Live or Die: Predicting Outcome When Horses Colic

SI 618: Final Project

Valyn P. Dall (vdall)

MOTIVATION

The purpose of this project is to explore data about colic in horses provided by the Ontario Veterinary College in 1989. The goal of this exploration is to discover if mortality in horses afflicted with severe colic can be predicted using a set of data analysis skills implemented by the author.

A diagnosis of colic refers to gastrointestinal pain in the horse, and is considered more of a symptom and not so much a diagnosis. There are numerous causes of colic, and although most cases can be resolved on their own or at home under the care of the horse's primary veterinarian, more severe cases may require surgical intervention. It is assumed that because the horses represented in the dataset are under the care of the Ontario Veterinary College, these cases of colic are severe and could not be resolved at home. By investigating this dataset, the author hopes to gain insight into if the vitals taken during hospital admission provide any predictive value on whether or not the horse will survive its colic episode. This information could be useful for equine veterinarians and owners who are considering surgery over euthanasia, and could prompt further investigation over a wider set of datasets.

The following four questions are addressed in this project:

1. Surgery is a risky proposition and done as a last resort. Of the horses that underwent surgery, how many recovered? Also, is there a relationship between the horse's age and likelihood it will recover from surgery?

2. Veterinarians record numerous vitals when a horse presents with colic. Can certain vital measurements (e.g. temperature, respiratory rate, pulse, capillary refill time) predict if a horse will recover from colic?

3. Some vitals, like pain; temperature of extremities; and characteristics of the abdomen, are subjective. Do these ratings have any correlation to if a horse will require surgery?

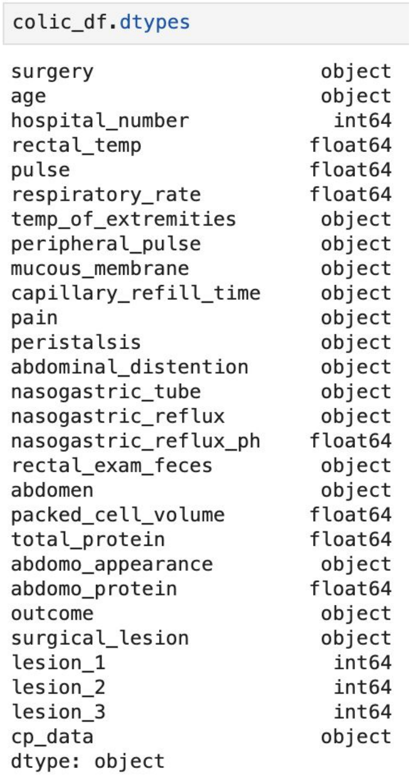
4. Using the numerical values provided throughout the dataset, can the implementation of a random forest accurately predict if a horse survives colic?

DATA SOURCE

Dataset url: https://www.kaggle.com/uciml/horse-colic

UCI Machine Learning released the dataset on 06-06-2017. Although the dataset is relatively new, it appears the data is from 1989. The creators are Mary McLeish & Matt Cecile from the University of Guelph Department of Computer Science. The Ontario Veterinary College is part of the University of Guelph, and it is assumed the data was collected from there since no other specific veterinary hospitals are mentioned.

The dataset contains 299 records with 28 columns and is 59KB. The data file contains a .csv file representing the raw data and a .txt file that serves to describe the definitions for the variable names. The .txt file was digested prior to working on the .csv file, and it was determined cp_data (presence of pathology data) will not be useful since pathology data is not included in this dataset. It is anticipated the remaining values will be important for the analysis because they serve to measure the state of the horse when it was presented to the hospital team. However, some variables, like surgical_lesion and lesion_1; lesion_2; lesion_3, could only be indicated if the horse was either operated on or underwent necropsy. These variables could be considered retroactively, but will not provide any valuable information to the veterinarian or the owner prior to surgery. Therefore, they were excluded from any predictive analysis.



```
colic_df.dtypes
```

```
surgery                   object
age                       object
hospital_number            int64
rectal_temp              float64
pulse                    float64
respiratory_rate         float64
temp_of_extremities       object
peripheral_pulse          object
mucous_membrane           object
capillary_refill_time     object
pain                      object
peristalsis               object
abdominal_distention      object
nasogastric_tube          object
nasogastric_reflux        object
nasogastric_reflux_ph    float64
rectal_exam_feces         object
abdomen                   object
packed_cell_volume       float64
total_protein            float64
abdomo_appearance         object
abdomo_protein           float64
outcome                   object
surgical_lesion           object
lesion_1                   int64
lesion_2                   int64
lesion_3                   int64
cp_data                   object
dtype: object
```

Figure 1: There are 28 variables that are either objects, floats, or integers.

METHODS

Question 1: Of the horses that underwent surgery, how many recovered? Also, is there a relationship between the horse's age and likelihood it will recover from surgery?

