Platform 2: R

In R, the rpart package is used to create the Decision Tree model, and rpart.plot is used for visualizing the tree structure. The Gini criterion, which measures the impurity of splits, is used in the Decision Tree model by default in R.

```
knitr::opts_chunk$set(echo = TRUE)

library(rpart)
library(rpart.plot)
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(ggplot2)
```

Data Loading

The cleaned dataset from the Jupyter Notebook is used to maintain consistency across both platforms.

```
data <- read.csv("cleaned_data.csv")
```

Data Preprocessing

The dataset is split into training and testing sets to allow for model evaluation. 70% of the data is used for training, while 30% is reserved for testing. The set.seed(42) ensures reproducibility.

```
# Splitting the dataset
set.seed(42)
trainingIndex <- createDataPartition(data$HadHeartAttack, p = 0.7, list = FALSE)
train <- data[trainingIndex, ]
test <- data[-trainingIndex, ]
```

Model Building

A Decision Tree is built using the rpart function with a maximum depth of 5, consistent with the Python model, and the Gini criterion (default in R). The rpart.plot function is used to visualize the tree, showing significant risk factors for heart disease.

```
dt_model <- rpart(HadHeartAttack ~ .,
                  data = train,
                  method = "class",
                  control = rpart.control(maxdepth = 5, cp = 0.001, minsplit = 20))

# Print summary of the model to check variable importance and splits
summary(dt_model)
```

```
## Call:
## rpart(formula = HadHeartAttack ~ ., data = train, method = "class",
##     control = rpart.control(maxdepth = 5, cp = 0.001, minsplit = 20))
##   n= 311564
##
##            CP nsplit rel error    xerror        xstd
## 1 0.016200045      0 1.0000000 1.0000000 0.007296337
## 2 0.010894107      2 0.9675999 0.9722285 0.007200329
## 3 0.001608715      4 0.9458117 0.9498194 0.007121662
## 4 0.001000000      6 0.9425943 0.9466019 0.007110277
##
## Variable importance
##         HadAngina               Sex         HadStroke    HeightInMeters
##                92                 2                 2                 1
##      RemovedTeeth WeightInKilograms
##                 1                 1
##
## Node number 1: 311564 observations,    complexity param=0.01620005
##   predicted class=0  expected loss=0.05686151  P(node) =1
##     class counts: 293848 17716
##    probabilities: 0.943 0.057
##   left son=2 (292896 obs) right son=3 (18668 obs)
##   Primary splits:
##       HadAngina         < 0.5      to the left,  improve=6174.2150, (0 missing)
##       HadStroke         < 0.5      to the left,  improve=1145.3790, (0 missing)
##       GeneralHealth     < 2.5      to the right, improve=1035.7050, (0 missing)
##       DifficultyWalking < 0.5      to the left,  improve= 848.2347, (0 missing)
##       AgeCategory       < 8.5      to the left,  improve= 808.7820, (0 missing)
##
## Node number 2: 292896 observations
##   predicted class=0  expected loss=0.0317314  P(node) =0.9400829
##     class counts: 283602  9294
##    probabilities: 0.968 0.032
##
## Node number 3: 18668 observations,    complexity param=0.01620005
##   predicted class=0  expected loss=0.4511463  P(node) =0.05991706
##     class counts: 10246  8422
##    probabilities: 0.549 0.451
##   left son=6 (15610 obs) right son=7 (3058 obs)
##   Primary splits:
##       HadStroke     < 0.5      to the left,  improve=148.9518, (0 missing)
##       Sex           < 0.5      to the left,  improve=141.0019, (0 missing)
##       RemovedTeeth  < 1.5      to the right, improve=112.8562, (0 missing)
##       GeneralHealth < 2.5      to the right, improve=112.5565, (0 missing)
##       SmokerStatus  < 0.25     to the left,  improve=105.8464, (0 missing)
##
## Node number 6: 15610 observations,    complexity param=0.01089411
##   predicted class=0  expected loss=0.4231903  P(node) =0.05010207
##     class counts:  9004  6606
##    probabilities: 0.577 0.423
##   left son=12 (6312 obs) right son=13 (9298 obs)
##   Primary splits:
##       Sex           < 0.5      to the left,  improve=123.67710, (0 missing)
##       SmokerStatus  < 0.25     to the left,  improve= 95.99606, (0 missing)
##       RemovedTeeth  < 1.5      to the right, improve= 82.19056, (0 missing)
##       GeneralHealth < 2.5      to the right, improve= 66.37540, (0 missing)
##       HadDiabetes   < 0.5      to the left,  improve= 47.60241, (0 missing)
##   Surrogate splits:
##       HeightInMeters        < 1.682609 to the left,  agree=0.840, adj=0.603, (0 split)
##       WeightInKilograms     < 69.925   to the left,  agree=0.681, adj=0.212, (0 split)
##       DifficultyErrands     < 