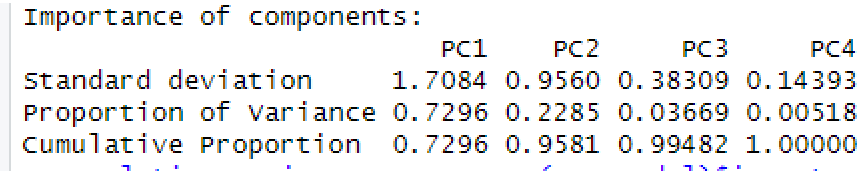
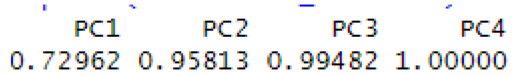
Principal Component Analysis

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(caret)
library(openxlsx)
library(scales)
library(factoextra)
library(nnet)

# ----------------------
# 1. Load and Inspect
# ----------------------
data(iris)
df <- iris

# ----------------------
# 2. Min-Max Normalization
# ----------------------
normalize <- function(x) {
  return((x - min(x)) / (max(x) - min(x)))
}

df_norm <- df %>%
  mutate(across(where(is.numeric), normalize))

# ----------------------
# 3. Standardization (Z-score)
# ----------------------
df_scaled <- df_norm %>%
  mutate(across(where(is.numeric), scale))

# ----------------------
# 4. PCA (all components)
# ----------------------
pca_model <- prcomp(df_scaled[, 1:4], center = FALSE, scale. = FALSE)
summary(pca_model)
```



```
Importance of components:
                          PC1    PC2     PC3     PC4
Standard deviation     1.7084 0.9560 0.38309 0.14393
Proportion of Variance 0.7296 0.2285 0.03669 0.00518
Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

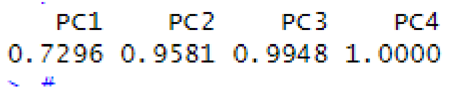
```
cumulative_variance <- summary(pca_model)$importance[3, ]
print(cumulative_variance)
```



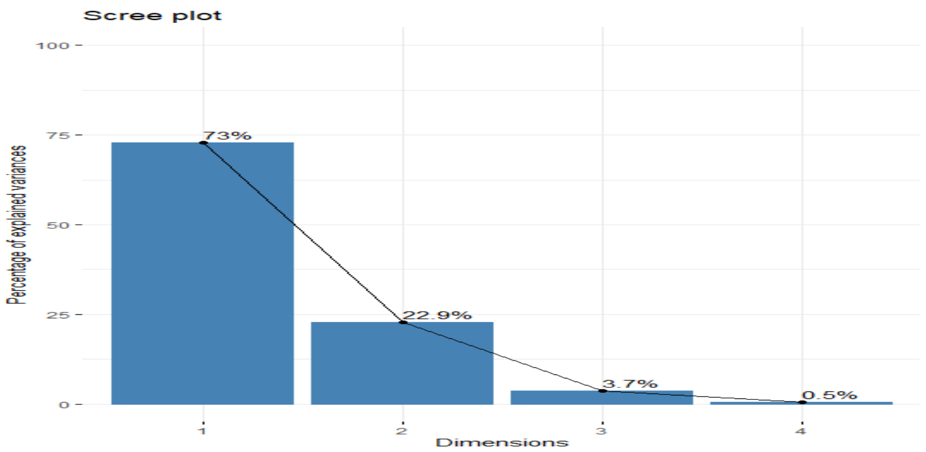
```
    PC1     PC2     PC3     PC4 
0.72962 0.95813 0.99482 1.00000 
```

```
fviz_eig(pca_model, addlabels = TRUE, ylim = c(0, 100))


# Get PC scores
```

```
pca_scores <- as.data.frame(pca_model$x)
pca_scores$Species <- df$Species

# ----------------------
# 5. Export to Excel
# ----------------------
output <- cbind(df, df_norm[, 1:4], df_scaled[, 1:4], pca_scores)
write.xlsx(output, "pca_output_R.xlsx", rowNames = FALSE)
cat("✅ Exported to pca_output_R.xlsx\n")
# Explained variance (importance matrix)
importance_matrix <- summary(pca_model)$importance

# Save variance ratio and cumulative variance
explained_variance_ratio <- importance_matrix[2, ]
cumulative_variance <- importance_matrix[3, ]

# Display in console
cat("\nExplained Variance Ratio:\n")
print(round(explained_variance_ratio, 4))

cat("\nCumulative Variance:\n")
print(round(cumulative_variance, 4))
```