I created a dataframe that only included fields relevant to the question, named q1_df. After looking at sample(), I looked at value_counts() for each outcome, age, and surgery to get an idea of the numbers involved. I ran a groupby() keeping “surgery,” “outcome,” and “age” while assigning a new field “count.” I used this dataframe to create a set of visualizations which compared mortality based on raw data. I was also interested in how the data could be investigated by percentage, so I created another dataframe, surgery_stat, that assigned a percentage to each row. I merged this with the surgery dataframe in order to answer the question and createl visualizations based on the percentages.

Due to the decision to parse the dataset using only fields relevant to the question, handling of the data was pretty straightforward. There were no missing fields or noisy data for q1_df.

I used pandas and seaborn. I started with a full dataset; selected the appropriate fields; performed sample(); performed value_counts() on relevant fields; performed groupby() on relevant fields while adding count and resetting the index; created 2 visualizations using seaborn (both bar charts showing outcome(x-axis) and count(y-axis), and the second set was based on age category). I then created surgery_stat dataframe to include a breakdown based on percentages and merged that with surgery dataframe using the index. I repeated my visualizations using “percentage” instead of “count.”

The biggest challenge was coming up with the idea to merge on the index. Originally, I was going to come up with abbreviations to represent each row and rename the index, but I think merge was a better call. One thing that I could not resolve was adding a title to my catplots. When set, the title would bleed onto the age titles; I could not find any documentation on how to change the padding to raise the title above the age labels.

Question 2: Can certain vital measurements (e.g. temperature, respiratory rate, pulse, total) predict if a horse will recover from colic?

I created a dataframe, q2_df, utilizing fields from the original dataset relevant to the question and dropping Nan. This got me through the seaborn visualizations, but then I created a new dataframe, new_q2, where I mapped integers over each outcome variable in order to move forward with the statistical analysis. After that, I categorized whether a horse lived or died (died/euthanized) under a new field “isalive” before moving on to the OLS regressions and ANOVA tables.
There are quite a few rows where a numerical value was NaN, especially in diagnostics that might not have occurred until treatment became more aggressive. After playing around with interpolate() and

filling NaN with the mean, I decided to just drop them. This seemed to have the least amount of negative impact on the data as analysis progressed.

I used pandas, seaborn, and statsmodels. I created a dataframe containing the fields relevant to the question, dropped values that were NaN, and looked at a sample of the data as well as the mean of the variables. I used pairplot to compare the numerical variables and set hue to outcome to flag which horses lived, died, or were euthanized. Based on some promising variables from the pairplot, I created kde jointplots to look at the relationships between the chosen variables as well as jointgrids. To prepare for the regression/ANOVA analysis, I mapped each outcome to a variable and then created a new field ("isalive") to state whether a horse was "alive" or "dead" (died/euthanized) and ran the mean on new_q2. I then ran OLS regressions and ANOVA tables on the chosen fields based on the pairplot insights.

My biggest challenge was interpreting some of the statistical reports. I was surprised there seemed to be little predictability among these fields, and caught myself trying to overly apply meaning to the results. I had a difficult time deciding which visualizations to use in addition to the pairplot. I decided on the kde jointpolots because I think they are pretty, and I believe the movement of the lines and saturation of the color imparts a lot of important information. Additionally, I feel the jointgrids we covered in class using pearsonr provide big picture snapshots of the information of interest.

Question 3: Some vitals, like pain; temperature of extremities; and characteristics of the abdomen, are subjective. Do these ratings have any correlation to if a horse will require surgery?

I created dataframe q3_df using fields relevant to the question and ran a sample. I broke this question into four parts, each based on the categorical data fields found in the new dataframe. For each part, I performed a crosstab between the field in question and "surgery." This provided me with a dataframe to use for chi-squared, mosaic plots, and heatmaps.

Challenges were not much of an issue. Fields that contained NaN were not considered for analysis, so the question covered less than the 299 original cases.

I used pandas, seaborn, statsmodels for mosaic, and scipy for chi2_contingency. After making a dataframe using fields relevant to the question, I looked at a data sample(). I created a process where I would perform a crosstab on the field in question and "surgery", and use the resulting dataframe to supply to chi2_contingency. This created chi-squared, p-values, and degrees of freedom for the field in question. The next step of the process was to create a mosaicplot between the target field and "surgery." Duplicated from lecture, lambda was implemented to change the color properties when creating the mosaic. The final process was to create a heatmap from the resulting chi-sq dataframe. This process was repeated four times (Abdominal Distention; Peristalsis; Pain; Extremity Temperature).

For the most part, this question was pretty straightforward once I decided to break it down into four sections. Originally, I tried to complete this question without having the structure and process in place, and I ended up wasting time due to my confusion.

As I was writing this report, I performed additional analysis on “abdominal_distention” and “peristalsis.” The code is located near the top of the notebook under Header and Scrap. The additional work was done as scrap_df and scrap2_df and they were created to provide the numbers I referenced for horses that did not undergo surgery.