0.5      to the right, agree=0.626, adj=0.075, (0 split)
##      HadDepressiveDisorder < 0.5      to the right, agree=0.621, adj=0.063, (0 split)
##      HadAsthma             < 0.5      to the right, agree=0.616, adj=0.050, (0 split)
##
## Node number 7: 3058 observations,    complexity param=0.001608715
##   predicted class=1  expected loss=0.4061478  P(node) =0.009814998
##     class counts:  1242  1816
##    probabilities: 0.406 0.594
##   left son=14 (1406 obs) right son=15 (1652 obs)
##   Primary splits:
##       Sex            < 0.5      to the left,  improve=30.12201, (0 missing)
##       GeneralHealth  < 3.5      to the right, improve=17.38165, (0 missing)
##       HadDiabetes    < 0.5      to the left,  improve=14.59119, (0 missing)
##       HeightInMeters < 1.639789 to the left,  improve=13.52388, (0 missing)
##       RemovedTeeth   < 1.5      to the right, improve=10.98247, (0 missing)
##   Surrogate splits:
##       HeightInMeters    < 1.679839 to the left,  agree=0.828, adj=0.625, (0 split)
##       WeightInKilograms < 70.535   to the left,  agree=0.655, adj=0.249, (0 split)
##       DifficultyErrands < 0.5      to the right, agree=0.594, adj=0.117, (0 split)
##       HadAsthma         < 0.5      to the right, agree=0.587, adj=0.102, (0 split)
##       MentalHealthDays  < 1.5      to the right, agree=0.585, adj=0.097, (0 split)
##
## Node number 12: 6312 observations
##   predicted class=0  expected loss=0.3467997  P(node) =0.02025908
##     class counts:  4123  2189
##    probabilities: 0.653 0.347
##
## Node number 13: 9298 observations,    complexity param=0.01089411
##   predicted class=0  expected loss=0.4750484  P(node) =0.02984299
##     class counts:  4881  4417
##    probabilities: 0.525 0.475
##   left son=26 (6358 obs) right son=27 (2940 obs)
##   Primary splits:
##       RemovedTeeth    < 1.5      to the right, improve=70.58009, (0 missing)
##       GeneralHealth   < 2.5      to the right, improve=56.16869, (0 missing)
##       SmokerStatus    < 0.25     to the left,  improve=51.66284, (0 missing)
##       AgeCategory     < 5.5      to the left,  improve=29.08122, (0 missing)
##       AlcoholDrinkers < 0.75     to the right, improve=27.27180, (0 missing)
##   Surrogate splits:
##       HadCOPD       < 0.5      to the left,  agree=0.696, adj=0.038, (0 split)
##       SmokerStatus  < 2.5      to the left,  agree=0.692, adj=0.027, (0 split)
##       GeneralHealth < 1.5      to the right, agree=0.688, adj=0.014, (0 split)
##       BMI           < 17.255   to the right, agree=0.685, adj=0.004, (0 split)
##       SleepHours    < 10.5     to the left,  agree=0.685, adj=0.002, (0 split)
##
## Node number 14: 1406 observations,    complexity param=0.001608715
##   predicted class=1  expected loss=0.4822191  P(node) =0.004512716
##     class counts:   678   728
##    probabilities: 0.482 0.518
##   left son=28 (839 obs) right son=29 (567 obs)
##   Primary splits:
##       HadDiabetes        < 0.5      to the left,  improve=11.143310, (0 missing)
##       GeneralHealth      < 3.5      to the right, improve= 8.274080, (0 missing)
##       MentalHealthDays   < 15.5     to the left,  improve= 8.262859, (0 missing)
##       PhysicalHealthDays < 22       to the left,  improve= 7.916400, (0 missing)
##       DifficultyErrands  < 0.5      to the left,  improve= 7.768530, (0 missing)
##   Surrogate splits:
##       HadKidneyDisease  < 0.5      to the left,  agree=0.633, adj=0.090, (0 split)
##       BMI               < 30.415   to the left,  agree=0.624, adj=0.069, (0 split)
##       WeightInKilograms < 85.8669  to the left,  agree=0.614, adj=0.042, (0 split)
##       HighRiskLastYear  < 0.75     to the left,  agree=0.602, adj=0.012, (0 split)
##       SleepHours        < 1.5      to the right, agree=0.599, adj=0.005, (0 split)
##
## Node number 15: 1652 observations
##   predicted class=1  expected loss=0.3414044  P(node) =0.005302281
##     class counts:   564  1088
##    probabilities: 0.341 0.659
##
## Node number 26: 6358 observations
##   predicted class=0  expected loss=0.4331551  P(node) =0.02040672
##     class counts:  3604  2754
##    probabilities: 0.567 0.433
##
## Node number 27: 2940 observations
##   predicted class=1  expected loss=0.4343537  P(node) =0.009436263
##     class counts:  1277  1663
##    probabilities: 0.434 0.566
##
## Node number 28: 839 observations
##   predicted class=0  expected loss=0.466031  P(node) =0.002692866
##     class counts:   448   391
##    probabilities: 0.534 0.466
##
## Node number 29: 567 observations
##   predicted class=1  expected loss=0.4056437  P(node) =0.001819851
##     class counts:   230   337
##    probabilities: 0.406 0.594
```

