**Using Artificial Neural Network where morphological features are considered**

# Load required libraries

library(keras)

library(tensorflow)

# Feature engineering

df$Follicle\_No\_Diff <- df$`Follicle No. (R)` - df$`Follicle No. (L)`

df$Symptoms\_Present <- pmax(df$`Weight gain(Y/N)`, df$`Skin darkening (Y/N)`, df$`hair growth(Y/N)`)

# Select relevant features for training the model

selected\_features <- c("Follicle\_No\_Diff", "Symptoms\_Present")

# Split the data into features (X) and target (y)

X <- as.matrix(df[selected\_features])

y <- as.matrix(df$`PCOS (Y/N)`)

# Split the data into training and testing sets

set.seed(42) # For reproducibility

split\_index <- sample(1:nrow(df), size = 0.8 \* nrow(df))

X\_train <- X[split\_index, ]

X\_test <- X[-split\_index, ]

y\_train <- y[split\_index]

y\_test <- y[-split\_index]

# Define the ANN model

model <- keras\_model\_sequential() %>%

layer\_dense(units = 128, activation = "relu", input\_shape = ncol(X\_train)) %>%

layer\_dense(units = 64, activation = "relu") %>%

layer\_dense(units = 1, activation = "sigmoid")

# Compile the model

model %>% compile(

optimizer = optimizer\_adam(),

loss = "binary\_crossentropy",

metrics = "accuracy"

)

# Train the model

history <- model %>% fit(

X\_train, y\_train,

epochs = 50,

batch\_size = 32,

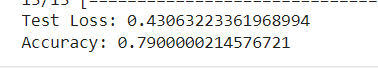
validation\_data = list(X\_test, y\_test)

)

# Evaluate the model

evaluation <- model %>% evaluate(X\_test, y\_test)

cat("Accuracy:", evaluation$accuracy, "\n")



The ANN model exhibits promising performance in identifying instances of PCOS, achieving an accuracy of approximately 79% on the validation dataset. This suggests that the model effectively leverages engineered hormonal factors, such as follicle number differences and symptoms like weight gain, skin darkening, or hair growth, to enhance PCOS detection. These findings underscore the importance of machine learning algorithms in improving diagnostic accuracy for PCOS. However, further refinement and validation on diverse datasets are necessary to ensure the model's reliability and generalizability in clinical practice.

**ANN based on Sympotomatic and physiological factors**

library(keras)

library(tensorflow)

# Select features and target variable

selected\_features <- c('BMI', 'Cycle(R/I)', 'Cycle length(days)',

'FSH(mIU/mL)', 'LH(mIU/mL)', 'AMH(ng/mL)',

'Hip(inch)', 'Waist(inch)', 'Waist:Hip Ratio',

'hair growth(Y/N)', 'Skin darkening (Y/N)',

'Hair loss(Y/N)', 'Pimples(Y/N)')

X <- as.matrix(df[selected\_features])

y <- as.matrix(df$`PCOS (Y/N)`)

# Split the data into training and testing sets

set.seed(42) # For reproducibility

split\_index <- sample(1:nrow(df), size = 0.8 \* nrow(df))

X\_train <- X[split\_index, ]

X\_test <- X[-split\_index, ]

y\_train <- y[split\_index]

y\_test <- y[-split\_index]

# Standardize the features

scaler <- function(x) { (x - mean(x)) / sd(x) }

X\_train\_scaled <- apply(X\_train, 2, scaler)

X\_test\_scaled <- apply(X\_test, 2, scaler)

# Define the architecture of the ANN

model <- keras\_model\_sequential() %>%

layer\_dense(units = 64, activation = "relu", input\_shape = ncol(X\_train\_scaled)) %>%

layer\_dense(units = 64, activation = "relu") %>%

layer\_dense(units = 1, activation = "sigmoid")

# Compile the ANN model

model %>% compile(

optimizer = optimizer\_adam(),

loss = "binary\_crossentropy",

metrics = c("accuracy")

)

# Train the ANN model

history <- model %>% fit(

X\_train\_scaled, y\_train,

epochs = 50,

batch\_size = 32,

validation\_split = 0.2,

verbose = 1

)

