Application of Machine Learning for Stellar Astronomy

By Jan van Roestel

Machine learning

“Machine learning” is a collection of methods (algorithms) that perform a function (classification, regression, clustering, ranking,...), but of instead of being programmed by a human, are optimized using the data itself.

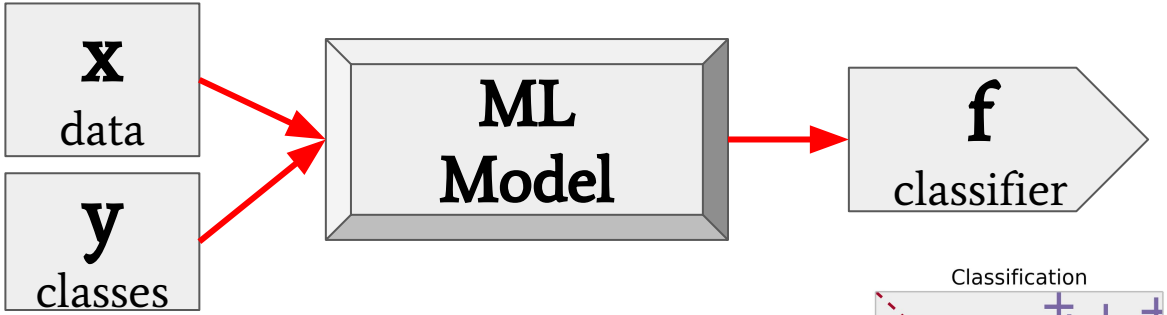
Machine learning can be separated into two types:

1. **Supervised machine learning**
   A subset of the data is known (**‘training data’**) and is used to create the ML-model. This is the most common type (in astronomy)
2. **Unsupervised machine learning**
   Nothing is known about the data; the ML-model is created from all data. Examples are principal component analysis (PCA), clustering, data-compression.

Supervised Machine Learning Classification

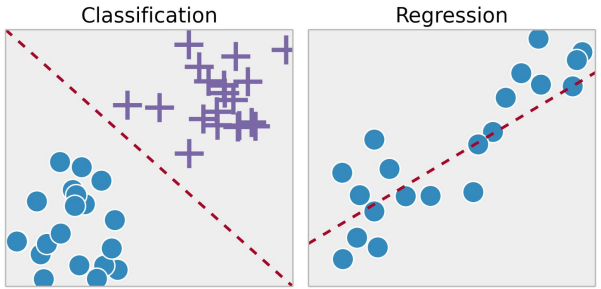
Function 'f', which maps input **'x'** to **'y'.**

