Decision Tree and Entropy Assignment

Robert Stephen and Yinghao Tan

1. Variables selected:

**Age:** Because there could be a chance that an older person would have a higher blood pressure when compared to the higher audience.

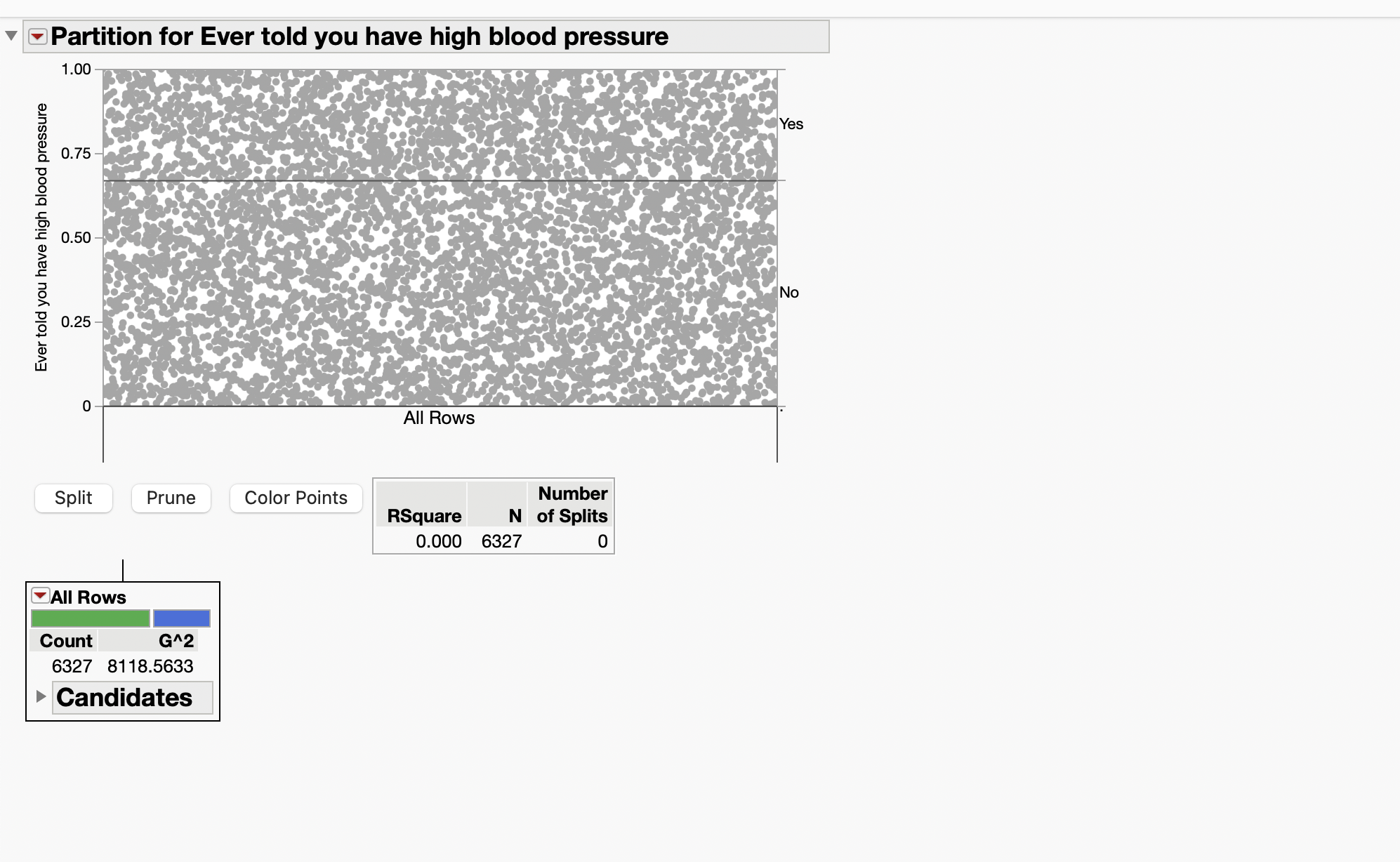
**Sleep Hours:** Sleep can cause inflammations in the human body including blood pressure, so a person with fewer hours of sleep might have a higher blood pressure than a person that sleeps enough hours.

**Weight:** Having higher weight can lead to higher insulin in the bloodstream that could cause an individual to have higher blood pressure.

**Body Mass Index**: Because higher BMI can lead to higher hypertension in the blood which leads to higher blood pressure.

**Minutes Vigorous-Intensity work**: Because the more vigorous activities and individual does the less likehood for that individual to have a higher blood pressure. This is because intense work can lower the insulin within the body, reducing blood pressure.

**General Health Condition**: This is a categorical variable and we decided to choose these variables, because if an individual has a poor health condition, then it would be easier for this individual to get higher blood pressure even at a younger age group.



In the first split, the factor “Sleep Hours” has been used.

