Name:

Giovanni Zavalza

Netid:

gzava3

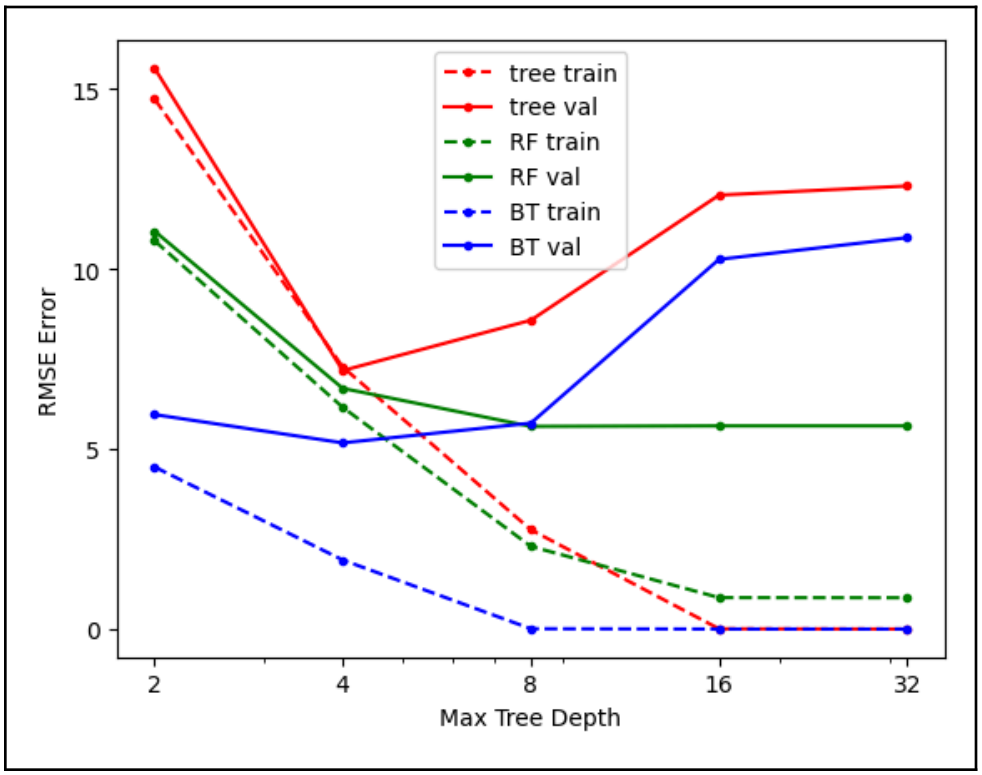
CS 441 - HW 4: Trees and MLPs

Complete the sections below. You do not need to fill out the checklist. **Do select all relevant pages in Gradescope.**

Total Points Claimed	**[ ] / 170**
1. Model Complexity with Tree Regressors	
a. Depth vs. Error plot	[ ] / 10
b. Analysis	[ ] / 20
2. MLPs with MNIST	
a. Loss Curves	[ ] / 20
b. Model Selection and Results	[ ] / 20
3. Species Prediction	
a. Feature Analysis	[ ] / 10
b. Simple Rule	[ ] / 10
c. Model Design	[ ] / 10
4. Stretch Goals	
a. Improve MNIST classification	[ ] / 30
b. A second simple rule	[ ] / 10
c. Positional encoding of RGB Image	[ ] / 30

1. Model Complexity with Tree Regressors

a. Include your plot below.



b. Analyze your results:

1. For a given max tree depth, which of regressor model (single tree, random forest, boosted tree) has the lowest bias (or most powerful)?

Boosted Tree

2. For single regression trees, what tree depth achieves minimum validation error?

4

3. A model “overfits” when increasing the complexity increases the validation error. Which model is least prone to overfitting? Why?

