## Data Exploration

This data has already been cleaned up and is largely usable because it was taken from the UCI Machine Learning Repository. Since there are no missing values, the only tidying up I had to do was go around and adjust the variable classes, assigning factor labels to some of the encoded variables so they more accurately mirror the original integer coding. The table of contraceptive method was then explored to see the proportion with more informative labels as shown below:

```
no-use  long-term short-term
    42.7       22.6       34.7
```

The data-set has been inspected to check whether the features and descriptions outlined in the documentation are aligning with the data-set. Dimension of the data set is (1473,10), 1473 representing rows and there were 10 columns present.

Boxplots are a valuable tool for exploring and understanding the characteristics of the data before proceeding with model building in machine learning tasks. The boxplot of our dataset is shown in Fig.1 below

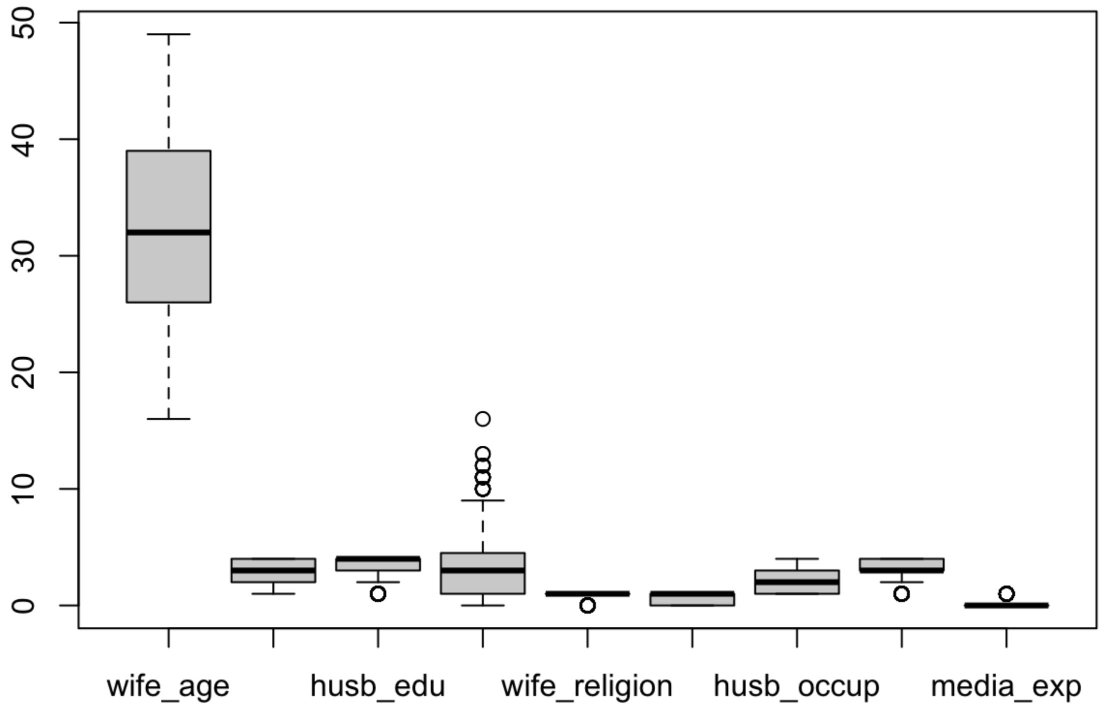


Fig.1: Boxplot of the dataset

The boxplot of the data set shows that the median wife age is 32, 25% of the wives are older than 39 and 25% are younger than 27. There are no outliers in the wifes' ages. All other variables are relatively normally distributed.

In machine learning, the process of changing the features to have a comparable scale is referred to as scaling a dataset. This is significant because similar-scale features yield superior results for many machine learning techniques. The Fig.2 below is the boxplot of the dataset after scaling.

