Horse colic dataset

- The main aim of this project is to classify if adult horses will succumb after being diagnosed with colic. Finding what attributes have a high correlation with the horse dying or being euthanized.

- The aim is to get the smallest tree, choosing the attribute that produces the purest nodes, that is the greatest information gain.
- Information gain, amount of information gained by knowing the value of the attribute
- Entropy of distribution before the split vs the entropy after the split

Horse colic dataset

- 300 samples, 28 features
  - 21 float64
  - 4 int 64
  - 3 objects (data about lesion types)
- I removed 9 features –
  - 50-82% missing data in one
  - Not relevant according to the data description
  - Hospital number (284 unique values of 300 total)
- I also removed noisy data
  - 8% of the data is from 6 month old horses or younger, they have a much higher death rate, the values are extremely fluctuating compared to adult horses. (Higher pulse, respiratory rate, narrower temperature range)
- I discretisized values to simplify classification in the tree
  - Temperature, low, normal, high
  - Pulse, low, normal, high
  - Protein levels in blood
  - Mucous membrane (tongue color)

Data exploring

- I performed the data cleaning in Python.
- I also decided to try doing PCA because of the high dimensionality, to find the variance in the data using Scikit package in python

Data prep

A lot of missing data
Numbers are the ratio of
Missing data, 82.3% is the most

```
[183]
# Pre processing

# Had to add the column names to the data file
# Forcing data types to str to further examine data, forcing lesion attributes as strings because leading zero's will disappear

data = pd.read_csv('horse-colic.data', delim_whitespace=True, na_values="?", dtype={'lesion_1':str, 'lesion_2':str, 'lesion_3':str})

nullRatio = data.isna().sum() / len(data)*100
nullRatio.sort_values(ascending=False)
```

```
nasogastric_reflux_ph    82.333333
abdomo_protein           66.000000
abdomo_appearance        55.000000
abdomen                  39.333333
nasogastric_reflux       35.333333
nasogastric_tube         34.666667
rectal_exam_feces        34.000000
peripheral_pulse         23.000000
rectal_temp              20.000000
respiratory_rate         19.333333
temp_of_extremities      18.666667
abdominal_distention     18.666667
pain                     18.333333
mucous_membrane          15.666667
peristalsis              14.666667
total_protein            11.000000
capillary_refill_time    10.666667
packed_cell_volume        9.666667
pulse                     8.000000
surgery                   0.333333
outcome                   0.333333
lesion_3                  0.000000
surgical_lesion           0.000000
lesion_1                  0.000000
lesion_2                  0.000000
hospital_number           0.000000
age                       0.000000
cp_data                   0.000000
dtype: float64
```

Data examination – Before data prep

High standard deviation in pulse, respiratory rate and protein levels
High range of values, mostly because of (8% young horses in the dataset)

	rectal_temp	pulse	respiratory_rate	peripheral_pulse	total_protein	packed_cell_volume
count	240.000000	276.000000	242.000000	231.000000	267.000000	271.000000
mean	38.167917	71.913043	30.417355	2.017316	24.456929	46.295203
std	0.732289	28.630557	17.642231	1.042428	27.475009	10.419335
min	35.400000	30.000000	8.000000	1.000000	3.300000	23.000000
25%	37.800000	48.000000	18.500000	1.000000	6.500000	38.000000
50%	38.200000	64.000000	24.500000	2.000000	7.500000	45.000000
75%	38.500000	88.000000	36.000000	3.000000	57.000000	52.000000
max	40.800000	184.000000	96.000000	4.000000	89.000000	75.000000