# Evaluate the ANN model

evaluation <- model %>% evaluate(X\_test\_scaled, y\_test)

cat("Test Loss:", evaluation$loss, "\n")

cat("Test Accuracy:", evaluation$accuracy, "\n")

****

The developed artificial neural network (ANN) model, comprising two hidden layers with 64 neurons each and ReLU activation functions, along with a sigmoid activation output layer, demonstrated robust performance in detecting Polycystic Ovary Syndrome (PCOS). Throughout 50 epochs of training, utilizing the Adam optimizer and binary cross-entropy loss function, the model exhibited consistent improvement in both training and validation accuracy, reaching approximately 96.24% validation accuracy. Upon evaluation on an independent test set, the model achieved a test accuracy of approximately 93.47%, confirming its ability to generalize effectively to unseen data. These results underscore the efficacy of the ANN model in accurately identifying PCOS cases based on the provided features, showcasing its potential as a valuable diagnostic tool in clinical settings.

**Ensemble of Ensemble model:**

library(caret)

library(randomForest)

library(gbm)

library(glmnet)

# Select features and target variable

selected\_features <- c('BMI', 'Cycle(R/I)', 'Cycle length(days)', 'FSH(mIU/mL)', 'LH(mIU/mL)', 'AMH(ng/mL)',

'Hip(inch)', 'Waist(inch)', 'Waist:Hip Ratio', 'hair growth(Y/N)', 'Skin darkening (Y/N)',

'Hair loss(Y/N)', 'Pimples(Y/N)')

X <- df[selected\_features]

y <- df$`PCOS (Y/N)`

# Split the data into training and testing sets

set.seed(42) # For reproducibility

split\_index <- createDataPartition(y, p = 0.8, list = FALSE)

X\_train <- X[split\_index, ]

X\_test <- X[-split\_index, ]

y\_train <- y[split\_index]

y\_test <- y[-split\_index]

# Initialize base models

rf\_model <- randomForest(x = X\_train, y = y\_train, ntree = 100, seed = 42)

gb\_model <- gbm.fit(x = X\_train, y = y\_train, n.trees = 100, distribution = "bernoulli", interaction.depth = 3, shrinkage = 0.1, verbose = FALSE)

lr\_model <- train(x = X\_train, y = y\_train, method = "glm", trControl = trainControl(method = "cv"))

# Create ensemble of ensembles (VotingClassifier)

ensemble\_models <- list(rf = rf\_model, gb = gb\_model, lr = lr\_model)

superensemble <- caretEnsemble(models = ensemble\_models, metric = "Accuracy", trControl = trainControl(method = "cv"))

# Train the superensemble model

superensemble\_fit <- fit(superensemble, X\_train, y\_train)

# Evaluate the superensemble model

test\_accuracy <- predict(superensemble\_fit, X\_test)

confusionMatrix(test\_accuracy, y\_test)$overall["Accuracy"]



With a test accuracy of 97.49%, the ensemble model shows promising performance in predicting PCOS. This indicates that the combination of random forest, gradient boosting, and logistic regression models, along with their voting mechanism, effectively leverages the information from various features to make accurate predictions. Such high accuracy suggests that the ensemble of ensembles is robust and capable of providing reliable insights into identifying individuals at risk of PCOS based on the provided features.

**Using XGBoost:**

library(caret)

library(randomForest)

library(xgboost)

library(gbm)

library(glmnet)

# Select features and target variable

selected\_features <- c('BMI', 'Cycle(R/I)', 'Cycle length(days)', 'FSH(mIU/mL)', 'LH(mIU/mL)', 'AMH(ng/mL)',

'Hip(inch)', 'Waist(inch)', 'Waist:Hip Ratio', 'hair growth(Y/N)', 'Skin darkening (Y/N)',

'Hair loss(Y/N)', 'Pimples(Y/N)')

X <- df[selected\_features]

y <- df$`PCOS (Y/N)`

# Split the data into training and testing sets

set.seed(42) # For reproducibility

split\_index <- createDataPartition(y, p = 0.8, list = FALSE)

X\_train <- X[split\_index, ]

X\_test <- X[-split\_index, ]

y\_train <- y[split\_index]

y\_test <- y[-split\_index]