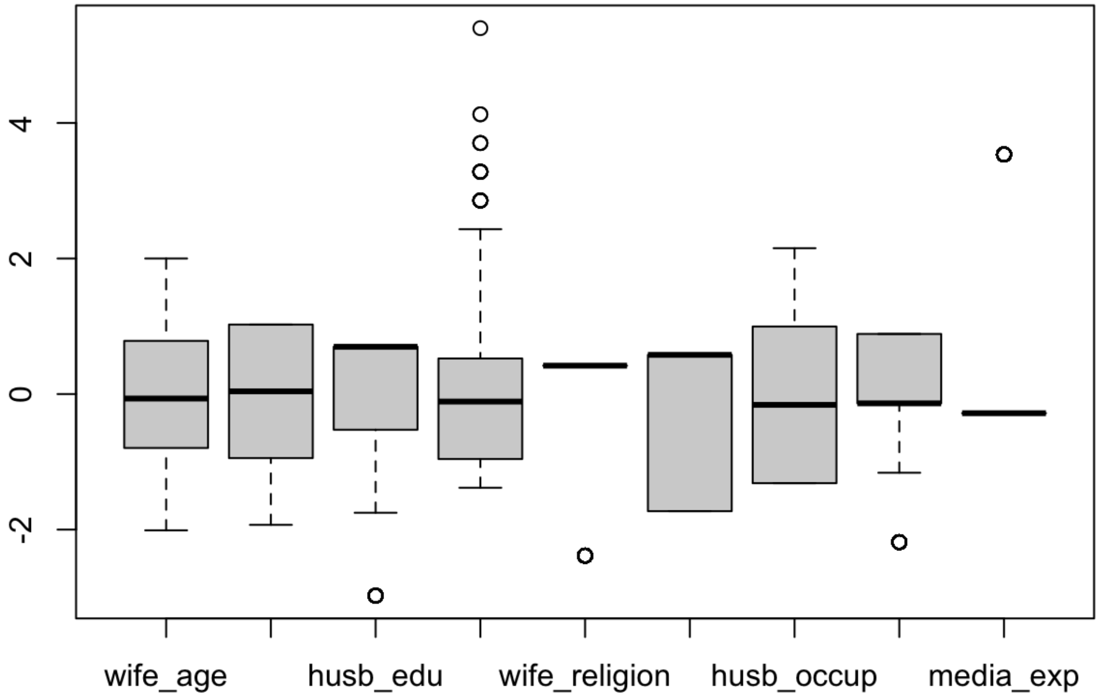


Fig.2: Boxplot of the dataset after scaling

Histograms are one of the best and accurate methods to visualize numerical data. It shows the frequency distribution within an attribute.

I wanted to view this data first in order to understand what I had and what engineering aspects would be worth investigating before I built any models on it. Although the number of people in each education level varies, our data first led me to believe that the likelihood of a husband and wife using either long- or short-term contraception increased with their educational attainment. I plotted the frequency of each education level for both men and women in the data set to make sure.