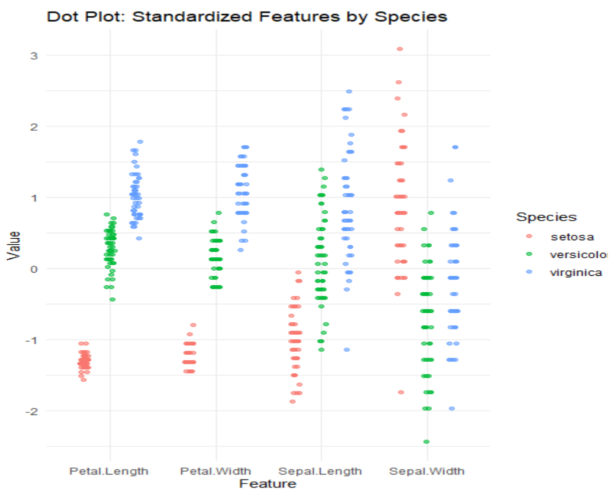
```
   PC1    PC2    PC3    PC4
0.7296 0.9581 0.9948 1.0000
```

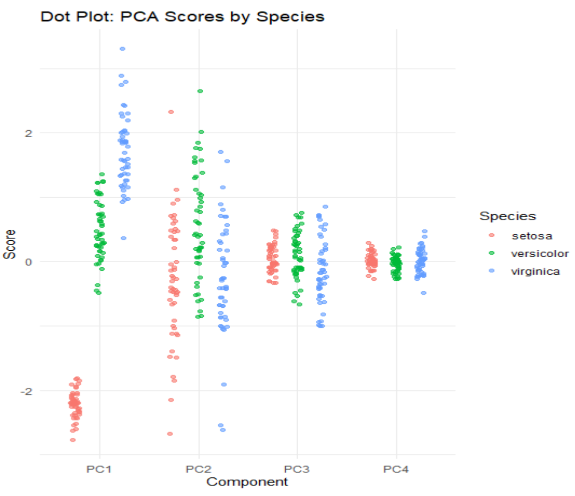


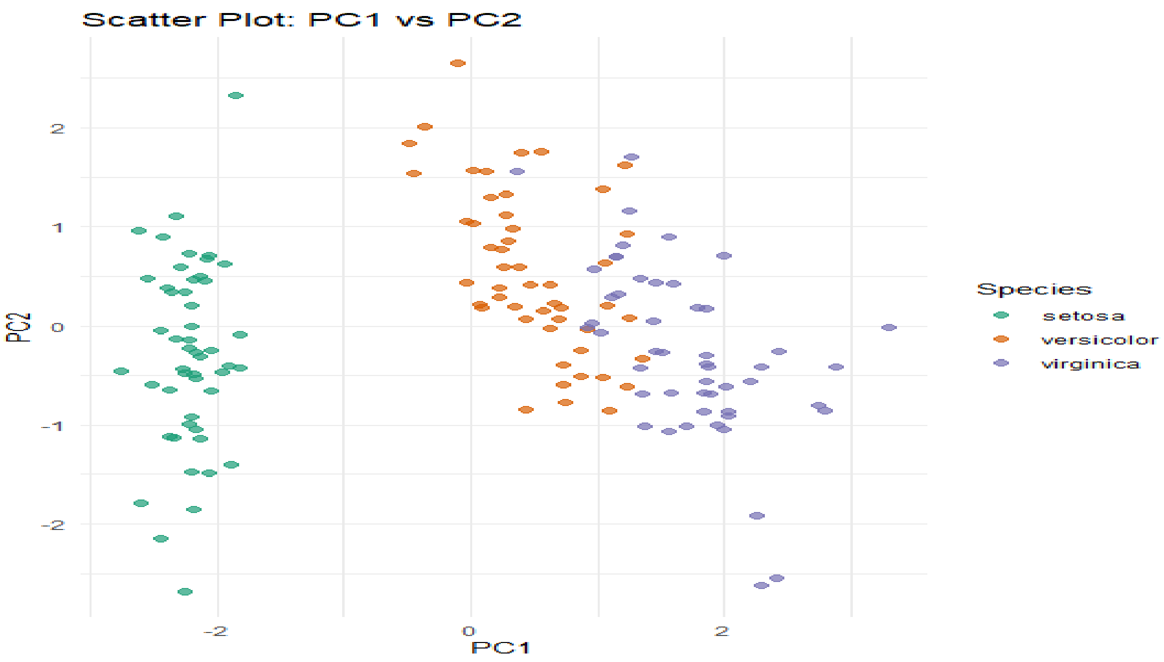
The PCA results revealed that the first two principal components (PC1 and PC2) together explained approximately 95.81% of the total variance in the dataset, indicating that a large proportion of the data's structure could be captured in just two dimensions. This dimensionality reduction facilitated clear visualization and efficient classification.

```
# ----------------------
# 6. Dot Plot: Standardized Features
# ----------------------
df_scaled$Species <- df$Species
df_scaled$sample <- 1:nrow(df_scaled)

df_scaled_long <- pivot_longer(df_scaled, cols = 1:4, names_to = "Feature", values_to = "Value")

ggplot(df_scaled_long, aes(x = Feature, y = Value, color = Species)) +
  geom_jitter(position = position_jitterdodge(jitter.width = 0.2), alpha = 0.6) +
  theme_minimal() +
  ggtitle("Dot Plot: Standardized Features by Species")

df_pca_long <- pca_scores %>%
  mutate(sample = 1:nrow(.)) %>%
  pivot_longer(cols = starts_with("PC"),
               names_to = "Component",
               values_to = "Score")
pca_scores <- as.data.frame(pca_model$x)
pca_scores$Species <- df$Species  # <-- this line must be included



# ----------------------
# 7. Dot Plot: PCA Components
# ----------------------
df_pca_long <- pca_scores %>%
  mutate(sample = 1:nrow(.)) %>%
  pivot_longer(cols = starts_with("PC"), names_to = "Component", values_to = "Score")

ggplot(df_pca_long, aes(x = Component, y = Score, color = Species)) +
  geom_jitter(position = position_jitterdodge(jitter.width = 0.2), alpha = 0.6) +
  theme_minimal() +
  ggtitle("Dot Plot: PCA Scores by Species")

table(pca_scores$Species)
pca_scores$Species <- df$Species
```