Question 4: Using the numerical values provided throughout the dataset, can the implementation of a random forest accurately predict if a horse survives colic?

I created a dataframe, q4_df, using numerical fields relevant to the question and then ran interpolate(). Interpolate was not my first choice, but dropna would not work once the dataframe made it to the RandomForestClassifier. Even after interpolate, there was a null value, so I filled it with 0. Next, I set x to locate not “outcome” and y was set to “outcome.” I then ran the process for the random forest.

I think because my first row contained a NaN value, my original approach for dropping NaN did not work. I decided to interpolate instead, as that seemed to have more integrity than replacing it with mean() when I played around with the data. After interpolate(), I still had to fillna() with 0 for the “abdomo_protein” field.

This required pandas, matplotlib, and a lot of sklearn imports. After creating q4_df, and addressing the NaN values, I set x not equal to the column “outcome” and y equal to the column “outcome” to prepare for training. I then ran train_test_split on x and y and set test_size=0.3 and random_state=0. The lens of the outputs matched appropriately. Next was the RandomForestClassifier to create the random forest model. This model was set to fit my training fields. Then I created predicted labels based on my x_test in order to gauge the test’s accuracy. In order to see best_params, best_estimator, and best_score, I ran the model through GridSearchCV. I also viewed the dataframe that was the result of GridSearchCV. After seeing the accuracy, I wanted to see if I could improve it by changing the folds and hyperparameters. The last piece was to look at feature_importances and create a visualization showing which values impact “outcome” the most.

My biggest challenge was trying to figure out why I could not dropna() from the q4 dataset. Changing to interpolate did not seem to impact the data that much, but I cannot speak to its impact on the accuracy of training.

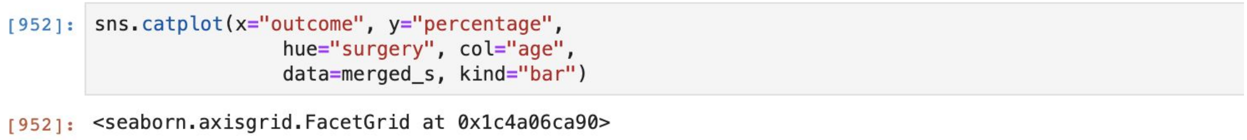
ANALYSIS & RESULTS

Question 1: Of the horses that underwent surgery, how many recovered? Also, is there a relationship between the horse's age and likelihood it will recover from surgery?

Of the 299 cases in the dataset, 180 underwent surgery and 119 did not. Of the 180 cases that underwent surgery, 26 adults and 1 foal were euthanized; 48 adults and 10 foals died; and 88 adults and 7 foals recovered (lived).

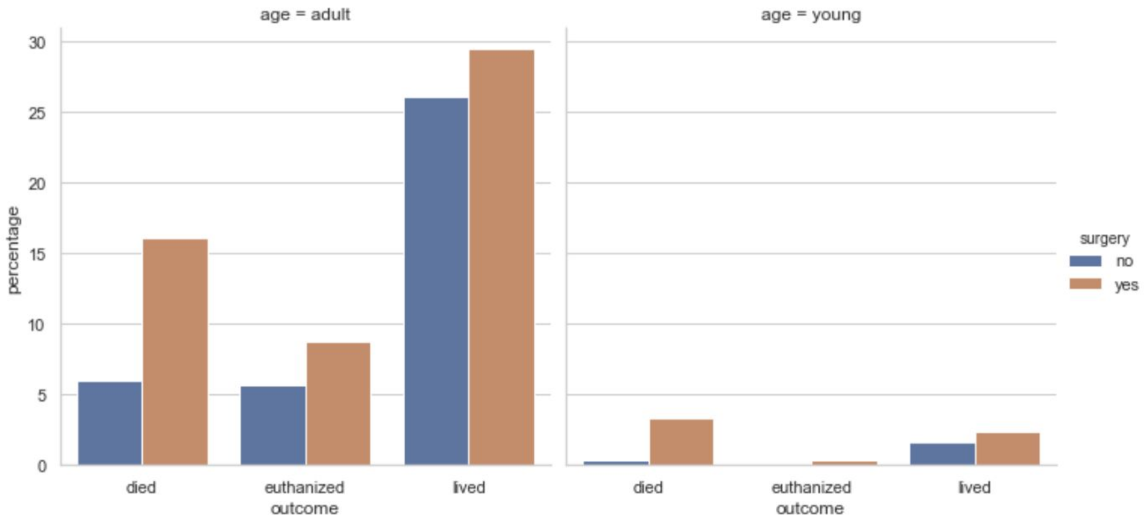
Overall, 29.43% of horses presented to the clinic were adults, underwent surgery and lived. Foals that underwent surgery and lived made up 2.34% of the casework.

