We try to find the most relevant questions that predict whether a child will finish the program. In our analysis, 0 corresponds to a child not finishing the program, while 1 corresponds to a child finishing the program. For question answers: Not at all true = 1, Not really true = 2, Unclear = 3, Sort of true = 4, True = 5, and Very true = 6. VHemA1 corresponds to question 1, VHemA2 corresponds to question 2, etc.

All data is shown in FullAnalysis.pdf. The methods used to identify these questions yielded the following:

Linear Regression: (Estimate is trend, positive estimate is positive correlation, negative estimate is negative correlation; Pr(>|t|) is percentage that the observation has no correlation)

Significant Questions: Question 22, Question 33, Question 34, Question 34, Question 45, Question 47.

# Coefficients:

# Estimate Std. Error t value Pr(>|t|)

# VHemA34 -0.0959147 0.0428579 -2.238 0.0275 \* (significant, **positive** correlation)

# VHemA39 0.0895486 0.0450175 1.989 0.0495 \* (significant, **negative** correlation)

# VHemA45 -0.0976137 0.0433340 -2.253 0.02656 \* (significant, **negative** correlation)

# VHemA50 0.1063727 0.0510369 2.084 0.0398 \* (significant, **negative** correlation)

# VHemA56 -0.1586908 0.0625639 -2.536 0.0128 \* (significant, **negative** correlation)

# ---

# Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

#

# Residual standard error: 0.3127 on 96 degrees of freedom

# Multiple R-squared: 0.5784, Adjusted R-squared: 0.3281

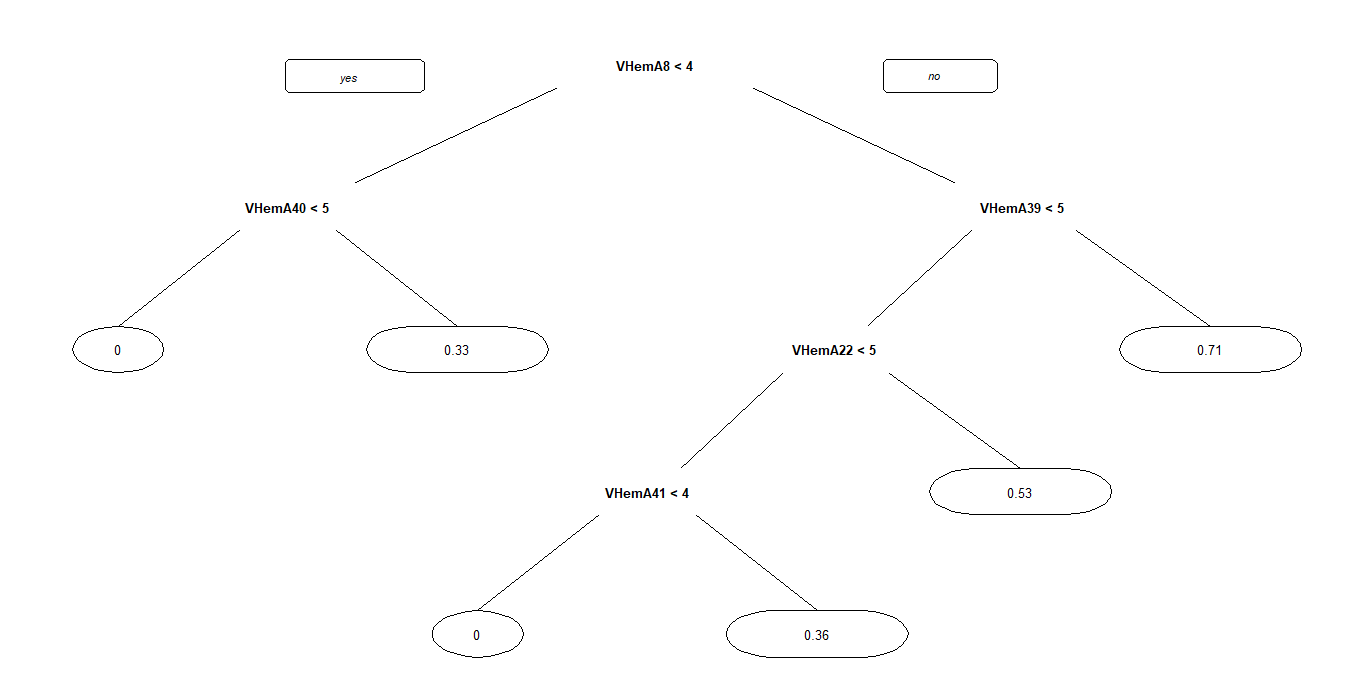
# F-statistic: 2.311 on 57 and 96 DF, p-value: 0.0001434

Logistic Regression: Proven to be ineffective due to the number of questions and will not be used to determine which questions are significant.

