ASDS – 5303

Final Project Report

NASA Exoplanet Classification

Hareen Maddipatla (1002233075)
Shalini Dronadula (1002207495)
Ananya Joseph (1002238425)

THE UNIVERSITY OF TEXAS
AT ARLINGTON

1. INTRODUCTION:

1.1. Problem Statement:

The growing catalogue of exoplanets demands efficient and accurate methods for their classification based on physical and orbital properties. Traditional astrophysical approaches often fail to generalize effectively for the vast diversity in planetary characteristics. With over 5,000 exoplanets discovered to date, characterized by varying masses, radii, distances, and discovery methods, a systematic classification method is crucial for advancing astrophysical research and identifying Earth-like planets. The challenge lies in handling missing data, standardizing diverse features, and determining the most suitable classification models for such a multifaceted dataset.

1.2. Objective:

The primary objective of this study is to classify exoplanets into distinct types—Gas Giant, Neptune-like, Super Earth, and Terrestrial—using advanced data analysis and machine learning techniques. This involves cleaning and preprocessing a dataset of over 5,000 exoplanets, conducting exploratory data analysis to uncover patterns, and evaluating the effectiveness of classification models. The study aims to achieve high accuracy in predictions while ensuring interpretability, contributing to a more structured understanding of planetary systems and their properties.

1.3. Motivation:

The exploration and classification of exoplanets are pivotal for understanding the universe and identifying potential Earth-like planets. Advances in observational technology have exponentially increased the discovery rate of exoplanets, but the sheer volume and complexity of the data require efficient analytical methods. This study is motivated by the opportunity to apply machine learning to enhance classification accuracy and uncover insights into planetary formation and evolution. Moreover, bridging the gap between data-driven techniques and astrophysical research can significantly impact future explorations and scientific discoveries.

1.4. Scope:

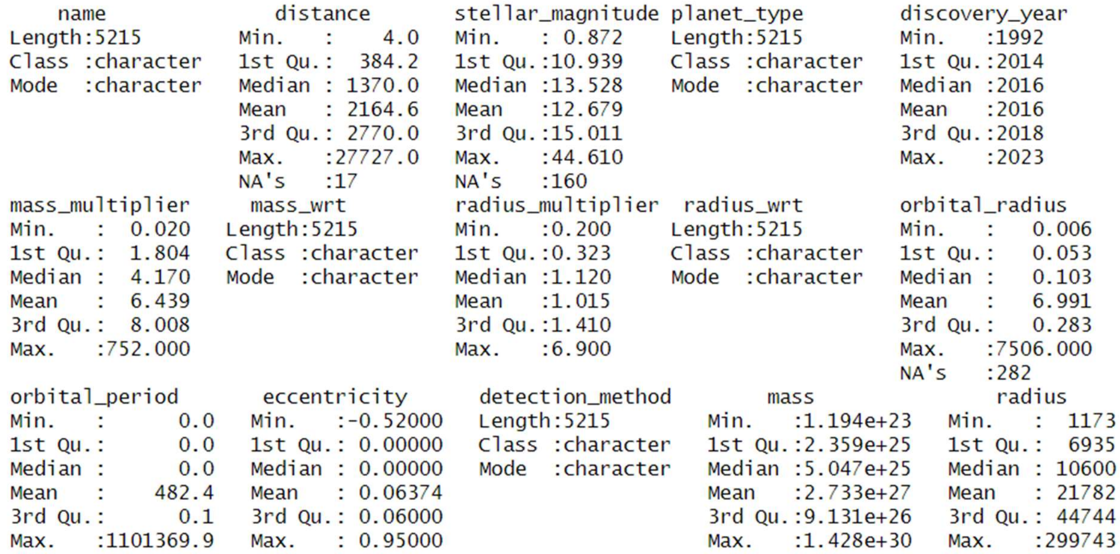
This project focuses on the classification of exoplanets based on their physical (mass and radius) and orbital (distance, orbital radius) characteristics. The scope includes data preprocessing, feature engineering, exploratory data analysis, and the application of machine learning models—Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA)—to classify exoplanets. While the study emphasizes the efficiency and interpretability of these models, it also paves the way for integrating more advanced algorithms and additional planetary attributes in future analyses, making it a foundational step in computational astrophysics.

2. DESCRIPTIVE ANALYSIS:

The dataset analyzed in this project comprises 5,215 rows and 15 columns, each capturing diverse characteristics of exoplanets. These variables include both numeric and non-numeric data. The numeric variables, such as mass, radius, distance, and stellar magnitude, provide quantifiable insights into the physical and orbital properties of the planets, while non-numeric variables like discovery method and planet type categorize and contextualize the findings. The dataset reflects the extraordinary diversity of exoplanets, with key attributes ranging from minuscule terrestrial planets to massive gas giants, spanning distances of just a few light-years to thousands.

The post-2016 surge in the discovery of exoplanets is notably linked to the second phase of NASA's Kepler mission, known as the K2 mission. This phase involved a broader exploration of different regions of space compared to the mission's earlier years. With its enhanced observational scope, the K2 mission not only expanded the volume of planetary data collected but also allowed for discoveries around stars previously beyond its reach. This marked an era of collaborative data analysis, enabling scientists worldwide to access and interpret Kepler's vast trove of exoplanet data. The collaborative nature of the K2 mission, combined with advancements in detection techniques, led to an unprecedented rise in the number of confirmed exoplanets, reflecting a significant milestone in astrophysical research.

```
summary(main_data)
```



```
     name              distance       stellar_magnitude planet_type        discovery_year
 Length:5215        Min.   :    4.0   Min.   : 0.872    Length:5215        Min.   :1992
 Class :character   1st Qu.:  384.2   1st Qu.:10.939    Class :character   1st Qu.:2014
 Mode  :character   Median : 1370.0   Median :13.528    Mode  :character   Median :2016
                    Mean   : 2164.6   Mean   :12.679                       Mean   :2016
                    3rd Qu.: 2770.0   3rd Qu.:15.011                       3rd Qu.:2018
                    Max.   :27727.0   Max.   :44.610                       Max.   :2023
                    NA's   :17        NA's   :160
 mass_multiplier     mass_wrt         radius_multiplier  radius_wrt         orbital_radius
 Min.   :  0.020   Length:5215        Min.   :0.200     Length:5215        Min.   :   0.006
 1st Qu.:  1.804   Class :character   1st Qu.:0.323     Class :character   1st Qu.:   0.053
 Median :  4.170   Mode  :character   Median :1.120     Mode  :character   Median :   0.103
 Mean   :  6.439                      Mean   :1.015                        Mean   :   6.991
 3rd Qu.:  8.008                      3rd Qu.:1.410                        3rd Qu.:   0.283
 Max.   :752.000                      Max.   :6.900                        Max.   :7506.000
                                                                           NA's   :282
 orbital_period        eccentricity      detection_method        mass                radius
 Min.   :      0.0   Min.   :-0.52000   Length:5215        Min.   :1.194e+23   Min.   :  1173
 1st Qu.:      0.0   1st Qu.: 0.00000   Class :character   1st Qu.:2.359e+25   1st Qu.:  6935
 Median :      0.0   Median : 0.00000   Mode  :character   Median :5.047e+25   Median : 10600
 Mean   :    482.4   Mean   : 0.06374                      Mean   :2.733e+27   Mean   : 21782
 3rd Qu.:      0.1   3rd Qu.: 0.06000                      3rd Qu.:9.131e+26   3rd Qu.: 44744
 Max.   :1101369.9   Max.   : 0.95000                      Max.   :1.428e+30   Max.   :299743
```

3. DATA PREPROCESSING:

Data cleaning and preprocessing are critical steps in ensuring that the dataset is dependable and suitable for analysis. For this project, significant efforts were made to clean the data, address quality issues, and preprocess the variables to extract meaningful insights.