Supervised machine learning: use a known sample (**x**&**y**) to make **'f'**



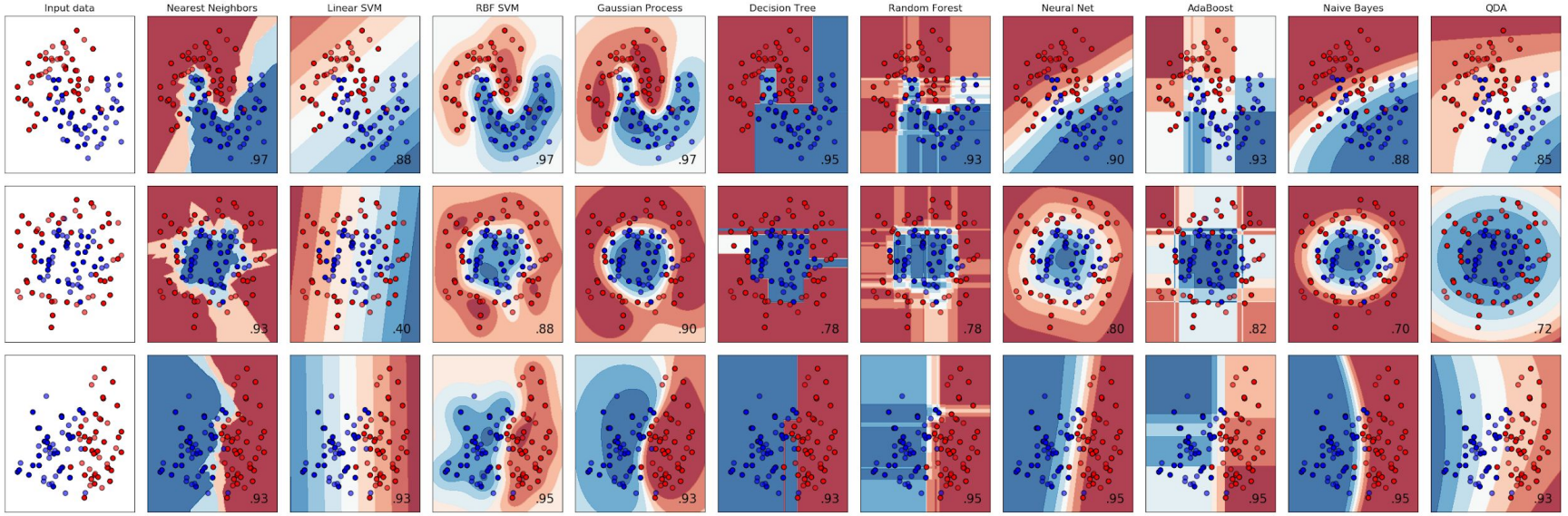
If **y** is discrete; a task is called **'classification'**

If **y** is continuous; it is called **'regression'**



A set of very flexible tools

Different types of machine learning methods exists. The choice of algorithm depends on the goal and type of input data, but also on how much it can be used as a black box. Many common algorithms are implemented in the **python package Scikit-learn.**



Two most popular algorithms: Random Forest

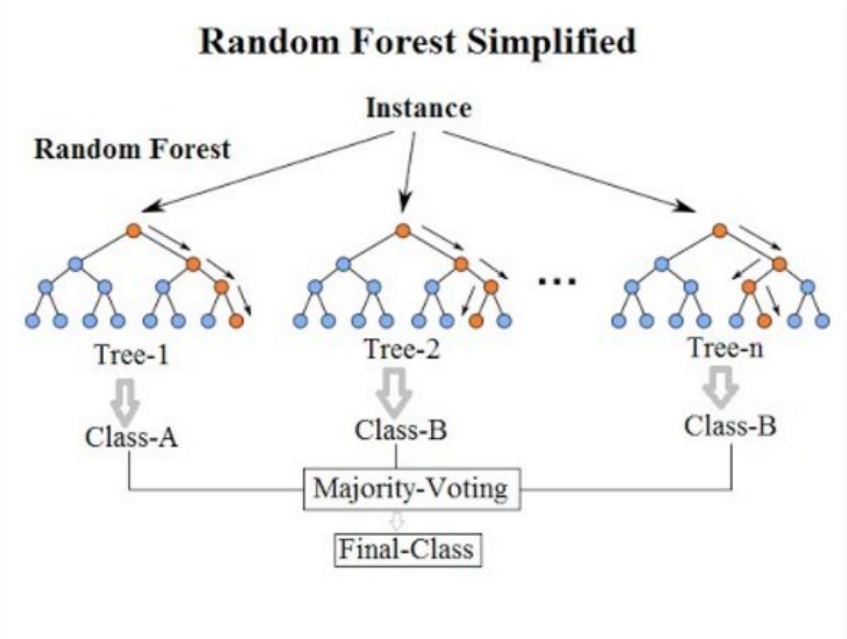
A collection of decision ‘trees’

Advantages:

- Fast to train and apply
- Easy to understand
- Works out of the box
- Works with small data samples

