

Machine Learning report

Analysis on Horse colic dataset



Student: Monica Soares Brandao

Subject: Data Mining

Supervisor: Manoela Kohler

Summary

This report shows the process of machine learning and results in the classification of Horse colic dataset Basically, the Rapidminer tool was used for this work.

The goal is to predict whether a horse can live or die. based on past medical conditions, using one or more Machine Learning models.

As it is a classification problem with categorical labels, supervised learning models will be applied here.

About the database

The base used was the Horse colic base, which contains 28 attributes, categorical or continuous, that describe the animal's health status.

The label corresponds to three classes of output that indicate what happened to the horse, whether it died, lived or euthanized.

The base contains a large number of missing values, approximately 25% of the total values.

The dataset was divided into 299 records for training and 89 records for testing.

A data dictionary (1) was provided indicating the meaning and importance of each attribute, which was also considered in this analysis (2).

For this project, the basic scheme for Data Mining was followed.

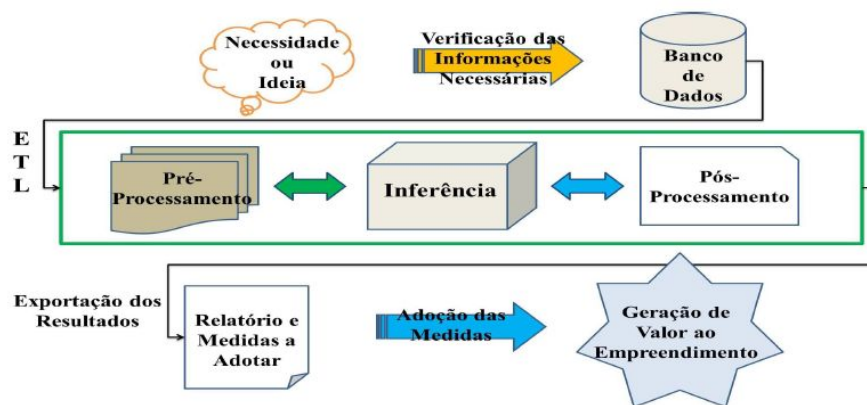


Fig. Basic scheme of a data mining project

1 Data dictionary is found in Annex I

2 For this analysis it was decided to keep the attributes lesson_1, lesson_2, lesson_3 in the original form, without dividing them.

1. Exploratory analysis of the data

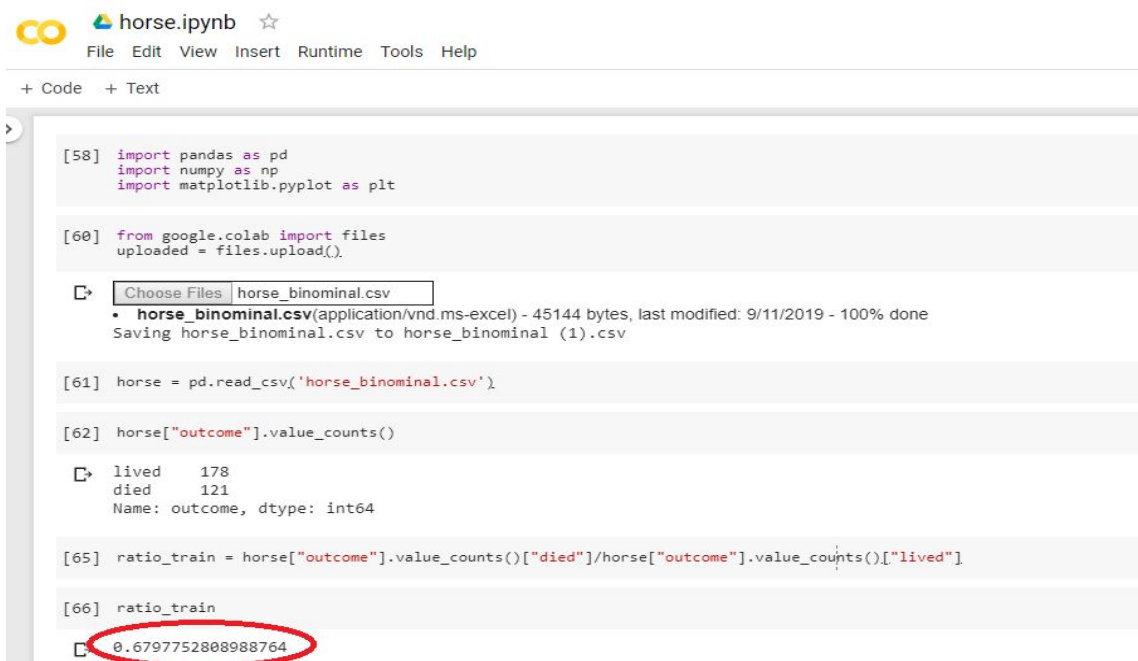
Having the exploratory analysis can reveal how the variables are related to each other. Analyzes made in a training database are shown below.

1.1. Label analysis

We estimate that the proportion of the dead class over the number of live to the training base is $121/178 = 0.67$. And for test basis it is $36/56 = 0.64$. For that, a Phyton Notebook was used with the Pandas library.

Training Data set		
Outcome?	Lived	Died
	178	121
Test Data set		
Outcome?	Lived	Died
	53	36

Fig 1.1.1 - Analysis of the label



```
[58] import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

[60] from google.colab import files
uploaded = files.upload()

Choose Files horse_binominal.csv
• horse_binominal.csv(application/vnd.ms-excel) - 45144 bytes, last modified: 9/11/2019 - 100% done
Saving horse_binominal.csv to horse_binominal (1).csv

[61] horse = pd.read_csv('horse_binominal.csv')

[62] horse["outcome"].value_counts()
lived    178
died     121
Name: outcome, dtype: int64

[65] ratio_train = horse["outcome"].value_counts()["died"]/horse["outcome"].value_counts()["lived"]

[66] ratio_train
0.6797752808988764
```

Fig 1.1.2. - Example of the proportion calculation (ratio) in Phyton / Pandas for the training base

Three techniques were used to calculate the weight of the attributes, Chi-squared statistics, Gini index and Information gain