One constraint of this dataset is that it is small. Another is that age is defined as being either less than 6 months old (young) or more than 6 months old (adult). This leaves a lot of room for misinterpretation since horses can easily live into their 30s and many would consider horses at least less than a year in age "young." For the purpose of this question, there were very few foal cases to include in the investigation. From looking at the visualization, it appears that foals that underwent surgery and died is the largest group within the age category. The largest group within the adult age range is horses that underwent surgery and lived.

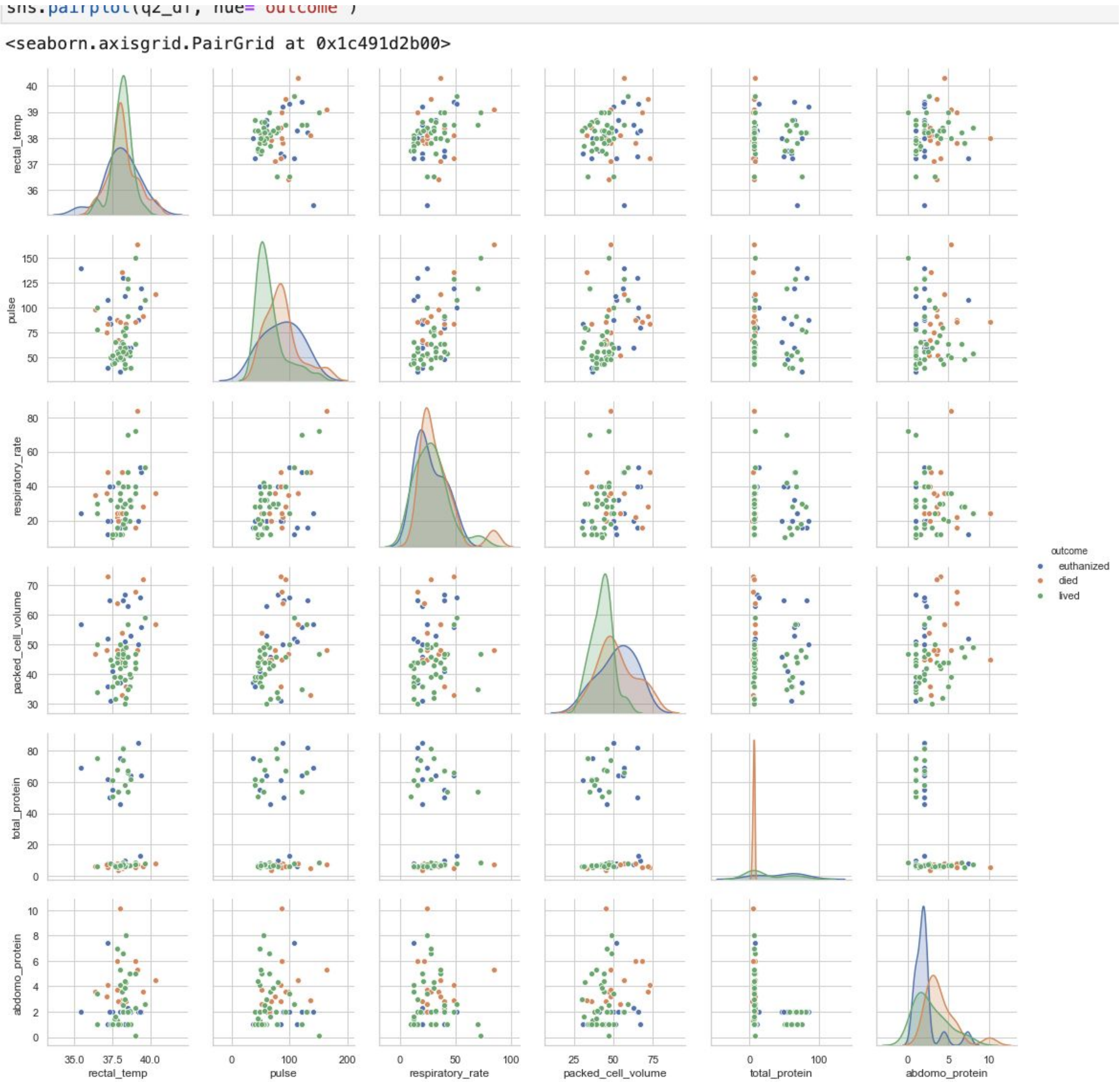


```
[952]: sns.catplot(x="outcome", y="percentage",
                       hue="surgery", col="age",
                       data=merged_s, kind="bar")

[952]: <seaborn.axisgrid.FacetGrid at 0x1c4a06ca90>
```



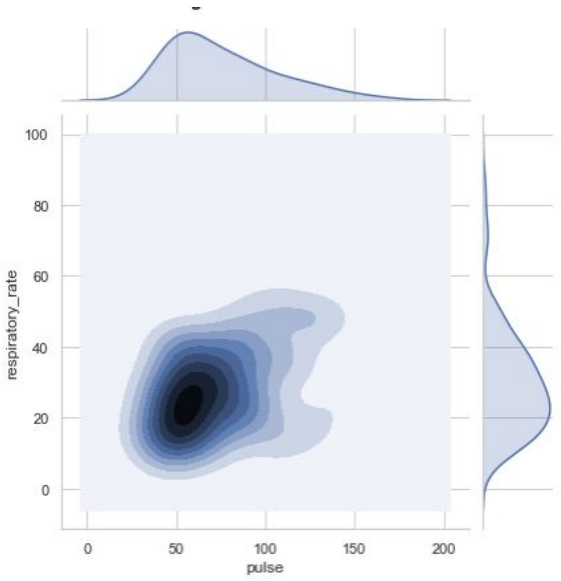
Question 2: Can certain vital measurements (e.g. temperature, respiratory rate, pulse, capillary refill time) predict if a horse will recover from colic?