Disadvantages

- Not very flexible

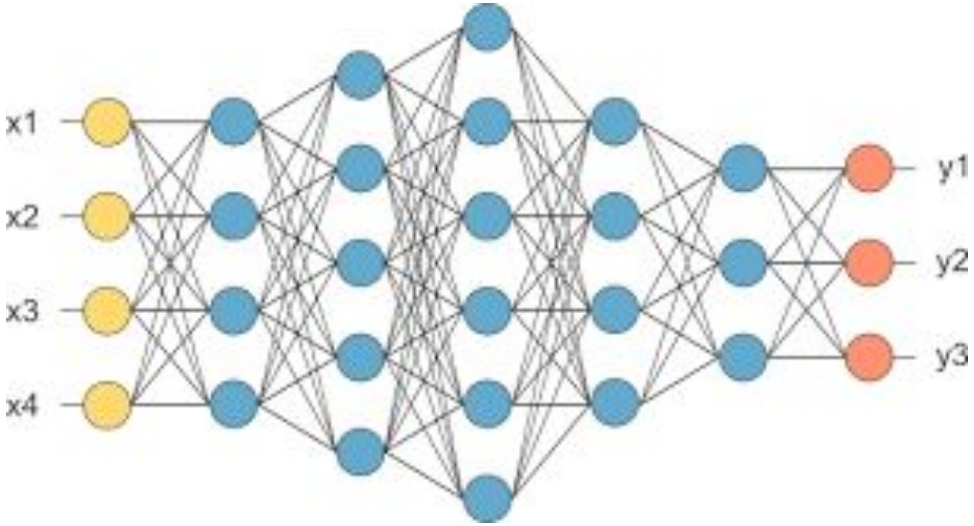


Two most popular algorithms: Neural Networks

A set of connected ‘neurons’

Advantages:

- Infinitely flexible: `neurons’ can be connected in many different way
- Works with images, timeseries, video, ...



Disadvantages

- Deciding on the the right architecture is difficult
- Needs a lot of data to ‘train’

When can machine learning be useful?

Machine learning has been very popular in the last few years. While it can do a lot of things, it is **not a silver bullet**. Typical problems were ML could be useful are:

- Task that a human can do in a split-second, but are hard to program.
- Tasks that take a (very) long time to compute can be approximated using ML methods.
- Any input data that hasn’t been seen yet in the training set, machine learning cannot deal with.

Disadvantages of machine learning: ‘BlackBox’

- **Training data is required** to construct the algorithms, which can be a challenge. In addition, **training data needs to representative** of the overall dataset.
- **Biases** and **completeness** can be difficult to quantify. This can be a problem when the output of machine learning methods are compared to predictions. For example; is the difference between the number of expected RR Lyrae stars versus what a ML-classifier finds an error in the prediction or the classifier?

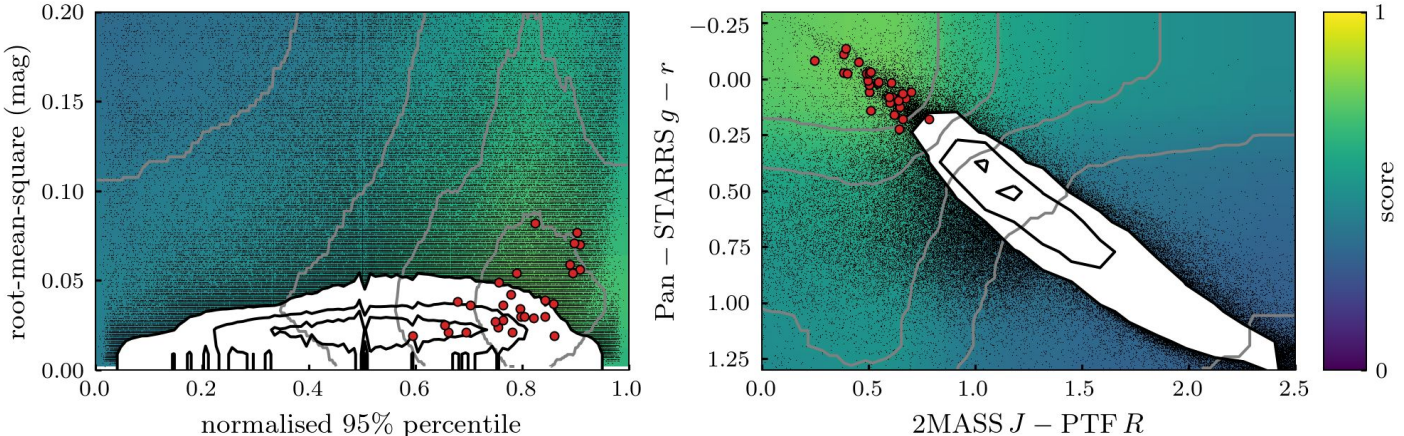
Examples of machine learning in stellar astronomy