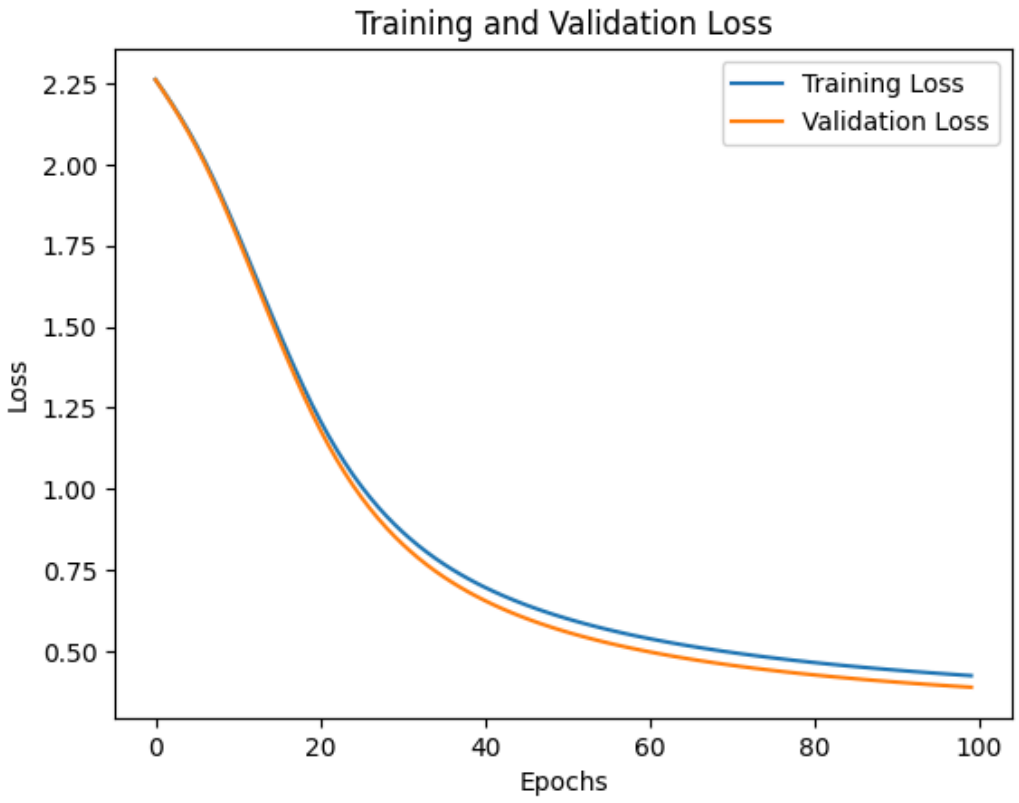
Random Forest
Given that each tree is trained independently on a single portion of the subset, it is very difficult for the model to memorize answers. Feature distinction at each split breaks this down further leading to a very generalized aggregation that is robust against overfitting.

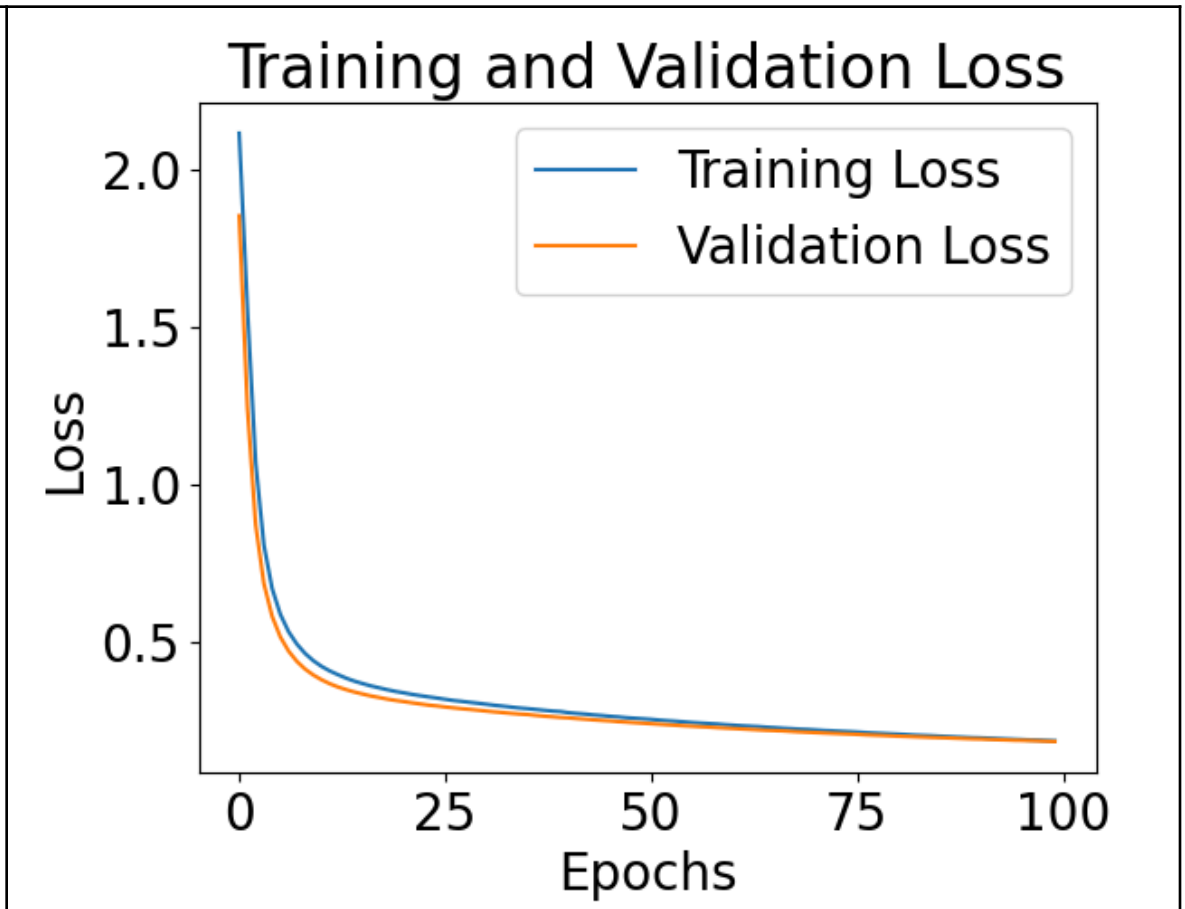
4. Do boosted trees seem to perform better with smaller or larger trees? Why?