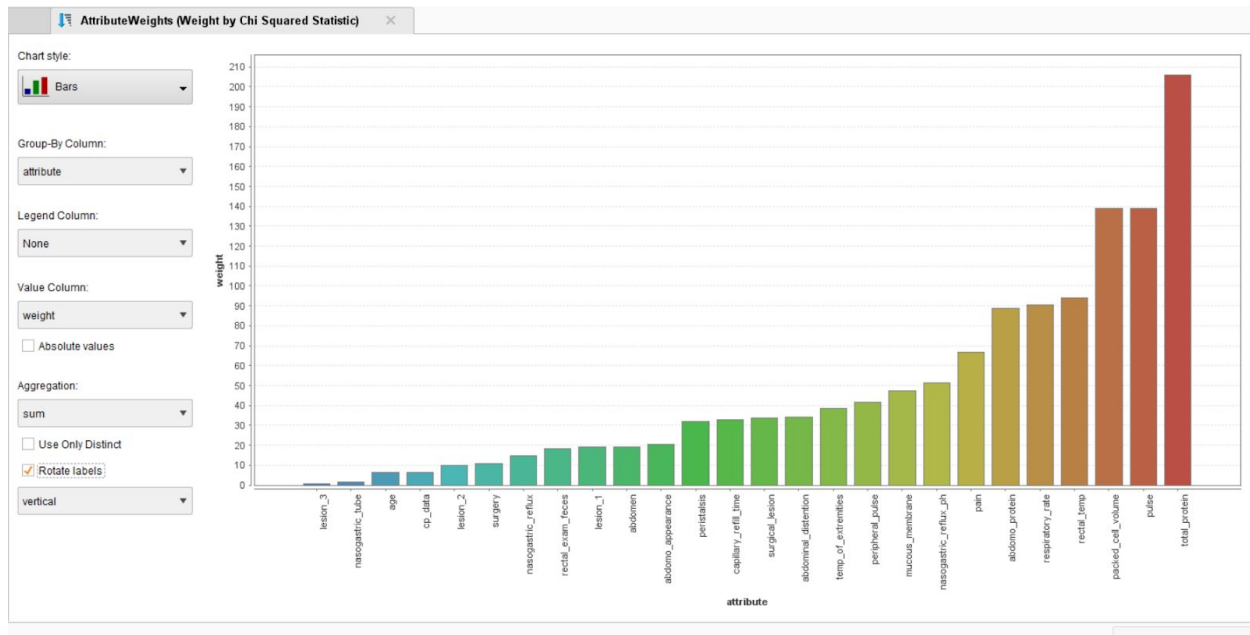


Fig1.1.3. - Attribute weights using the Chi-squared method

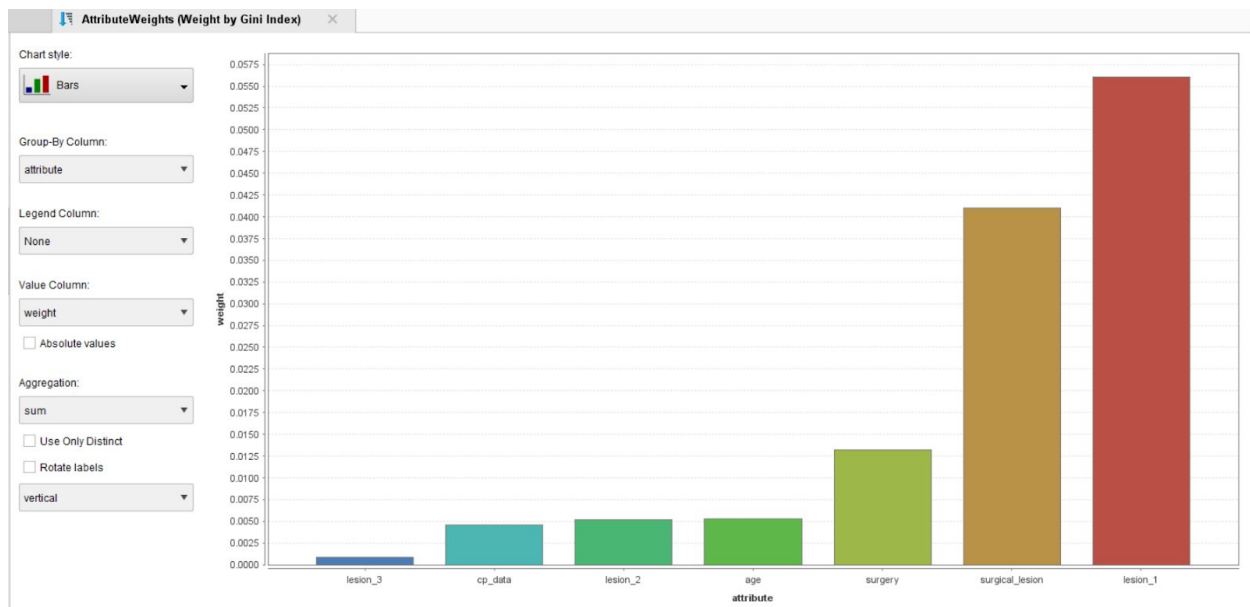


Fig1.1.4. - Attribute weights using the Gini index method

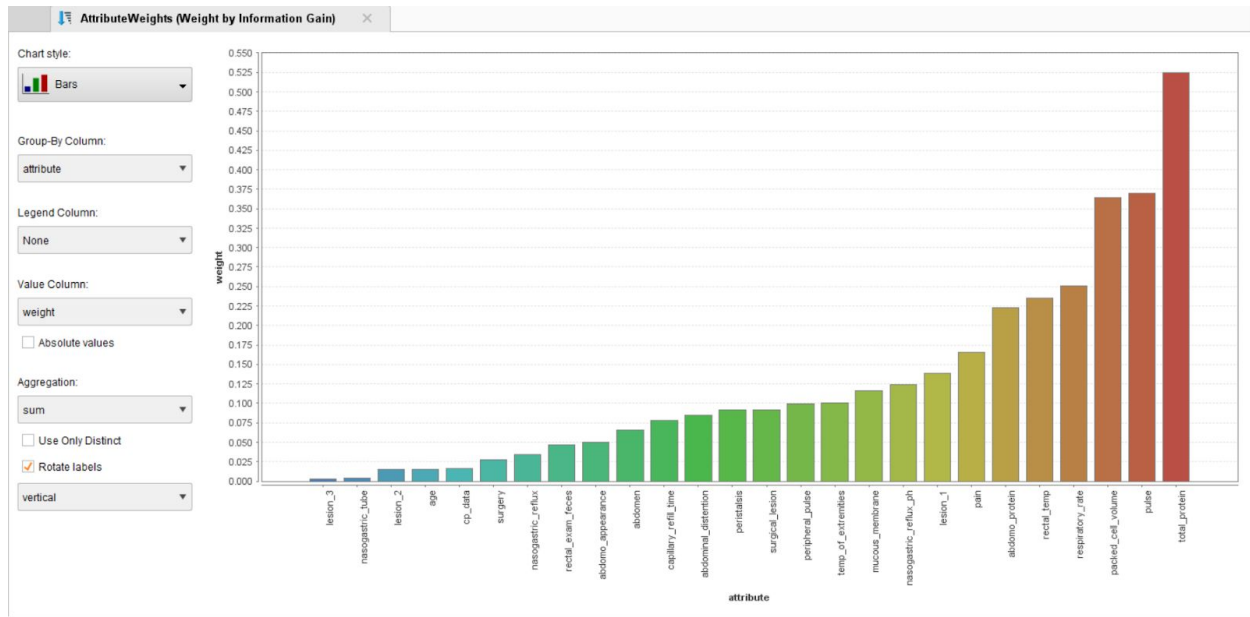


Fig 1.1.5 - Attribute weights by the Information gain method

The calculations for Information Gain and Chi-squared present similar weights for the attributes, differently from the Gini index.

Although the weights of the Information Gain and Chi-squared algorithms are similar, for future analyzes we used the weight by Information Gain because it has an accuracy result better than that of Chi-squared.

