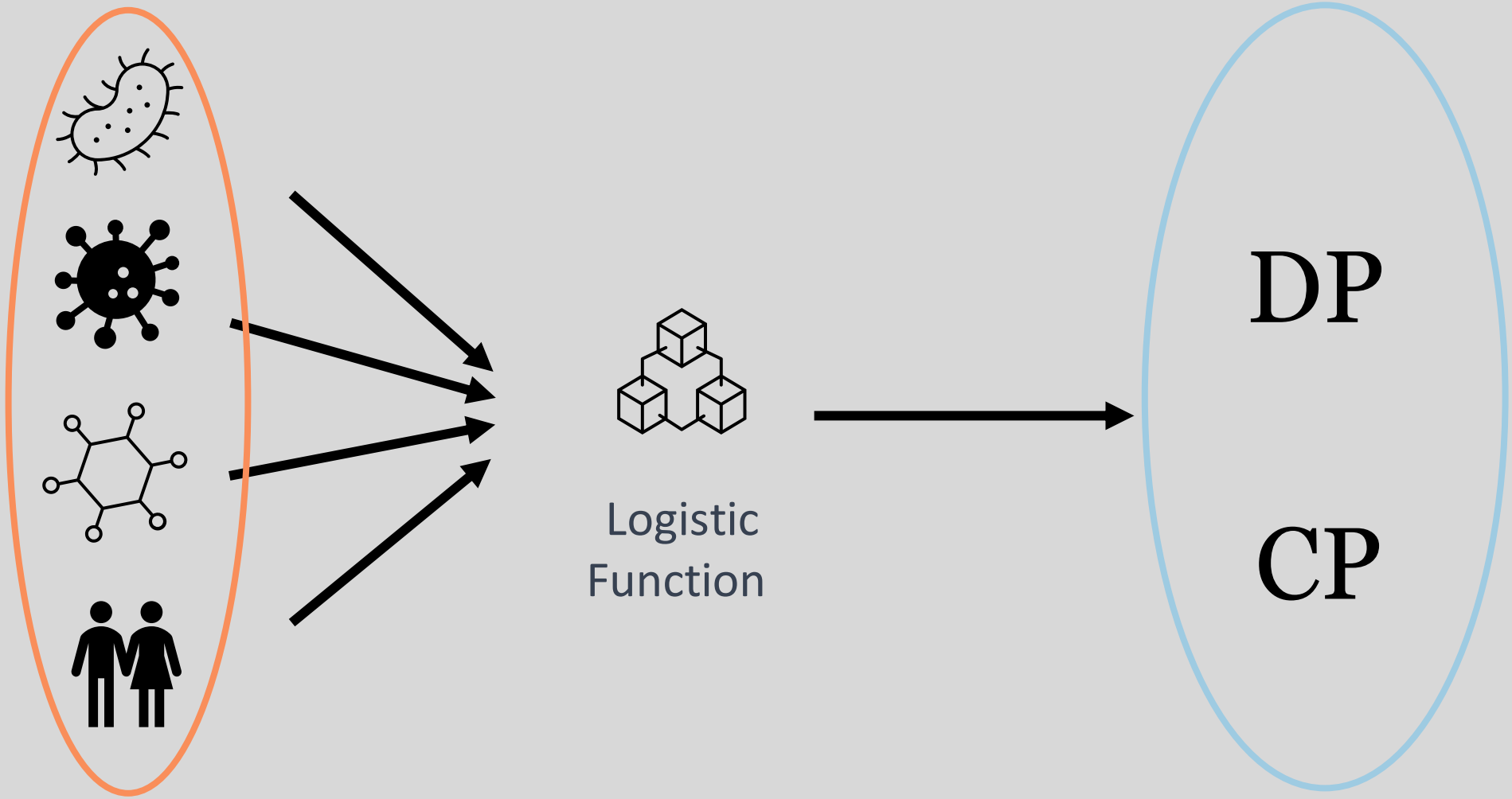


LOGISTIC REGRESSION

Model I


1
2
3
4
Logistic
Function
1010 1010
1010 1010
1010 1010
1010 1010
1010 1010
1010 1010


1
2
3
4
Logistic
Function
1010 1010
1010 1010
1010 1010
1010 1010
1010 1010
1010 1010

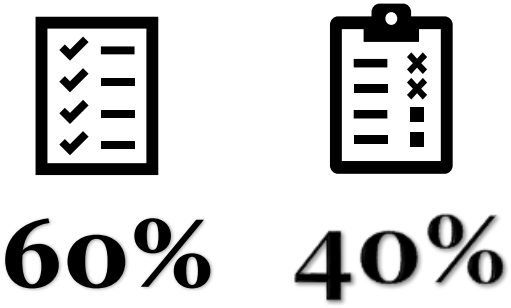

Logistic
Function
DP
CP

Assumptions

- **Binary outcome:** The outcome predicted has only two possible values
- **Linearity:** Straight-line relationship between the predictors and log-odds of predicted responses
- **Independence:** Each observation in the dataset should be independent of all other observations.
- **No multicollinearity:** The predictors should not be too strongly correlated with each other.
- **Large sample size:** The accuracy of the logistic regression model improves with a larger sample size.

MODEL

Train-test partitioning



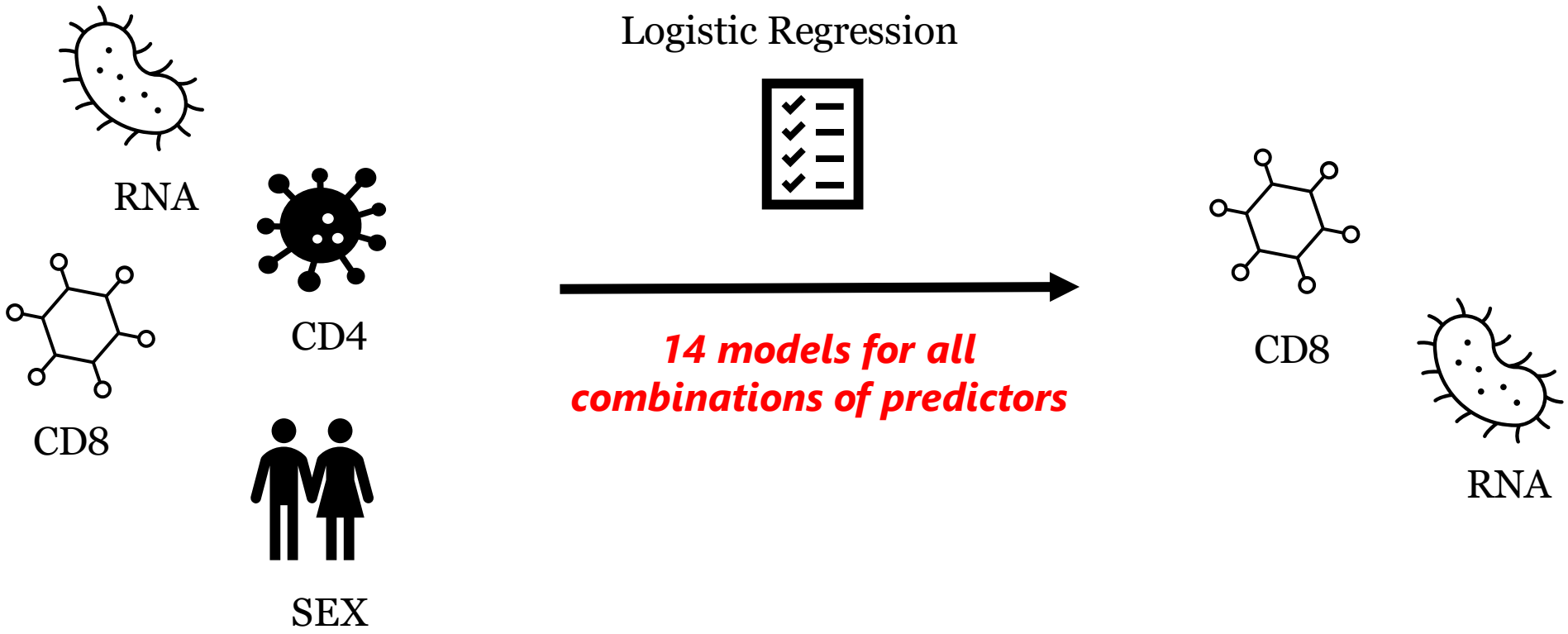
Logistic Regression Train-set

CHOSEN MODEL

Logistic Regression

14 models for all combinations of predictors

CHOSEN MODEL



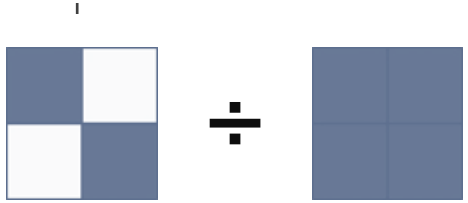
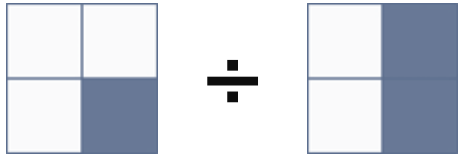
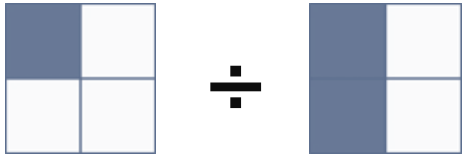
CONFUSION MATRIX DEFINITIONS

- **Accuracy:** The overall correctness of a predictive model in terms of its ability to correctly classify data into their true classes.
- **Sensitivity:** The ability of a predictive model to correctly identify positive cases.
- **Specificity:** The ability of a predictive model to correctly identify negative cases.

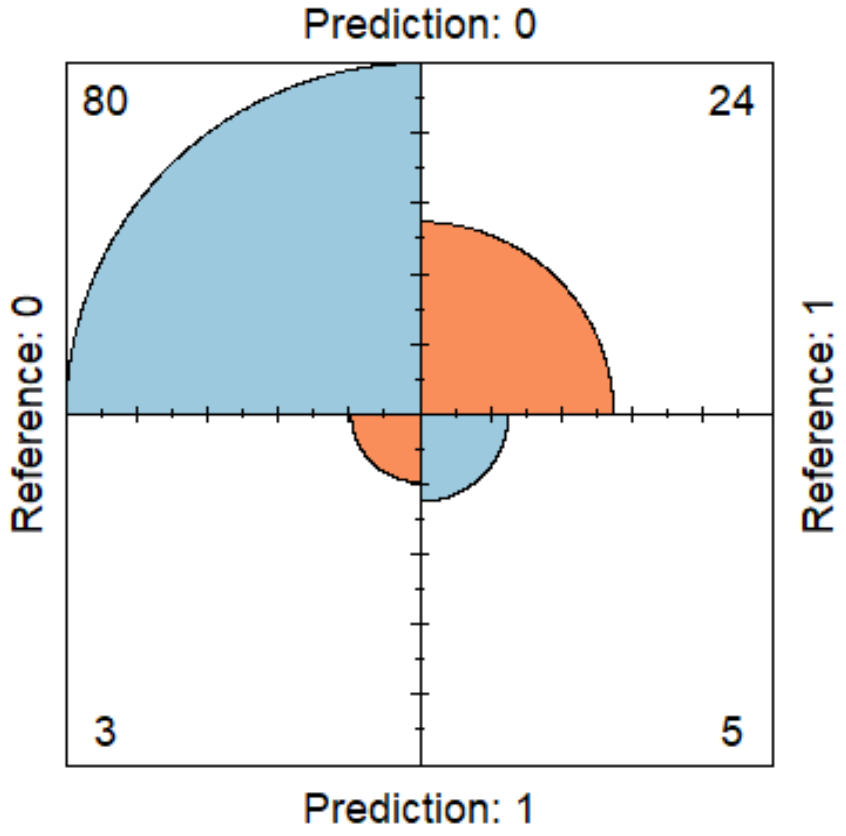
- **Type I Error:** Classifying a record as CP instead of DP
- **Type II Error:** Classifying a record as DP instead of CP

CONFUSION MATRIX

True Negative (TN)	**False Positive (FP)** *Type I Error*
False Negative (FN) *Type II Error*	**True Positive (TP)**

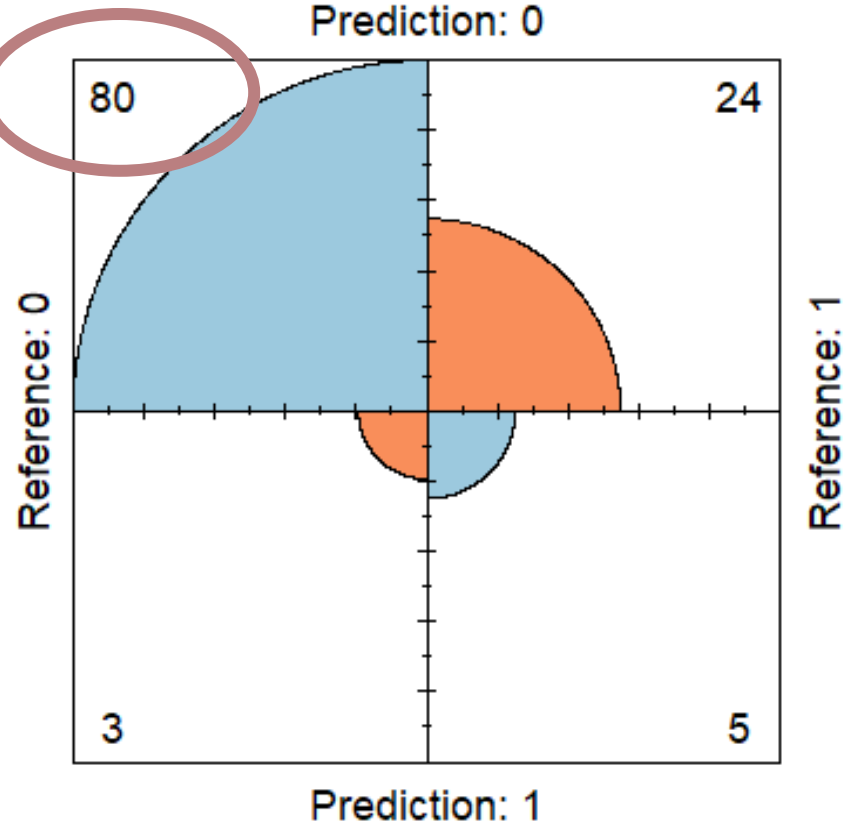
- Accuracy: 
- Sensitivity: 
- Specificity: 