Smaller. The point of boosted trees is to aggregate a bunch of simple trees to produce a generalization. If each tree is very complex and has high depth, it leads to severe overfitting as it memorizes the noise rather than the overall feature patterns.

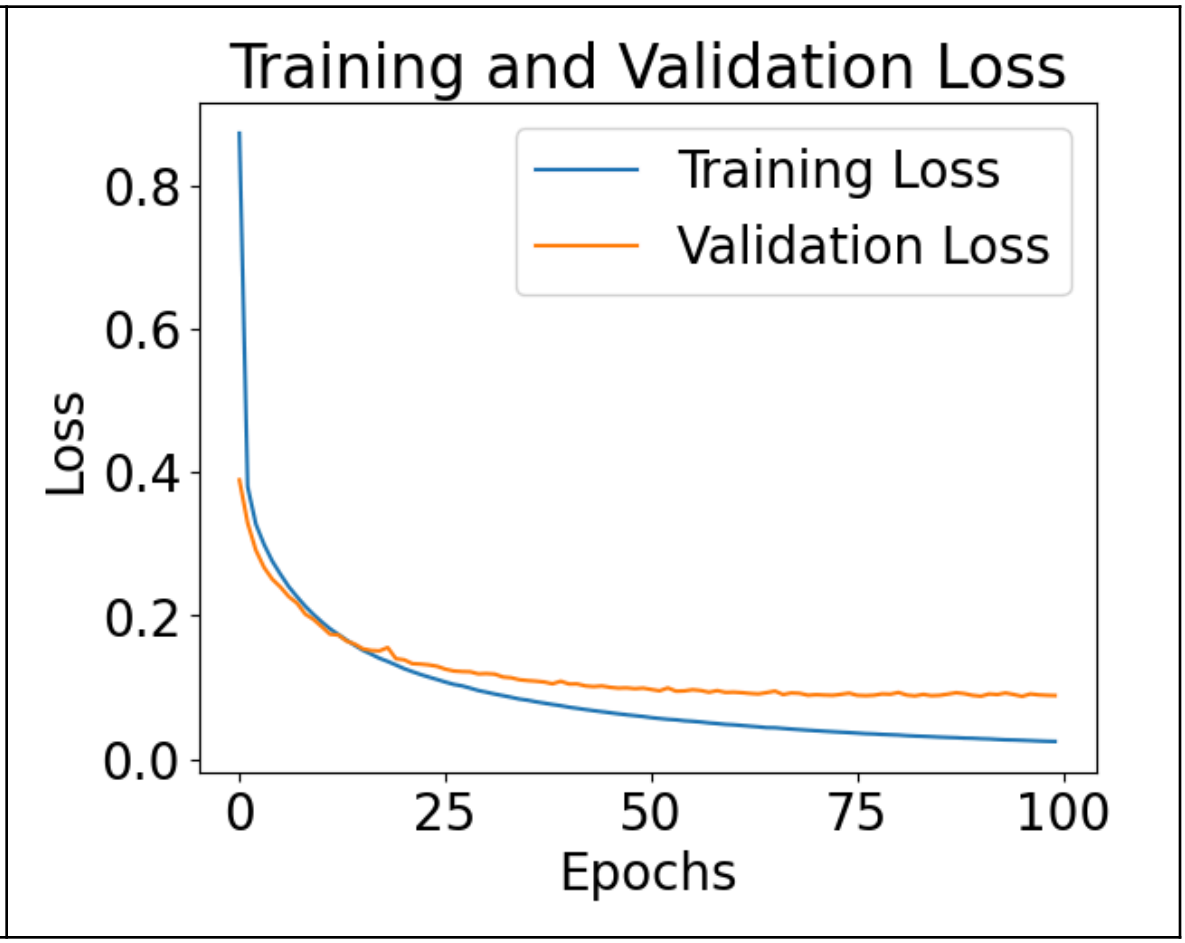
2. MLPs with MNIST

a. **Show the loss curves** for 3 learning rates (1E-2, 1E-1, 1E1) training for 100 epochs. An example of the loss curves is shown for LR=0.001.

