CSC8004 - Data Mining - Major Project Report

Analysis of Clinical Data from Horses with Colic

Name: Diana Kither

Table of Contents

i) Project Topic

For the major project, option three was selected and a data mining research project was conducted on a data set containing information relating to the clinical findings in horses suffering from colic. The data set, Horse Colic Data, that is publicly available from the UCI Machine Learning Repository, was utilised to investigate the relationship between signs on clinical examination and the underlying cause of colic. The question this project aims to answer is,

'Which clinical findings from examination of a horse suffering from colic are the strongest predictors for the animal requiring surgery?'

ii) Background

a) Introduction

Colic is defined as abdominal pain in a horse and is one of the most common clinical presentations in these animals for which a veterinarian will be consulted (Cook & Hassel 2014; Freeman 2018). Colic is a clinical sign rather than a specific disease and has a number of underlying causes that can range in severity. Colic can result from mild causes, for example gas distension of the bowel, where animals recover quickly and uneventfully with medical treatment. However, colic can also result from very serious underlying disease for example bowel obstructions or torsions and can be fatal. These cases require rapid referral to specialist equine veterinary hospitals for surgery, hospitalisation and intensive care. Colic is one of the leading causes for premature death in horses (*Colic Emergencies* 2020). Approximately 7-10% of cases of colic require surgical treatment (Cook & Hassel 2014; Fraiwan & Abutarbuch 2020). In some cases these patients are in clear need of referral or surgery however this may be less clear in other cases (Cook & Hassel 2014).

Veterinarians are most commonly assessing and treating cases of colic in the field rather than in a hospital setting. Thus decisions with respect to treatment at home or in hospital need to be made and if referral or surgery is required it needs to be arranged quickly. Depending on the location of equine hospital services, animals may or may not need to be transported significant distances to hospital. One of the most important decisions being made by the veterinarian in a case of a horse with colic is whether the animal may require surgery (Cook & Hassel 2014; Fraiwan & Abutarbuch 2020).

b) The Clinical Examination

To create a clinical picture as to the severity of disease and potential underlying causes of colic in a horse, the veterinarian must take a detailed history and perform a thorough clinical examination in the field. There are multiple important elements to the clinical examination of a horse with colic. An examination starts with measurement of body temperature, heart rate and respiratory rate or what is otherwise known as a TPR measurement. The colour of an animal's gums (mucous membranes) and strength of pulses should also be assessed. Abdominal sounds are auscultated with a stethoscope to evaluate intestinal motility. The veterinarian will also pass a nasogastric tube via the nose into the stomach to assess for the presence of gas or gastrointestinal reflux/fluid.

Some additional procedures as part of an examination may be performed by the veterinarian depending on the initial clinical appearance of an animal and facilities that are available where the animal is being examined. Rectal examination enables a veterinarian to assess for the presence of faeces in the rectum and palpate some aspects of the abdomen. A rectal examination may be conducted by the veterinarian if indicated in a clinical case and if facilities available in the field allow for this procedure to be conducted safely for both the patient and veterinarian.

Abdominocentesis is a procedure which involves obtaining abdominal or peritoneal fluid from the abdomen if present via a needle. It is not performed routinely in cases of colic unless a case is appearing more severe. Fluid is tested for levels of protein and cellular content which may direct to the level of severity or underlying cause of colic.

iii) Motivation

As a practicing veterinarian myself, I have treated colic in horses in the past and one of the defining factors in successfully treating life threatening causes of colic is prompt referral (*Colic Emergencies* 2020; Freeman 2018). Assessments of an animal's clinical status are based on clinical examination findings as well as the results of minor procedures conducted in the field. This enables a clinician to create a clinical picture of how unwell an animal is and to determine a list of potential diagnoses of the underlying cause.

As discussed in the introduction, one of the most foremost decisions being made by the veterinarian in a case of colic is whether the animal may require surgery, a decision that needs to be made as soon as possible (Cook & Hassel 2014; Fraiwan & Abutarbuch 2020, Thoefner et al 2003). Rapid referral of cases which require surgery can significantly impact an animal's survival and the development of further complications (*Colic Emergencies* 2020; Fraiwan & Abutarbuch 2020; Freeman 2018). There are multiple clinical findings that would give significant clinical concern regarding a case and would result in a veterinarian's opinion being that referral was required and surgery may be indicated. However this decision is not always clear cut and the status of a clinical case can change quickly (Cook & Hassel 2014).

iv) Problem Formulation

This project is investigating the following question,

'Which clinical findings from examination of a horse suffering from colic are the strongest predictors for the animal requiring surgery?'

To examine this question further, two data mining techniques were selected, decision tree classification and the Apriori algorithm for association rule mining.

There are some potential challenges to consider with this project. There are both numerical and categorical variables within the data set. For association rule mining using the Apriori algorithm, these variables will need to be appropriately binned and changed to categorical variables. Approximately 30% of the data values within the data set are missing. Decisions made on how to handle these missing values will influence the results of data mining.

Furthermore, in using these methods, associations may be found which are considered significant based on results of the algorithm but not in context of a clinical situation. Therefore results will need to be interpreted with respect to prior knowledge of the clinician.

v) Literature Review

Equine colic is one of the most common clinical presentations to equine veterinarians in the field (Fraiwan & Abutarbuch 2020; Freeman 2018,). By taking a detailed history and performing a thorough clinical examination, the veterinarian then forms a clinical picture to determine the animal's clinical status, the severity of colic, potential underlying causes, treatment measures and most importantly whether the underlying cause may require referral or surgery. Delay in referral if required can significantly affect prognosis (Curtis et al. 2015; Thoefner et al 2003)

There are several clinical indicators that would prompt a veterinarian to refer a case of colic. These include severe abdominal pain, recurrent abdominal pain that is not responsive to analgesia, if heart rate exceeds 60 beats per minute, if no intestinal sounds are present, if the horse's abdomen appears distended, abnormal findings on rectal examination, severe gastrointestinal reflux and abnormalities in abdominal fluid analysis (Cook & Hassel 2014; Freeman 2016). If an animal presents with any of these clinical signs, referral would be recommended.

Research has been conducted into both the significance of clinical signs in relation to the severity of a case of colic as well as development of computer modelling to aid decision making in clinical cases (Curtis et al 2015; Fraiwan and Abutarbuch 2020; Thoefner et al 2003). Fraiwan and Abutarbuch (2020) further highlighted that the use of artificial intelligence and machine learning as part of assessment of clinical cases is increasing in human medical fields compared to very limited use in the field of veterinary medicine.

Durchame et al (1989) developed a decision tree model for classifying of cases of colic as surgical or non-surgical based on clinical findings. In this study, clinical signs that ranked in highest order of importance in classifying these cases were abdominal distension, rectal examination findings and abdominal fluid characteristics (Durchame et al, 1989).

Thoefner et al (2003) performed a study to develop a classification and regression model for predicting whether cases of colic required surgery. In this model severity of pain was the initial classifier (Thoefner et al 2003). Body temperature, abdominal fluid analysis and rectal examination findings were also included in the model (Thoefner et al 2003). The study highlighted that the model developed did predict a larger number of false positives, horses requiring surgery that were in fact not surgical cases (Thoefner et al 2003). However if model accuracy could be improved then there may be a place for modelling in clinical decision making (Thoefner et al 2003).

In a case review conducted by Curtis et al (2015), logistic regression analysis was used in an effort to determine which clinical signs were important in predicting more severe cases of colic. The model developed by this study indicated that heart rate, pain, pulse character, the colour of a patient's gums (mucous membranes) and intestinal sounds as important parameters in a veterinarian's examination to determine the severity of a clinical case (Curtis et al 2015). Reeves et al (1991) also used regression modelling to determine the importance of clinical signs in assessment of horses with colic, finding that severity of abdominal pain, pulse character, rectal examination findings and abdominal sounds were more significant in decision making regarding the need for surgery.

Fraiwan and Abutarbuch (2020) used the findings from clinical cases of colic presented to a veterinary teaching hospital to test the efficacy of machine learning algorithms in determining whether surgery was indicated in these cases. This study found that the most significant clinical signs in the decision making process were whether the animal had experienced previous episodes of colic, if the animal had been referred by another veterinarian compared to if being presented by the owner for a first opinion consultation, rectal examination findings, abdominal fluid analysis and the

presence of gastrointestinal fluid reflux (Fraiwan & Abutarbuch, 2020). Furthermore the study showed that machine learning and artificial intelligence models were accurate in their predictions and that these techniques may be useful in the veterinary medicine in the future (Fraiwan & Abutarbuch, 2020).

As is evident from the above examples, there is some consistency as well as some slight variation in clinical signs that were determined to be more significant in deciding if surgery was indicated in cases of colic from the various modelling techniques.

vi) System Design and Architecture

There are three proposed steps in the investigation of this problem, data pre-processing and exploratory data analysis, applying data mining techniques and knowledge presentation.

a) Data Pre-Processing and Exploratory Data Analysis

The data set is publicly available with one file containing data for 300 cases, a second file containing a further 68 cases and a third file containing variable names and descriptions. The two case files will be combined and variable names added. This single text file and will then be uploaded into a data mining platform. The chosen platform for this project is RapidMiner.

The first initial step in processing the data set will be to conduct exploratory data analysis (EDA) to visualise, understand and gain initial insights from the data set prior to mining. This will be conducted in RapidMiner. EDA will also aid data cleaning through identification of outliers and missing values so a plan can be formulated to determine how these values will be managed in the data set prior to mining. Microsoft Excel will also be used to aid data cleaning.

Data reduction will involve further assessment of variables to determine which variables are relevant to the analysis as well as correlation and chi square analyses to investigate the strength of variables and determine if there is redundancy between variables. This will enable the final selection of variables for data mining.