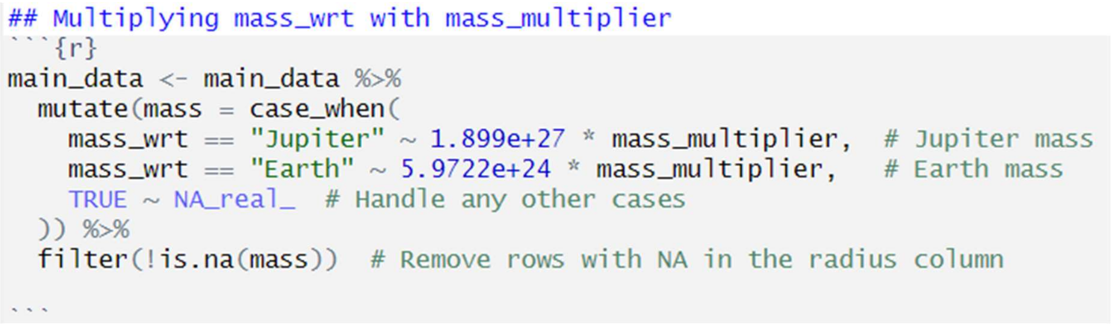
3.1. Handling Missing Values:

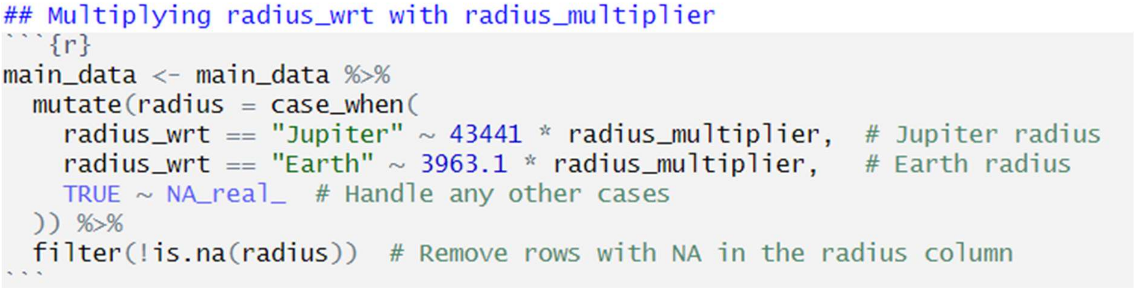
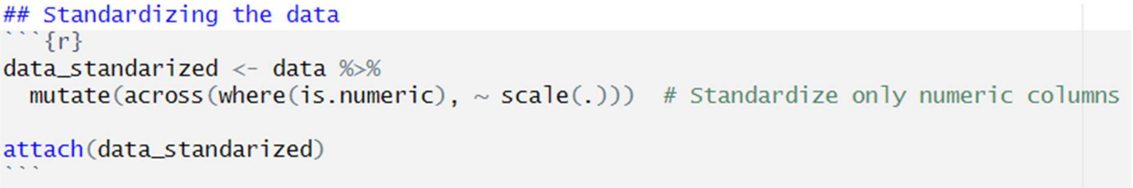
One of the critical preprocessing steps involved managing missing values. Several columns, such as mass, radius, and orbital_radius, contained NA values, which could potentially bias or weaken the model's performance. Rows with missing values in critical variables were removed using the ‘na.omit’ function. This ensured that only complete cases were considered for analysis, minimizing noise, and improving the reliability of subsequent computations.

```
main_data <- na.omit(main_data)
```

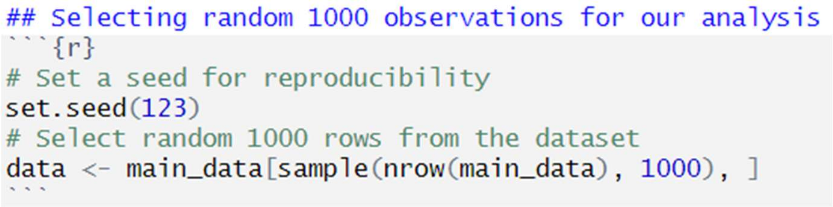
3.2. Feature Engineering:

Feature engineering played a pivotal role in this analysis. The planetary mass and radius were recalculated by multiplying their respective multipliers with constants for Earth and Jupiter, depending on the reference planet. These engineered features better represented the physical characteristics of the planets and formed the basis for meaningful comparisons and predictions.



~~~
## Multiplying mass_wrt with mass_multiplier
```{r}
main_data <- main_data %>%
  mutate(mass = case_when(
    mass_wrt == "Jupiter" ~ 1.899e+27 * mass_multiplier,  # Jupiter mass
    mass_wrt == "Earth" ~ 5.9722e+24 * mass_multiplier,   # Earth mass
    TRUE ~ NA_real_  # Handle any other cases
  )) %>%
  filter(!is.na(mass))  # Remove rows with NA in the radius column

~~~
```



```
## Multiplying radius_wrt with radius_multiplier
```{r}
main_data <- main_data %>%
  mutate(radius = case_when(
    radius_wrt == "Jupiter" ~ 43441 * radius_multiplier,  # Jupiter radius
    radius_wrt == "Earth" ~ 3963.1 * radius_multiplier,   # Earth radius
    TRUE ~ NA_real_  # Handle any other cases
  )) %>%
  filter(!is.na(radius))  # Remove rows with NA in the radius column
```
```

3.3. Data Standardization:

To address the wide range of values across numeric variables, standardization was applied using the ‘scale’ function. This step transformed numeric features to have a mean of zero and a standard deviation of one, ensuring that all features were on a comparable scale. Standardization was essential for machine learning models like Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), which are sensitive to differences in feature magnitude.



```
## Standardizing the data
```{r}
data_standarized <- data %>%
  mutate(across(where(is.numeric), ~ scale(.)))  # Standardize only numeric columns

attach(data_standarized)
```
```

3.4. Random Sampling:

Given the generous size of the cleaned dataset, a random sample of 1,000 observations was selected for efficient and representative analysis. This step ensured computational feasibility while retaining the dataset's diversity. A seed was set to ensure reproducibility, allowing other researchers to replicate the analysis under identical conditions.