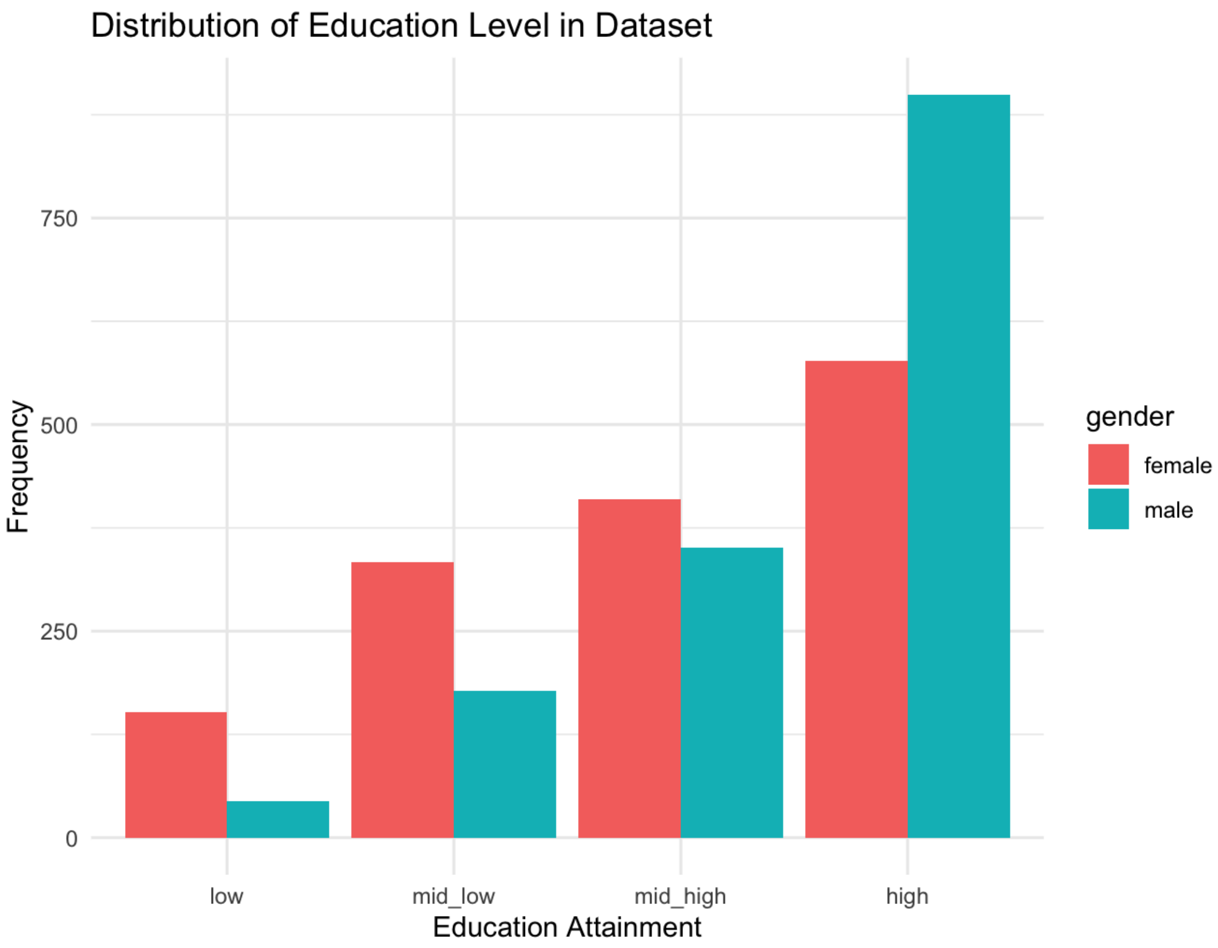


Fig.3: Distribution of Educational Level

Given that the Indonesians surveyed educational levels are clearly not distributed uniformly as shown in Fig.3 above, I plotted the husband's and wife's education levels independently versus the type of contraception used.