Classification of variable stars

To classify variable stars, three types of information are used; variability, colour, and intrinsic luminosity.

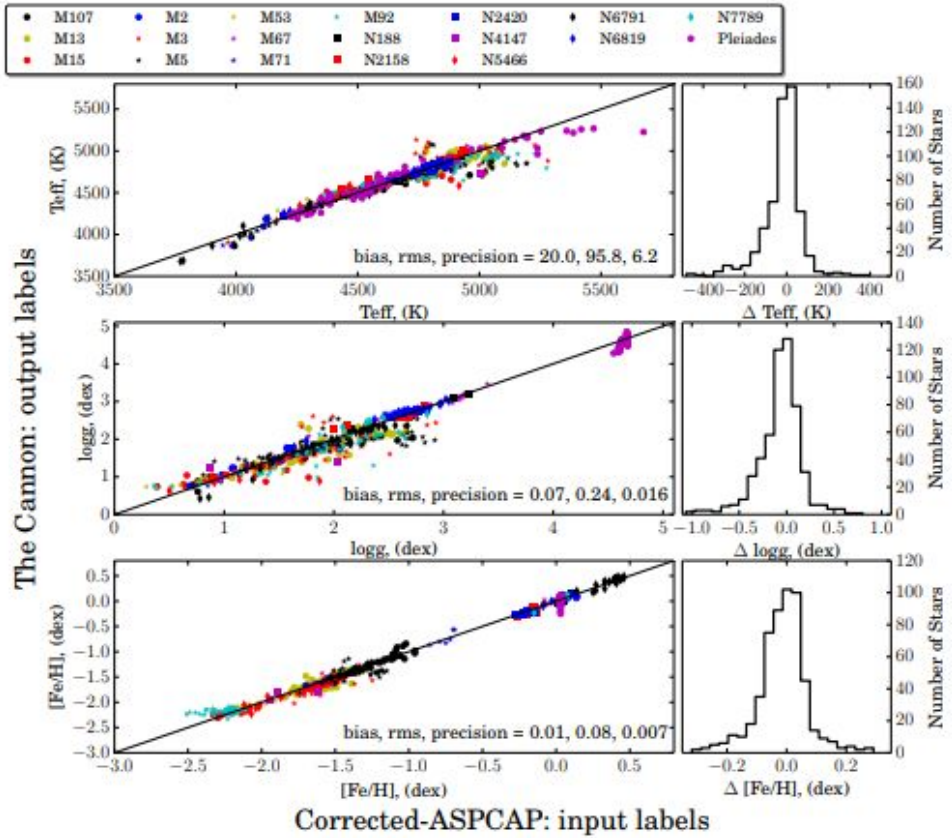
Typically, **lightcurve statistics** (median, RMS, period,...) are calculated, and combined with static colours and a parallax (or proper motion). These are then fed to a ML-classification algorithm (usually RandomForest). More advanced techniques which directly use the ligthtcurves are being investigated.

The figure below shows and example of how the ML-score depends on variability parameters (left) and colours (right). Examples: Richards 2012, Pashchenko 2017



Estimating stellar parameters from spectra

Ness et al 2017 developed a neural network that can quickly estimate stellar parameters (Temperature, surface gravity, [Fe/H]) from a 1D stellar spectrum. It is faster and more accurate than traditional model fitting approaches.



Faster modeling of physics of pulsating stars

Machine learning can also be used to **approximate physics simulations**.

