

EGU21-9382

<https://doi.org/10.5194/egusphere-egu21-9382>

EGU General Assembly 2021

© Author(s) 2022. This work is distributed under the Creative Commons Attribution 4.0 License.



## Enhancing accuracy and interpretability of machine learning models using super learning and permutation feature importance techniques in digital soil mapping

**Ruhollah Taghizadeh-Mehrjardi**<sup>1,2,3</sup>, Nikou Hamzehpour<sup>4</sup>, Maryam Hassanzadeh<sup>4</sup>, Karsten Schmidt<sup>5</sup>, and Thomas Scholten<sup>1,2</sup>

<sup>1</sup>Eberhard Karls University Tübingen, Institute of Geography, Soil Science and Geomorphology, Tübingen, Germany (ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de)

<sup>2</sup>CRC 1070 ResourceCultures, University of Tübingen, Gartenstr. 29, Tübingen, Germany

<sup>3</sup>Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran

<sup>4</sup>Soil Science Department, Faculty of Agriculture, University of Maragheh, Maragheh, Iran

<sup>5</sup>eScience Center, University of Tübingen, 72070 Tübingen, Germany

The digital soil mapping (DSM) approach predicts soil characteristics based on the relationship between soil observations and related covariates using machine learning (ML) models. In this research, we applied a wide range of machine learning models (12 base learners) to predict and map soil characteristics. To enhance accuracy and interpretability we combined the base learner predictions using super learning strategy. However, a major problem of using super learning and complex models is that the explicit share of individual covariates persons in the overall result cannot be explicitly quantified. To overcome this restriction and make the super learning models interpretable, we employed model-agnostic interpretation tools, for example, permutation feature importance. Particularly, we integrated the weight assigned to each ML base learner obtained by super learning and the ranked ML base learner's covariates obtained by permutation feature importance to explore the contribution of covariates on the final prediction. We tested our super learning and permutation feature importance techniques to predict and mapping physicochemical soil characteristics of Urmia Playa Lake (UPL) sediments in Iran. As expected, our results indicated that super learning could significantly improve the ML accuracies for predicting soil characteristics of single base learners. In terms of root mean square error, super learning improved over the performance of the linear regression by an average of 45.7%. Furthermore, the permutation feature importance allowed us to interpret our results better and prove the significant contribution of geomorphological features and groundwater data in predicting soil characteristics of UPL sediments.