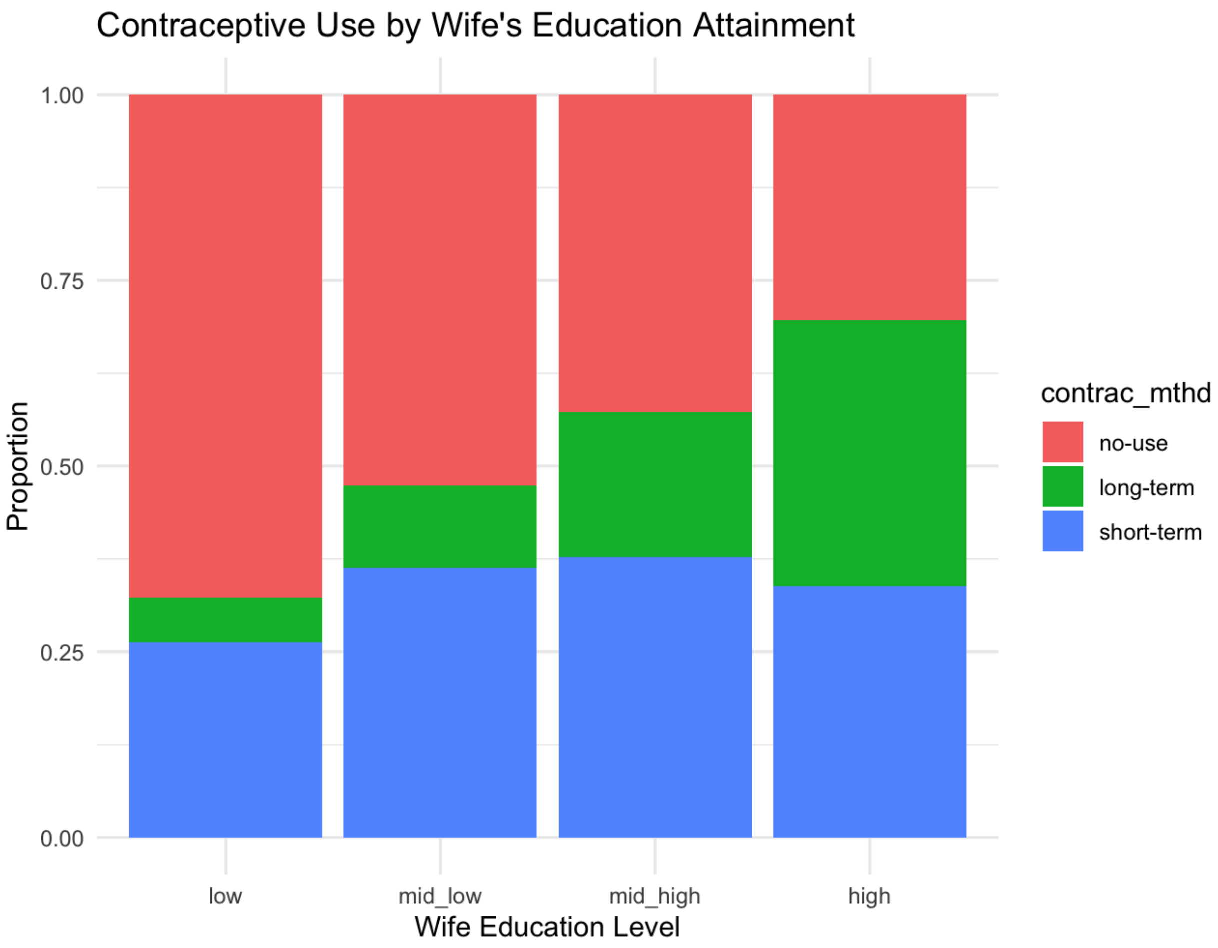


Fig.4: Contraceptive use by wife’s education level

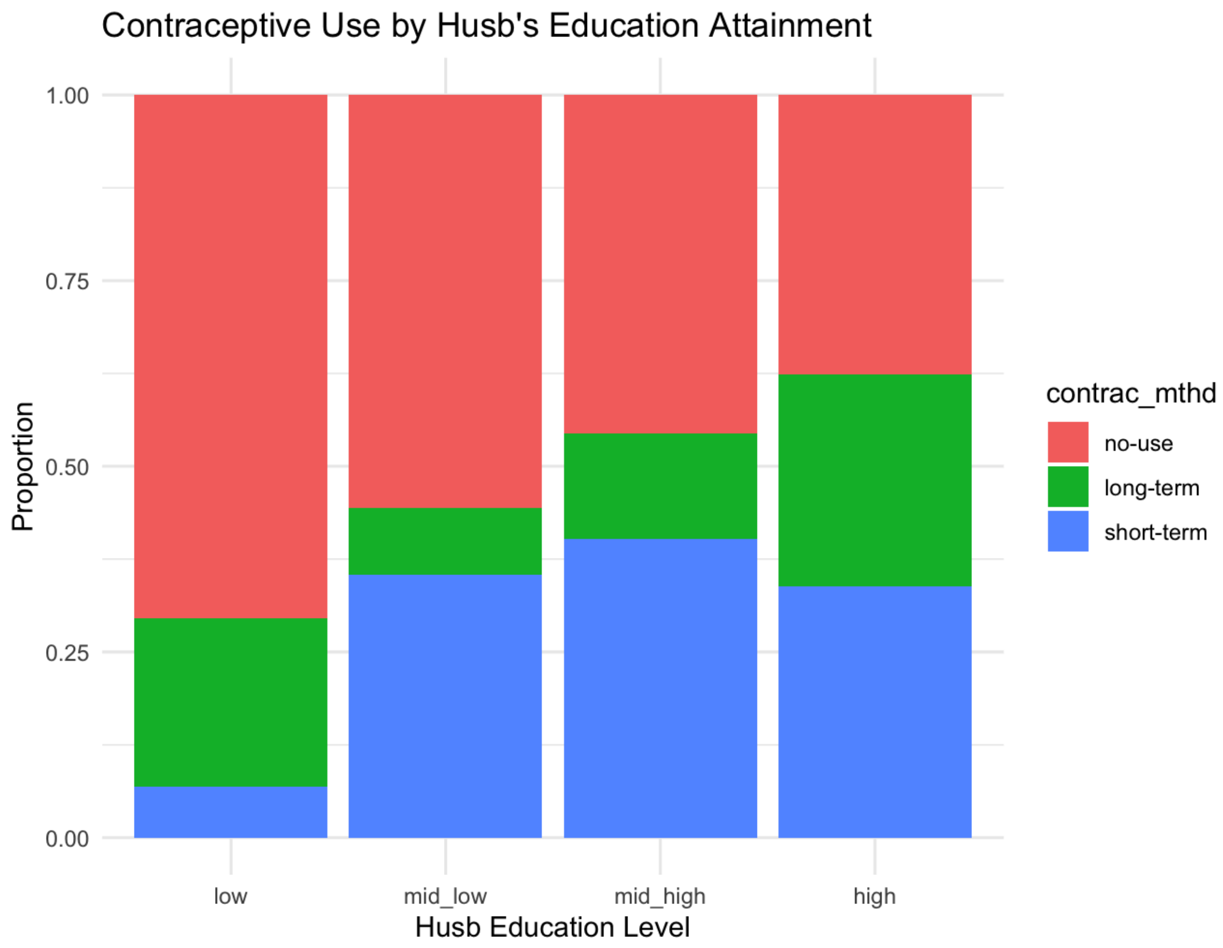


Fig.5: Contraceptive use by husband's education level

Looking at these plots in Fig.4 and Fig.5 above, a few different relationships appear:

A woman's tendency to utilize long-term contraceptive techniques rises with her degree of education, the likelihood that a couple will not use any form of birth control at all increases with both husband and wife's educational attainment, the percentage of couples that use short-term contraceptives appears to be least impacted by educational attainment.