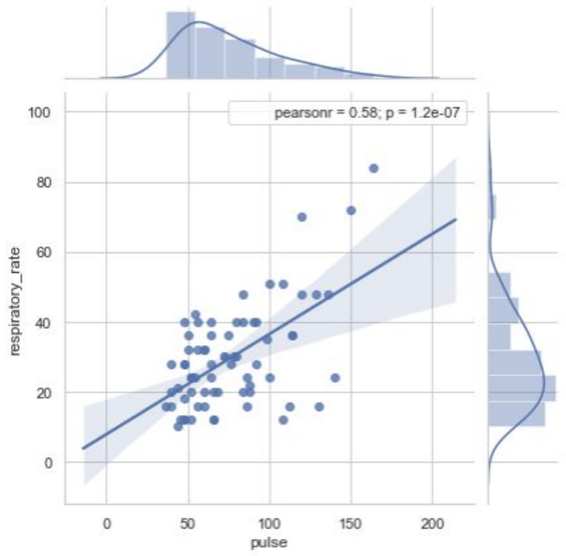


```
sns.pairplot(q2_df, hue= 'outcome')
```

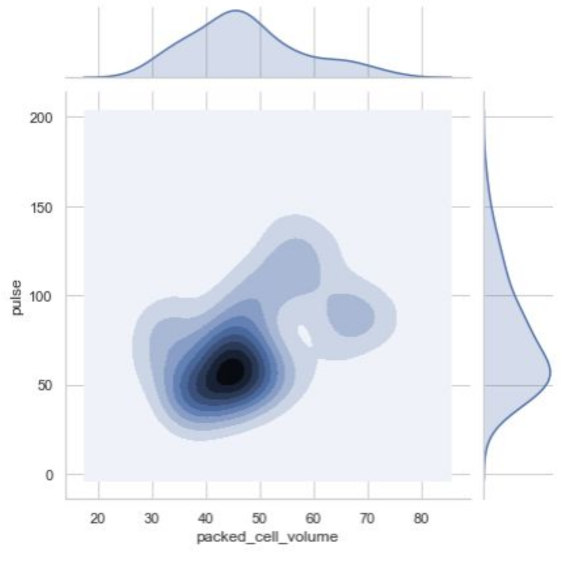
```
<seaborn.axisgrid.PairGrid at 0x1c491d2b00>
```

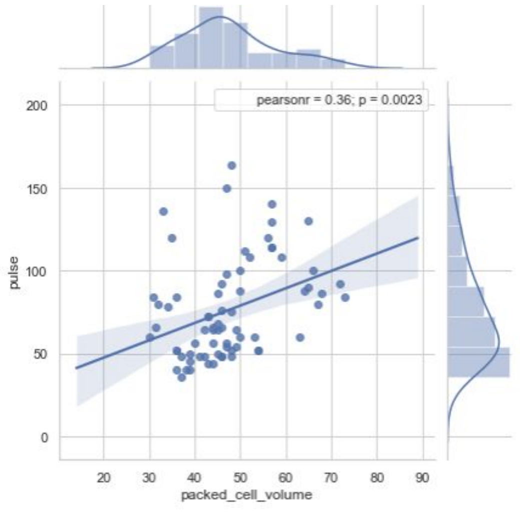
From looking at the pairplot, I was able to see some interesting trends between the different variables and outcomes. I selected “respiratory_rate”/”pulse” and “pulse”/”packed_cell_volume” based on the appearance of a relationship. I created a kde jointplot and jointgrid for each pairing.





A normal pulse is 28-44 beats per minute; a normal respiratory rate is 8-10 per minute. It appears most of the horses upon presentation have a slightly elevated pulse (roughly around 55), possibly due to the stress of not feeling well as well as travelling to the clinic. The respiration rates also seem elevated, most likely for the same reason as the pulse. The p-value between these relationships is less than 0.05, so we can reject the null hypothesis and suggest there is a relationship between these two values. From looking at the jointgrid, the two variables appear to be positively correlated. Referring back to the pairplot, it appears horses with higher respiratory rates and higher pulses had more occasions of death/euthanasia.