1.2 Analysis of the attributes

The attributes analysis reveals that the Hospital Number attribute is an **id**, so in the operator **Read csv** we define this attribute to ID so it does not go to the model.

As a label used was the attribute **outcome** that defines what happened to the animal, whether it lived or died (or died due to the process of euthanasia).

Therefore, it was considered that two categorical classes were sufficient to answer the proposed problem, to know if the animal would live (lived) or die (died), without specifying whether it was death by euthanasia (euthanasia).

So all records that were euthanasia (euthanasia) were changed to dead (died) in the database. And the label attribute changed from polynomial to binomial (change also made in the Read csv operator).

Through the bar graph we can see the balance of data for the base with polynomial and binomial label.

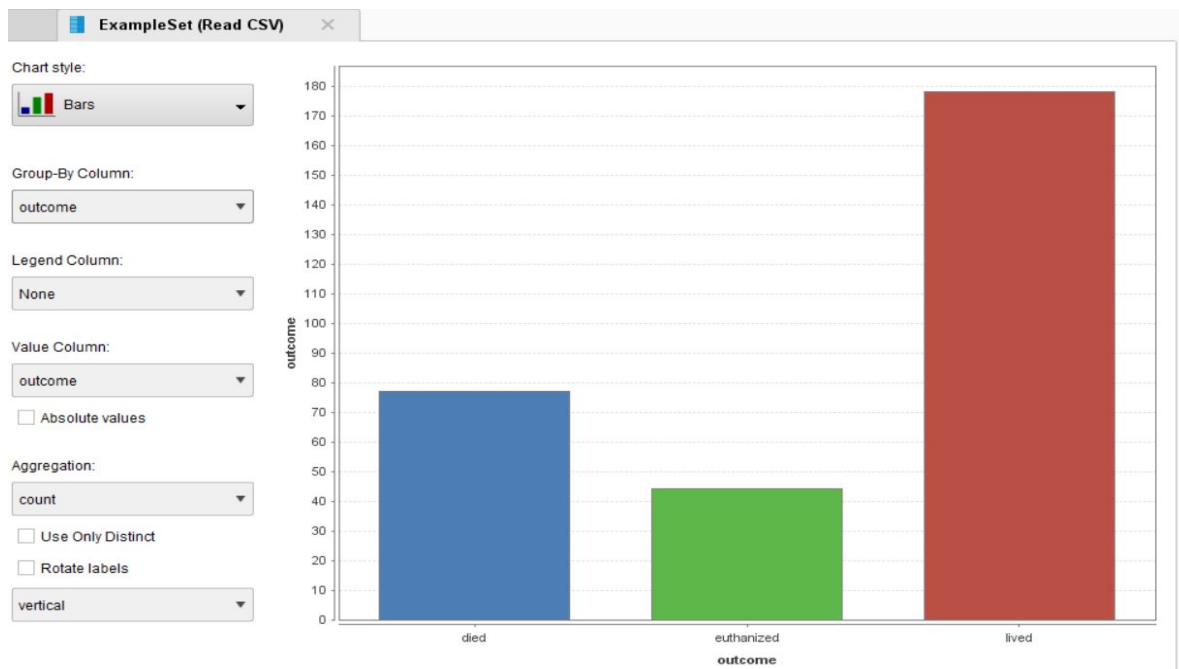


Fig1.2.1 – Original database – polynomial label

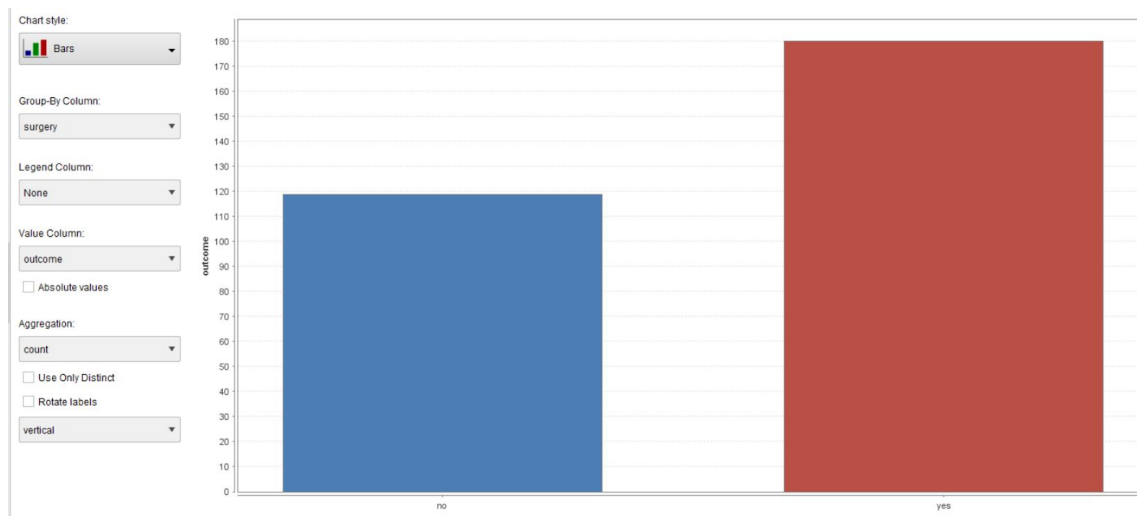


Fig 1.2.2 – Database with binomial label

1.3 Treatment of missing values

The base has many 'missing values', approximately 25%. That is why it is important to check the best way to replace these values and use them in the models.

First hand, using only the operator **Replace Missing Values** was not enough to eliminate them, since the “NA” value was not recognized by the operator as a “missing value” but as a nominal attribute value.

To work around this problem the operator was used **Declare Missing Value** to indicate that “NA” should be considered as a missing value.

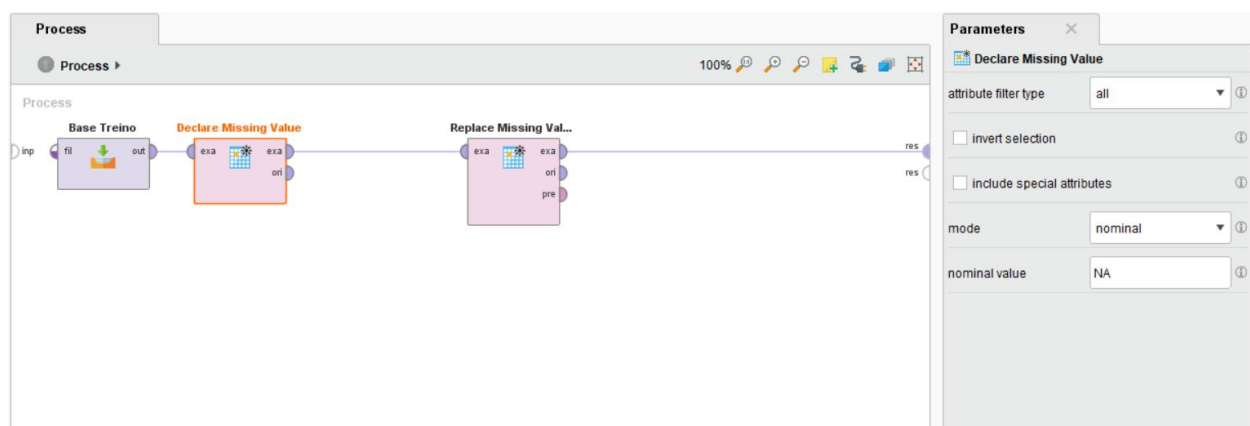


Fig 1.2.3 - Declare Missing Values and Replace Missing values

For the operator **Replace Missing Value**, the default value **Average** was used and the substitution of the values for Mean metric.

