Data analysis and machine learning have become an integrative part of the modern scientific methodology, offering automated procedures for the prediction of a phenomenon based on past observations, unraveling underlying patterns in data and providing insights about the problem. Yet, caution should avoid using machine learning as a black-box tool, but rather consider it as a methodology, with a rational thought process that is entirely dependent on the problem under study. In particular, the use of algorithms should ideally require a reasonable understanding of their mechanisms, properties and limitations, in order to better apprehend and interpret their results.

Accordingly, the goal of this thesis is to provide an in-depth analysis of random forests, consistently calling into question each and every part of the algorithm, in order to shed new light on its learning capabilities, inner workings and interpretability. The first part of this work studies the induction of decision trees and the construction of ensembles of randomized trees, motivating their design and purpose whenever possible. Our contributions follow with an original complexity analysis of random forests, showing their good computational performance and scalability, along with an in-depth discussion of their implementation details, as contributed within Scikit-Learn.

In the second part of this work, we analyze and discuss the interpretability of random forests in the eyes of variable importance measures. The core of our contributions rests in the theoretical characterization of the Mean Decrease of Impurity variable importance measure, from which we prove and derive some of its properties in the case of multiway totally randomized trees and in asymptotic conditions. In consequence of this work, our analysis demonstrates that variable importances as computed from non-totally randomized trees (e.g., standard Random Forest) suffer from a combination of defects, due to masking effects, misestimations of node impurity or due to the binary structure of decision trees.

Finally, the last part of this dissertation addresses limitations of random forests in the context of large datasets. Through extensive experiments, we show that subsampling both samples and features simultaneously provides on par performance while lowering at the same time the memory requirements. Overall this paradigm highlights an intriguing practical fact: there is often no need to build single models over immensely large datasets. Good performance can often be achieved by building models on (very) small random parts of the data and then combining them all in an ensemble, thereby avoiding all practical burdens of making large data fit into memory.