

Random Forest Documentation for 'Darter Richness'

1. Modeling Procedure

We used classification Random Forests (RF) analysis to explore associations among SC Stream Assessment response variables and NFHAP spatial predictor variables (Breiman 2001, Cutler et al. 2007). Machine learning techniques such as RF provide an alternative modeling paradigm to traditional statistics, where no a priori model is defined, and complex data structures (non-normal distributions, interactions) are accommodated. Machine learning techniques use an algorithm to learn the relationship between the response and its predictors by identifying dominant patterns in the dataset (Breiman 2001, Elith et al. 2008). Random Forests represent an advance in machine learning techniques that have increased the accuracy and prediction power of single classification and regression trees by the creation of an ensemble of trees (Breiman 2001). Random forests are non-parametric, can handle both categorical and continuous data as either predictor and/or response variables, can handle high-order interactions, are insensitive to outliers, and can accommodate missing data by using surrogates (Breiman 2001, De'ath and Fabricius 2000, Urban 2002). Categorical Random Forests fit an ensemble of trees to a dataset, where each individual tree in the forest is built using a randomly selected bootstrap sample of the training dataset. In addition, only a random subset of predictor variables is considered for node and splitpoint selection (Amit and German 1997). In this way, two elements of randomness are injected into the procedure. Observations not included in the bootstrap samples are passed down their respective trees, and each tree's terminal nodes contain a predicted categorical response to different combinations of observed values among predictor variable pathways. Each tree has a 'vote' in the most important predictive variables to split on, and on the categorical responses of different values of input combinations; and the majority of votes among the ensemble of trees 'wins'. Therefore, we can a) predict and rank variables that most strongly influence an outcome (variable importance plot), and b) isolate and examine the behavior of individual predictors on the outcome, while holding the effect of all other predictive variables constant (partial dependence plots).

RF modeling was conducted by building 5000 trees using default values for other parameters in the randomForest package in the R programming environment (R Core Team 2012). RF models have known biases in variable importance selection for highly correlated predictor variables; therefore we conducted a preliminary screening of our abiotic variables to eliminate highly correlated variables. Correlations remaining in models are listed in documentation item 6.

Literature Cited

Amit, Y., and D. German. 1997. Shape quantization and recognition with randomized trees. *Neural Computation* **9**:1545-1588.

Breiman, L. 2001. Random Forests. *Machine Learning* **45**:5-32.

Cutler, D., T. Edwards, K. Beard, A. Cutler, K. Hess, J. Gibson, and J. Lawler. 2007. Random forests for classification in ecology. *Ecology* **88**: 2783 – 2792.

De'ath, G., and K.E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* **81**:3178-3192.

Elith, J., Leathwick, J. R., Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* **77**:802-813.

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

Urban, D. L. 2002. Classification and regression trees. *in* B. a. J. B. G. McCune, editor. Analysis of Ecological Communities. MjM Software Design, Oregon

2. Error Estimates Procedure

The RF algorithm builds trees based on repeated randomized samples of the dataset, hence it is not essential to hold back data for testing after model creation to obtain an unbiased estimate of error. Model performance was evaluated with three accuracy measures calculated using the resubstitution method (Theodoridis and Kourtroubas 2006). The three measures were the Proportion Correctly Classified (PCC), Cohen's weighted Kappa statistic (weighted κ), and the area under the receiver operating curve (AUC). Both PCC and weighted κ are derived from the model confusion matrix, which gives the number of actual versus predicted classifications of group membership. PCC performance measures are given in two forms: 1) an overall PCC percentage (accuracy) representing the number of correctly classified cases divided by the total number of cases across all outcome classes, and 2) a measure of accuracy for a specific outcome class (precision). Weighted κ corrects the overall PCC for agreement caused by chance, and gives a value ranging from -1 to 1 (Cohen 1968). A positive value indicates greater agreement between modeled and measured classifications than expected by chance alone, and a negative value indicates less agreement than expected by chance alone (Table 2). Cohen's weighted κ was calculated using the vcd package in R (R Core Team 2012). The AUC is derived from plotting the true positive rate (sensitivity) against the false positive rate (specificity), with each point plotted representing a sensitivity/specificity pair. The area under the resulting plot is a measure of how well the model correctly classifies groups. AUC values range from 0 to 1, with values > 0.5 indicating better model performance than expected by chance alone (Swets 1988). We used the ordROC function in the nonbinROC R package in R to calculate AUC values (<http://cran.rproject.org/web/packages/nonbinROC/index.html>).