Data prep

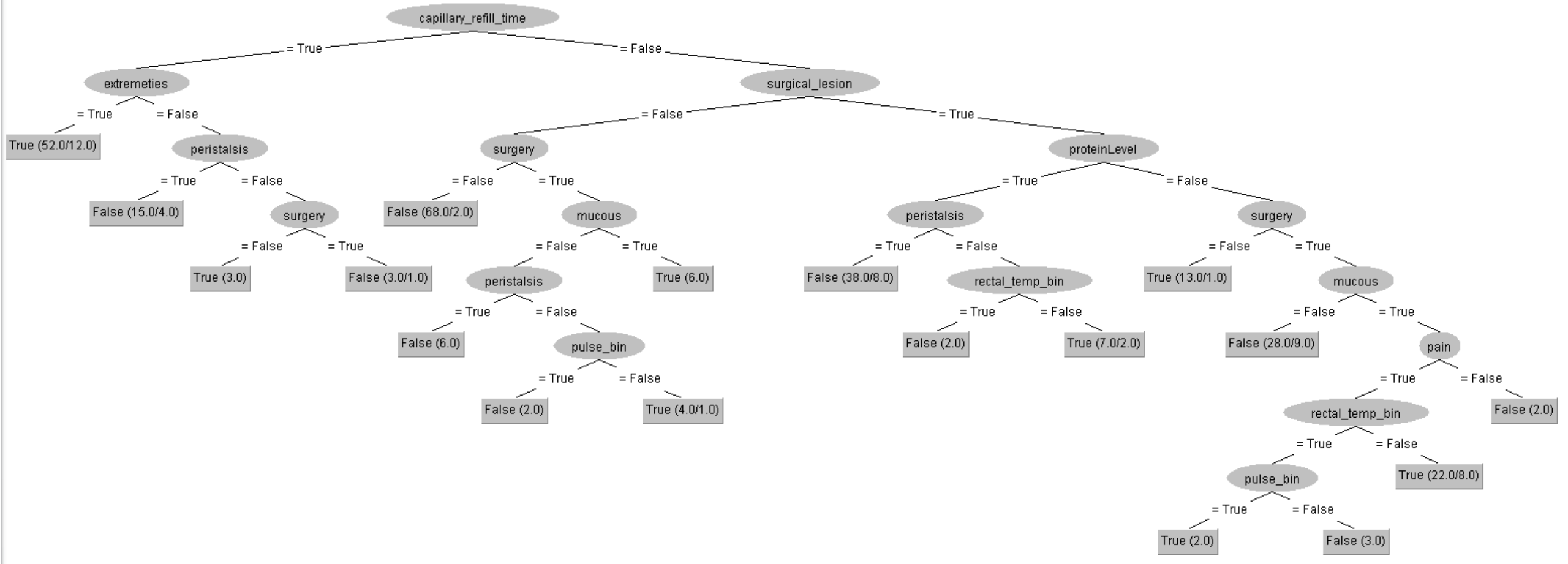
Converted to True/False values.
If the horse had :

- Surgery
- Long breathing
- Pain
- Peristalsis
- Abdominal distension
- Surgical lesions
- High rectal temp
- High pulse
- Heat in extremities
- Coloration of mucous
- High cell count
- High protein levels

```
surgery,capillary_refill_time,pain,peristalsis,abdominal_distention,surgical_lesion,rectal_temp_bin,pulse_bin,extremeties,mucous,packedCell,proteinLevel,death
False,True,True,True,True,False,True,True,True,False,True,True,True
True,False,True,True,True,False,True,True,False,True,True,True,True
False,False,True,True,False,False,False,False,False,True,True,False,False
False,True,False,False,False,False,False,True,False,True,True,False,True
False,False,True,True,True,False,False,False,False,True,False,False,False
True,False,True,True,True,True,False,True,False,False,True,False,False
True,False,False,True,True,True,False,True,True,False,True,True,True
False,False,True,True,True,True,False,True,True,True,True,False,True
True,False,True,True,False,True,False,True,True,True,True,False,False
False,False,True,True,True,True,True,False,False,True,True,False
True,False,True,True,True,False,False,False,False,False,False,False,False
True,False,True,True,False,True,False,True,True,False,True,True,False
True,False,True,True,True,True,False,True,True,True,True,False,True
False,False,False,False,False,False,False,True,False,False,False,False,False
True,False,True,True,False,True,False,True,False,False,True,False,False
False,False,False,True,True,False,True,True,True,True,True,True,False
True,True,True,True,True,True,True,True,False,True,True,False,False
False,False,False,True,False,False,True,True,False,True,False,False
True,False,True,True,False,True,True,False,False,True,True,False,False
True,False,False,True,True,True,False,True,True,True,True,True,False
False,False,False,False,False,False,False,True,False,False,False,False,False
True,True,True,True,True,True,False,True,True,True,True,False,False
True,False,False,True,True,False,False,True,True,False,True,False,False
True,False,False,False,False,False,False,False,False,False,False,False,True
False,False,False,True,False,False,False,True,False,False,True,True,False
False,True,True,True,True,True,False,True,True,True,True,False,True
False,True,True,True,True,True,False,True,True,True,True,True,True
True,False,True,True,True,True,False,True,False,False,True,False,False
True,False,True,True,False,True,False,True,False,False,True,True,False
True,True,True,True,True,True,False,True,True,True,True,False,False
False,True,True,True,True,True,False,True,True,True,True,True,True
False,True,False,False,False,False,False,True,False,True,True,False,True
True,False,True,True,False,True,False,True,False,False,True,True,False
False,False,True,True,True,False,True,True,False,False,True,False,False
True,True,True,True,True,True,False,True,True,True,True,False,True
False,False,True,True,False,False,False,True,True,True,True,False,False
True,False,True,True,True,True,False,True,True,True,True,True,True
True,True,True,True,False,True,False,True,True,True,True,True,True
False,True,True,True,True,True,False,True,True,True,True,False,True
```