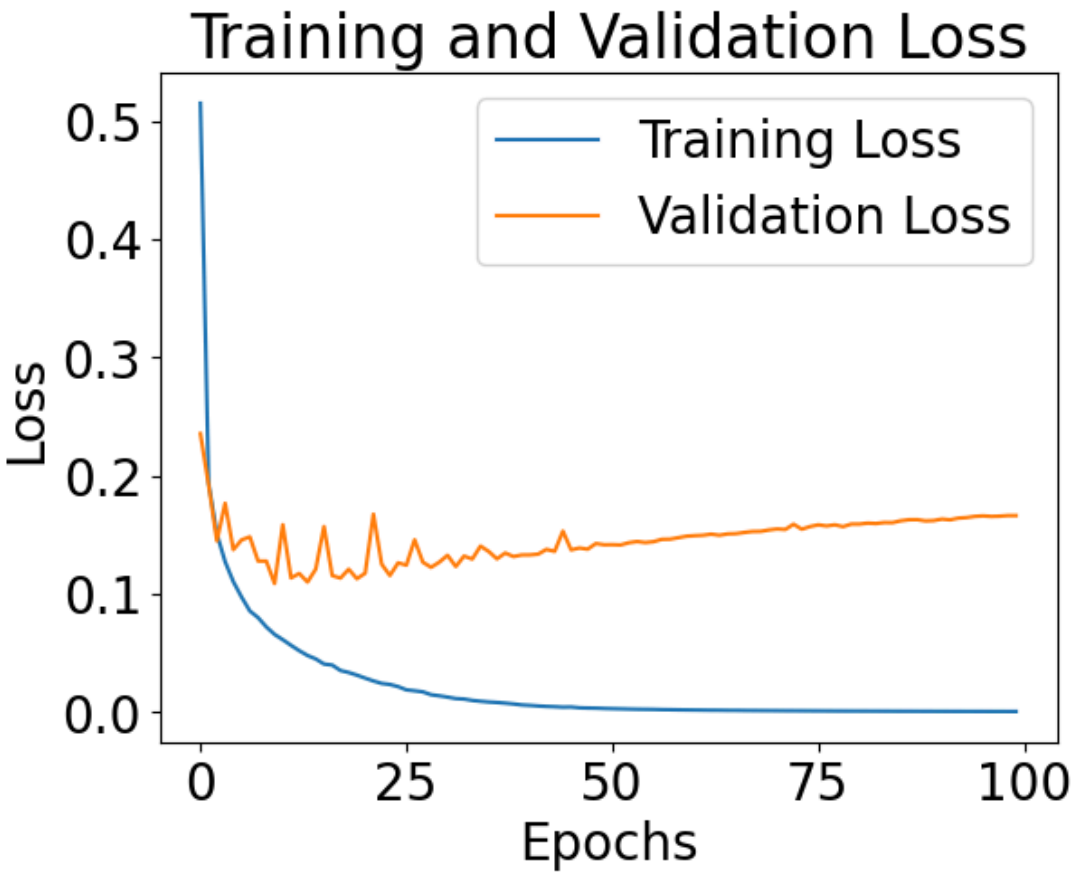
Learning Rate	Loss Curve Plot
0.001	

0.01	

Training and Validation Loss
Training Loss
Validation Loss
Loss
2.0
1.5
1.0
0.5
0
25
50
75
100
Epochs

0.1	 Training and Validation Loss Training Loss Validation Loss Loss 0.0 0.2 0.4 0.6 0.8 0 25 50 75 100 Epochs

1



b. Model selection and results

Select the best hyperparameters (learning rate and number of epochs up to 100) based on minimizing the validation loss.

Learning Rate

0.1

Epochs

100 (96 was min(valid_loss) but there was still a downwards trend so 100 will provide better results for test dataset)

Report the losses and errors for the model trained with these hyperparameters:

Use scientific notation with one decimal place, e. 1.5E-3

Training **Loss**	Validation **Loss**	Test **Loss**
2.5E-2	8.7E-2	8.0E-2

Show two decimal places for percent

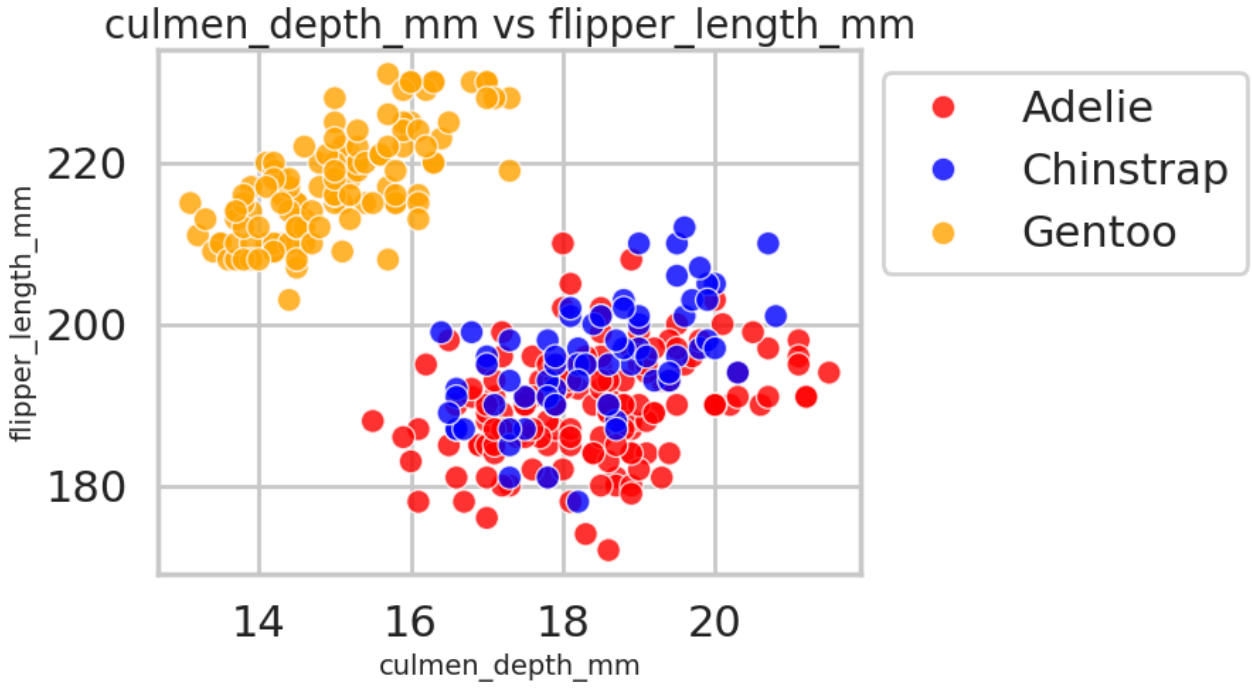
Training **Error** (%)	Validation **Error** (%)	Test **Error** (%)
0.47%	2.53%	2.50%