These preprocessing steps collectively ensured that the dataset was clean, consistent, and ready for advanced analysis. They addressed challenges such as missing data, feature
```

variability, and computational efficiency, laying a sturdy foundation for exploratory data analysis and machine learning applications.



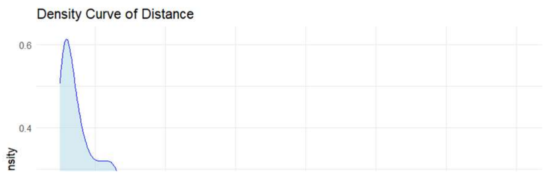
~~~
## Selecting random 1000 observations for our analysis
```{r}
# Set a seed for reproducibility
set.seed(123)
# Select random 1000 rows from the dataset
data <- main_data[sample(nrow(main_data), 1000), ]
```
~~~

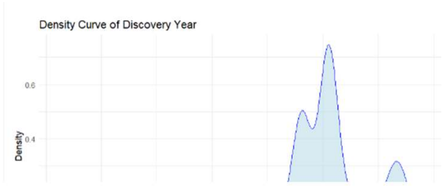
4. EXPLORATORY DATA ANALYSIS:

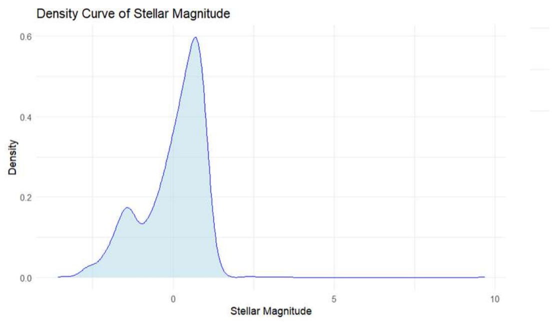
The exploratory data analysis phase was crucial in understanding the dataset's structure, distributions, and relationships between variables. Various statistical and graphical techniques were employed to uncover patterns and insights.

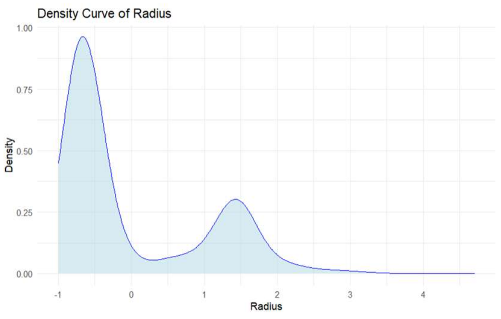
The distributions of key variables such as distance, stellar magnitude, radius, and mass were analyzed using density plots. These plots revealed several patterns:

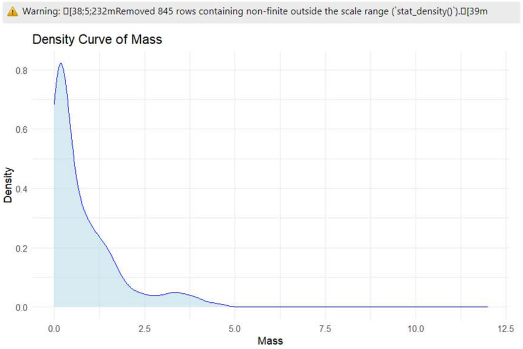
- **Distance:** The density curve showed a concentration of planets located at shorter distances, with the density rapidly decreasing for farther planets. This reflects an observational bias, as closer planets are easier to detect with current technology.
- **Stellar Magnitude and Radius:** The stellar magnitude distribution displayed a concentrated peak, indicating most planets orbit stars with similar brightness. Similarly, the radius distribution exhibited a primary peak dominated by smaller planets and a secondary peak corresponding to larger gas giants.
- **Discovery Year:** A significant surge in discoveries was noted post-2000, particularly after 2016, aligning with advancements in detection methods and the second phase of the Kepler mission (K2). This trend reflects the increasing effectiveness of space telescopes and the growing global collaboration in astrophysical research.





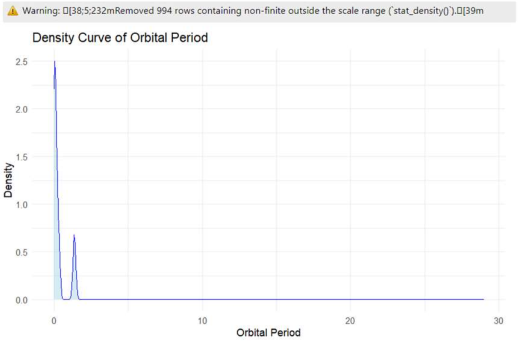






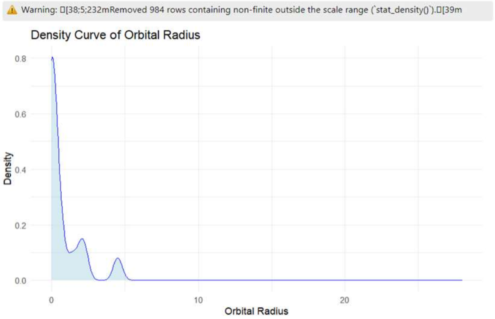
```
⚠ Warning: □[38;5;232mRemoved 845 rows containing non-finite outside the scale range (`stat_density()`).□[39m
```

Density Curve of Mass



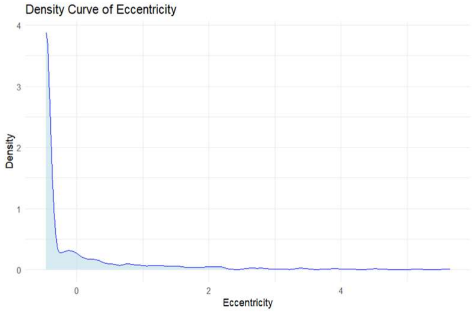
```
⚠ Warning: □[38;5;232mRemoved 994 rows containing non-finite outside the scale range (`stat_density()`).□[39m
```

Density Curve of Orbital Period



