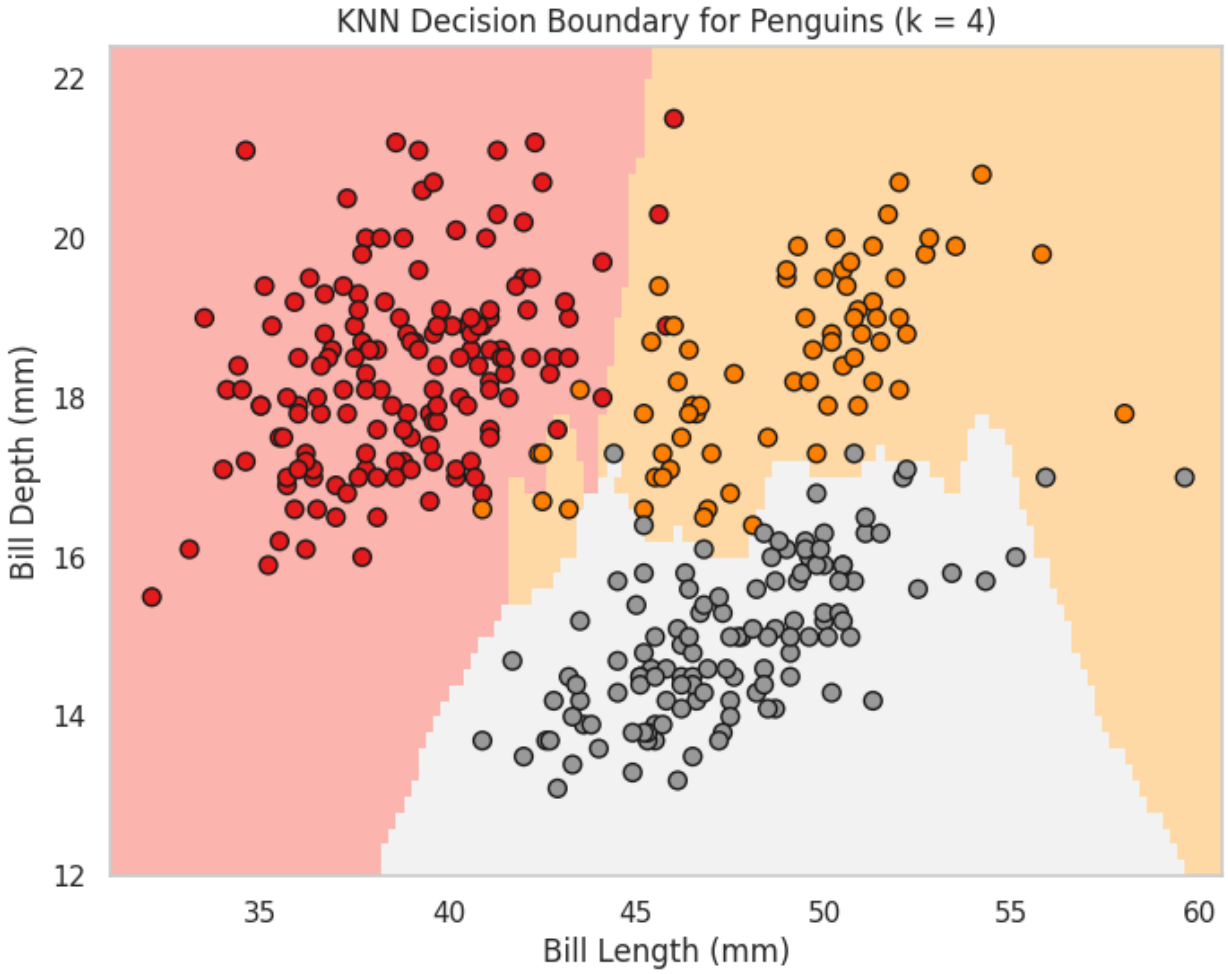
Jonathan Lopez

Homework 1: Analysis and Model Evaluation

K-Nearest Neighbors (KNN) with Penguins Dataset:

The objective was to use KNN to predict the penguin species based on bill length and bill depth. After loading the dataset and dropping rows with missing values, I encoded species into numeric values to fit the KNN classifier. After doing a 70/30 train-test split, GridsearchCV was used with cross-validation to find the optimal K value. After determining the best k of 4, model performance indicated strong accuracy with a score of 95%.

