

**NANYANG
TECHNOLOGICAL
UNIVERSITY**
SINGAPORE

SC4000 Machine Learning

Predict Health Outcomes of Horses

| Name | Matriculation Number | Contributions |
|-----------------------|---------------------------------|---|
| Bong Jia Hui | U2121174D | Data pre-processing, modelling, report and presentation |
| Jamie Cheng Jia Ping | U2121421K | Data pre-processing, modelling, report and presentation |
| Leung Chun Kit Andrew | U2123332H | Data pre-processing, modelling, report and presentation |
| Pachigulla Ramtej | U2120398K | Data pre-processing, modelling, report and presentation |

| | |
|--|----|
| 1. Introduction..... | 3 |
| 1.1 Objective..... | 3 |
| 1.2 Challenges..... | 3 |
| 1.3 Ranking..... | 4 |
| 2. Data Pipeline..... | 5 |
| 2.1 Dataset Description..... | 5 |
| 2.2 Exploratory Data Analysis..... | 5 |
| 2.2.1 Numerical Variables..... | 6 |
| 2.2.2 Categorical Variables..... | 9 |
| 2.3 Data Cleaning..... | 14 |
| 2.3.1 Numerical Variables..... | 14 |
| 2.3.2 Categorical Variables..... | 14 |
| 2.4 Feature Engineering..... | 14 |
| 2.5 Transformation (Encoding and Scaling)..... | 15 |
| 2.5.1 Binary..... | 15 |
| 2.5.2 Categorical..... | 15 |
| 2.5.3 Numerical..... | 15 |
| 3. Modelling..... | 16 |
| 3.1 Choice of Models..... | 16 |
| 3.1.1 Background..... | 17 |
| 3.1.2 Algorithm of LightGBM..... | 17 |
| 3.1.3 Comparison of Decision Tree-Based ensembles..... | 17 |
| 3.2 Hyperparameter Tuning..... | 19 |
| 3.3 Stacking..... | 19 |
| 4. Experimental Study..... | 19 |
| 5. Conclusion..... | 20 |
| 6. References..... | 20 |

1. Introduction

1.1 Objective

This project revolves around the analysis and prediction of survival outcomes in horses using a dataset containing detailed medical information. The dataset includes a series of attributes for each horse, from general characteristics like age to clinical metrics such as temperature and respiratory rate, as well as qualitative observations, including the level of pain the horse experienced. Each entry is also paired with an outcome label indicating whether the horse lived, euthanised or died, providing a foundation for supervised learning in this classification task.

Our team's objective is to leverage this dataset to develop a machine learning model that can accurately predict survival outcomes of horses. To do so, we aim to explore how significant each feature is in predicting a horse's likelihood of recovery and survival. This competition not only allows us to apply and refine advanced data processing and modelling techniques but also presents an opportunity to contribute insights that may aid in veterinary care and decision-making in equine health.

1.2 Challenges

The primary challenge with handling this dataset revolved around managing the complexity and quality of the dataset. The dataset consisted of 27 features with a mix of numerical and categorical variables, and many of them contained missing or inconsistent values. Handling these missing values was particularly challenging as improper imputation could introduce bias or distort the balance in the dataset. Some missing values might also not be random due to specific medical conditions, hence requiring thoughtful strategies to tackle this problem.

Another significant challenge was the heterogeneity of the data types as it included ordinal and nominal categorical variables, as well as continuous numerical variables, and each of these required different preprocessing approaches. Encoding categorical variables without losing essential information was critical. For ordinal variables, preserving the inherent order was necessary, while nominal variables required methods like one-hot encoding, which could increase dataset's dimensionality.

Understanding the domain-specific content of the features added another layer of difficulty as many of these variables were medical measurements or observations specific to equine health, such as the colour of "mucous_membrane" or "peristalsis" activity. Since we lacked proper domain expertise, it was challenging to interpret the feature accurately and how they affected the horse survival prediction. This required us to make informed decisions during feature engineering and model selection.

1.3 Ranking


We achieved a public score of **0.85365**, which corresponds to **76th** place in the public leaderboard position. There are a total of **1541** submissions and thus our percentile achieved is **76/1541 \approx 4.93%**

Submission and Description

Private Score ⓘ

Public Score ⓘ

Selected



submission_LGBMClassifier_final.csv

Complete (after deadline) · 19s ago

0.73939

0.85365

☐

Predict Health Outcomes of Horses

Late Submission

Overview

Data

Code

Models

Discussion

Leaderboard


Rules

Team

Submissions

71

Subhajit




0.85975

4

1y

72

Andriy_uru




0.85975

10

1y

73

NIKHIL




0.85975

29

1y

74

Lyudmila Akh




0.85975

5

1y

75

AkshatSG




0.85975

6

1y

76

mist665




0.85365

2

1y

77

gaba42



0.85365

1

1y


We will be inserted here

Our score: 0.85365

1539

▼ 1

Raúl Reaño Araya



0.13484


3

1y

1540

—

Team +1



0.00000


1

1y

1541

—

Subhojyoti Khastagir



0.00000

1

1y

Figure 1: Submission scores and ranking

2. Data Pipeline

2.1 Dataset Description

We utilised two datasets for model training and evaluation. The primary dataset given was generated using a portion of another dataset called the Horse Survival Dataset. Hence to further enhance our analysis, we also incorporated the original Horse Survival Dataset, which was made available through a link provided in the competition's overview. Using both datasets allowed us to explore the differences between them and assess whether integrating the original dataset into our training process would improve the model's performance. More importantly, by concatenating the two datasets together, we can increase the size of our training data which will reduce the difference between training and testing errors and enable the development of a more robust and accurate predictive model for our use case.

