Homework # 3 (200 points)

Due on April 09, 11:59pm EST

For this assignment, we will be working with two different datasets. For problem # 1, we will still be working with the Penguin dataset. For problem # 2-4, we will be working with the attached dataset on loan approval status.

1. Please use K-nearest neighbor (KNN) algorithm to predict the species variable. Please be sure to walk through the steps you took. (40 points)

2. Please use the attached dataset on loan approval status to predict loan approval using Decision Trees. Please be sure to conduct a thorough exploratory analysis to start the task and walk us through your reasoning behind all the steps you are taking. (40 points)

3. Using the same dataset on Loan Approval Status, please use Random Forests to predict on loan approval status. Again, please be sure to walk us through the steps you took to get to your final model. (50 points)

4. Using the Loan Approval Status data, please use Gradient Boosting to predict on the loan approval status. Please use whatever boosting approach you deem appropriate; but please be sure to walk us through your steps. (50 points)

5. Model performance: please compare the models you settled on for problem # 2 – 4. Comment on their relative performance. Which one would you prefer the most? Why? (20 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/ Under Graduate)
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)