1.3 *Graphs analyzed in the data exploratory phase*

An important part of data exploration is to understand how different attributes are related to each other. So using different types of graphs helps us in this analysis.

In this report we include three charts that were considered to best describe this relationship with the attributes, Deviation, Quartile Color Matrix and Bubble Chart.

Below we can see the difference of each one and what they reveal to us in relation to the attributes.

a.Deviation

This chart shows attributes relevant to the classification. In it we can analyze the attribute *abdominal_distention*, indicated in the data dictionary as an important attribute for the classification.

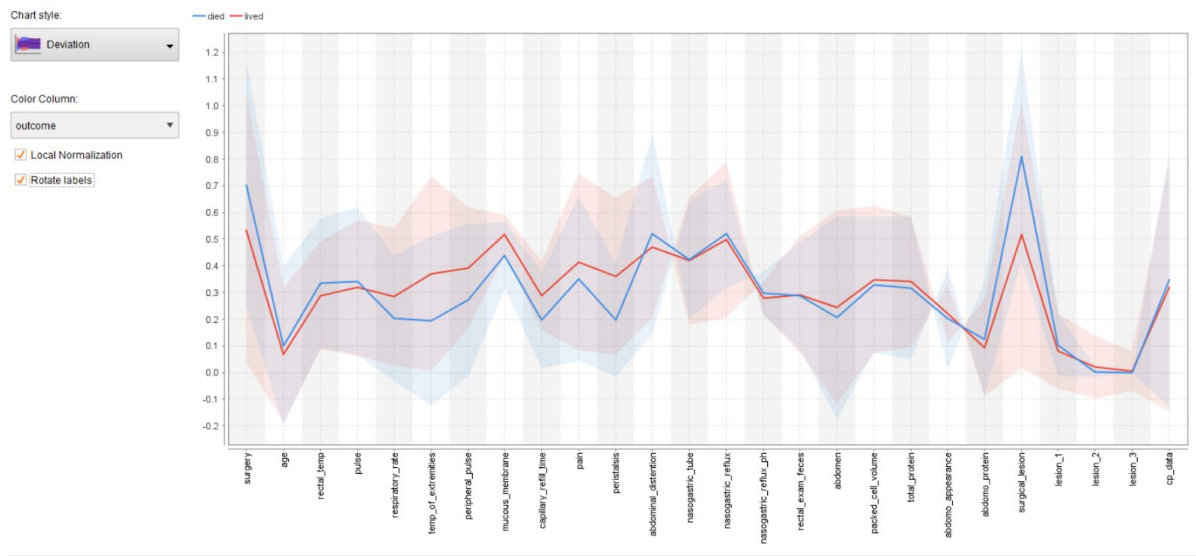


Fig 1.3.1 - Deviation Chart

b.Quartile Color Matrix

In this graph we can see the relationship of the label with other attributes.

This graph shows a marked difference in distribution when separated by the label.

It allows us to see that the attribute *peristalsis* has almost no overlap.

On the other hand, in other attributes there is an overlap, but there is almost no difference between them. This shows that they have a low relationship with the label.

The Attributes *abdominal_distention*, *respiratory_rate*, *abdomen* and *total_protein* also discriminate well the label, but there is a greater variance between classes.

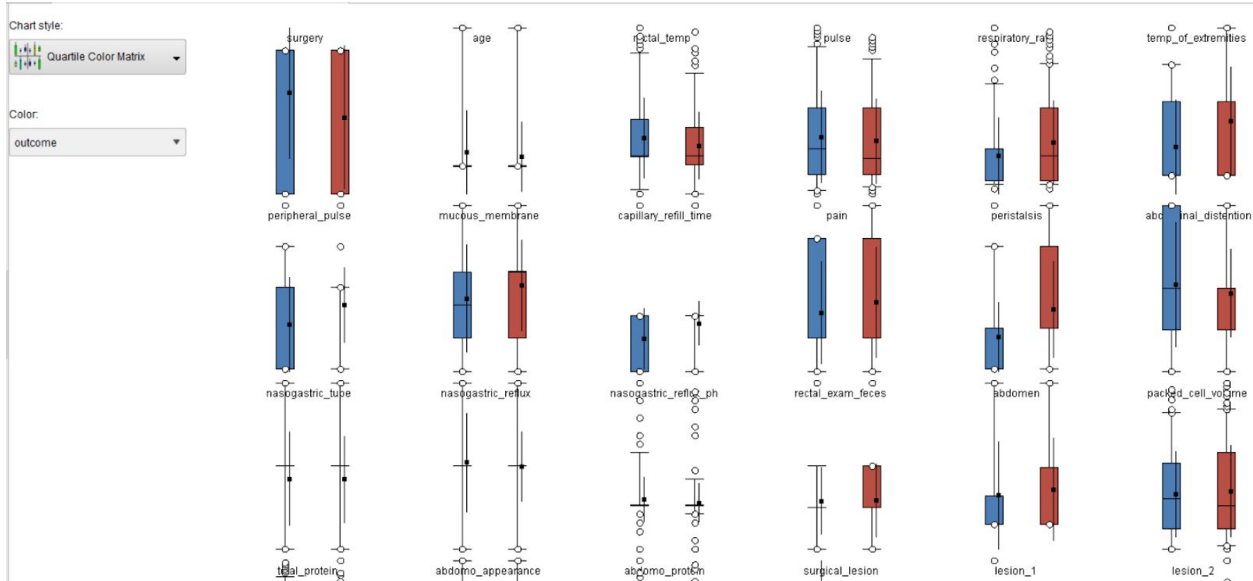


Fig 1.5.2 – Quartile Color Matrix graph

c. Bubble Chart

In the bubble chart, we can view four features at the same time, using the graphical tools to specify the x and y axes, as well as a third dimension expressed as the size of the bubble that represents the feature. The target class is indicated by the color. This chart also helps us to analyze outliers.