The concatenated dataset has a size of 1531 samples, consisting of 28 columns - 27 input features and 1 outcome variable indicating whether the horse has died, lived or was euthanised. The columns can be categorised into categorical variables and numerical variables. Descriptions of these are found in the sections below.

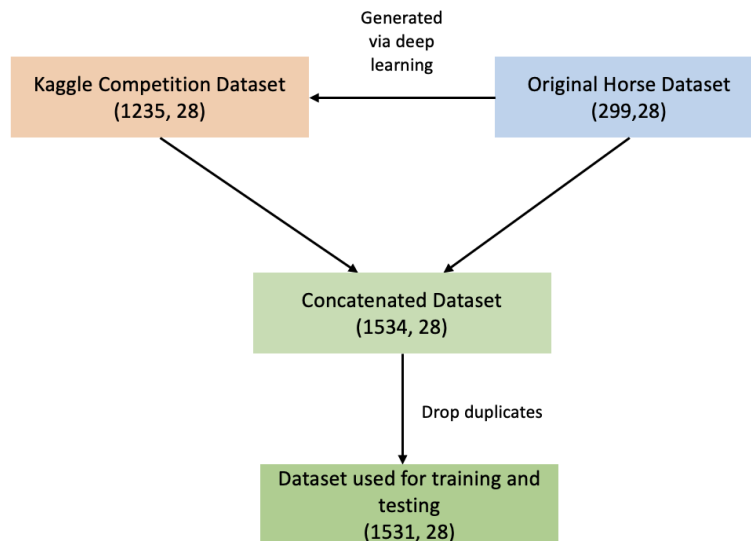


Figure 2: Datasets used for training and testing of model

2.2 Exploratory Data Analysis

In this section, we examine both numerical and categorical variables to gain insights into their distributions and underlying patterns, as well as their relationships with the target variable. This analysis will facilitate us in identifying redundant features, highlight key attributes, and uncover insights useful for feature engineering.

2.2.1 Numerical Variables

Feature Description

Each of the numerical column names, description and relevant information about these columns are listed in the table shown below.

| Column | Description | Additional Information |
|---------------------------------------|---|---|
| hospital_number | Numeric id representing the case number assigned to the treated horse | May not be unique as the same horse being treated multiple times will be assigned the same hospital_number as the first. |
| rectal_temp | The temperature in degrees celsius | The normal temperature is 37.8. An elevated temperature may occur due to infection, while an unusually low temperature indicates that Horse is in late shock. |
| pulse | The heart rate in beats per minute | Horses with painful lesions suffering from circulatory shock may have an elevated heart rate. |
| respiratory_rate | The respiratory rate of horse | Normal respiratory rate for horses is 8 to 10. Respiration rates may increase with medical conditions |
| nasogastric_reflux_ph | The pH of the nasogastric tube that is refluxed | Scale ranges from 0 to 14 with 7 being neutral |
| packed_cell_volume | The number of red cells by volume in the blood | The level rises as the circulation becomes comprised or as the animal becomes dehydrated |
| total_protein | The concentration of protein in the horse | The higher the value the greater the dehydration |
| abdomo_protein | The protein concentration in the abdominal cavity | The higher the level of protein the more likely it is to have a compromised gut |
| Lesions: lesion_1, lesion_2, lesion_3 | Numbers indicating the type of lesions. | The first number is site of lesion, second number is type, third number is subtype and fourth number is specific code |

Statistical Summary

The statistical summary of all the numerical variables (mean, count, quartiles) are summarised in the DataFrame shown below.

| | hospital_number | rectal_temp | pulse | respiratory_rate | nasogastric_reflux_ph | packed_cell_volume | total_protein | abdomo_protein | lesion_1 | lesion_2 | lesion_3 |
|-------|-----------------|-------------|-------------|------------------|-----------------------|--------------------|---------------|----------------|--------------|-------------|-------------|
| count | 1.531000e+03 | 1471.000000 | 1507.000000 | 1473.000000 | 1285.000000 | 1502.000000 | 1498.000000 | 1333.000000 | 1531.000000 | 1531.000000 | 1531.000000 |
| mean | 9.782395e+05 | 38.196941 | 78.145985 | 30.105906 | 4.396109 | 49.017976 | 21.883044 | 3.269017 | 3801.345526 | 29.467015 | 4.328543 |
| std | 1.389757e+06 | 0.779873 | 29.077196 | 16.597691 | 1.940040 | 10.596155 | 26.799768 | 1.617140 | 5434.895232 | 337.075782 | 97.720282 |
| min | 5.184760e+05 | 35.400000 | 30.000000 | 8.000000 | 1.000000 | 23.000000 | 3.300000 | 0.100000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 5.288040e+05 | 37.800000 | 52.000000 | 18.000000 | 2.000000 | 42.000000 | 6.600000 | 2.000000 | 2124.000000 | 0.000000 | 0.000000 |
| 50% | 5.298270e+05 | 38.200000 | 72.000000 | 28.000000 | 4.500000 | 48.000000 | 7.500000 | 3.000000 | 2209.000000 | 0.000000 | 0.000000 |
| 75% | 5.341970e+05 | 38.600000 | 96.000000 | 36.000000 | 6.200000 | 55.000000 | 13.000000 | 4.300000 | 3205.000000 | 0.000000 | 0.000000 |
| max | 5.305629e+06 | 40.800000 | 184.000000 | 96.000000 | 7.500000 | 75.000000 | 89.000000 | 10.100000 | 41110.000000 | 7111.000000 | 2209.000000 |

Figure 3: Statistical Summary of numerical variables

Visualisations and Insights

The density distribution plots between each numerical variables segmented according to their corresponding target variable are shown below.