Data transformation is the final step of data pre-processing. As discussed below, two data mining techniques will be used in the project. The data set comprises both numerical and categorical variables. Prior to application of the Apriori algorithm to the data set, numerical variables will be transformed into categorical variables through binning techniques. For some numerical variables data transformation will also be performed in relation to techniques used for the management of missing values.

b) Data Mining

Decision tree classification and the Apriori algorithm for association rule mining have been selected to investigate associations between clinical findings from examination of horses with colic and clinical cases requiring surgery for this project. This will be performed using the RapidMiner platform.

c) Knowledge Presentation

The final results of the project and insights gained from analysis will then be presented.

A diagram of the problem solving design and system architecture for the project is presented in Appendix A, Figure 4.

vii) The Data Set

a) Data Set Description

A publicly available data set was sourced from the UCI Machine Learning Repository. The data set was donated by the University of Guelph, Canada and collated the clinical examination findings and initial diagnostic results of 368 horses who presented with colic. In all cases a final diagnosis of the underlying cause of disease and whether surgery was required is known.

The data set however is an older data set having been published in 1989. Despite this, the clinical presentation of colic in horses has not greatly changed during this time. What has advanced however are knowledge, skills and technology in the areas of anaesthesia, advanced diagnostics, surgical technique, intensive care and medical treatment of these cases (*Colic Emergencies* 2020; Freeman 2018). Thus it was deemed to still be appropriate to use this data set. This data set can be accessed at https://archive.ics.uci.edu/ml/datasets/Horse+Colic.

b) Data Pre-Processing and Exploratory Data Analysis

Data pre-processing and exploratory data analysis (EDA) were performed on the data set in RapidMiner and Microsoft Excel prior to data mining in RapidMiner. After the training and test data files were downloaded from the repository they were then opened in text editor to view the files in their original form. The test data was then copied into the training data file to combine all results. The data file was then converted to a comma separated values (.csv) file entitled horse-colic_final.csv for ease of opening in RapidMiner and Excel. Variable columns in the data file required labelling. A third file, the variable name file, contains the names and descriptions of all the variables in the data set. This variable name file was downloaded and horse-colic_final.csv was opened in Excel so headings could be added to the variable columns. The descriptions of each variable in the data set and the values denoting these variables are presented in Appendix B.

Horse-colic_final.csv was uploaded into Rapid Miner so the original data set could be viewed and analysed. When initially uploading the data set, some variable types were altered so these were appropriately adjusted during file upload. Missing values in the file were denoted by question marks which was accounted for in file upload. Further data pre-processing and exploratory data analysis was conducted in RapidMiner and a diagram of this process is presented in Appendix C, Figure 5.

The original data set contained 368 cases and 28 variables. The next steps taken with data pre-processing were to identify identification and label variables. The label or predictor variable in this case is Surgical_Lesion. This variable denotes if a colic case resulted from surgical or non-surgical causes based on pathology results. The purpose of this analysis is to determine which variables have the strongest association with the value of the variable Surgical_Lesion being surgical. Hospital ID (Hosp_ID) was set as an identification variable.

At this point necessary data reduction was conducted by removing variables that were deemed not relevant to this analysis including the variables indicating lesion type (Lesion_Type_1, Lesion_Type_2, Lesion_Type_3), other pathology data (CP) and whether the animal did have surgery (Surgery), thus reducing the number of variables. The Select Attributes operator was used to perform this task. What is also clearly evident from looking at initial data statistics in Rapid Miner is the number of missing values present for almost all variables so the next steps taken were to assess each variable, their distribution and to manage missing values before completing further exploratory data analysis.

After uploading the data set to RapidMiner, statistics and distribution of variables can be viewed in the results window. The variable Age is denoted by two numerical values, 1 indicating an adult horse

and 2 indicating a horse under 6 months of age. When looking at the distribution of Age as shown in Appendix D, Figure 6, there are 340 cases with a value of 1 and no cases with a value of 2. However, 28 cases had a value of 9, an unknown value. If these cases were removed from the data set, all the remaining cases would have the same age thus the variable would have no effect on analysis and was therefore it was removed from the data set.

The next variables assessed were those representing the first aspects of the clinical examination, rectal temperature (Temp), heart rate (HR), respiratory rate (RR), mucous membrane colour (MM) and capillary refill time (CRT). Given that these variables are considered essential parts of an initial examination of an animal, any cases where these values were missing where removed from the data set. The Filter Examples operator was used and the total number of cases in the data set was reduced to 214.

The next variables considered represent further points of assessment in the process of a clinical examination, the presence of gut sounds (Peristalsis), peripheral pulse assessment (PP), severity of pain (Pain), degree of abdominal distension (Abdo_Dist) and temperature of the extremities (Temp_Ext). 7.5% of cases in the data set had missing values for Peristalsis. These values were replaced by the mode for Peristalsis, 3 = hypomotile, which accounted for 46% of cases in the data set. The next most common value was normal, accounting for 21% of cases. Peripheral pulse had 15% of cases with missing values and these were also replaced by the mode value, normal = 1 which was shared by 52% of cases. The variable Pain had 22 cases with missing values. Given the frequency of the mode for Pain was very close to that of the second most common value as seen its distribution (Appendix D, Figure 7), these 22 cases were removed. Cases with missing values for Temp_Ext were also removed for the same reasoning. Abdo_Dist is considered an important parameter and at this stage of analysis 12 cases were missing this value. When looking at the distribution for this variable (Appendix D, Figure 8), the mode is evident however there is not a great difference between it and the next most frequent values. Given the importance of this parameter these 12 cases where removed from the data set rather than the value being replaced. The Filter Examples operator was again used to remove cases and the Replace operator to replace missing values. After this next step in data pre-processing case numbers were reduced to 167.

The next two variables addressed are the simple diagnostic tests of packed cell volume (PCV) and total protein (TP). On initial examination of the distribution of TP as shown Appendix D, Figure 9, there is a very wide distribution of values for TP also seen with the mean for TP calculated as 24.7 that is less than its standard deviation of 27.7. In Figure 9 it is also evident that there is a division in values when viewing the histogram which is also clear when looking at ordered values for TP, with an absence of values between 13 and 46. Total protein can be measured in units of g/L and g/dL which differ by a factor of 10. Thus it was assumed this distribution of results occurred due to differences in units when total protein was measured in some cases. Thus values greater than 46 were assumed to be g/L and as such converted to g/dL for consistency in measurements. This variable was then renamed to TP_gdL. This adjustment was performed in Excel prior to analysis of the data set in RapidMiner. The new distribution is shown in Appendix D, Figure 10. Given that TP_gdL and also PCV have skewed distributions when shown graphically, their missing values were replaced with their median value. The Aggregate operator was used to calculate the median values and the Replace Missing Values operator was used to replace values. The distribution for PCV is shown in Appendix D, Figure 11.

The final group of variables assessed related to minor procedures that may be performed in the field as part of an examination. Rectal examination (Rectal_Ex) may not always be performed with every case of colic. Rectal examination is often not performed in mild cases of colic. In addition facilities may not be available at the location of an examination which enable rectal examination to be

performed safely. Thus a new value for the variable was created, 0 = Unknown, to replace missing values. Abdominal assessment (Abdomen) which is done via rectal palpation again may not always be performed. For this variable the value 2 = Other represents other cases, thus it was used to replace missing values for this variable. The Replace operator was used.

Nasogastric tubing (NGT) should ideally be performed as part of a clinical exam however may not be performed in mild cases of colic. The value of 1 = no gas, indicating no gas was present at nasogastric intubation, was used to replace missing values as cases that may not have been tubed would also fit into this category. Nasogastric fluid pH (NG_pH) had a significant number of missing values at 123. These values were replaced with a value of 0 assuming that the pH level was either not measured as not all cases will have reflux or if present levels may not have been measured in the field. Again the Replace operator was used. After replacing missing values, to aid interpretation of NG_pH, values were assigned to bins at a unit of 1. The Discretize by User Specification operator was used.

Nasogastric reflux or presence of fluid when a nasogastric tube is passed (NGR) had 46 missing values. When compared to NG_pH in a contingency table, 43 cases had no pH measured so these cases were assumed to have no reflux and were allocated an NGR value of 1 = no reflux. The Replace Missing Values operator was used. 3 cases however had NG_pH measurements indicating the presence of reflux even though the value for NGR was missing. These cases thus cannot be considered to have no reflux and therefore were assigned the mode value of the variable, 3 = < 1 Litre of reflux. The Pivot operator was used to create contingency tables, the cases and respective row numbers were located in the data set with their NG_pH values and the Set Data operator was used to replace missing values.

Abdominocentesis is a procedure where fluid is obtained from the abdomen via a needle for analysis. This procedure is not routinely performed for all cases of colic so the large number of missing values related to the two variables associated with this procedure was expected. 103 out of the remaining 167 cases had missing values for the variable Centesis_Protein which represents the total protein level measured in abdominal fluid. A value for 0 was allocated to cases where protein levels were missing or unknown, assuming it was not measured. This variable was also binned into units of 1gm/dL for ease of understanding. The Discretize by User Specification and Replace operators were used. The appearance of abdominal fluid (Centesis) had 70 missing values. When compared to Centesis_Protein in a contingency table, 68 values had no protein measurements. These cases were given a Centesis value of 0 = unknown, assuming that abdominocentesis was not performed. Two cases had Centesis_Protein measurements but no value for fluid appearance. These two cases were assigned the mode value for this variable, 3 = serosanguinous. The Pivot and Set Data operators were used.

This cleaned data set now consists of 167 cases with 15 categorial variables, 5 numerical variables and 2 special variables (label and ID).

After data cleaning, the data set was then further examined with exploratory data analysis. When looking at the distribution of each variable some important points were noted. The distribution of the label or response variable Surgical_Lesion had an even distribution between surgical and non-surgical cases with 98 surgical cases and 69 non-surgical cases. The incidence of surgical causes of colic in the field is between 7 and 10% thus the data set is inconsistent with an expected case distribution (Cook & Hassel 2014; Fraiwan & Abutarbuch 2020). However to investigate the problem of this project with data mining techniques an even distribution of cases is required.

