Dataset

The Adult Dataset (https://archive.ics.uci.edu/ml/datasets/Adult) was chosen for this task. It contains various basic data about people, namely
age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

The target column stores information on income: whether it exceeds $50K/yr.

Predicting quality

An F1-score was used to determine the quality of the algorithm. There is a class imbalance in the dataset:

```
True     745
False    255
```

This tells not to use accuracy. Therefore the choice is left to use F1-score. In addition, it takes into account cases where predicting false negatives is a significant disadvantage.

Binarisation of data

In order to determine the binarisation strategy, we need to understand data structure in dataset. The dataset contains: numeric, categorical data.

The categorical case:

In this case, it is sufficient to create a new column for each individual category, which will be True if the original column had this value, otherwise False.

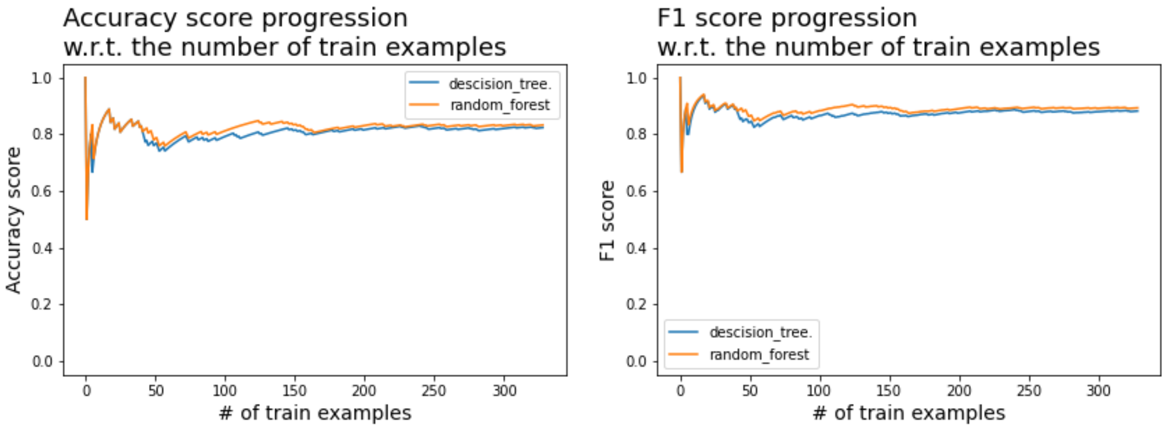
If there were only two categories, they would change to True or False.

Numerical case:

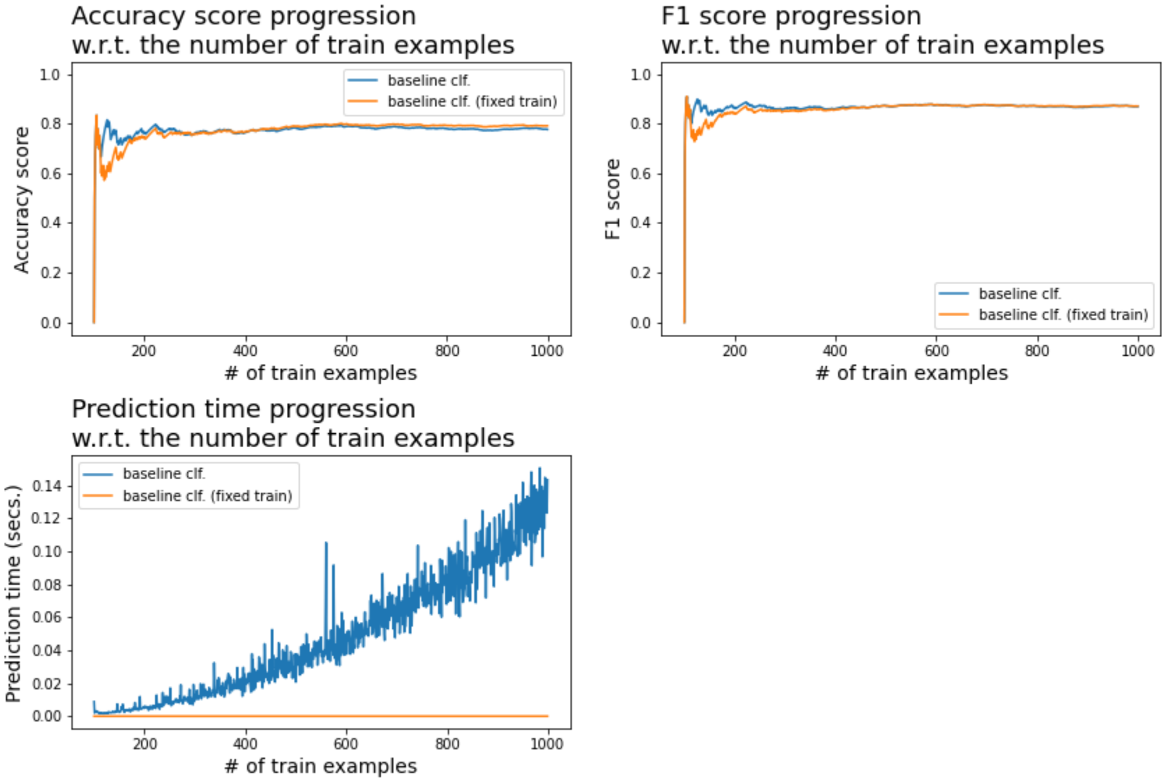
Create a separate column for each value, which would be True if the original column had a given value, otherwise False. However, this solution may slow down the algorithm, as it will eventually need to go through a large number of rows.

Comparison with existing algorithms

Decision trees, random forests were used for the analysis.



Results of Decision trees and Random forests



Results of Lazy_fca

It can be seen that random forest performs better than all 3 algorithms, although the decision tree is not far behind. Lazy_fca is in the same range of values as the two known algorithms, but the final result is lower (about 5%). This can be connected to the fact that the algorithm has not been optimised and there is room for improvement.