Final Project

1. **Exploratory Analysis:**

In the initial exploratory analysis of the “voles” dataset the I used the GGpairs() function to examine the distribution of data points for each of the two groups. As indicated in the boxplots shown for each variable, the variables that do not overlap could be potential candidates for becoming significant predictors in the created models. After examining the variables, the correlation between variables was examined next. The result was that many of the molar measurements were highly correlated, supporting either the acceptance of one of the correlated variables or fitting a mixed model to account for the “spatially correlated variables”, those being the molars’ proximity in the mouth.

1. **Modeling Section:**

Six models were fitted to assess different methods for predicting the vole species.

1. **Logistic Regression all variables (Kappa = 0.7780):** This model served as the baseline model to evaluate the performance of all other models.
2. **Logistic Regression (Kappa = 0.5925):** In this model, I manually removed variables that were not significant. The thought being it may perform well against a penalty model.
3. **Logistic Regression with Lasso (Kappa = 0.7400):** For this model, automatic variable selection was used which should lead to higher performance as opposed to manual variable selection.
4. **Rpart Tree (CART) (Kappa = 0.7385):** This model was used as a baseline for the bagged tree and random forest models.
5. **Bagged Tree (glmnet) (Kappa = 1.0000):** This model was used to increase complexity to determine the response by utilizing a bagging technique.
6. **Random Forest Tree (Kappa = 1.0000): T**he random forest model was utilized to see if there was value in adding random variables to the modeling strategy.
7. **Model Selection:** I settled on using the Kappa statistic for my analysis. My thinking was it was compensate for the variance found during my initial AUC calculations. Based on that statistic, I selected 3 models to pass through to the test data set. I chose the Lasso GLM, Bagged Tree, and Random Forrest models. I selected the Lasso GLM because I felt that the automatic variable selection feature of the model would allow for an ideal final model compared to the tree models. As for the Tree models, I selected the Bagged Tree and the Random Forest models. While both models had a perfect score (1.0000), I still wanted to pass the test data through both models to see how each would perform.
8. **Final Model:** I selected the Lasso GLM Model (Kappa = 0.7400) to make the final predictions. Based on the Kappa scores of the 3 models, the tree models over-fit the data, as shown by their lower scores on test data than the training dataset.