Group Project

library(tidyverse)

Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'stringr' was built under R version 4.3.2

Warning: package 'lubridate' was built under R version 4.3.2

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(here)

Warning: package 'here' was built under R version 4.3.3

here() starts at C:/Users/Leonel/Desktop/MSDA/MSDA/MS Data Analytics/Current Class/DA 6213/final project

# Read the dataset  
df <- read.csv("student\_depression\_dataset.csv")  
print(head(df))

id Gender Age City Profession Academic.Pressure Work.Pressure CGPA  
1 2 Male 33 Visakhapatnam Student 5 0 8.97  
2 8 Female 24 Bangalore Student 2 0 5.90  
3 26 Male 31 Srinagar Student 3 0 7.03  
4 30 Female 28 Varanasi Student 3 0 5.59  
5 32 Female 25 Jaipur Student 4 0 8.13  
6 33 Male 29 Pune Student 2 0 5.70  
 Study.Satisfaction Job.Satisfaction Sleep.Duration Dietary.Habits  
1 2 0 '5-6 hours' Healthy  
2 5 0 '5-6 hours' Moderate  
3 5 0 'Less than 5 hours' Healthy  
4 2 0 '7-8 hours' Moderate  
5 3 0 '5-6 hours' Moderate  
6 3 0 'Less than 5 hours' Healthy  
 Degree Have.you.ever.had.suicidal.thoughts.. Work.Study.Hours  
1 B.Pharm Yes 3  
2 BSc No 3  
3 BA No 9  
4 BCA Yes 4  
5 M.Tech Yes 1  
6 PhD No 4  
 Financial.Stress Family.History.of.Mental.Illness Depression  
1 1.0 No 1  
2 2.0 Yes 0  
3 1.0 Yes 0  
4 5.0 Yes 1  
5 1.0 No 0  
6 1.0 No 0

print(str(df))

'data.frame': 27901 obs. of 18 variables:  
 $ id : int 2 8 26 30 32 33 52 56 59 62 ...  
 $ Gender : chr "Male" "Female" "Male" "Female" ...  
 $ Age : num 33 24 31 28 25 29 30 30 28 31 ...  
 $ City : chr "Visakhapatnam" "Bangalore" "Srinagar" "Varanasi" ...  
 $ Profession : chr "Student" "Student" "Student" "Student" ...  
 $ Academic.Pressure : num 5 2 3 3 4 2 3 2 3 2 ...  
 $ Work.Pressure : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ CGPA : num 8.97 5.9 7.03 5.59 8.13 5.7 9.54 8.04 9.79 8.38 ...  
 $ Study.Satisfaction : num 2 5 5 2 3 3 4 4 1 3 ...  
 $ Job.Satisfaction : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ Sleep.Duration : chr "'5-6 hours'" "'5-6 hours'" "'Less than 5 hours'" "'7-8 hours'" ...  
 $ Dietary.Habits : chr "Healthy" "Moderate" "Healthy" "Moderate" ...  
 $ Degree : chr "B.Pharm" "BSc" "BA" "BCA" ...  
 $ Have.you.ever.had.suicidal.thoughts..: chr "Yes" "No" "No" "Yes" ...  
 $ Work.Study.Hours : num 3 3 9 4 1 4 1 0 12 2 ...  
 $ Financial.Stress : chr "1.0" "2.0" "1.0" "5.0" ...  
 $ Family.History.of.Mental.Illness : chr "No" "Yes" "Yes" "Yes" ...  
 $ Depression : int 1 0 0 1 0 0 0 0 1 1 ...  
NULL

# Remove any rows with NA values  
df\_clean <- na.omit(df)  
  
# Fit the model on clean data  
logistic\_model <- glm(Depression ~ Gender + Age + Academic.Pressure + CGPA +   
 Study.Satisfaction + Work.Study.Hours + Financial.Stress +   
 `Have.you.ever.had.suicidal.thoughts..` + Family.History.of.Mental.Illness,  
 data = df\_clean, family = binomial)  
  
# Get predictions  
predicted\_probs <- predict(logistic\_model, type = 'response')  
predicted\_classes <- ifelse(predicted\_probs > 0.5, 1, 0)  
  
# Create confusion matrix  
conf\_matrix <- table(Predicted = predicted\_classes, Actual = df\_clean$Depression)  
print("Confusion Matrix:")

[1] "Confusion Matrix:"

print(conf\_matrix)