Literature Cited

Cohen, J. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 70:213

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. Science **240**:1285-1293.

Theodoridis, S., and K. Kourtroubas. 2009. Pattern Recognition. 4th edition. Elsevier, New York.

3. Response Variable Definition

Darters dominantly utilize benthic habitats to complete all of their life history requirements. Due to their specificity for feeding, reproducing and seeking shelter in benthic habitats, darters are particularly sensitive to the degradation of benthic stream habitats. Benthic habitats become degraded from the effects of siltation, flow modification, and reduction in dissolved oxygen levels from the accumulation of organic matter. Streams with reduced numbers of darter species may reflect degraded ecological conditions.

Fish data used to calculate this metric were collected as part of the South Carolina Stream Assessment. This metric was calculated using data from 167 Wadeable Freshwater Stream Sample Locations in South Carolina's upstate, a region which includes four major drainage basins (Catawba/Watauga Rivers, Broad River, Saluda River, Savannah River), and four Level IV ecoregions (Blue Ridge, Inner Piedmont, Outer Piedmont, Slate Belt; Omernik 1987). Darter Richness is initially calculated as a count (richness) of darter species at a given sample location. Several subsequent steps were taken to transform Darter Richness into a three-category metric, with designations for 1) low darter richness, 2) intermediate darter richness, and 3) high darter richness. We first had to account for the fact that streams with larger drainage areas have naturally greater species richness than streams with smaller drainage areas. To incorporate this trend in metric scoring, we developed Maximum Species Richness (MSR) graphs. MSR graphs were created by plotting Darter Richness against the log (base 10) transformed values of the corresponding drainage basin area. Lines delineating the 95th and 5th percentiles (where allowed) were drawn. The area between the two lines was trisected based on a method developed by Lyons (1992). Data points falling above the middle trisection scored 'high', those falling within the middle trisection scored 'intermediate', and those falling below the middle trisection scored 'low' (Figure 1).

Literature Cited

Lyons, J. 1992. Using the index of biotic integrity (IBI) to measure environmental quality in warmwater streams of Wisconsin. U.S. Department of Agriculture, Forest Service, General Technical Report NC-149. St. Paul, Minnesota: U.S. Department of Agriculture, Forest Service, North Central Forest Experiment Station. 51p.

Omernik, J. M. 1987. Ecoregions of the conterminous United States. *Annals of the Association of American Geographers* 77(1): 118-125.