Decision Tree: Decides whether a child finished the program based on their question responses.

The higher the question on the branches, the most significant it was viewed by the program. To interpret the diagram, if a child’s answer to question 6 was less than 5.5 (less than 6, which corresponds to anything but Very true), then the tree would continue to the left. If that same child had an answer which was greater than 5.5 for question 47 (that is, they answered Very true), the tree would finish by taking the right branch, which ends on 0.57. That is, the probability that the child finished the program, as predicted by the decision tree, is 0.57.

Significant questions: Question 8, Question 39, Question 22, Question 48, Question 6, Question 47, Question 24, Question 56.



Random Forest: Shown here is node purity from question to question. More useful variables will have higher purity values.

Significant Questions: Question 39, Question 40, Question 8, Question 20.

# VHemA39 1.39139176

# VHemA40 1.35367497

# VHemA8 1.28774432

# VHemA20 1.03858004

# VHemA10 0.87624209

# VHemA34 0.78578905

# VHemA47 0.76016285

# VHemA50 0.70745407

# VHemA1 0.68565575

# VHemA6 0.56988831

# VHemA26 0.55500039

# VHemA2 0.5447657

# VHemA22 0.53557431

# VHemA51 0.53083424

# VHemA38 0.48583843

# VHemA4 0.45350114

# VHemA18 0.39140931

# VHemA30 0.37850037

# VHemA13 0.3611777

# VHemA7 0.35989265

# VHemA11 0.32595554

# VHemA36 0.29300878

# VHemA16 0.28403418

# VHemA27 0.28230003

# VHemA29 0.27709945

# VHemA52 0.27694413

# VHemA55 0.27087457

# VHemA54 0.24915745

# VHemA31 0.24802985

# VHemA35 0.24012515

# VHemA45 0.22894488

# VHemA41 0.22131756

# VHemA21 0.21941575

# VHemA57 0.21909235

# VHemA19 0.21561636

# VHemA37 0.20153494

# VHemA17 0.19263189

# VHemA5 0.18913594