CONFUSION MATRIX

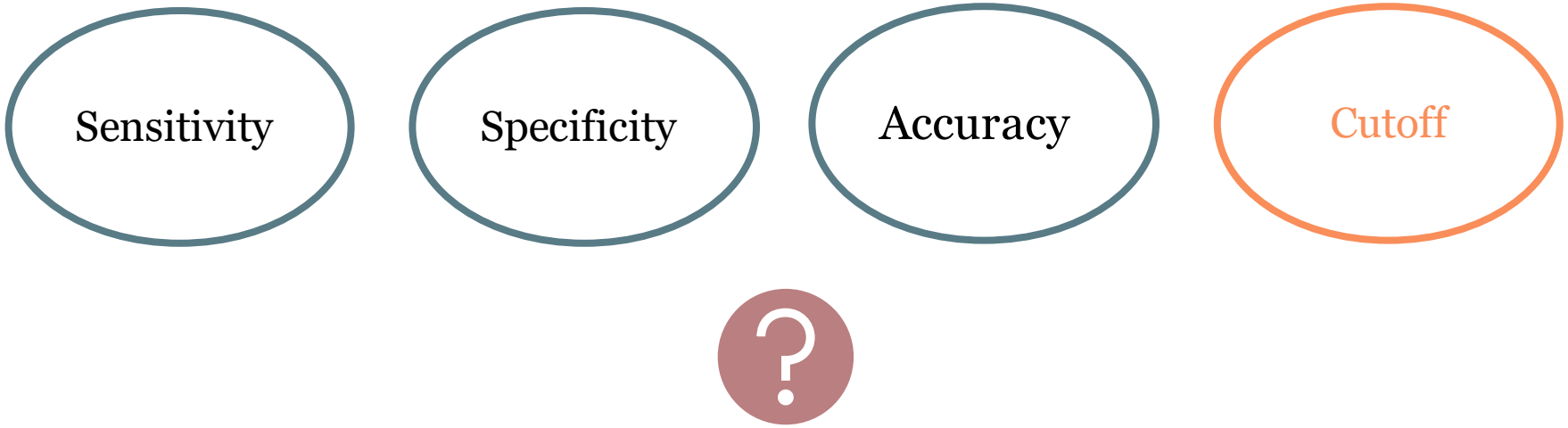


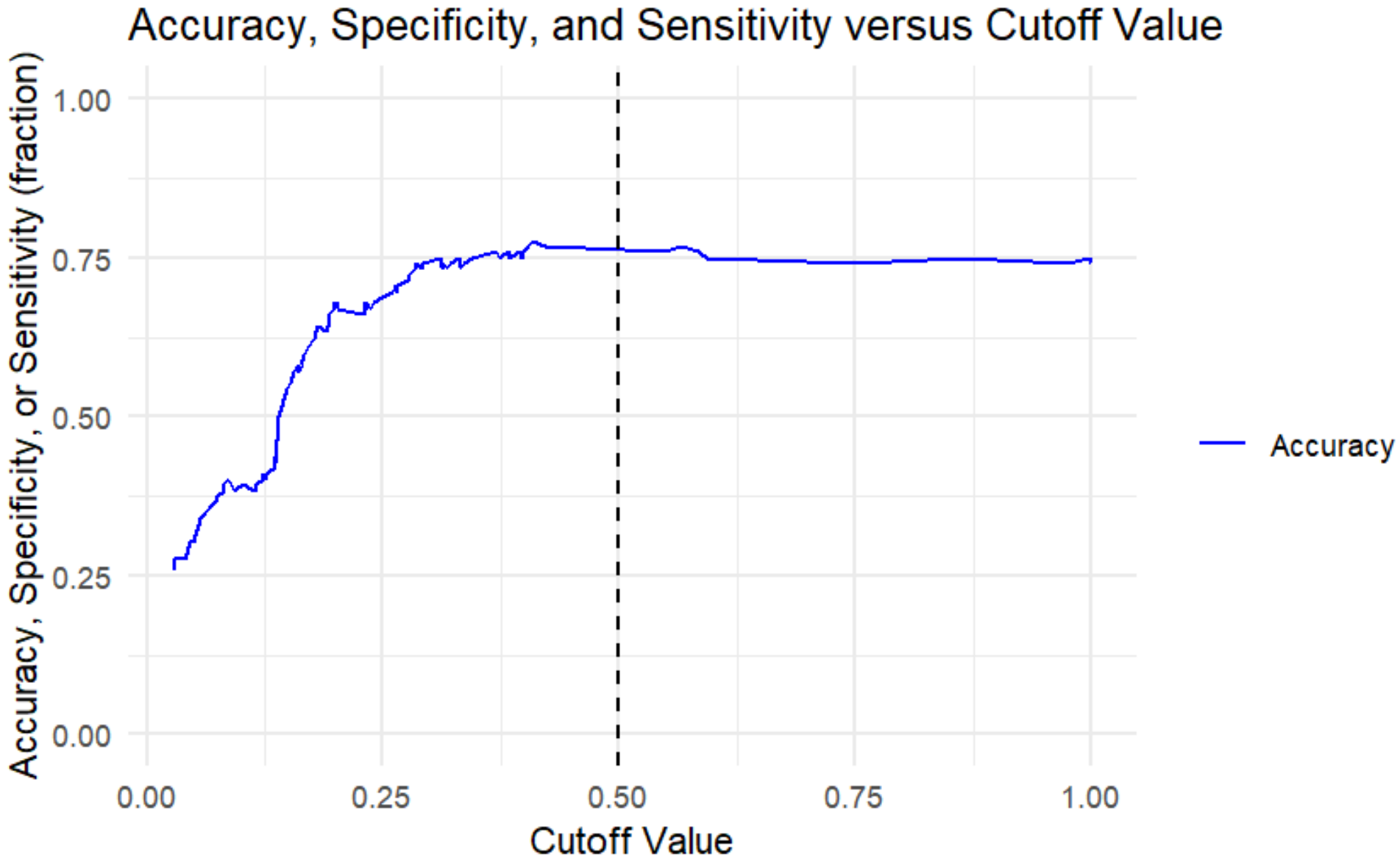
- Accuracy: **75.9%**
- Sensitivity: **17.5%**
- Specificity: **96.4%**

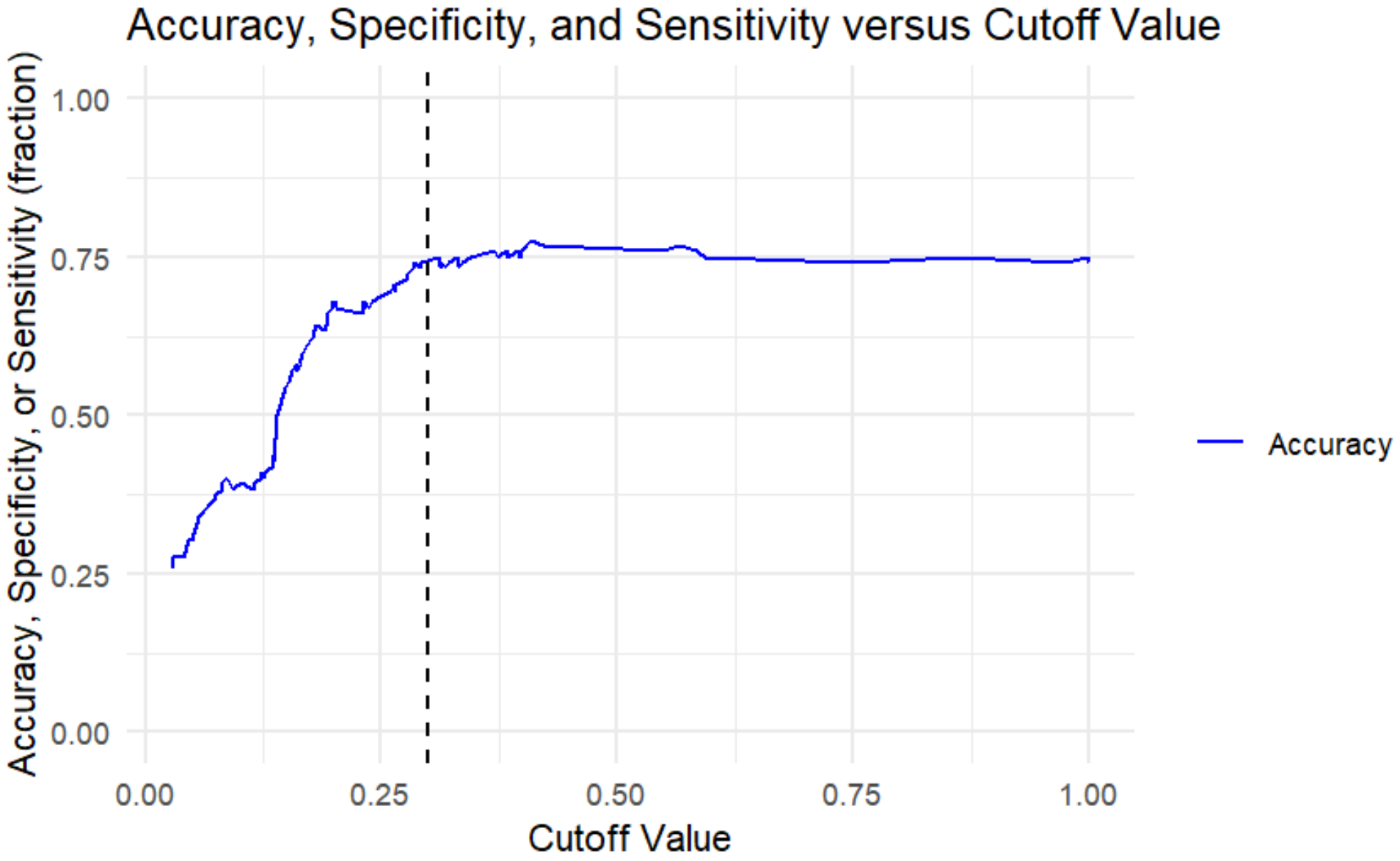
CONFUSION MATRIX

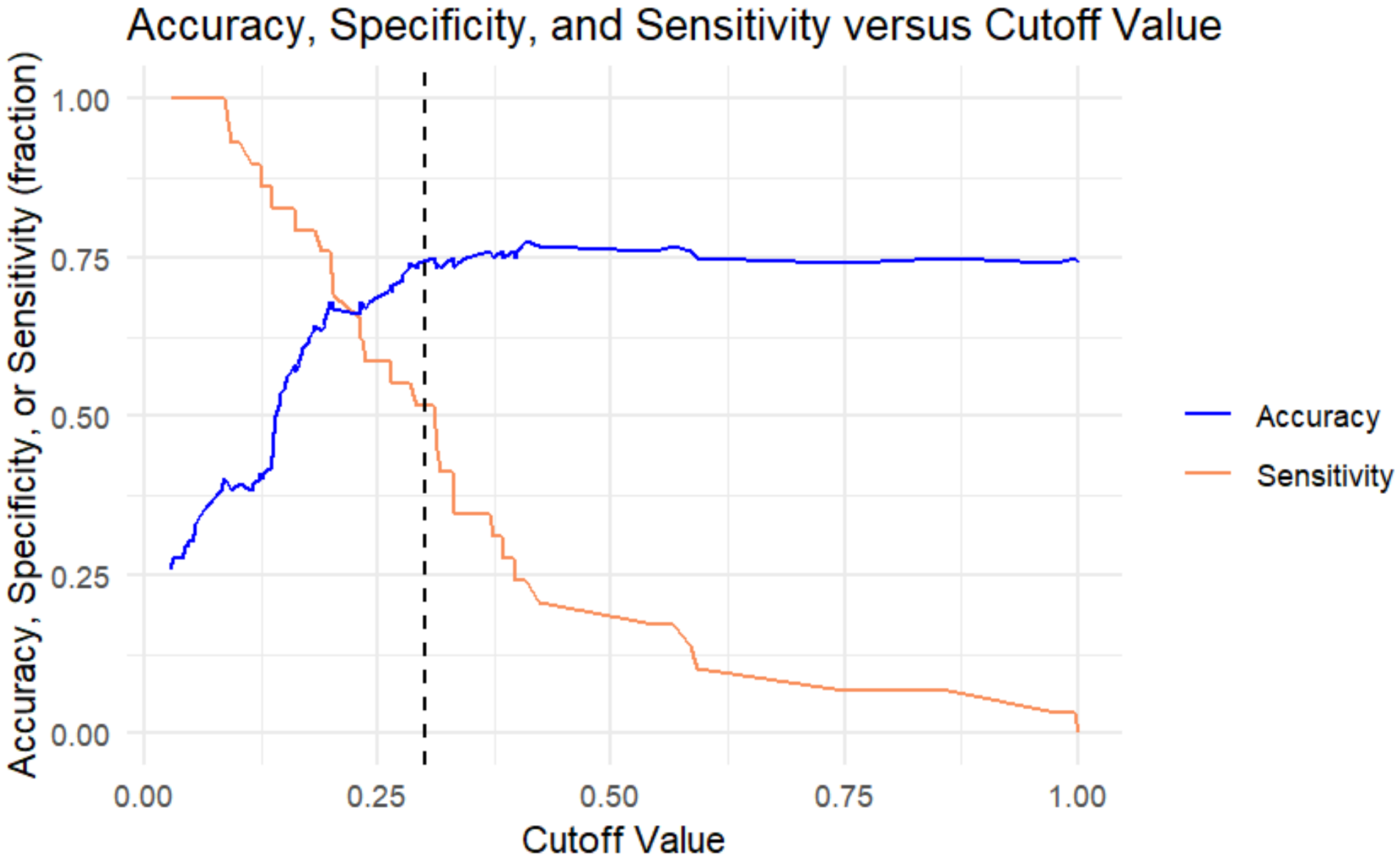


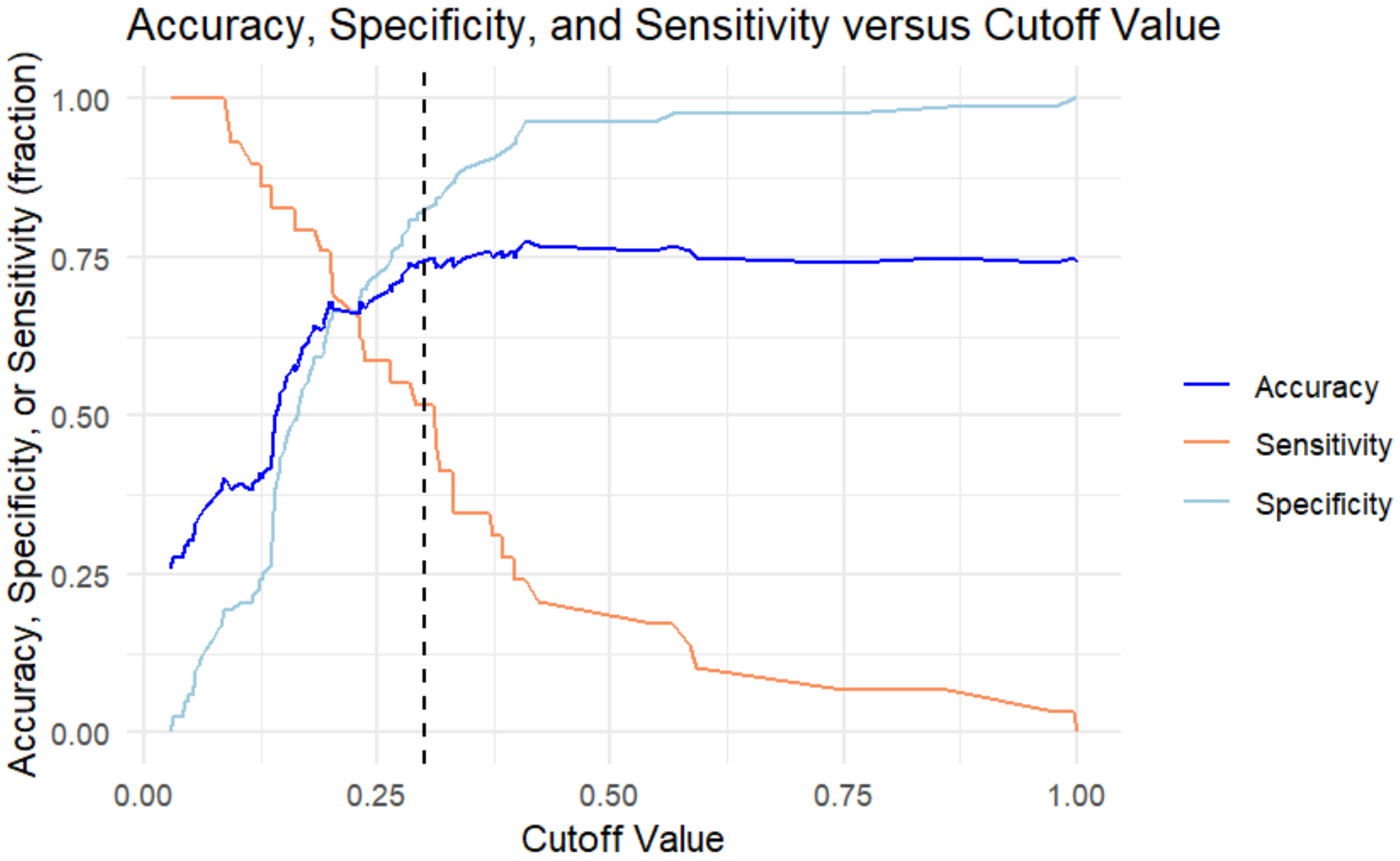
- Fewer positive samples than negative samples


Sensitivity
Specificity
Accuracy
Cutoff
?