Across variables for initial clinical examination findings, distributions where skewed with higher measures of clinical findings closer to the normal range and much lower rates of extreme findings as would be expected. The distributions of both heart rate and respiratory rate were skewed to the right by larger values with a maximum 184 and 80 respectively (Appendix D, Figures 12, 13). These very high these values are physiologically possible in a severe case of colic and thus these cases were not removed as they may be important outliers. Total protein levels had a similar skewed distribution with a large maximum value of 13. Rectal temperature however had a normal distribution.

Variables denoting procedures that were performed in the field were skewed towards unknown or not performed as they are more likely to be performed with more severe cases of colic which are less frequently seen. Rectal examination however had a very even distribution across possible values.

The data set was further explored with correlation analysis of numerical variables and weight of chi-square was also calculated to investigate the strength of the relationship between variables and the label variable, Surgical_Lesion. This was conducted in RapidMiner with the Correlation and Weight by Chi-Square operators with the default settings. The RapidMiner process and results of these analyses are shown in Appendix E (Figures 14, 15 and 16). There was not a high degree of correlation between numerical variables. As expected there was a moderate correlation between heart rate and respiratory rate otherwise the remaining numerical variables had low levels of correlation. When examining weight by chi-square, pain had the highest degree of weight on surgical outcome followed by abdominal examination findings, degree of abdominal distension, heart rate and appearance of abdominal fluid.

viii) Snapshots – User Interface

The platform RapidMiner was used for the data mining project. In addition, Microsoft Excel and Text Editor were utilised initially as part of data pre-processing. Snap shots of all three of these user interfaces are shown below.

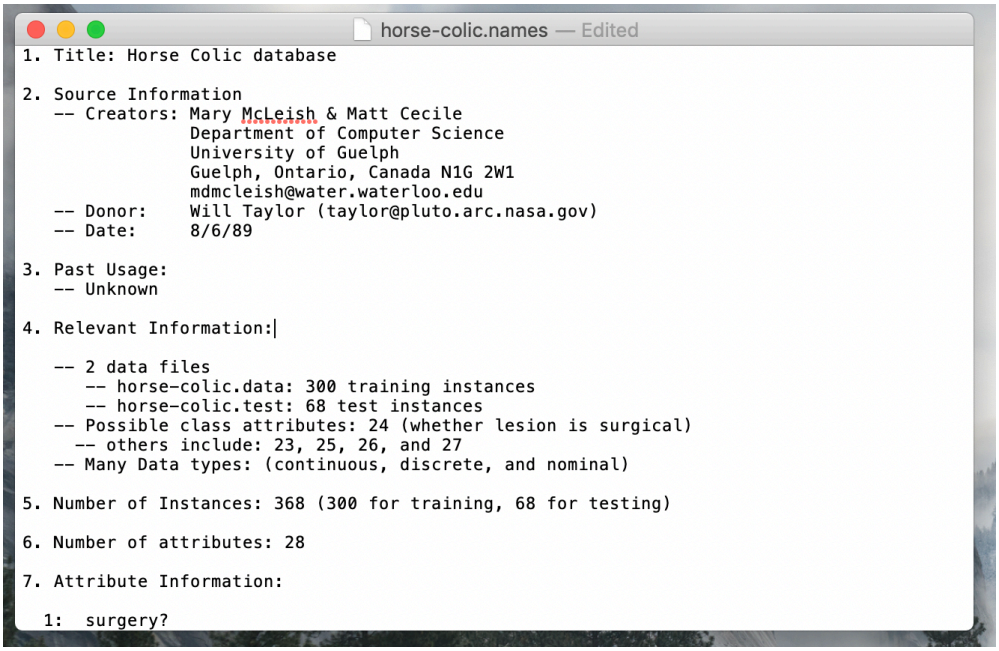


Figure 1a - Original Data Files in Text Editor (1)

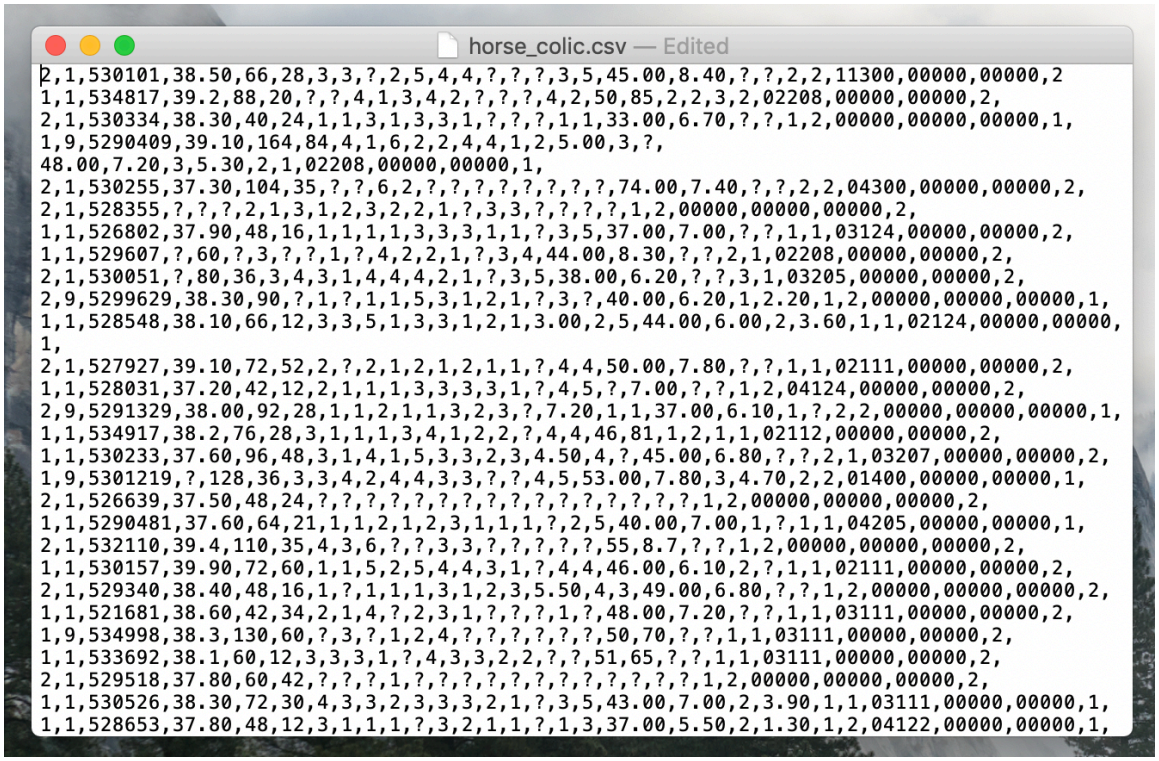


Figure 1b - Original Data Files in Text Editor (2)