qq

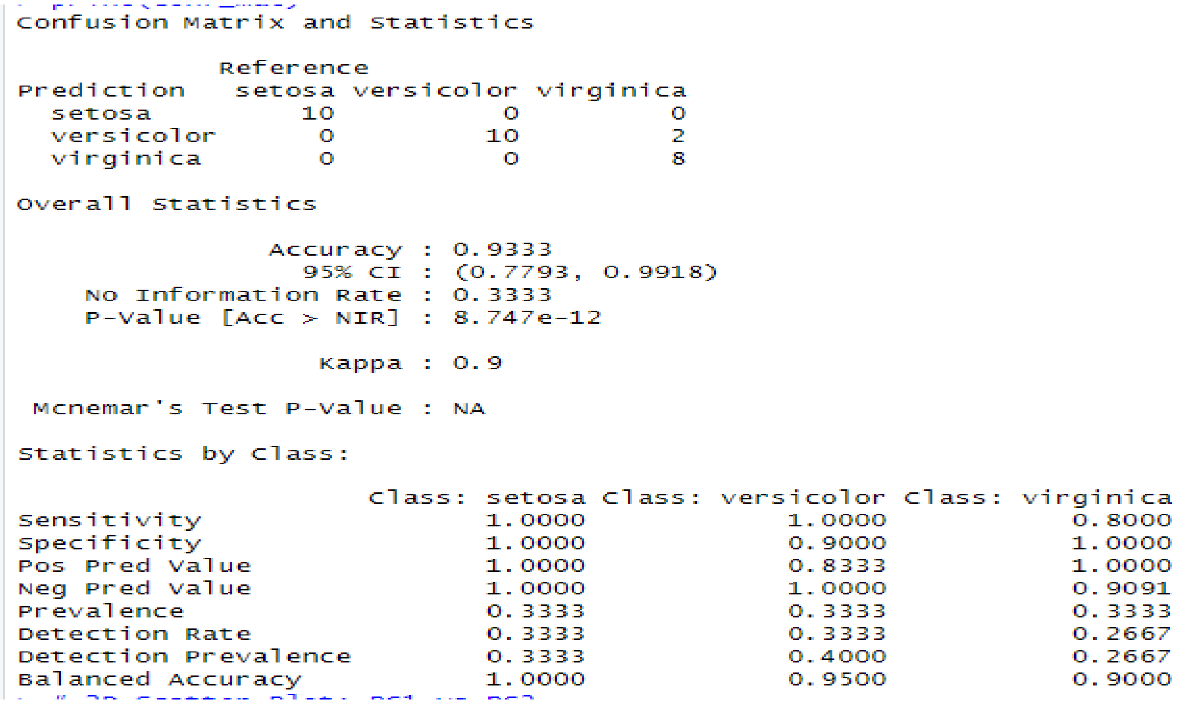






A dot plot of the standardized features grouped by species showed clear separation of Setosa from the other two species, while some overlap remained between Versicolor and Virginica, which is typical for this dataset. The scores plot from PCA confirmed these findings: Setosa was distinctly separated along the first principal component, whereas Versicolor and Virginica exhibited some overlap but could still be partially distinguished along the second component. This visualization validated the effectiveness of PCA in capturing inter-species variation.