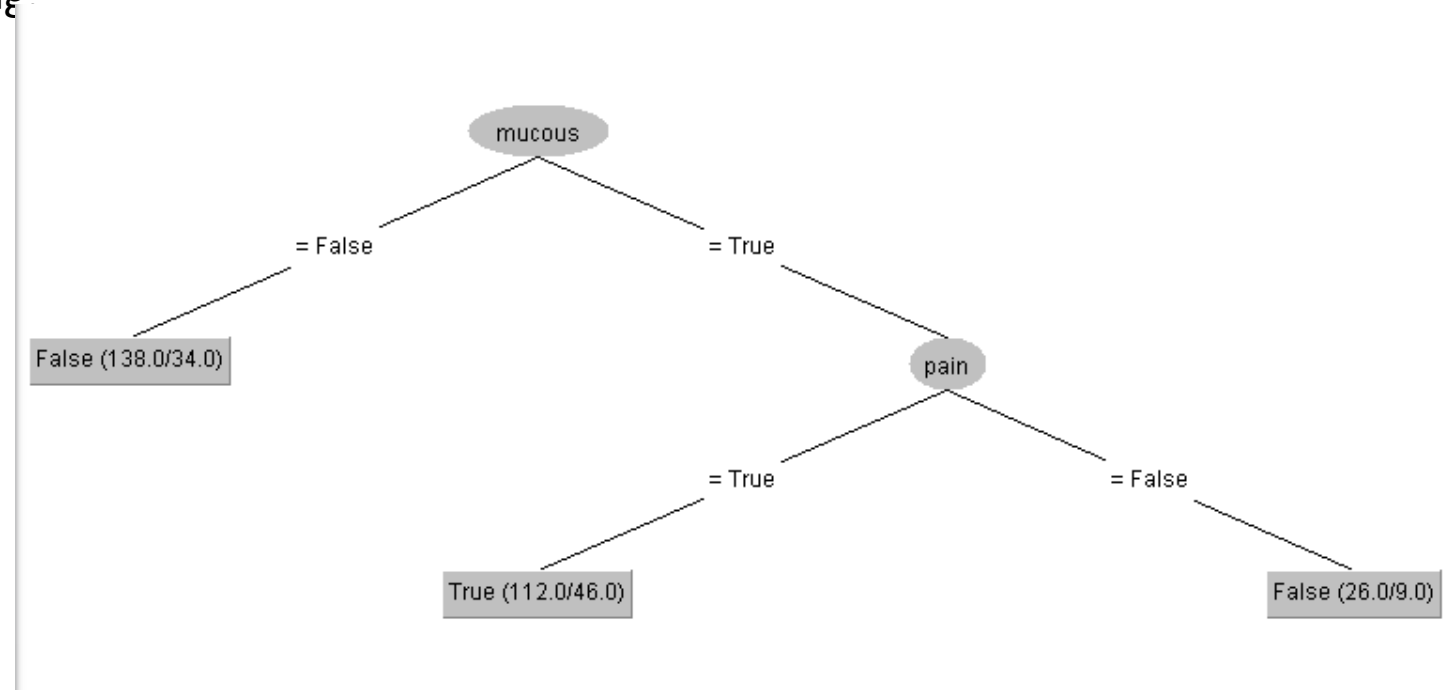
C4.5 tree generated with 13 attributes

A lot of leaf nodes with few cases, although good classifications but this may lead to over classification
By using this tree I got 74.6% correct classification.