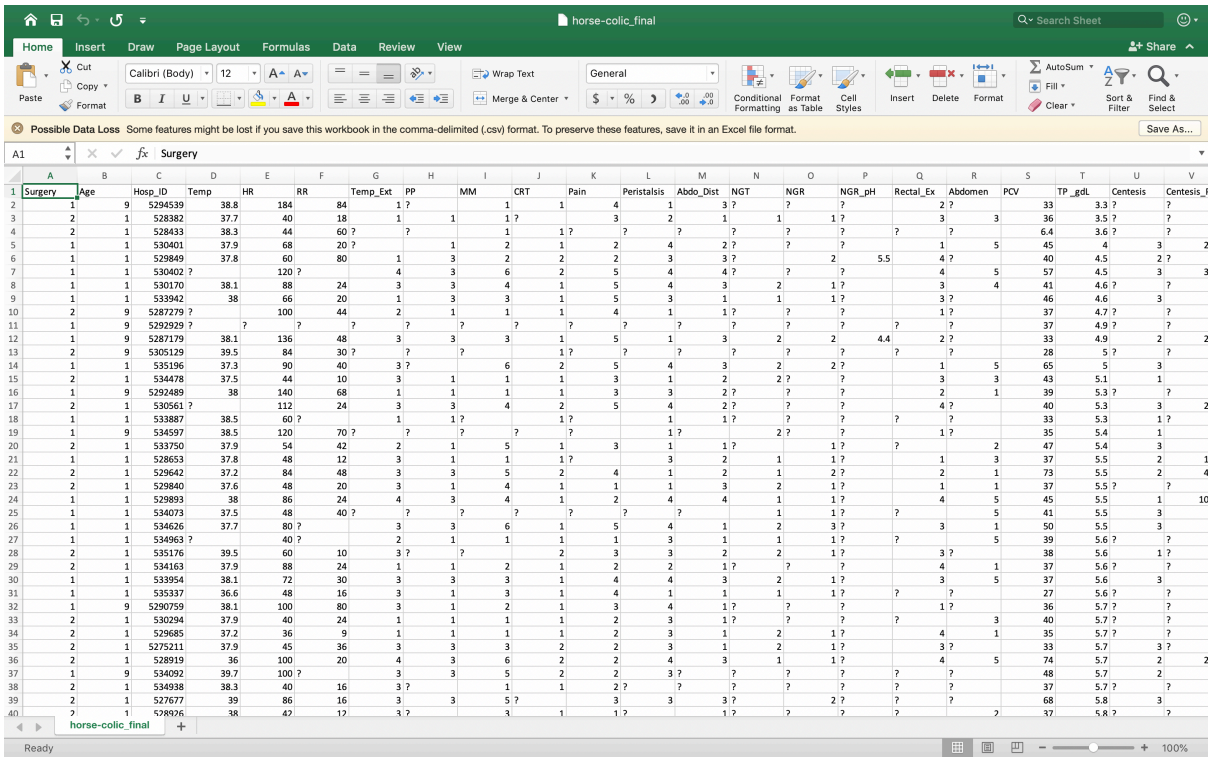


Figure 2 – User Interface Microsoft Excel

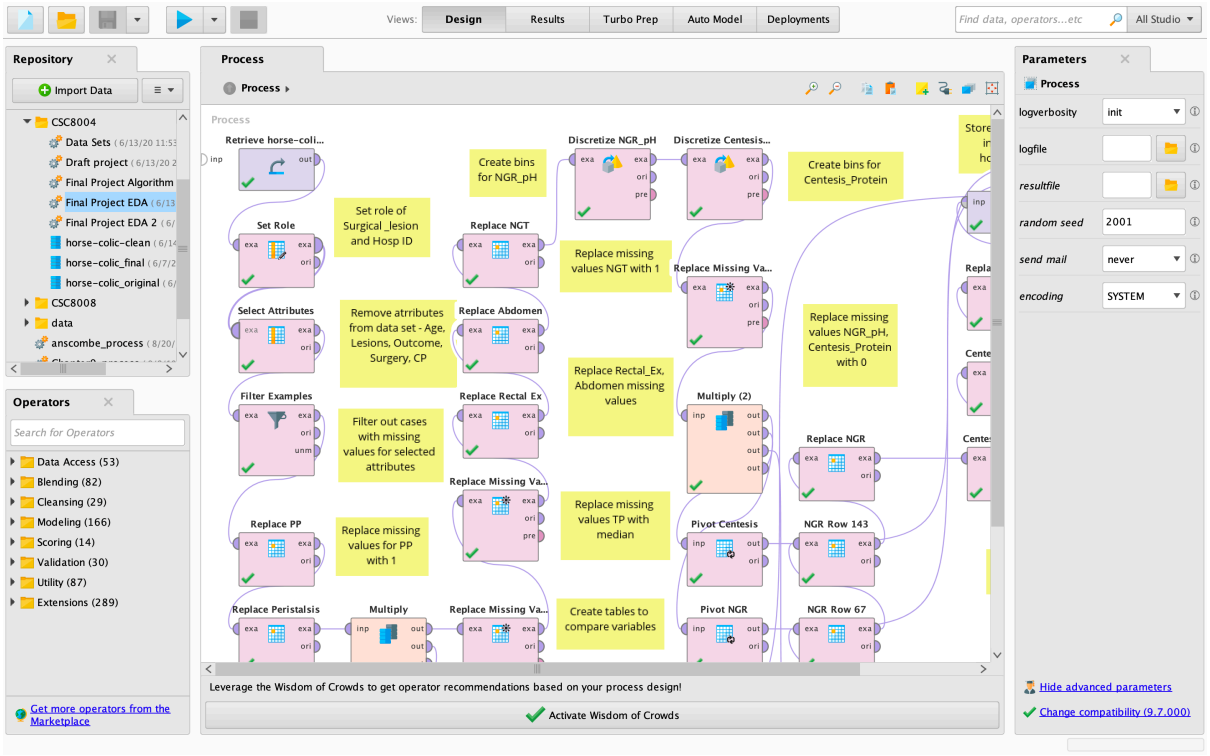


Figure 3a – User Interface RapidMiner (1)

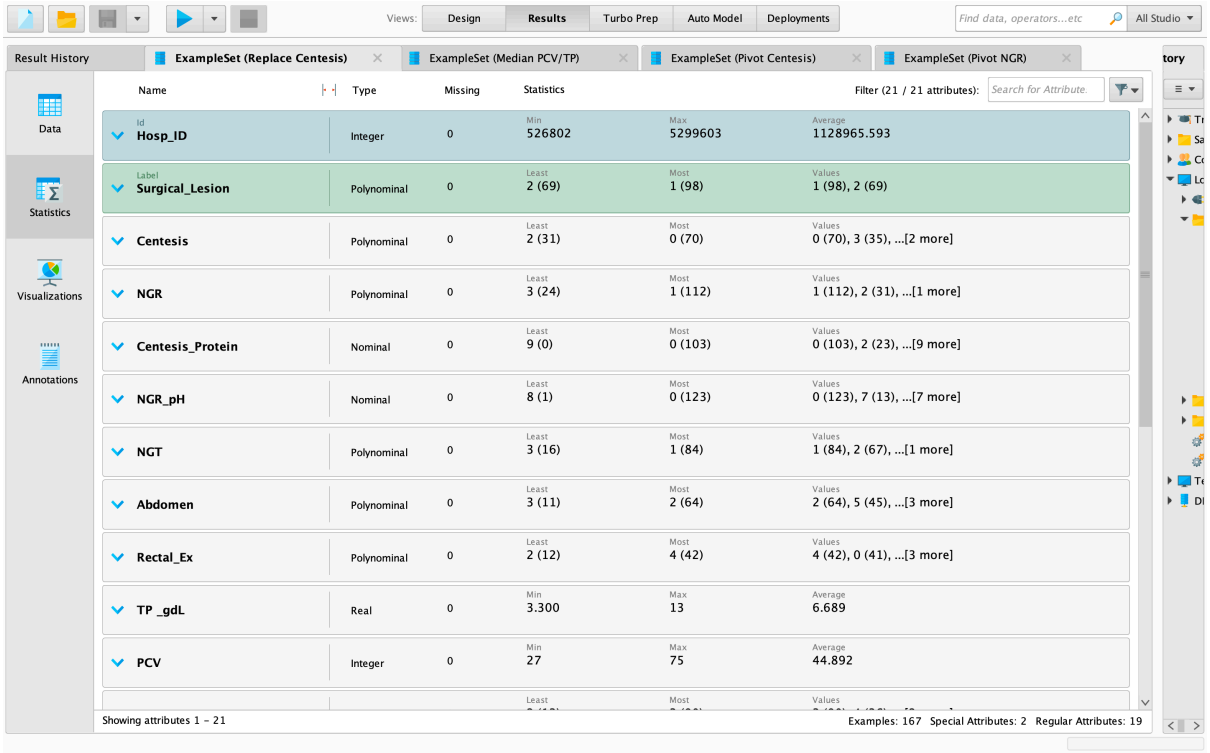


Figure 3b – User Interface RapidMiner (2)

iv) Data Mining User Instructions - The Apriori Algorithm and Decision Tree Classification

To examine the strength of the relationships between different clinical findings and surgical causes of colic, two data mining techniques were applied. The first was the Apriori algorithm for association rule mining and the second was decision tree classification. The RapidMiner processes for each algorithm are shown in Appendix F, Figure 17 and 18.

a) The Apriori Algorithm and Association Rule Mining

The Apriori algorithm was applied to the data set in RapidMiner to identify frequent item sets and generate strong association rules from them. The W-Apriori operator from the WEKA package in RapidMiner was used to perform this task. Prior to running the algorithm, all numerical variables were classified into bins using the Discretize by User Specification operator which enables bins to be classified by the user. Variables were grouped into equal width bins based on the variable's units. Bins were named with the lower margin value of the bin. Bin names for each variable are shown in the table below.

Table 1 – Bin names of numerical variables in the data set

Variable	**Bins**
HR	40, 60, 80, 100, 120, 140, 160, 180
RR	10, 20, 30, 40, 50, 60, 70, 80
Temp	36, 37, 38, 39, 40
PCV	30, 40, 50, 60,70
TP_gdL	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

The final step in data modification prior to running the algorithm is to set the variable Surgical_Lesion to a regular variable using the operator Set Role so it is included in the Apriori algorithm.

As a first step the W-Apriori operator was run on the data set with the default settings of minimum support of 0.1 and minimum confidence of 0.9. This however did not yield any frequent item sets. The minimum support was then lowered to 0.5 which yielded 20 best rules. As relationships between animals requiring surgery is being investigated association rules containing the variable Surgical_Lesion in particular were examined.

b) Decision Tree Classification

The second data mining technique used with the data set was decision tree classification. A decision tree model was built using RapidMiner in an effort to classify clinical cases of colic as surgical or non-surgical. The decision tree was built through an internal process within the Cross-Validation operator. This enables the data set to be split into training and test sets for the purpose of building and testing the model. The decision tree operator was used in the internal process to build the model. The Gini-index was selected as the separation criterion with a confidence level of 0.1. The Apply operator then used to apply the model and the performance operator was used to measure its performance. The model produced is shown in Appendix G, Figure 19 and 20.

x) Findings, Lessons and Experiences

a) Findings – The Apriori Algorithm and Association Rule Mining

The results of the Apriori algorithm run in RapidMiner are displayed below.

W-Apriori

```
Apriori
=======

Minimum support: 0.1 (17 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 39

Size of set of large itemsets L(2): 51

Best rules found:

 1. Pain=1 32 ==> Surgical_Lesion=2 28    conf:(0.88)
 2. Peristalsis=1 28 ==> Surgical_Lesion=2 18    conf:(0.64)
 3. Peristalsis=4 36 ==> PP=3 23    conf:(0.64)
 4. Rectal_Ex=3 39 ==> NGT=2 23    conf:(0.59)
 5. Centesis=1 31 ==> Surgical_Lesion=2 18    conf:(0.58)
 6. Rectal_Ex=1 33 ==> Surgical_Lesion=2 19    conf:(0.58)
 7. Abdo_Dist=2 42 ==> Surgical_Lesion=2 24    conf:(0.57)
 8. Centesis=3 35 ==> PP=3 20    conf:(0.57)
 9. Abdo_Dist=3 48 ==> PP=3 27    conf:(0.56)
10. Centesis=2 31 ==> NGT=2 17    conf:(0.55)
11. Centesis=1 31 ==> NGT=2 17    conf:(0.55)
12. Centesis=1 31 ==> Temp_Ext=1 17    conf:(0.55)
13. RR=10 44 ==> Surgical_Lesion=2 24    conf:(0.55)
14. RR=30 33 ==> Abdomen=5 18    conf:(0.55)
15. CRT=2 39 ==> PP=3 21    conf:(0.54)
16. PCV=30 56 ==> Surgical_Lesion=2 29    conf:(0.52)
17. Rectal_Ex=1 33 ==> PCV=30 17    conf:(0.52)
18. RR=30 33 ==> PP=3 17    conf:(0.52)
19. Rectal_Ex=1 33 ==> Temp_Ext=1 17    conf:(0.52)
20. Centesis=3 35 ==> CRT=2 18    conf:(0.51)
```

When lowering the minimum confidence to 0.5, 20 best rules were obtained. The strongest association rule was found to be between Pain = 1 (alert, no pain) and Surgical_Lesion = 2 (non-surgical case). This would be expected as it would be unlikely that an animal with minimal or no pain would have a surgical cause of colic, more pronounced signs of pain would be expected. The second strongest rule was between Peristalsis = 1 (hypermotile gut) and Surgical_Lesion = 2 (non-surgical case). Again this was expected as horses with increased gut sounds or increased gut movement are unlikely to have surgical causes of colic. In surgical cases gut motility is most likely reduced or absent.