Actual  
Predicted 0 1  
 0 9081 1900  
 1 2484 14436

# Calculate accuracy, sensitivity, and specificity  
accuracy <- sum(diag(conf\_matrix))/sum(conf\_matrix)  
sensitivity <- conf\_matrix[2,2]/(conf\_matrix[2,2] + conf\_matrix[1,2])  
specificity <- conf\_matrix[1,1]/(conf\_matrix[1,1] + conf\_matrix[2,1])  
  
print(paste("Accuracy:", round(accuracy, 3)))

[1] "Accuracy: 0.843"

print(paste("Sensitivity:", round(sensitivity, 3)))

[1] "Sensitivity: 0.884"

print(paste("Specificity:", round(specificity, 3)))

[1] "Specificity: 0.785"

# ROC curve  
library(pROC)

Warning: package 'pROC' was built under R version 4.3.3

Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'

The following objects are masked from 'package:stats':  
  
 cov, smooth, var

roc\_curve <- roc(df\_clean$Depression, predicted\_probs)

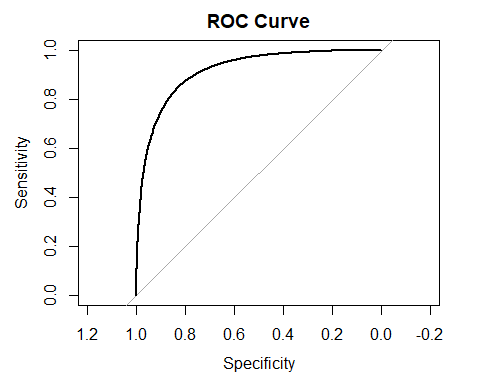
Setting levels: control = 0, case = 1

Setting direction: controls < cases

auc\_value <- auc(roc\_curve)  
print(paste("AUC:", round(auc\_value, 3)))

[1] "AUC: 0.915"

# Plot ROC curve  
plot(roc\_curve, main="ROC Curve")



# Calculate odds ratios by exponentiating the logistic regression coefficients   
odds\_ratios <- exp(coef(logistic\_model))   
print("Odds Ratios:")

[1] "Odds Ratios:"

print(odds\_ratios)

(Intercept)   
 0.4140814   
 GenderMale   
 1.0604371   
 Age   
 0.8958115   
 Academic.Pressure   
 2.3090100   
 CGPA   
 1.0615796   
 Study.Satisfaction   
 0.7858413   
 Work.Study.Hours   
 1.1227590   
 Financial.Stress1.0   
 0.1727618   
 Financial.Stress2.0   
 0.2555127   
 Financial.Stress3.0   
 0.4999881   
 Financial.Stress4.0   
 0.7922738   
 Financial.Stress5.0   
 1.5657520   
Have.you.ever.had.suicidal.thoughts..Yes   
 12.3005101   
 Family.History.of.Mental.IllnessYes   
 1.2719489

# Calculate 95% confidence intervals for the odds ratios   
conf\_int <- exp(confint(logistic\_model))

Waiting for profiling to be done...

print("95% Confidence Intervals for Odds Ratios:")

[1] "95% Confidence Intervals for Odds Ratios:"

print(conf\_int)

2.5 % 97.5 %  
(Intercept) 0.01890855 4.4452242  
GenderMale 0.98845991 1.1376599  
Age 0.88923420 0.9024025  
Academic.Pressure 2.24446971 2.3760499  
CGPA 1.03668260 1.0870954  
Study.Satisfaction 0.76564137 0.8065133  
Work.Study.Hours 1.11212349 1.1335354  
Financial.Stress1.0 0.01631793 3.7516510  
Financial.Stress2.0 0.02413702 5.5481963  
Financial.Stress3.0 0.04723902 10.8557104  
Financial.Stress4.0 0.07485769 17.2013944  
Financial.Stress5.0 0.14793378 33.9954227  
Have.you.ever.had.suicidal.thoughts..Yes 11.41701752 13.2605834  
Family.History.of.Mental.IllnessYes 1.18610520 1.3641044