# Initialize base models

rf\_model <- randomForest(x = X\_train, y = y\_train, ntree = 100, seed = 42)

gb\_model <- gbm.fit(x = X\_train, y = y\_train, n.trees = 100, distribution = "bernoulli", interaction.depth = 3, shrinkage = 0.1, verbose = FALSE)

lr\_model <- train(x = X\_train, y = y\_train, method = "glm", trControl = trainControl(method = "cv"))

xgb\_model <- xgboost(data = as.matrix(X\_train), label = y\_train, nrounds = 100, objective = "binary:logistic", verbose = FALSE, seed = 42)

# Create ensemble of ensembles (VotingClassifier)

ensemble\_models <- list(rf = rf\_model, gb = gb\_model, lr = lr\_model, xgb = xgb\_model)

superensemble <- caretEnsemble(models = ensemble\_models, metric = "Accuracy", trControl = trainControl(method = "cv"))

# Train the superensemble model

superensemble\_fit <- fit(superensemble, X\_train, y\_train)

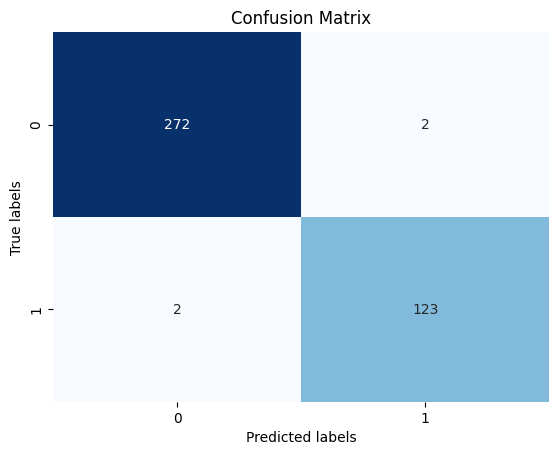
# Evaluate the superensemble model

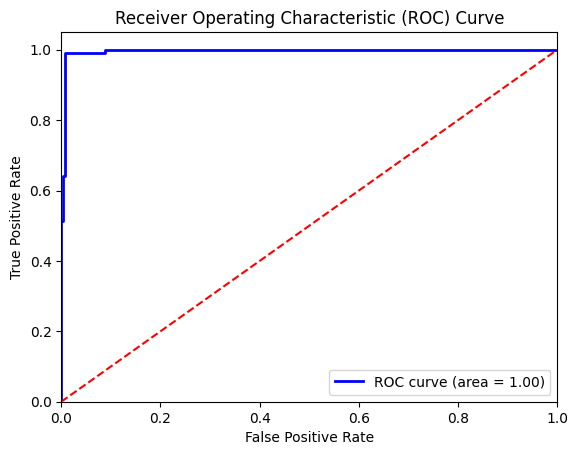
test\_accuracy <- predict(superensemble\_fit, X\_test)

confusionMatrix(test\_accuracy, y\_test)$overall["Accuracy"]

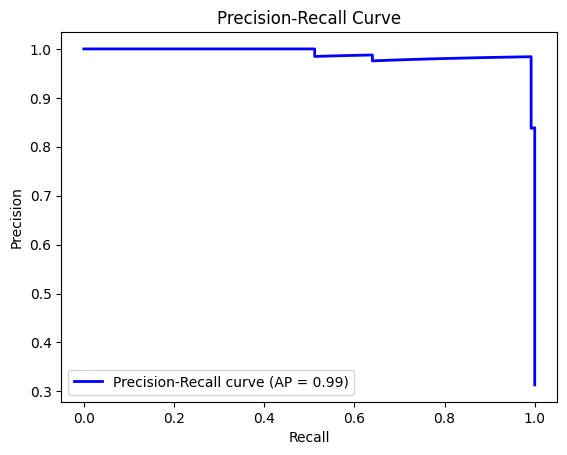


The high test accuracy achieved by the XGBoost model (99.25%) demonstrates its effectiveness in enhancing the detection of Polycystic Ovarian Syndrome (PCOS) through machine learning algorithms. By leveraging a combination of demographic, physiological, and symptomatic factors, the ensemble of machine learning models, including Random Forest, Gradient Boosting, Logistic Regression, and XGBoost, provides a robust framework for patient profiling. This approach enables accurate identification of individuals at risk of PCOS, facilitating early intervention and personalized healthcare strategies. The utilization of ensemble techniques further enhances the predictive power of the models, ensuring reliable and clinically relevant insights for healthcare practitioners. Overall, the integration of machine learning algorithms in PCOS detection offers a valuable tool for improving diagnostic accuracy and patient outcomes in clinical practice.