The next association rules related to Surgical_Lesion were between Centesis = 1 (clear abdominal fluid appearance, normal), Rectal_Ex = 1 (normal) and Abdomen = 2 (other findings). Again these cases did not require surgery and had clinical examination findings that were normal or in the case of the variable Abdomen, the category of other, which includes cases where this was likely not examined due to a case being less severe. Several other examination findings were present in association rules including NGT, PCV, PP, CRT and Temp. Again the associations between these other clinical signs were as expected. No association rules however were found with Surgical_Lesion = 1 or that is cases where surgery was required. Furthermore no frequent item sets that containing more than one item were found.

b) Findings - Decision Tree Classification

The decision tree model produced by RapidMiner from the data set had a sensitivity of 74.49% in predicting surgical cases of colic however had quite a high false positive rate with a specificity of just 56.52%. The overall accuracy of the model in predicting whether a case of colic was surgical was 67.21%.

The first clinical finding considered when classifying cases of colic was pain. If pain levels were high at 4 or 5 then a case immediately classified as surgical which is to be expected as many surgical causes of colic are very painful. If pain levels were lower, the appearance of abdominal fluid and abdominal examination findings from rectal examination were the next clinical signs considered when classifying cases. In mild cases of colic, temperature was the next classifying variable. Additional variables included in the decision tree where rectal examination findings, degree of abdominal distension, temperature of body extremities, heart rate and total protein levels. As expected animals with abnormal findings on rectal examination and those developing abdominal distension were likely to be surgical cases. Animals with more severe abnormalities of temperature of body extremities, heart rate and total protein levels were also more likely to be surgical cases which is expected as it is these clinical signs that alter with dehydration and later shock which can ensue with surgical causes of colic.

c) Lessons and Experiences

This project enabled me to further develop my understanding of data mining algorithms and enabled me to compare the results of two algorithms when applied to the data set. The results from the Apriori algorithm were both expected as well as unexpected. The association rules found between clinical findings and non-surgical cases are what I would anticipate as a clinician if I was examining a mild or medical case of colic that was unlikely to require surgery. However I did not expect that no association rules would be found between clinical signs and cases that did require surgery. In addition other association rules between clinical findings were found that were expected however these were not considered relevant to the problem being examined in this analysis, highlighting that association rules can be strong but not necessarily interesting and that prior knowledge of the data set is also important in interpreting results of data mining techniques.

The decision tree classifier did yield more relevant results to the problem being investigated highlighting the information gained from applying different mining techniques to a data set. The clinical signs identified by the decision tree model in this analysis coincide with the findings from other studies discussed in the earlier literature review. In my own clinical experience these clinical findings would be included in the reasons why a veterinarian would be concerned the underlying cause of colic is surgical and advise referral. However one clinical sign that was not included in the decision tree model which would prompt concern for a surgical cause of colic is the presence of nasogastric reflux. The presence of reflux raises clinical concern for the potential of an obstruction of the bowel and it would also prompt a veterinarian to recommend referral of a clinical case.

Decision tree classification proved to be a more effective technique for investigating the problem however the accuracy of the model needs to be improved. In addition, the data set used was small and in the future using a larger data may assist with further modelling and improving model accuracy. The high level of missing values and decisions in managing these values would also influence results.

xi) Conclusions

The analysis highlighted that clinical findings related to severity of abdominal pain, intestinal motility, rectal examination findings, presence of abdominal distension, peritoneal fluid examination, heart rate, body temperature and temperature of body extremities are more strongly associated to cases of surgical colic.

These findings are similar to those highlighted in the literature however one clinical finding that would also cause concern that a case of colic was surgical that was not highlighted by this analysis is the presence of nasogastric reflux.

Decision tree classification proved to be more effective in investigating the problem than the Apriori algorithm and association rule mining however the accuracy of the model needs to be improved. The data set used in this analysis had a small number of cases and in future analyses a larger number of cases would be preferred which would likely also influence results.

xii) References

1. *Colic Emergencies* 2020, Large Animal Hospital College of Veterinary medicine University of Florida, USA, viewed 13/4/20, https://largeanimal.vethospitals.ufl.edu/hospital-services/surgery/colic/

2. Cook, VL, Hassel DM, 2014, 'Evaluation of Colic in Horses: Decision for Referral', *Veterinary Clinics of North America Equine Practice*, Vol 30, No 2, pp 383-398

3. Curtis L, Burford JH, Thomas JSM, Curran ML, Bayes TC, England GCW, Freeman SL, 2015, 'Prospective study of the primary evaluation of 1016 horses with clinical signs of abdominal pain by veterinary practitioners and the differentiation of critical and non-critical cases', *Acta Veterinaria Scandinavica,* Vol 57, No 69, https://doi.org/10.1186/s13028-015-0160-9

4. Durchame NG, Pascoe PJ, Lumsden JH, Durchame GR (1989), 'A computer-derived protocol to aid in selecting medical v surgical treatment in horses with colic', *Equine Veterinary Journal*, Vol 21, No 6, pp 447-450

5. Fraiwan MA, Abutarbush SM, 2020 'Using Artificial Intelligence to Predict the Survivability Likelihood and Need for Surgery in Horses Presented with Acute Abdomen (Colic)', *Journal of Equine Veterinary Science'* Vol 90, July 2020, https://www.sciencedirect.com/science/article/pii/S0737080620300642

6. Freeman DE, 2016, '*What not to do when referral is indicated'*, Pacific Veterinary Conference, viewed 6/6/20, https://www.vin.com/members/cms/project/defaultadv1.aspx?id=7351464&pid=15054&

7. Freeman, DE 2018, 'Fifty Years of Colic Surgery' *Equine Veterinary Journal*, Vol. 50, No. 4, pp 423-435

8. Guibersen J, 2017, *'Equine Colic: Medical and Surgical Cases'*, Wild West Veterinary Conference, Veterinary Information Network(VIN), viewed 6/6/20, https://www.vin.com/members/cms/project/defaultadv1.aspx?id=8216549&pid=19026&

9. Reeves MJ, Curtis CR, Salman MD, Stashak TS, Reif JS, 1991, 'Multivariate Prediction Model for the Need for Surgery with Horses with Colic' *American Journal of Veterinary Research,* Vol 52, No 11, 1903-7

10. Thoefner MB, Ersboll BK, Jansson N, Hesselholt M, 2003 'Diagnostic decision rule for support in clinical assessment of the need for surgical intervention in horses with acute abdominal pain', *Canadian Journal of Veterinary Research,* Vol 67, No 1 pp 20-29

xiii) Appendices

Appendix A

Figure 4: System Architecture for Data Mining Project

Appendix B

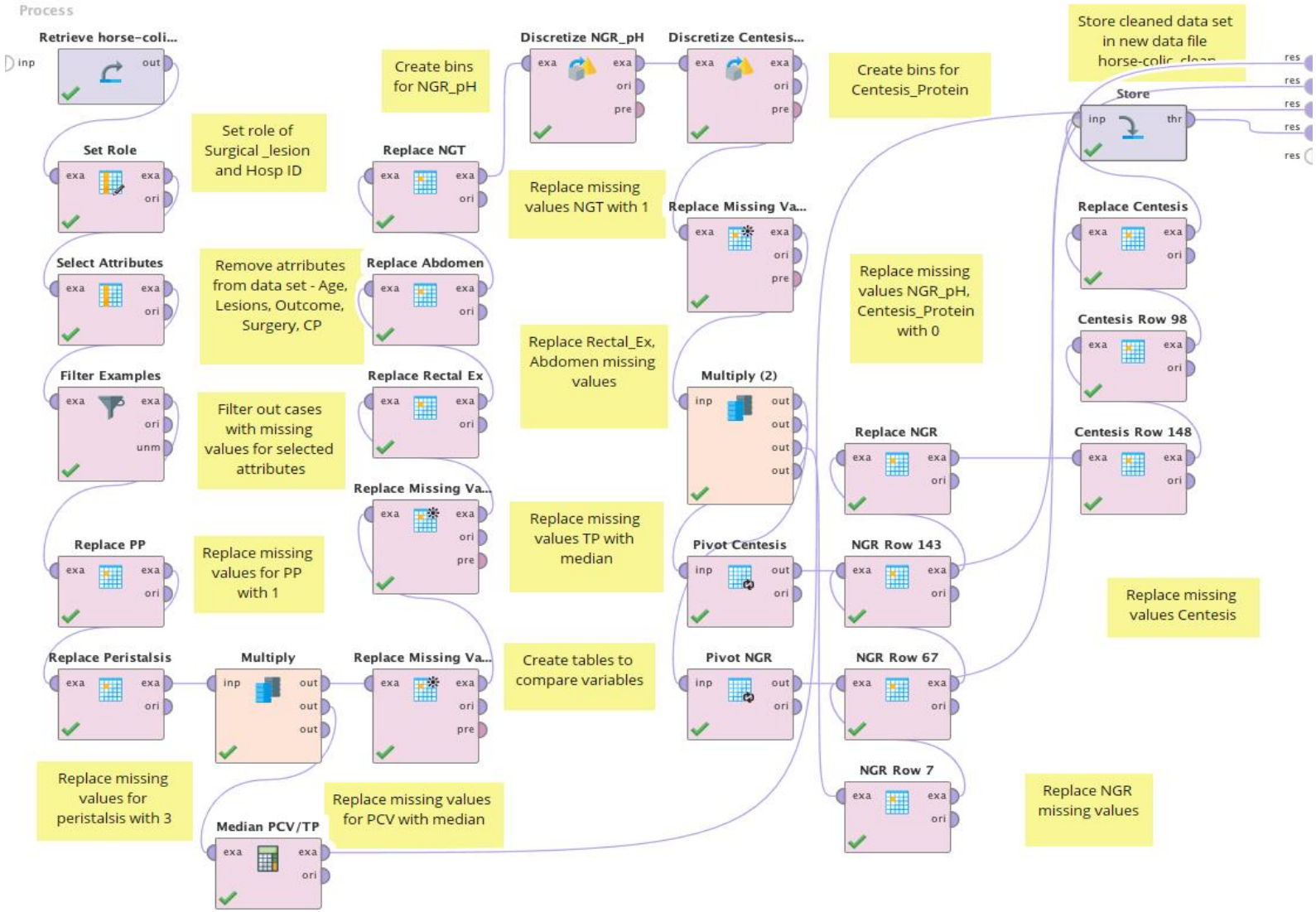
Descriptions of variables used in analysis from Horse Colic Database.
Source information: Created by Mary McLeish and Matt Cecile (1989) Department of Computer Science, University of Guelph, Guelph, Ontario, Canada N1G 2W1