```
⚠ Warning: □[38;5;232mRemoved 984 rows containing non-finite outside the scale range (`stat_density()`).□[39m
```

Density Curve of Orbital Radius



Density Curve of Eccentricity

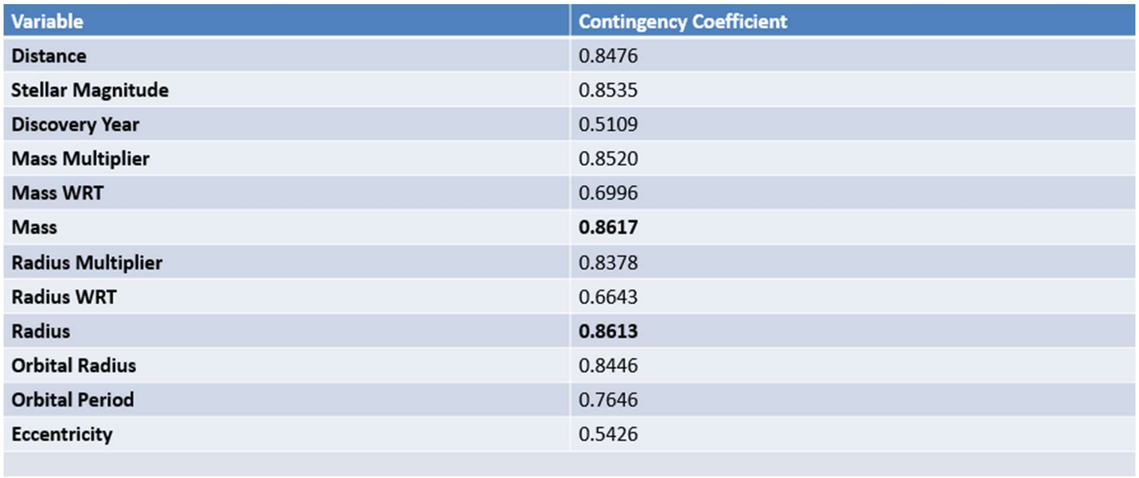
The relationship between mass and radius and other planetary characteristics was explored to identify the most relevant predictors for classification. These analyses showed that larger masses and radii were associated with gas giants, while smaller values corresponded to terrestrial planets. Orbital characteristics, such as eccentricity and orbital radius, also showed clustering around specific planet types but were less distinctive compared to mass and radius.

The decision to use radius and mass as predictors was quantitatively supported by their high contingency coefficients. Contingency coefficients measure the strength of association between categorical variables (planet type) and continuous predictors (mass, radius). Both mass (0.8617) and radius (0.8613) exhibited the highest coefficients among all variables analyzed, indicating a robust relationship with planet type.

This strong association arises because mass and radius are fundamental characteristics that directly determine a planet's classification:

- **Gas Giants:** Large mass and radius values are characteristic of gas giants, reflecting their massive atmospheres and significant gravitational pulls.
- **Terrestrial Planets:** Smaller mass and radius values are linked to terrestrial planets, which are composed primarily of rock and metal.
- **Intermediate Types:** Neptune-like and Super Earth planets exhibit intermediate values, reinforcing the predictive power of mass and radius.

By selecting these two features, the analysis focused on variables with the highest discriminatory power, ensuring the models were both accurate and interpretable. The high contingency coefficients validated the choice, confirming that radius and mass were the most reliable predictors for distinguishing between exoplanet types.



Variable	Contingency Coefficient
Distance	0.8476
Stellar Magnitude	0.8535
Discovery Year	0.5109
Mass Multiplier	0.8520
Mass WRT	0.6996
Mass	**0.8617**
Radius Multiplier	0.8378
Radius WRT	0.6643
Radius	**0.8613**
Orbital Radius	0.8446
Orbital Period	0.7646
Eccentricity	0.5426

5. MODEL

5.1. Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis (LDA) is a classification technique that assumes the data features follow a multivariate normal distribution with identical covariance matrices across classes. This assumption results in linear decision boundaries that separate the classes. In the context of the exoplanet dataset, LDA was chosen because it effectively models the relationships between key features, such as mass and radius, and planet types. These features showed strong linear associations with the classification labels, making them ideal for LDA.

Another significant advantage of LDA is its simplicity and interpretability. The linear decision boundaries it generates are easy to understand and visualize, which is valuable for explaining classification results in scientific contexts. Additionally, LDA is computationally efficient, making it well-suited for datasets with a moderate number of features and observations, like the exoplanet dataset.

5.2. Quadratic Discriminant Analysis (QDA):

Quadratic Discriminant Analysis (QDA) extends LDA by allowing each class to have its own covariance matrix, thereby relaxing the assumption of identical covariance across classes. This results in quadratic decision boundaries that can better capture non-linear relationships. In the exoplanet dataset, this flexibility was crucial as some planet types exhibited more complex distributions of mass and radius that could not be adequately modelled with linear boundaries.

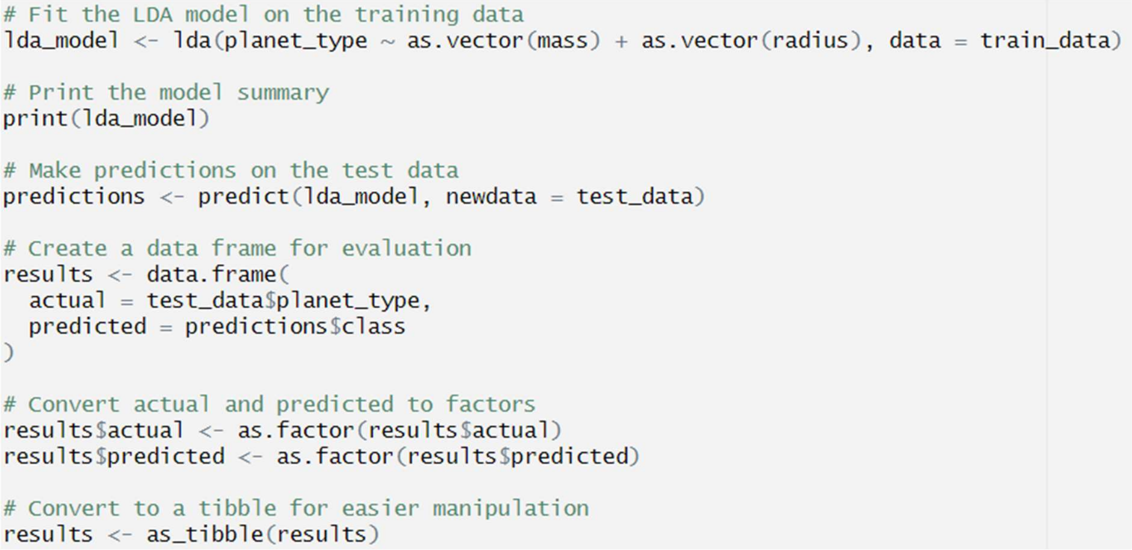
QDA's strength lies in its ability to handle non-linear separations in the feature space. This makes it particularly valuable for datasets where the relationships between features and classes are not strictly linear. However, this flexibility comes with additional computational complexity and a higher risk of overfitting, especially in smaller datasets or those with noise. For this analysis, careful preprocessing and standardization mitigated these risks, allowing QDA to fully utilize its potential to model non-linear relationships.

5.3. Results:

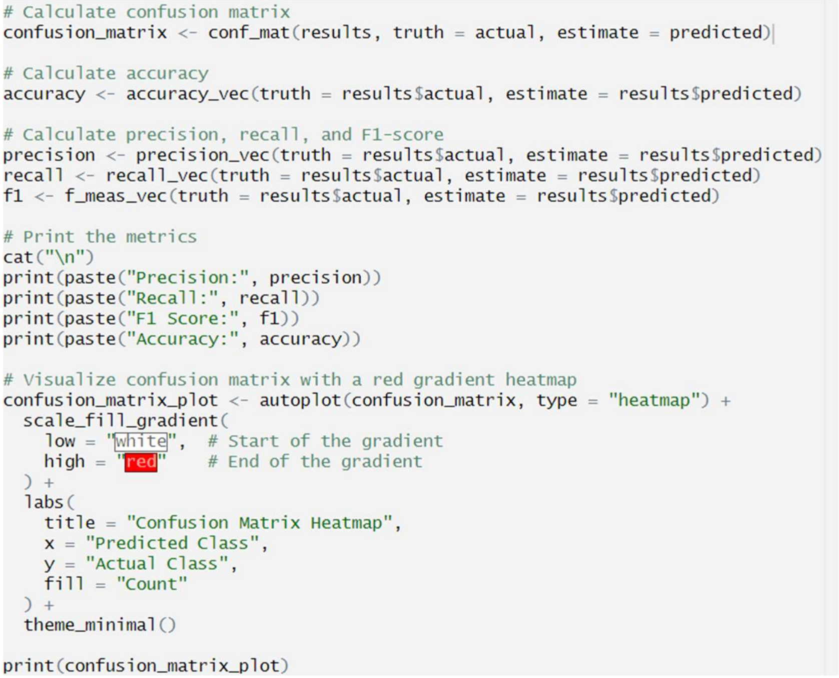
5.3.1. Linear Discriminant Analysis (LDA):

The Linear Discriminant Analysis (LDA) model achieved an accuracy of 87.63%, demonstrating its strong performance in classifying exoplanets based on mass and radius. The model effectively distinguished between planet types, leveraging the linear relationships between the features and the classification labels.

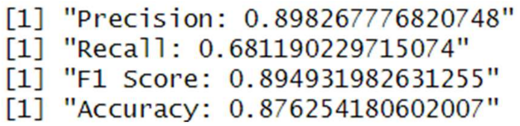
The decision boundaries produced by LDA were linear, reflecting the model's assumption of identical covariance matrices across classes. These boundaries clearly separated the exoplanet categories, providing a straightforward interpretation of how mass and radius contribute to planetary classification. The simplicity of the decision boundaries makes LDA particularly valuable for scientific applications, where understanding the model's logic is essential.



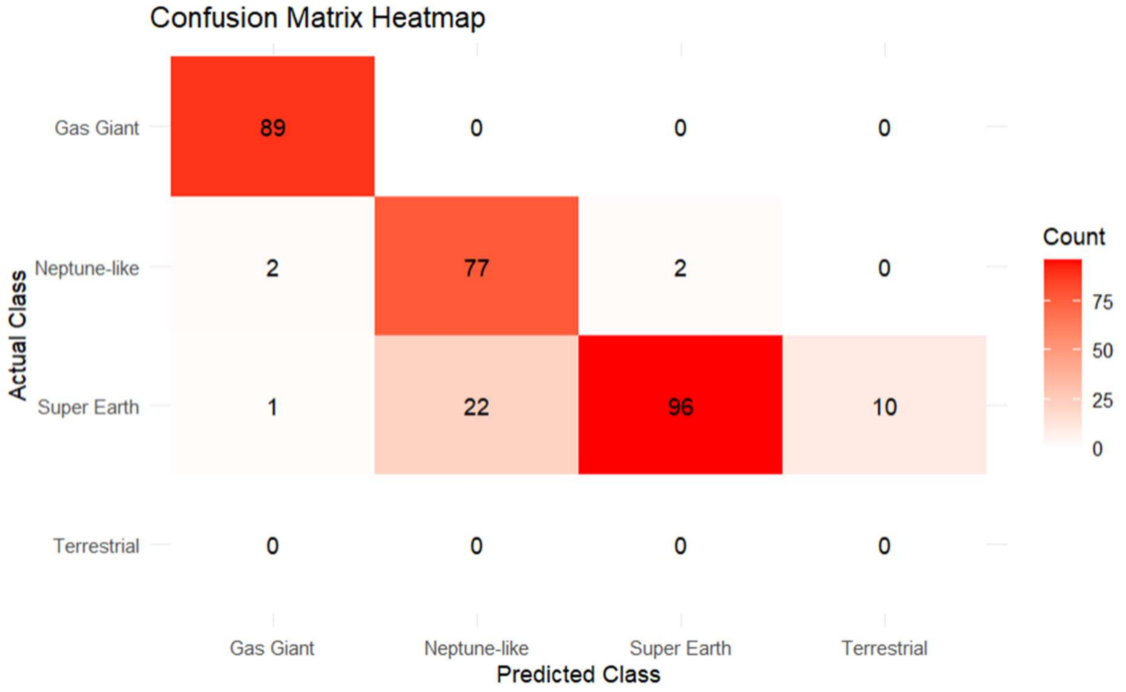
```
# Fit the LDA model on the training data
lda_model <- lda(planet_type ~ as.vector(mass) + as.vector(radius), data = train_data)