A normal packed cell volume is considered to be between 30-50. Most horses seem to be right around 45 with a pulse of 75(elevated). The p-value between these relationships is less than 0.05, so we can reject the null hypothesis and suggest there is a relationship of significance between these two variables. From looking at the jointgrid, the two variables appear to be positively correlated. Referring back to the pairplot, horses that have an elevated packed cell volume and elevated pulse have more occasions of death/euthanasia.

Four OLS regressions and ANOVAs were conducted once the dataframe was properly mapped and assigned values for dead or alive under the field “isalive.”

Pulse

Both indicate an F-statistic of 9.399 and a p-value based on that F-statistic of 0.003. This is less than 0.05, so the null hypothesis is rejected and a relationship of significance can be assumed. Additionally, the p-value for the t-statistic for C(isalive)[T.dead] is also less than 0.05 (0.003) so this also supports rejecting the null hypothesis.

Respiratory Rate

Both indicate an F-statistic of 0.1947 and a p-value based on that F-statistic of 0.660. This is more than 0.05, so the null hypothesis fails to be rejected and a relationship of significance cannot be assumed. Additionally, the p-value for the t-statistic for C(isalive)[T.dead] is also more than 0.05 (0.660) so this also supports failing to reject the null hypothesis.

Packed Cell Volume

Both indicate an F-statistic of 18.9 and a p-value based on that F-statistic of 4.55e-05. This is less than 0.05, so the null hypothesis is rejected and a relationship of significance can be assumed. Additionally, the p-value for the t-statistic for C(isalive)[T.dead] is also less than 0.05 (0.000) so this also supports rejecting the null hypothesis.

Total Protein

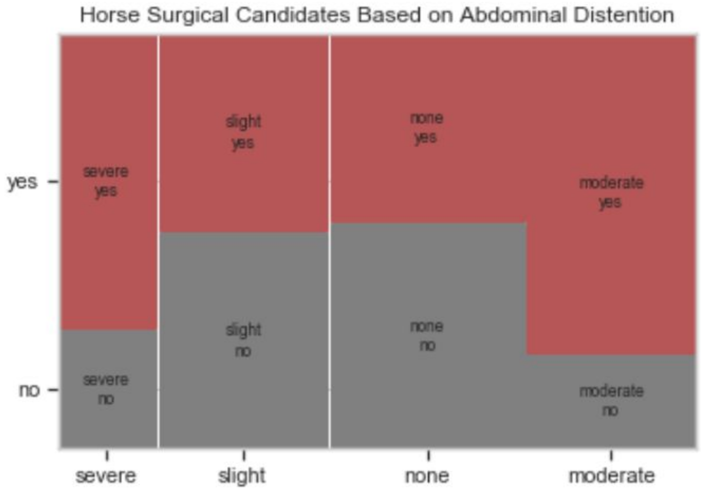
Both indicate an F-statistic of 0.019 and a p-value based on that F-statistic of 0.890. This is more than 0.05, so the null hypothesis fails to be rejected and a relationship of significance cannot be assumed. Additionally, the p-value for the t-statistic for C(isalive)[T.dead] is also more than 0.05 (0.890) so this also supports failing to reject the null hypothesis.

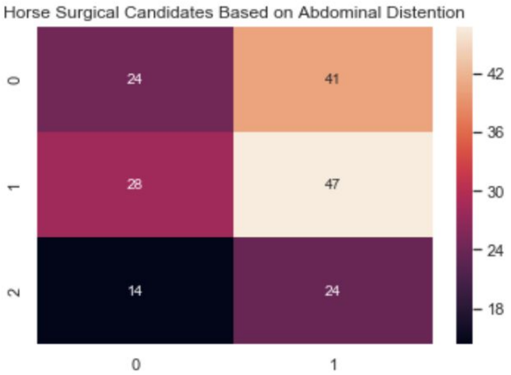
It does appear that certain individual values (like packed cell volume and pulse) can predict if a horse will recover from colic or not. However, as demonstrated by the pairplot and additional visualizations, how certain values interact with others provide a clearer picture as to whether a horse has a better chance of recovery or not.

Question 3: Some vitals, like pain; temperature of extremities; and characteristics of the abdomen, are subjective. Do these ratings have any correlation to if a horse will require surgery?

Abdominal Distention

The chi2 p-value for abdominal distention is less than 0.05 (0.0002) so the null hypothesis is rejected, suggesting the relationship between abdominal distension and surgery is significant.

