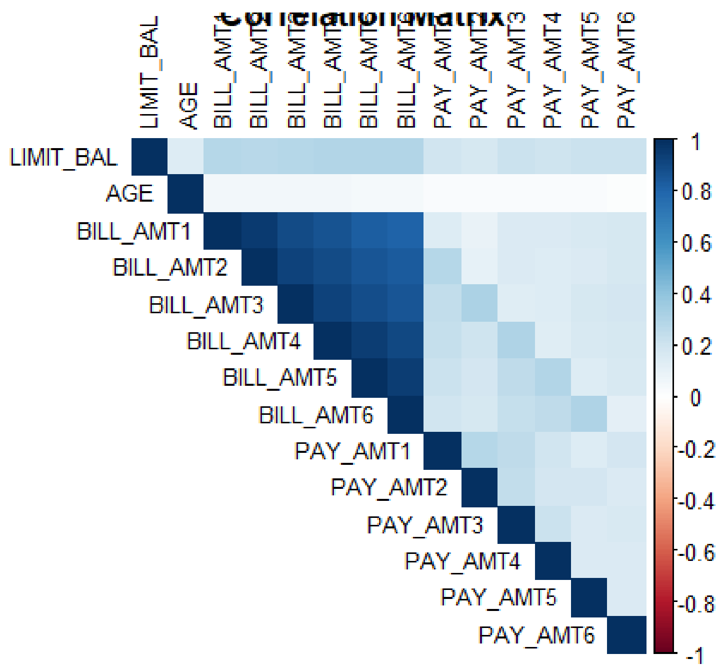
Final Project

IMT 572: Introduction to Data Science

Apeksha Tejwani

SUMMARY STATISTICS

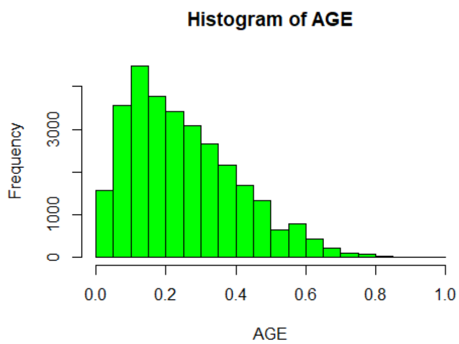
```
[1] "Summary Statistics After Transformation:"
> print(summary_transformed)
   LIMIT_BAL          SEX       EDUCATION MARRIAGE       AGE               PAY_0             PAY_2             PAY_3      
 Min.   :0.0000   1:11888   0:   14   0:   54   Min.   :0.0000   0      :14737   0      :15730   0      :15764  
 1st Qu.:0.0404   2:18112   1:10585   1:13659   1st Qu.:0.1207   -1     : 5686   -1     : 6050   -1     : 5938  
 Median :0.1313             2:14030   2:15964   Median :0.2241   1      : 3688   2      : 3927   -2     : 4085  
 Mean   :0.1591             3: 4917   3:  323   Mean   :0.2497   -2     : 2759   -2     : 3782   2      : 3819  
 3rd Qu.:0.2323             4:  123             3rd Qu.:0.3448   2      : 2667   3      :  326   3      :  240  
 Max.   :1.0000             5:  280             Max.   :1.0000   3      :  322   4      :   99   4      :   76  
                            6:   51                              (Other):  141   (Other):   86   (Other):   78  
     PAY_4             PAY_5             PAY_6           BILL_AMT1        BILL_AMT2        BILL_AMT3         BILL_AMT4     
 0      :16455   0      :16947   0      :16286   Min.   :0.0000   Min.   :-69777   Min.   :-157264   Min.   :-170000  
 -1     : 5687   -1     : 5539   -1     : 5740   1st Qu.:0.1497   1st Qu.:  2985   1st Qu.:   2666   1st Qu.:   2327  
 -2     : 4348   -2     : 4546   -2     : 4895   Median :0.1663   Median : 21200   Median :  20089   Median :  19052  
 2      : 3159   2      : 2626   2      : 2766   Mean   :0.1918   Mean   : 49179   Mean   :  47013   Mean   :  43263  
 3      :  180   3      :  178   3      :  184   3rd Qu.:0.2059   3rd Qu.: 64006   3rd Qu.:  60165   3rd Qu.:  54506  
 4      :   69   4      :   84   4      :   49   Max.   :1.0000   Max.   :983931   Max.   :1664089   Max.   : 891586  
 (Other):  102   (Other):   80   (Other):   80                                                                     
   BILL_AMT5        BILL_AMT6         PAY_AMT1          PAY_AMT2          PAY_AMT3         PAY_AMT4     
 Min.   :-81334   Min.   :-339603   Min.   :0.000000   Min.   :      0   Min.   :     0   Min.   :     0  
 1st Qu.:  1763   1st Qu.:   1256   1st Qu.:0.001145   1st Qu.:    833   1st Qu.:   390   1st Qu.:   296  
 Median : 18105   Median :  17071   Median :0.002404   Median :   2009   Median :  1800   Median :  1500  
 Mean   : 40311   Mean   :  38872   Mean   :0.006483   Mean   :   5921   Mean   :  5226   Mean   :  4826  
 3rd Qu.: 50191   3rd Qu.:  49198   3rd Qu.:0.005731   3rd Qu.:   5000   3rd Qu.:  4505   3rd Qu.:  4013  
 Max.   :927171   Max.   : 961664   Max.   :1.000000   Max.   :1684259   Max.   :896040   Max.   :621000  
                                                                                                           
    PAY_AMT5           PAY_AMT6         default.payment.next.month
 Min.   :     0.0   Min.   :     0.0   0:23364                   
 1st Qu.:   252.5   1st Qu.:   117.8   1: 6636                   
 Median :  1500.0   Median :  1500.0                             
 Mean   :  4799.4   Mean   :  5215.5                             
 3rd Qu.:  4031.5   3rd Qu.:  4000.0                             
 Max.   :426529.0   Max.   :528666.0                             
```