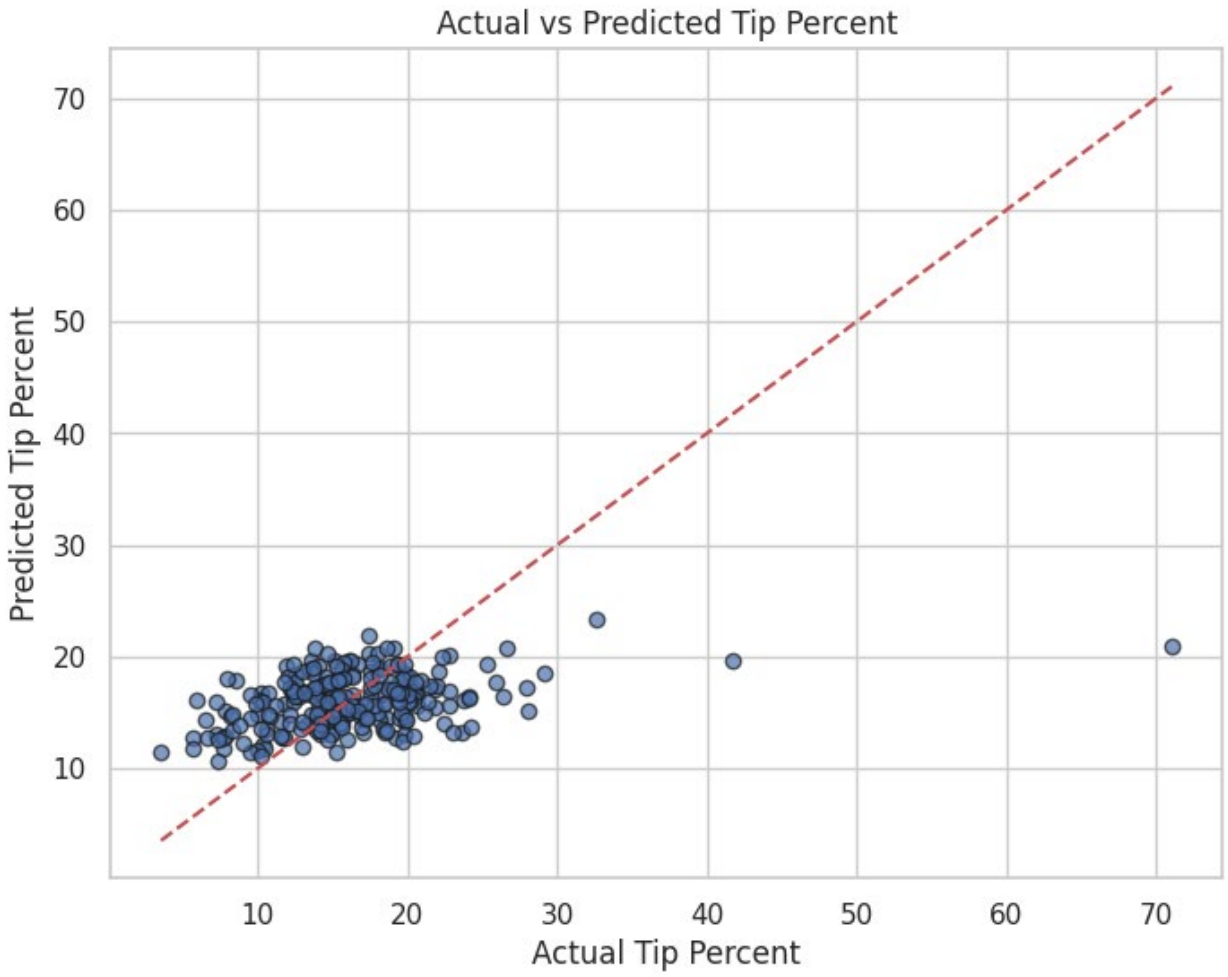


The grid plot shows the KNN decision boundaries separated by the three penguin species. This showed distinct areas for each species with minor overlapping. A K value of 4 yielded accuracy and effectively balanced bias and variance. The clearly defined boundaries visually confirm the accuracy of the KNN model.

Jonathan Lopez

Polynomial vs. Linear Regression on Tips Dataset:

The objective was to determine if a polynomial or linear model would be better in predicting the tip percentage based on total bill and party size. I calculated tip percentage as the tip amount divided by the total bill and then multiplied by 100. For fitting the models, degree 1 was used for linear regression and degree 2 for polynomial regression. When comparing the R^2 score, the linear model had .120 score and the polynomial model had a slightly higher score of .149. The underlying function for the linear model was tip_percent = 19.89 - .270(total_bill) + .598(size). For the polynomial model, tip_percent = 25.503 - .751(total_bill) + .069(size) + .008(total_bill^2) + .024(total_bill x size) + .006(size^2).

