Polycystic Ovary Syndrome (PCOS)Risk Prediction: Analysis using machine learning models

Harshitha Ponugoti

2024-11-09

Introduction

Polycystic Ovary Syndrome (PCOS) is a common endocrine disorder affecting women of reproductive age, marked by a range of physical and biochemical symptoms. It impacts not only reproductive health but also metabolic, cardiovascular, and mental well-being. Early diagnosis and management are essential to mitigate long-term health risks and improve quality of life.

Key Components of PCOS Health Monitoring:

- **Demographic Indicators**: Basic demographic information, such as age, height, weight, and Body Mass Index (BMI), provides foundational insights into patient profiles and helps assess obesity or underweight status, which are critical in PCOS risk assessment.
- **Medical History**: Detailed records of menstrual cycle regularity, hair growth patterns, and acne presence give insight into hormonal imbalances and other symptoms associated with PCOS. Menstrual irregularities and androgen-related symptoms (like excess hair growth) are often primary indicators of PCOS.
- **Blood Tests**: Hormonal levels (including FSH, LH, and AMH), glucose levels, and cholesterol levels are essential to understanding the metabolic and endocrine profiles of individuals. These indicators reveal underlying metabolic risks such as insulin resistance, which is commonly associated with PCOS and can lead to diabetes if unmanaged.
- **Physical Characteristics**: Waist and hip circumference measurements help assess fat distribution, an important factor in evaluating metabolic health and risks associated with PCOS. An increased waist-to-hip ratio is often correlated with a higher likelihood of metabolic complications.
- **Ultrasound Examination**: Results of ultrasound examinations, specifically examining the presence and size of ovarian follicles and ovarian volume, are central to PCOS diagnosis. The detection of polycystic ovaries is one of the diagnostic criteria for PCOS and provides insight into ovarian health.

Importance of PCOS Health and Risk Prediction

PCOS risk prediction enables healthcare professionals to assess the likelihood of PCOS and prioritize early interventions. Given the variety of symptoms and potential complications associated with PCOS, predictive modeling allows for personalized treatment plans, ultimately reducing the risks of long-term health issues like diabetes, cardiovascular disease, and infertility.

Project Goals: Predictive Modeling for PCOS Diagnosis and Risk Factor Analysis

This project leverages demographic, medical, and clinical data to develop predictive models that can accurately diagnose PCOS and assess risk factors associated with the condition. By testing multiple classification models, we aim to determine the best-performing model for PCOS diagnosis, using metrics such as accuracy, precision, and recall to compare effectiveness. The project also highlights key health indicators, guiding healthcare providers on factors to prioritize for early detection and treatment, contributing to improved outcomes for women with PCOS.

Load Libraries

```r
library(caTools)

## Warning: package 'caTools' was built under R version 4.4.1

library(class)

## Warning: package 'class' was built under R version 4.4.1

library(e1071)

## Warning: package 'e1071' was built under R version 4.4.1

library(rpart)

## Warning: package 'rpart' was built under R version 4.4.1

library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 4.4.1

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.4.1

library(randomForest)

## Warning: package 'randomForest' was built under R version 4.4.1

## randomForest 4.7-1.2

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'
## The following object is masked from 'package:ggplot2':
## 
##     margin
```

```r
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.4.1
```

