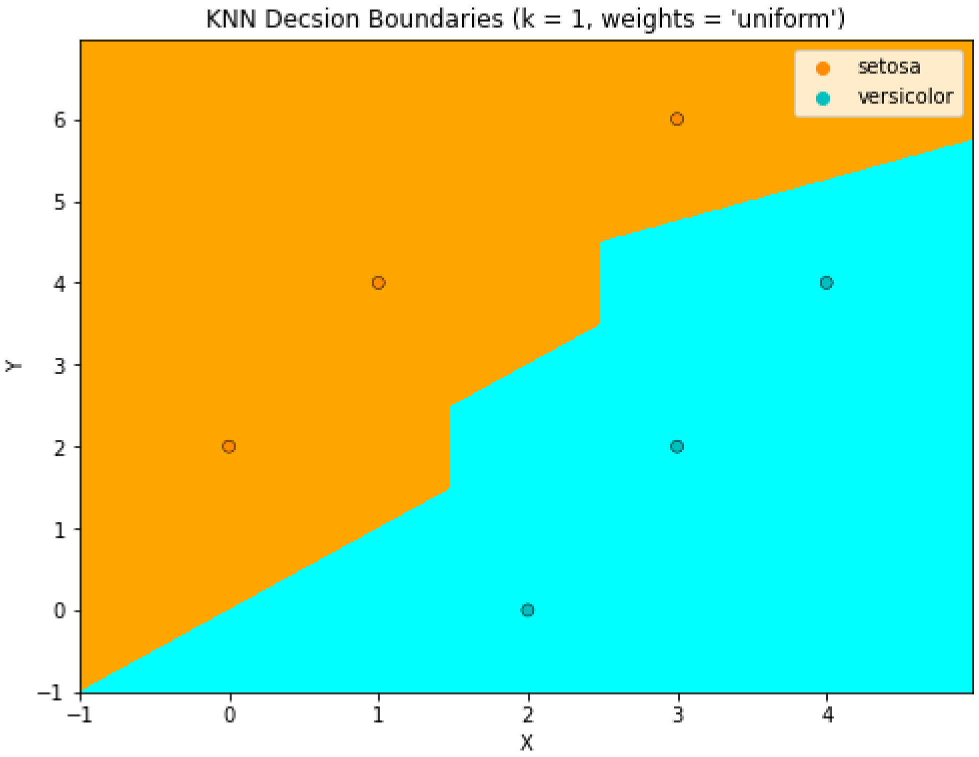
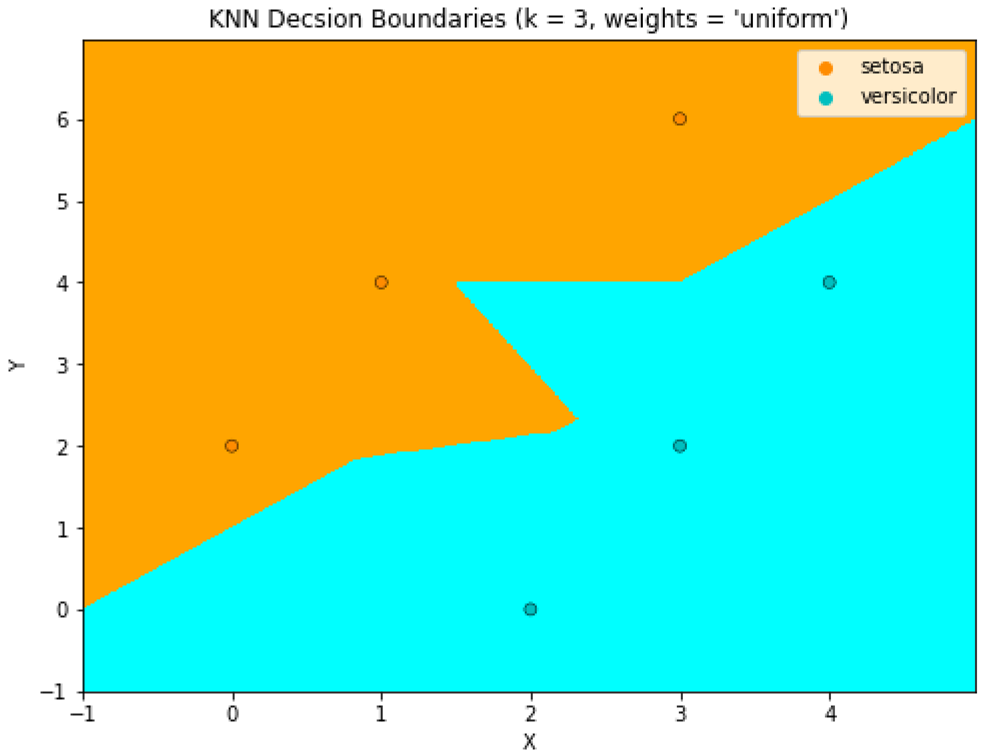
Devin Yerasi
UID: 305 167 818

1. K-Nearest Neighbors for Classification(link to collaboratory notebook: https://colab.research.google.com/drive/1O9OqepQ9D3UJyre---X_kWr3ScXNTOWa?usp=sharing) Also, I understand the concept and how to use python for it, BUT I’m not too sure how to go about this by hand, and would love a demo in class or discussion please :)
    a.



    b.