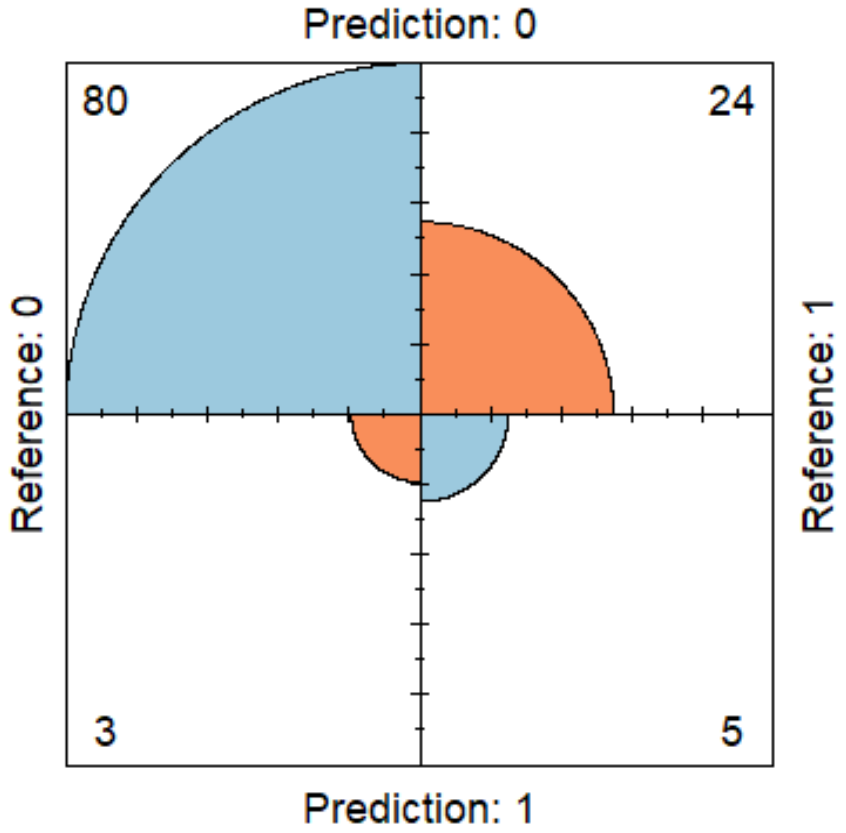

Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, or Sensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, or Sensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, or Sensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy
Sensitivity


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, or Sensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy
Sensitivity
Specificity

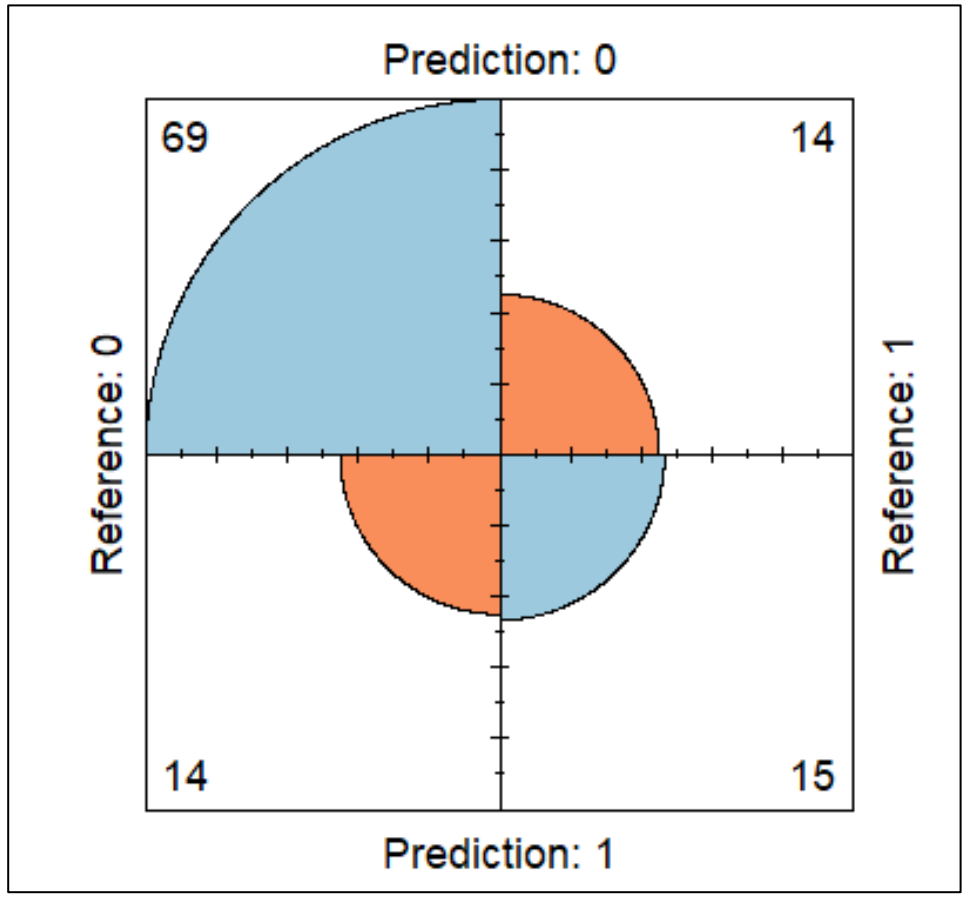
CONFUSION MATRIX



Cutoff: 0.5

- Accuracy: **75.9%**
- Sensitivity: **17.5%**
- Specificity: **96.4%**

CONFUSION MATRIX



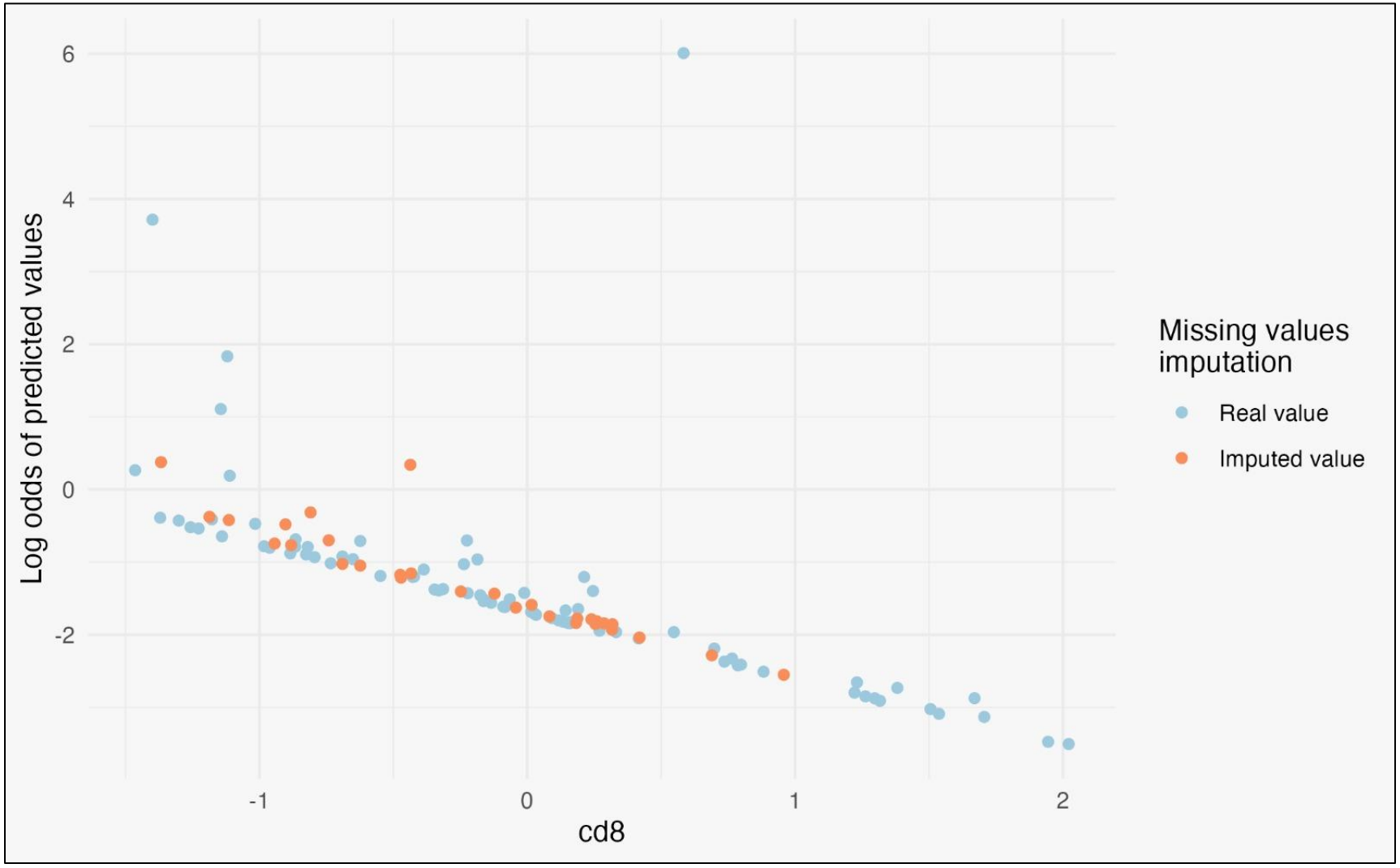
Cutoff: 0.3

- Accuracy: ~~75.9%~~ **75.0%**
- Sensitivity: ~~17.5%~~ **51.7%**
- Specificity: ~~96.4%~~ **83.1%**

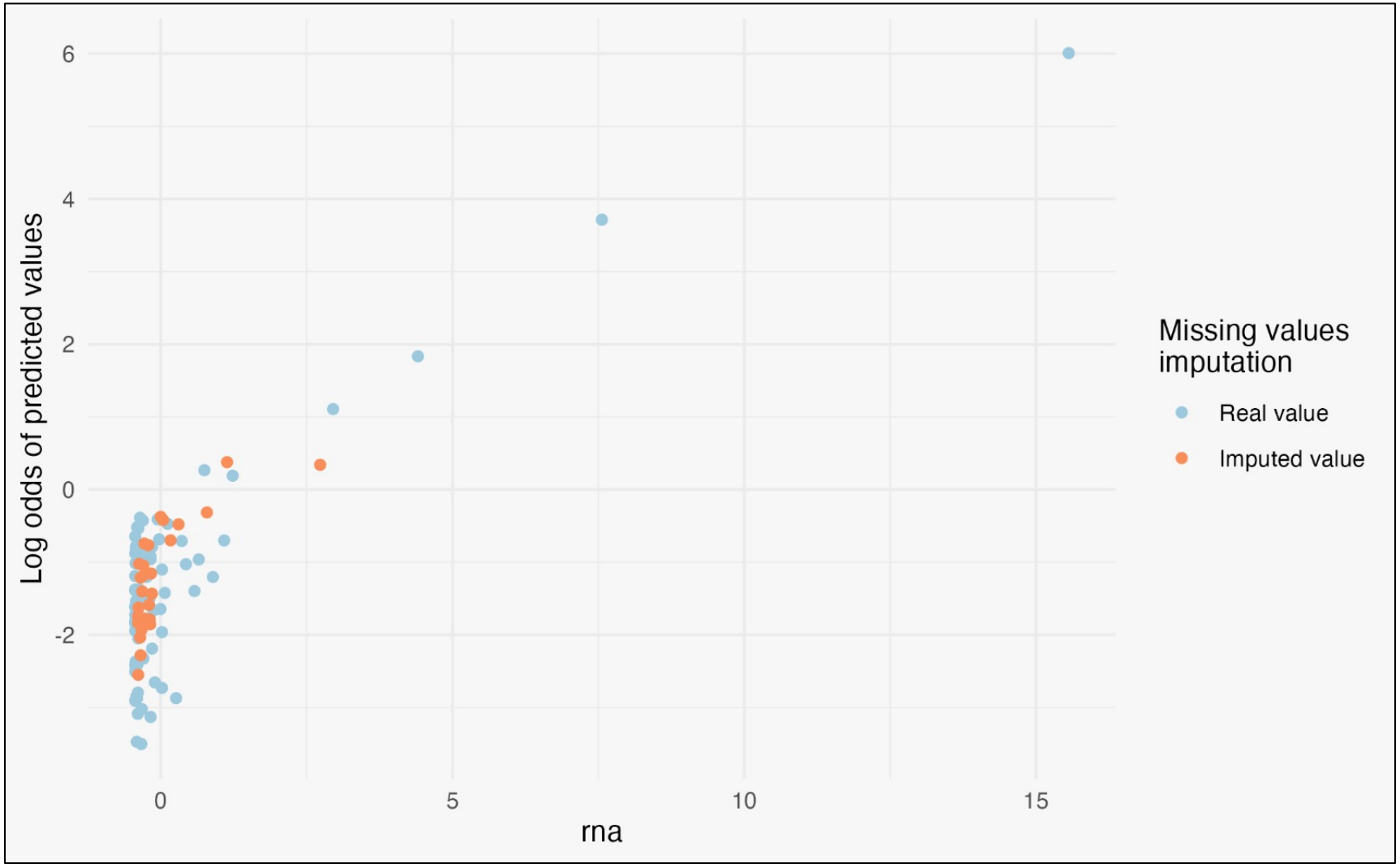
Assumptions

- **Binary outcome:** The outcome predicted has only two possible values
- **Linearity:** Straight-line relationship between the predictors and Log-odds of predicted responses
- **Independence:** Each observation in the dataset should be independent of all other observations.
- **No multicollinearity:** The predictors should not be too strongly correlated with each other.
- **Large sample size:** The accuracy of the logistic regression model improves with a larger sample size.