Load Dataset

```r
dataset <- read.csv(file.choose())
str(dataset)
```

```
## 'data.frame':    541 obs. of  43 variables:
##  $ Sl..No               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Patient.File.No.     : int  10001 10002 10003 10004 10005 10006 10007 1
0008 10009 10010 ...
##  $ PCOS..Y.N.           : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Age..yrs.            : int  28 36 33 37 25 36 34 33 32 36 ...
##  $ Weight..Kg.          : num  44.6 65 68.8 65 52 74.1 64 58.5 40 52 ...
##  $ Height.Cm.           : num  152 162 165 148 161 ...
##  $ BMI                  : num  19.3 24.9 25.3 29.7 20.1 ...
##  $ Blood.Group          : int  15 15 11 13 11 15 11 13 11 15 ...
##  $ Pulse.rate.bpm.      : int  78 74 72 72 72 78 72 72 72 80 ...
##  $ RR..breaths.min.     : int  22 20 18 20 18 28 18 20 18 20 ...
##  $ Hb.g.dl.             : num  10.5 11.7 11.8 12 10 ...
##  $ Cycle.R.I.           : int  2 2 2 2 2 2 2 2 2 4 ...
##  $ Cycle.length.days.   : int  5 5 5 5 5 5 5 5 5 2 ...
##  $ Marraige.Status..Yrs.: num  7 11 10 4 1 8 2 13 8 4 ...
##  $ Pregnant.Y.N.        : int  0 1 1 0 1 1 0 1 0 0 ...
##  $ No..of.aborptions    : int  0 0 0 0 0 0 0 2 1 0 ...
##  $ FSH.mIU.mL.          : num  7.95 6.73 5.54 8.06 3.98 3.24 2.85 4.86 3.7
6 2.8 ...
##  $ LH.mIU.mL.           : num  3.68 1.09 0.88 2.36 0.9 1.07 0.31 3.07 3.02
1.51 ...
##  $ FSH.LH               : num  2.16 6.17 6.3 3.42 4.42 ...
##  $ Hip.inch.            : int  36 38 40 42 37 44 39 44 39 40 ...
##  $ Waist.inch.          : int  30 32 36 36 30 38 33 38 35 38 ...
##  $ Waist.Hip.Ratio      : num  0.833 0.842 0.9 0.857 0.811 ...
##  $ TSH..mIU.L.          : num  0.68 3.16 2.54 16.41 3.57 ...
##  $ AMH.ng.mL.           : chr  "2.07" "1.53" "6.63" "1.22" ...
##  $ PRL.ng.mL.           : num  45.2 20.1 10.5 36.9 30.1 ...
##  $ Vit.D3..ng.mL.       : num  17.1 61.3 49.7 33.4 43.8 52.4 42.7 38 21.8
27.7 ...
##  $ PRG.ng.mL.           : num  0.57 0.97 0.36 0.36 0.38 0.3 0.46 0.26 0.3
0.25 ...
##  $ RBS.mg.dl.           : num  92 92 84 76 84 76 93 91 116 125 ...
##  $ Weight.gain.Y.N.     : int  0 0 0 0 0 1 0 1 0 0 ...
##  $ hair.growth.Y.N.     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Skin.darkening..Y.N. : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hair.loss.Y.N.       : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ Pimples.Y.N.        : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ Fast.food..Y.N.     : int  1 0 1 0 0 0 0 0 0 0 ...
##  $ Reg.Exercise.Y.N.   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ BP._Systolic..mmHg. : int  110 120 120 120 120 110 120 120 120 110 ...
##  $ BP._Diastolic..mmHg.: int  80 70 80 70 80 70 80 80 80 80 ...
##  $ Follicle.No...L.    : int  3 3 13 2 3 9 6 7 5 1 ...
##  $ Follicle.No...R.    : int  3 5 15 2 4 6 6 6 7 1 ...
##  $ Avg..F.size..L...mm.: num  18 15 18 15 16 16 15 15 17 14 ...
##  $ Avg..F.size..R...mm.: num  18 14 20 14 14 20 16 18 17 17 ...
##  $ Endometrium..mm.    : num  8.5 3.7 10 7.5 7 8 6.8 7.1 4.2 2.5 ...
##  $ X                   : chr  "" "" "" "" ...
```

```
summary(dataset)
```

```
##      Sl..No    Patient.File.No.   PCOS..Y.N.       Age..yrs.
##  Min.   :  1   Min.   :10001    Min.   :0.0000   Min.   :20.00
##  1st Qu.:136   1st Qu.:10136    1st Qu.:0.0000   1st Qu.:28.00
##  Median :271   Median :10271    Median :0.0000   Median :31.00
##  Mean   :271   Mean   :10271    Mean   :0.3272   Mean   :31.43
##  3rd Qu.:406   3rd Qu.:10406    3rd Qu.:1.0000   3rd Qu.:35.00
##  Max.   :541   Max.   :10541    Max.   :1.0000   Max.   :48.00
##
##   Weight..Kg.       Height.Cm.         BMI         Blood.Group
##  Min.   : 31.00   Min.   :137.0   Min.   :12.42   Min.   :11.0
##  1st Qu.: 52.00   1st Qu.:152.0   1st Qu.:21.64   1st Qu.:13.0
##  Median : 59.00   Median :156.0   Median :24.24   Median :14.0
##  Mean   : 59.64   Mean   :156.5   Mean   :24.31   Mean   :13.8
##  3rd Qu.: 65.00   3rd Qu.:160.0   3rd Qu.:26.63   3rd Qu.:15.0
##  Max.   :108.00   Max.   :180.0   Max.   :38.90   Max.   :18.0
##
##  Pulse.rate.bpm. RR..breaths.min.    Hb.g.dl.       Cycle.R.I.
##  Min.   :13.00   Min.   :16.00    Min.   : 8.50   Min.   :2.00
##  1st Qu.:72.00   1st Qu.:18.00    1st Qu.:10.50   1st Qu.:2.00
##  Median :72.00   Median :18.00    Median :11.00   Median :2.00
##  Mean   :73.25   Mean   :19.24    Mean   :11.16   Mean   :2.56
##  3rd Qu.:74.00   3rd Qu.:20.00    3rd Qu.:11.70   3rd Qu.:4.00
##  Max.   :82.00   Max.   :28.00    Max.   :14.80   Max.   :5.00
##
##  Cycle.length.days. Marraige.Status..Yrs. Pregnant.Y.N.    No..of.aborptio
ns
##  Min.   : 0.000     Min.   : 0.000        Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 4.000     1st Qu.: 4.000        1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 5.000     Median : 7.000        Median :0.0000   Median :0.0000
##  Mean   : 4.941     Mean   : 7.681        Mean   :0.3808   Mean   :0.2884
##  3rd Qu.: 5.000     3rd Qu.:10.000        3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :12.000     Max.   :30.000        Max.   :1.0000   Max.   :5.0000
##                     NA's   :1
##   FSH.mIU.mL.        LH.mIU.mL.          FSH.LH            Hip.inch.
##  Min.   :   0.21   Min.   :   0.02   Min.   :   0.0021   Min.   :26.00
```