Model Evaluation

The model is used to predict the outcomes in the test set, and a confusion matrix is generated to evaluate the accuracy of the model, comparing predicted and actual outcomes□

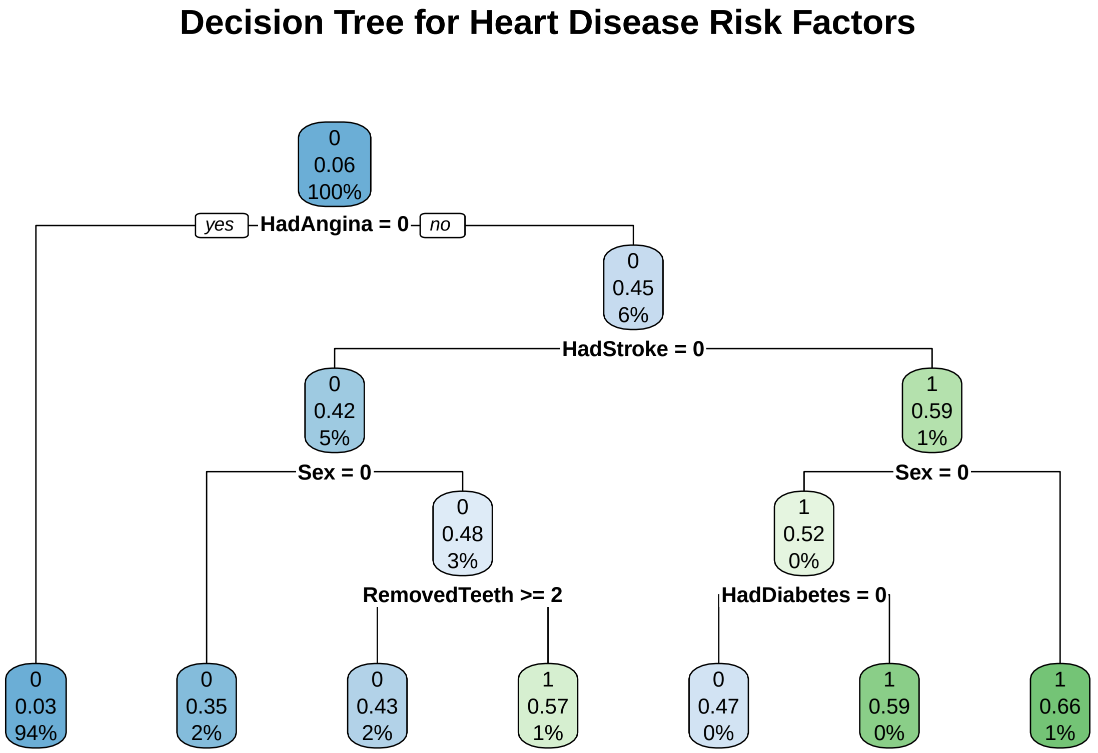
```r
# Make predictions
predictions <- predict(dt_model, newdata = test, type = "class")

# Convert both to factors and ensure they have the same levels
predictions <- as.factor(predictions)
test$HadHeartAttack <- as.factor(test$HadHeartAttack)

# Ensure they have the same levels
levels(predictions) <- levels(test$HadHeartAttack)

# Evaluate the model
confusionMatrix(predictions, test$HadHeartAttack)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##           0 125170    6134
##           1    966    1257
##
##                Accuracy : 0.9468
##                  95% CI : (0.9456, 0.948)
##     No Information Rate : 0.9446
##     P-Value [Acc > NIR] : 0.0002356
##
##                   Kappa : 0.2421
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9923
##             Specificity : 0.1701
##          Pos Pred Value : 0.9533
##          Neg Pred Value : 0.5655
##              Prevalence : 0.9446
##          Detection Rate : 0.9374
##    Detection Prevalence : 0.9834
##       Balanced Accuracy : 0.5812
##
##        'Positive' Class : 0
##
```