## Odds Ratio Analysis from Logistic Regression

| Predictor | Odds Ratio (OR) | 95% CI Lower | 95% CI Upper | Significance (CI ≠ 1) |
| --- | --- | --- | --- | --- |
| (Intercept) | 0.4141 | 0.0199 | 4.4452 | No |
| Gender (Male vs. Female) | 1.0604 | 0.9885 | 1.1377 | No |
| Age (per year increase) | 0.8958 | 0.8892 | 0.9024 | Yes (protective) |
| Academic Pressure | 2.3090 | 2.2445 | 2.3760 | Yes (risk ↑) |
| CGPA (per unit increase) | 1.0616 | 1.0366 | 1.0870 | Yes (risk ↑) |
| Study Satisfaction | 0.7858 | 0.7656 | 0.8065 | Yes (protective) |
| Work‐Study Hours (per hour increase) | 1.1228 | 1.1121 | 1.1335 | Yes (risk ↑) |
| Financial Stress = 1 | 0.1728 | 0.0163 | 3.7516 | No |
| Financial Stress = 2 | 0.2555 | 0.0241 | 5.5482 | No |
| Financial Stress = 3 | 0.4999 | 0.0472 | 10.8570 | No |
| Financial Stress = 4 | 0.7923 | 0.0749 | 17.2010 | No |
| Financial Stress = 5 | 1.5658 | 0.1479 | 33.9950 | No |
| Ever had suicidal thoughts (Yes vs. No) | 12.3005 | 11.4172 | 13.2679 | Yes (risk ↑) |
| Family History of Mental Illness (Yes vs. No) | 1.2719 | 1.1861 | 1.3641 | Yes (risk ↑) |

### Interpretations

* **Odds Ratio (OR):**
  + Values greater than 1 indicate increased odds of the outcome for that predictor.
  + Values less than 1 indicate decreased odds (protective effect).
* **Significant Predictors:**
  + **Risk Factors (OR > 1 and statistically significant):**
    - **Academic Pressure (OR = 2.31):** Higher academic pressure increases the odds of the outcome.
    - **CGPA (OR = 1.06):** Increases in CGPA are associated with slightly higher odds of the outcome.
    - **Work‐Study Hours (OR = 1.12):** More work-study hours increase the odds.
    - **Ever had suicidal thoughts (OR = 12.30):** Strongly increases the odds.
    - **Family History of Mental Illness (OR = 1.27):** Increases the odds.
  + **Protective Factors (OR < 1 and statistically significant):**
    - **Age (OR = 0.90):** Older age decreases the odds.
    - **Study Satisfaction (OR = 0.79):** Higher satisfaction is associated with lower odds.
* **Non-significant Predictors:**
  + **Gender (Male vs. Female) and Financial Stress levels (1 to 5):** Their 95% confidence intervals include 1, implying no statistically significant effect.

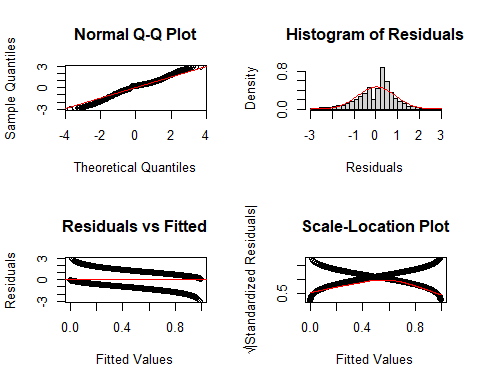
# Take a random sample of 5000 observations for normality testing  
set.seed(123)  
sample\_size <- 5000  
sample\_indices <- sample(length(residuals(logistic\_model)), sample\_size)  
  
# Extract residuals and fitted values for the sample  
residuals\_sample <- residuals(logistic\_model, type="deviance")[sample\_indices]  
fitted\_sample <- fitted(logistic\_model)[sample\_indices]  
  
# Normality Tests on sample  
shapiro\_test <- shapiro.test(residuals\_sample)  
print("Shapiro-Wilk Test Results (on 5000 sample):")

[1] "Shapiro-Wilk Test Results (on 5000 sample):"

print(shapiro\_test)

Shapiro-Wilk normality test  
  
data: residuals\_sample  
W = 0.97803, p-value < 2.2e-16

# Visual tests for normality  
par(mfrow=c(2,2))  
  
# Q-Q Plot  
qqnorm(residuals\_sample, main="Normal Q-Q Plot")  
qqline(residuals\_sample, col="red")  
  
# Histogram  
hist(residuals\_sample,   
 main="Histogram of Residuals",   
 xlab="Residuals",   
 freq=FALSE,  
 breaks=30)  
curve(dnorm(x, mean=mean(residuals\_sample), sd=sd(residuals\_sample)),   
 add=TRUE, col="red")  
  