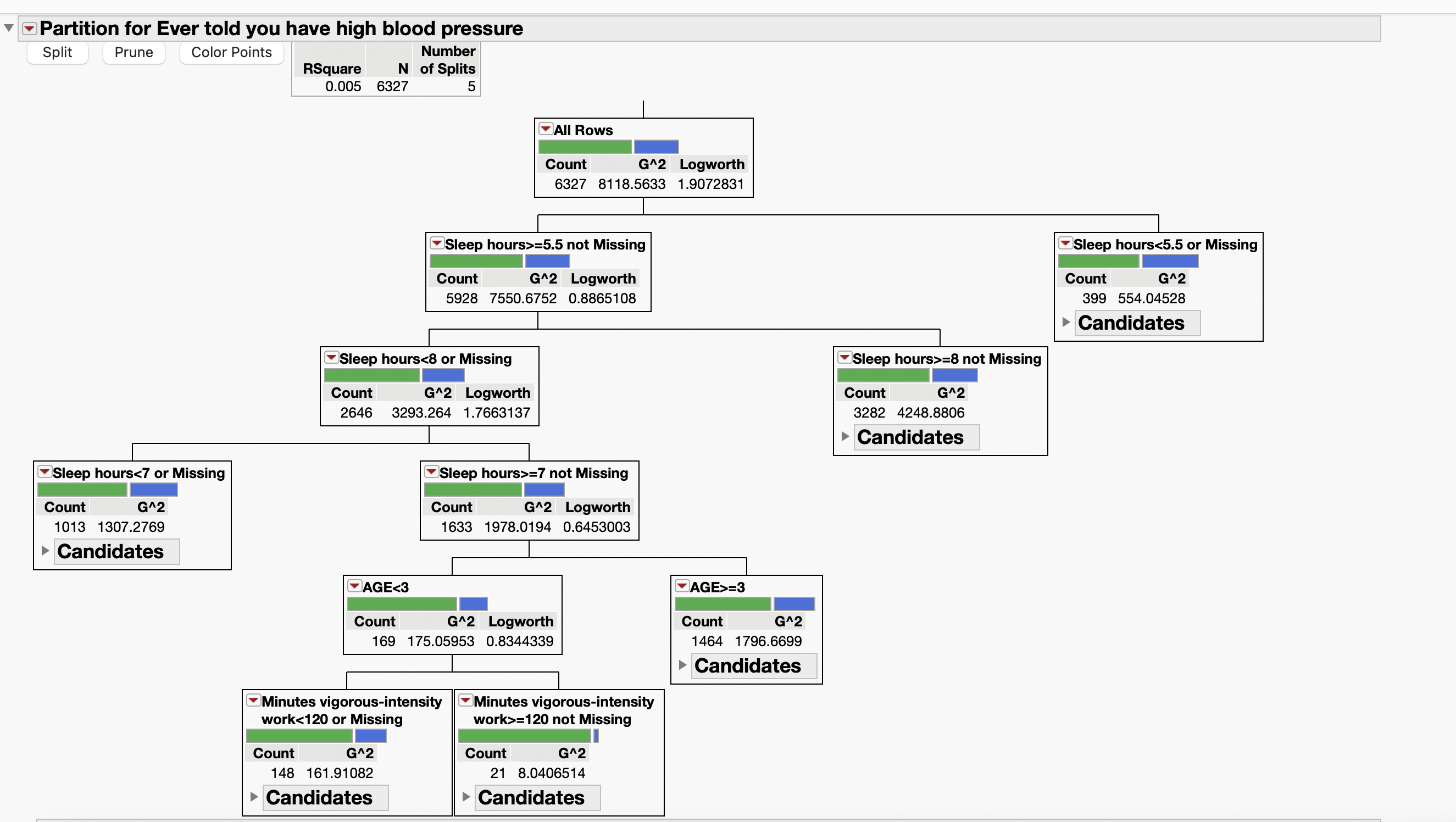
Entropy Calculation:

Entropy = -1 \* ((P (Type I objects)) \* log2(P (Type I objects)) + (P (Type II objects)) \* log2(P (Type II  
objects))

Entropy = (-1 \* ((0.6744) \* log2(0.5842)) + ((0.3247) \* log2(0.4133)))

= 0.937

Entropy is a measure of disorder. More specifically, entropy is a scientific concept and a measurable physical property mostly associated with a state of disorder, randomness, and uncertainty. Looking at the entropy value of the leaf nodes, it is noticeable that the leaf nodes are impure.



When performing multiple splits in the decision tree, we observe that the R square value does not increase significantly as the number of splits increases. The R square value increases only by 0.001 per split. Factors like “Sleep Hours,” “Age” and “Minutes Vigorous-Intensity work” are being used for the split.

In the first split the factor: “Minutes Vigorous- Intensity work” is being used.

