Executive Summary - Ensemble Species Distribution Models

Although, the literature focuses on the Maxent algorithm (Bicknell et al. 2017; Miller 2010; Phillips et al. 2004), algorithm ensembles that include RF, GLM and Maxent provide more conservative species distribution outputs than ensembles that include GLM, SVM and Maxent. The objective for the following project is to determine whether the RF -Maxent – GLM algorithm ensemble, or the Maxent – SVM – GLM algorithm ensemble provides the best species distribution model output. The following R packages were used to conduct the ensemble species distribution models: Stacked Species Distribution Models (SSDM) (https://cran.r-project.org/web/packages/SSDM/vignettes/SSDM.html#model_algorithms), raster and rgdal. The SSDM package allows for the combination of several species distribution model (SDMs) outputs, each using a different modeling algorithm. Ensemble species distribution models are the best choice for SDMs, because single algorithms can lead to uncertainty in species distributions, and a combination of algorithms allows for an agreement on outputs. In addition to the species distribution outputs, model accuracy assessment was used to determine ensemble accuracy, and which ensemble is the best fit for my capstone project goal. Environmental variable importance values were calculated to determine the mean environmental variable governing the species distribution models. Mean area under ROC curve (AUC) was used to determine test accuracy for the two different algorithm ensembles. Mean AUC for the Maxent – SVM – GLM algorithm ensembles for the 100 fish species subset was 95.93%. Mean AUC for the RF – Maxent – GLM algorithm ensembles for the 100 fish species subset was 98.23% (Table 1). A student's T-Test was showed that there was no significant difference between two algorithm ensembles when comparing test accuracy (p= 0.0887). For the Maxent – SVM – GLM algorithm ensemble, the environmental variable with the highest importance was the hydrography variable (hydrography_fac_bin) which had the mean importance of 40.69%. The next highest environmental variable for the SDMs created using the Maxent – SVM – GLM algorithm ensembles was flow accumulation (Guyana_fac) at 8.76%. For the RF – Maxent – GLM algorithm ensemble, the environmental variable with the highest importance was the hydrogeography variable at 30.64%. Flow accumulation was the variable with the next highest importance at 9.53%. Overall, both algorithm ensembles provide similar outputs and high accuracies with the same driving environmental variables. The RF – Maxent – GLM algorithm provides the highest accuracies and most concentrated SDMs (Figure 3), which are the desired results to implement these results into MARXAN analysis to determine priority conservation areas in Guyana.

Algorithm Ensemble	**Mean AUC (%)**	**Mean Sensitivity (True Positive Rate) (%)**	**Mean Specificity (True Negative Rate) (%)**	**Mean Omission Rate (%)**	**Proportion Correct (%)**	**Mean Kappa Coefficient (%)**
Maxent – SVM – GLM	95.93	95.6	94.8	5.4	95.0	69.6
RF – MAXENT – GLM	98.23	98.0	97.0	2.1	97.0	71.0

Table 1: Mean Confusion Matrix for the two algorithm ensembles used for species distribution models.

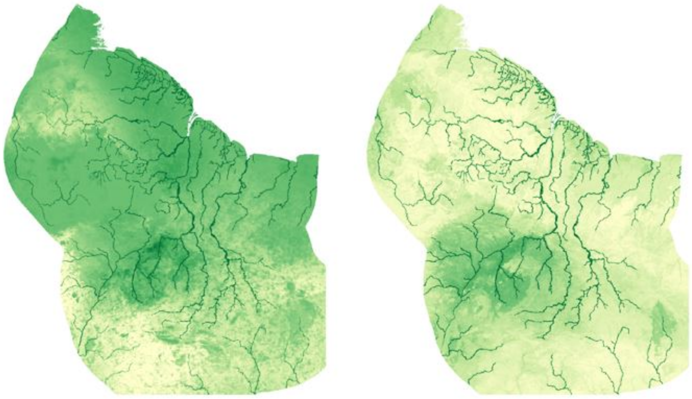

Figure 3: *Brycon falcatus* species distribution models. The map to the left shows the species distribution created by the Maxent – SVM – GLM algorithm ensemble. The map on the right shows the species distribution created by the RF – Maxent – GLM algorithm ensemble. Areas of bright great indicate the most likely occurrences of *Brycon falcatus.* Areas of yellow indicate areas where the species will least likely be.