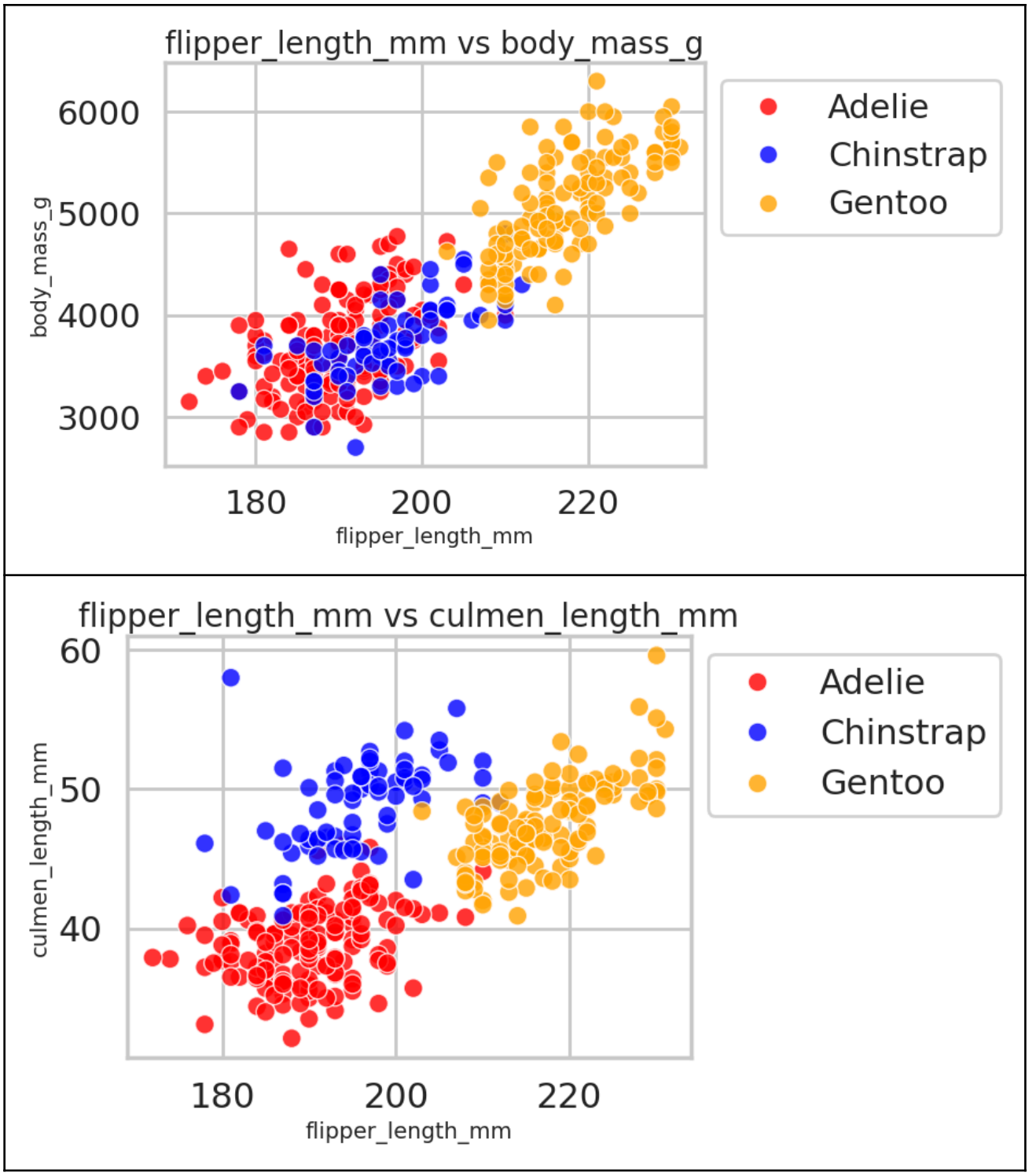
3. Species Prediction

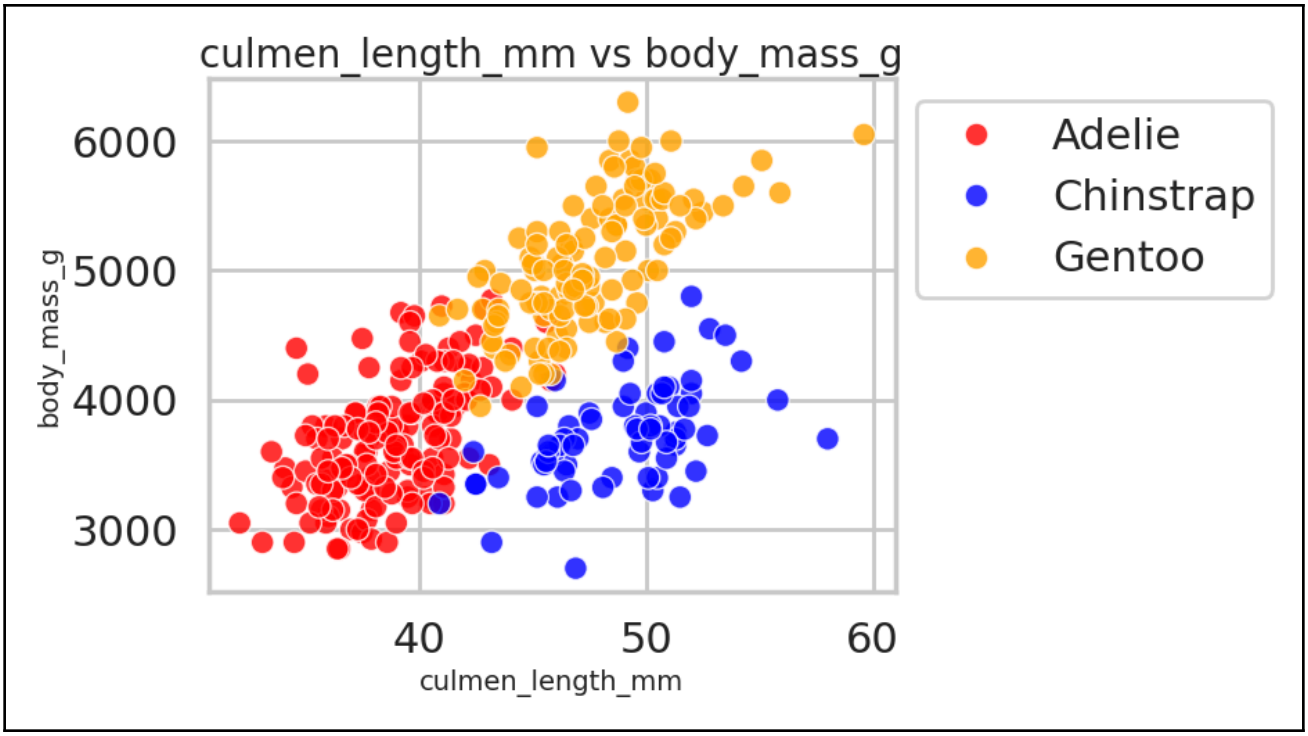
a. Visualization of Features

Include at least two scatterplots of pairs of features.

Visualization (labels should make clear which features are used)




flipper_length_mm vs body_mass_g
body_mass_g
6000
5000
4000
3000
180
200
220
flipper_length_mm
Adelie
Chinstrap
Gentoo
flipper_length_mm vs culmen_length_mm
culmen_length_mm
60
50
40
180
200
220
flipper_length_mm
Adelie
Chinstrap
Gentoo



You may extend the table if you have more results