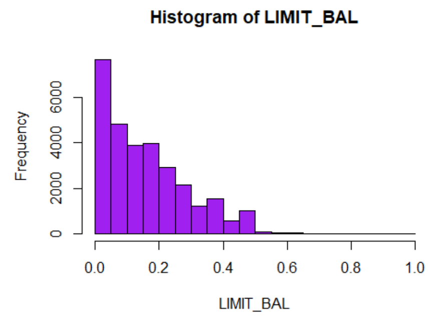


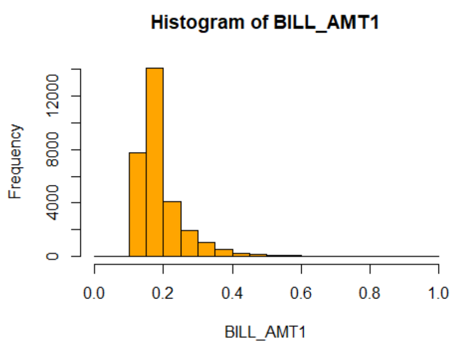
This correlation matrix visualization helps to refine the variable selection process for further modeling. Conclusions:

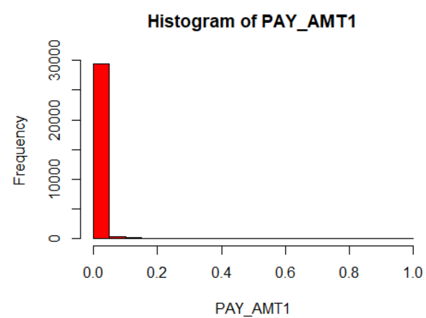
- BILL_AMT1 to BILL_AMT6 are highly correlated with each other (close to 1). Including all of them in a regression model might cause multicollinearity, which can distort the coefficients and reduce model meaning. Hence, I would keep only BILL_AMT1
- Keep LIMIT_BAL (low correlation with most variables).
- Payment amounts (PAY_AMT1 to PAY_AMT6) have lower correlations with other variables, showing they might provide unique information. They are however correlated with one another, hence I would choose the recent payment amounts (PAY_AMT1, PAY_AMT3)

Exploratory Analysis on dataset in R:

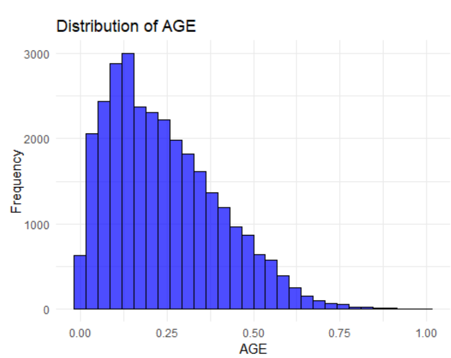


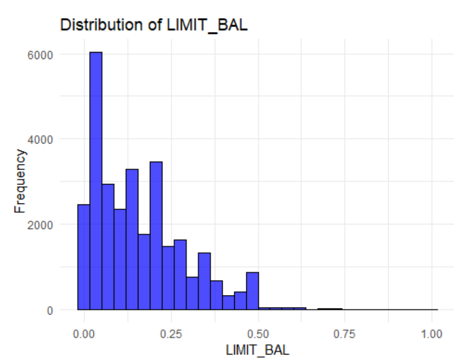


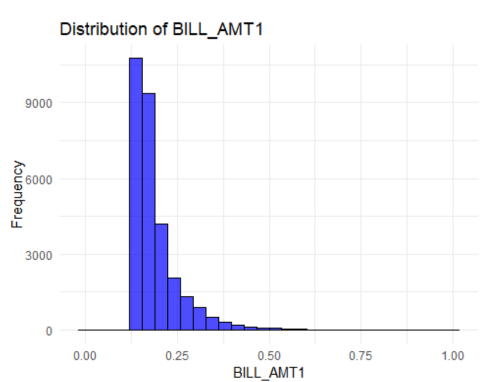


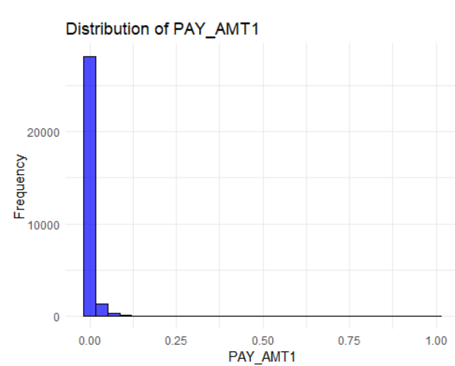


Normalized variables:



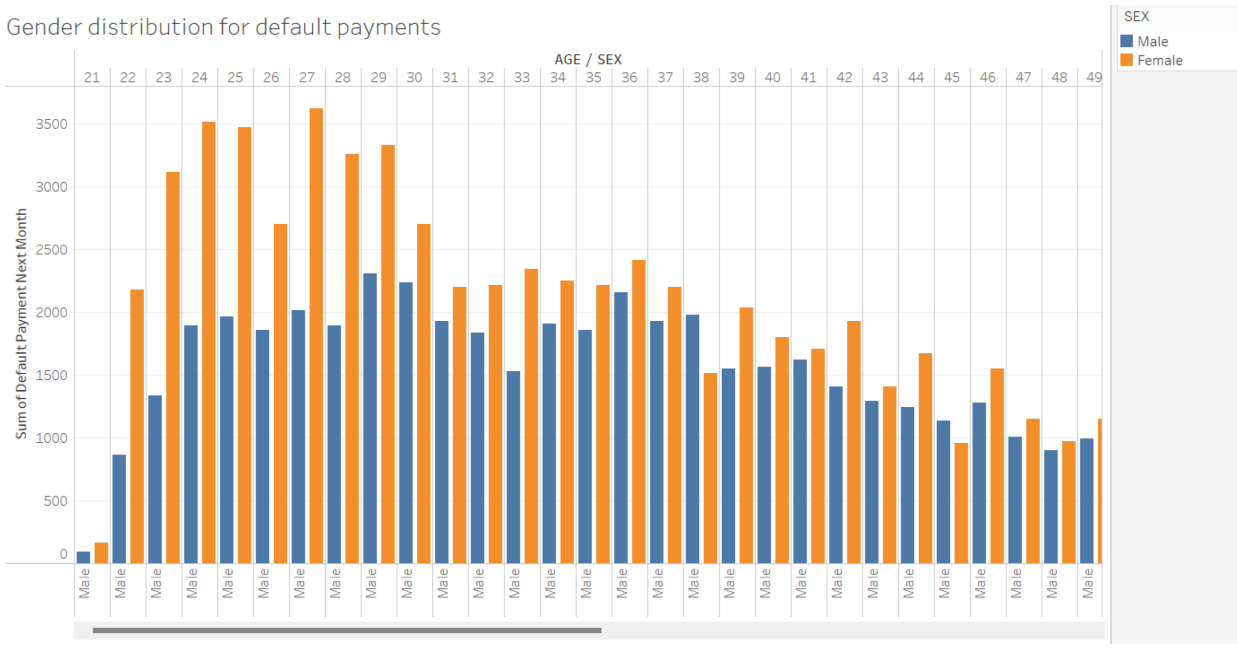






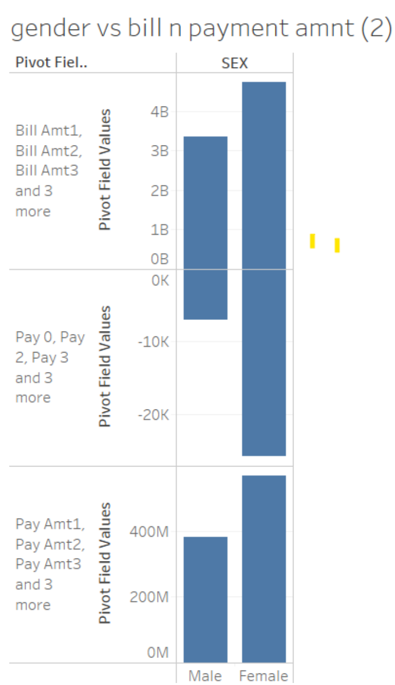
Exploratory Analysis on dataset in Tableau:

- For many age groups, females (orange bars) appear to have higher defaults compared to males (blue bars), especially in the 25–30 age range.



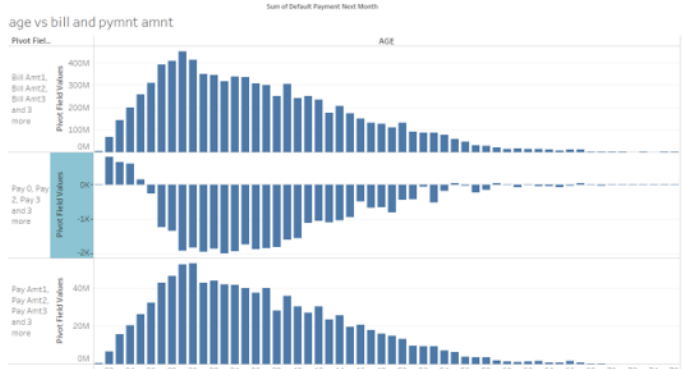
- **Credit Usage**: Females have both higher bill amounts and payment amounts, indicating higher credit usage compared to males.

  Payment Behaviour: Despite the higher payments, females also show a higher negative payment status, suggesting delayed or inconsistent payments.



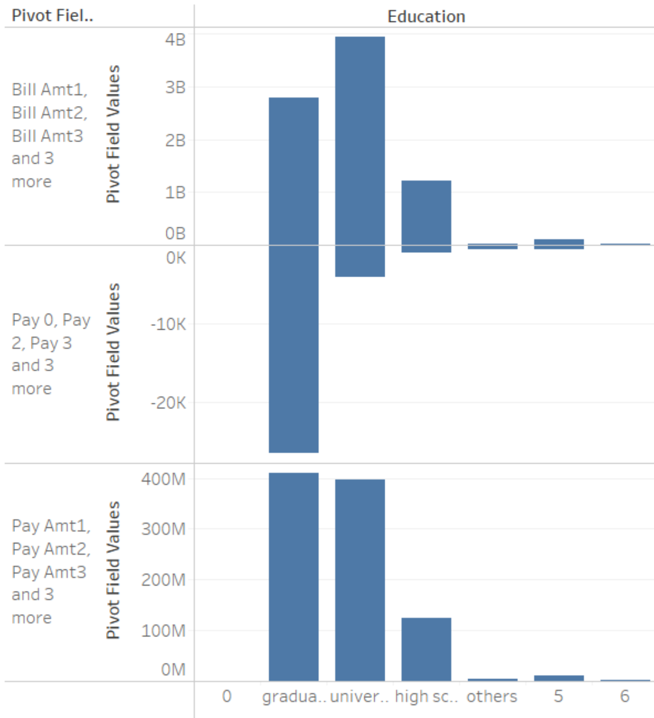
- **Single individuals** exhibit slightly higher delayed payments and total payments, suggesting they might be at a higher risk of defaults compared to married individuals.