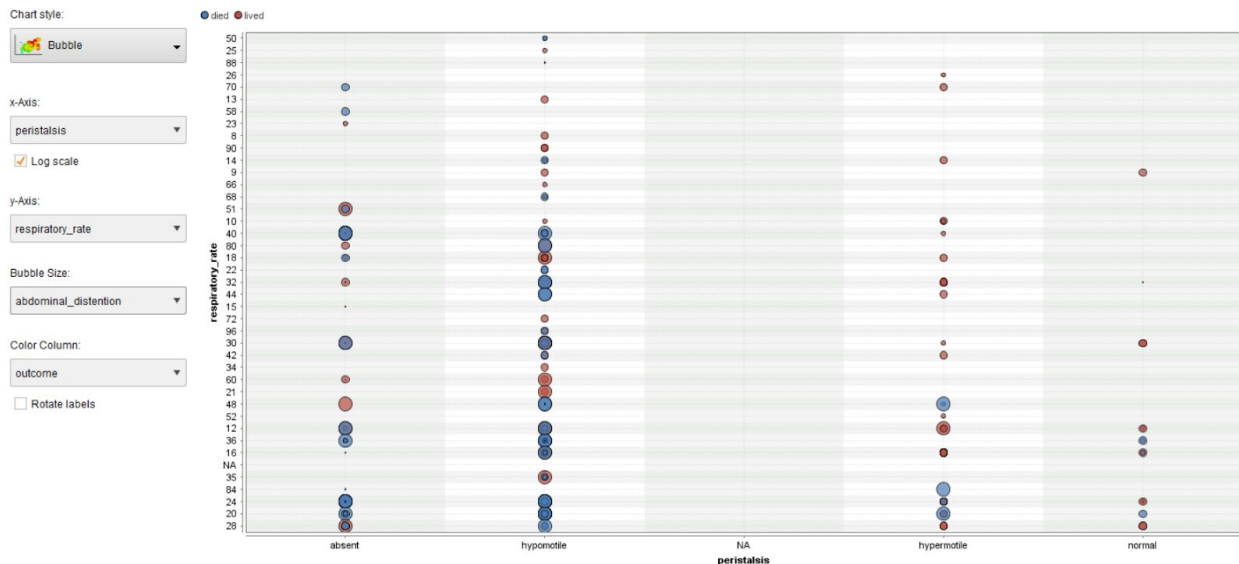


Fig 1.3.3 – Bubble chart

d. Decision Tree

A decision tree was applied to see the important and relevant attributes for the model. With that, we realized that the suggested attribute that was *lesion_1*, that is, the type of injury that the animal had, would be the most important attribute.

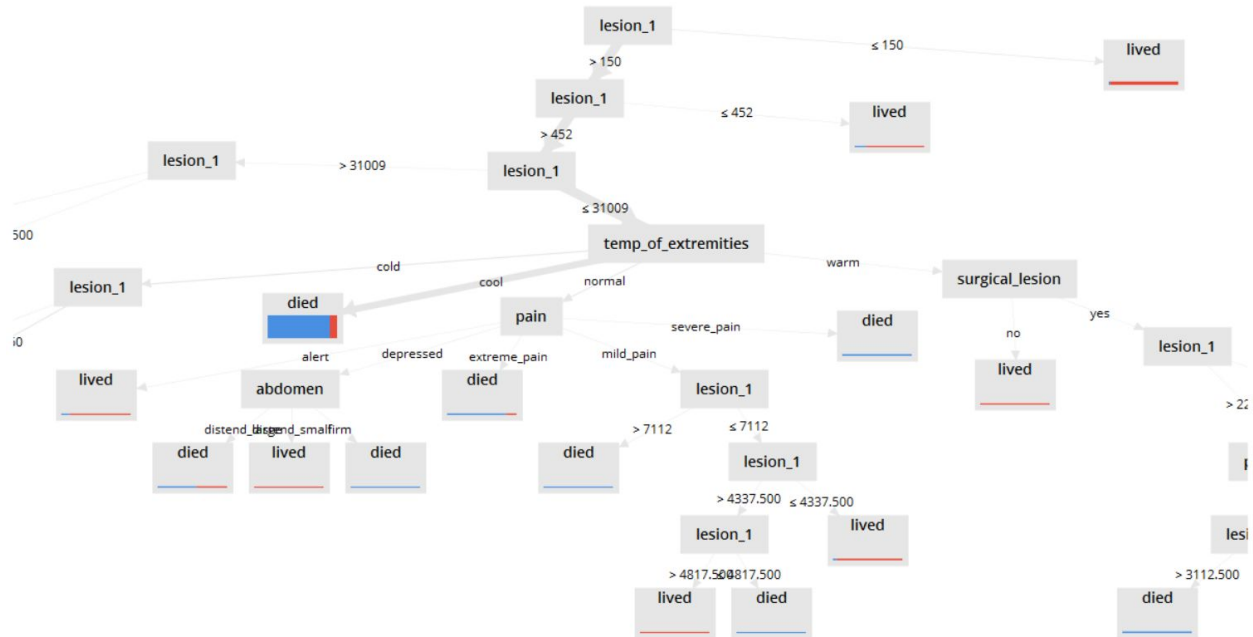


Fig 1.5.4 - Decision Tree

AttributeWeights (Decision Tree)

attribute	weight ↓
lesion_1	0.694
surgical...	0.088
peripher...	0.066
lesion_2	0.059
abdomen	0.048
pain	0.036
temp_of...	0.010

Fig 1.5.5 - Decision Tree - weights of attributes

2. Pre-processing

The base was normalized to give equal weights to the attributes and to decrease the convergence time of the algorithms. For this, the operator used was **Normalize** in RapidMinner.

Also to reduce the dimensionality of the attributes, a selection of attributes was made by filters (feature weights) in order to decrease the cost of learning, increase the accuracy of the algorithm and generate compact models that are easier to interpret.

Wrappers were not used as they are computationally more expensive and have the risk of overfitting.

In this step we also use 3 different models, k-NN, Random Forest and SVM.

2.1. Model k-NN

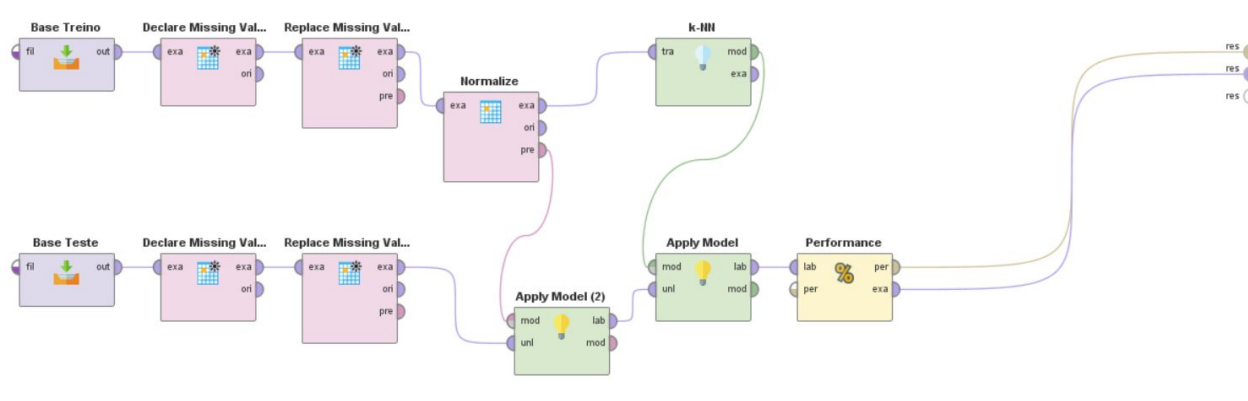


Fig 2.1.1 -. Model k-NN

For this first analysis, $N = 5$ was used.