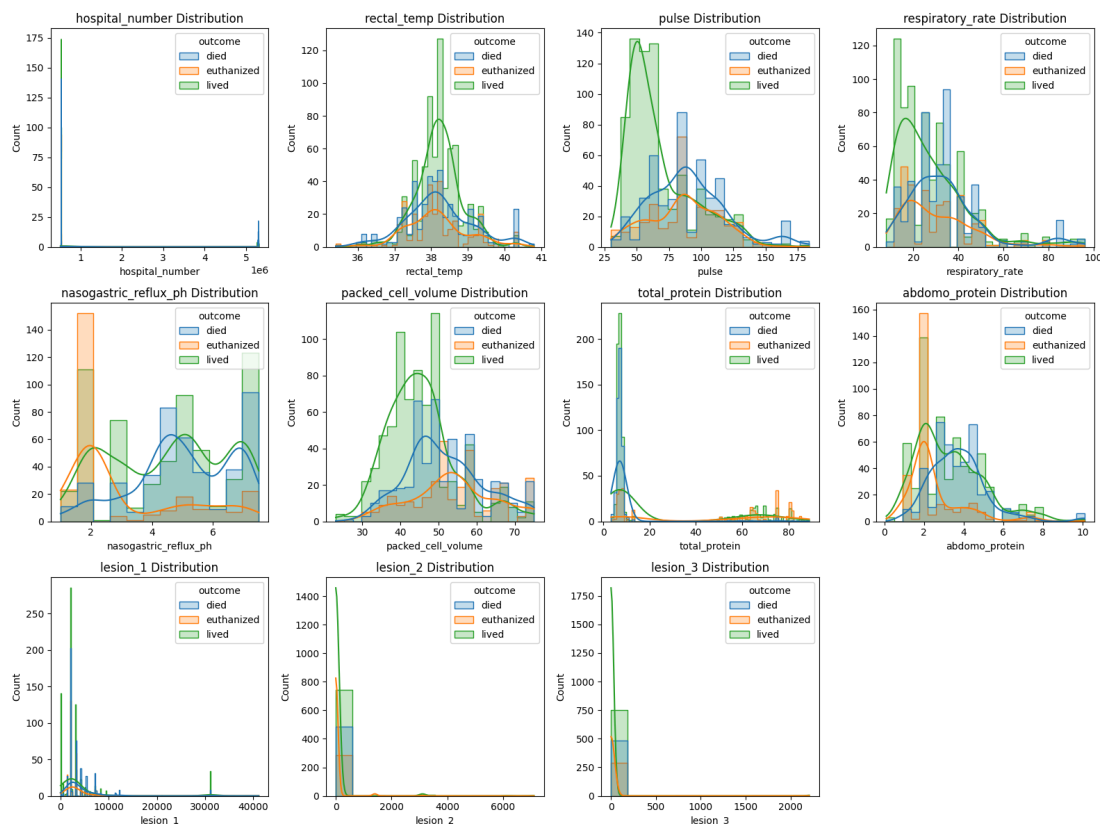


Figure 4: Plots of Numerical Variables

Since *lesion_2* and *lesion_3* show no distinct numerical distribution (as their values are clustered around just a few points), we will include them in our categorical variables' statistical analysis below.

A heatmap is created to examine the correlation across numerical variables, allowing us to effectively clean or process the interdependence of these variables. The removal of highly correlated features can also reduce the dimensionality of our training data.

