# Technical Report

Group 3: Isabelle Caldwell, Ananke Krishnan, Gillian McGinnis

11/15/2021

## Abstract

In this analysis we are interested in how lifestyle factors affect the likelihood of stroke. Strokes are the second leading cause of death worldwide and are caused by a combination of physiological and lifestyle factors. A boosted tree model and a penalized classification tree model were used to infer which variables, from a set of eight lifestyle and other categorical variables, had the largest impact on occurrence of a stroke. Hypertension, heart disease, smoking status, and marital status were found to have the greatest effect on stroke likelihood.

## Introduction

Strokes (also known as cerebrovascular accidents, or CVAs) are the second leading cause of death worldwide, according to the World Health Organization. The CDC reports that every 40 seconds someone in the United States has a stroke and every 4 minutes someone dies of a stroke. Johns Hopkins' reports that there are many risk factors for stroke, which generally include conditions beyond one's control (such as age, pre-existing conditions, and family history), as well as lifestyle choices (such as living environment and physical activity).

In this project, we will examine and determine the statistical relationship between lifestyle choices on likelihood of stroke. Health research is often done on the measurable physiological variables that can be collected in your average yearly checkup. However, we are more interested in how lifestyle and everyday choices affect the likelihood of stroke. The data of interest includes categorical variables of both physiological and lifestyle information. Data accessed includes both physiological information (such as age and BMI [Body Mass Index]) in addition to personal information (such as work status and sector, marital status, and history of smoking).

# Methods

## Data

The data set used was the “Stroke Prediction Dataset” downloaded from Kaggle. This data set has 5110 observations and 11 predictors for the presence or absence of a stroke. Out of the 5110 observations, 201 of them have a value of `NA` for the BMI column (`bmi`).