Data Preprocessing

```r
# Drop unnecessary columns
dataset <- dataset[, !(names(dataset) %in% c("X"))]
dataset$AMH.ng.mL. <- as.numeric(as.character(dataset$AMH.ng.mL.))
dataset <- na.omit(dataset)
# Check for missing values
null_counts <- colSums(is.na(dataset))
print(null_counts[null_counts > 0])
```

```
## named numeric(0)
```

Data Splitting

```r
# Split the data (80% training, 20% testing)
set.seed(123)
split <- sample.split(dataset$PCOS..Y.N., SplitRatio = 0.8)
training_set <- subset(dataset, split == TRUE)
test_set <- subset(dataset, split == FALSE)
train_label <- training_set$PCOS..Y.N.
test_label <- test_set$PCOS..Y.N.
```

K-Nearest Neighbors (KNN)

```r
# KNN Model
knn_class <- knn(train = training_set[, -which(names(training_set) == "PCOS..
Y.N.")],
                 test = test_set[, -which(names(test_set) == "PCOS..Y.N.")],
                 cl = train_label, k = 5)
cm_knn <- table(test_label, knn_class)
acc_knn <- sum(diag(cm_knn)) / sum(cm_knn)
paste("Accuracy KNN: ", round(acc_knn * 100, 2), "%")
```

```
## [1] "Accuracy KNN:  73.83 %"
```