Of these three options, which two features (by themselves) are best able to classify the penguin species?

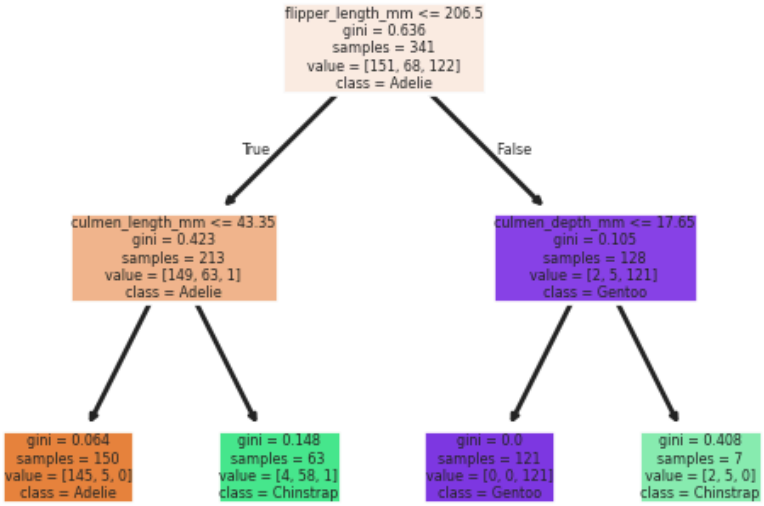
1. Culmen Depth + Flipper Length
2. Flipper Length + Culmen Length
3. Flipper Length + Body Mass

Flipper Length + Culmen Length

b. Simple rule to identify Gentoo

Display your decision tree with labeled features and classes.

Decision Tree for Identifying Gentoo Penguins



Write down the simple two-part rule to identify Gentoo. For example, the format should be "If Mass > 3000 and Culmen Depth < 17, then species is Gentoo".

