# Meeting Pieter Gijsbers – TPOT

A meeting was planned to discuss the TPOT evolutionary algorithm and using it with biomedical data. Some brief topics were discussed and a very promising result came out of this talk.

**Theory Evolutionary algorithms**

TPOT works on the principles of evolutionary algorithm, a kind of algorithm optimization combined with algorithm selection. Several aspects are important, such as preprocessing, feature selection and thirdly the chosen algorithms. Several pipelines are made to test multiple techniques which are either branched out when promising results occur or discarded when not.

Evolutionary algorithms are discussed in the Business Intelligence course from Anna Wilbrik. She can be contacted to find more information of this topic.

**Tested biomedical data**

A small selection of biomedical datasets are tested with TPOT. A set on heart diseases and diabetes was present and possible a DNA dataset as well. No problematic results came out of those, however these also do not span all possible data sets.

**Preprocessing**

For biomedical data sets, preprocessing is very important. The acquired data sets usually are very heterogeneous, complex and of low quality. Entries for values are often absent and outliers are present very often. At last looking at populations is important, as behavior can be very different for different groups.

These three topics are only briefly addressed in TPOT. Missing values are given the median value, which is a quick solution to the problem, however better solutions exist. Outlier detection is not available yet and is something TPOT would benefit from. At last sub population checking is done in several algorithms, such as decision trees. However no real focus is done on it. Extended implementation would be beneficial for the algorithm.

**Feature selection**

Feature selection is another important aspect for biomedical data. It consists of large quantities of features, for example the number of genes in DNA or the pixels in an MRI image. Feature selection often must be done to tackle that. No validation of feature selection is done before, however some algorithms, for example Principle Component Analysis is present in TPOT. This would be worth looking into if TPOT handles feature selection efficiently.

**Current algorithms**

The main focus has been on classification for TPOT. Regression is available but in literature all research focuses on classification. A regression evaluation would be very useful as regression is also a big part in biomedical data.