Linear Assumption Test Set: CD8



Linear Assumption Test Set: RNA



CLASSIFICATION TREES

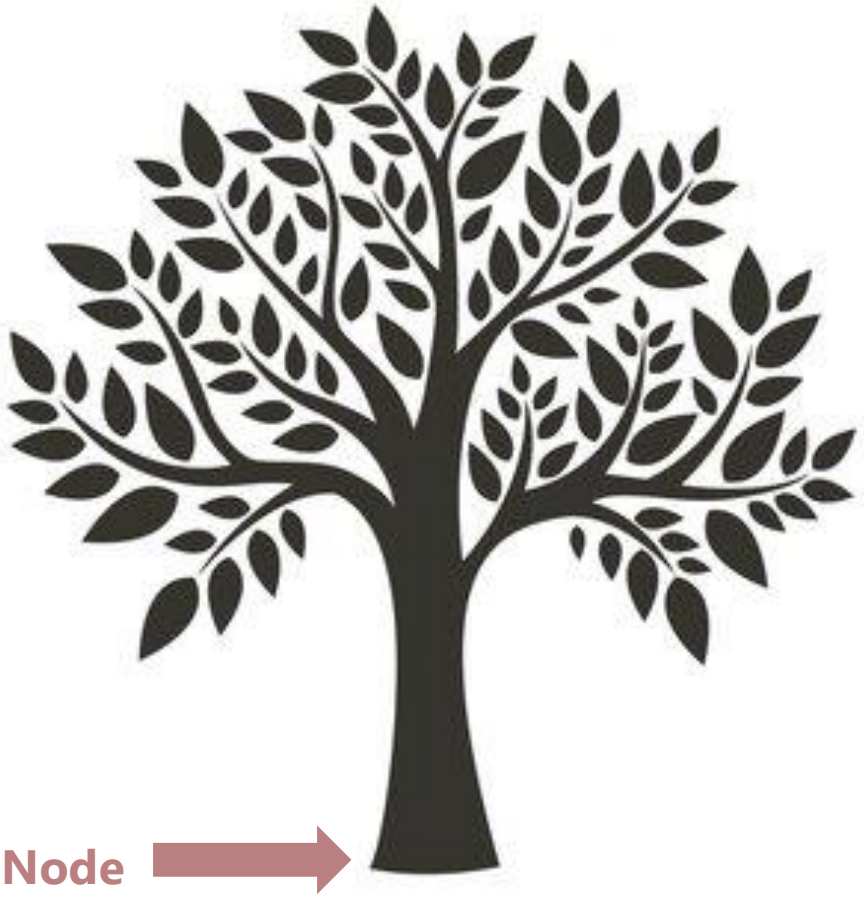
Model II

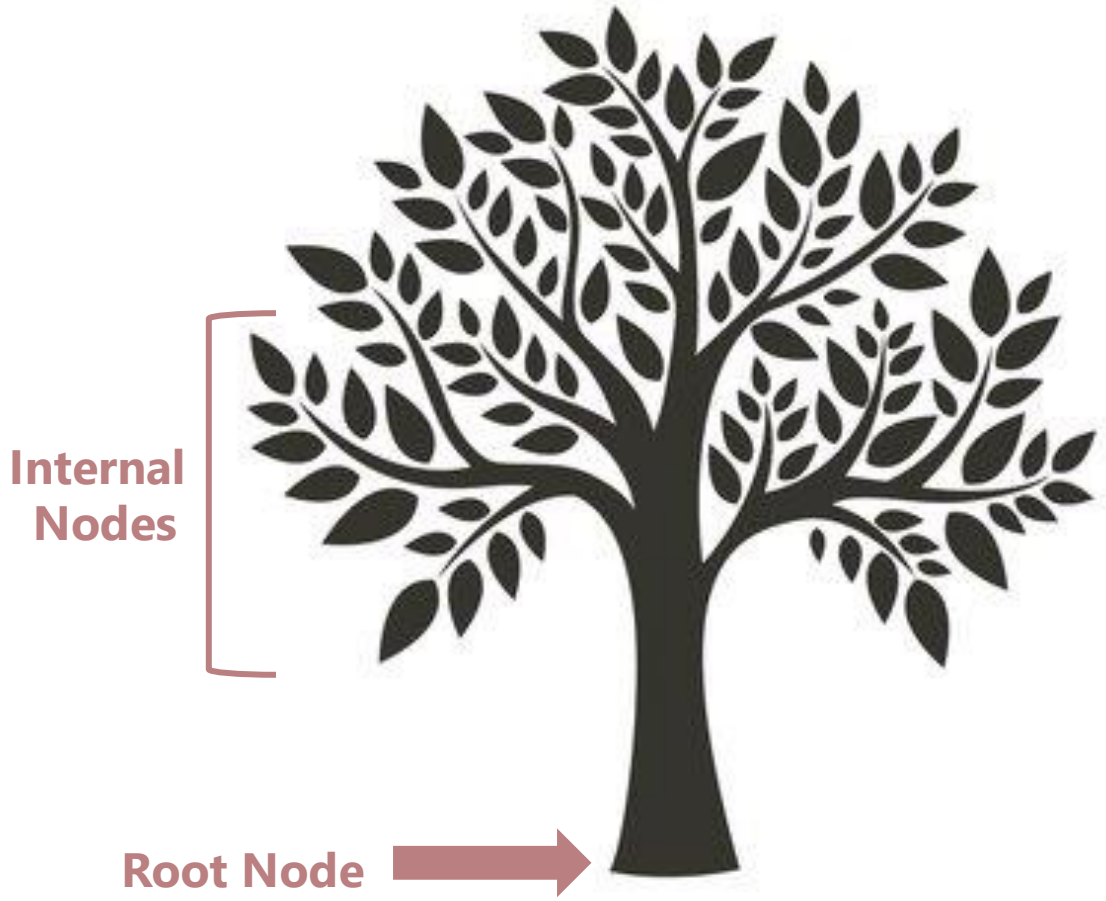
TREE OVERVIEW

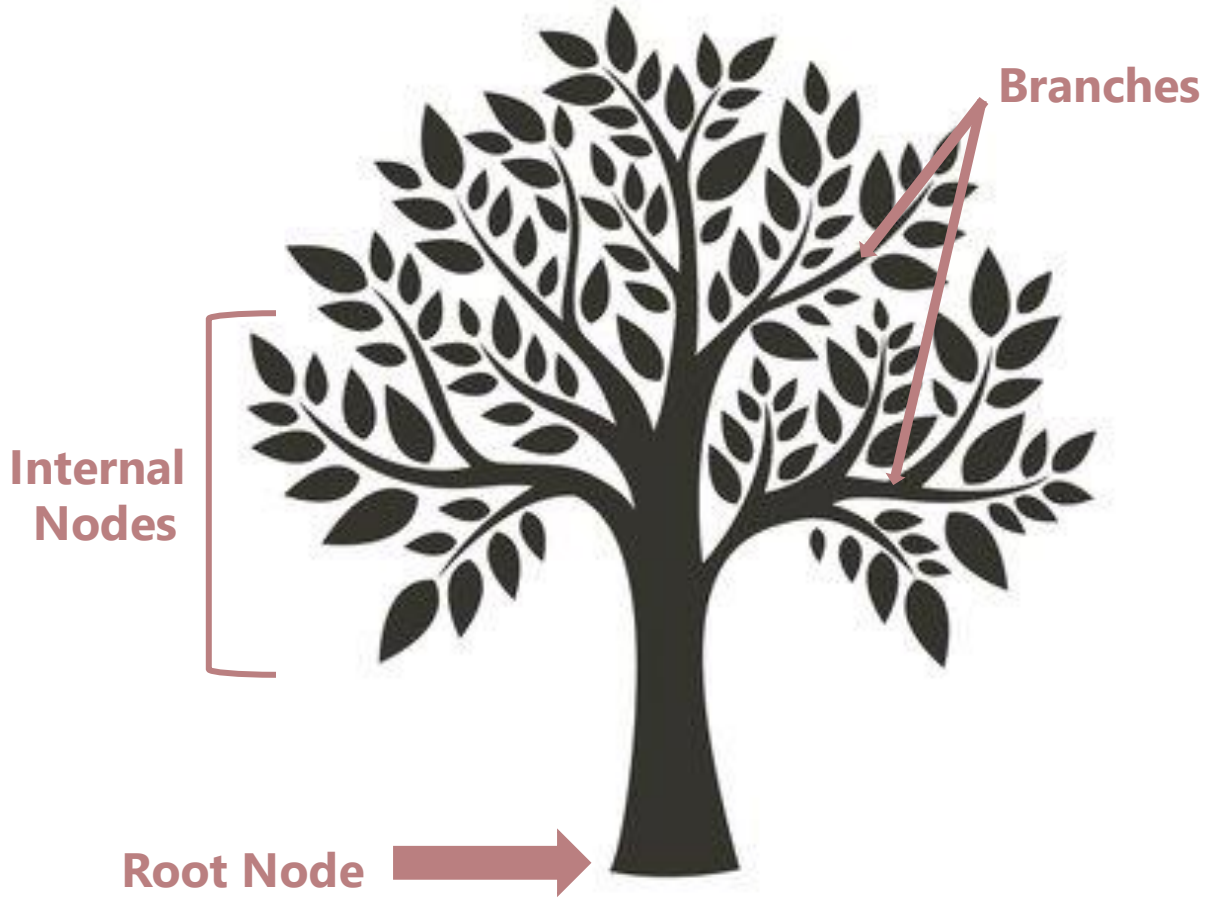
- **Advantages**
  - Trees are incredibly simple to communicate to individuals, even more so than linear regression.
  - They resemble human decision-making processes.
  - Trees can be graphically represented and understood without expertise.
  - They can accommodate qualitative predictors without requiring dummy variables, if there are not too many levels.

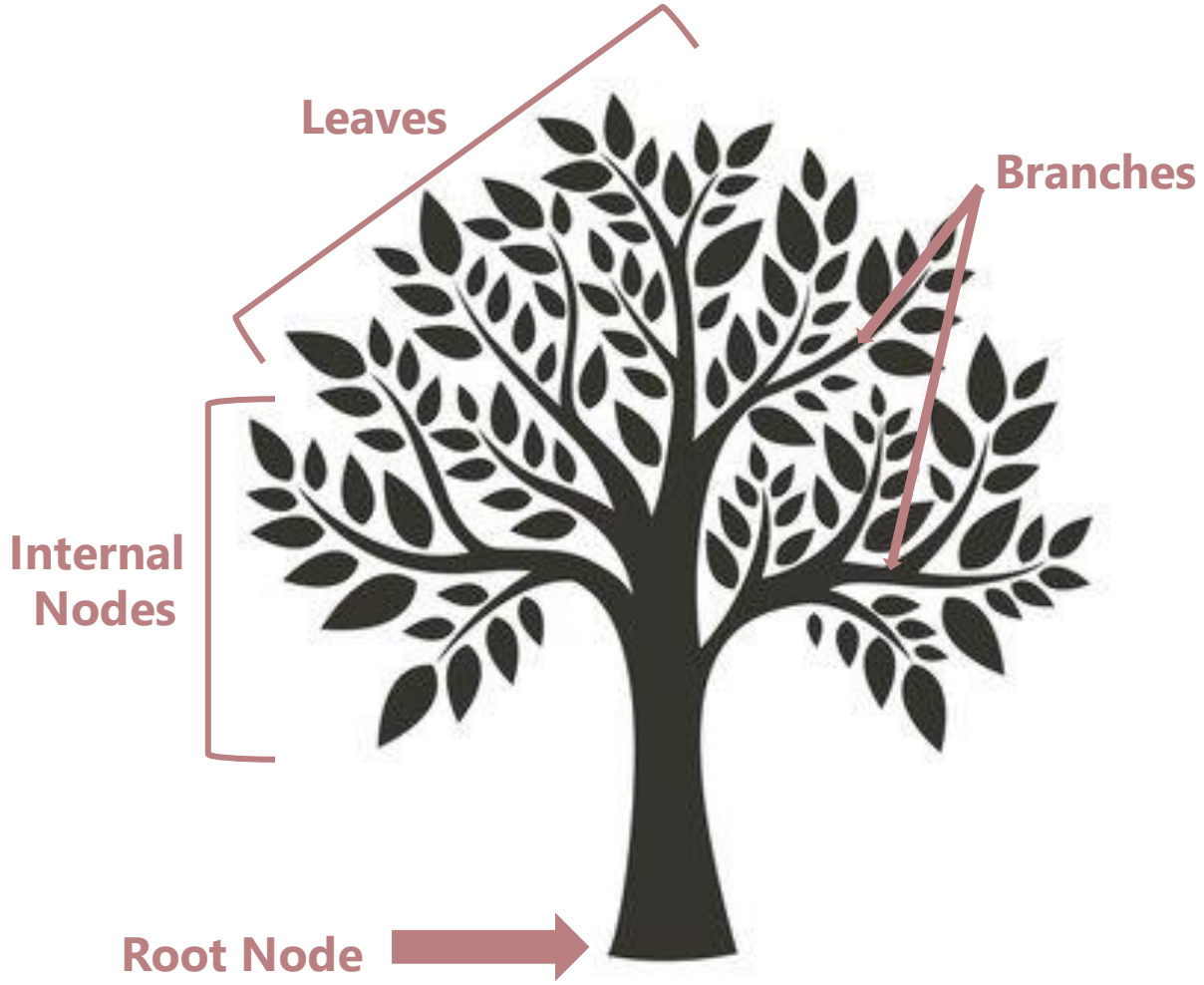
TREE OVERVIEW

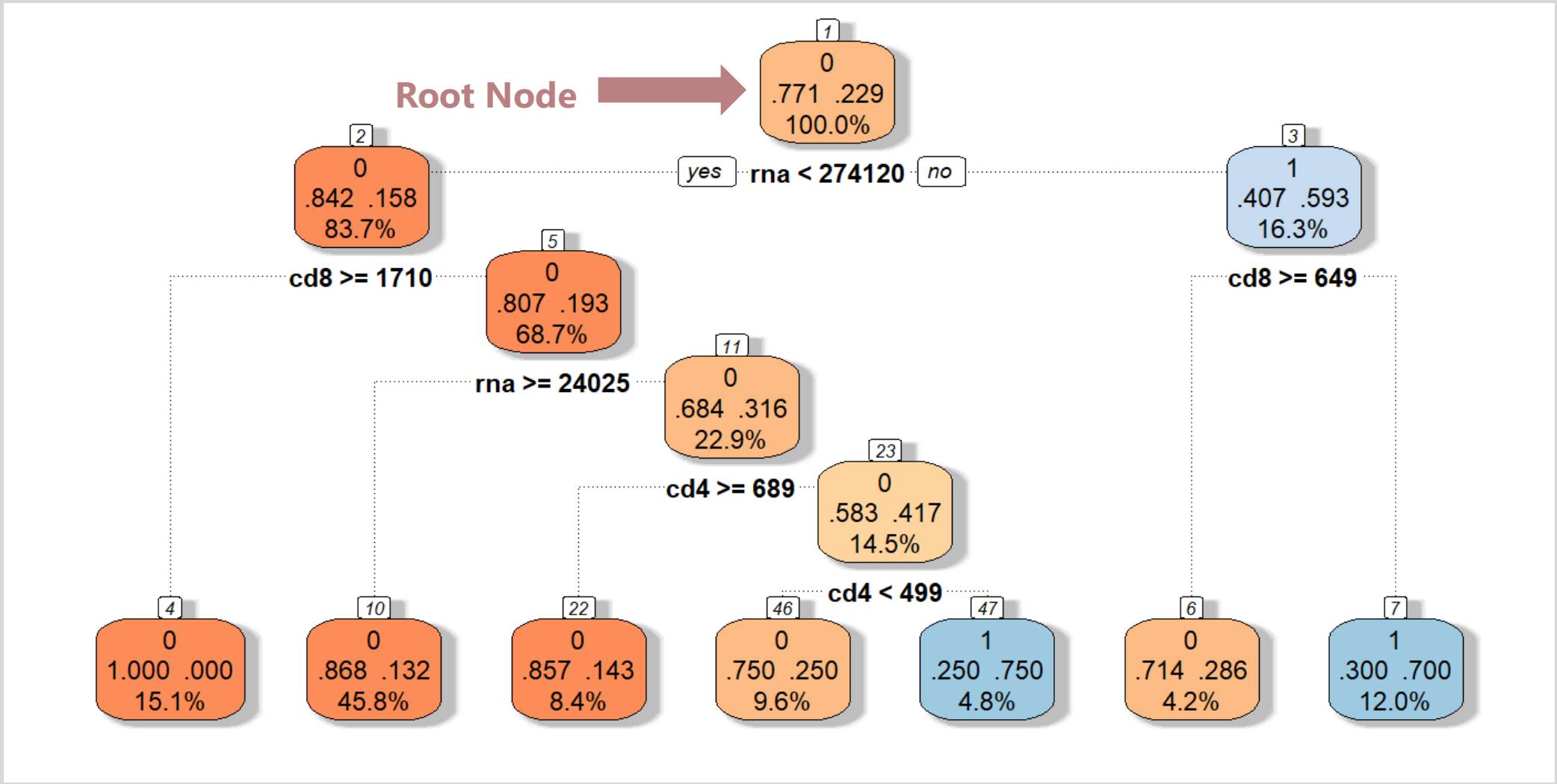
- **Disadvantages**
  - Trees generally have poor predictive performance because they are not so flexible. Can be improved by using a combination of different trees .
  - A small change in the data can completely change the tree.