For example, Hendriks & Aerts used a neural net to approximate stellar pulsations models (by MESA and GYRE). The neural net is used to find an approximate solution; **speeding up the 'fitting' process**.

Other examples are: parameter estimation of eclipsing binaries and determining the internal structure of Red Giants.

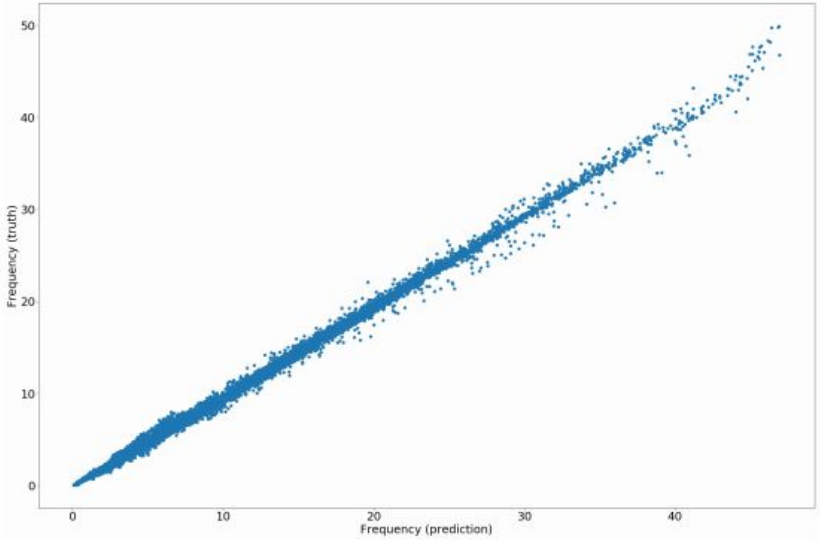
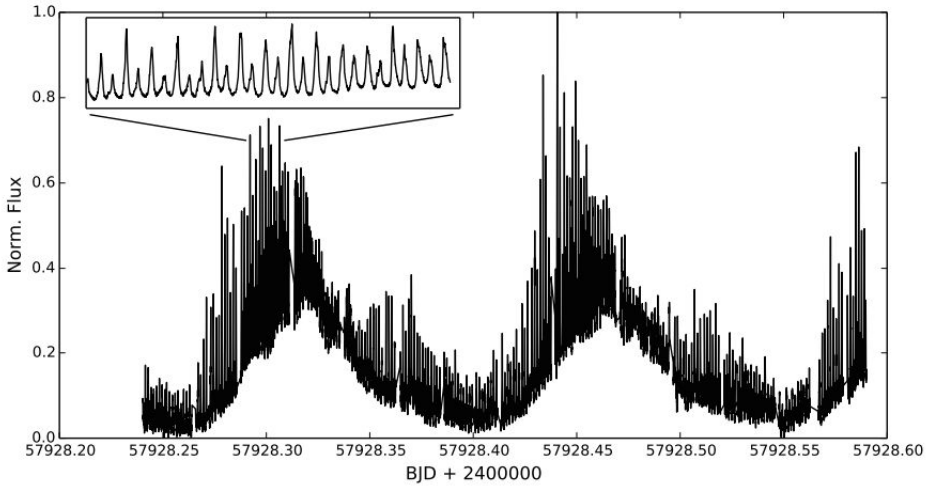


Figure 4. Visualization of the performance of the network. Plotted are the network predictions and the true values (from the dataset), for all stellar parameters and frequency modes. In this plot, 100,000 points from the validation data are shown.

Remaining challenges for machine learning in astronomy

- Outlier and novelty detection (e.g. how to identify new types of objects/events?)
- Uncertainty propagation in data
- How to handle irregular time series directly (e.g. lightcurves)
- How to handle sparse datasets, and how to use informative missing data (limits)



Novelty detection is difficult:
AR Sco (white dwarf ‘pulsar’) has a unique lightcurve but was misclassified as a Delta Scuti variable in the 70s. In 2014, an amature astronomer noticed it’s unique nature.