Open in

Turbo Prep

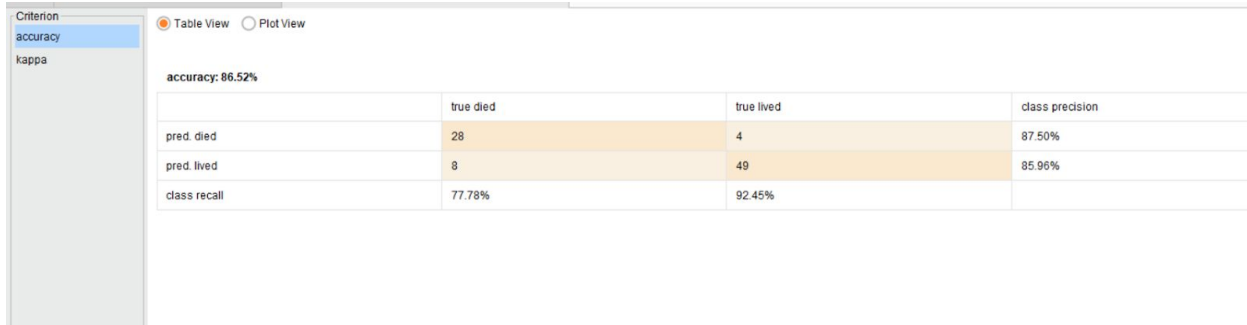
Auto Model

Filter (89 / 89 examples): all

Row No.	hospital_nu...	outcome	prediction(o...	confidence(died)	confidence(lived)	lesion_1	lesion_2	lesion_3	surgery	age	rectal_temp	pulse
1	1	died	died	0.608	0.392	0.118	-0.139	-0.058	no	adult	37.3	104
2	2	lived	lived	0.386	0.614	-0.286	-0.139	-0.058	no	adult	39.1	72
3	3	lived	lived	0.387	0.613	0.086	-0.139	-0.058	yes	adult	37.2	42
4	4	died	lived	0.250	0.750	-0.677	-0.139	-0.058	no	young	38	92
5	5	lived	lived	0.383	0.617	0.101	-0.139	-0.058	yes	adult	37.6	64
6	6	lived	lived	0.388	0.612	-0.101	-0.139	-0.058	yes	adult	38.6	42
7	7	lived	lived	0	1	-0.101	-0.139	-0.058	yes	young	38.3	130
8	8	lived	lived	0.192	0.808	0.085	-0.139	-0.058	yes	adult	37.8	48
9	9	lived	lived	0.367	0.633	-0.469	-0.139	-0.058	yes	adult	38	100
10	10	died	died	0.808	0.192	0.638	-0.139	-0.058	no	adult	38	104
11	11	died	died	0.806	0.194	0.286	-0.139	-0.058	no	adult	38.3	112
12	12	died	died	0.612	0.388	-0.084	-0.139	-0.058	yes	adult	38	120
13	13	died	died	0.826	0.174	0.638	-0.139	-0.058	yes	adult	38.9	80
14	14	lived	lived	0	1.000	-0.101	-0.139	-0.058	no	adult	38.6	46
15	15	died	died	0.612	0.388	0.638	-0.139	-0.058	yes	young	38.6	160
16	16	died	died	1	0	-0.083	-0.139	-0.058	yes	adult	38	64
17	17	died	died	1	0	0.101	-0.139	-0.058	no	adult	38	96
18	18	lived	lived	0.377	0.623	-0.286	-0.139	-0.058	yes	adult	38	64

Fig 2.1.2 reliable analysis

Analyzing the confusion matrix we see that for $N = 5$ the accuracy was 86.52%

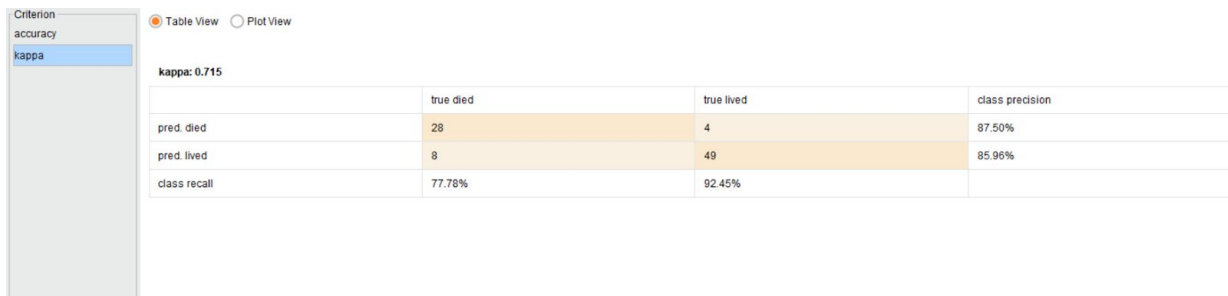


	true died	true lived	class precision
pred. died	28	4	87.50%
pred. lived	8	49	85.96%
class recall	77.78%	92.45%	

accuracy: 86.52%

Fig 2.1.3 - Analysis of the Accuracy

And the Kappa of 0.715.



	true died	true lived	class precision
pred. died	28	4	87.50%
pred. lived	8	49	85.96%
class recall	77.78%	92.45%	

kappa: 0.715

Fig 2.2.4 - Kappa

analysis Continuing the analysis, the K was changed to 8. With this, the accuracy increases to 88.76% and the Kappa to 0.763. Increasing K to 10, the accuracy and Kappa do not change.

2.2) Random Forest

Random Forest was the second model applied.

For this case, although it is not so necessary to balance the data set, once the proportion between parts were not so different, we conclude that the result was better with the balanced base.

In this case, we apply the operator **SMOTE upsampling**.

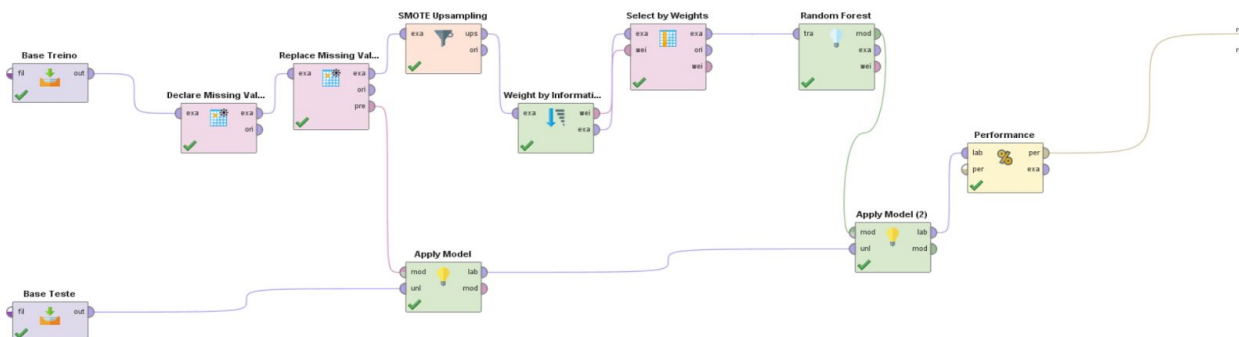


Fig 2.2.1 - Random Forest Model

The parameters applied in the Random Forest model were, *Tree = 100* and *criterion = gain_ratio*. For these parameters, the accuracy was 87.64, slightly higher than the accuracy given for the k-NN model.