See ‘Exploratory Data Analysis’ for more information about specific variables and distributions. We subset the data to only include the categorical or lifestyle variables, and excluded the numeric variables (BMI, Age, Average Glucose Level).

```
## Rows: 5,110
## Columns: 12
## $ id                <dbl> 9046, 51676, 31112, 60182, 1665, 56669, 53882, 10434~
## $ gender            <fct> Male, Female, Male, Female, Female, Male, Male, Fema~
## $ age               <dbl> 67, 61, 80, 49, 79, 81, 74, 69, 59, 78, 81, 61, 54, ~
## $ hypertension      <fct> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1~
## $ heart_disease     <fct> 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0~
## $ ever_married      <fct> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ work_type         <fct> Private, Self-employed, Private, Private, Self-emplo~
## $ residence_type    <fct> Urban, Rural, Rural, Urban, Rural, Urban, Rural, Urb~
## $ avg_glucose_level <dbl> 228.69, 202.21, 105.92, 171.23, 174.12, 186.21, 70.0~
## $ bmi               <dbl> 36.6, NA, 32.5, 34.4, 24.0, 29.0, 27.4, 22.8, NA, 24~
## $ smoking_status    <fct> formerly smoked, never smoked, never smoked, smokes,~
## $ stroke            <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

### Dataset Attribute Descriptions

- `id`: (numeric) a unique patient identifier
- `gender`: (factor w/ 3 levels) gender identity of the patient (“Male”, “Female”, or “Other”)
- `age`: (numeric) the age of the patient in years (Range: 0.08-82)