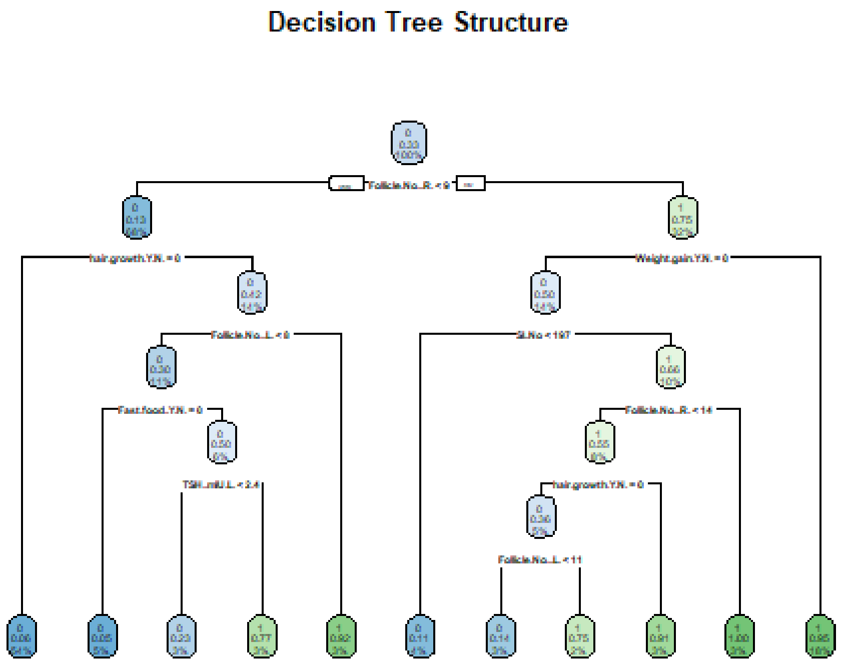
Naive Bayes Model

```r
# Naive Bayes Model
model_naive <- naiveBayes(PCOS..Y.N. ~ ., data = training_set)
predict_naive <- predict(model_naive, newdata = test_set)
cm_naive <- table(test_label, predict_naive)
acc_naive <- sum(diag(cm_naive)) / sum(cm_naive)
paste("Accuracy Naive Bayes: ", round(acc_naive * 100, 2), "%")
```

```
## [1] "Accuracy Naive Bayes:  82.24 %"
```

Decision Tree Model

```r
# Decision Tree Model
dt_model <- rpart(PCOS..Y.N. ~ ., data = training_set, method = "class")
rpart.plot(dt_model, main = "Decision Tree Structure")
```


Decision Tree Structure

```
predict_dt <- predict(dt_model, test_set, type = "class")
cm_dt <- table(test_label, predict_dt)
acc_dt <- sum(diag(cm_dt)) / sum(cm_dt)
paste("Accuracy Decision Tree: ", round(acc_dt * 100, 2), "%")

## [1] "Accuracy Decision Tree:  85.05 %"
```

Random Forest Model

```
# Random Forest Model
training_set$PCOS..Y.N. <- as.factor(training_set$PCOS..Y.N.)
test_label <- factor(test_label, levels = levels(training_set$PCOS..Y.N.))
rf_model <- randomForest(PCOS..Y.N. ~ ., data = training_set)
predict_rf <- predict(rf_model, test_set, type = "response")
cm_rf <- table(test_label, predict_rf)
acc_rf <- sum(diag(cm_rf)) / sum(cm_rf)
paste("Accuracy Random Forest: ", round(acc_rf * 100, 2), "%")

## [1] "Accuracy Random Forest:  94.39 %"
```