# Print the model summary
print(lda_model)

# Make predictions on the test data
predictions <- predict(lda_model, newdata = test_data)

# Create a data frame for evaluation
results <- data.frame(
  actual = test_data$planet_type,
  predicted = predictions$class
)

# Convert actual and predicted to factors
results$actual <- as.factor(results$actual)
results$predicted <- as.factor(results$predicted)

# Convert to a tibble for easier manipulation
results <- as_tibble(results)
```

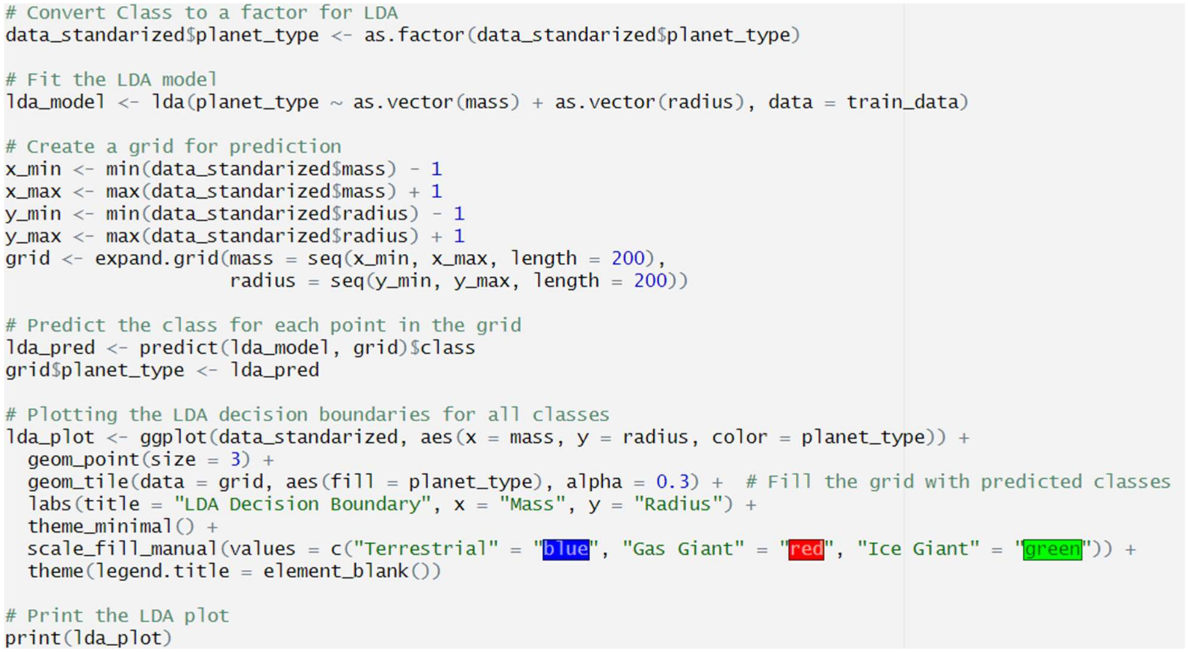


```
# Calculate confusion matrix
confusion_matrix <- conf_mat(results, truth = actual, estimate = predicted)

# Calculate accuracy
accuracy <- accuracy_vec(truth = results$actual, estimate = results$predicted)

# Calculate precision, recall, and F1-score
precision <- precision_vec(truth = results$actual, estimate = results$predicted)
recall <- recall_vec(truth = results$actual, estimate = results$predicted)
f1 <- f_meas_vec(truth = results$actual, estimate = results$predicted)

# Print the metrics
cat("\n")
print(paste("Precision:", precision))
print(paste("Recall:", recall))
print(paste("F1 Score:", f1))
print(paste("Accuracy:", accuracy))

# Visualize confusion matrix with a red gradient heatmap
confusion_matrix_plot <- autoplot(confusion_matrix, type = "heatmap") +
  scale_fill_gradient(
    low = "white",  # Start of the gradient
    high = "red"    # End of the gradient
  ) +
  labs(
    title = "Confusion Matrix Heatmap",
    x = "Predicted Class",
    y = "Actual Class",
    fill = "Count"
  ) +
  theme_minimal()