Variable Name	Variable Description	Values
Surgery	- Whether the animal had surgery as part of treatment	1 = Had surgery 2 = Did not have surgery
Age		1 = Adult horse 2 = < 6 mths
Hosp_ID	- ID number assigned to each case	Numeric
Temp	- Rectal Temperature	Degrees Celsius Normal = 37.8
HR	- Heart Rate	Beats per minute Normal = 30-40
RR	- Respiratory Rate	Breaths per minute Normal = 8-10
Temp_Ext	- Temperature of extremities	1 = Normal 2 = Warm 3 = Cool 4 = Cold
PP	- Peripheral pulse assessment	1 = Normal 2 = Increased 3 = Reduced 4 = Absent
MM	- Colour of mucous membranes	1 = Normal pink 2 = Bright pink 3 = Pale pink 4 = Pale cyanotic 5 = Bright red, injected 6 = Dark Cyanotic
CRT	Capillary Refill Time	1 = < 3 secs 2 = > 3 seconds
Pain	Pain assessment of animal	1 = No pain 2 = Depressed 3 = Intermittent mild pain 4 = Intermittent severe pain 5 = Continuous severe pain

Peristalsis	Assessment of gut sounds and motility	1 = Hypermotile 2 = Normal 3 = Hypomotile 4 = Absent
Abdo_Dist	- Presence of degree of abdominal distension	1 = None 2 = Slight 3 = Moderate 4 = Severe
NGT	- Presence of gas coming from nasogastric tube	1 = None 2 = Slight 3 = Significant
NGR	- Presence of fluid reflux from nasogastric tube	1 = None 2 = < 1 Litre 3 = > 1 Litre
NGR_pH	- pH of fluid reflux	Measured on pH scale 1-14 0 = Unknown
Rectal_Ex	- Presence of faeces on rectal examination findings	0 = Unknown 1 = Normal 2 = Increased 3 = Decreased 4 = Absent
Abdomen	- Abdomen exam findings from rectal palpation	1 = Normal 2 = Other 3 = Firm faeces in colon 4 = Distended small intestine 5 = Distended large intestine
PCV	- Packed cell volume or number of red blood cells in blood	Integer value
TP	- Total protein level in blood	gms/dL
Centesis	- Appearance of fluid obtained on abdominocentesis	0 = Unknown 1 = Clear 2 = Cloudy 3 = Serosanguinous
Centesis_Protein	- Protein level of abdominal fluid	gms/dL 0 = Unknown
Outcome	- Outcome of case	1 = Lived 2 = Died 3 = Euthanised
Surgical_Lesion	- Was the lesion surgical, this is known retrospectively about each case due to pathology findings	1 = Yes 2 = No
Lesion_Type	- Cause of colic	Numerical values used to denote lesion type
CP	- Pathology data	1 = Yes 2 = No

Appendix C

Figure 5 - RapidMiner Process for Data Cleaning and initial Exploratory Data Analysis