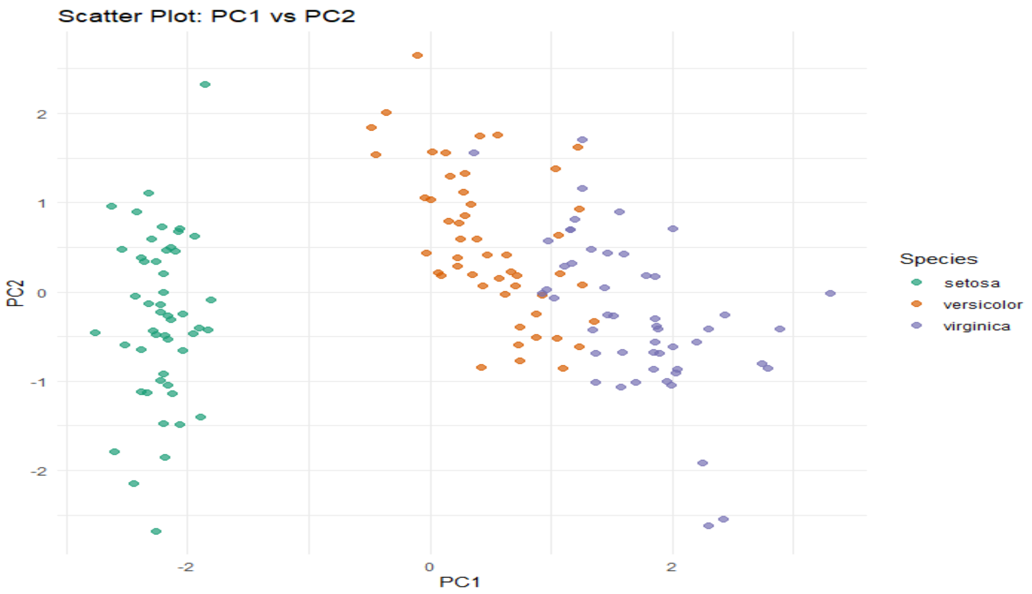
```
# ----------------------
# 8. Classification + Confusion Matrix (Logistic Regression)
# ----------------------
set.seed(123)
train_index <- createDataPartition(pca_scores$Species, p = 0.8, list = FALSE)
train_data <- pca_scores[train_index, ]
test_data <- pca_scores[-train_index, ]

# Multinomial logistic regression
model <- multinom(Species ~ PC1 + PC2 + PC3 + PC4, data = train_data)

# Prediction
pred <- predict(model, newdata = test_data)
conf_mat <- confusionMatrix(pred, test_data$Species)

print(conf_mat)
```



```
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         10          0         0
  versicolor      0         10         2
  virginica       0          0         8

Overall Statistics
                                          
               Accuracy : 0.9333          
                 95% CI : (0.7793, 0.9918)
    No Information Rate : 0.3333          
    P-Value [Acc > NIR] : 8.747e-12       
                                          
                  Kappa : 0.9             
                                          
 Mcnemar's Test P-Value : NA              

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            1.0000           0.8000
Specificity                 1.0000            0.9000           1.0000
Pos Pred Value              1.0000            0.8333           1.0000
Neg Pred Value              1.0000            1.0000           0.9091
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.3333           0.2667
Detection Prevalence        0.3333            0.4000           0.2667
Balanced Accuracy           1.0000            0.9500           0.9000
```

To assess classification performance, a multinomial logistic regression model was fitted using the principal component scores as predictors. The model achieved a high overall accuracy of 93.3%, with a strong Kappa statistic of 0.9, indicating substantial agreement beyond chance. Sensitivity and specificity for Setosa were perfect, while Versicolor and Virginica showed slightly reduced but still strong performance, with minor confusion between them.

```
# 2D Scatter Plot: PC1 vs PC2
ggplot(pca_scores, aes(x = PC1, y = PC2, color = Species)) +
  geom_point(size = 2, alpha = 0.7) +
  theme_minimal() +
  labs(title = "Scatter Plot: PC1 vs PC2", x = "PC1", y = "PC2") +
  scale_color_brewer(palette = "Dark2")
# Dot Plot of PCA Scores (PC1–PC4)

ggplot(df_pca_long, aes(x = Component, y = Score, color = Species)) +
  geom_jitter(position = position_jitterdodge(jitter.width = 0.2), alpha = 0.6) +
  theme_minimal() +
  labs(title = "Dot Plot: PCA Components by Species", y = "PCA Score") +
  scale_color_brewer(palette = "Set2")
```



The PC1 vs. PC2 scatter plot further confirmed that PCA effectively compressed the data into a lower-dimensional space, preserving the key discriminatory information among the species.