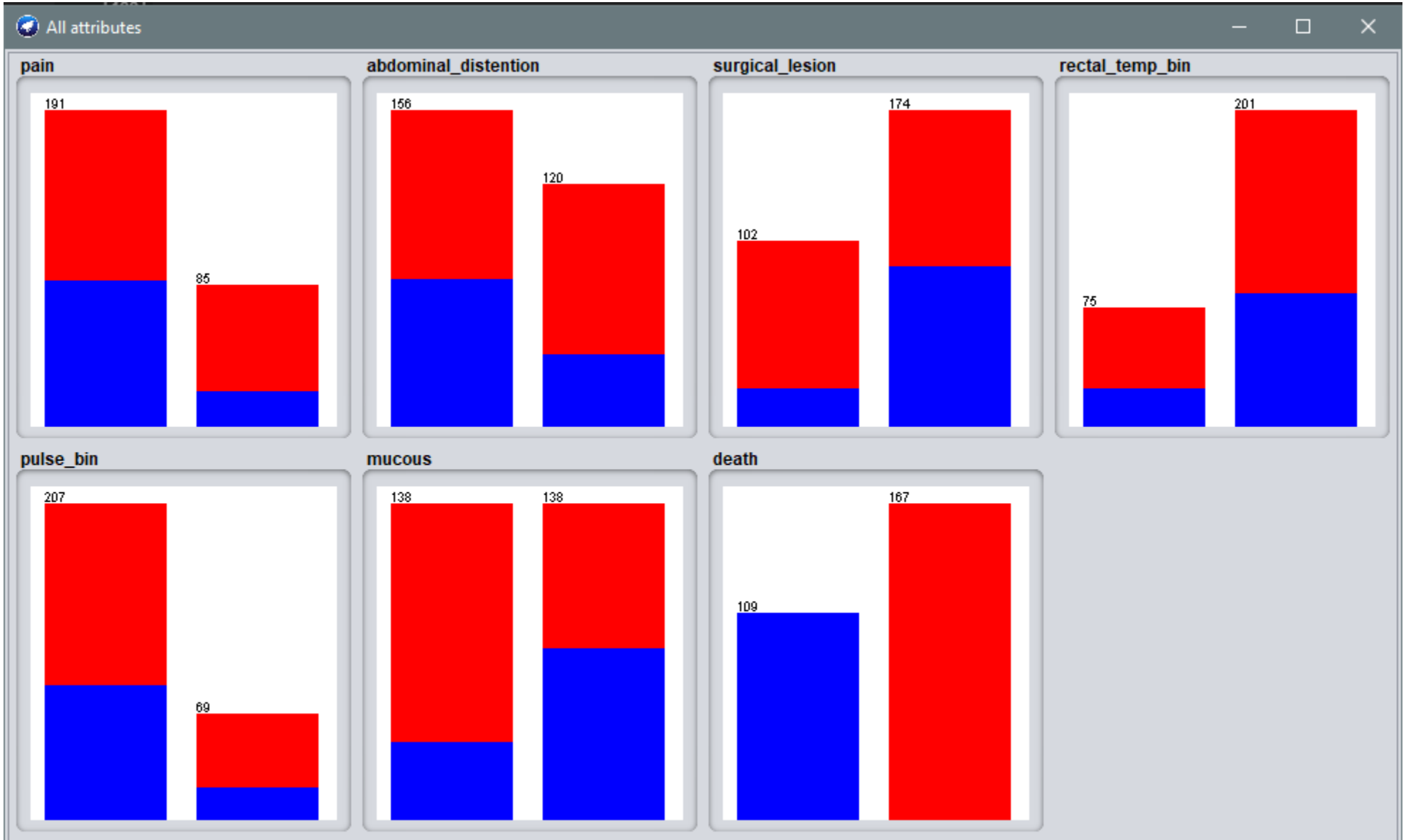


C4.5 tree generated with 7 attributes

The algorithm ignored quite a few of the attributes but still gets a good classification. It only uses two of the 7 total attributes.
Few nodes with many cases, somewhat accurate clasifications but many misclassifications.
By using this tree I got 67% correct classification.
However I feel that using only two attributes is way to little data to rely on.
I tried playing with the variables, removing one or two from 13. But that always seemed to lead to overpruning
I would chose the first model because of the accuracy. Fewer attributes are omitted, but I would like to have time to play a little bit more with the pruning

