Mary Morkos

Dr. Davis

MTH 2151 Statistical Modeling

23 April 2024

Final Project Write-Up: Bird Classification

For the final project, I undertook a bird species classification task using machine learning techniques, aiming to predict bird species qualitatively and estimate Humerus Length quantitatively based on morphological features like the length of the humerus, ulna, femur, etc. The dataset was pre-processed and cleaned to prepare for prediction tasks. Tasks involved handling missing values by filling them with the mean of each numerical column, checking for duplicates, and dropping unnecessary columns. Additionally, visualizations including box plots and count plots were created to understand data distributions. For Prediction 1, the problem is to qualitatively predict bird species based on morphological features. The target variable is 'type' representing bird species, while predictors are morphological features. A Random Forest Classifier model was built and evaluated, achieving an accuracy score of 82.14%. Interpretation of results includes analyzing the confusion matrix to understand model performance across different bird species and identifying useful predictors contributing to model performance. In Prediction 2, the goal is to quantitatively predict Humerus Length. The target variable is 'huml', while predictors include other morphological features. A Linear Regression model was trained and evaluated, resulting in a Mean Squared Error (MSE) of 40.23 and an R-squared (R^2) Score of 0.988. Analysis involves interpreting where the model performs well and where it does not, identifying significant predictors influencing Humerus Length prediction. Cross-validation was used to validate model performance, ensuring robustness and generalization. Feature importance

analysis was conducted to identify crucial predictors, using Recursive Feature Elimination (RFE) for feature selection. Hyperparameter optimization was performed using grid search to tune model parameters. Visualizations were used to interpret feature importance and grid search results.