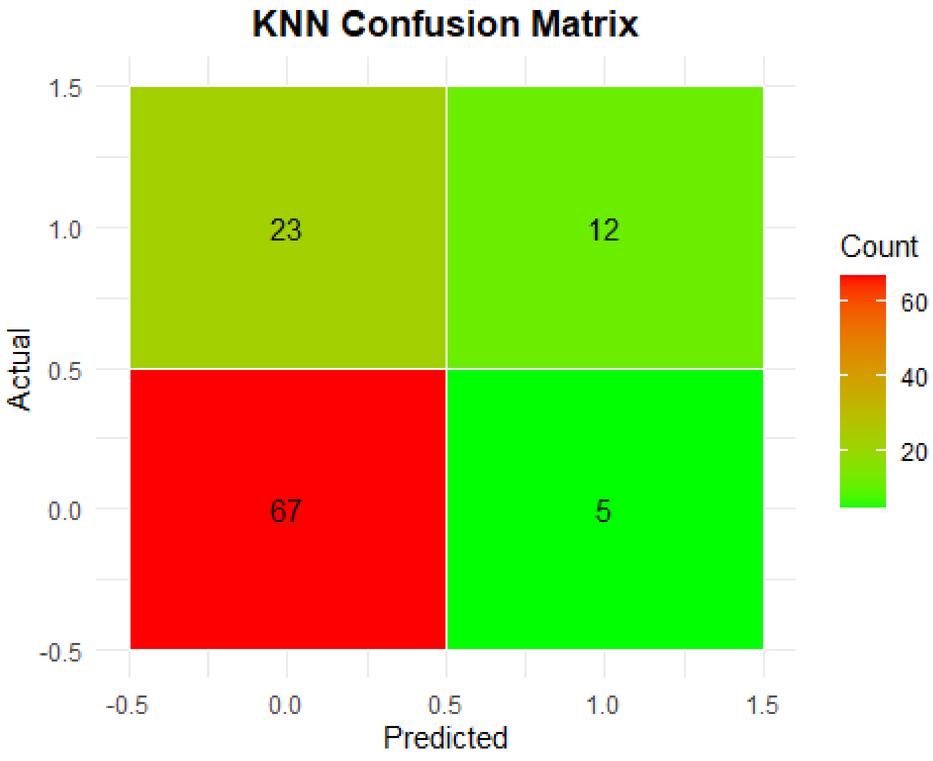
Confusion Matrix Plotting

```
plot_cm <- function(cm, title) {
  cm_df <- melt(as.table(cm))
  colnames(cm_df) <- c("Actual", "Predicted", "Count")

  ggplot(cm_df, aes(x = Predicted, y = Actual, fill = Count)) +
    geom_tile(color = "white") +
    scale_fill_gradient(low = "green", high = "red") +
```

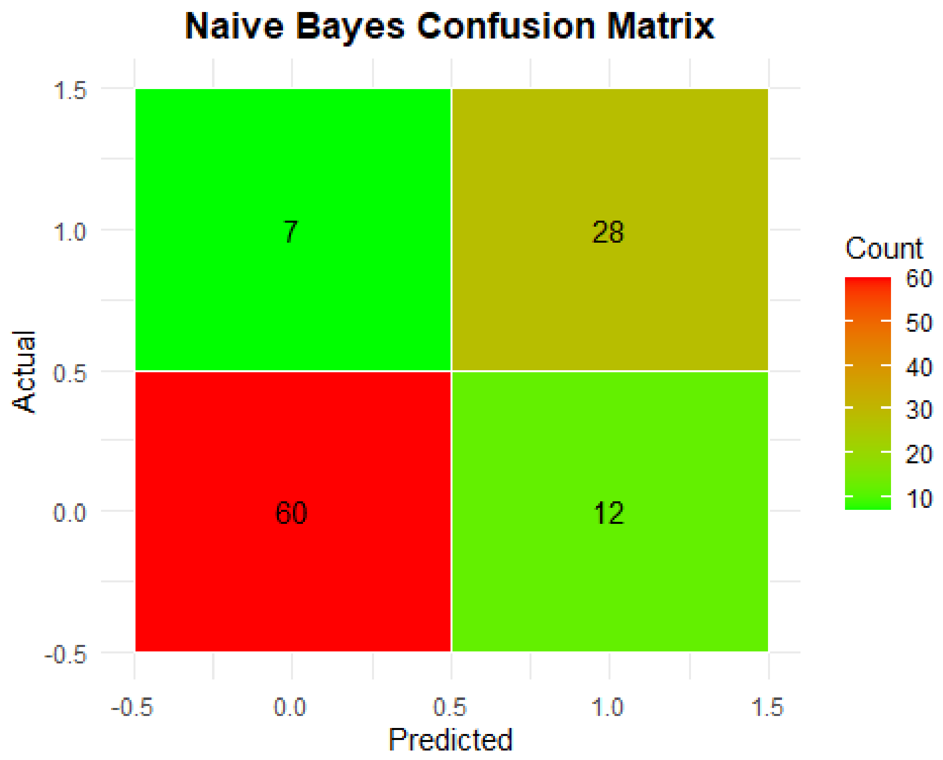
```r
    geom_text(aes(label = Count), color = "black") +
    labs(title = title, x = "Predicted", y = "Actual", fill = "Count") +
    theme_minimal() +
    theme(plot.title = element_text(hjust = 0.5, face = "bold"))
}

# Plot Confusion Matrices
if (dev.cur() != 1) dev.off()
```

```
## null device
##           1
```