- `hypertension`: (factor w/ 2 levels) whether the patient has hypertension (0 for No, 1 for Yes)
- `heart_disease`: (factor w/ 2 levels) whether the patient has heart disease (0 for No, 1 for Yes)
- `ever_married`: (factor w/ 2 levels) whether the patient has been married or not (0 for No, 1 for Yes)
- `work_type`: (factor w/ 5 levels) type of work (“children”, “Govt_jov”, “Never_worked”, “Private” or

"Self-employed", where children is for age <18)

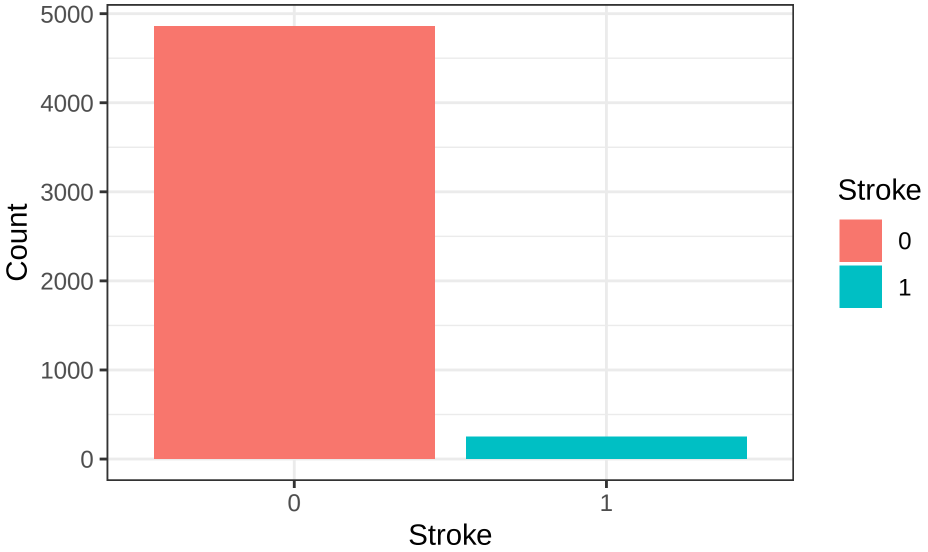
- `residence_type`: (factor w/ 2 levels) place of residence ("Rural", "Urban")
- `avg_glucose_level`: (numeric) average glucose blood level in mmol/L (Range: 55.12-271.74)
- `bmi`: (numeric) ration measuring height to weight (Range: 10.3-97.6 with 201 `NA` values)
- `smoking_status`: (factor w/ 4 levels) status of patient in terms of smoking ("formerly smoked", "never smoked", "smokes" or "Unknown"). Unknown indicates that smoking data was not available for patient.
- `stroke`: (factor w/ 2 levels) whether the patient had a stroke (0 for No, 1 for Yes)

### Pre-processing

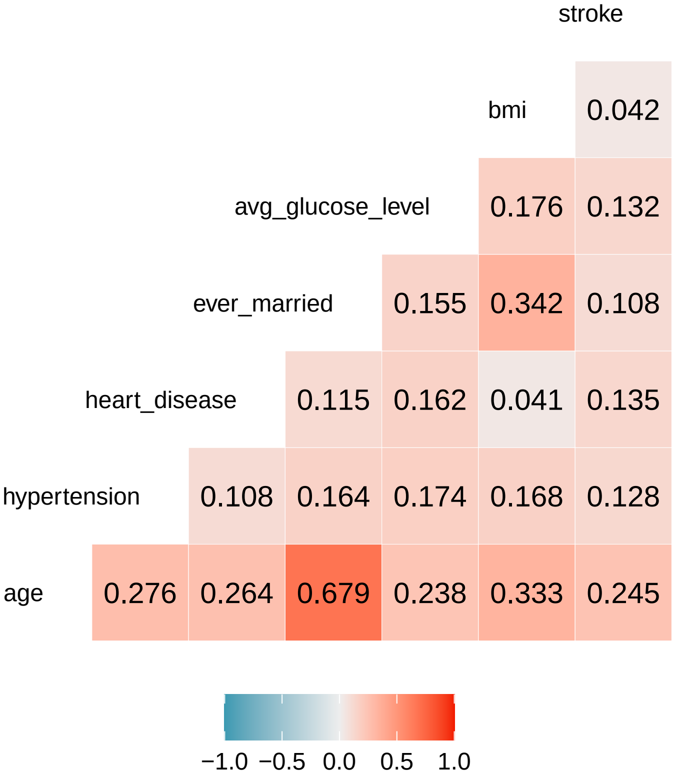
We subset the data to only include the categorical or lifestyle variables and excluded the numeric variables (BMI, Age, Average Glucose Level). All categorical variables were modified to be levelled factors. Additionally, observations with NA values were dropped.

## Exploratory Data Analysis

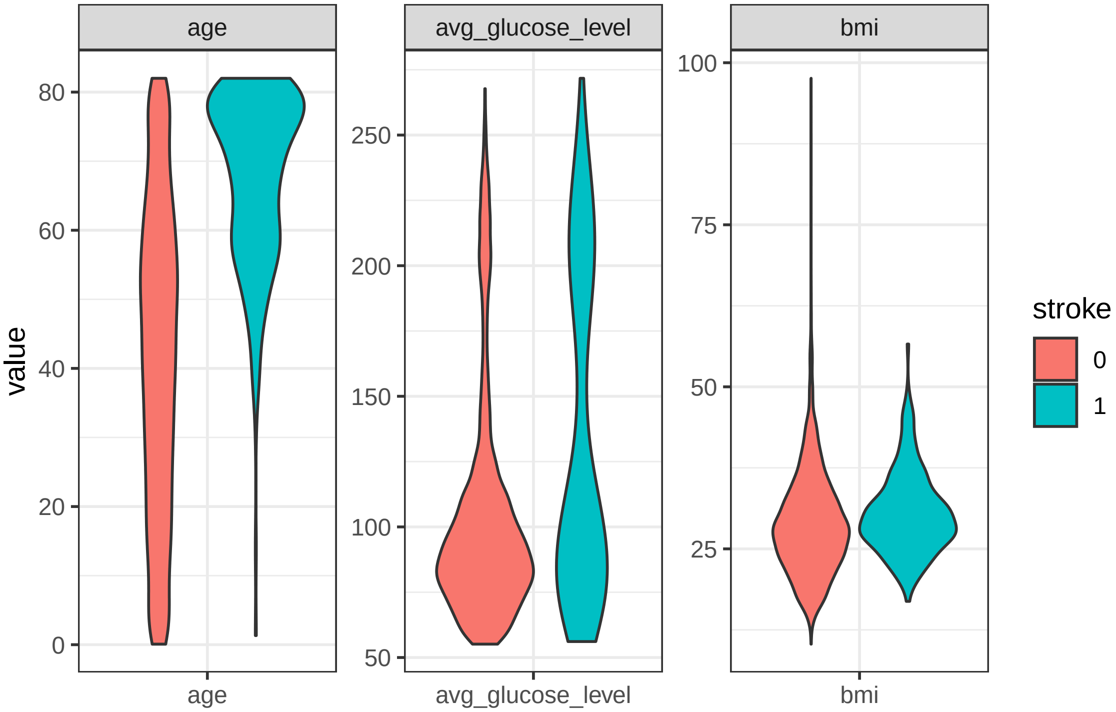
Distribution of `stroke` variable in the data set:



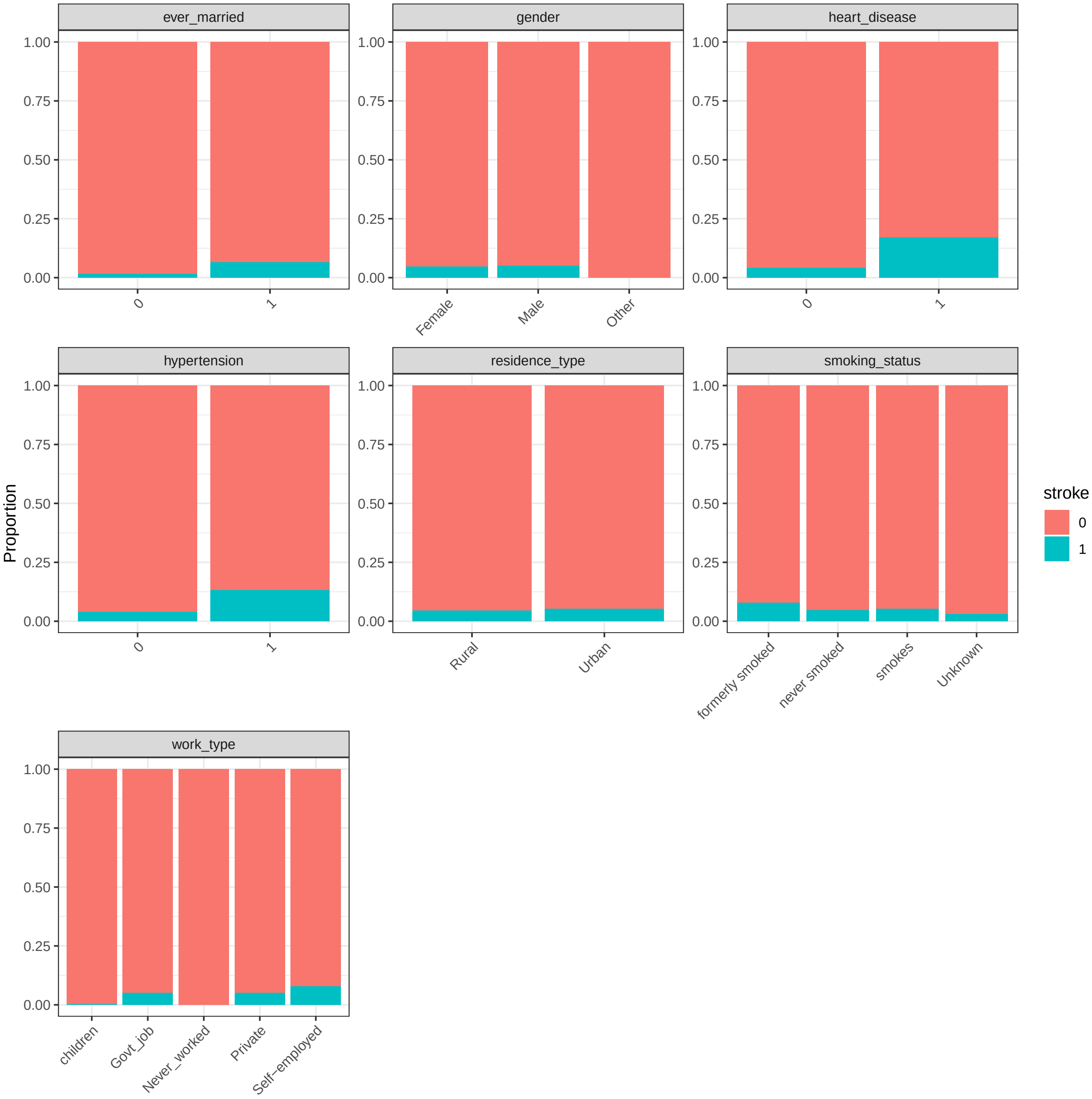
Correlation diagram of binary and numeric variables:



Violin plot demonstrating distribution of continuous variables:



Percentile distributions of variables by `stroke`:



Graphical analysis was done on all categorical variables to see the distribution of strokes within each variable. While it may seem like the difference in stroke classifications is small, a 5% increase in chances of having a stroke is large on the population scale. From the graphs above we predict that the variables `smoking_status`, `work_type`, `heart_disease`, `ever_married`, and `hypertension` to have the greatest effect on `stroke`.

This data set seems to lend itself well to Linear or Quadratic Discriminant Analysis due to the predominantly categorical, specifically binary categorical variables. However, since we want to focus primarily on inference, we will favor techniques that allow us to assess models rather than test them out, namely decision trees.

# Results

Using decision trees, we found that heart disease and hypertension were the most important factors associated with likelihood of having a stroke. We used two different types of decision tree models to make this inference.

## Models

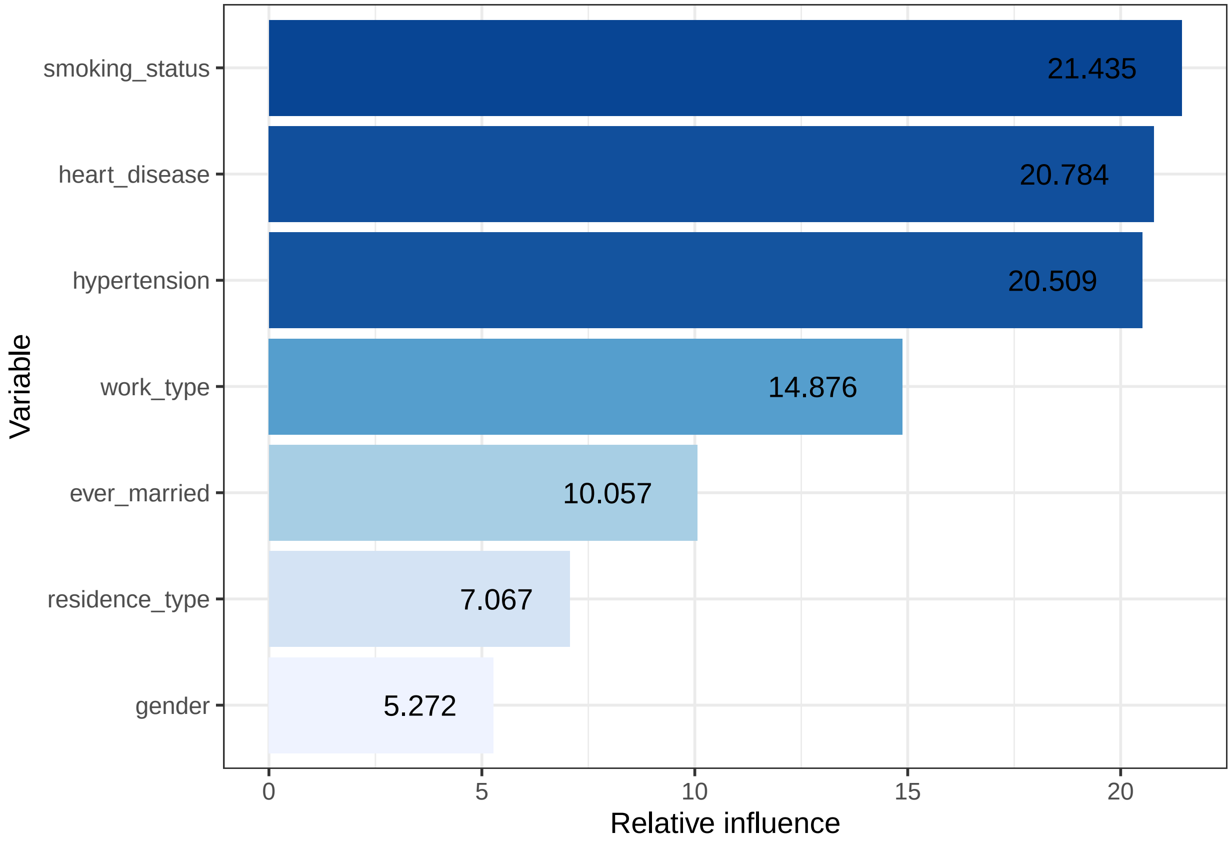
Decision trees are a powerful statistical model that is capable of classifying outcomes in a highly interpretable way. They use a splitting strategy based on given predictors handset parameters that classify the likelihood that a certain observation falls into a certain outcome. They are highly interpretable because of the way that you can follow the predictor thresholds that result in splits and classification. However, they are a highly flexible model that is susceptible to overfitting to the training dataset, which makes predicting difficult.

We chose to use a decision tree for this project because we were extremely interested in creating an interpretable model and less so in predicting whether an individual observation will have a stroke.

### Boosted Tree

Boosting is a learning algorithm that builds a stronger model on many weaker models, here the weaker model is the decision tree. In boosting the decision trees are created sequentially, each one building on the tree before it. For this reason, boosted tree models often have a lower mean squared error rate also known as the misclassification error rate. The misclassification error rate is the number of times a model incorrectly classifies an observation. This could be predicting an individual will have a stroke when they won't, or predicting they will not have a stroke and when they will. A boosted tree will highlight the most important variables in our dataset which will allow us to see which variables have the largest impact on stroke likelihood. It is easily susceptible to overfitting and a highly variable model, because it is built on sequential decision trees.

Our boosted tree model ranked the provided predictors according to relative influence on the model as follows:



```
##              var   rel.inf
## 1 smoking_status 21.435097
## 2  heart_disease 20.783528
## 3   hypertension 20.509150
## 4      work_type 14.876206
## 5   ever_married 10.057281
## 6 residence_type  7.066596
## 7         gender  5.272142
```

On the test set, if assuming the threshold of positive classification as 10% (i.e., probability above 0.1 will be categorized as a stroke), the boosted tree model predicted 47 false negatives and 110 false positives, with a misclassification rate of 0.1228482:

```
##           Truth
## Prediction    0    1
##          0 1102   47
##            1  110   19
```

This can be improved by increasing the number of trees generated in the boosted model at a cost to computational time.

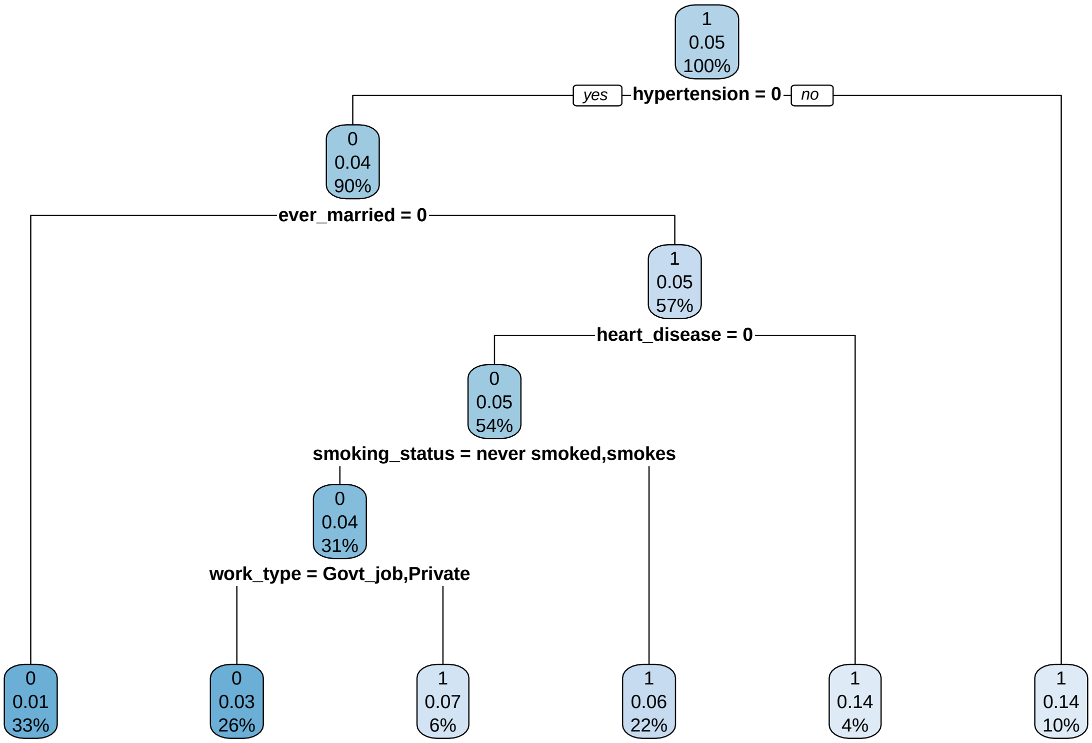
### Penalized Classification Tree

Assigning a penalty to misclassified observations is one way to get around the problem of overfitting and improve the predictive accuracy of trees. We do this by incorporating the relative loss of misclassification via the Gini Index:

$$G = \sum_i \sum_j L(i, j) p_i p_j$$

where $L(i, j)$ is the relative loss of predicting $j$ if the truth is $i$. This is especially useful when a misclassification can result in actual harm, such as predicting a stroke will not happen when it will. In R, this is accomplished by constructing a penalty matrix that is taken into account when building the final tree. The penalized classification tree was introduced into our analysis in an attempt to reduce the high type II error found in the boosted tree model. Type II error is when an individual is predicted to not have a stroke, when in fact they do. This is much worse than predicting an individual has a stroke when they don't.

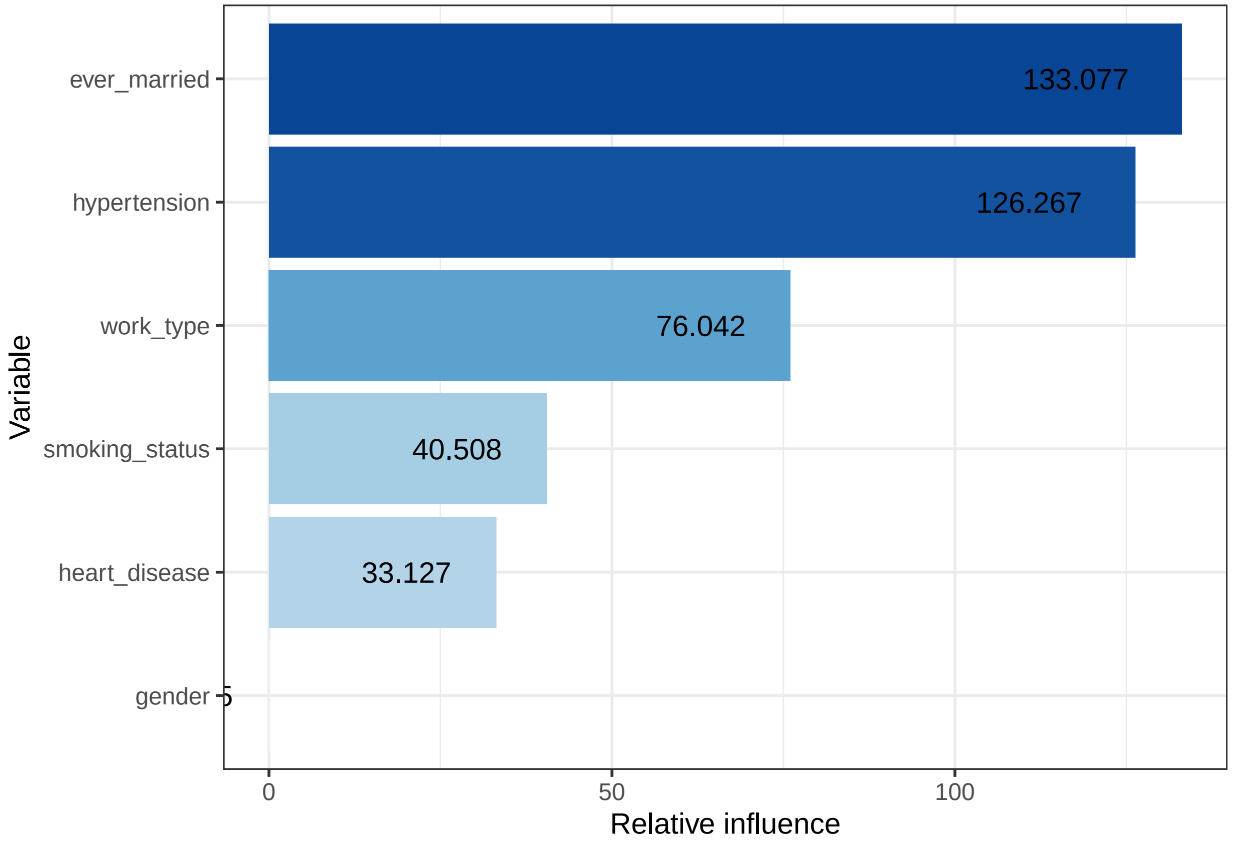
Our penalty matrix had a high penalty of 20 for false negatives and 1 for false positives. These penalties were weighted accordingly due to the high cost of predicting false negatives in this context. The final splits for this decision tree have decent class purity, with the highest value of 0.14.



At a misclassification rate of 0.1486698, this tree performs poorly in terms of false positives, but has decreased in false negatives, which could be improved by imposing a higher penalty on false positive classifications. Its confusion matrix reports as follows:

```
##           Truth
## Prediction    0    1
##          0 1063   41
##          1  149   25
```

Since this is a singular tree (as opposed to a boosted model), the variable importance also changes:



```
##               var   rel.inf
## 1   ever_married 133.07664
## 2   hypertension 126.26708
## 3      work_type  76.04173
## 4 smoking_status  40.50790
## 5  heart_disease  33.12749
## 6         gender   0.10495
```

The boosted model predicted the most important variable to be smoking_status (with a relative influence of 21.435097) and the least as gender (5.2721421), while this penalized model reported its most influential variable to be ever_married (influence: 133.0766416) and the least as gender (0.10495).

## Discussion

In this project, we aimed to use decision trees to make inferences about non-medical variables that could influence an individual's likelihood of suffering from a stroke. We decided to use decision trees to do so. Using boosted and penalized trees, we were able to identify hypertension, heart disease, smoking, and marital status as important predictors in the likelihood of a stroke.

While we first struggled with highly inaccurate models with high type II errors, resulting in models that glossed over individuals who were actually at risk of a stroke, this was resolved by strengthening the decision tree algorithm with boosted trees and penalized regression. We still had difficulty in the penalized classification tree in terms of having a high rate of false positives, but were able to resolve the concern surrounding false negatives, which is arguably more dangerous to a patient's health.

We do have some concerns with our conclusions, namely with the reliability of the data source. The .csv file was obtained from Kaggle, and the author who uploaded the file did not reveal the source of the data due to confidentiality concerns. Additionally, just under 5% (4.873%) of the observations were for individuals with a stroke, meaning that there might not be enough data to infer stroke presence as accurately as we would like.

Further research might extend this model to be able to predict stroke presence or absence, which would provide a useful tool to predict strokes using factors in addition to a patient's medical history. In order to use this model to predict stroke presence, the tree parameters would have to be fine-tuned to reduce type I and II error.

# Code Appendix

## Setting hyperparameters