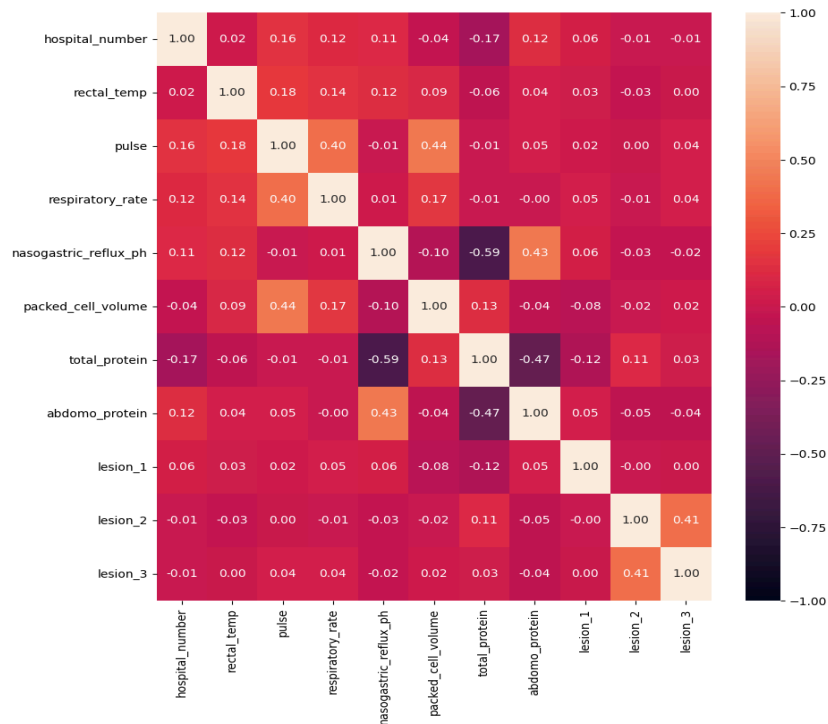


Figure 5: Heatmap of numerical variables interdependence

Notably, there seems to be little to no correlation between the numerical variables. Therefore, no further tests were conducted to explore collinearity or dependence, as it was not deemed essential.

Statistical Test of Significance

By Central Limit Theorem, we approximate the sample means to be normally distributed as the sample size is large. Hence, we conducted parametric tests to determine the significance of the features.

For numerical variables, we performed either One-way ANOVA or Welch's Analysis of Variance (ANOVA) test, depending on whether there is a violation of the assumption that variances across groups are equal.

H0: The means for all outcomes are equal

H1: The mean for at least one of the outcome is different

The results of the test showed that all other features except rectal_temp and lesion_3 have a p-value<0.05. Thus, at a significance level of 0.05, we do not reject the null hypothesis that the mean rectal_temp for all outcomes are equal.


```
Welch's ANOVA
data: rectal_temp and outcome
F Statistic: 0.4110194364389573
P-value: 0.6631328443127569
```

2.2.2 Categorical Variables

Feature Description

Each of the categorical variable column name, description and possible values are listed in the table shown below.

| Column | Description | Type and Possible Values |
|-----------------------|--|---|
| surgery | Variable indicating if horse had surgery | A Binary variable with two values: 'Yes' and 'No' – where 'Yes' signifies that the horse underwent surgery, and 'No' indicates that it did not. |
| age | The age of the horse | A categorical variable with two values: 'Adult' and 'Young' – where 'Adult' refers to a horse of adult age, and 'Young' refers to a horse of a younger age. |
| temp_of_extremities | An indication of peripheral indication | A categorical variable with four values: 'Normal,' 'Warm,' 'Cool,' and 'Cold' – where each value indicates the temperature of the horse's extremities, reflecting peripheral circulation. |
| peripheral_pulse | The measure of blood flow through peripheral arteries | A categorical variable with four values: 'Normal,' 'Increased,' 'Reduced,' and 'Absent' – where each value represents the measure of blood flow through the horse's peripheral arteries. |
| mucous_membrane | Measurement of colour of mucous | A categorical variable with six values: 'Normal pink,' 'Bright pink,' 'Pale pink,' 'Pale cyanotic,' 'Bright red,' and 'Dark cyanotic' – where each value represents the colour of the horse's mucous membranes. |
| capillary_refill_time | The longer the refill, the poorer the circulation of the horse | A categorical variable with three values: 'Less than 3s,' '3s,' and 'More than 3s' – where longer refill times indicate poorer circulation, with '3s' |

| | | |
|----------------------|--|--|
| | | representing an intermediate level of circulation. |
| pain | Horse pain level | A categorical variable with five values: 'Alert,' 'Depressed,' 'Intermittent mild pain,' 'Intermittent severe pain,' and 'Continuous severe pain' – where each value represents the level of pain observed in the horse. |
| peristalsis | Activity of horse's gut | A categorical variable with four values: 'Hypermotile,' 'Normal,' 'Hypomotile,' and 'Absent' – where each value indicates the activity level of the horse's gut. |
| abdominal_distention | Indication of horse's gut motility | A categorical variable with four values: 'None,' 'Slight,' 'Moderate,' and 'Severe' – where each value indicates the level of abdominal distention in relation to the horse's gut motility. |
| nasogastric_tube | Gas coming out of the tube | A categorical variable with three values: 'None,' 'Slight,' and 'Significant' – where each value indicates the amount of gas coming out of the nasogastric tube. |
| nasogastric_reflux | Obstruction of fluid passage | A categorical variable with three values: 'None,' 'More than 1L,' and 'Less than 1L' – where each value indicates the extent of nasogastric reflux, reflecting the obstruction of fluid passage. |
| rectal_exam_feces | Faeces of horse; absent of faeces indicated obstruction | A categorical variable with four values: 'Normal,' 'Increased,' 'Decreased,' and 'Absent' – where each value indicates the condition of the horse's faeces, with 'Absent' suggesting a potential obstruction. |
| abdomen | Areas of obstruction in abdomen | A categorical variable with five values: 'Normal,' 'Other,' 'Firm,' 'Distend_small,' and 'Distend_large' – where each value indicates the presence of areas of obstruction in the horse's abdomen. |
| abdomo_appearance | Description of fluid obtained from horse' abdominal cavity | A categorical variable with three values: 'Clear,' 'Cloudy,' and 'Serosanguinous' – where each value describes the appearance of fluid obtained from the horse's abdominal cavity. |
| surgical_lesion | Binary Variable indicating whether the | A binary variable with two values: 'Yes' and 'No' – where 'Yes' indicates that the problem is caused |

| | | |
|---------|--|--|
| | problem is caused by lesions | by surgical lesions, and 'No' signifies that it is not. |
| cp_data | Binary variable indicating whether pathology data is present for this case | A binary variable with two values: 'Yes' and 'No' – where 'Yes' indicates that pathology data is present for the case, and 'No' signifies that it is absent. |