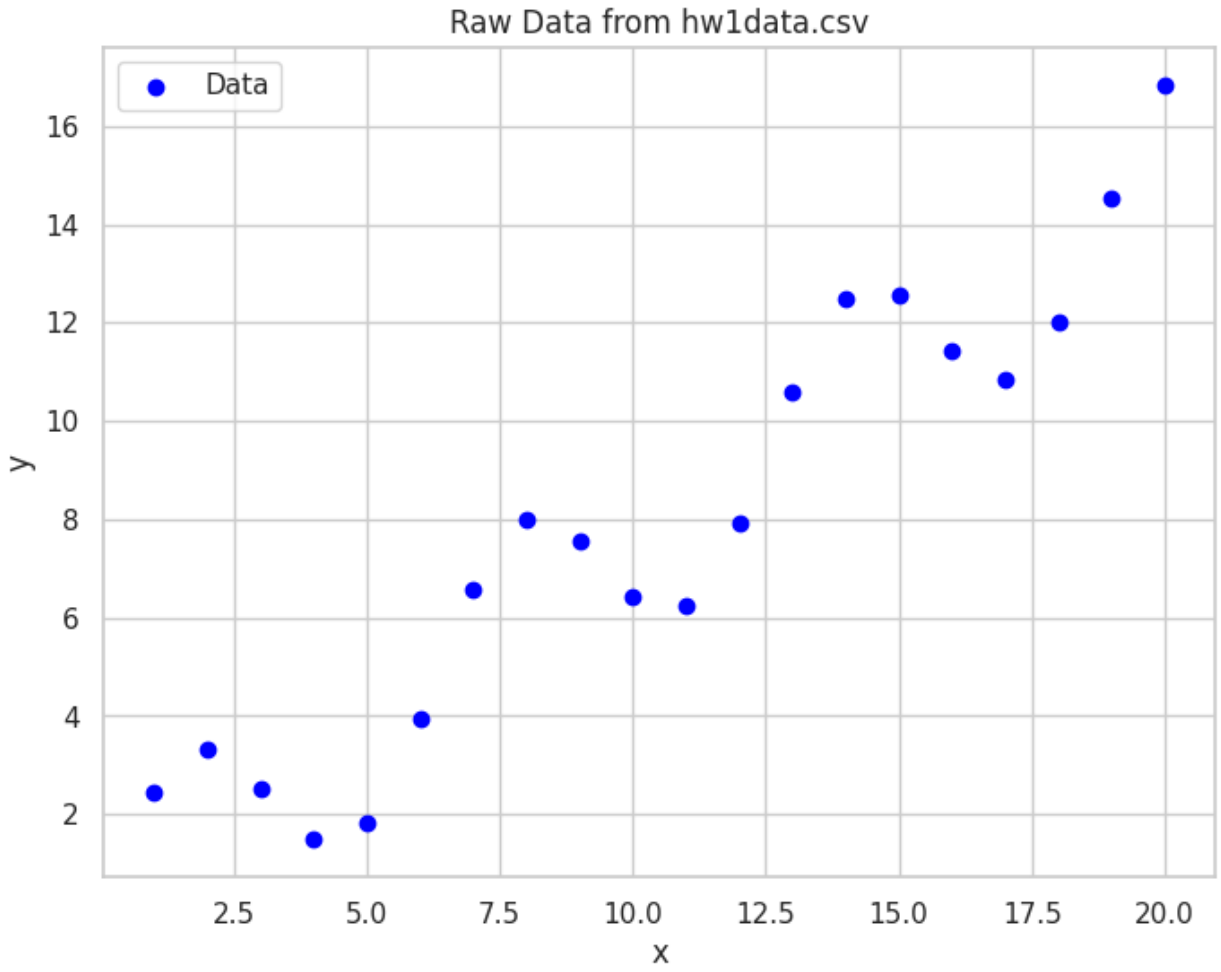


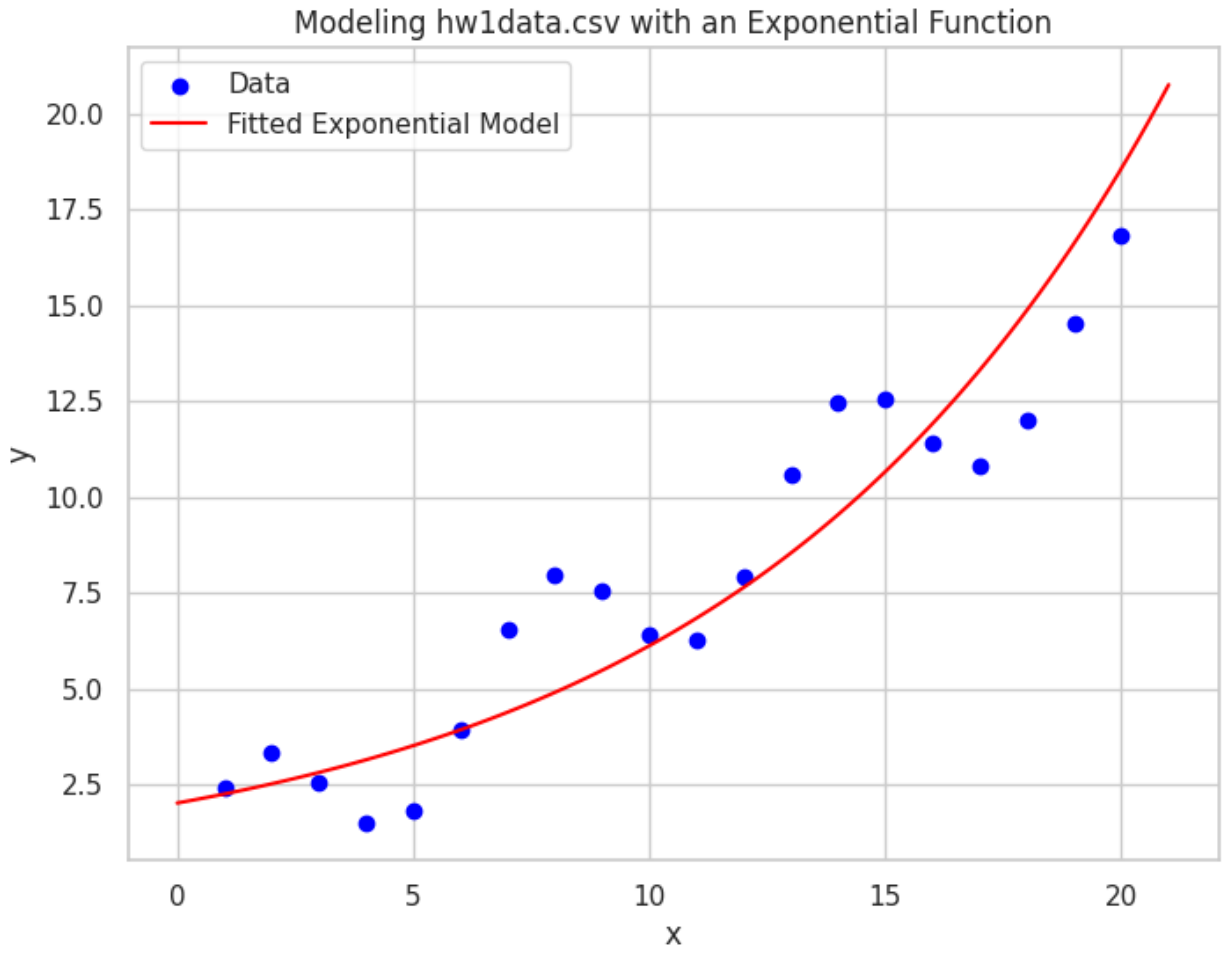
The plot shows the comparison of the actual vs predicted tip percentage, showing moderate predictive capabilities with noticeable outliers. Although the polynomial model slightly improved the R^2, the improvement (2.9) was not significant.

Jonathan Lopez

Linear Regression modeling hw1data after Transformation:

The objective was to analyze the underlying mathematical relationship in the provided data and construct a model. After plotting the data and observing a clear nonlinear relationship, log transformation was applied to y to linearize the data, then fitting a linear regression. Linear regression on log transformed the data yielding an intercept of 0.702 and a slope of 0.111, resulting in an underlying function of $y = e^{.702+.111xX}$.