accuracy: 87.64%

	true died	true lived	class precision
pred. died	36	11	76.60%
pred. lived	0	42	100.00%
class recall	100.00%	79.25%	

Fig 2.2.3 - Confusion matrix - Random Forest - criterion = gain_ratio

The criterion was changed to *Gini index*, this increased the accuracy to 91.01%.

accuracy: 91.01%

	true died	true lived	class precision
pred. died	36	8	81.82%
pred. lived	0	45	100.00%
class recall	100.00%	84.91%	

Fig 2.2.4 - Parameter confusion matrix - Gini Index

2.3) SVM

The third model applied was the SVM.

Since SVM only accepts binomial values, the nominal operator to numeric was used both on the training and testing basis. For this case there was no balancing or use of filters or wrappers.

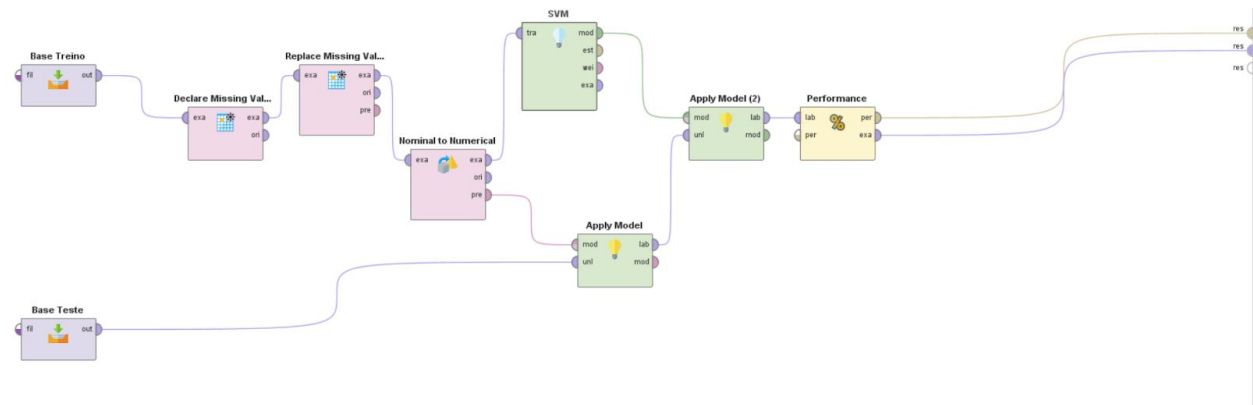


Fig 2.3.1 - SVM

model This model presented a confusion matrix with an improvement in the accuracy of its analyzes, 91.01%.

<div> <div>Criterion</div> <div>accuracy</div> </div> <div> <div>Table View</div> <div>Plot View</div> </div>			
accuracy: 91.01%			
	true died	true lived	class precision
pred. died	31	3	91.18%
pred. lived	5	50	90.91%
class recall	86.11%	94.34%	

Fig 2.3.2 - Confusion matrix

Making adjustments to the complexity parameters, L pos and L neg, there was an improvement in accuracy to 98.88%.

SVM (Support Vector Machine)

kernel cache

200

ⓘ

C

0.0

ⓘ

convergence epsilon

0.01

ⓘ

max iterations

100000

ⓘ

☒ scale

ⓘ

L pos

8.0

ⓘ

L neg

8.0

ⓘ

Fig 2.3.3 - Adjustment of SVM parameters

accuracy: 98.88%			
	true died	true lived	class precision
pred. died	35	0	100.00%
pred. lived	1	53	98.15%
class recall	97.22%	100.00%	

Fig 2.3.4 - Adjusted confusion matrix

2.4) Naive Bayes

The fourth model used was the Naive Bayes.

Also indicated for classification problems, this model is based on probability (Bayes' theorem).

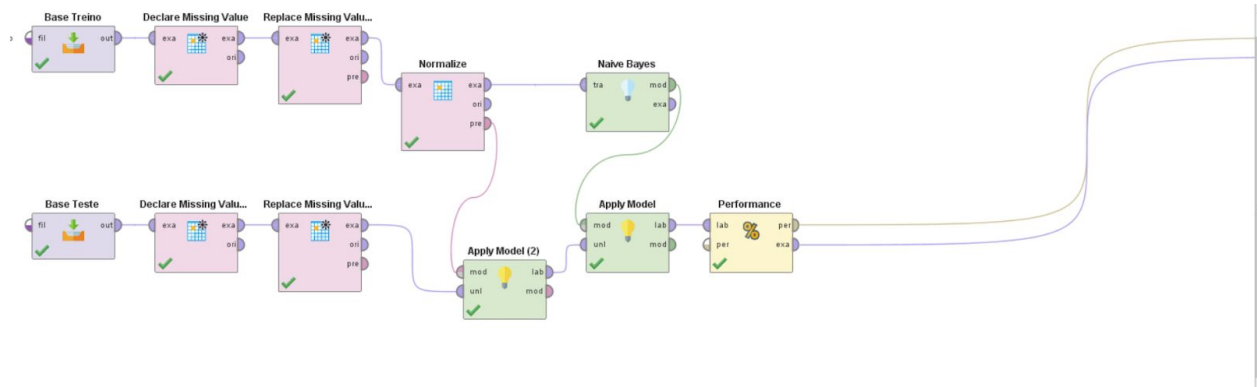


Fig 2.3.3 - Naive Bayes model

Using the default parameter (laplace correction) we obtained the confusion matrix below, with 71.91% accuracy and Kappa 0.475.

accuracy: 71.91%

	true died	true lived	class precision
pred. died	36	25	59.02%
pred. lived	0	28	100.00%
class recall	100.00%	52.83%	

Fig 2.3.4 - Naive Bayes confusion matrix - laplace correction = yes

When the laplace correction parameter was not used, the accuracy increased to 82.02% and Kappa 0.652

accuracy: 82.02%

	true died	true lived	class precision
pred. died	36	16	69.23%
pred. lived	0	37	100.00%
class recall	100.00%	69.81%	

Fig 2.3.5 - Naive Bayes confusion matrix - laplace correction = no

3. Others considerations

For the purposes of assessing the improvement in accuracy, a comparison was made between the application or not of weights for the k-NN, SVM and Naive Bayes models, but as there was no difference in the accuracy result, the original operators shown were maintained in this report. Weights were only applied in the Random Forest model.