Appendix D

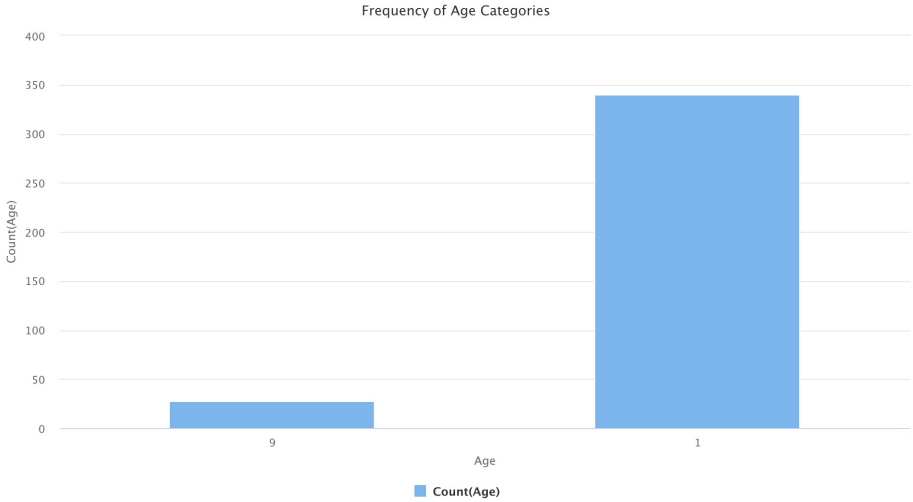


Figure 6 – Frequency of Age Categories in Horse Colic Data

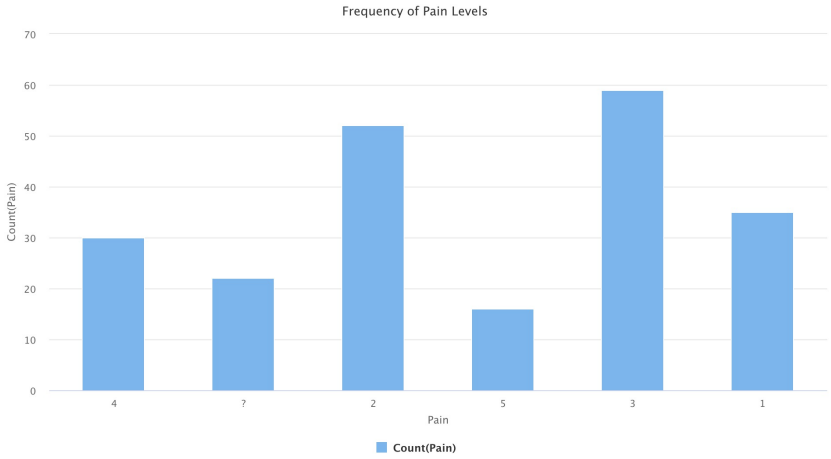


Figure 7 – Frequency of Pain Levels

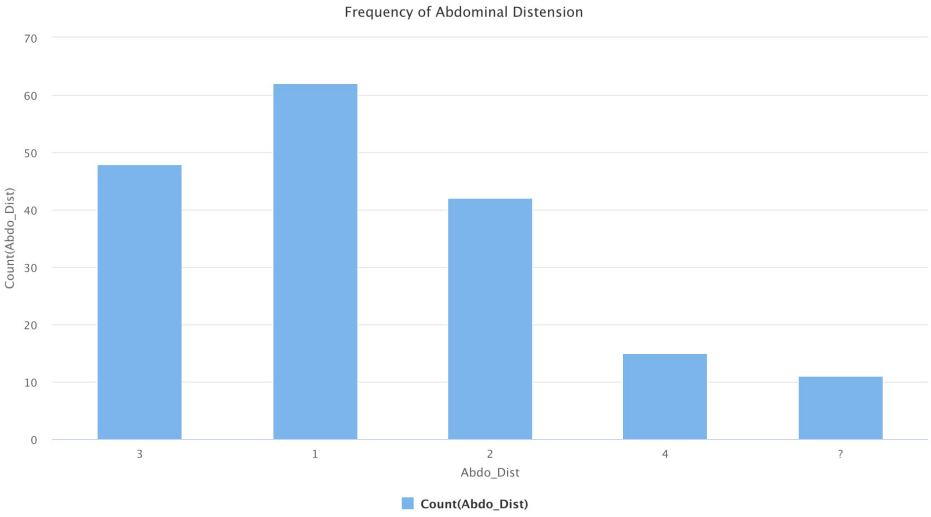


Figure 8 – Frequency of Abdominal Distension

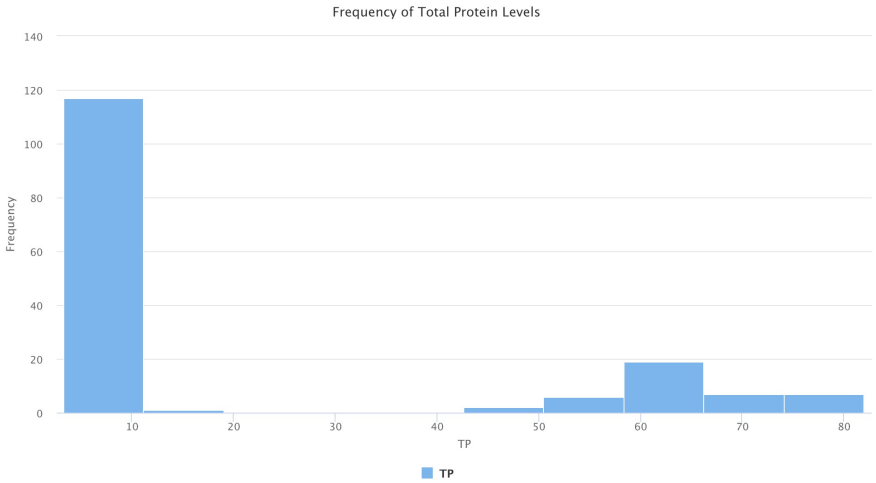


Figure 9 – Frequency of Total Protein Levels prior to adjustment of units

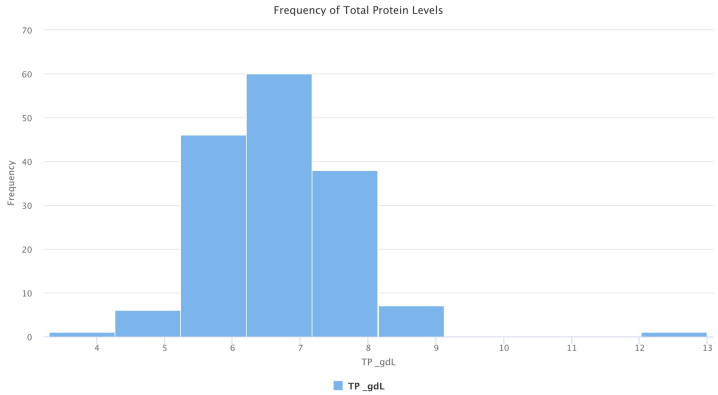


Figure 10 – Frequency of Total Protein Levels after unit adjustment

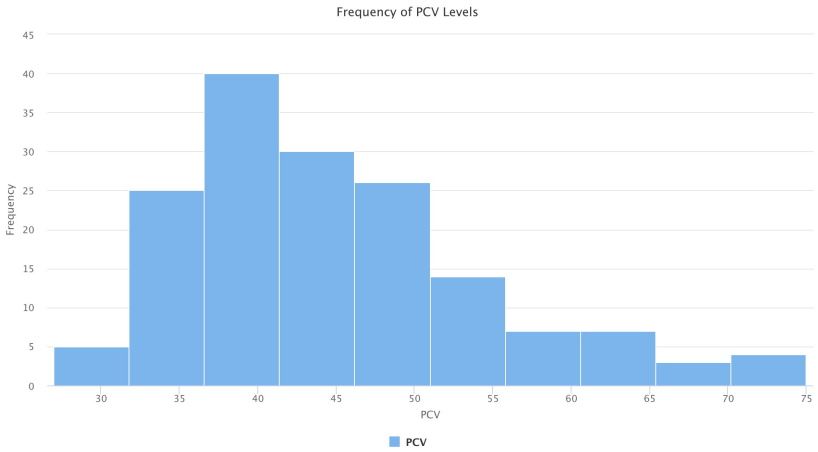


Figure 11 – Distribution of PCV Levels

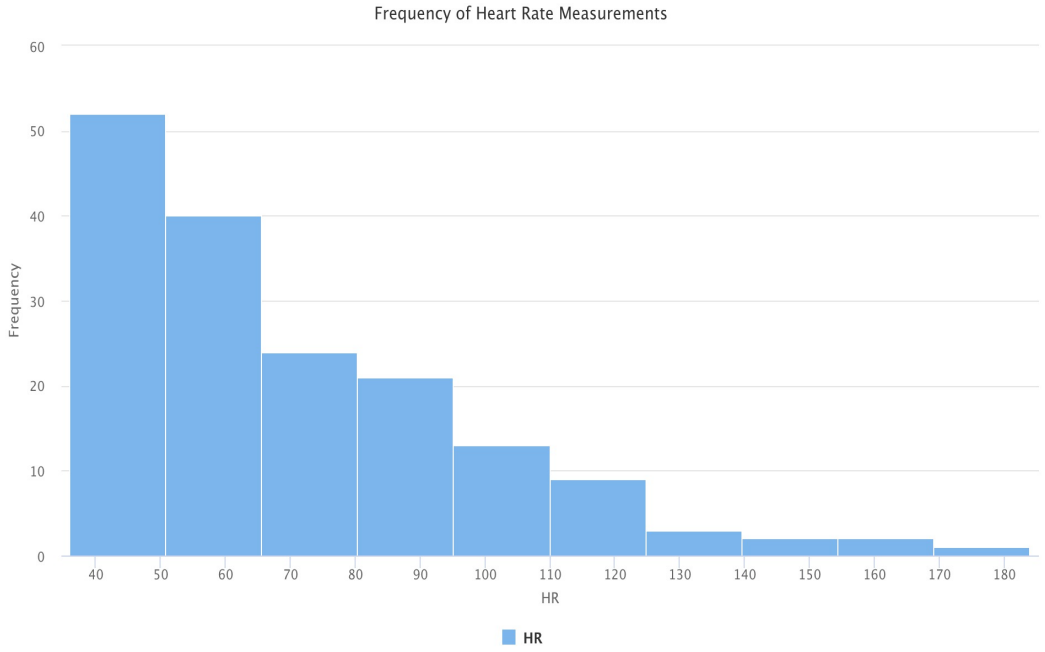


Figure 12 – Frequency of Heart Rate Measurements

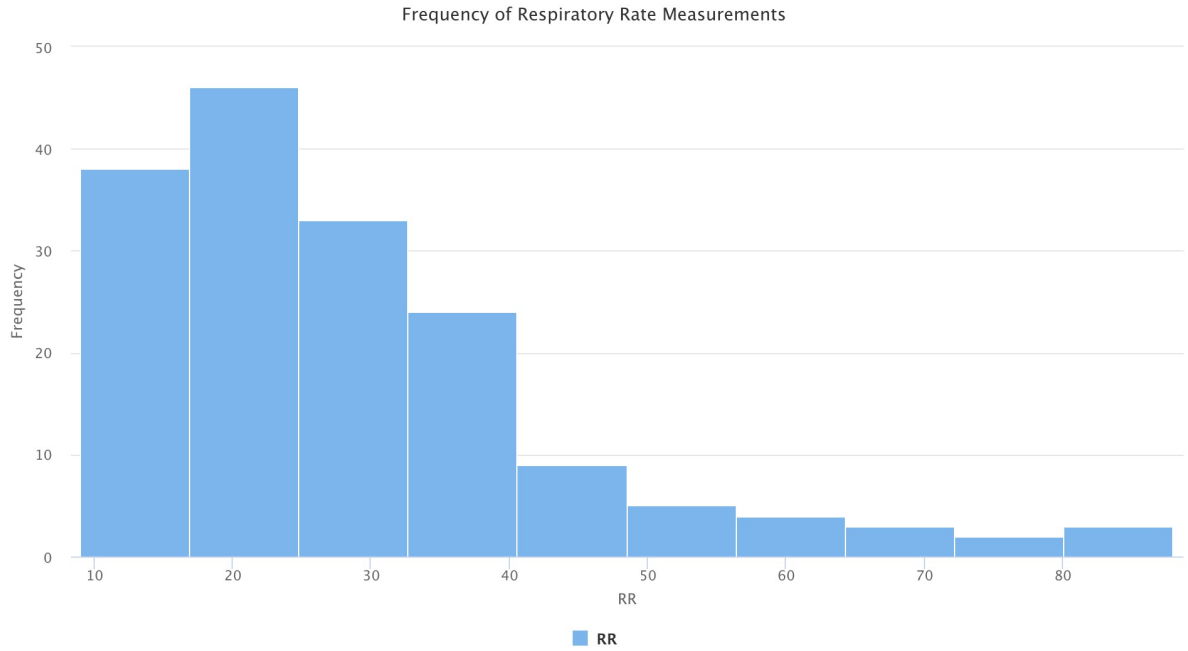


Figure 13 – Frequency of Respiratory Rate Measurements

Appendix E

Figure 14 - Process from further Exploratory Data Analysis in RapidMiner

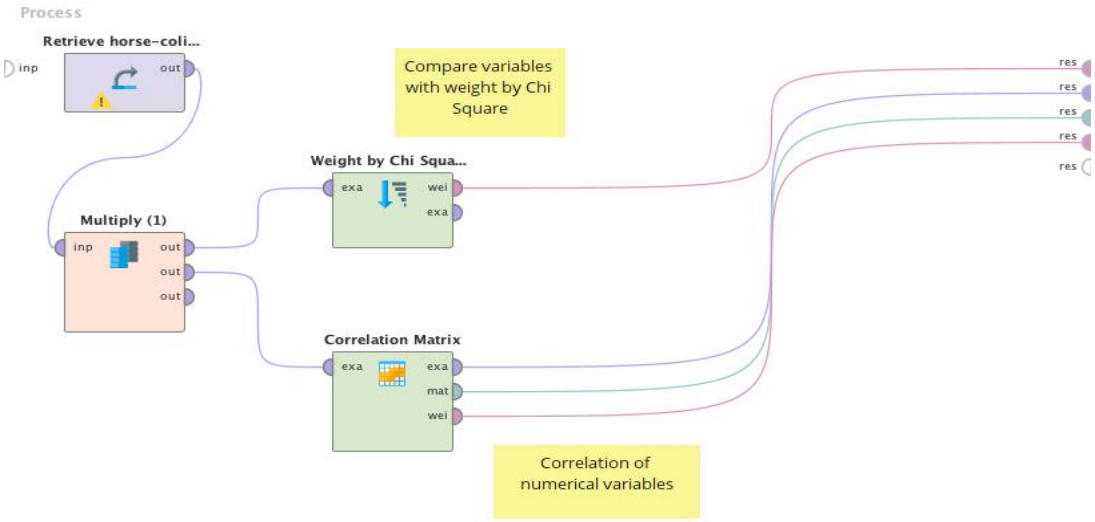


Figure 15 – Correlation Matrix Numerical Variables

Attribu...	TP _gdL	PCV	Temp	HR	RR
TP _gdL	1	0.379	0.200	0.033	-0.070
PCV	0.379	1	0.111	0.348	0.119
Temp	0.200	0.111	1	0.184	0.244
HR	0.033	0.348	0.184	1	0.517
RR	-0.070	0.119	0.244	0.517	1

