Co-financed by the Connecting Europe
Facility of the European Union

xAIM

Introduction

Trustworthy AI - Lecturer: Emanuela Raffinetti; Python instructor: Alex Gramegna
E-mail: emanuela.raffinetti@unipv.it; alex.gramegna01@universitadipavia.it

Contents and assessment

The module aims at providing:

- an overview of the main Machine and Statistical Learning models, both on the theoretical and practical view point;
- specific criteria and methods will be discussed, bsed on the recent research developments, in order to fulfill the requirement of trustworthy Artificial Intelligence (AI).

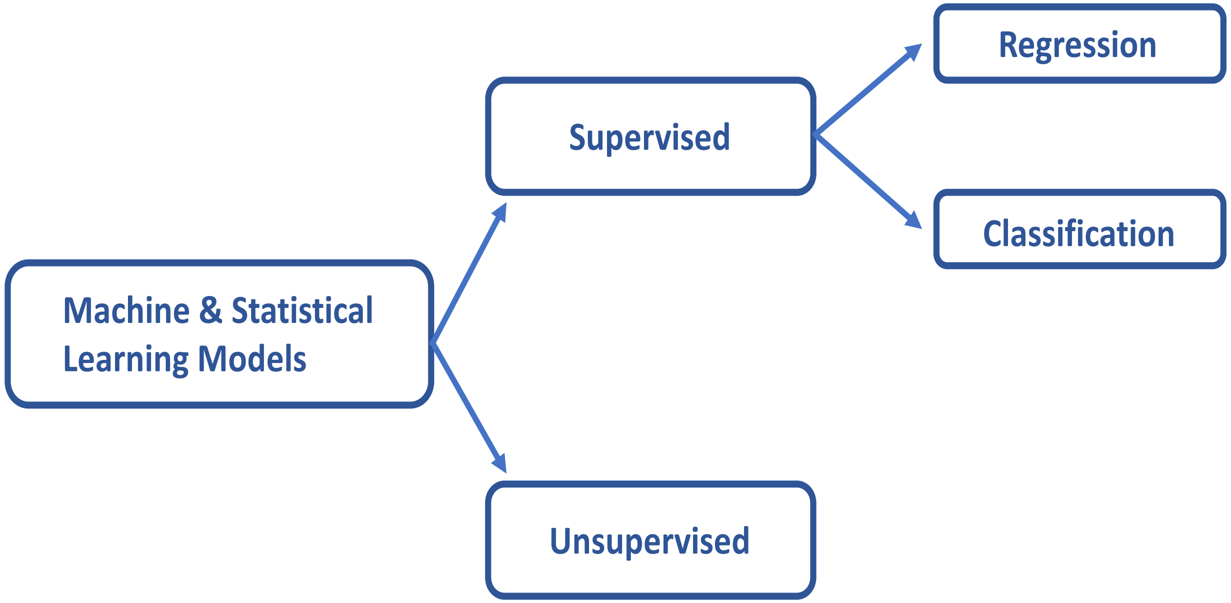
The examination procedure is addressed to evaluate the skills in interpretating the results of a Python output and in adapting the technical tools to real case-studies.

Co-financed by the Connecting Europe Facility of the European Union

Syllabus

SESSIONS (PRACTICAL PART)	TOPICS
SESSION 1 (2h) - Recorded	• Introduction • Linear regression models
SESSION 2 (1h) - Recorded	• Logistic regression models
SESSION 3 (2h) - Recorded	• Model selection
SESSION 4 (2h) - Recorded	• Neural network and tree models
SESSION 5 (2h) - Recorded	• Python Practice 1 • Statistical and Machine Learning Models
SESSION 6 (2.5h) - Live	• ***Explainability*** of Artificial Intelligence methods • Approaches to evaluate Explainability
SESSION 7 (2h)	• First reading: ***“Giudici P., Raffinetti E.: Lorenz Model Selection, Journal of Classification, 37(3), pp 754-768 (2020)”*** • Reflect on the content and prepare a brief report on a specific issue, problem and how to manage it. The report has to be delivered by the date of the last session.
SESSION 8 (2h) - Recorded	• ***Accuracy*** of Artificial Intelligence methods • Approaches to evaluate Accuracy
SESSION 9 (2.5h) - Live	• Python Practice 2 • Explainability and Accuracy of AI
SESSION 10 (2h)	• Second reading: “***Giudici P., Raffinetti E.: Explainable AI methods in cyber risk management, Quality and Reliability Engineering International, 38(3), pp 1318-1326 (2022)”*** • Reflect on the content and prepare a brief report on a specific issue, problem and how to manage it. The report has to be delivered by the date of the last session.
SESSION 11 (2h) - Recorded	• S.A.F.E. Artificial Intelligence • Metrics to evaluate AI safety
SESSION 12 (2h) - Recorded	• Python Practice 3 • S.AF.E. AI

Machine and Statistical Learning Models



- Supervised models are characterised by the presence of a target variable to be predicted.
- The main purpose of the supervised models is to derive predictions which are function of specific factors (explanatory variables).
- Examples of supervised models are: linear and logistic models, neural networks, tree models.

- Unsupervised models are not built on the presence of a target variable to be predicted.
- The main purpose of the supervised models is to discover the structure of data, by finding patterns from input data without references to outcomes.
- An example of supervised models is the cluster analysis.

Regression models

In regression models, the target variable is continuous.