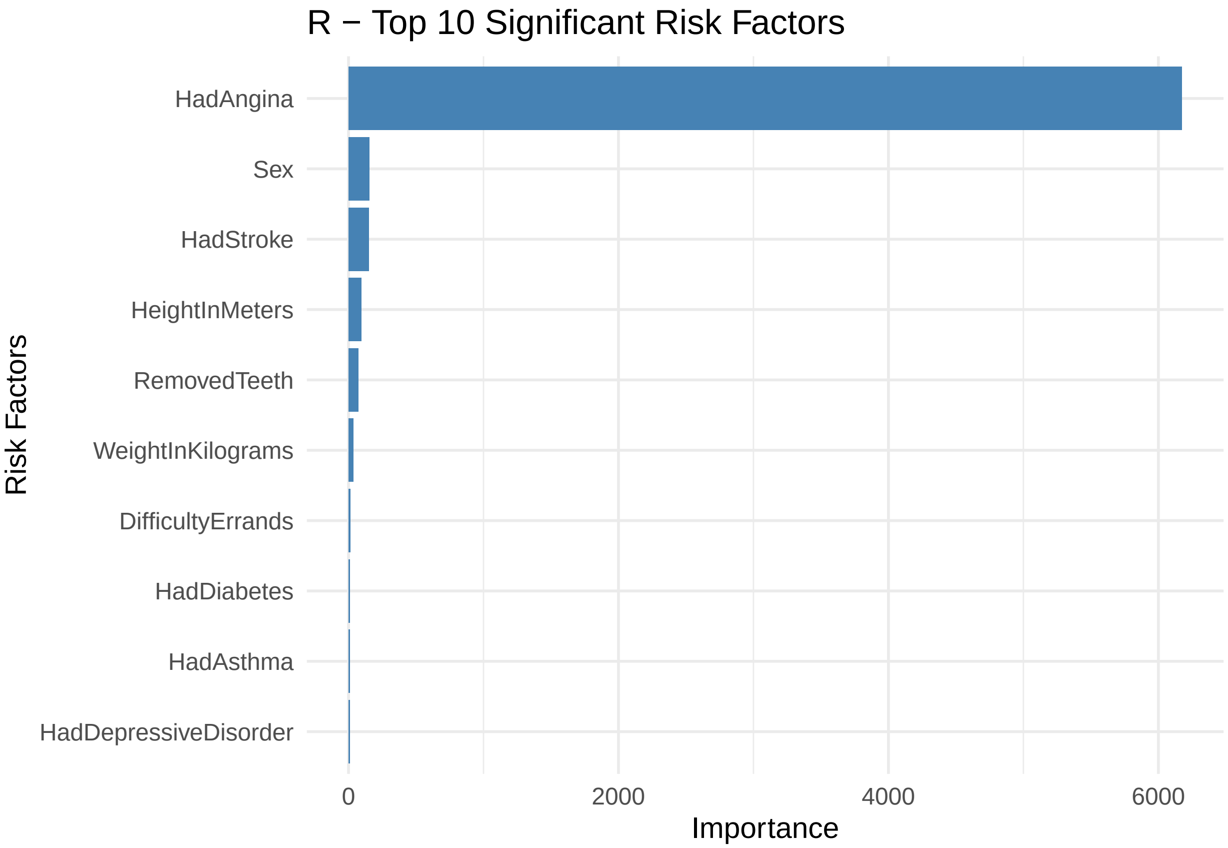
Model Visualization

The rpart.plot function is used to visualize the tree, showing significant risk factors for heart dise

```
# Plot the decision tree with enhanced visualization
rpart.plot(dt_model, main = "Decision Tree for Heart Disease Risk Factors", extra = 106)
```



```r
# Extract variable importance
importance <- as.data.frame(dt_model$variable.importance)
# Convert row names to a proper column for plotting
importance$Variable <- rownames(importance)
# Rename the columns for clarity
colnames(importance) <- c("Importance", "Variable")
# Sort the factors by importance in descending order
importance <- importance[order(-importance$Importance), ]
# Select the top 10 most important factors
top_10_importance <- importance[1:10, ]
# Plot using ggplot2
ggplot(top_10_importance, aes(x = reorder(Variable, Importance), y = Importance)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +  # Flip the coordinates to have a horizontal bar plot
  theme_minimal() +
  labs(title = "R - Top 10 Significant Risk Factors", x = "Risk Factors", y = "Importance")
```


R − Top 10 Significant Risk Factors
Risk Factors
HadAngina
Sex
HadStroke
HeightInMeters
RemovedTeeth
WeightInKilograms
DifficultyErrands
HadDiabetes
HadAsthma
HadDepressiveDisorder
Importance