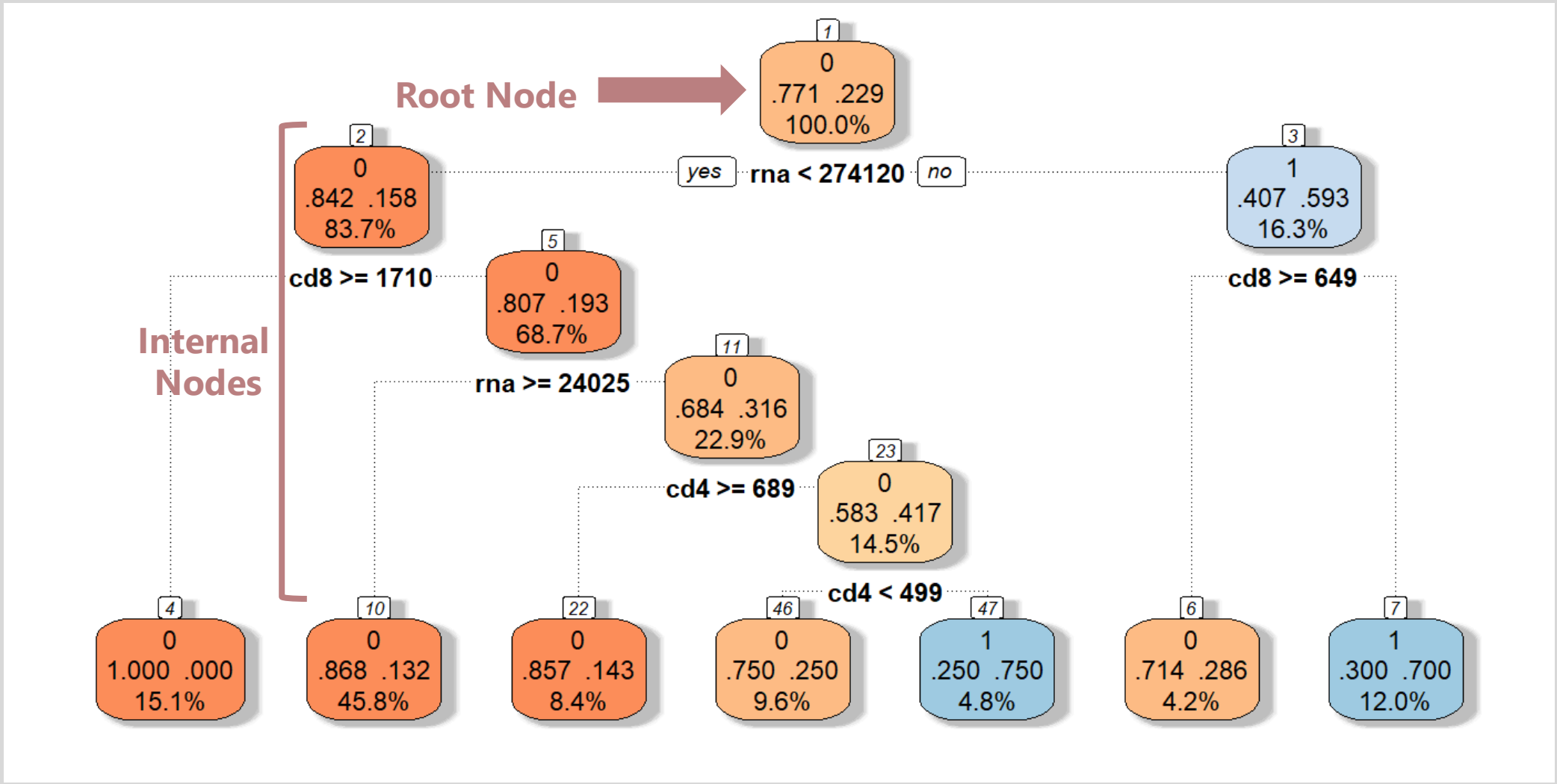

Root Node

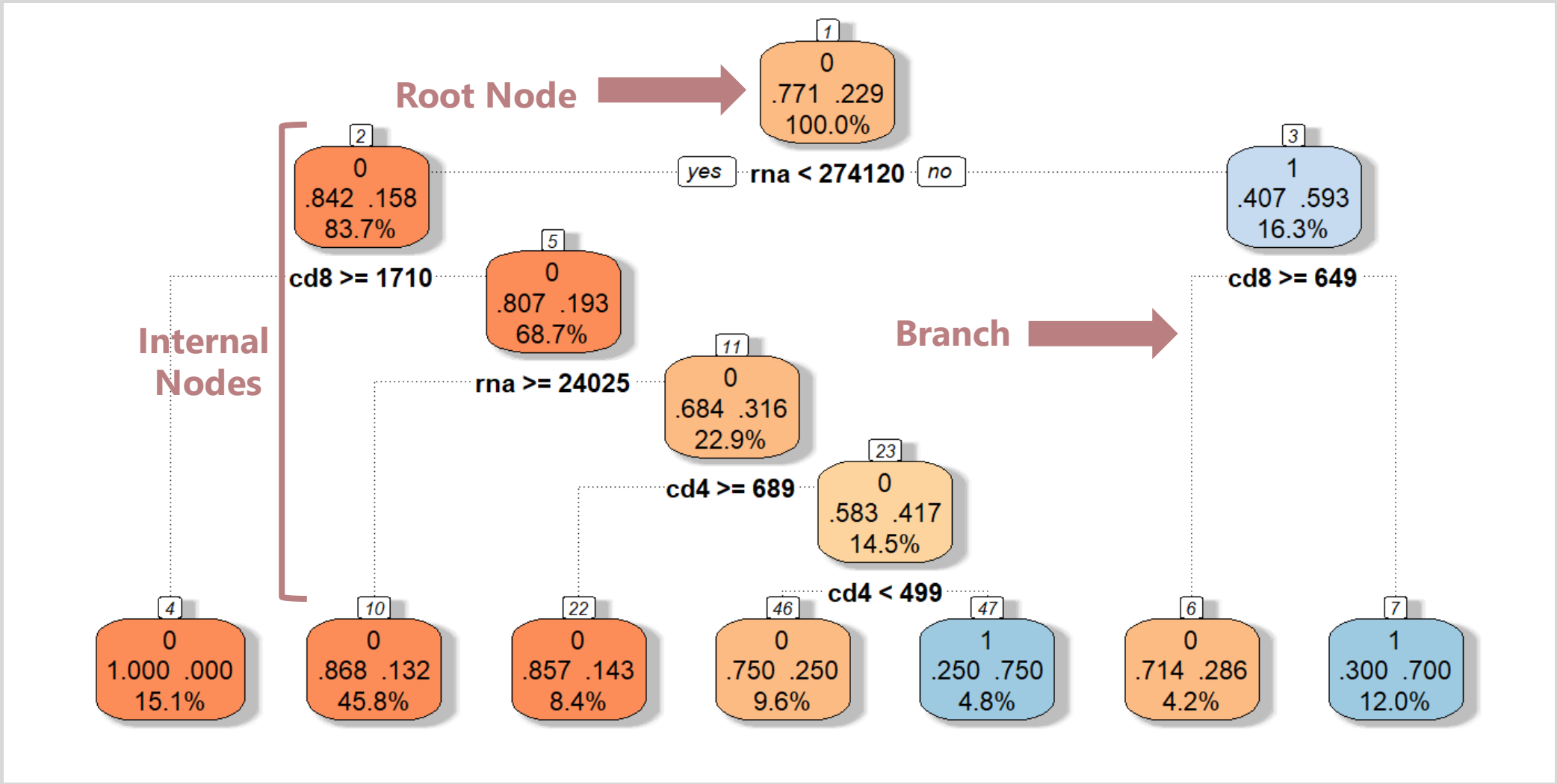

Internal
Nodes
Root Node

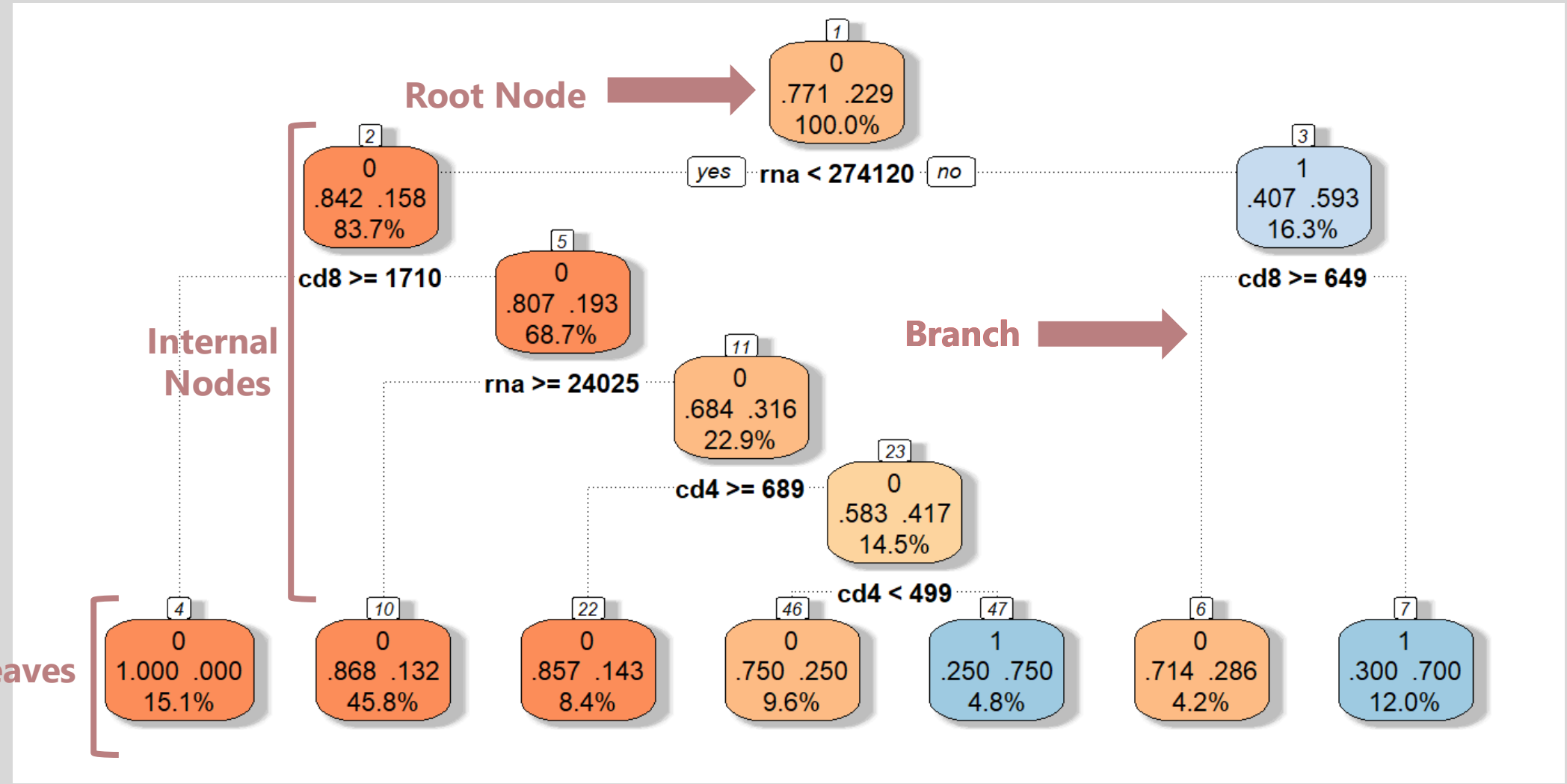

Branches
Internal Nodes
Root Node


Leaves
Branches
Internal Nodes
Root Node

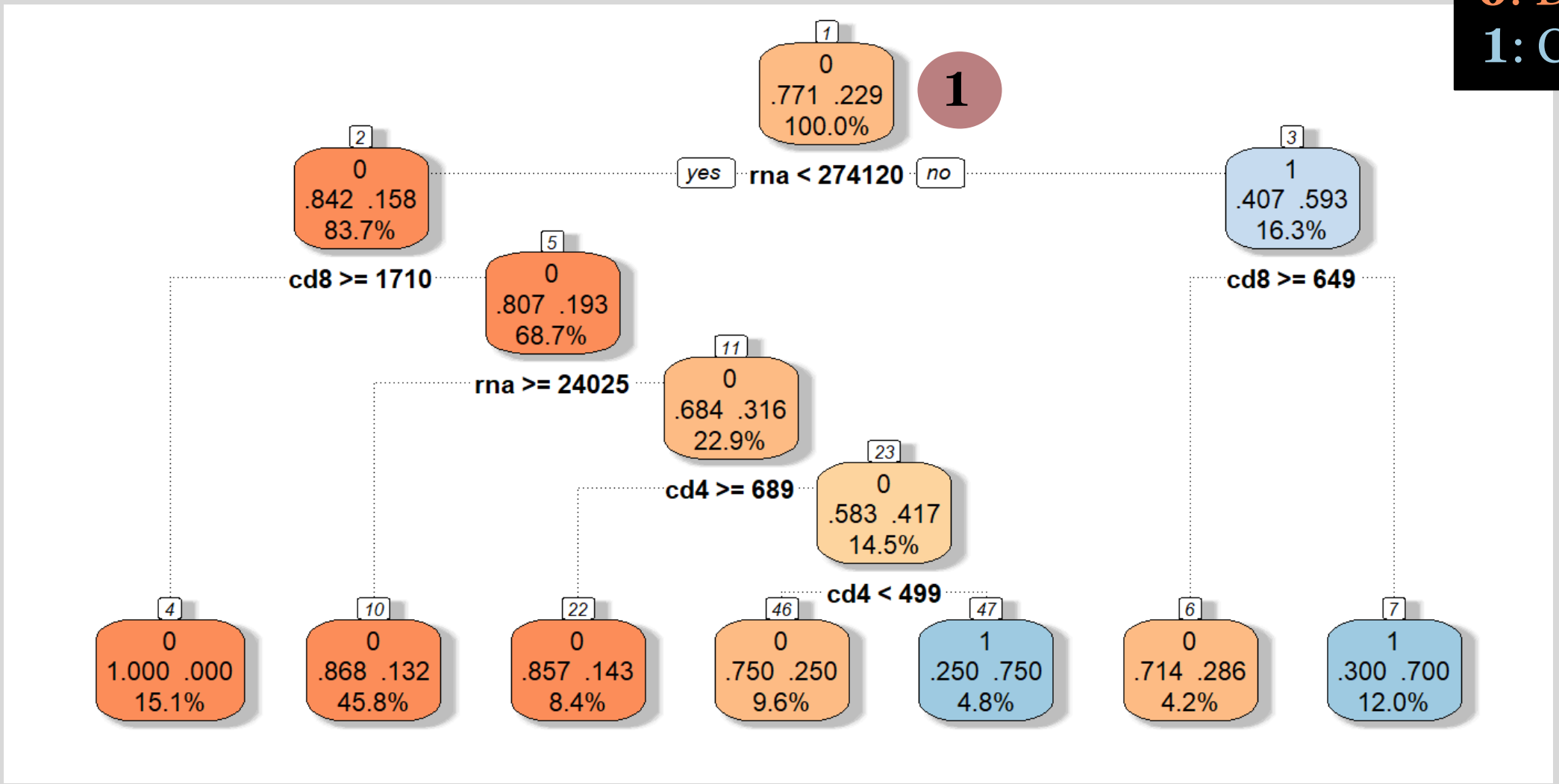

Root Node
1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
5
0
.807 .193
68.7%
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%


Root Node
Internal Nodes
1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
cd8 >= 649
5
0
.807 .193
68.7%
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

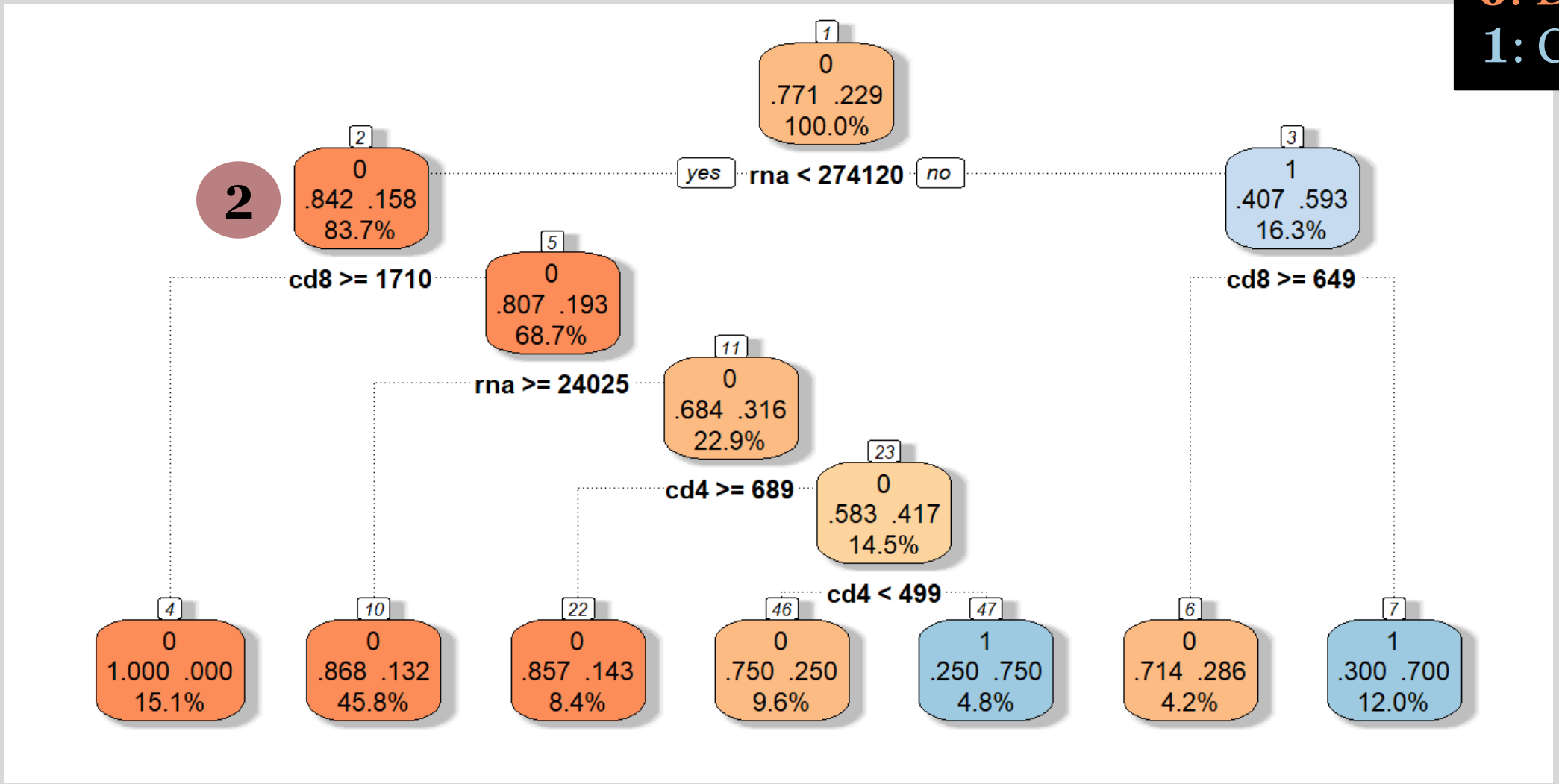

Root Node
1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
cd8 >= 649
5
0
.807 .193
68.7%
Internal Nodes
Branch
11
0
.684 .316
22.9%
rna >= 24025
23
0
.583 .417
14.5%
cd4 >= 689
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%