print(confusion_matrix_plot)
```

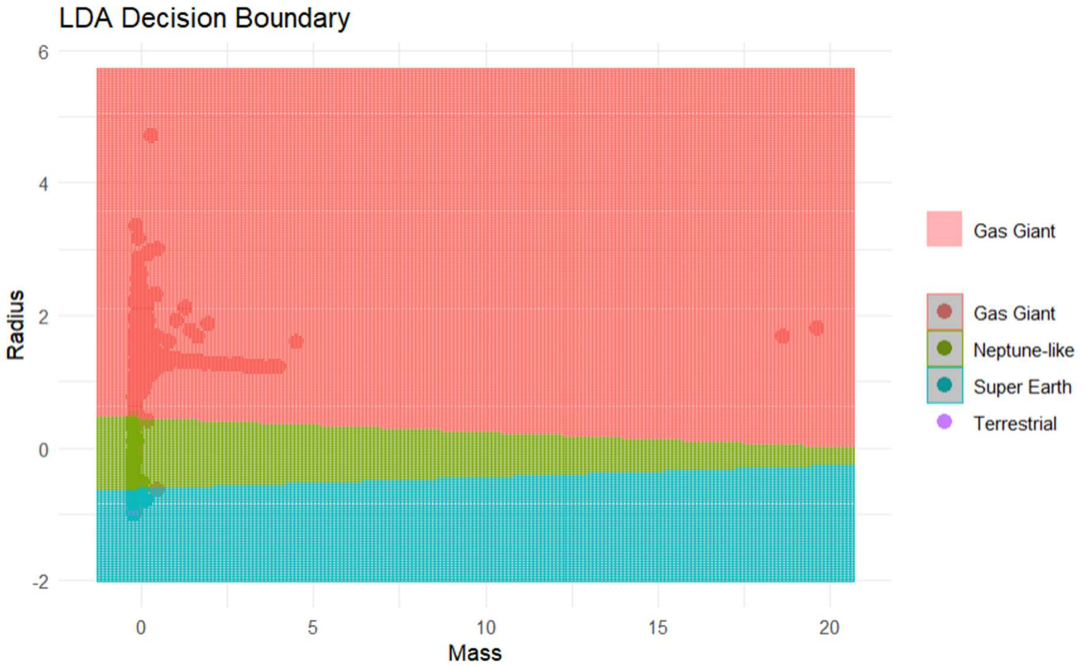


```
[1] "Precision: 0.898267776820748"
[1] "Recall: 0.681190229715074"
[1] "F1 Score: 0.894931982631255"
[1] "Accuracy: 0.876254180602007"
```





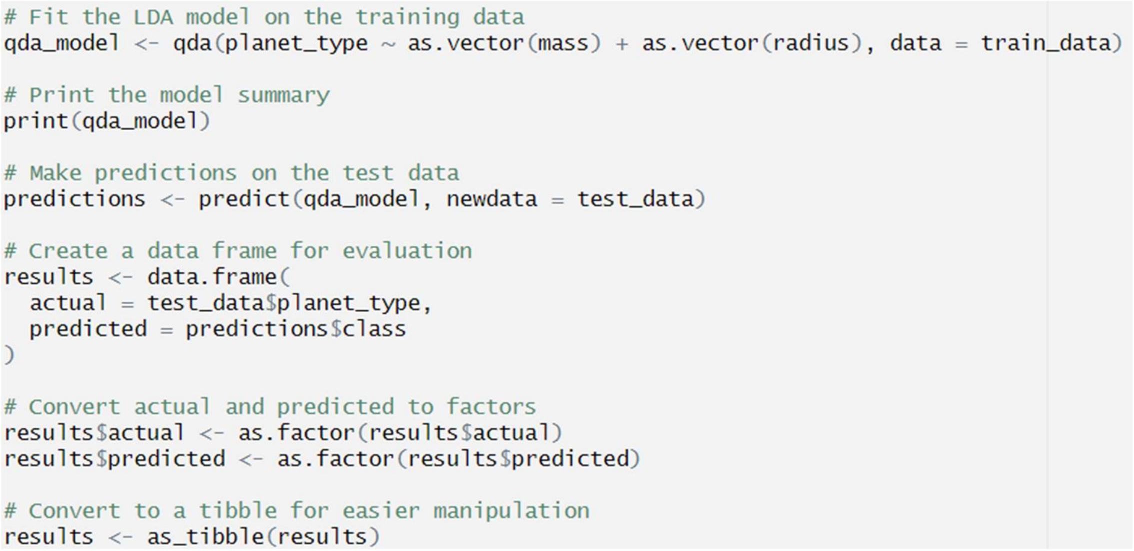
```
# Convert Class to a factor for LDA
data_standarized$planet_type <- as.factor(data_standarized$planet_type)

# Fit the LDA model
lda_model <- lda(planet_type ~ as.vector(mass) + as.vector(radius), data = train_data)

# Create a grid for prediction
x_min <- min(data_standarized$mass) - 1
x_max <- max(data_standarized$mass) + 1
y_min <- min(data_standarized$radius) - 1
y_max <- max(data_standarized$radius) + 1
grid <- expand.grid(mass = seq(x_min, x_max, length = 200),
                    radius = seq(y_min, y_max, length = 200))

# Predict the class for each point in the grid
lda_pred <- predict(lda_model, grid)$class
grid$planet_type <- lda_pred

# Plotting the LDA decision boundaries for all classes
lda_plot <- ggplot(data_standarized, aes(x = mass, y = radius, color = planet_type)) +
  geom_point(size = 3) +
  geom_tile(data = grid, aes(fill = planet_type), alpha = 0.3) +  # Fill the grid with predicted classes
  labs(title = "LDA Decision Boundary", x = "Mass", y = "Radius") +
  theme_minimal() +
  scale_fill_manual(values = c("Terrestrial" = "blue", "Gas Giant" = "red", "Ice Giant" = "green")) +
  theme(legend.title = element_blank())