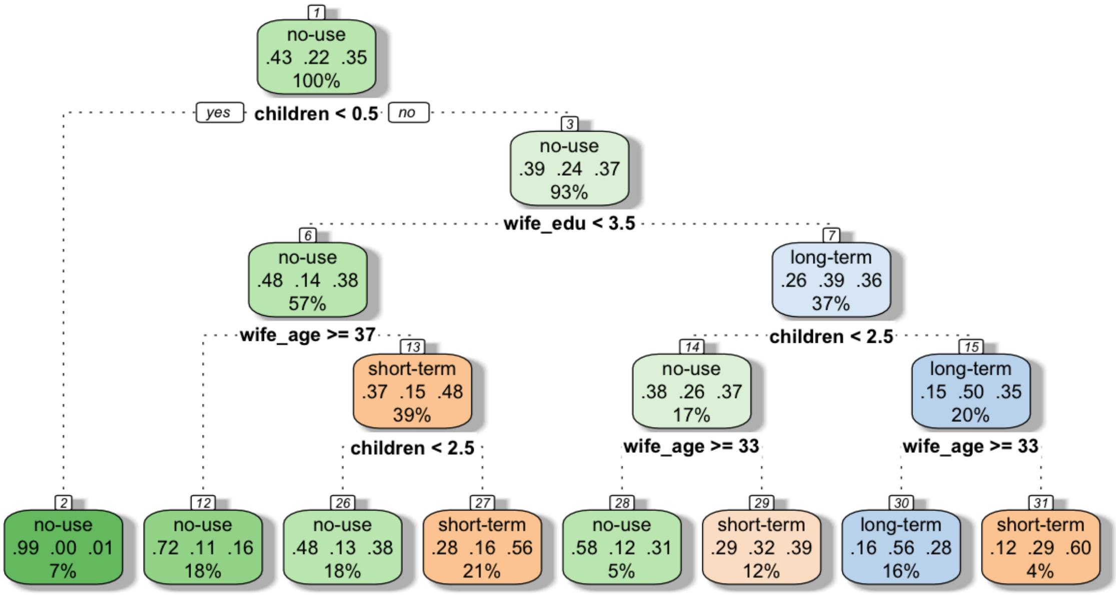
Naturally, these are only correlations, and at this time, we cannot be certain that education degree influences the choice of contraceptive technique.

## Results

The best parameter for the kNN model was with 27 nearest neighbors. I tried out different values of k and this proved to be the best with an accuracy of 51.36% and kappa value of 26.71%. The sensitivity(True Positive) values by class was 62.30% for no-use, 35.85% for long-term and

52.94% for short-term. The specificity(True Negative) values by class was 70.27% for no-use, 85.71% for short-term, and 70.39% for long-term.