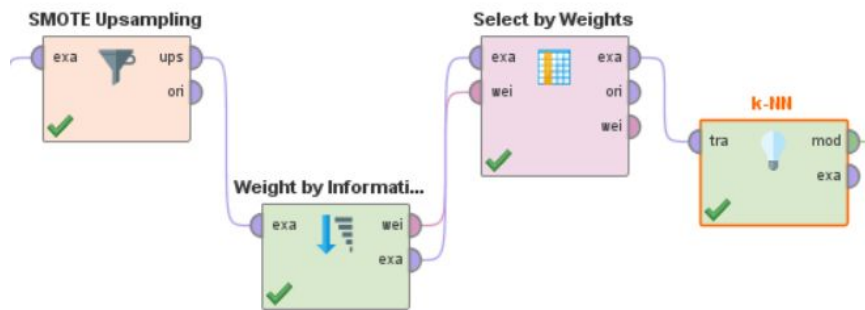


Fig 3. Example of application of the selection by weights (filter) in the k-NN

4. Conclusion

We can conclude through the matrix below, that based on the accuracy we see that the best model that applies in the Horse colic case would be the **SVM**.

Modelo	Acurácia	Kappa
k-NN	88.76 %	0.763
Random Forest	91.01 %	0.755
SVM	98.88 %	0.977
Naive Bayes	82.02 %	0.652

5. Acknowledgments

My thanks to Professor Manoela and also the coordination of the course for teaching us and presenting this discipline.

Annex I - Data dictionary

Attribute Information:

1: *surgery?*

1 = Yes, it had surgery

2 = It was treated without surgery

2: *Age*

1 = Adult horse

2 = Young (<6 months)

3: *Hospital Number*

- numeric id

- the case number assigned to the horse
(may not be unique if the horse is
treated > 1 time)

4: *rectal temperature*

- linear

- in degrees celsius.

- An elevated temp may occur due to
infection.

- temperature may be reduced when
the animal is in late shock

- normal temp is 37.8

- this parameter will usually change as
the problem progresses, eg. may start
out normal, then become
elevated because of the injury, passing
back through the normal range as the
horse goes into shock

5: *pulse*

- linear

- the heart rate in beats per minute

- is a reflection of the heart condition:

30 -40 is normal for adults

- rare to have a lower than normal rate
although athletic horses may have a
rate of 20-25

- animals with painful lesions or
suffering from circulatory shock may
have an elevated heart rate

6: *respiratory rate*

- linear

- normal rate is 8 to 10

- usefulness is doubtful due to the great
fluctuations

7: *temperature of extremities*

- a subjective indication of peripheral
circulation

- possible values:

1 = Normal

2 = Warm

3 = Cool

4 = Cold

- cool to cold extremities indicate
possible shock

- hot extremities should correlate with
an elevated rectal temp.

8: *peripheral pulse*

- subjective

- possible values are:

1 = normal

2 = increased

3 = reduced

4 = absent

- normal or increased pp are indicative of adequate circulation while reduced or absent indicate poor perfusion

9: ***mucous membranes***

- a subjective measurement of color

- possible values are:

1 = normal pink

2 = bright pink

3 = pale pink

4 = pale cyanotic

5 = bright red / injected

6 = dark cyanotic

- 1 and 2 probably indicate a normal or slightly increased circulation

- 3 may occur in early shock

- 4 and 6 are indicative of serious circulatory compromise

- 5 is more indicative of septicemia

10: ***capillary refill time***

- a clinical judgment. The longer the refill, the poorer the circulation

- possible values

1 = <3 seconds

2 = > 3 seconds

11: ***pain - a subjective judgment of the horse's pain level***

- possible values:

1 = alert, no pain

2 = depressed

3 = intermittent mild pain

4 = intermittent severe pain

5 = continuous severe pain

- should NOT be treated as a ordered or discrete variable!

12: ***peristalsis***

- an indication of the activity in the horse's gut. As the gut becomes more distended or the horse

becomes more toxic, the activity decreases

- possible values:

1 = hypermotile

2 = normal

3 = hypomotile

4 = absent

13: ***abdominal distension***

- An IMPORTANT parameter.

- possible values

1 = none

2 = slight

3 = moderate

4 = severe

- an animal with abdominal distension is likely to be painful and have reduced gut motility.

- a horse with severe abdominal distension is likely to require surgery just to relieve the pressure

14: ***nasogastric tube***

- this refers to any gas coming out of the tube

- possible values:

1 = none

2 = slight

3 = significant

- a large gas cap in the stomach is likely to give the horse discomfort

15: ***nasogastric reflux***

- possible values

1 = none

2 = > 1 liter

3 = <1 liter

- the greater amount of reflux, the more likelihood that there is some serious obstruction to the fluid passage from the rest of the intestine

16: ***nasogastric reflux PH***

- linear

- scale is from 0 to 14 with 7 being neutral

- normal values are in the 3 to 4 range

17: ***rectal examination - feces***

- possible values

1 = normal

2 = increased

3 = decreased

4 = absent

- absent feces probably indicates an obstruction

18: ***abdomen***

- possible values

1 = normal

2 = other

3 = firm feces in the large intestine

4 = distended small intestine

5 = distended large intestine

- 3 is probably an obstruction caused by a mechanical impaction a Normally na is treated medically

- 4 and 5 Indicate the surgical lesion

19: ***packed cell volume***

- straight

- the # by volume of red cells in the blood

-normal range is 30 to 50. The level rises to the compromised circulation or the passe the animal becomes dehydrated.

20: ***total protein***

- linear

- normal values lie in the 6-7.5 (gms / dL) range

- the higher the value the greater the dehydration

21: ***abdominocentesis appearance***

- a needle is put in the horse's abdomen and fluid is obtained from the abdominal cavity

- possible values:

1 = clear

2 = cloudy

3 = serosanguinous

- normal fluid is clear while cloudy or serosanguinous indicates a compromised gut

22: ***abdominocentesis total protein***

- linear

- the higher the level of protein the more likely it is to have a compromised gut. Values are in gms / dL

23: **outcome**

- what eventually happened to the horse?

- possible values:

1 = lived

2 = died

3 = was euthanized

24: **surgical injury?**

- retrospectively, was the problem (injury) surgical?

- all cases are either operated upon or autopsied so that this value and the lesion type are always known

- possible values:

1 = Yes

2 = No

25, 26, 27: **type of lesion**

- first number is site of lesion

1 = gastric

2 = sm intestine

3 = lg colon

4 = lg colon and cecum

5 = cecum

6 = transverse colon

7 = retum / descending colon

8 = uterus

9 = bladder

11 = all intestinal sites

00 = none

- second number is type

1 = simple

2 = strangulation

3 = inflammation

4 = other

- third number is subtype

1 = mechanical

2 = paralytic

0 = n / a

- fourth number is specific code

1 = obturation

2 = intrinsic

3 = extrinsic

4 = adynamic

5 = volvulus / torsion

6 = intussuption

7 = thromboembolic

8 = hernia

9 = lipoma / splenic incarceration

10 = displacement

0 = n / a

28: **cp_data**

- is pathology data present for this case?

1 = Yes

2 = No