- Younger individuals, particularly those in their late 20s and early 30s, are more likely to have higher bill amounts, payment amounts, and default rates.
- Older individuals tend to have lower bill amounts and payments, with significantly lower default rates, likely reflecting better financial stability and experience in managing debt.

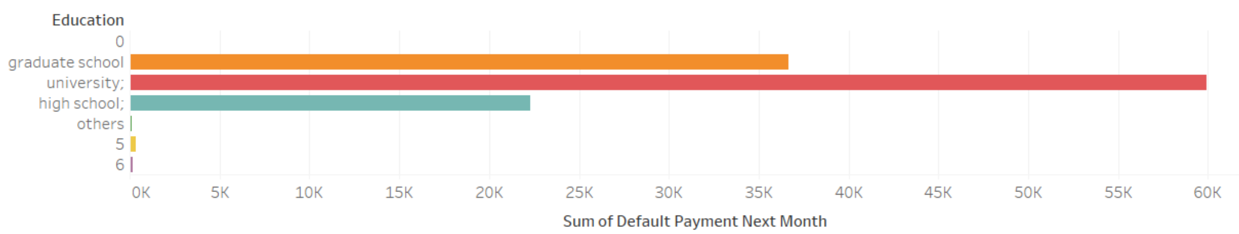


- Individuals with **university-level education** are a significant focus for default analysis due to their higher bill amounts and greater default rates.
- The **graduate school group** may represent a more financially stable segment within the dataset.

education vs bill n payment amnt



education



REGRESSION BASED EXPLORATORY ANALYSIS

Logit Model Summary

1. **Coefficients and Significance:**
   - The coefficients represent the log-odds change for a one-unit change in the predictor variable, holding other variables constant.
   - A smaller P- value indicates statistical significance:
     - Variables such as LIMIT_BAL, PAY_AMT1, PAY_AMT3, BILL_AMT1, and various levels of PAY_0 and PAY_2 have **significant effects** (***).
     - These significant predictors are important drivers of the outcome (default).
2. **Marginal Effects:**
   - logitmfx computes the marginal effects (dF/dx), representing the change in the probability of default for a one-unit change in the predictor variable.
   - Key findings:
     - LIMIT_BAL has a significant **negative effect**, indicating that higher credit limits decrease the probability of default.
     - PAY_AMT1 and PAY_AMT3 have negative effects, meaning larger payments reduce the likelihood of default.
     - PAY_0, PAY_2, and PAY_3 (payment history indicators) show positive effects, meaning higher overdue payments increase the probability of default.

Probit Model Summary

1. **Coefficients and Significance:**
   - Similar to the Logit model, LIMIT_BAL, PAY_AMT1, PAY_AMT3, BILL_AMT1, and PAY_0, PAY_2, and PAY_3 have statistically significant effects (***).
2. **Marginal Effects:**
   - Key observations:
     - LIMIT_BAL still has a strong negative marginal effect, confirming its protective role against default.
     - Payment history (PAY_0, PAY_2, PAY_3) continues to show positive effects, highlighting its critical importance in predicting default.

Overall Fit: The Probit model slightly outperforms the Logit model based on AIC.

Summary of different trials for various sets of predictors

Predictor sets	Logit AIC	Probit AIC
LIMIT_BAL + PAY_AMT1 + PAY_AMT3 + BILL_AMT1 + PAY_0 + PAY_2 + PAY_3 + SEX + EDUCATION + MARRIAGE	26382.43	26378.6
LIMIT_BAL, PAY_AMT1, PAY_AMT3, PAY_0, PAY_2, PAY_3	26502.85	26498.33
LIMIT_BAL, BILL_AMT1, PAY_0, PAY_2, SEX, EDUCATION, MARRIAGE	26596.13	26591.75