# Print the LDA plot
print(lda_plot)
```



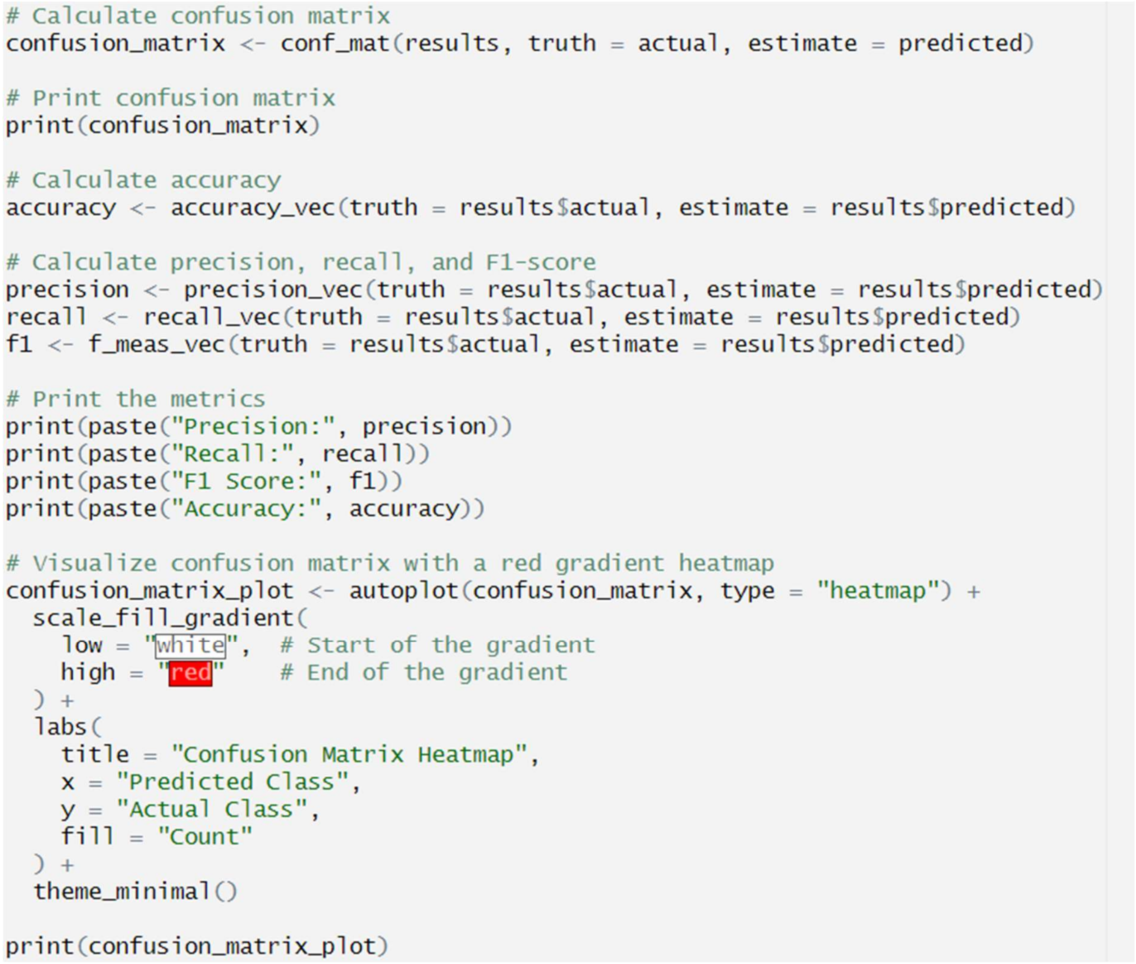
5.3.2. Quadratic Discriminant Analysis (QDA):

The Quadratic Discriminant Analysis (QDA) model achieved an impressive accuracy of 95.32%, significantly outperforming LDA in this analysis. This high accuracy highlights QDA's ability to effectively classify exoplanets by capturing the non-linear relationships between the features of mass and radius and the planet types.

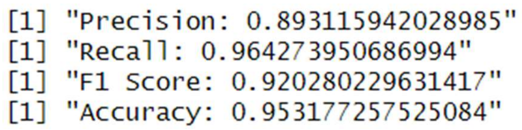
Unlike LDA, QDA produced quadratic decision boundaries, allowing it to adapt to more complex distributions in the dataset. These nuanced boundaries provided a better fit for the diverse characteristics of the exoplanet data, such as variations in mass and radius across different planet types. This flexibility enabled QDA to accurately distinguish between categories that LDA's linear boundaries may have struggled to separate.



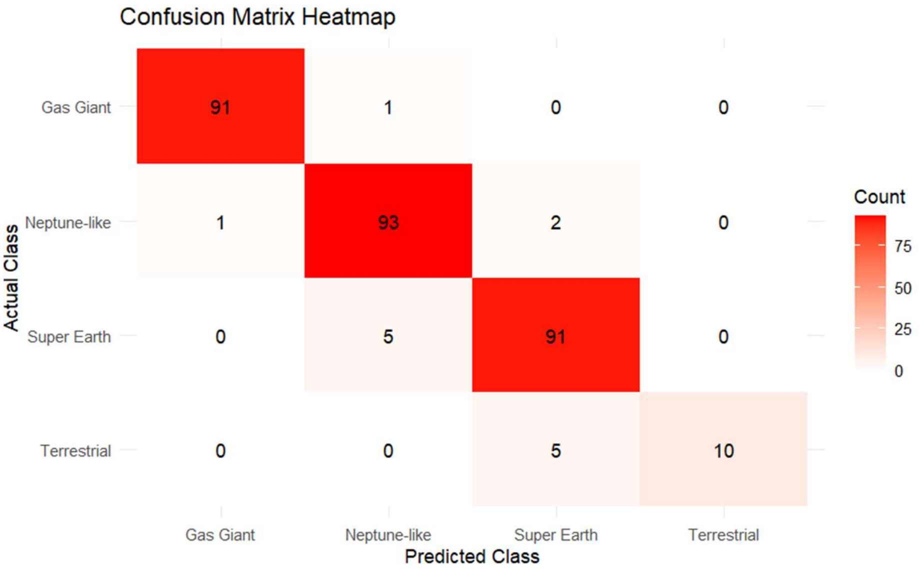
```
# Fit the LDA model on the training data
qda_model <- qda(planet_type ~ as.vector(mass) + as.vector(radius), data = train_data)

# Print the model summary
print(qda_model)

# Make predictions on the test data
predictions <- predict(qda_model, newdata = test_data)

# Create a data frame for evaluation
results <- data.frame(
  actual = test_data$planet_type,
  predicted = predictions$class
)

# Convert actual and predicted to factors
results$actual <- as.factor(results$actual)
results$predicted <- as.factor(results$predicted)

# Convert to a tibble for easier manipulation
results <- as_tibble(results)
```



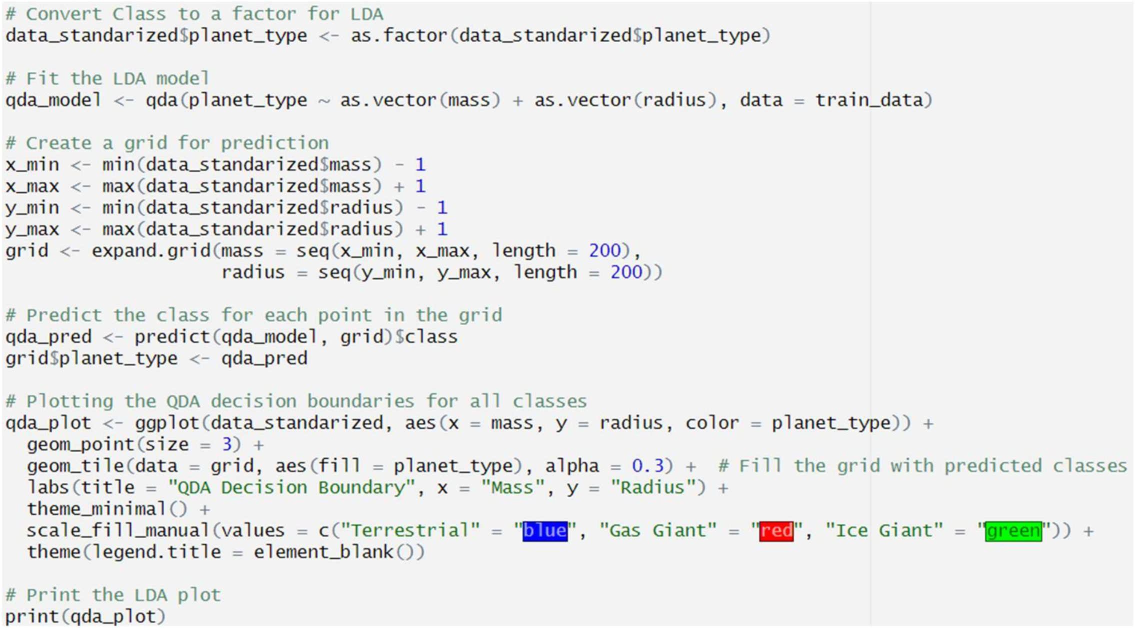
```
# Calculate confusion matrix
confusion_matrix <- conf_mat(results, truth = actual, estimate = predicted)

# Print confusion matrix
print(confusion_matrix)

# Calculate accuracy
accuracy <- accuracy_vec(truth = results$actual, estimate = results$predicted)

# Calculate precision, recall, and F1-score
precision <- precision_vec(truth = results$actual, estimate = results$predicted)
recall <- recall_vec(truth = results$actual, estimate = results$predicted)
f1 <- f_meas_vec(truth = results$actual, estimate = results$predicted)

# Print the metrics
print(paste("Precision:", precision))
print(paste("Recall:", recall))
print(paste("F1 Score:", f1))
print(paste("Accuracy:", accuracy))