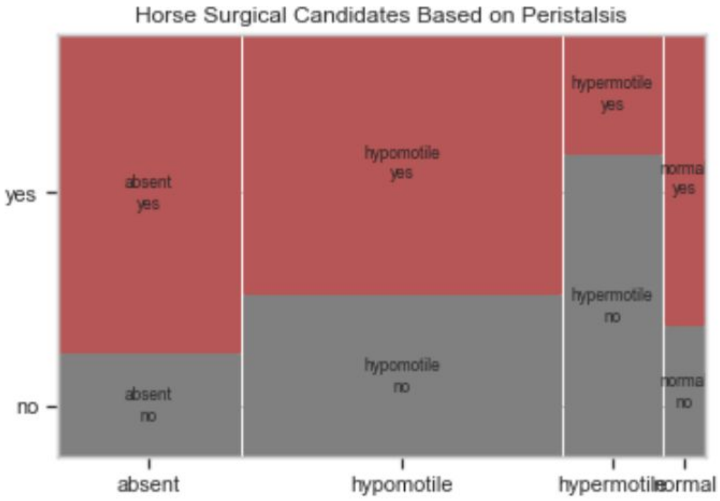


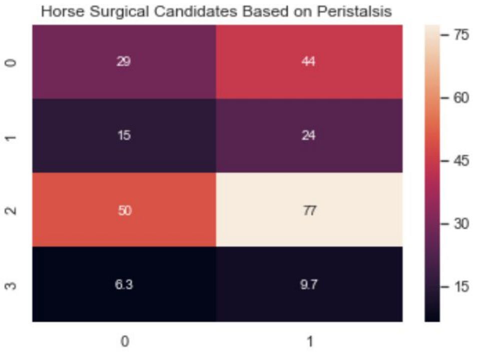


Horses classified as having moderate to severe abdominal distention had more incidents of undergoing surgery. This is confirmed by the projections displayed from the heatmap using the chi2 information where 50 surgery cases had moderate abdominal distention (41 chi2) and 24 were severe (27 chi2). For comparison, there were 26 cases of horses being classified as having moderate to severe cases of abdominal distention that did not undergo surgery. Of those, only 5 lived and 21 died or were euthanized.

Peristalsis

The chi2 p-value for peristalsis is less than 0.05 (7.006045856226024e-06) so the null hypothesis is rejected, suggesting the relationship between peristalsis and surgery is significant.

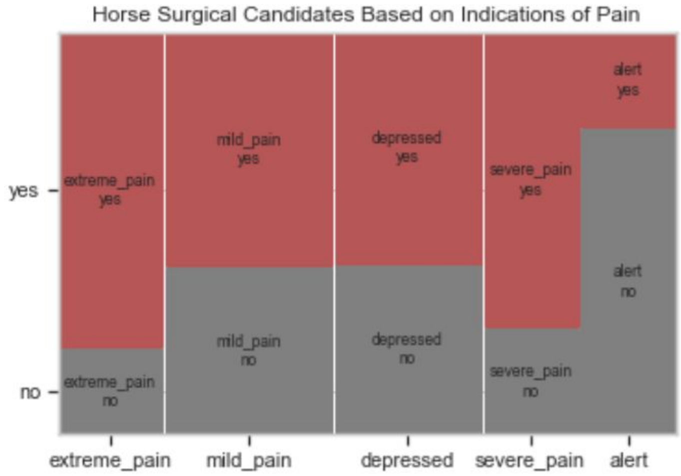


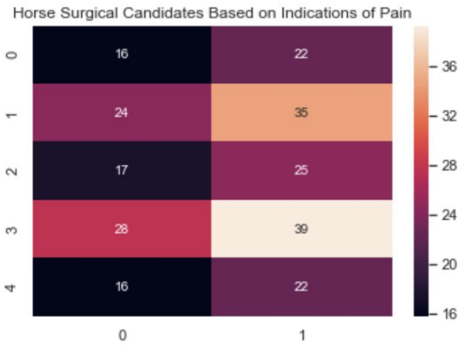


Horses classified as having hypomotile to absent peristalsis (gut movement) had more incidents of undergoing surgery. This is confirmed by the projections displayed from the heatmap using the chi2 information where 55 surgery cases had absent peristalsis (44 chi2) and 78 were hypomotile (77 chi2). For comparison, there were 67 cases of horses being classified as having absent to hypomotile peristalsis that did not undergo surgery. Of those, 41 lived and 26 died or were euthanized.

Pain

The chi2 p-value for pain is less than 0.05 (1.5724375484624656e-05) so the null hypothesis is rejected, suggesting the relationship between pain and surgery is significant.