# VHemA48 0.18630651

# VHemA25 0.18479105

# VHemA32 0.17725446

# VHemA56 0.16557068

# VHemA24 0.16003654

# VHemA53 0.15778443

# VHemA12 0.15396226

# VHemA3 0.14708521

# VHemA46 0.13707764

# VHemA28 0.13453587

# VHemA9 0.13214555

# VHemA23 0.12672899

# VHemA42 0.12659343

# VHemA15 0.12439152

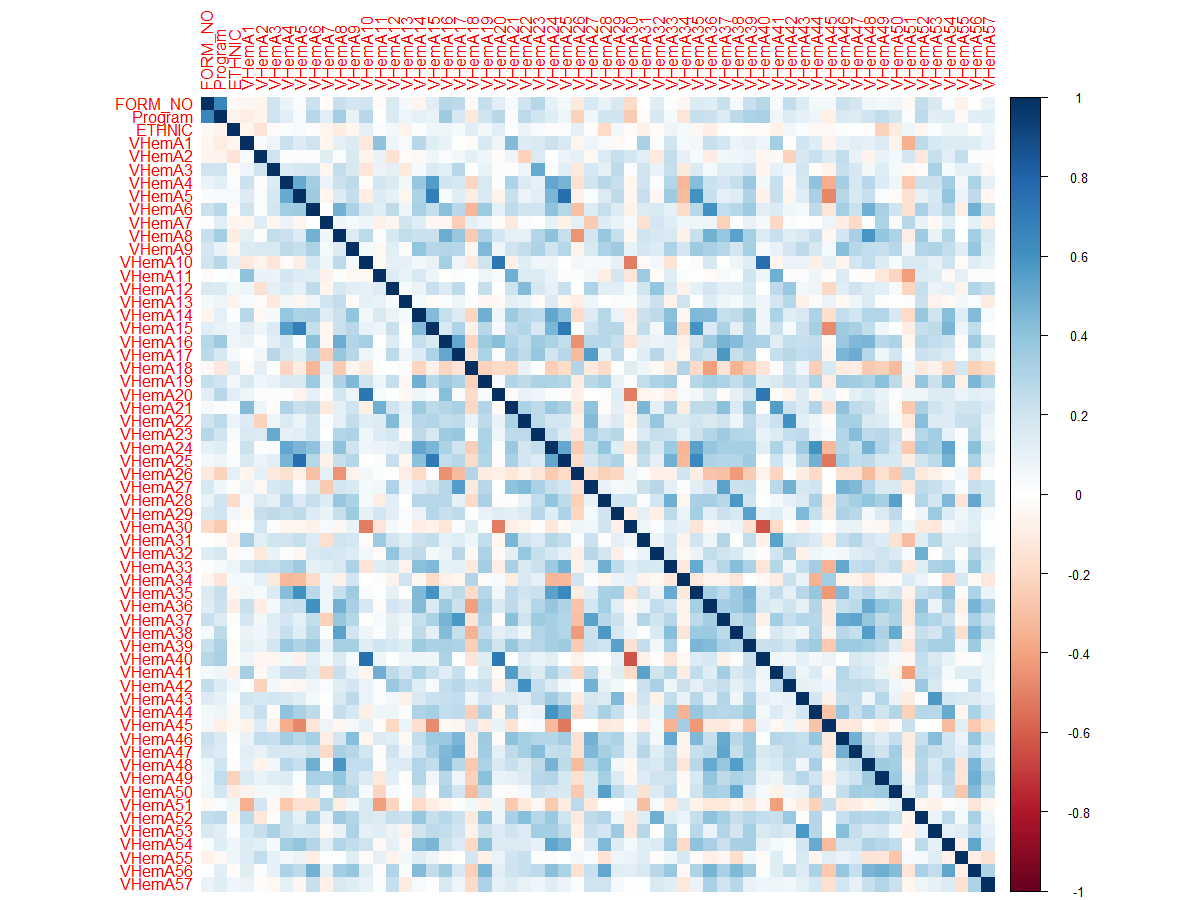
# VHemA33 0.11141616

# VHemA14 0.0971341

# VHemA43 0.09459412

# VHemA49 0.08239974

# VHemA44 0.0691158

Correlation Plot: Less used for finding significant questions and more for finding relationships between questions.

Significant Questions: This table shows which questions were marked as significant by each method. Though we would weigh the importance of Linear Regression more, and the importance of the Decision Tree less.

|  |  |  |  |
| --- | --- | --- | --- |
| Question | Linear Regression | Decision Tree | Random Forest |
| 1 |  |  |  |
| 2 |  |  |  |
| 3 |  |  |  |
| 4 |  |  |  |
| 5 |  |  |  |
| 6 |  | • |  |
| 7 |  |  |  |
| 8 |  | • | • |
| 9 |  |  |  |
| 10 |  |  |  |
| 11 |  |  |  |
| 12 |  |  |  |
| 13 |  |  |  |
| 14 |  |  |  |
| 15 |  |  |  |
| 16 |  |  |  |
| 17 |  |  |  |
| 18 |  |  |  |
| 19 |  |  |  |
| 20 |  |  | • |
| 21 |  |  |  |
| 22 |  | • |  |
| 23 |  |  |  |
| 24 |  | • |  |
| 25 |  |  |  |
| 26 |  |  |  |
| 27 |  |  |  |
| 28 |  |  |  |
| 29 |  |  |  |
| 30 |  |  |  |
| 31 |  |  |  |
| 32 |  |  |  |
| 33 |  |  |  |
| 34 | • |  |  |
| 35 |  |  |  |
| 36 |  |  |  |
| 37 |  |  |  |
| 38 |  |  |  |
| 39 | • | • | • |
| 40 |  |  | • |
| 41 |  |  |  |
| 42 |  |  |  |
| 43 |  |  |  |
| 44 |  |  |  |
| 45 | • |  |  |
| 46 |  |  |  |
| 47 |  | • |  |
| 48 |  | • |  |
| 49 |  |  |  |
| 50 | • |  |  |
| 51 |  |  |  |
| 52 |  |  |  |
| 53 |  |  |  |
| 54 |  |  |  |
| 55 |  |  |  |
| 56 | • |  |  |
| 57 |  |  |  |