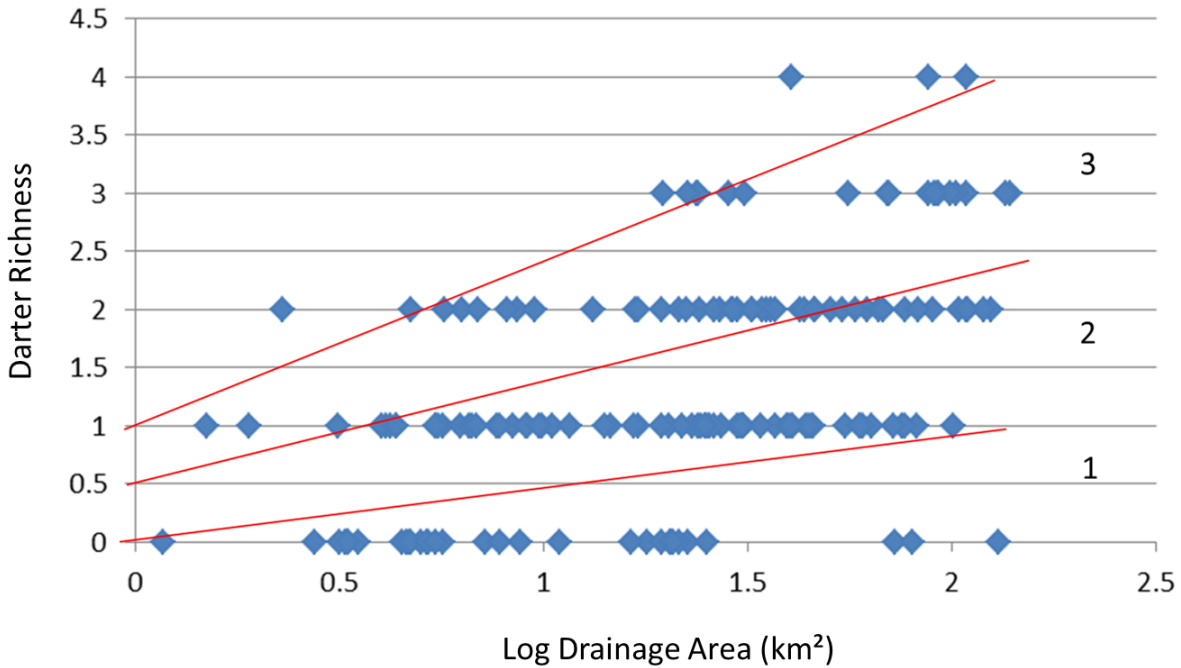


Figure 1. MSR graph for Darter Richness. Number of darter species plotted against the log (base 10) transformed value of the drainage basin area (km²). N=167. Numbers along right side of figure denote score values (classification values) for different regions of the plot.

4. Variables Retained in Model

[1] "AreasqkmC"	"Dam_coC"	"C_H2O_01"	"C_URBAN_01"
[5] "C_BARREN_01"	"C_FOREST_01"	"C_GRASSHRUB_01"	"C_WETLAND_01"
[9] "C_ROWOCROP_01"	"SLOPE"	"SOILHYGRP"	"SOILPERM"
[13] "ELEV_MEAN"	"U_Score_PercRich"		

5. Model Call

```
* randomForest(formula = U_Score_PercRich ~ ., data = PCS[, -1], keep.forest = TRUE, importance =
TRUE, ntree = 5000, do.trace = TRUE, type = 2)
Type of random forest: classification
Number of trees: 5000
No. of variables tried at each split: 3
```

* The original call to randomForest (R package 'randomForest'; v4.5-36)

6. Correlations Remaining in Model

"i= 1 variables C_URBAN_01 and C_FOREST_01 correlation= -0.702056051893182"

"i= 2 variables AreasqkmC and Dam_coC correlation= 0.623552895411633"

"i= 3 variables SLOPE and ELEV_MEAN correlation= 0.610697335060513"

"i= 4 variables C_BARREN_01 and C_GRASSHRUB_01 correlation= 0.508999880546235"

"i= 5 variables C_FOREST_01 and SLOPE correlation= 0.500785079845662"

"i= 6 variables C_URBAN_01 and C_GRASSHRUB_01 correlation= -0.433818381241452"

"i= 7 variables C_WETLAND_01 and C_ROWOCROP_01 correlation= 0.410295771088565"

"i= 8 variables C_BARREN_01 and C_WETLAND_01 correlation= 0.382913116254108"

"i= 9 variables C_WETLAND_01 and SLOPE correlation= -0.370434112992506"

"i= 10 variables SOILHYGRP and ELEV_MEAN correlation= -0.358361891710235"

7. List of Important Variables (Mean Decrease Accuracy)

Table 1. List of important variables in descending order of importance. Table denotes mean decrease in accuracy, and a weighting factor based on the percentage of importance explained relative to the most important predictor variable (Dam_coC).