Statistical Summary

The statistical summary of all the categorical variables are summarised in the DataFrame shown below.

| | count | unique | top | freq |
|-----------------------|-------|--------|----------------|------|
| surgery | 1531 | 2 | yes | 1065 |
| age | 1531 | 2 | adult | 1433 |
| temp_of_extremities | 1436 | 4 | cool | 806 |
| peripheral_pulse | 1402 | 4 | reduced | 826 |
| mucous_membrane | 1463 | 6 | pale_pink | 341 |
| capillary_refill_time | 1493 | 3 | less_3_sec | 1020 |
| pain | 1432 | 6 | depressed | 486 |
| peristalsis | 1467 | 5 | hypomotile | 791 |
| abdominal_distention | 1452 | 4 | moderate | 608 |
| nasogastric_tube | 1348 | 3 | slight | 859 |
| nasogastric_reflux | 1404 | 4 | more_1_liter | 641 |
| rectal_exam_feces | 1239 | 5 | absent | 572 |
| abdomen | 1202 | 5 | distend_small | 525 |
| abdomo_appearance | 1318 | 3 | serosanguinous | 614 |
| surgical_lesion | 1531 | 2 | yes | 1117 |
| cp_data | 1531 | 2 | no | 766 |

Figure 6: Statistical Summary of categorical variables

Visualisations and Insights

Some of the descriptions of the categorical variables in the table indicate potential ordinal relationships as the values follow a linear progression (eg. cold, normal, warm). To investigate these further, pie charts were generated to visualise each categorical feature's relationship with the target variable. One such plot of a feature named capillary_refill_time is shown below.

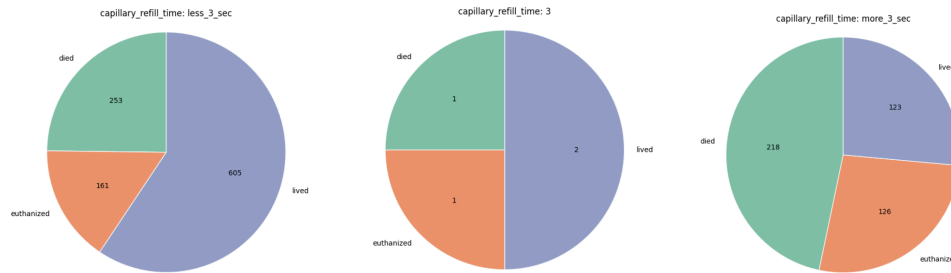


Figure 7: Ordinal variable: *capillary_refill_time* vs target variable

In the case of *capillary_refill_time*, three pie charts were plotted corresponding to the different categories (less than 3s, 3s, and more than 3s) present in *capillary_refill_time*. For the category of 'more than 3s', only 26% of horses lived, while 47% died and 27% were euthanised. In contrast, for the 'less than 3s' category, a significant majority of 60% of horses lived, while 20% died and 20% were euthanised. For the '3s' category, the outcomes were evenly split, with 50% of horses living, 25% dying, and 25% being euthanised. This analysis allows for the ordering of the variable from less to more severe: 'less than 3s', '3s', and 'more than 3s', effectively capturing the positive correlation between the capillary refill time and the mortality of horses.

However, for the variable *pain*, the pie charts do not reveal a linear trend among the outcome distributions, even though the distributions differ widely across the different 'pain' values. This inconsistency suggests that there may be no inherent order and it may be more appropriate to treat it as a nominal variable.

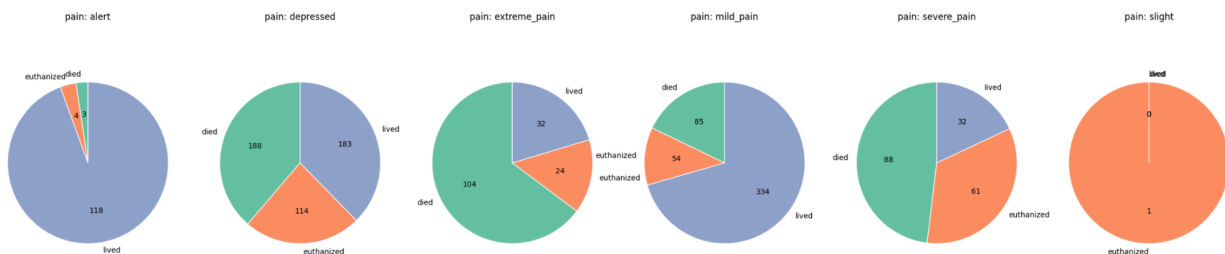


Figure 8: Nominal variable: *pain* vs target variable

Statistical Test of Significance

For categorical variables, we performed Pearson's Chi-Squared test of independence on each categorical feature and the target variable (outcome). As mentioned earlier, *lesion_2* and *lesion_3* will be treated as categorical variables due to their distribution patterns, and therefore, we are including them in the chi-squared tests.