# Residuals vs Fitted plot  
plot(fitted\_sample, residuals\_sample,  
 xlab="Fitted Values",   
 ylab="Residuals",  
 main="Residuals vs Fitted")  
abline(h=0, col="red")  
  
# Scale-Location Plot  
sqrt\_abs\_res <- sqrt(abs(residuals\_sample))  
plot(fitted\_sample, sqrt\_abs\_res,  
 xlab="Fitted Values",  
 ylab="√|Standardized Residuals|",  
 main="Scale-Location Plot")  
lines(lowess(fitted\_sample, sqrt\_abs\_res), col="red")



# Test for homoscedasticity using Breusch-Pagan test  
library(lmtest)

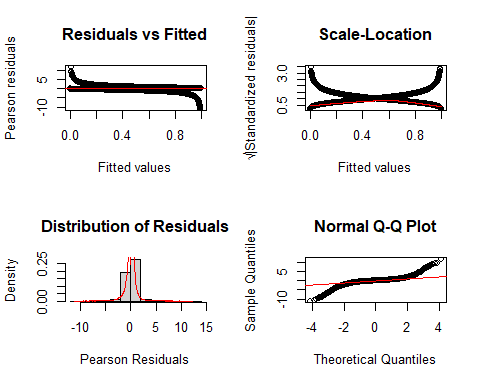
Warning: package 'lmtest' was built under R version 4.3.3

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

residuals\_squared <- residuals(logistic\_model, type="pearson")^2  
bp\_model <- lm(residuals\_squared ~ fitted(logistic\_model))  
bp\_test <- bptest(bp\_model)  
  
# Create diagnostic plots  
par(mfrow=c(2,2))  
  
# 1. Residuals vs Fitted  
plot(fitted(logistic\_model), residuals(logistic\_model, type="pearson"),  
 xlab="Fitted values",  
 ylab="Pearson residuals",  
 main="Residuals vs Fitted")  
abline(h=0, col="red")  
  
# 2. Scale-Location Plot  
sqrt\_abs\_res <- sqrt(abs(residuals(logistic\_model, type="pearson")))  
plot(fitted(logistic\_model), sqrt\_abs\_res,  
 xlab="Fitted values",  
 ylab="√|Standardized residuals|",  
 main="Scale-Location")  
lines(lowess(fitted(logistic\_model), sqrt\_abs\_res), col="red")  
  
# 3. Residuals Distribution  
hist(residuals(logistic\_model, type="pearson"),  
 main="Distribution of Residuals",  
 xlab="Pearson Residuals",  
 freq=FALSE)  
lines(density(residuals(logistic\_model, type="pearson")), col="red")  
  
# 4. QQ Plot  
qqnorm(residuals(logistic\_model, type="pearson"))  
qqline(residuals(logistic\_model, type="pearson"), col="red")



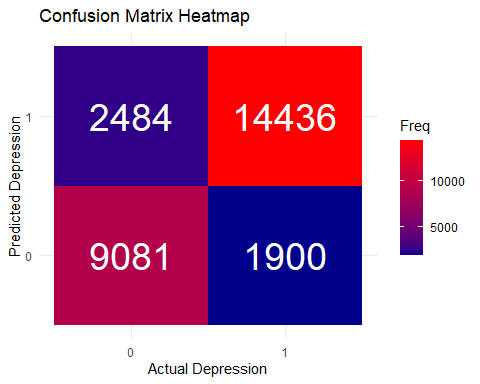
# Print Breusch-Pagan test results  
print("Breusch-Pagan Test for Homoscedasticity:")

[1] "Breusch-Pagan Test for Homoscedasticity:"

print(bp\_test)

studentized Breusch-Pagan test  
  
data: bp\_model  
BP = 0.3459, df = 1, p-value = 0.5564

# Create confusion matrix visualization  
library(ggplot2)  
  
# Get predicted probabilities and classes  
pred\_prob <- predict(logistic\_model, type = "response")  
pred\_class <- ifelse(pred\_prob > 0.5, 1, 0)  
conf\_matrix <- table(Predicted = pred\_class, Actual = df\_clean$Depression)  
  
# Convert confusion matrix to data frame for plotting  
conf\_df <- as.data.frame(conf\_matrix)  
colnames(conf\_df) <- c("Predicted", "Actual", "Freq")  
  