Predictor Variable	MeanDecreaseAccuracy	Weighting Factor
Dam_coC	11.724	1.000
SOILHYGRP	11.123	0.949
C_BARREN_01	11.094	0.946
C_ROWOCROP_01	11.049	0.942
AreasqkmC	10.265	0.876
SOILPERM	10.155	0.866
C_WETLAND_01	9.192	0.784
C_URBAN_01	8.681	0.740
C_GRASSHRUB_01	8.269	0.705
C_H2O_01	8.138	0.694
C_FOREST_01	6.653	0.568
SLOPE	5.166	0.441
ELEV_MEAN	3.164	0.270

8. Variable Importance Plot

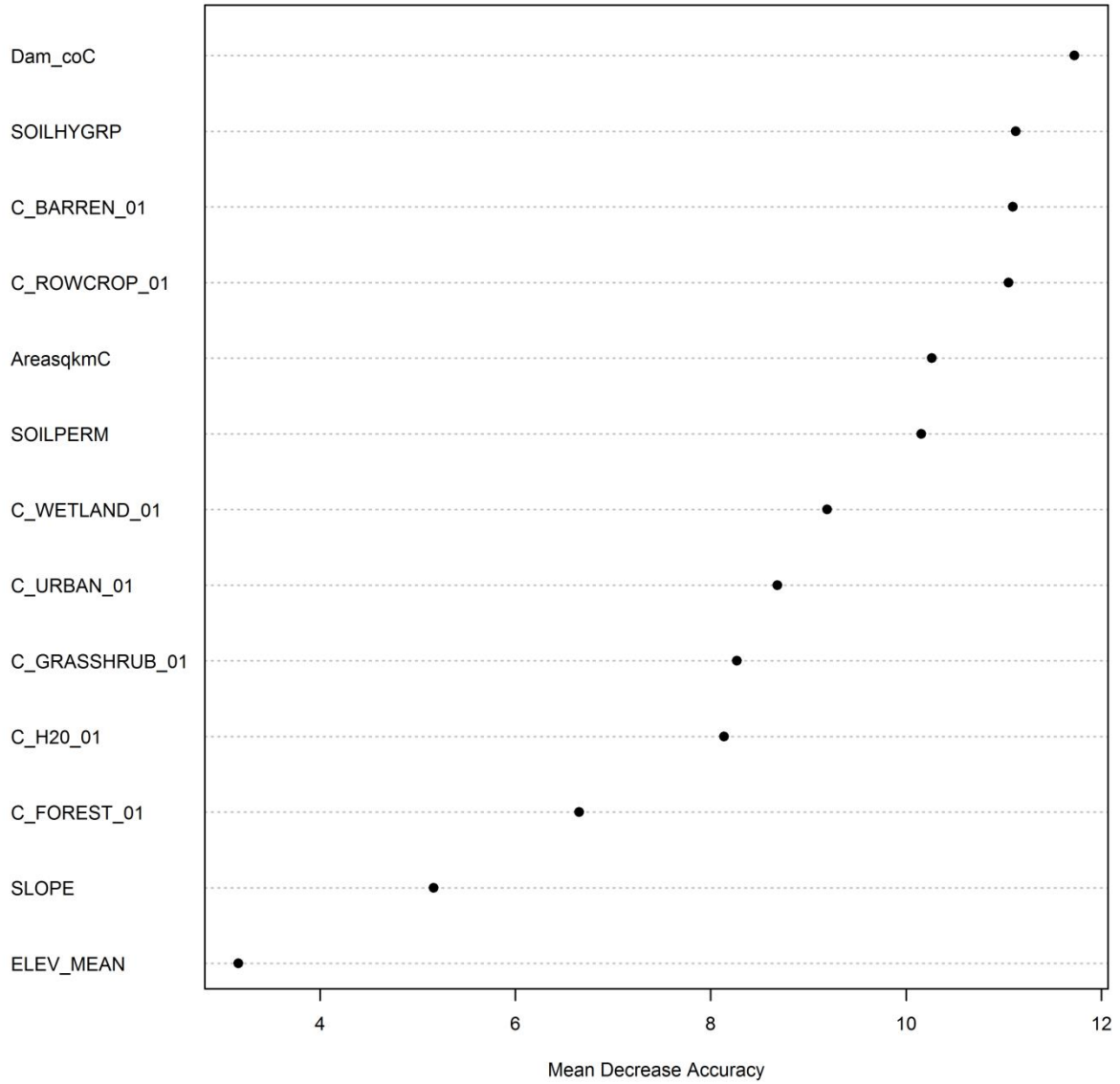
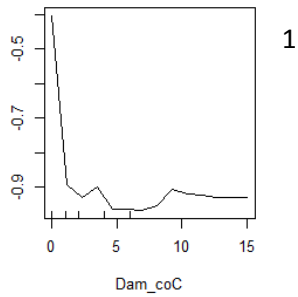


Figure 2. Top ranked variables from random forests classification for predicting Darter Richness in Upstate South Carolina.