H0: There is no association between the feature and the outcome
H1: There is a significant association between the feature and the outcome

The results of the test showed that except lesion_3, all features have a p-value<0.05. Thus, at a significance level of 0.05, we do not reject the null hypothesis that there is no association between lesion_3 and outcome.

```
Pearson's Chi-squared Test of Independence  
data: lesion_3 and outcome  
Chi2 Statistic: 1.4896471400457827  
P-value: 0.4748180646906728
```

Based on the statistical tests above, the variables *lesion_3* and *rectal_temp* will either be dropped or feature engineered to produce more useful information.

2.3 Data Cleaning

In this section, we discuss how we deal with missing values.

2.3.1 Numerical Variables

The missing values in numerical variables were replaced with K-Nearest Neighbours Imputer which utilises the mean from the top k most similar data points.

2.3.2 Nominal and Ordinal Variables

Initially, we filled the missing values in each categorical variable with the most common value. However, the results were unsatisfactory. Consequently, we explored other imputation methods, such as median-based and class-based imputations, which are more appropriate for our ordinal variables. The class-based imputations yielded the best results, likely because labelling the horses as 'neutral' or 'normal' maintains the characteristics of the respective classes. This approach avoids making assumptions about the horses' health by using modes, which can often represent extreme values (Alam et al., 2023).

2.4 Feature Engineering

Upon inspecting some of the variables within the dataset, we notice some variables which require special treatment and feature engineering.

Lesion

Upon closer inspection and looking at the Data Card for the dataset, we realise that the columns for lesions, lesion_1 is encoded in a 3-6 digit format, where it encapsulates data regarding the lesion. The data includes the site of the lesion, type and subtype of the lesion and the specific condition. Hence, we use a short algorithm to extract the data into 4 separate columns, and treat the data as nominal categorical data during training. Lesion_2 was encoded as a binary variable upon further exploration as the outcomes of records with non-zero values were either "euthanized" and "lived" i.e. none of them "died". Lastly Lesion_3 was dropped due to results of the chi-squared test.

Mucous Membranes

The variable is a subjective measurement of colour of a horse's mucous membranes, with values describing the colour from 1-6, increasing in colour intensity. However, upon further inspection of the data description, the increasing colour intensity does not imply any linear trend in the severity of symptoms; instead, it refers to different categories. Hence, we considered this as a nominal variable and performed one-hot encoding.

Pain

We initially assumed that this categorical variable could be ordinal in nature due to the increasing levels of pain observed from the categories. However, there are limitations in which prior treatments of pain may mask pain level to some extent. Hence, we treat the categorical variable as nominal and create the respective columns for the respective categories.

Rectal Temperature

The variable is a measurement taken, in degrees celsius. The data card mentions that an elevated temperature may occur due to infections or that temperatures may be reduced when the animal is in late shock. It is also mentioned that the normal temperature of a horse is 37.8. Hence, we can take the deviation of the recorded temperature from 37.8 to normalise data for better training.

2.5 Transformation (Encoding and Scaling)

2.5.1 Binary

Firstly, we consider the binary variables. All of the variables have only two possible values and hence we can simply perform Binary Encoding with values 1 or 0 to represent the data.

2.5.2 Nominal and Ordinal

Next, we move on to categorical variables. Most of the variables are ordinal with the exception of *mucous_membrane* and *pain* which is nominal. Hence, the ordinal variables are encoded manually with respect to their ranked order.

The variables *mucous_membrane* and *pain* however are nominal, and hence we performed One Hot Encoding.

2.5.3 Numerical

Lastly, the numerical variables were scaled using Standard scaler.

Summary of Data Preprocessing:

| Transformation | Columns | Number of columns |
|--|---|-------------------|
| Input missing values | All | 27 |
| Dropped | <i>lesion_3, rectal_temp, hospital_number, cp_data</i> | 4 |
| Ordinal Encoding | <i>peristalsis, rectal_exam_feces, nasogastric_reflux, temp_of_extremities, peripheral_pulse, capillary_refill_time, peristalsis, abdominal_distention, nasogastric_tube, nasogastric_reflux, rectal_exam_feces, abdomen, abdomo_appearance</i> | 13 |
| Binary Encoding | <i>age, surgery, surgical_lesion</i> | 3 |
| One-Hot Encoding | <i>mucous_membrane, pain</i> | 2 |
| Special Case 1: Lesion 1 transformation | lesion_1 | 1 |
| Special Case 2: Transformation of nominal to binary variable | lesion_2 | 1 |

3. Modelling

3.1 Choice of Models

Since our dataset is tabular with 28 feature columns and multi-class target variables, we initially employed classic machine learning models for classification, including SVM, logistic regression and decision trees. However, given the high dimensionality of the data, we found that SVM's hyperplanes did not perform well, and logistic regression was not optimal for multi-class classification.

| | Model | Accuracy | Precision | Recall | F1 Score |
|---|---------------------|----------|-----------|----------|----------|
| 0 | SVM | 0.501085 | 0.337398 | 0.353622 | 0.501085 |
| 1 | Logistic Regression | 0.657267 | 0.656060 | 0.599346 | 0.657267 |
| 2 | Decision Tree | 0.665944 | 0.652142 | 0.639943 | 0.665944 |

Figure 9: Model performances across different performance metrics

Decision trees demonstrated the most consistent performance across the different models we tested. To further enhance their robustness, we plan to leverage ensemble learning techniques, a concept emphasised in our lecture content. By applying methods like bagging and boosting, specifically through algorithms like LightGBM, we aim to improve the decision trees' accuracy.