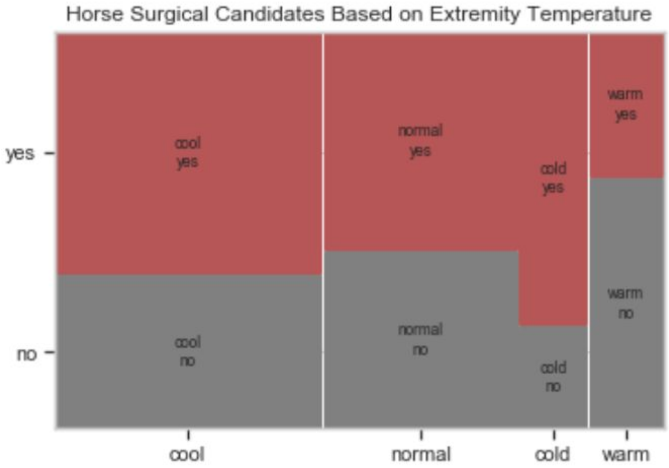


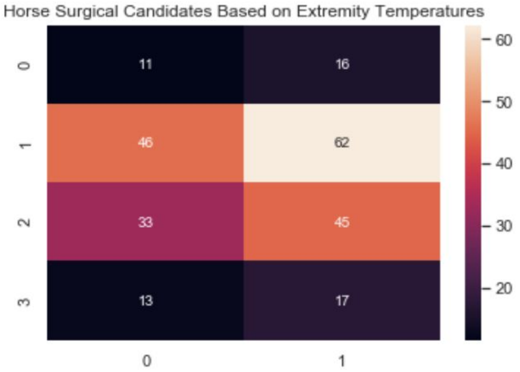


Horses classified as having extreme to severe pain were more likely to undergo surgery, even though they are outnumbered as surgical candidates in comparison to horses who are depressed or even in mild pain. The discrepancy is supported by the projections of the heatmap using chi2. 33 cases of extreme pain underwent surgery (24 chi2) and 28 cases of severe pain underwent surgery (22 chi2) as did 39 cases of mild pain (39 chi2) and 34 cases of depressed (34). It seems although pain can indicate if a horse needs to undergo surgery, the designation might be the most up for interpretation by the assigning clinician.

Extremity Temperature

The chi2 p-value for extremity temperature is less than 0.05 (0.027) so the null hypothesis is rejected, suggesting the relationship between extremity temperature and surgery is significant.





Horses classified as having cold extremities were more likely to undergo surgery, and horses with warm extremities were less likely to be surgical candidates. These findings are supported by the projections of the heatmap using chi2. 20 cases of cold extremities underwent surgery (15 chi2).

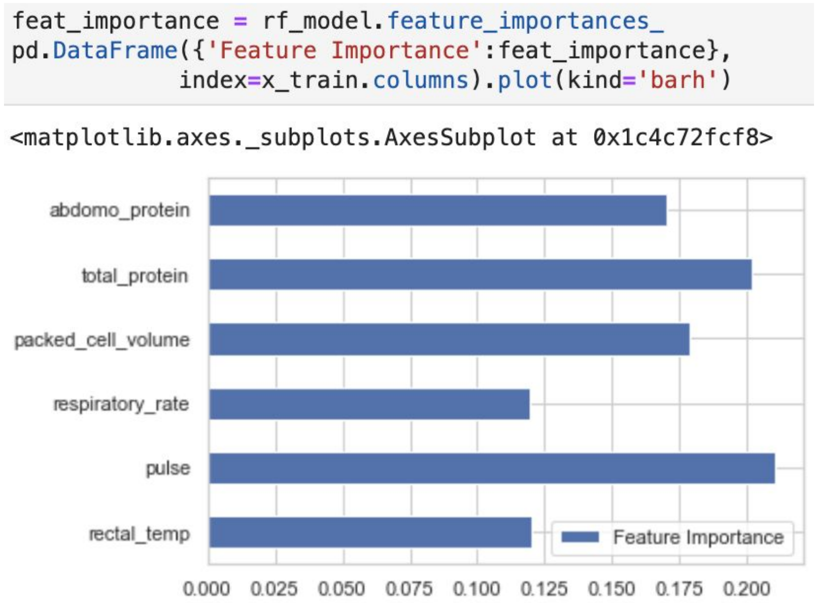
It seems that even though judgement can be subjective for many of these fields, negative diagnosis surrounding abdominal distention and peristalsis seem to indicate a higher frequency of surgeries performed. Although comparisons were drawn to mortality of horses that did not undergo surgery for similar diagnosis, those outcomes could have been influenced by the financial burden surgery places on the owner just as much as the physical deterioration of the horse.

Question 4: Using the numerical values provided throughout the dataset, can the implementation of a random forest accurately predict if a horse survives colic?

Given this dataset, the best accuracy I could achieve was around 72%. Although this seems disappointingly low, survival predictions provided by veterinary experts often aren't as high when it comes to colic in horses.

The first accuracy I obtained using the original code provided in class achieved roughly 61% with a best_score of roughly 69% after running GridSearchCV. Best_params were a max_depth: 5 and n_estimators: 25. I changed the fold from 10 to 5 and that improved best_score a little (69%). I tried to tune the forest further and changed the hyperparameters. This brought best_score to about 72% and best_params to max_depth:7 and n_estimators:16. I think this is an appropriate target as this is not a very large dataset and having higher params could cause needless splits with max_depth and too many trees with n_estimators.

According to feature_importance, pulse and total_protein were consistently among the top two features for colic and outcomes. This did not surprise me, as similar trends were visible in the pairplot from question 2. Even still, according to the visualization, the measure of importance seems low.



```
feat_importance = rf_model.feature_importances_
pd.DataFrame({'Feature Importance':feat_importance},
            index=x_train.columns).plot(kind='barh')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1c4c72fcf8>
```

.