**Problem description**

The NFI has previously developed a machine learning model to predict the number of contributors (NOC) to a DNA sample based on features derived from Short Tandem Repeat (STR) data. This Random Forest classifier currently uses 19 of the top features and achieves an accuracy of about 83%. However, the only output experts receive is the predicted NOC. No information about how the model came to this conclusion is provided, therefore not allowing experts to use this tool effectively as support for their decision making.

XAI has been recognized as a tool to help humans understand the *why* of outcomes in Machine Learning (ML) applications. Especially in domains where ML is used to influence people’s lives, it is pertinent to know factors that influence certain decisions. In recent years, counterfactual explanations have been at the forefront of development for generating insights into individual predictions. Within a classification context, instead of answering the question “*Why class A?”*, counterfactuals answer “*Why not class B?”*. This way of reasoning is underpinned by the social sciences to be effective.

With the addition of XAI, the NFI hopes to improve the value of their prediction tool for experts in determining the number of contributors. The workflow of the experts consists of first determining the NOC without any support, after which they validate their own ideas with the current NOC model. If these outcomes do not align, the expert should be able to consult informative explanations of the NOC model to determine why it does not generate the same outcome as the expert had envisioned. In this way, the expert can make an informed decision to stick with their own conclusion if the model does not seem to have learned the correct distinctions, or choose the predicted value if the model brings up a good argument.

*Expert predicts a sample to be a 4-person mixture based on the MAC of 7, a TAC of 115 and 40 peaks being below the RFU threshold.*

*NOC model predicts that sample to be a 5-person mixture*

*Expert wonders why not 4*

*Counterfactual explanations states that if there were 10 less peaks below the RFU threshold, it would be classified as a 4-person mixture*

*Expert looks at the*

The data obtained from the NFI has many features; each data sample has 23 loci, for which all alleles are measured and their peak height. In contrast, the data is low in volume. There are roughly 590 different profiles available.

**Main related works**

Since the field is new, not many of these methods have been rigorously tested or compared. A recent study has classified existing techniques for generating counterfactuals according to certain properties [1]. For instance, whether they exploit parts of the underlying model, the data distribution, or how many feature changes are permitted. However, these methods have not been actually put to the test based on quantitate measures, or have been submitted to user study.

* Some methods
* SHAP?
* Together?

**Contribution**

*How can we generate informative counterfactual explanations for predictions of the number of contributors (NOC)?*

* What do experts look at when determining the NOC?
* What information helps experts decide between two NOC values?
* What are features of the data?
* Which counterfactual generation techniques could be applied to this type of data?

**Planning**

First, a thorough literature study must be performed to identify the current counterfactual techniques and any desiderata that were used.

Methodology, project objectives, organization, planning, risks and mitigation plan

**References**

1. Verma, S., J.P. Dickerson, and K. Hines, *Counterfactual Explanations for Machine Learning: A Review.* ArXiv, 2020. **abs/2010.10596**.