```r
n_trees <- 100
shrinkage_penalty <- 0.1
interaction_depth <- 3
split_ratio <- 0.75

percent_classification <- 0.1
```

## Splitting data

```r
set.seed(666)
strokes_mini <- strokes %>% select_if(is.factor)

train_index <- sample(1:nrow(strokes_mini), nrow(strokes_mini) * .75)
test_index <- (1:nrow(strokes_mini))[-(train_index)]
strokes_trn <- strokes_mini %>% slice(train_index)
strokes_tst <- strokes_mini %>% slice(test_index)

# Cleaning environment
remove(train_index, test_index)
```

## Creating a boosted tree

```r
set.seed(666)
boost_strokes <- gbm(
  stroke ~.,
  data = strokes_trn,
  distribution = "multinomial",
  n.trees = n_trees,
  shrinkage = shrinkage_penalty,
  interaction.depth = interaction_depth
)
```

```
boost_var_imp <- summary(boost_strokes, plotit=F) %>%
  data.frame() %>%
  rownames_to_column() %>%
  select(!rowname)


boost_var_imp_plot <- boost_var_imp %>%
  ggplot(aes(rel.inf, reorder(var, rel.inf))) +
  geom_col(aes(fill = rel.inf)) +
  scale_fill_distiller(direction=1) +
  geom_text(aes(label = as.character(round(rel.inf, 3))), hjust = 1.5) +
  theme_bw() +
  labs(
    x = "Relative influence",
    y = "Variable"
  ) +
  theme(legend.position = "none")
```

```
boosted_preds <- predict(
  boost_strokes,
  strokes_tst,
  type = "response",
  n.trees = n_trees
) %>%
  data.frame() %>%
  pull(2)


boosted_preds <- as.factor(ifelse(boosted_preds >= percent_classification, 1, 0))
```

```
results_boosted <- data.frame(
  obs = strokes_tst$stroke,
  preds = boosted_preds
)
```

```r
# Conf
boost_conf_mat <- conf_mat(results_boosted, truth = obs, estimate = preds)

# False negatives
boost_false_neg <- boost_conf_mat$table %>%
  data.frame() %>%
  filter(Prediction == 0 & Truth == 1) %>%
  pull(Freq)

# False positives
boost_false_pos <- boost_conf_mat$table %>%
  data.frame() %>%
  filter(Prediction == 1 & Truth == 0) %>%
  pull(Freq)

# Misclasss
boost_misclass <- 1 - (accuracy(results_boosted, truth = obs, estimate = preds) %>% pull(".estimate"))
```

### Penalized tree

```r
set.seed(666)
penalty_matrix <- matrix(c(0, 1, 20, 0), byrow = T, nrow = 2)

penalized_tree <- rpart(stroke ~ ., data = strokes_trn, parms = list(loss = penalty_matrix))
```

```r
penalized_var_imp <- penalized_tree$variable.importance %>%
  data.frame() %>%
  rownames_to_column() %>%
  rename(var = rowname, rel.inf = ".")

penalized_var_imp_plot <- penalized_var_imp %>%
  ggplot(aes(rel.inf, reorder(var, rel.inf))) +
  geom_col(aes(fill = rel.inf)) +
  scale_fill_distiller(direction=1) +
  geom_text(aes(label = as.character(round(rel.inf, 3))), hjust = 1.5) +
  theme_bw() +
  labs(
    x = "Relative influence",
    y = "Variable"
  ) +
  theme(legend.position = "none")
```

```
penalized_preds <- predict(
  penalized_tree,
  strokes_tst)

penalized_preds <- as.factor(ifelse(penalized_preds[,2] >= percent_classification, 1, 0))
```

```
results_penalized <- data.frame(
  obs = strokes_tst$stroke,
  preds = penalized_preds
)

# Conf
penalized_conf_mat <- conf_mat(results_penalized, truth = obs, estimate = preds)

# Misclasss
penalized_misclass <- 1 - (accuracy(results_penalized, truth = obs, estimate = preds) %>% pull(".estima
```

## References

"Risk Factors for Stroke." n.d. Accessed December 5, 2021. (https://www.hopkinsmedicine.org/health/conditions-and-diseases/stroke/risk-factors-for-stroke).

"Stroke | NHLBI, NIH." n.d. Accessed December 5, 2021. (https://www.nhlbi.nih.gov/health-topics/stroke).

Palacios, Soriano Federico. 2021. "Stroke Prediction Dataset." Accessed December 12, 2021. https:

//kaggle.com/fedesoriano/stroke-prediction-dataset.