TRAINING A CLASSIFIER ON OUTCOME DEFAULT

1. **KNN:**

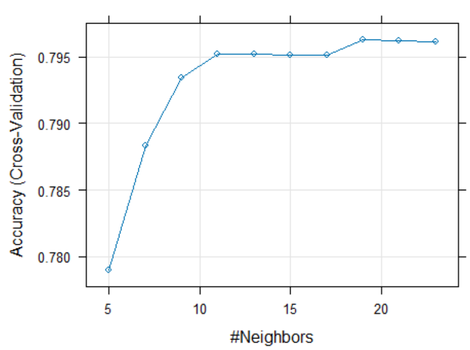
```
> knn_caret
k-Nearest Neighbors

30000 samples
   10 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 24000, 24001, 24000, 23999, 24000
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   5  0.7789667  0.2212978
   7  0.7882999  0.2260289
   9  0.7934000  0.2272785
  11  0.7952333  0.2221839
  13  0.7952334  0.2116732
  15  0.7951667  0.2041995
  17  0.7951001  0.1971845
  19  0.7963000  0.1974175
  21  0.7962000  0.1925579
  23  0.7961668  0.1887470

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 19.
```



Explanation: The optimal k value was determined to be **19**, as it corresponds to the highest accuracy 79.6%. This means the best-performing kNN model uses 19 neighbors to make predictions.

2. SVM Linear

```
> print(svm_linear_caret)
Support Vector Machines with Linear Kernel

30000 samples
   10 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 24000, 24000, 24000, 23999, 24001
Resampling results:

  Accuracy   Kappa
  0.8173666  0.3357327

Tuning parameter 'C' was held constant at a value of 1
```

Explanation: **Accuracy of 81.74%**, is a good result, indicating that the linear kernel was effective for this dataset.

3. Random Forest

```
Random Forest

30000 samples
   10 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 24001, 23999, 24000, 24000, 24000
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.8082000  0.2559178
   6    0.8199001  0.3608883
  11    0.8173334  0.3618985
  16    0.8116667  0.3506241
  20    0.8089666  0.3472920
  25    0.8070334  0.3434435
  30    0.8062333  0.3434632
  34    0.8063999  0.3439704
  39    0.8058999  0.3427293
  44    0.8058999  0.3427685

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 6.
```

Explanation: The Random Forest classifier performs well on this dataset with an optimal accuracy of **81.99%.**

4. SVM Radial

```
Support Vector Machines with Radial Basis Function Kernel

30000 samples
   10 predictor
    2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 24001, 24000, 24000, 24000, 23999
Resampling results across tuning parameters:

  C       Accuracy   Kappa
    0.25  0.8160333  0.3148228
    0.50  0.8164999  0.3230074
    1.00  0.8162666  0.3295654
    2.00  0.8160000  0.3390262
    4.00  0.8145333  0.3404970
    8.00  0.8121333  0.3391421
   16.00  0.8108666  0.3395819
   32.00  0.8089332  0.3373690
   64.00  0.8055665  0.3312221
  128.00  0.8022665  0.3256194

Tuning parameter 'sigma' was held constant at a value of 0.03780675
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.03780675 and C = 0.5.
> gc()
           used  (Mb) gc trigger  (Mb)  max used   (Mb)
Ncells  5232583 279.5   10073775 538.0  10073775  538.0
Vcells 36652334 279.7  124668149 951.2 304310570 2321.8
```

Explanation: The SVM Radial model achieved an accuracy of **81.65**%

Comparison between models:

1. **Random Forest (RF)** achieves the highest accuracy of approximately **0.8199**. This suggests that the Random Forest algorithm is the most effective model among the four tested for predicting the target variable default.payment.next.month.
2. **SVM with Linear Kernel** comes in second, with an accuracy of **0.8174**, which is very close to the Random Forest.
3. **SVM with Radial Kernel** achieves an accuracy of **0.8165**, slightly lower than the SVM Linear Kernel and Random Forest. It is still competitive but slightly less effective for this dataset.
4. **kNN (k-Nearest Neighbors)** has the lowest accuracy of **0.7963**. While still a reasonable classifier, it does not perform as well as the other models.

Conclusion:

Based on the accuracies, **Random Forest** should be selected as the best model for this problem since it provides the highest accuracy.

SVM with Linear Kernel could also be a suitable alternative, if runtime or computational efficiency is a concern.