The decision tree performed better than the kNN with an accuracy of 55.88%, this indicates the overall effectiveness of the model. The model resulted in a 31.08% kappa value. The sensitivity(True Positive) values by class was 67.21% for no-use, 33.02% for long-term and 58.17% for short-term while the specificity(True Negative) values were 71.81% for no-use, 90.18% for short-term, and 69.20% for long-term. I proceeded to see what was going on under the hood with the decision tree by plotting a fancyRpartPlot in fig.6 below.



Rattle 2024-Apr-23 11:59:20 user

Fig.6: fancyRpartPlot of decision tree

As it happens, the crucial break in our decision tree is really the number of children born. The initial group of women is effectively divided into two groups by the model: those who have previously given birth and those who have not. It is quite likely that women who are childless are not utilizing any form of birth control. Following that first division, the wife's educational background becomes a significant factor in dividing options. The wife's age background and then number of children again are addressed next.

I tried to improve the model performance by using cross-validation but this didn’t help as the accuracy was lower compared to the typical decision tree model.

The Naive Bayes algorithm didn’t perform as well as the decision tree nor the kNN model, it had an accuracy of 47.51%, kappa value of 21.88%, sensitivity values of 42.62% for no-use, 51.89% for long-term and 50.33% for short-term. The specificity(True Negative) values by class was 81.85% for no-use, 70.24% for short-term, and 70.59% for long-term.

## Discussion

Based on the results, it's evident that the decision tree model outperformed both the KNN and Naive Bayes algorithms in terms of accuracy and overall effectiveness. With an accuracy of 55.88% and a kappa value of 31.08%, the decision tree model demonstrated robust performance in classifying contraceptive choices among the study cohort. The key insights gained from analyzing the decision tree model reveal that the number of children born emerged as a crucial factor in determining contraceptive choices. This finding underscores the importance of reproductive history in influencing contraceptive decisions among women. Additionally, the age of the wife, education level, and the number of children further refined the decision-making process within the model.

Attempts to enhance model performance through cross-validation did not yield significant improvements, indicating that the standard decision tree model already effectively captures the underlying patterns in the data. This suggests that the decision tree model's performance is inherently strong and may not benefit significantly from additional optimization techniques.

The kNN model performed relatively well with an accuracy of 52.71%, but it was outperformed by the decision tree model. Similarly, the Naive Bayes algorithm exhibited lower accuracy (47.51%) and kappa value (21.88%) compared to both the decision tree and kNN models. These results suggest that the decision tree algorithm is better suited for classifying contraceptive choices in this context.

Overall, the findings highlight the complexity of factors influencing contraceptive decisions and underscore the importance of utilizing robust modeling techniques, such as decision trees, to effectively analyze and predict contraceptive behavior among women. Further research could explore additional variables and ensemble methods to enhance predictive accuracy and deepen our understanding of contraceptive decision-making processes.

## Limitations

The accuracy of the models to represent the whole country of Indonesia may be relatively poor due to the significantly small size of the dataset. Furthermore, the tiny size of the dataset raises the possibility of bias. The fact that supervised machine learning needs a lot of data to produce a model that is relatively accurate is one of its main drawbacks. Thus, a more accurate model might have been created for this research if a bigger dataset had been available.

## Conclusion

The machine learning models in this study proved successful in helping us comprehend the underlying behaviors of Indonesians choosing contraceptive techniques than it is to guess the selection that a certain Indonesian woman would make. The Family Planning Program's ultimate objective is to enable women to make their own decisions regarding family planning and contraception by educating them about the use of contraceptives. There are other ways to achieve that objective, though, and the Indonesian government must choose the most practical one given its limited time and resources. We can determine the primary determinants of these women's behavior patterns by utilizing a decision tree model. The family planning program's policies and financial allocation should be shaped by this understanding. For instance, it is evident from the decision tree visualization's right side that young, less educated women are far more likely to use either no contraception at all or very little of it. The Indonesian government should focus on this group of women by educating them about long-term contraceptive options through educational initiatives in secondary schools and colleges. Indonesia has a considerably greater chance of regaining its reputation as a leader in population control and family planning if choices about the family planning program are data-driven.