Root Node
Internal Nodes
Branch
Leaves
1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
cd8 >= 649
5
0
.807 .193
68.7%
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

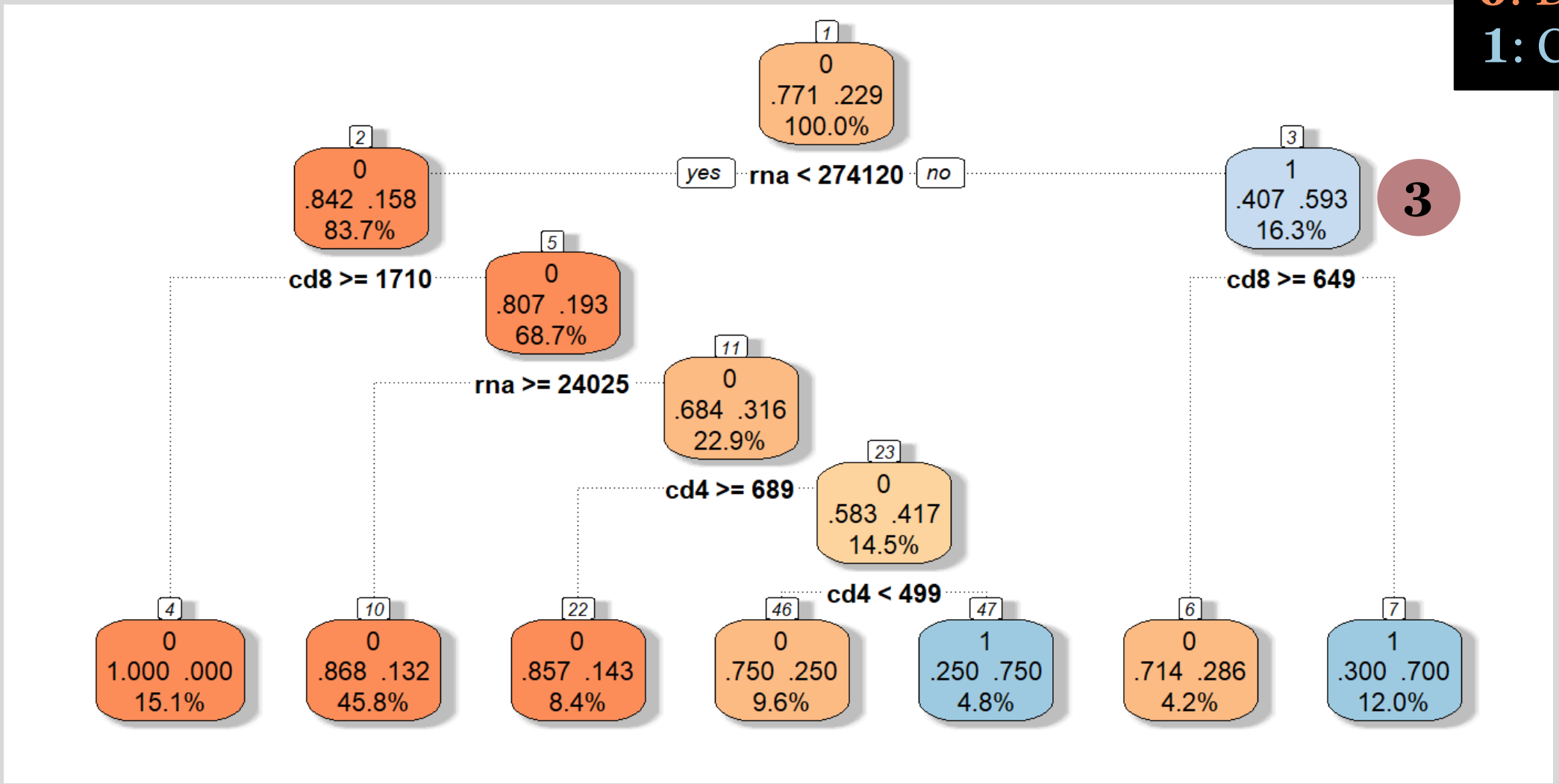
0: DP
1: CP


1
0
.771 .229
100.0%
1
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
5
0
.807 .193
68.7%
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

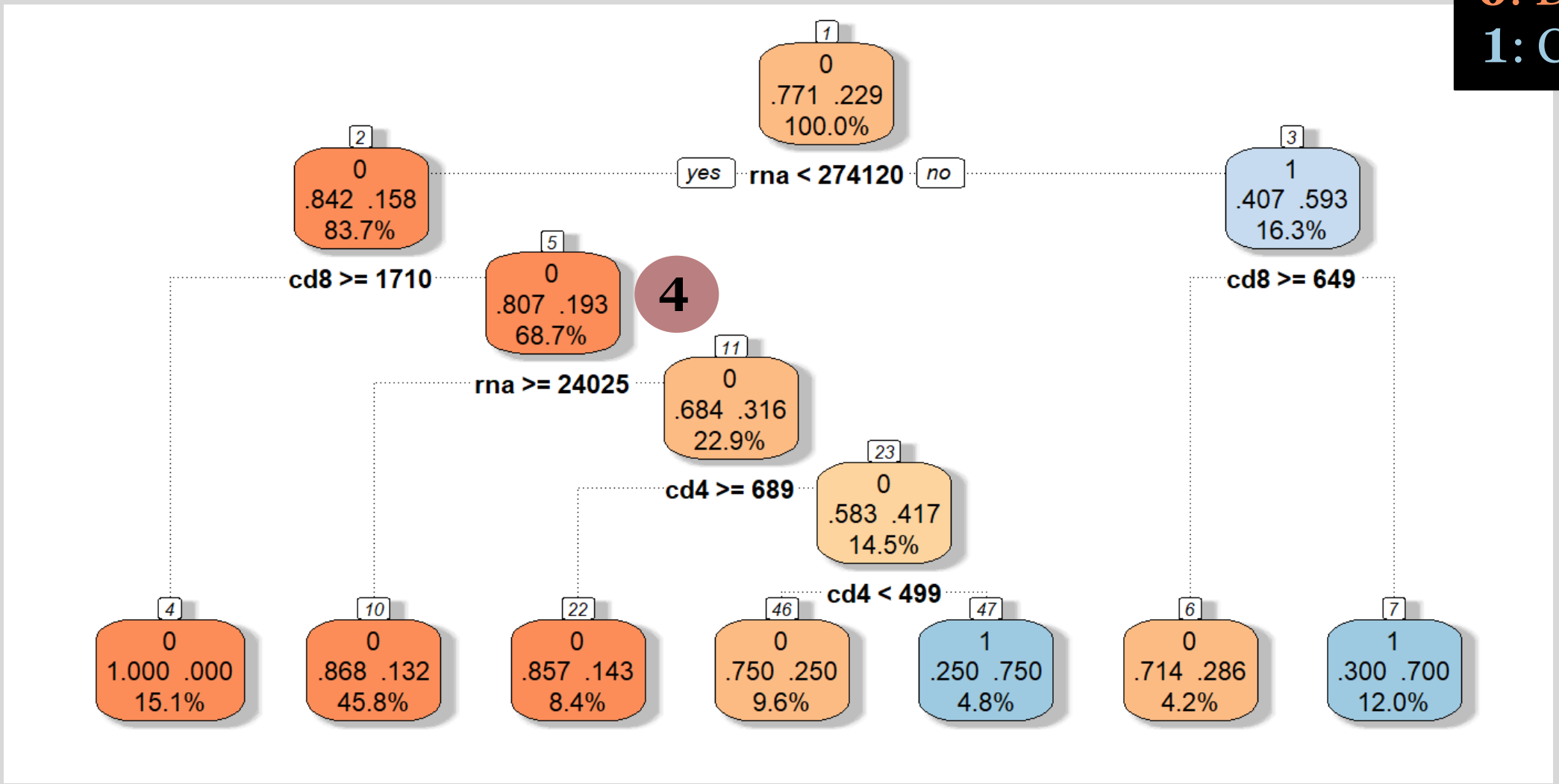
0: DP
1: CP


1
0
.771 .229
100.0%
yes
rna < 274120
no
2
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
5
0
.807 .193
68.7%
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

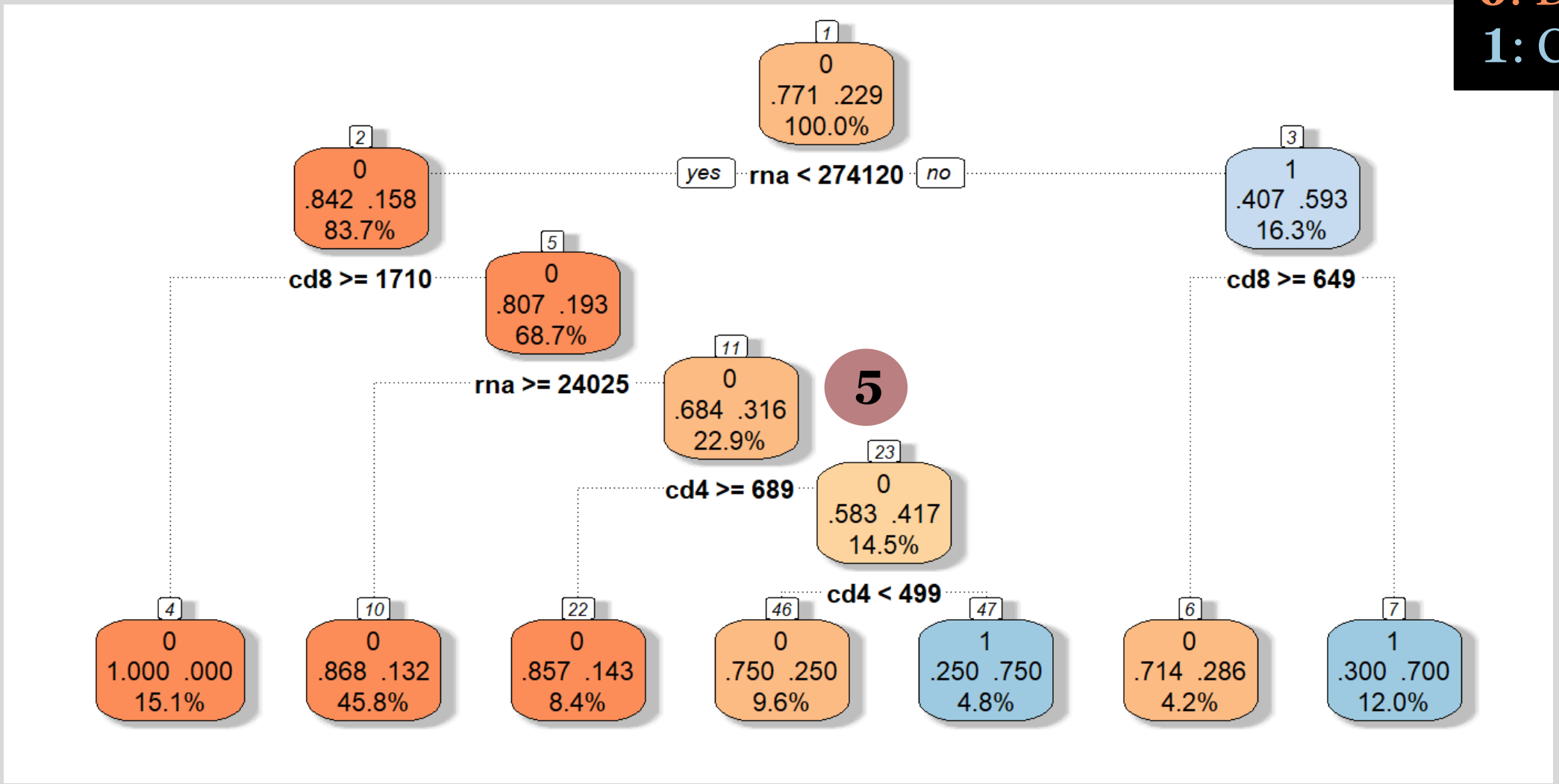
0: DP
1: CP


1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
3
cd8 >= 1710
5
0
.807 .193
68.7%
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
23
0
.583 .417
14.5%
cd4 >= 689
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

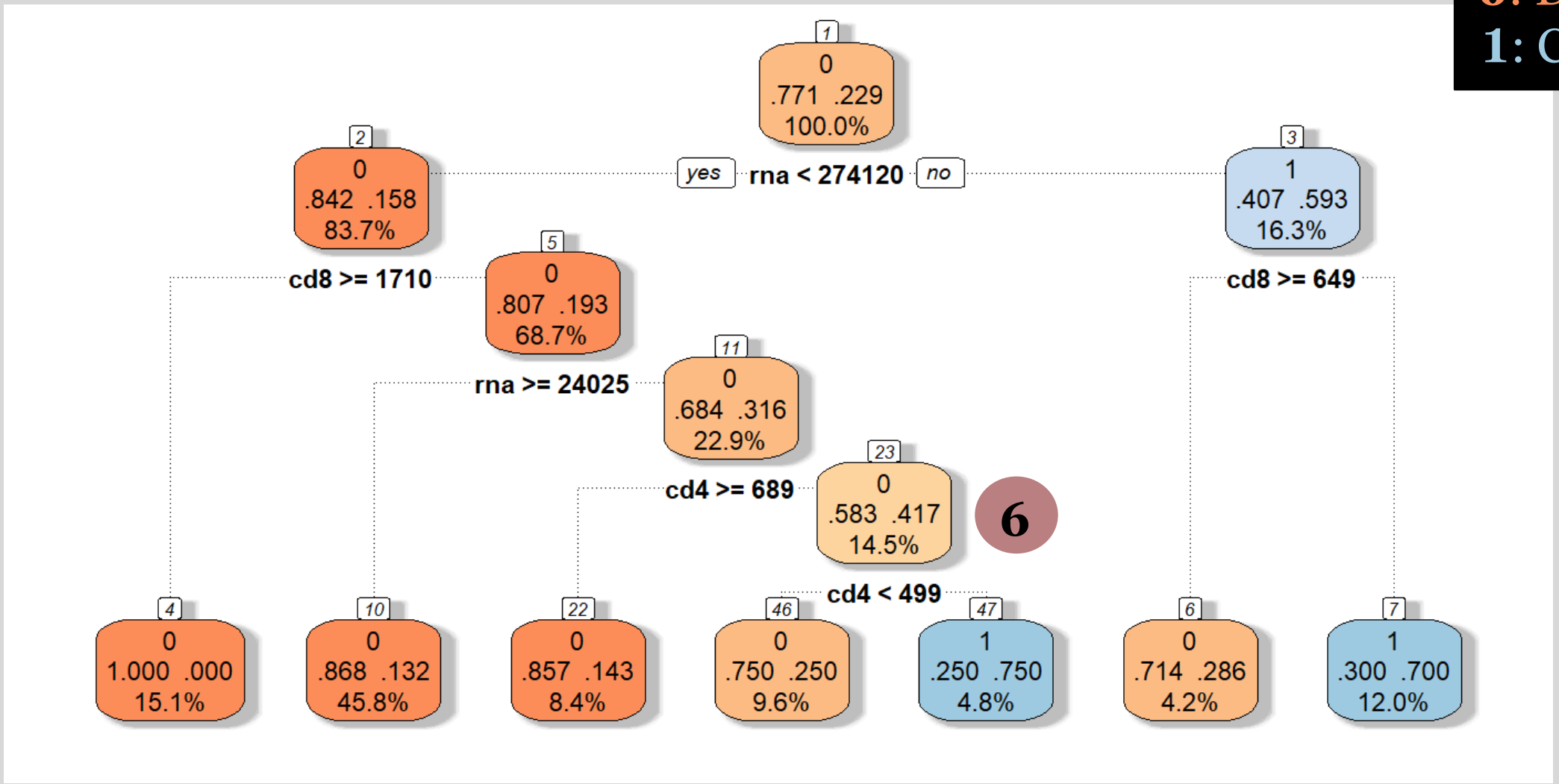
0: DP
1: CP


1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
5
0
.807 .193
68.7%
4
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