# Visualize confusion matrix with a red gradient heatmap
confusion_matrix_plot <- autoplot(confusion_matrix, type = "heatmap") +
  scale_fill_gradient(
    low = "white",  # Start of the gradient
    high = "red"    # End of the gradient
  ) +
  labs(
    title = "Confusion Matrix Heatmap",
    x = "Predicted Class",
    y = "Actual Class",
    fill = "Count"
  ) +
  theme_minimal()

print(confusion_matrix_plot)
```



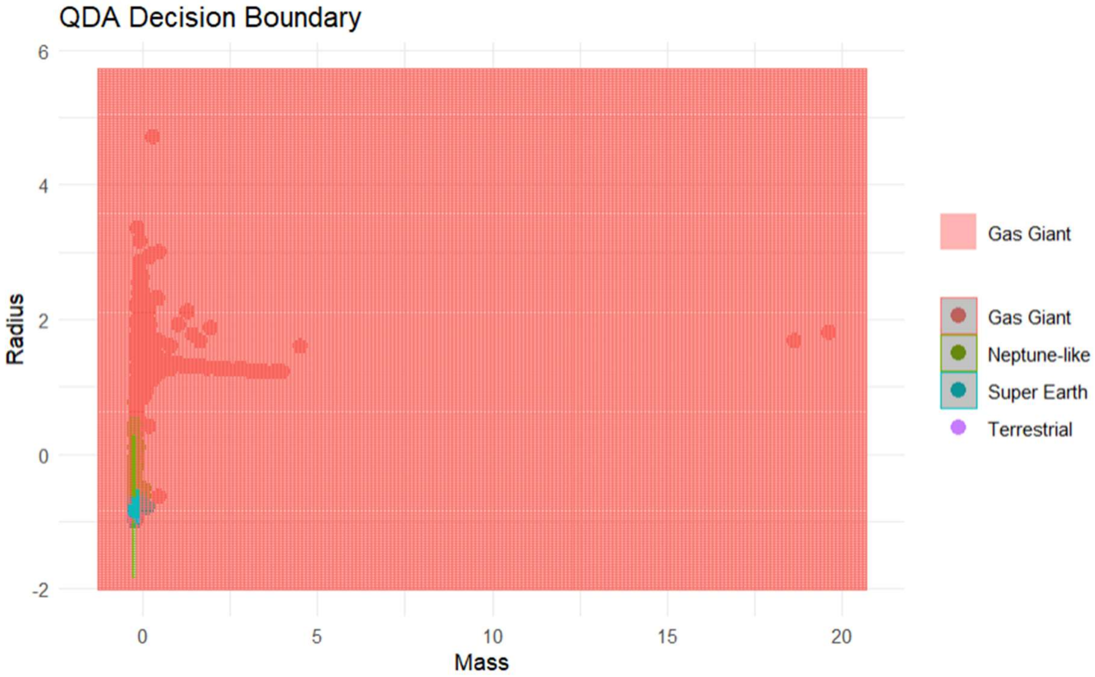
```
[1] "Precision: 0.893115942028985"
[1] "Recall: 0.964273950686994"
[1] "F1 Score: 0.920280229631417"
[1] "Accuracy: 0.953177257525084"
```



	Gas Giant	Neptune-like	Super Earth	Terrestrial
Gas Giant	91	1	0	0
Neptune-like	1	93	2	0
Super Earth	0	5	91	0
Terrestrial	0	0	5	10



```
# Convert Class to a factor for LDA
data_standarized$planet_type <- as.factor(data_standarized$planet_type)

# Fit the LDA model
qda_model <- qda(planet_type ~ as.vector(mass) + as.vector(radius), data = train_data)

# Create a grid for prediction
x_min <- min(data_standarized$mass) - 1
x_max <- max(data_standarized$mass) + 1
y_min <- min(data_standarized$radius) - 1
y_max <- max(data_standarized$radius) + 1
grid <- expand.grid(mass = seq(x_min, x_max, length = 200),
                    radius = seq(y_min, y_max, length = 200))

# Predict the class for each point in the grid
qda_pred <- predict(qda_model, grid)$class
grid$planet_type <- qda_pred

# Plotting the QDA decision boundaries for all classes
qda_plot <- ggplot(data_standarized, aes(x = mass, y = radius, color = planet_type)) +
  geom_point(size = 3) +
  geom_tile(data = grid, aes(fill = planet_type), alpha = 0.3) +  # Fill the grid with predicted classes
  labs(title = "QDA Decision Boundary", x = "Mass", y = "Radius") +
  theme_minimal() +
  scale_fill_manual(values = c("Terrestrial" = "blue", "Gas Giant" = "red", "Ice Giant" = "green")) +
  theme(legend.title = element_blank())

# Print the LDA plot
print(qda_plot)
```



6. CONCLUSION:

In this project, we compared the performance of Quadratic Discriminant Analysis (QDA) and Linear Discriminant Analysis (LDA) for classifying planetary types based on their physical characteristics, specifically mass and radius. The results reveal that QDA outperformed LDA due to its ability to handle non-linear separations in the data. QDA achieved a higher accuracy of 95.3% compared to LDA's 87.6%, as well as superior recall and F1 score, making it the better model for this dataset.

The confusion matrix further highlights that QDA had fewer misclassifications across all classes, while LDA struggled particularly with distinguishing Super Earths from Neptune-like planets due to its linear decision boundaries. These findings indicate that the dataset exhibits non-linear separability, favouring QDA's quadratic boundaries over LDA's linear ones. While

LDA showed comparable precision, its lower recall demonstrates its limitations with overlapping feature spaces.

Based on these results, QDA is recommended as the more suitable model for this dataset. Future work could involve exploring additional classifiers, such as Support Vector Machines or Random Forests, to benchmark performance further. This study underscores the importance of selecting models that align with the data's underlying structure to achieve optimal classification outcomes.

7. RECOMMENDATIONS:

- **Expand the Feature Set**: Adding additional features, such as orbital parameters, atmospheric composition, or discovery methods, could provide deeper insights and improve the classification models. These features may capture more complex relationships between exoplanet properties and their types.
- **Explore Advanced Models**: While LDA and QDA have proven effective, trying out advanced machine learning models like Random Forest, Support Vector Machines (SVM), or Neural Networks could further improve classification accuracy. These models could capture intricate patterns and dependencies in the data that linear or quadratic models might miss.
- **Handle Non-linear Relationships More Effectively**: For datasets with evident non-linear patterns, I recommend exploring ensemble methods, such as Gradient Boosting or XGBoosting. These models are powerful and could complement the strengths of QDA by capturing non-linearities in a systematic way.
- **Incorporate Time-Series Analysis**: Given the dataset's span from 1992 to 2023, incorporating time-series analysis could provide valuable insights into trends in exoplanet discoveries and their evolving characteristics over time. This could reveal how advancements in detection technologies have shaped the dataset.
- **Improve Data Quality**: Enhancing the quality of the data remains a priority. Addressing missing values through advanced imputation techniques or supplementing the dataset with additional reliable sources could significantly improve model performance and reduce bias.
- **Develop a Hybrid Model**: I think it would be worth exploring a hybrid approach that combines the strengths of LDA and QDA. For example, using LDA as a baseline for its simplicity and interpretability while applying QDA to datasets or subsets with more evident non-linear trends might yield a balanced solution.

8. REFERENCES:

- https://www.kaggle.com/datasets/adityamishraml/nasaexoplanets