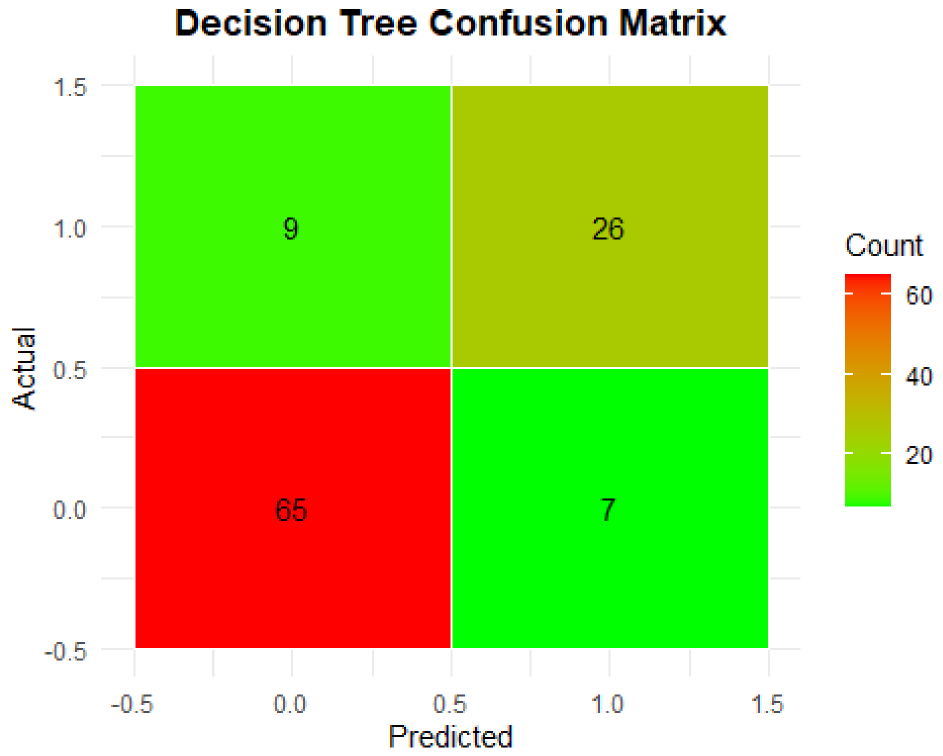
```r
plot_cm(cm_knn, "KNN Confusion Matrix")
```



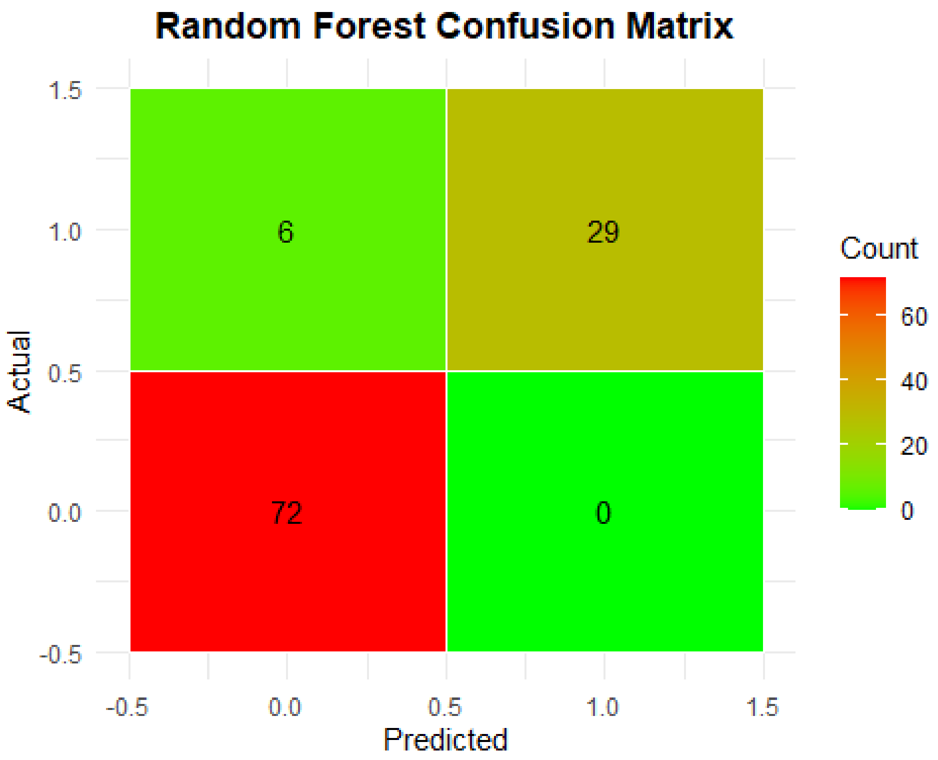
```r
plot_cm(cm_naive, "Naive Bayes Confusion Matrix")
```



```r
plot_cm(cm_dt, "Decision Tree Confusion Matrix")
```