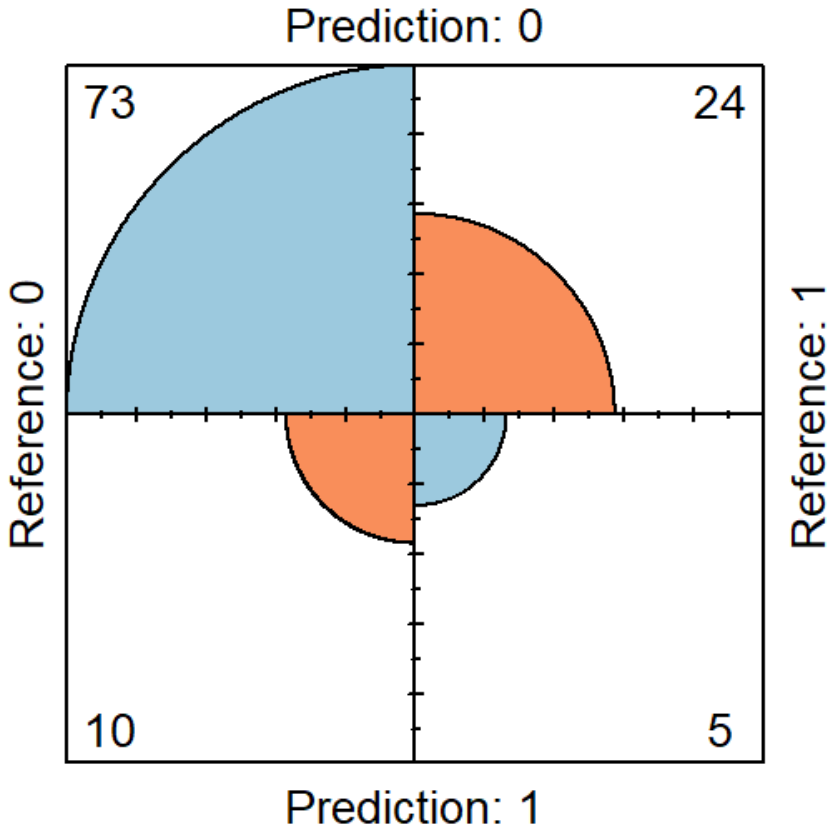
0: DP
1: CP


1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
5
0
.807 .193
68.7%
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
5
cd4 >= 689
23
0
.583 .417
14.5%
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

0: DP
1: CP

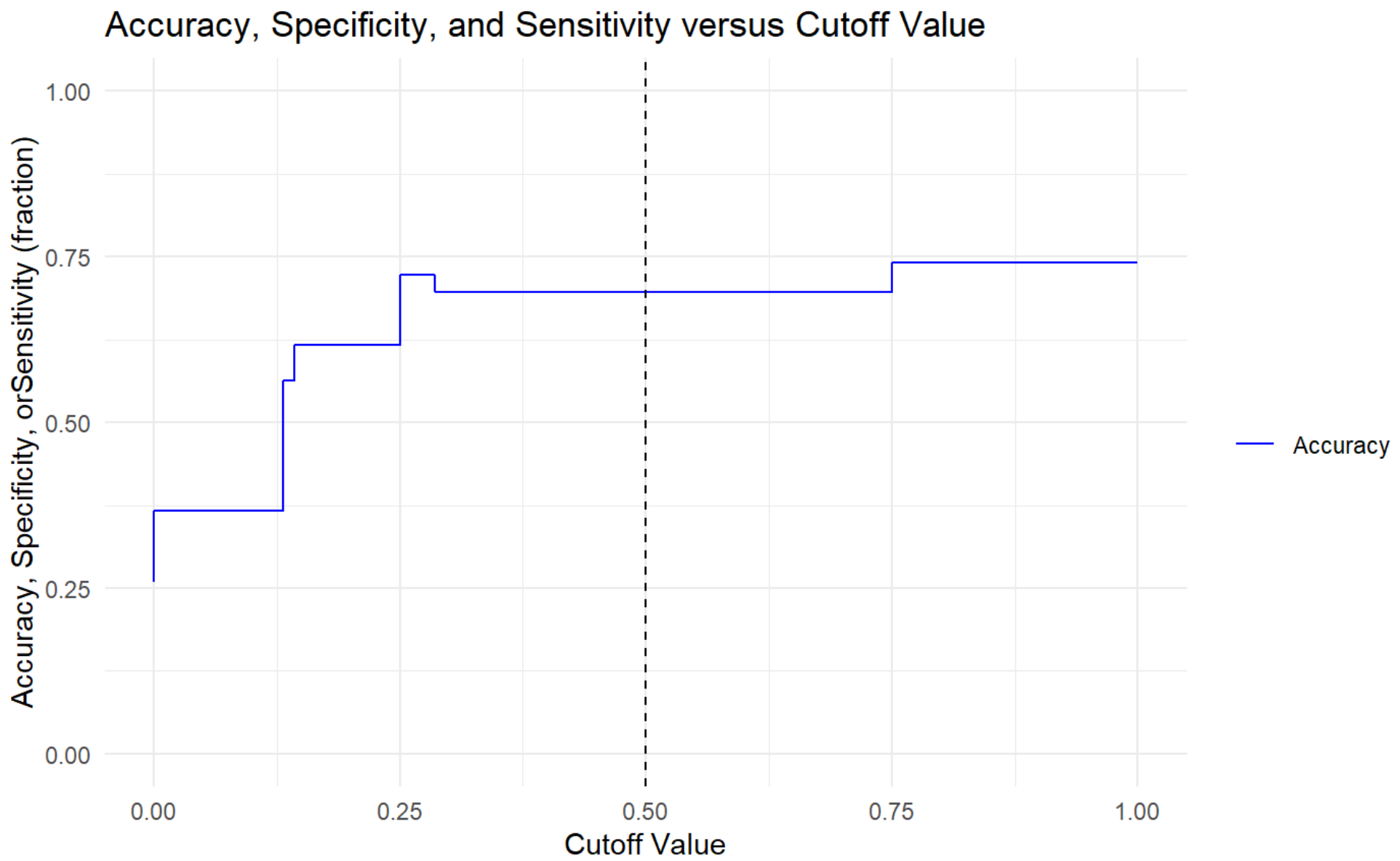

1
0
.771 .229
100.0%
yes
rna < 274120
no
2
0
.842 .158
83.7%
3
1
.407 .593
16.3%
cd8 >= 1710
5
0
.807 .193
68.7%
cd8 >= 649
rna >= 24025
11
0
.684 .316
22.9%
cd4 >= 689
23
0
.583 .417
14.5%
6
cd4 < 499
4
0
1.000 .000
15.1%
10
0
.868 .132
45.8%
22
0
.857 .143
8.4%
46
0
.750 .250
9.6%
47
1
.250 .750
4.8%
6
0
.714 .286
4.2%
7
1
.300 .700
12.0%