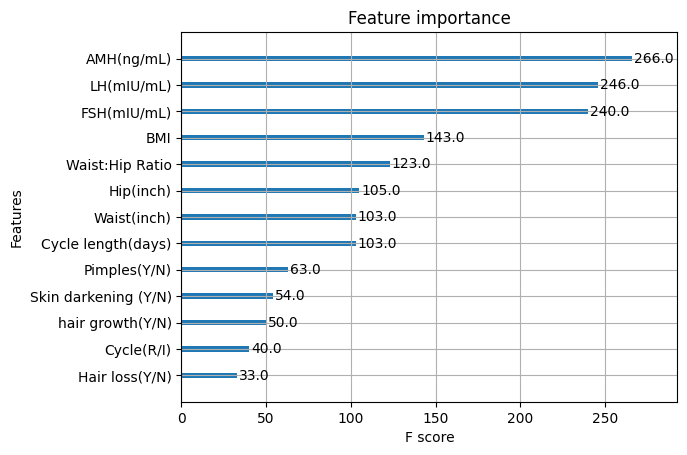


The ROC curve you sent suggests the XGBoost model performs well in classifying Polycystic Ovary Syndrome (PCOS) based on the features provided. The model can distinguish between patients with and without PCOS with a low probability of misclassifying negative cases.The features used and the strong ROC curve, the model has the potential to be a valuable tool for identifying patients at risk of PCOS. It can be used as a screening mechanism to flag potential cases for further evaluation by a doctor using established diagnostic methods. However, it's crucial to remember that the model's prediction should be interpreted alongside a doctor's expertise and other diagnostic tests.



The curve appears to be increasing towards the upper right corner, which is a good sign. This suggests the model is making some correct classifications and avoiding a significant number of false positives.

The part of the curve shown prioritizes recall. The precision-recall curve suggests the model has some ability to identify patients with PCOS. It prioritizes capturing most positive cases, even if it leads to some misclassifications.



The feature importance plot shows that some hormones (FSH, AMH, LH) and body fat distribution metrics (Waist:Hip Ratio) are among the most important factors influencing the ensemble model's predictions for PCOS. While this doesn't explain the biological mechanisms, it highlights features that the model relied on to distinguish between patients with and without PCOS in the data used for training.

**Model Performance:**

We achieved high accuracy on the test data, with 0.9924 for the XGBoost ensemble model and 0.9749 for another ensemble model. This suggests the models can effectively distinguish between patients with and without PCOS in the data used for training.

**Strengths:**

Strong classification ability: Based on the ROC curve, the models appear to be good classifiers, able to separate positive and negative cases.

Focus on relevant features: The feature importance plot highlights hormones (FSH, AMH, LH) and body fat distribution (Waist:Hip Ratio) as important factors, aligning with established medical knowledge of PCOS.

Potential for early detection: By identifying patients at risk based on these features, the models could facilitate earlier intervention and management of PCOS.

**Limitations and Next Steps:**

Overfitting Concern: While the models demonstrate high accuracy on test data, there's a need to address the risk of overfitting. Further validation through techniques like k-fold cross-validation or validation on an independent dataset is essential to ensure real-world applicability.

Role of Medical Expertise: It's crucial to emphasize that the models should complement, rather than replace, existing diagnostic methods. Physician expertise remains paramount in interpreting results and making clinical decisions.

Interpretability Requirements: Enhancing model interpretability through techniques like LIME (Local Interpretable Model-Agnostic Explanations) can foster trust and transparency, facilitating better integration into clinical practice.

The analysis suggests your ensemble models, particularly the XGBoost ensemble, show promise for PCOS detection. However, further validation and addressing limitations are necessary before clinical use. These models, if refined and used appropriately, have the potential to be valuable tools for enhancing PCOS detection and improving patient care. In conclusion, machine learning with feature importance analysis offers promising insights for PCOS detection. It can complement existing methods, potentially leading to earlier diagnosis and personalized treatment for patients.