Devin Yerasi
UID: 305 167 818



c. If our dataset changed to what was described, there would be incredibly high variance due to the large spread of y values over a limited dataset. One solution for this would be regularization.
d. I would run a for loop over multiple values from k ranging from 1 -> 1000 with small increments from 1-100 and increments of at least 10 for the rest. Inside the for-loop I would 'train' a knn model and compute accuracy, precision, recall, and f1 scores on the test set for each value of k. Then I would find where these values peak and further explore more values of k with smaller increments in this range till I found my ideal k.

2. True or False, Simple Explanations
    a. False: LASSO Regression will tend to give you sparse coefficients B, as it treats all same increment reduction of Bj's as worth the same, leading to more terms being zeroed out.
    b. False: As Ridge Regression weights same increment reductions of larger Bj's as worth more in the loss term, it will emphasize reducing these over smaller Bj's, resulting in a shrinking of coefficients, but not a zero-ing out of them.
    c. True: A larger lamda in Ridge Regression will greater penalize B (Bj^2), resulting in a smaller magnitude of B after minimizing the loss function with the L2 error term. It will treat the same increment reduction as worth more the greater magnitude Bj has, resulting in an emphasis in reducing larger Bj values.
    d. True: Ridge regression tends to 0 out coefficients, so a larger lamda will zero out more features, so the number of nonzero coefficients in B will decrease.

    e. False: Data Scientists must use a diverse set of tools for different stages of the data science process
    f. False: Data Engineering can often be an essential part of the data science pipeline, especially when collecting a new dataset.

3. Logistic Regression Boundary and Interpretation
    a. Xp = [log(P(Y=1)/(1-P(Y=1)) - B0 - B1X1-...B(p-1)X(p-1)]/Bp
    b. When p=2, X2 increases with a decrease in X1. X1 and X2 have an inverse relationship on the decision boundary.
    c. As B1 and B2 are both positive, the model will be shaped as a positive sloping S-curve. As B0 is negative, the curve will be shifted to the right by c=-B0/(B1+B2). As B1 and B2 are generally small, the slope of the curve will be larger. When X1=X2=0, P(Y=1) = 1/(1+e^(-B0)) = 1/(1+e^1) = 0.2689. A one unit increase in either X1 or X2 will increase P(Y=1) to be closer and closer to 1.
4. Confusion Table
    a. FPs, FNs, TPs, TNs
        i. False Positives: 66, predicted positive but truly labeled negative
        ii. False Negatives: 150, predicted negative but truly labeled positive
        iii. True Positives: 45, predicted and truly labeled positive
        iv. True Negatives: 801, predicted and truly labeled negative
    b. Compute TPR and FPR
        i. True Positive Rate: [TP/(TP+FN)] = 45/(45+150) = 0.23
        ii. False Positive Rate: [FP/(FP+TN)] = 66/(66+801) = 0.076
    c. If we increase the decision threshold, less entries will be predicted as positive and more will be predicted as negative. This is because we increase the threshold at which the model predicts positive. Most likely, this will decrease the False Positives by a larger amount than the True Positives decrease, resulting in a slightly lower(or equal) TPR and a lower FPR.
5. Support Vector Machine
    a. If we were to remove one of the uncircled points, our SVM decision boundary should not change, as SVM examines just the points nearest to the decision boundary while making its calculations.
    b. Hard margins provide strict constraints to correctly classify every datapoint, maximizing the minimum distance from the decision boundary to the training points. It works only if the data is linearly separable and is very sensitive to outliers. Soft margins are not sensitive to outliers and can work even if the data is not linearly separable. In this case, it will not matter, as there are no outliers, the data is linearly separable, and the outlining points all lie at the same margin.
    c. In the Radial Basis Function kernel SVM, C is the inverse strength of regularization, so as C increases, the model becomes overfitted. It does this by accepting smaller margins for larger values of C, provided the decision function is better at classifying the training data. The gamma parameter defines how far the influence of a single training exercise reaches, with low values meaning far. Gamma can be seen as the inverse of the radius of influence of select vectors.
6. Augmentation

Devin Yerasi
UID: 305 167 818

a. One pair of features I would be interested in would be a cross term of neighborhood_group and room_type. I think price is heavily dependent on both the neighborhood_group it is in and the type of room offered. By combining these two features, I believe the model can generate better price ranges for booking prices depending on the neighborhood, as similar room types in the same area seem to have similar prices. Another pair of features I would cross is reviews_per_month and price. This will give a feature of the minimum average monthly revenue for a place. It is a minimum as not everyone who books the place will write a review, but everyone who has written a review has booked the place. It seems that places with a greater average monthly revenue will have more demand and incentives for the owner to keep the place nice, resulting in an increased price.
b. One could segment the geographical area into a set number of rectangular blocks based on latitude and longitude ranges and bucket the cross of latitude and longitude features into one of these rectangular blocks to create a discrete geographical feature. The granularity of these rectangular blocks would have to be played around with, but at least be at the block level.