If…

Flipper_length_mm > 206.5

and

Culmen_depth_mm <=17.65

then species is Gentoo.

Rule precision: fraction of penguins that satisfy this rule that are Gentoos (# gentoo predicted / # predicted)

121 / 121

Rule recall: fraction of all Gentoo penguins that are identified as Gentoo using this rule (# gentoo predicted / # gentoo)

121 / 122

c. Model Design

Describe the model that achieves best 5-fold cross-validation accuracy:

Random Forest combines multiple 'weak' decision trees to get an average. It trains each tree on a random subset of the data and using a random selection of features for each split, making the trees unique. As mentioned before, this reduces overfitting by combining trees before they are complex enough to start memorizing noise.Thus, Random Forest generalizes by averaging the predictions of many distinct trees.

5-fold Cross-Validation Accuracy: (xx.x%)

99.1%

4. Stretch Goals

a. Improve MNIST Classification Performance using MLPs

Report the classification val and test errors and details of your best method. Describe your approach and parameters. Feel free to change the MLP batch size, optimizer (e.g. try Adam), learning rate, number of epochs, hidden layer size, activation layer, or anything else.

Description and key parameters

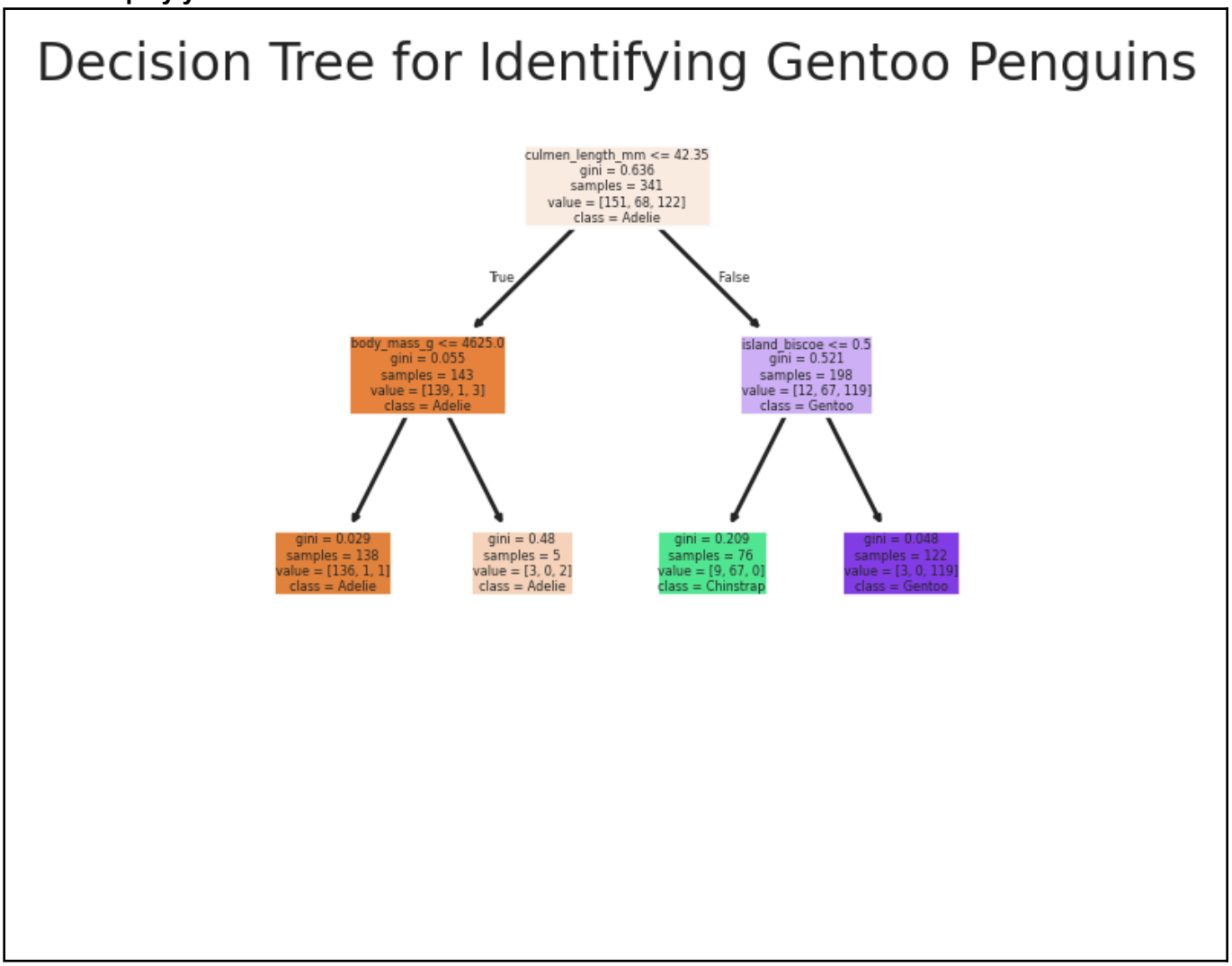
Optimizer =
Hidden layer(s) =
Learning rate =
Number of epochs =

Any other details:

Validation **Error** (%)	Test **Error** (%)

b. Find a second simple rule to identify Gentoo

Display your decision tree with labeled features and classes.



