Analysing features importance for identification of tiger beetles using machine learning

D. L. Abeywardhana[1], C. D. Dangalle[1]*, Anupiya Nugaliyadde[2], Y.W. Mallawarachchi[3]

[1]Department of Zoology and Environment Sciences, University of Colombo, Sri Lanka.
[2]Murdoch University, Perth, Australia.
[3]Sri Lanka Institute of Information Technology, Malabe, Sri Lanka.

Performance of machine learning models mainly rely on the quality of the input data fed into the model. Therefore, using all of the features/attributes in a dataset as input data may have a negative effect rather than a positive effect on the resulting model causing increase of training time and to model over-fitting. The present study was conducted to identify the most suitable features that can be used in a machine learning model developed to identify ground-dwelling tiger beetle species. As input data, habitat and morphometric data of tiger beetles collected from 2002 – 2017 from various locations of Sri Lanka were used. The data set comprised of 468 records with 12 features of 14 species. Each specimen collected was considered as a single record of the dataset, and climatic zone, GPS co-ordinates of location, habitat type, elevation, air temperature, solar radiation, relative humidity, wind speed, soil moisture, soil salinity, soil pH and body length of the specimen were considered as features. The dataset was pre-processed and fed into various algorithms: KNN, SVM, Naïve Bayes, Ensemble Extra Trees Classifier. From above, Ensemble Extra Trees Classifier yielded a test accuracy of 85.35% and was selected as the most suitable algorithm. Therefore, Ensemble Extra Trees Classifier was selected to evaluate the hierarchical importance of the features of the current dataset. The study revealed that body length, habitat type and elevation of the locations were the three most informative features in the dataset which supported species identification. However, using a fewer number of attributes which provide higher feature importance values reduced classification accuracy. The main reason for above scenario was that features except body length were more or less similar and had slight variation while body length had high variation that results in overfitting of the machine learning model. In order to prevent overfitting and increase validation accuracy combining all the features is necessary.

Keywords: Ensemble Extra Tree Classifier, feature importance, tabular data, tiger beetle dataset
* cddangalle@gmail.com

Acknowledgement: National Science Foundation of Sri Lanka (Grant No. RG/2017/EB/01).