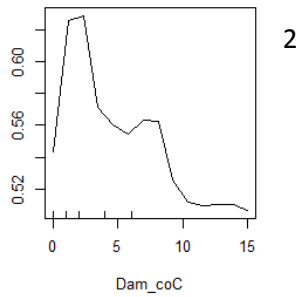
9. Partial Dependence Plots

Partial dependence plots isolate and examine the relationships between top ranked predictors and Darter Richness (1=low darter richness, 2=intermediate darter richness, 3=high darter richness) while holding the effect of all other predictive variables constant.

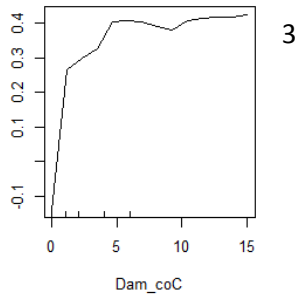
Partial Dependence on Dam_coC



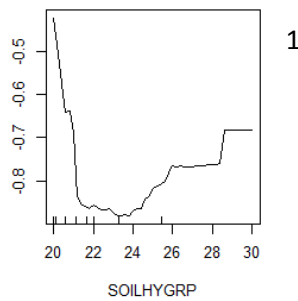
Partial Dependence on Dam_coC



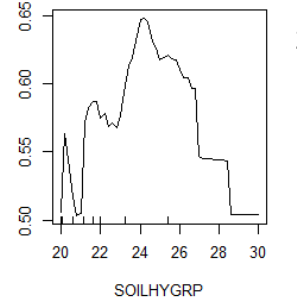
Partial Dependence on Dam_coC



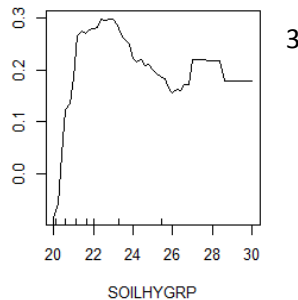
Partial Dependence on SOILHYGRP



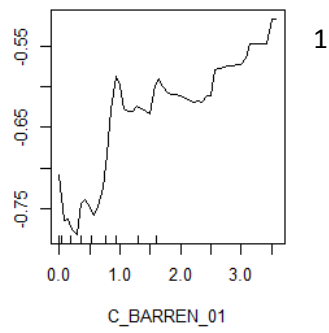
Partial Dependence on SOILHYGRP



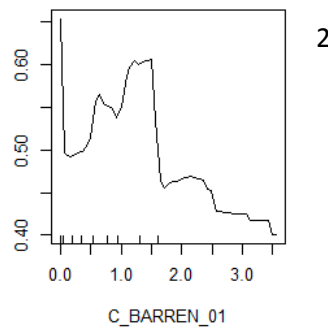
Partial Dependence on SOILHYGRP



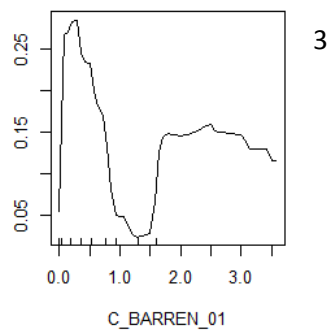
Partial Dependence on C_BARREN_01



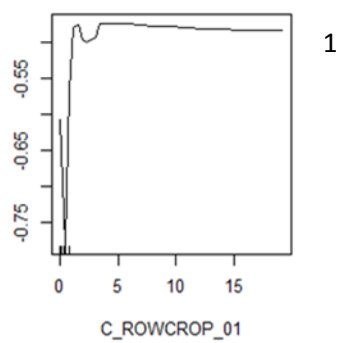
Partial Dependence on C_BARREN_01



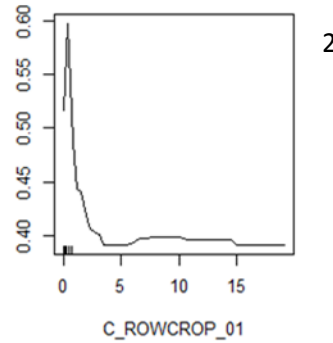
Partial Dependence on C_BARREN_01