Figure 16 – Variable Weights by Chi Square Statistic

attribute	wei... ↓
Pain	44.790
Abdomen	36.266
Abdo_Dist	34.243
HR	29.440
Centesis	19.868
Peristalsis	16.149
Temp_Ext	13.295
RR	12.922
PP	12.544
PCV	11.747
Centesi...	11.194
NGR_pH	10.537
Temp	9.890
Rectal_Ex	9.395
MM	8.354
NGR	5.350
CRT	4.851
TP _gdL	4.195
NGT	3.080

Appendix F

Figure 17 - Processes for W-Apriori Algorithm and Decision Tree Classification in RapidMiner

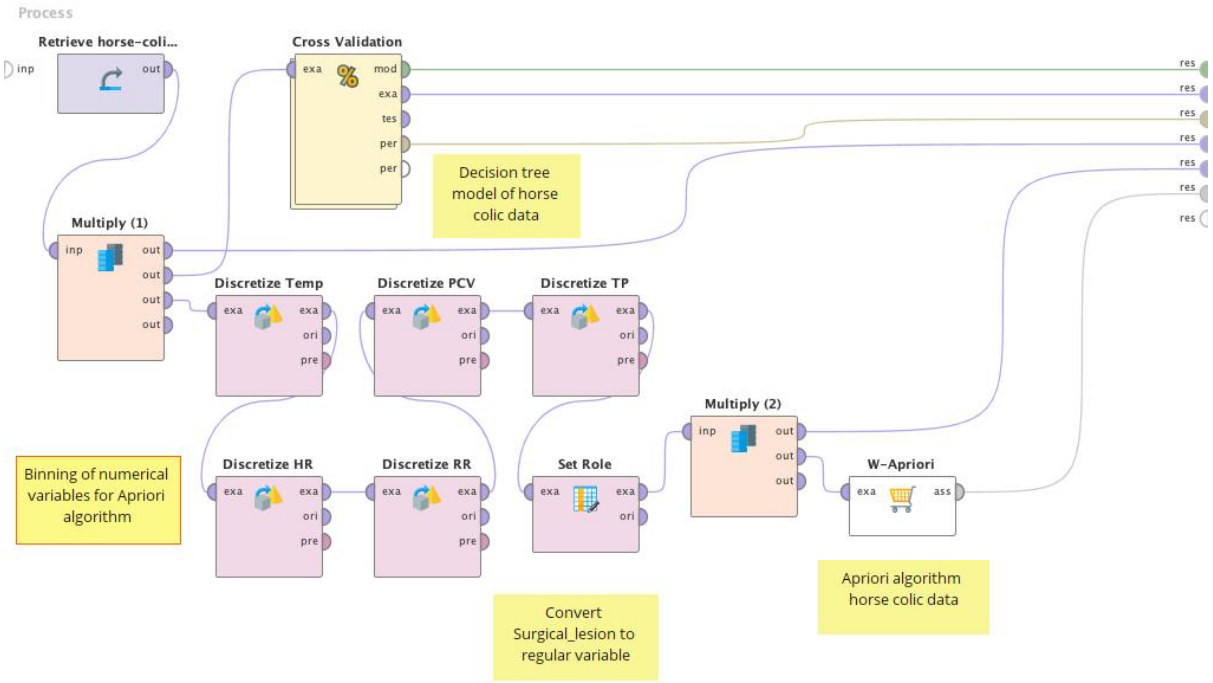
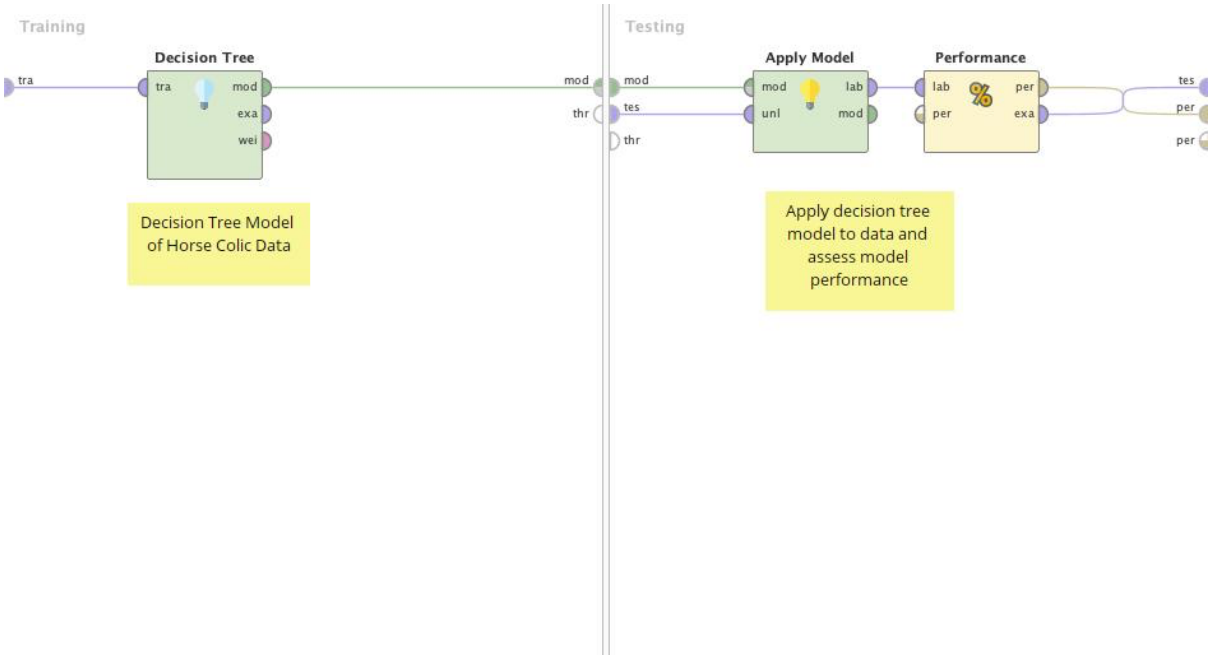


Figure 18 – Internal Process for the Cross Validation Operator and Decision Tree Classification in RapidMiner



Appendix G

Figure 19 - Decision Tree Classifier Horse Colic Data

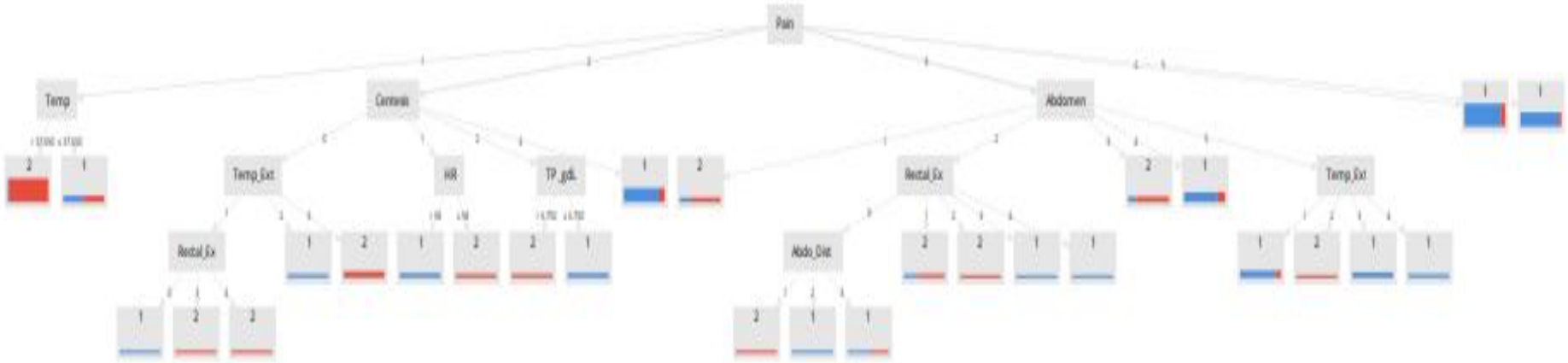


Figure 20 – Decision Tree Description Horse Colic Data

Tree

```
Pain = 1
|   Temp > 37.650: 2 {1=1, 2=25}
|   Temp ≤ 37.650: 1 {1=3, 2=3}
Pain = 2
|   Centesis = 0
|   |   Temp_Ext = 1
|   |   |   Rectal_Ex = 0: 1 {1=2, 2=0}
|   |   |   Rectal_Ex = 3: 2 {1=0, 2=2}
|   |   |   Rectal_Ex = 4: 2 {1=0, 2=2}
|   |   Temp_Ext = 2: 1 {1=3, 2=0}
|   |   Temp_Ext = 3: 2 {1=0, 2=6}
|   Centesis = 1
|   |   HR > 58: 1 {1=4, 2=0}
|   |   HR ≤ 58: 2 {1=0, 2=3}
|   Centesis = 2
|   |   TP _gdL > 6.750: 2 {1=0, 2=3}
|   |   TP _gdL ≤ 6.750: 1 {1=4, 2=0}
|   Centesis = 3: 1 {1=13, 2=2}
Pain = 3
|   Abdomen = 1: 2 {1=1, 2=3}
|   Abdomen = 2
|   |   Rectal_Ex = 0
|   |   |   Abdo_Dist = 1: 2 {1=0, 2=2}
|   |   |   Abdo_Dist = 2: 1 {1=2, 2=0}
|   |   |   Abdo_Dist = 3: 1 {1=1, 2=1}
|   |   Rectal_Ex = 1: 2 {1=1, 2=2}
|   |   Rectal_Ex = 2: 2 {1=0, 2=2}
|   |   Rectal_Ex = 3: 1 {1=2, 2=0}
|   |   Rectal_Ex = 4: 1 {1=2, 2=0}
|   Abdomen = 3: 2 {1=1, 2=4}
|   Abdomen = 4: 1 {1=8, 2=2}
|   Abdomen = 5
|   |   Temp_Ext = 1: 1 {1=6, 2=1}
|   |   Temp_Ext = 2: 2 {1=0, 2=2}
|   |   Temp_Ext = 3: 1 {1=5, 2=0}
|   |   Temp_Ext = 4: 1 {1=3, 2=0}
Pain = 4: 1 {1=22, 2=3}
Pain = 5: 1 {1=14, 2=1}
```