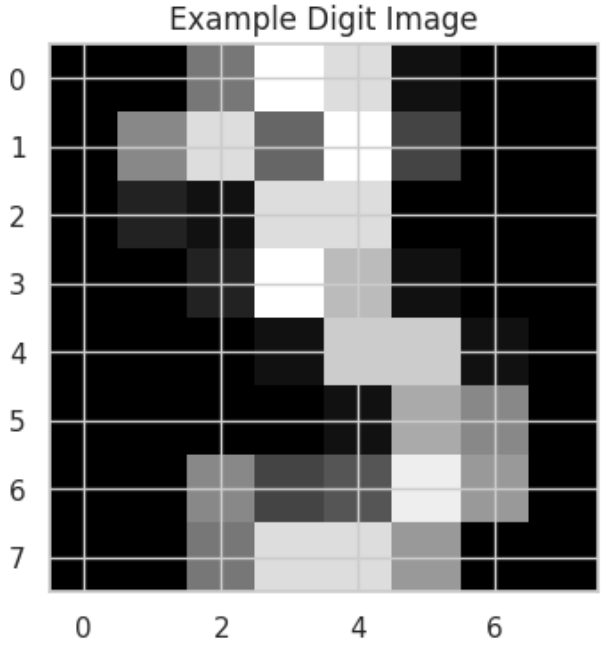
The original data showed exponential growth, and the transformed linear regression model effectively captured the exponential trend when plotted against original data points. The models predicted values beyond the original data range remained reasonable, confirming the model was valid. Overall, the exponential function appropriately describes the dataset, with the log-transform improving the linear regression models performance significantly.

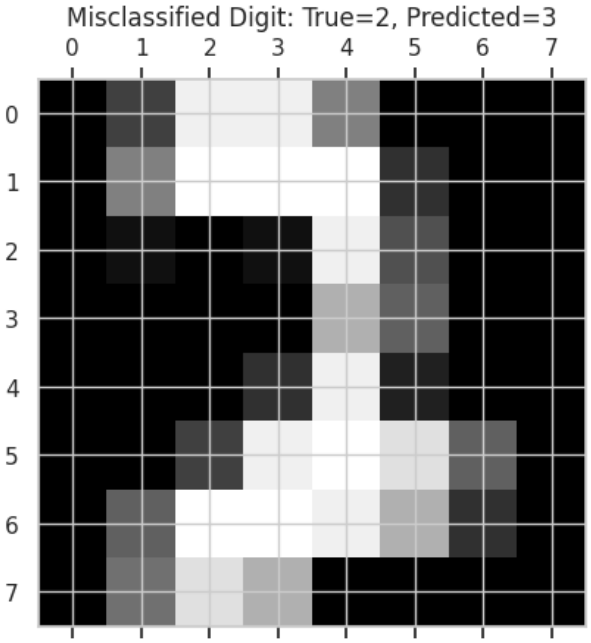
Logistic Regression on Diabetes Health Indicators Dataset:

The objective was to use BMI to predict whether a patient had diabetes. After loading the dataset, I converted diabetes categories into binary classes of 1 for diabetes and 0 for otherwise. After performing logistic regression using the BMI, the results for the coefficient was 0.078 and the intercept was -4.137. I proceeded to verify logistic regression with selection of BMI values of 31 and 31 and calculate the log-odds and odds ratio. The results revealed a log-odds at BMI 30: -1.782 and BMI 31: -1.704. The difference in log odds match up to the coefficient of 0.078. The odds ration (per unit BMI increase) was $e^{0.078}$= 1.082. Essentially each one unit increase in BMI increases the odds of diabetes diagnosed by 8.2%, which proves BMIs predictive importance.

Logistic Regression Classification of Digits Dataset:

The objective was to classify handwritten digit images using logistic regression. After splitting data into 75/25 split, I fitted a logistic regression classifier which resulted in a 97.3% accuracy score. Analysis of the misclassified numeral included a sample true digit 2 and predicted digit 3, with predicted probabilities of .518 for digit 3 and .476 for digit 2. The digits ambiguous shaped and overlapping features likely caused misclassification and low confidence suggests that the model reasonably identifies ambiguity.





Conclusion

In conclusion KNN on penguin data set provided an excellent prediction accuracy, clearly separating species based on beak measurements. The tips regression model however only slightly improved with polynomial regression when compared to linear regression. For the hw1 csv data, the linear regression on log-transformed data showed exponential growth. Performing logistic regression on BMI also proved to be a meaningful predictor of diabetes and proceeded to validate the log-odds and odds ratio. The digit classification demonstrated high accuracy but with some misclassification due to ambiguity.

References

https://www.w3schools.com/python/python_ml_grid_search.asp

https://www.geeksforgeeks.org/encoding-categorical-data-in-sklearn/