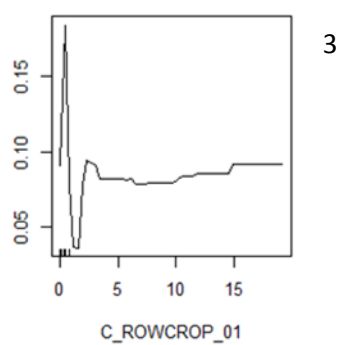
Partial Dependence on C_ROW CROP_01



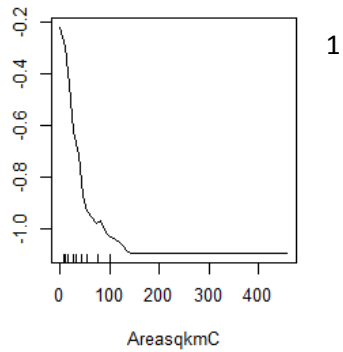
Partial Dependence on C_ROW CROP_01



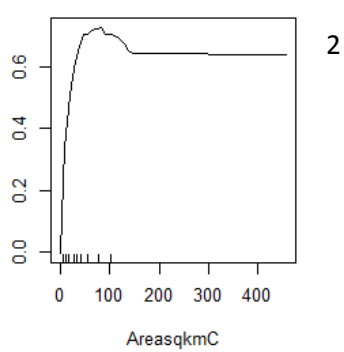
Partial Dependence on C_ROW CROP_01



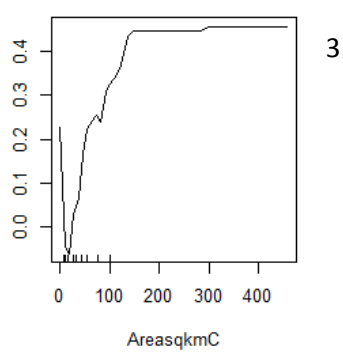
Partial Dependence on AreasqkmC



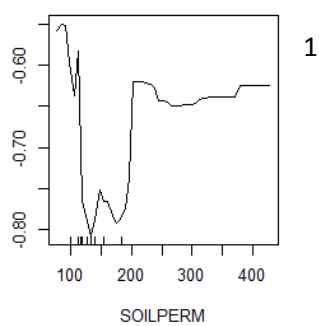
Partial Dependence on AreasqkmC



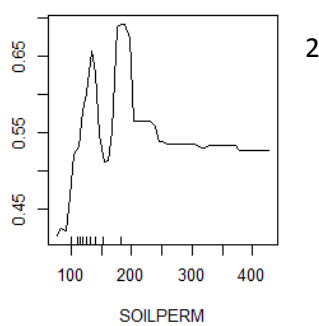
Partial Dependence on AreasqkmC



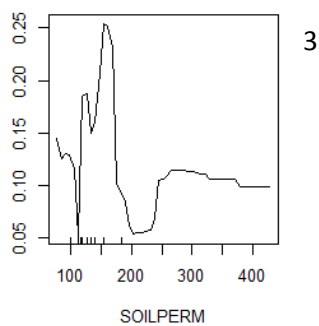
Partial Dependence on SOILPERM



Partial Dependence on SOILPERM



Partial Dependence on SOILPERM



10. Error Estimate: PCC Result and Confusion Matrix

Percent Correctly Classified: 45.51 %
Group 1 Accuracy: 32.35 %
Group 2 Accuracy: 59.74 %
Group 3 Accuracy: 33.93 %

Confusion Matrix			
	1	2	3
1	11	21	2
2	10	46	21
3	1	36	19

11. Error Estimate: Weighted K Result

$wK = 0.20$ (Slight Strength of Agreement)

12. Error Estimate: AUC Result

$AUC = 0.62$ (Model performance better than expected by chance alone)