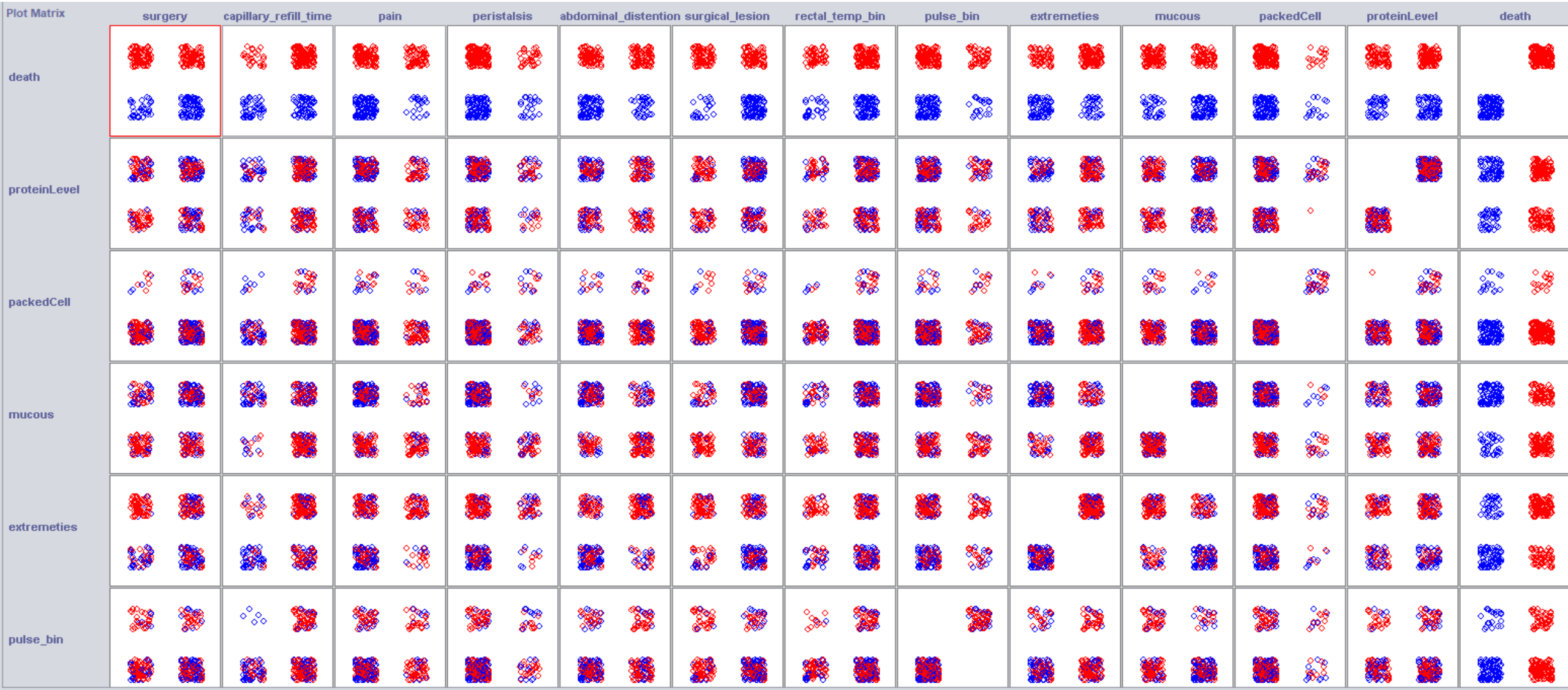


Attributes – blue = dead horse



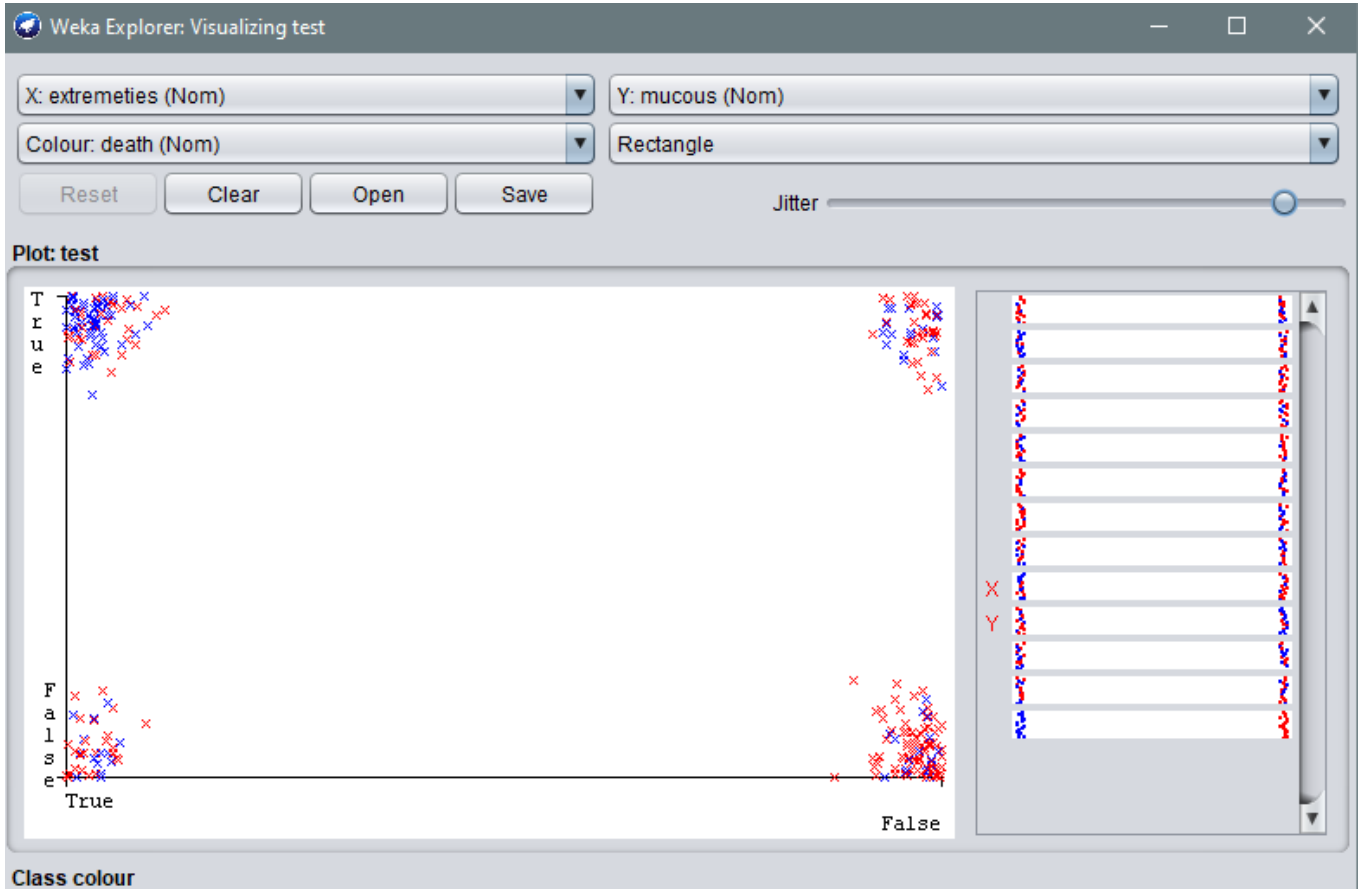
Blue indicates death

Lessons learned



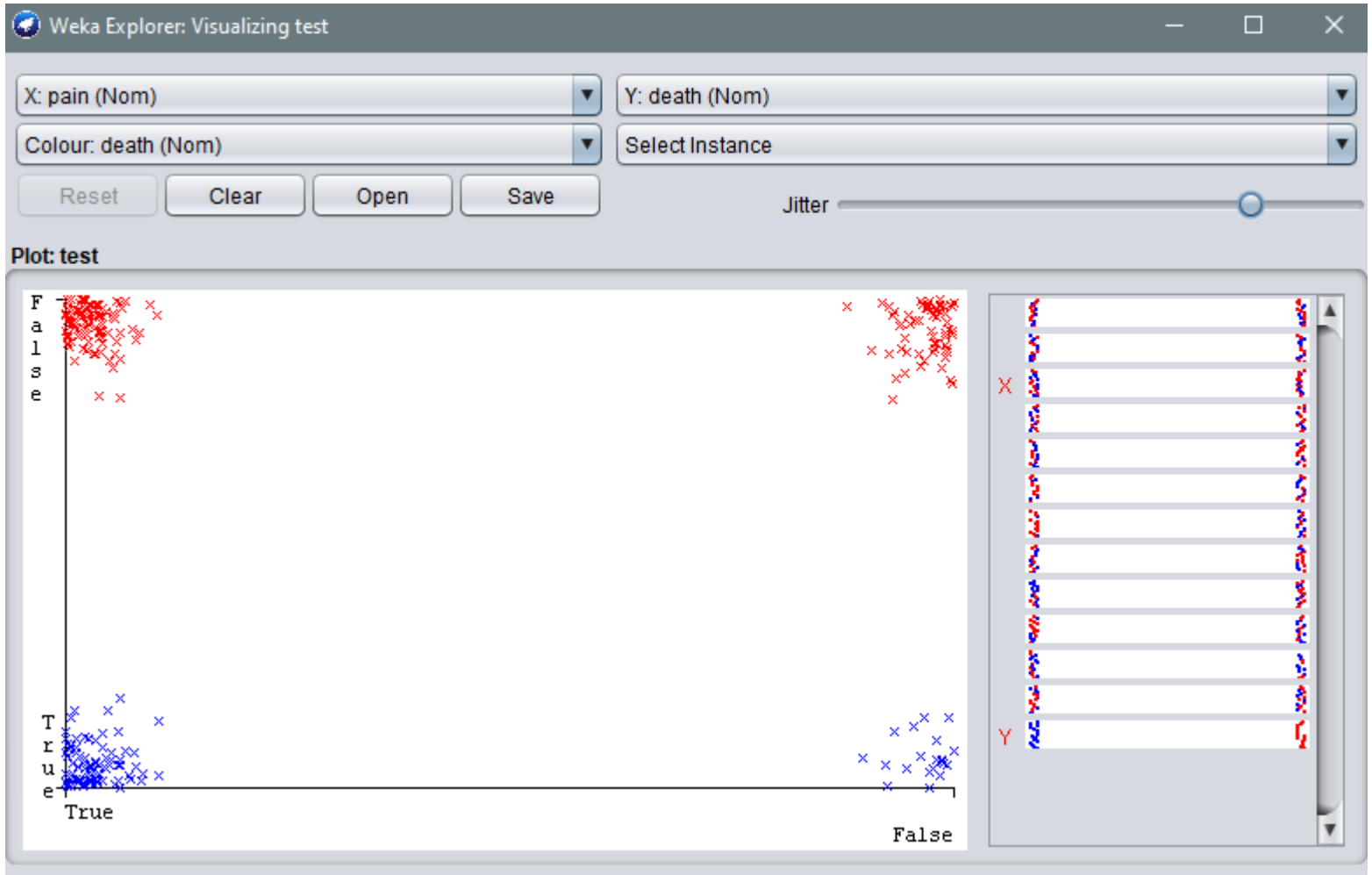
- Blue = Dead
- Correlation between temperature in extremities and coloration of mucous.
- If the horses have signs of discoloration of the mucous and heat in the extremities they are likely to be classified as dead

Lessons learned



- Blue = Dead (true)
- Correlation between pain and death, a lot of horses that show pain levels are dead

Lessons learned



fin

- All code is on github, including python and weka

https://github.com/andriOlafsson/horseColic

- Data cleaning was performed in a jupyter notebook that can be run online
- https://nbviewer.jupyter.org/github/andriOlafsson/horseColic/blob/master/jupyter.ipynb

Thank you

- Andri Ólafsson