# Plot confusion matrix heatmap  
ggplot(conf\_df, aes(x = Actual, y = Predicted, fill = Freq)) +  
 geom\_tile() +  
 geom\_text(aes(label = Freq), color = "white", size = 10) +  
 scale\_fill\_gradient(low = "darkblue", high = "red") +  
 theme\_minimal() +  
 labs(title = "Confusion Matrix Heatmap",  
 x = "Actual Depression",  
 y = "Predicted Depression")

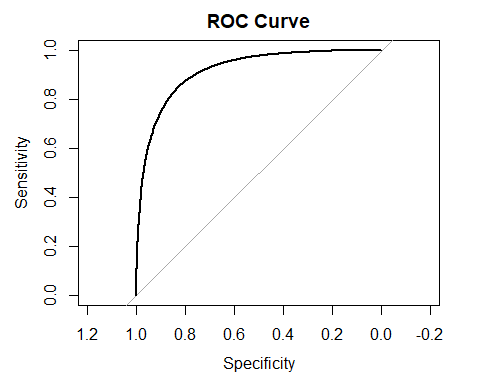


# ROC Curve  
library(pROC)  
roc\_obj <- roc(df\_clean$Depression, pred\_prob)

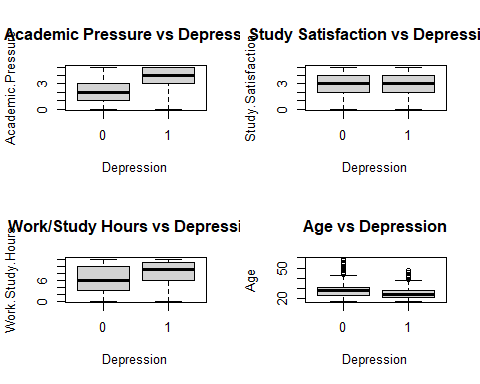
Setting levels: control = 0, case = 1

Setting direction: controls < cases

plot(roc\_obj, main = "ROC Curve")



auc\_value <- auc(roc\_obj)  
  
# Boxplots for key numerical predictors  
par(mfrow = c(2, 2))  
boxplot(Academic.Pressure ~ Depression, data = df\_clean,   
 main = "Academic Pressure vs Depression")  
boxplot(Study.Satisfaction ~ Depression, data = df\_clean,   
 main = "Study Satisfaction vs Depression")  
boxplot(Work.Study.Hours ~ Depression, data = df\_clean,   
 main = "Work/Study Hours vs Depression")  
boxplot(Age ~ Depression, data = df\_clean,   
 main = "Age vs Depression")



## Diagnostic Plot Analysis

### Visual Diagnostics

The diagnostic plots reveal:

* **Normal Q-Q Plot**: Shows some deviation from normality at the tails
* **Residuals vs Fitted**: Shows relatively even spread around zero
* **Scale-Location**: Shows relatively constant spread of standardized residuals
* **Distribution of Residuals**: Shows approximate symmetry but with some deviation from normal distribution

### Key Findings

1. While the residuals show some deviation from normality (as indicated by the Shapiro-Wilk test), this is common in large datasets and doesn’t necessarily invalidate our model.
2. The homoscedasticity assumption appears to be met (supported by the Breusch-Pagan test).
3. The diagnostic plots suggest that while there are some departures from ideal conditions, they’re not severe enough to invalidate the model’s conclusions.

## Impact on Business Models

Based on the diagnostic plots and model assumptions for the depression study, the findings can influence several business models:

* **Healthcare Platforms**:  
  The study’s insights may enable platforms offering mental health services to tailor interventions and predictive analytics, improving user outcomes through early detection and personalized therapy.
* **Employee Wellness Programs**:  
  Organizations investing in employee wellness might leverage these findings to implement mental health initiatives, screen for potential risks, and proactively offer support to improve productivity and wellbeing.
* **Education-Related Mental Health Services**:  
  Educational institutions and related service providers can develop targeted programs for student mental health support, using the model to identify high-risk groups and deploy timely preventive measures.
* **Insurance and Managed Care Models**:  
  Insurers may incorporate these predictive models to design better mental health coverage plans and proactively manage the costs associated with mental health treatments.

These business models can benefit from integrating predictive insights from the study, leading to improved service targeting, cost reduction, and enhanced overall outcomes.