3.1.1 Background

LightGBM, short for Light Gradient-Boosting Machine, is an ensemble learning framework which is based on decision trees. In each boosting iteration, a new tree, which will be trained on the residuals to minimise the loss function, is added to reduce the error of the previous ensemble. The prediction will be based on the weighted sum of all previous trees' predictions.

3.1.2 Algorithm of LightGBM

LightGBM uses histogram-based boosting. Instead of using the original continuous values for features, LightGBM divides the feature range into intervals by mapping these continuous values into a set of discrete values, thereby reducing the complexity of the model. The binned data is then represented as a histogram, where each bin contains the count of the data points which fall within its range. This enables LightGBM to efficiently find the optimal split points for the trees. When building the trees, the gain for each potential split will be calculated, making the computation faster.

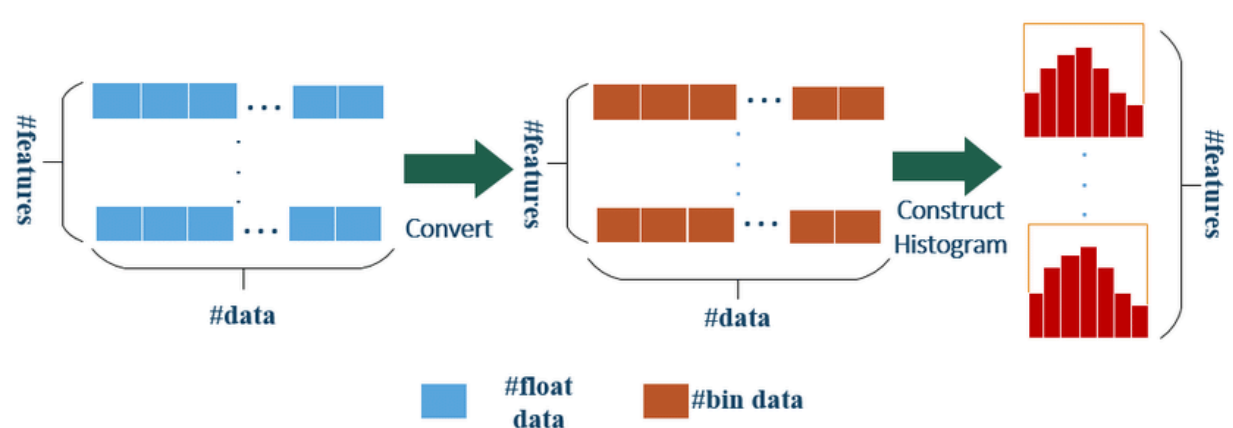


Figure 10: Diagram showing histogram-binning algorithm of LightGBM (Yin et al., 2021).

3.1.3 Comparison of Decision Tree-Based ensembles

Decision-tree based ensemble algorithms include Random Forest, XGBoost and LightGBM. The key differences between them are listed in Figure 10. One of the key differences between all 3 algorithms is the method of tree construction. Unlike other implementations of ensemble, decision-trees-based boosting or bagging algorithms such as XGBoost and Random Forests, LightGBM grows leaf-wise and not level-wise. LightGBM splits the leaf with the highest loss reduction, leading to deeper and unbalanced trees but achieving greater loss reduction and higher accuracy within fewer iterations. With our relatively large dataset (~1531 rows), it is relatively efficient but

could potentially cause overfitting if the dataset is small. This can be reduced using the *max-depth* parameter which specifies where the splitting would occur.

Random Forest and XGBoost on the other hand, grow level-wise. In each iteration, every node at the current level is split before the algorithm proceeds to the next level. All the nodes will be expanded simultaneously until the maximum depth or the minimum number of samples in a node is reached. Trees will be constructed in a more balanced, breadth-first approach but with less depth, resulting in a longer time complexity as compared to LightGBM (Suenaga et al., 2023).