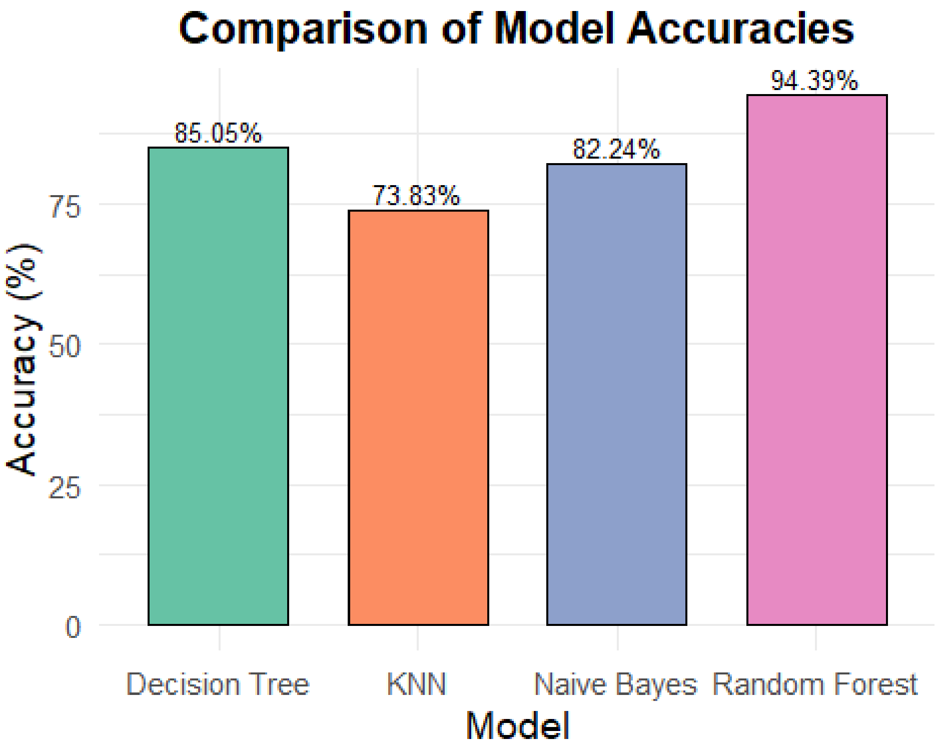


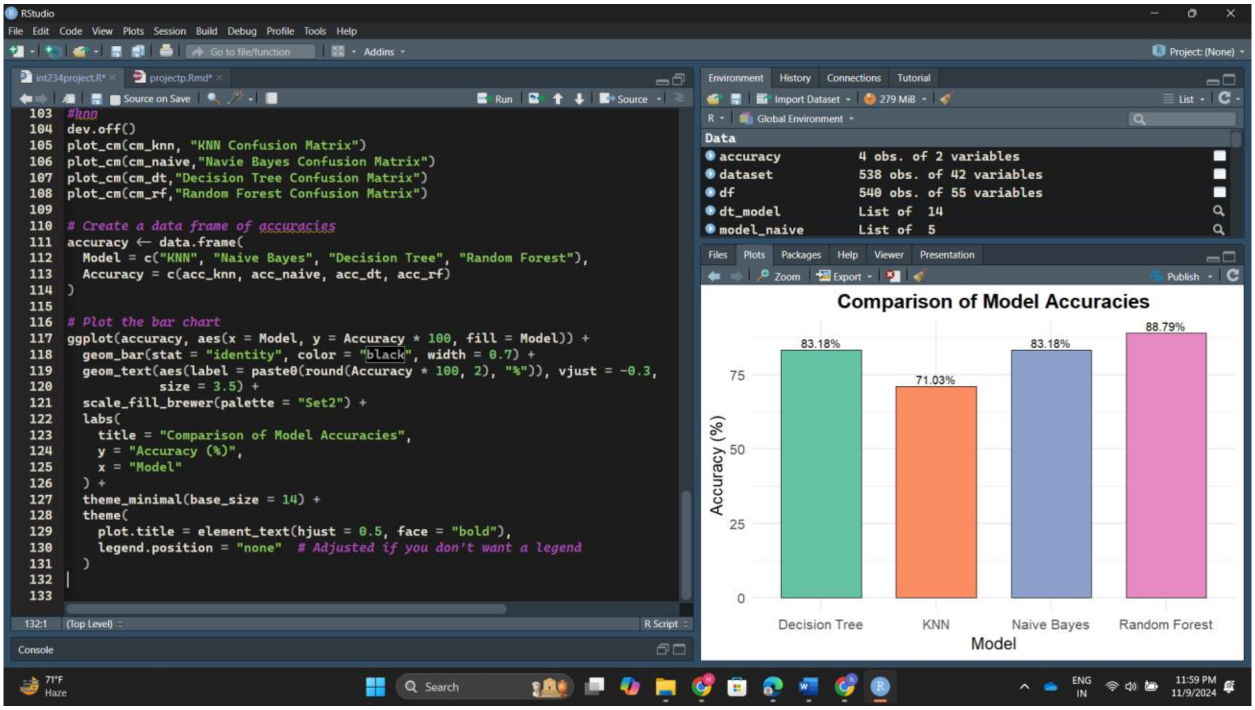
```r
plot_cm(cm_rf, "Random Forest Confusion Matrix")
```



Model Accuracy Comparison

```r
# Create a data frame of accuracies
accuracy <- data.frame(
  Model = c("KNN", "Naive Bayes", "Decision Tree", "Random Forest"),
  Accuracy = c(acc_knn, acc_naive, acc_dt, acc_rf)
)

# Plot the Bar Chart of Accuracies
ggplot(accuracy, aes(x = Model, y = Accuracy * 100, fill = Model)) +
  geom_bar(stat = "identity", color = "black", width = 0.7) +
  geom_text(aes(label = paste0(round(Accuracy * 100, 2), "%")), vjust = -0.3, size = 3.5) +
  scale_fill_brewer(palette = "Set2") +
  labs(title = "Comparison of Model Accuracies", y = "Accuracy (%)", x = "Model") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"), legend.position = "none")
```





```r
#knn
dev.off()
plot_cm(cm_knn, "KNN Confusion Matrix")
plot_cm(cm_naive,"Navie Bayes Confusion Matrix")
plot_cm(cm_dt,"Decision Tree Confusion Matrix")
plot_cm(cm_rf,"Random Forest Confusion Matrix")

# Create a data frame of accuracies
accuracy <- data.frame(
  Model = c("KNN", "Naive Bayes", "Decision Tree", "Random Forest"),
  Accuracy = c(acc_knn, acc_naive, acc_dt, acc_rf)
)

# Plot the bar chart
ggplot(accuracy, aes(x = Model, y = Accuracy * 100, fill = Model)) +
  geom_bar(stat = "identity", color = "black", width = 0.7) +
  geom_text(aes(label = paste0(round(Accuracy * 100, 2), "%")), vjust = -0.3,
            size = 3.5) +
  scale_fill_brewer(palette = "Set2") +
  labs(
    title = "Comparison of Model Accuracies",
    y = "Accuracy (%)",
    x = "Model"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    legend.position = "none"  # Adjusted if you don't want a legend
  )
```

```
Data
accuracy        4 obs. of 2 variables
dataset         538 obs. of 42 variables
df              540 obs. of 55 variables
dt_model        List of  14
model_naive     List of  5
```

Conclusion

In this analysis, we implemented four different machine learning models to predict PCOS. The accuracies of the models were compared, and the results were visualized using confusion matrices and a bar plot of accuracies.