The class of regression models includes:

- linear regression models: finding a line that best fits the data (simple linear regression); finding a plane that best fits data (multiple linear regression);
- regression trees: each variable represents a node of the tree. Based on the last nodes (leaves of the tree), a decision is made.
- random forest models: building multiple decision trees based on bootstrapped datasets of the original data and randomly selecting a subset of variables at each step of the decision tree. The mode of all the predictions of each decision tree is selected.
- neural network models: building a network of mathematical equations involving input variables and resulting in output variables.

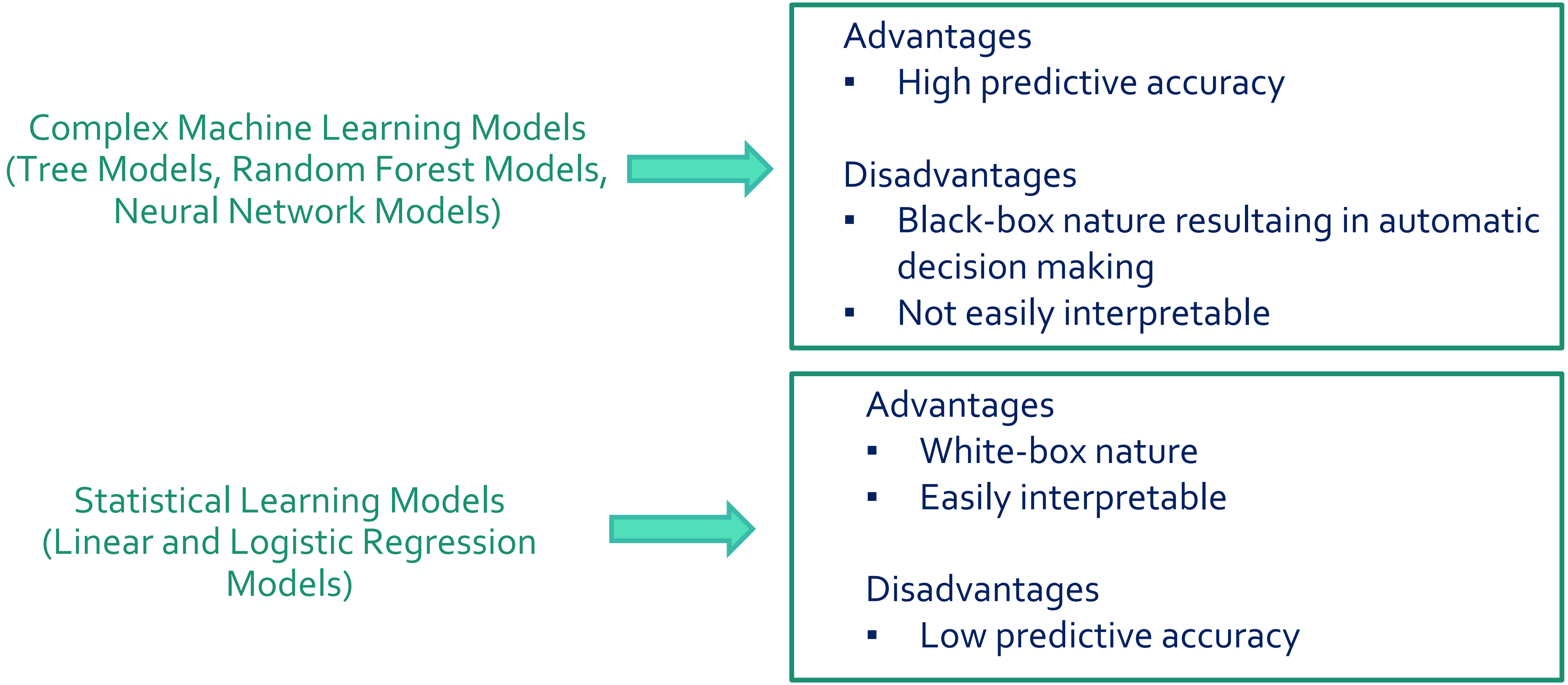
Classification models

In classification models, the target variable is discrete.

The class of classification models includes:

- logistic regression models: extension of the linear regression models used to model the probability of a finite number of outcomes;
- classification trees: an extension of the regression tree when the target variable is discrete;
- random forest models and neural network models can be equally used in the case of a discrete target variable.

Accuracy vs Explainability



Complex Machine Learning Models (Tree Models, Random Forest Models, Neural Network Models)

Advantages

- High predictive accuracy

Disadvantages

- Black-box nature resultaing in automatic decision making
- Not easily interpretable

Statistical Learning Models (Linear and Logistic Regression Models)

Advantages

- White-box nature
- Easily interpretable

Disadvantages

- Low predictive accuracy

How to make AI trustworthy

Recently, the European Commission has proposed a new regulation for a trustworthy AI.

In order to be trustworthy, AI methods have to fulfill four main key-principles:

- ***Sustainability***: AI methods should be robust, with respect to variations in terms of data and computations;
- ***Accuracy***: AI mehods should lead to accurate predictions;
- ***Fairness***: AI mehods should not discriminate by age, ethnicity, gender or other population groups;
- ***Explainability***: AI methods should be human interpretable in terms of its drivers

Along the module, the approaches and metrics to evluate the SAFEty of AI methodologies will be introduced and discussed to fulfill the basic requirements of Sustainability, Accuracy, Fairness and Explainability.