Data 622
Homework # 3 (175 points)

For this assignment, we will be working with the attached dataset on loan approval status. The loan approval status is the target variable here. For all the models here, don't forget to also provide performance statistics.

1. As you begin working with the dataset, at the beginning, please conduct a thorough exploratory data analysis. This step is necessary as you figure out which variables should be included in models. (10 points)

2. Use the LDA algorithm to predict the loan approval status. Please be sure to walk through the steps you took, this includes how you decided on the key variables. (40 points)

3. Use K-nearest neighbor (KNN) algorithm to predict the species variable. Please be sure to walk through the steps you took. This includes talking about what value for 'k' you settled on and why. (40 points)

4. Use Decision Trees to predict on loan approval status. (40 points)

5. Use Random Forests to predict on loan approval status. (40 points)

6. Model performance: please compare the models you settled on in problem # 2- 5. Comment on their relative performance. Which one would you prefer the most? Why? (5 points)

The loan approval status data dictionary:

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Undergraduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status (Target)	Loan approved (Y/N)