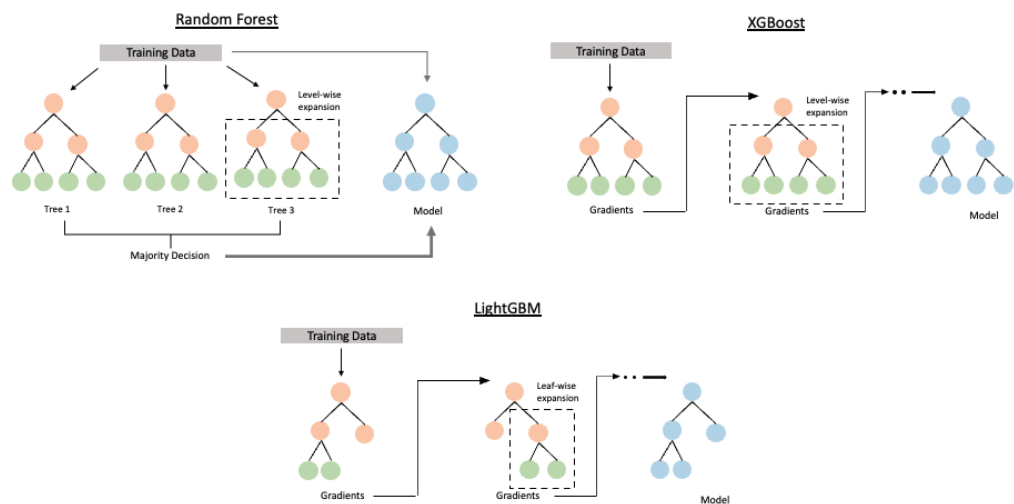


Figure 11: Diagram depicting differences between ensemble decision trees methods

| Feature | Random Forest | XGBoost | LightGBM |
|-----------------------|---|--|--|
| Algorithm Type | Ensemble of decision trees | Gradient boosting framework | Gradient boosting framework |
| Tree Construction | Level-wise; Builds trees in parallel (bagging) | Level-wise; Builds trees sequentially (boosting) | Leaf-wise; Builds trees sequentially (boosting) |
| Speed | Slower with large datasets | Faster than random forest | Significantly faster due to histogram-based algorithm |
| Hyperparameter tuning | Less complex with fewer hyperparameters to tune | Many hyperparameters for tuning | Many hyperparameters, but in general default ones work well. |

| | | | |
|-----------------|---|------------------------------|-------------------------------------|
| Most suited for | Structured data in domains where understanding of how nodes are split is required | Medium sized structured data | Large datasets with high dimensions |
|-----------------|---|------------------------------|-------------------------------------|

3.2 Hyperparameter Tuning

To find the best hyperparameters for the classifiers, GridSearchCV was used. An exhaustive search over a dictionary of parameter values was performed and optimised using a 3-fold cross validation. The tuned model showed poorer performance as evidenced by the lower f1 score in the validation set. This could be due to an overly complex model compared to the small training dataset available which resulted in overfitting.

3.3 Stacking

The 3 base classifiers trained - LGBMClassifier, XGBClassifier and RandomForestClassifier- attained desirable performance in terms of accuracy, precision, recall and f1 score. However, the predictions made by the classifiers on the validation set appear to be highly correlated. We attempted to stack the models using the predicted probability and as expected, there was a poorer performance compared to using just one classifier. Hence, stacking was not used in our final solution.

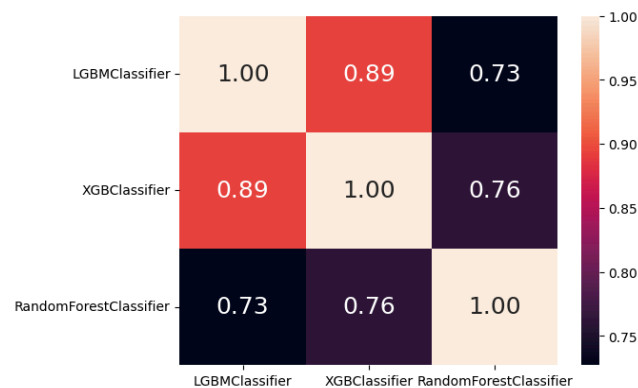


Figure 12 : Correlation between predictions on validation set

4. Experimental Study

In our experimental study, we performed statistical analyses such as One-way ANOVA and Pearson's Chi-Squared test to identify significant features. This analysis has also helped thus to figure out features that were deemed statistically insignificant, such as lesion_3 and rectal_temp and they were considered for removal or transformation using feature engineering to improve their utility to the model.

For modelling, we initially explored classic machine learning algorithms like SVM and logistic regression, but they did not perform well due to the high dimensionality of the

data and the multi-class nature of target variables. Decision trees showed better performance, which prompted us to employ ensemble learning techniques to help improve the robustness of the model. We implemented LightGBM, which leverages on histogram-based boosting and leaf-wise growth, and this leads to deeper trees and greater loss reduction. Even though we tried hyperparameter tuning using GridSearchCV, we found that the default parameters provided the best performance. This is most likely due to extensive tuning which led to overfitting for the given dataset size. We have also experimented with stacking various classifiers, but due to the high correlation among their predictions, this approach also did not improve performance over the single best model.

5. Conclusion

In conclusion, our team has successfully developed a robust machine learning model to predict horse survival outcomes by leveraging on advanced ensemble learning techniques. LightGBM was the most effective algorithm, outperforming other models due to its efficient computation and ability to handle high-dimensional data. The statistical analyses guided us in selecting key significant features which helped to enhance the model's predictions. Although hyperparameter tuning and stacking did not yield much improvements, we achieved a public score of 0.85365, which places us in the top 5% of the leaderboard.

Our findings also highlight the importance of feature engineering and the effectiveness of LightGBM in handling complex classification tasks in veterinary health data. Future work could focus on gathering more data to mitigate overfitting and exploring additional feature engineering strategies. This model does not only advance predictive capabilities but it also offers valuable insights in factors that influence horse survival, potentially aiding veterinarians in making informed decisions.

6. References

- Alam, S., Ayub, M. S., Arora, S., & Khan, M. A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, 9, 100341. <https://doi.org/10.1016/j.dajour.2023.100341>
- Suenaga, D., Takase, Y., Abe, T., Orita, G., & Ando, S. (2023). Prediction accuracy of Random Forest, XGBoost, LightGBM, and artificial neural network for shear resistance of post-installed anchors. *Structures*, 50, 1252–1263. <https://doi.org/10.1016/j.istruc.2023.02.066>
- Yin, L., Ma, P., & Deng, Z. (2021). JLGBMLoc—A Novel High-Precision Indoor Localization Method Based on LightGBM. *Sensors*, 21(8), 2722. <https://doi.org/10.3390/s21082722>