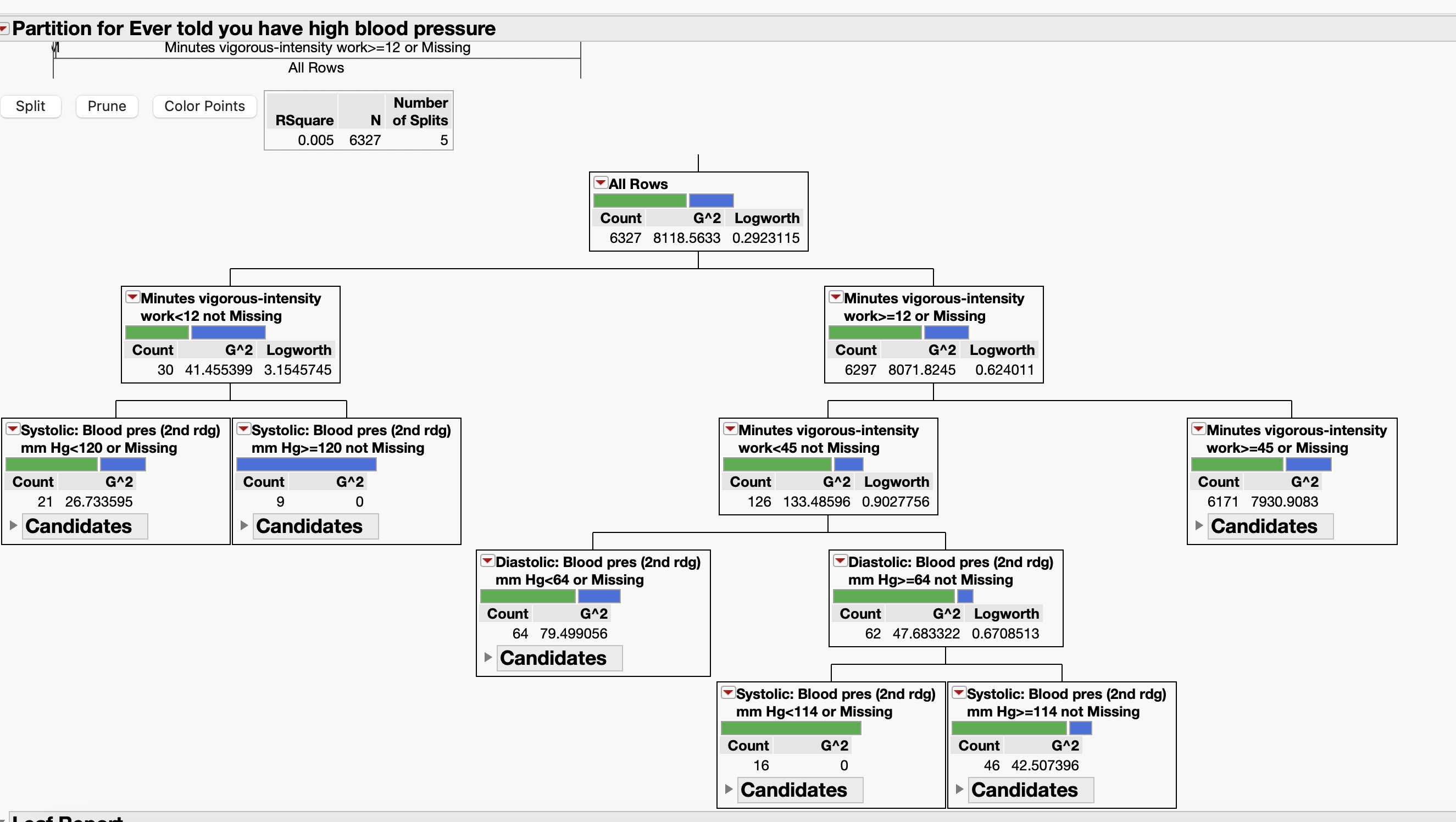
New Entropy Calculation with actual blood pressures:

Entropy = -1 \* ((P (Type I objects)) \* log2(P (Type I objects)) + (P (Type II objects)) \* log2(P (Type II objects))

Entropy= (-1 \* ((0.4732) + log2(0.6697)) + ((0.5268 \* log2(-0.3294)))

= 1.118

Entropy is a measure of disorder. More specifically, entropy is a scientific concept and a measurable physical property mostly associated with a state of disorder, randomness, and uncertainty. From the entropy value, it is noticed that the leaf nodes are still impure. The entropy value has increased from before, making the impurity of the leaf node even stronger.



The R value does not increase significantly. For every split, the R value increases by 0.001. Like what has been observed in the previous splits. Factors that have been used in the splits are: “Minutes Vigorous Intensity work,” “Systolic: Blood Pres (2nd rdg) mm Hg”, and Diastolic: Blood Pres (2nd rdg)”.

RESULTS CONCLUSION: The performance of the decision tree, as indicated by the slight increase in the R-squared value and the increase in entropy of the additional splits, suggests that although the selected variables are theoretically related to blood pressure, their predictive power in this model and dataset is limited. The presence of impurities in the leaf nodes and the minimal improvement in model fitting with more splits suggests the complexity of predicting blood pressure levels and that a more nuanced selection of variables or a different modeling approach may be required to more effectively capture underlying patterns.

In summary, the decision tree results highlight the complexity of modeling blood pressure levels using the selected variables, reflecting the challenges of achieving significant predictive accuracy and dealing with data impurities despite theoretically sound variable selection.