Provide the second two-part rule here (that is substantially different from your first rule).

If…

Culmen_length_mm > 42.35

and

Island_biscoe > 0.5

then species is Gentoo.

Rule precision: fraction of penguins that satisfy this rule that are Gentoos (# gentoo predicted / # predicted)

119 / 122

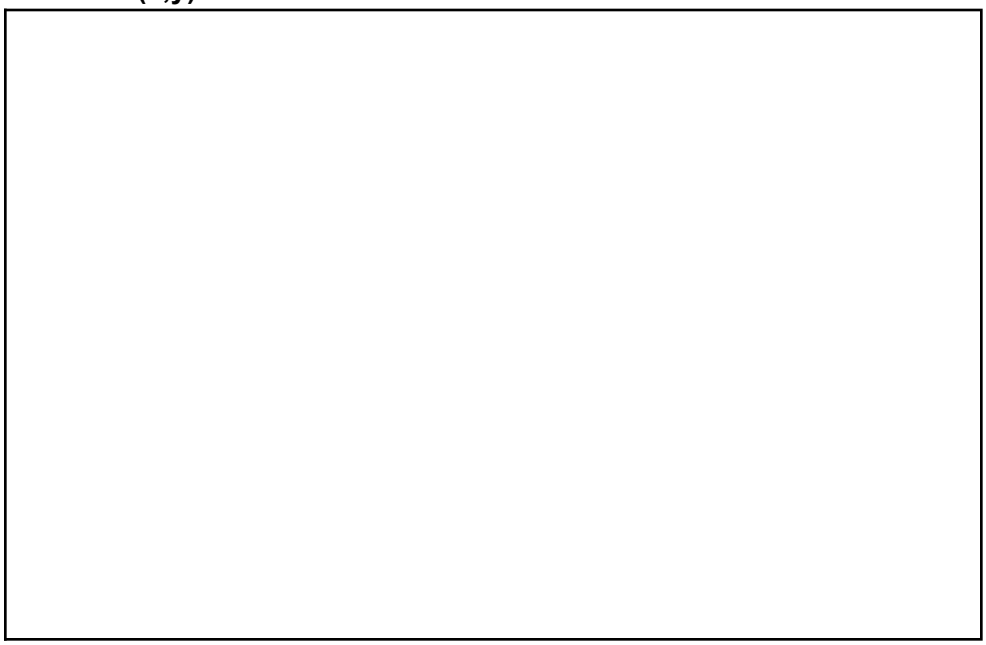
Rule recall: fraction of all Gentoo penguins that are identified as Gentoo using this rule (# gentoo predicted / # gentoo)

119 / 122

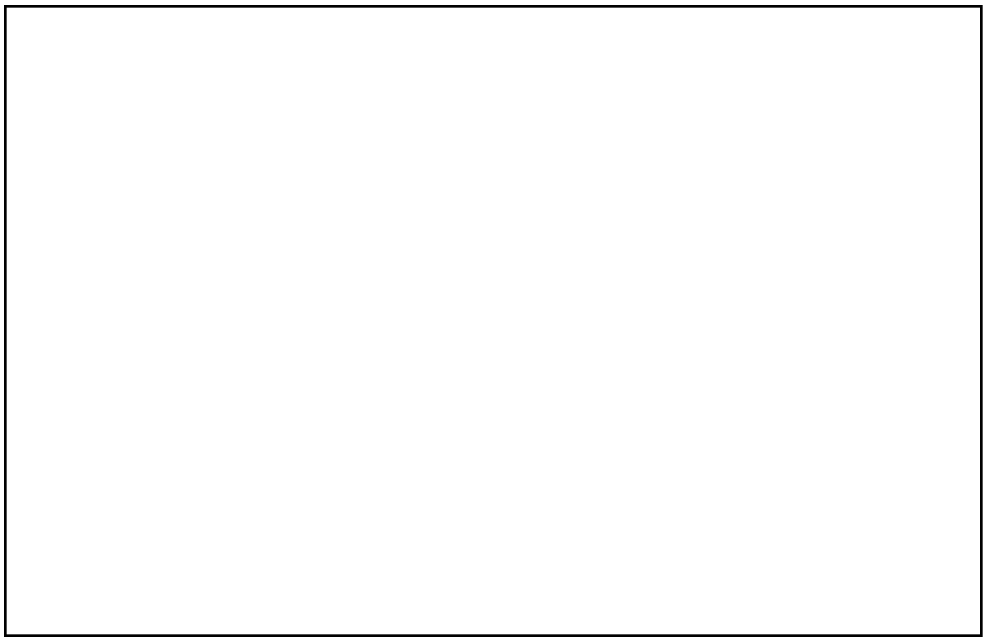
c. Positional encoding

Show the RGB image obtained by predicting directly from (x,y) and the image obtained by predicting from the positional encoding.

Input to network is (x,y)



Input to network is pos_enc(x, y)



Acknowledgments / Attribution

List any outside sources for code or improvement ideas or “None”.

ChatGPT, links, slides, and Python documentation.