Predictive Modeling of Diabetes Using M.L Techniques

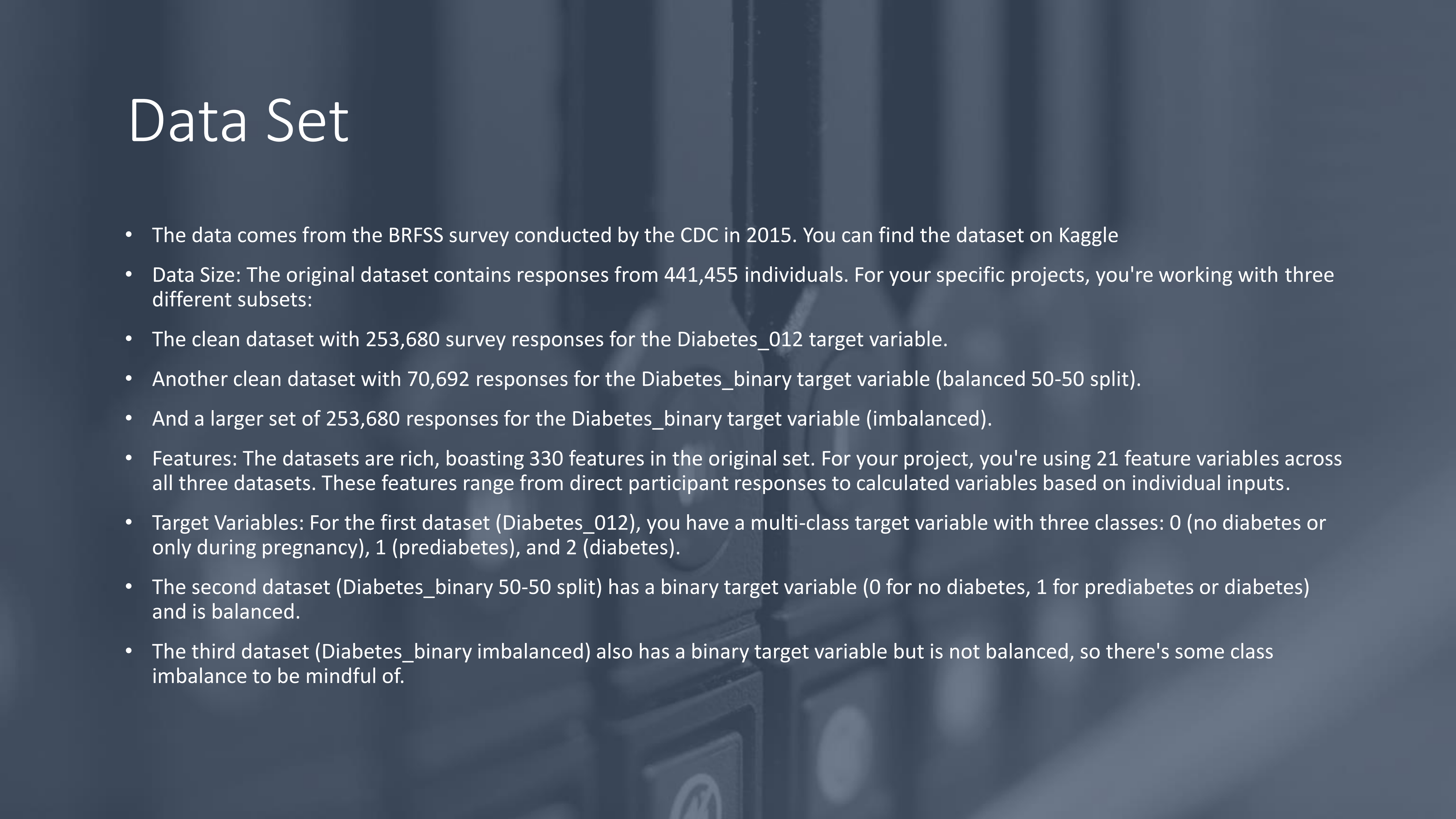
Group: 832

Sree Datta . P

Vijay Krishna . K

Durga Sai Sri . B

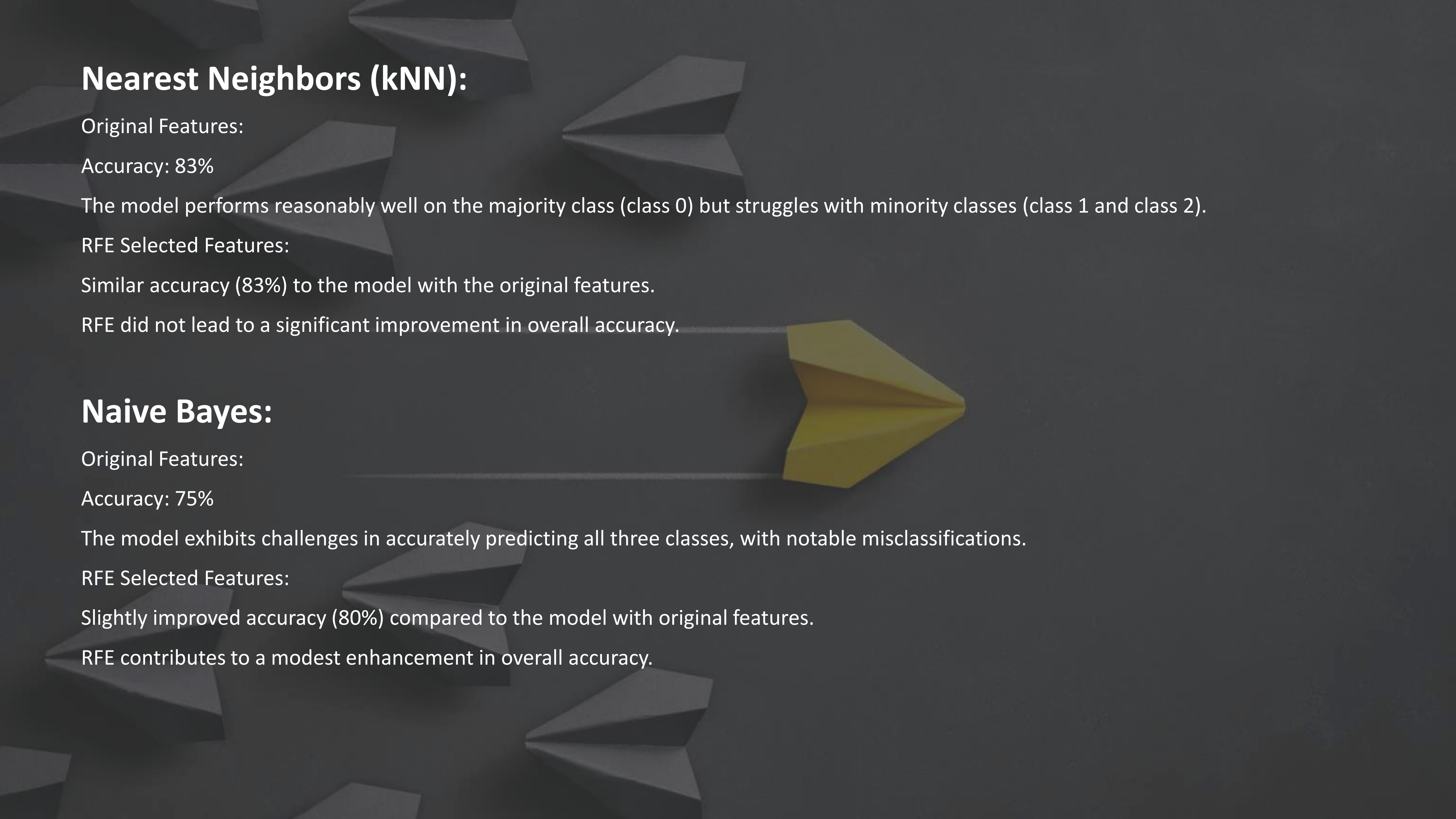
V.S.Tejaswi



Data Set

- The data comes from the BRFSS survey conducted by the CDC in 2015. You can find the dataset on Kaggle
- Data Size: The original dataset contains responses from 441,455 individuals. For your specific projects, you're working with three different subsets:
- The clean dataset with 253,680 survey responses for the Diabetes_012 target variable.
- Another clean dataset with 70,692 responses for the Diabetes_binary target variable (balanced 50-50 split).
- And a larger set of 253,680 responses for the Diabetes_binary target variable (imbalanced).
- Features: The datasets are rich, boasting 330 features in the original set. For your project, you're using 21 feature variables across all three datasets. These features range from direct participant responses to calculated variables based on individual inputs.
- Target Variables: For the first dataset (Diabetes_012), you have a multi-class target variable with three classes: 0 (no diabetes or only during pregnancy), 1 (prediabetes), and 2 (diabetes).
- The second dataset (Diabetes_binary 50-50 split) has a binary target variable (0 for no diabetes, 1 for prediabetes or diabetes) and is balanced.
- The third dataset (Diabetes_binary imbalanced) also has a binary target variable but is not balanced, so there's some class imbalance to be mindful of.

Conclusion



Nearest Neighbors (kNN):

Original Features:

Accuracy: 83%

The model performs reasonably well on the majority class (class 0) but struggles with minority classes (class 1 and class 2).

RFE Selected Features:

Similar accuracy (83%) to the model with the original features.

RFE did not lead to a significant improvement in overall accuracy.

Naive Bayes:

Original Features:

Accuracy: 75%

The model exhibits challenges in accurately predicting all three classes, with notable misclassifications.

RFE Selected Features:

Slightly improved accuracy (80%) compared to the model with original features.

RFE contributes to a modest enhancement in overall accuracy.