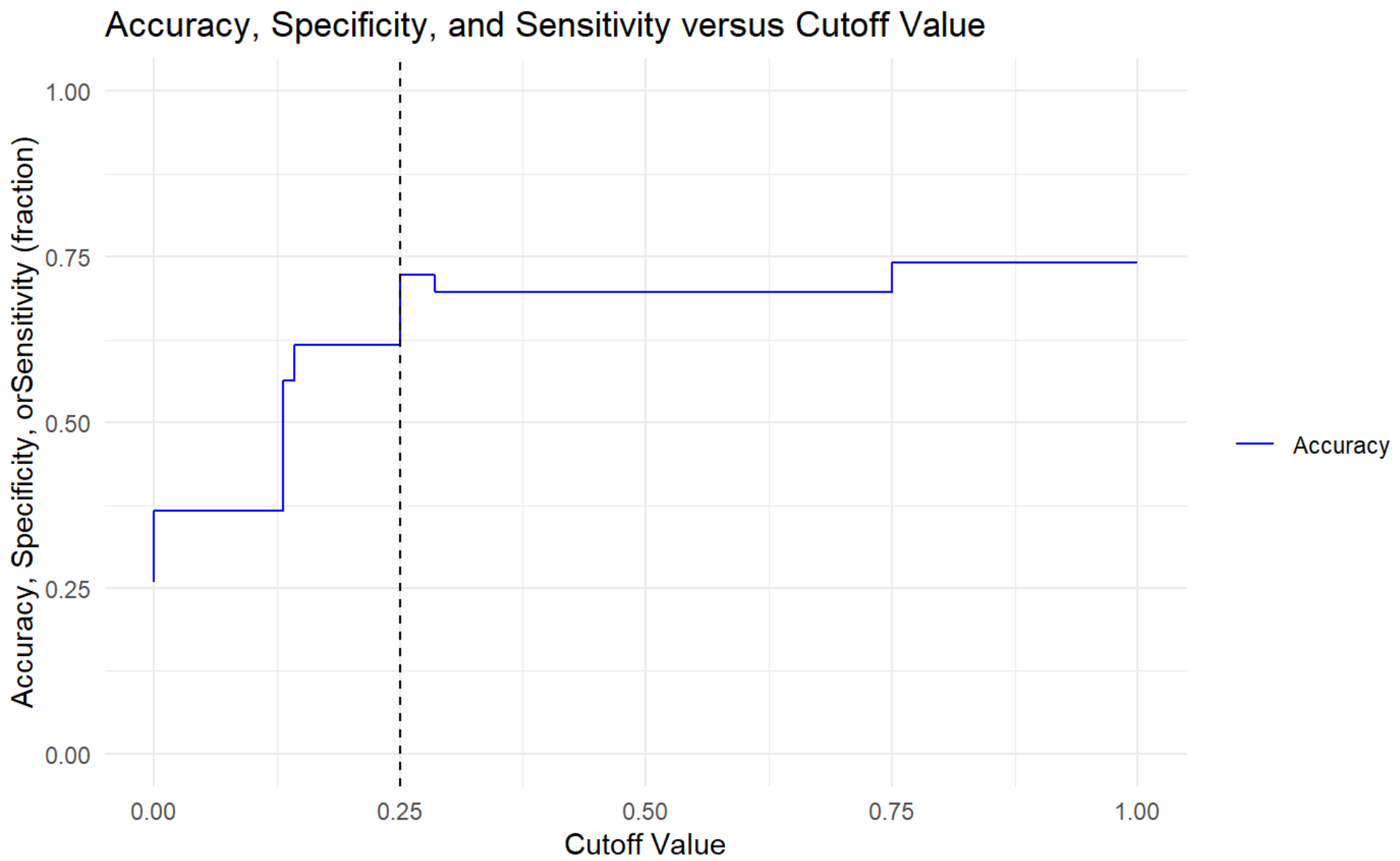
CONFUSION MATRIX

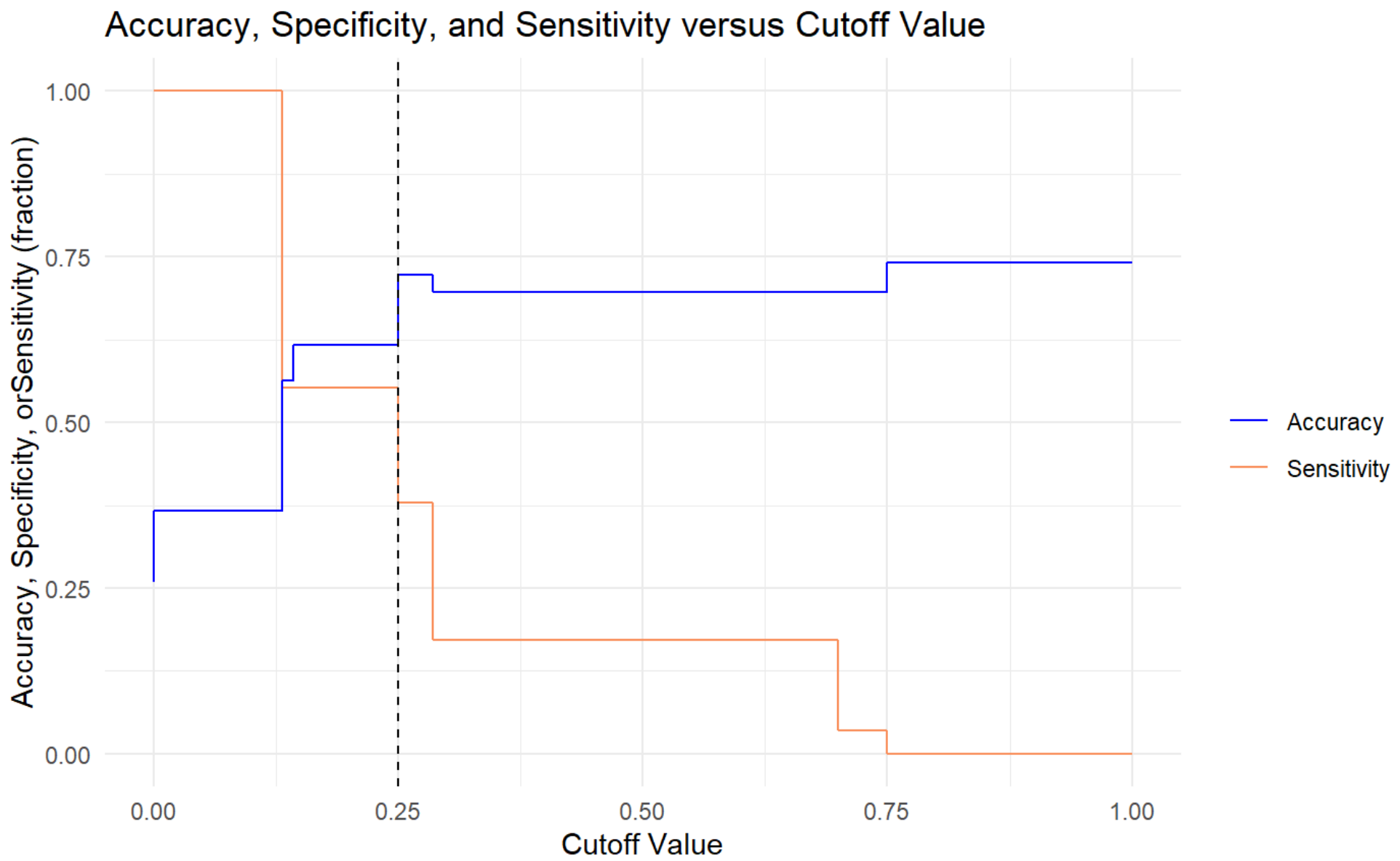


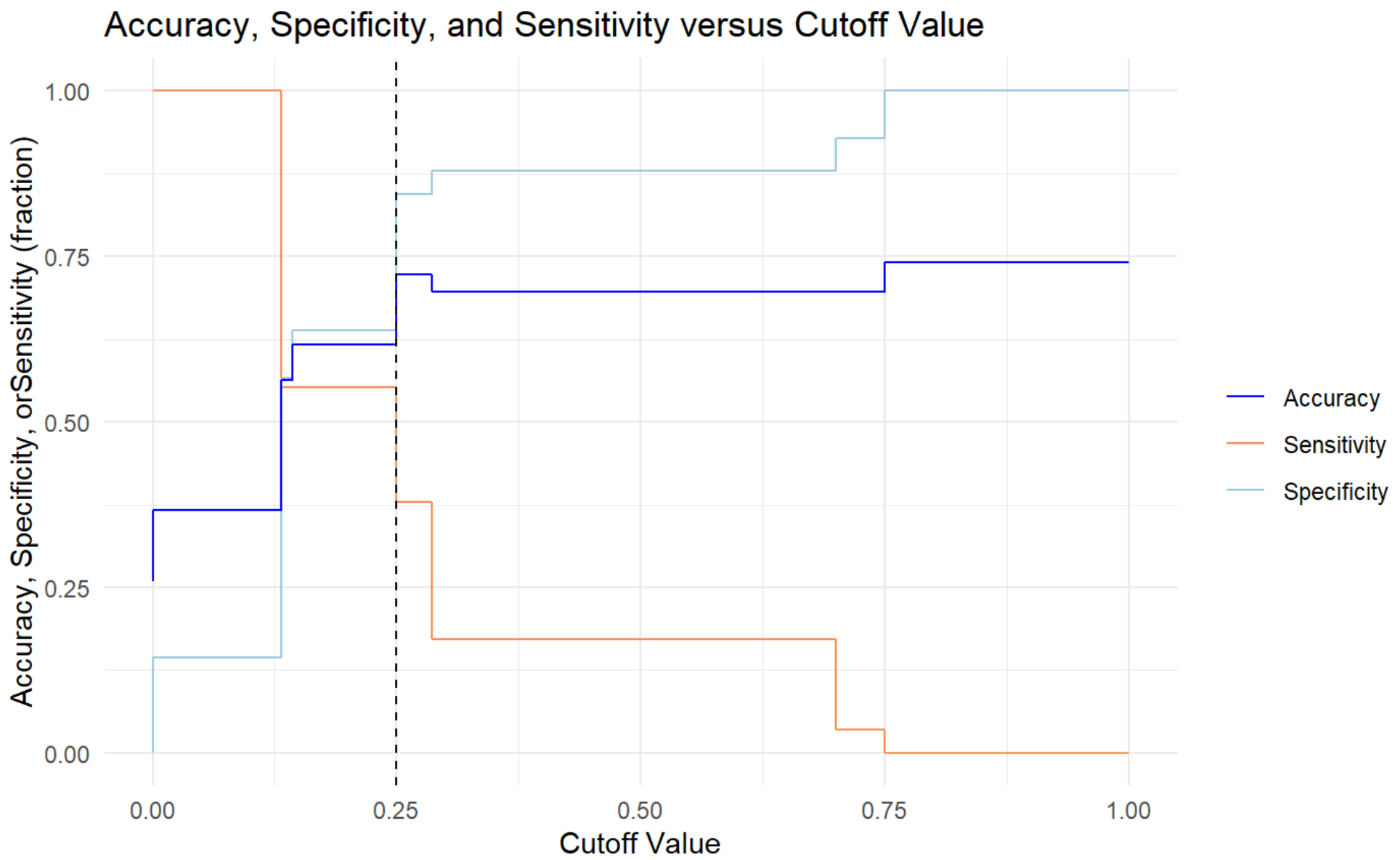
- Accuracy: **69.9%**
- Sensitivity: **17.2%**
- Specificity: **88.0%**

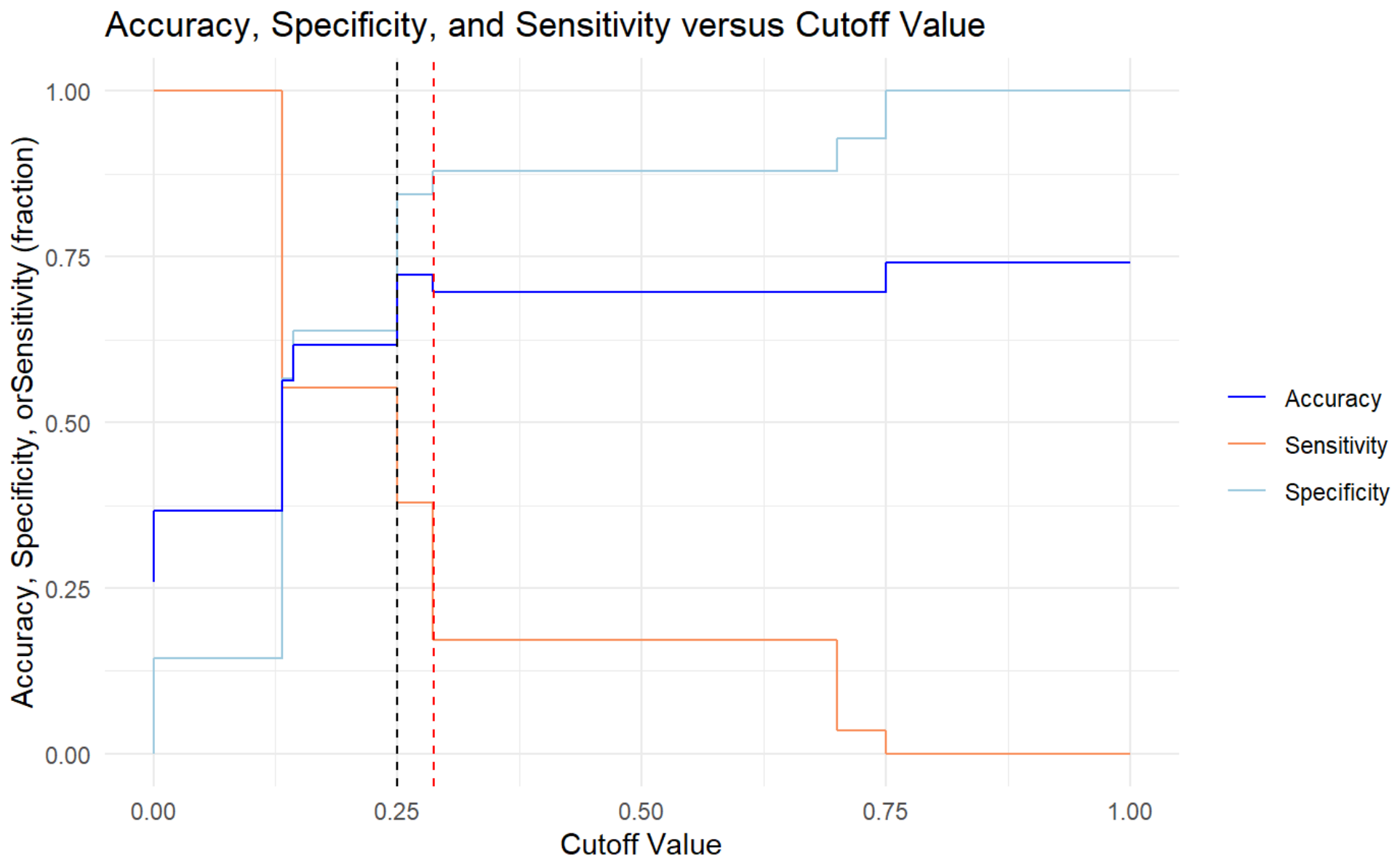
ACCURACY, SENSITIVITY & SPECIFICITY


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, orSensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy

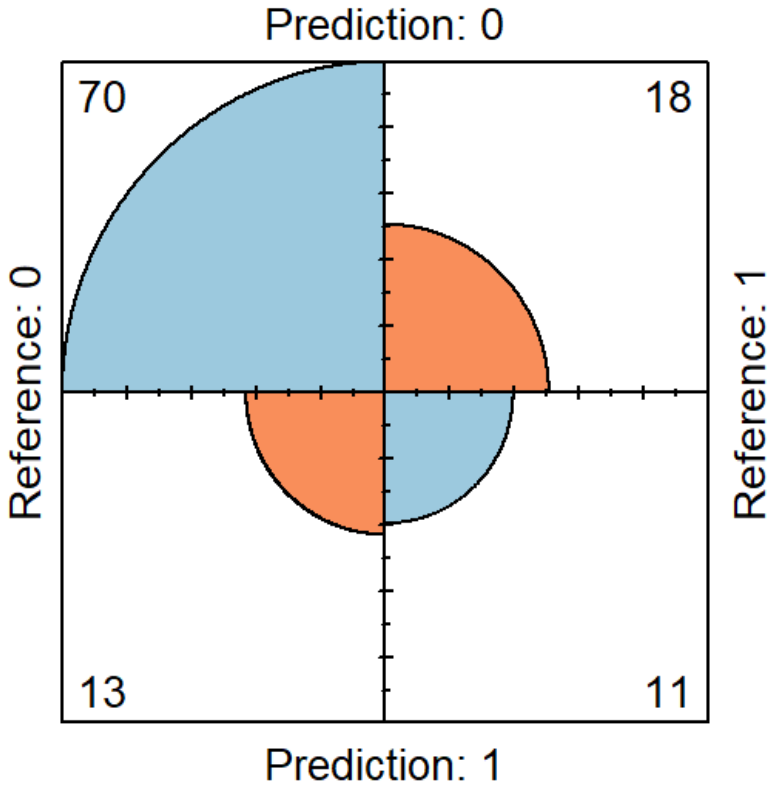

Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, orSensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, orSensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy
Sensitivity


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, orSensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy
Sensitivity
Specificity


Accuracy, Specificity, and Sensitivity versus Cutoff Value
Accuracy, Specificity, orSensitivity (fraction)
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Cutoff Value
Accuracy
Sensitivity
Specificity

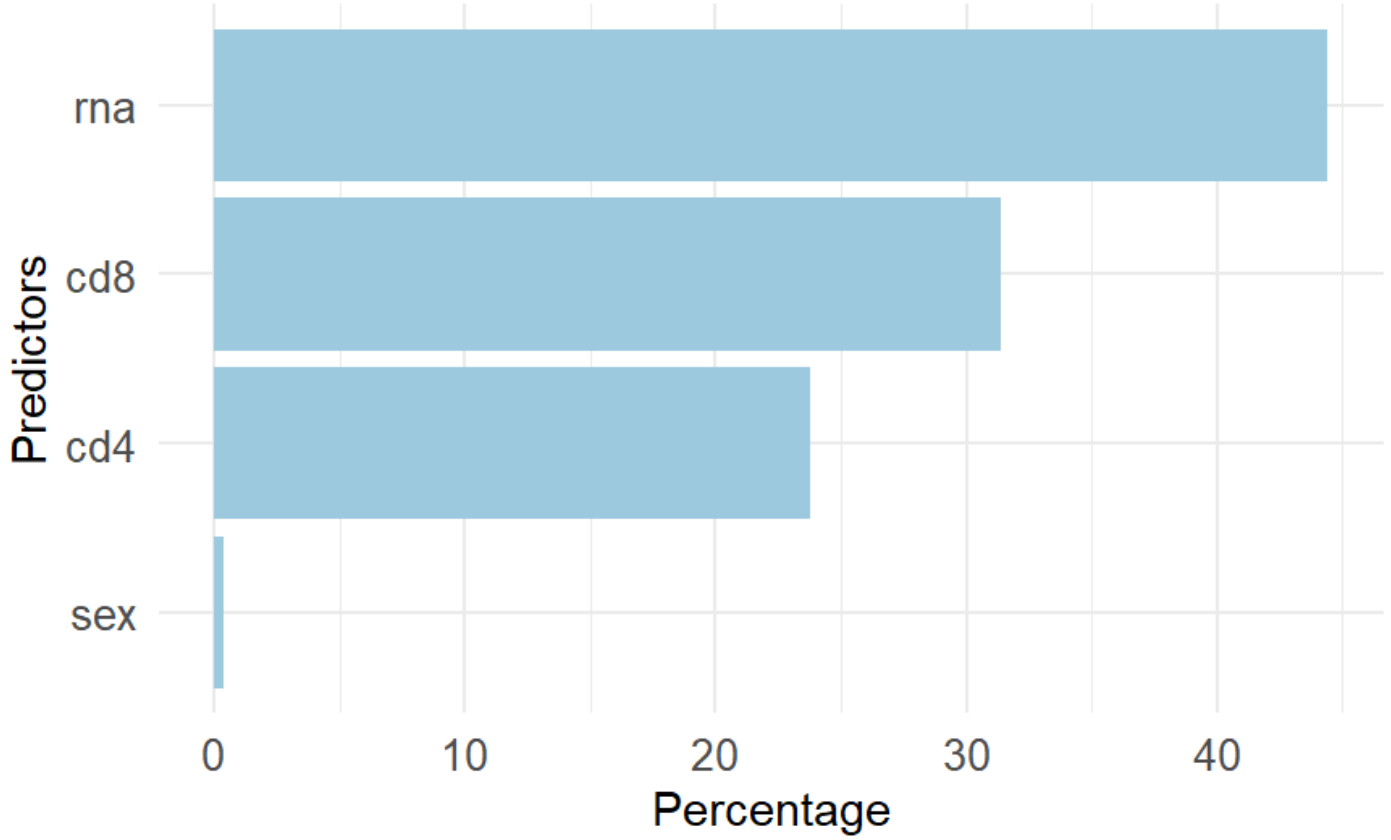
CONFUSION MATRIX



True Negative (TN)	**False Positive (FP)** *Type I Error*
False Negative (FN) *Type II Error*	**True Positive (TP)**

Accuracy: **72.3 %** | Sensitivity: **37.9%** | Specificity: **84.3%**

VARIABLE IMPORTANCE



MISSING DATA IMPUTATION

- Complete data is generated using a random forest-based algorithm (missForest package):
- Training Phase: n trees are grown using n random samples from existing data.
- Imputation Phase: These trees predict missing values.
- Iterative Process: The growth of the trees and imputation is done iteratively. Previous imputed values are utilized for imputing progressively better values.
- Final Imputations: The imputations are generated through averaging, these estimates make up the final imputations for the missing data.

- **WARNING:** No new information is being generated! The "new" values are simply computed from the existing records and carefully utilized until completion.

SUPPORTING RESEARCH

- Viral load among DP couples was lower than among CP couples
- CP showed elevated RNA viral load and decreased CD8 counts
- DP showed elevated CD8 counts and low RNA viral load
- Suggestive evidence of CD8 cells have different roles among DP and CP
  - DP = role in delay of disease progression
  - CP = destructive role

Conclusions

- CD4, CD8, and RNA viral load measurements are **INDEED able to distinguish** between DP and CP couples
  - **CD8** is the most important factor
  - RNA may be equally as important
    - Limitation → **30% missing** RNA viral load count
- In research, we checked correlation between RNA/CD8 to ensure that they wouldn't impact our models

Recommendations

- **Omit RNA viral load** if unable to retrieve original study measurements and rely on CD8 cell counts
- Include additional social and health **factors** in the dataset :
  - Age
  - Comorbidities
  - Treatments
  - Current stage in disease progression
  - Ethnic / genetic background
  - Heterosexual or homosexual relationship
- Improve **research methods** (30% of missing RNA data, 6.4% missing overall)

Thank you for your attention!

References

- Braunstein, S. L., Udeagu, C.-C., Bocour, A., Renaud, T., & Shepard, C. W. (2013). Identifying the correlates of membership in hiv-serodiscordant partnerships in new york city. *Sexually Transmitted Diseases*, *40*(10), 784–791. https://doi.org/10.1097/OLQ.0000000000000007
- *Cd4 cell count*. (n.d.). International Association of Providers of AIDS Care. Retrieved May 8, 2023, from https://www.iapac.org/fact-sheet/cd4-cell-count/
- *Cd4 lymphocyte count: Medlineplus medical test*. (n.d.). Retrieved May 8, 2023, from https://medlineplus.gov/lab-tests/cd4-lymphocyte-count/
- Denny, T. N., Skurnick, J. H., Palumbo, P., Perez, G., Monel, R., Stephens, R., Kennedy, C. A., & Louria, D. B. (1998). CD3+CD8+ cell levels as predictors of transmission in human immunodeficiency virus-infected couples: A report from the heterosexual HIV transmission study. *International Journal of Infectious Diseases*, *2*(4), 186–192. https://doi.org/10.1016/S1201-9712(98)90050-9
- Hambissa, Y. M., & Wolday, D. (2016). Immunological profile: Cd4, cd8, hiv cofactors and viral load in hiv discordant couples when compared with concordant couples. *Journal of Clinical & Cellular Immunology*, *07*(06). https://doi.org/10.4172/2155-9899.1000468
- *Hiv/aids*. (n.d.). Retrieved May 8, 2023, from https://www.who.int/health-topics/hiv-aids
- Mehra, B., Bhalla, P., Rawat, D., & Kishore, J. (2015). A study of Hiv-concordant and -discordant couples attending voluntary counselling and testing services at a tertiary care center in North India. *Indian Journal of Public Health*, *59*(4), 306. https://doi.org/10.4103/0019-557X.169664
- *Viral load*. (n.d.). International Association of Providers of AIDS Care. Retrieved May 8, 2023, from https://www.iapac.org/fact-sheet/viral-load/