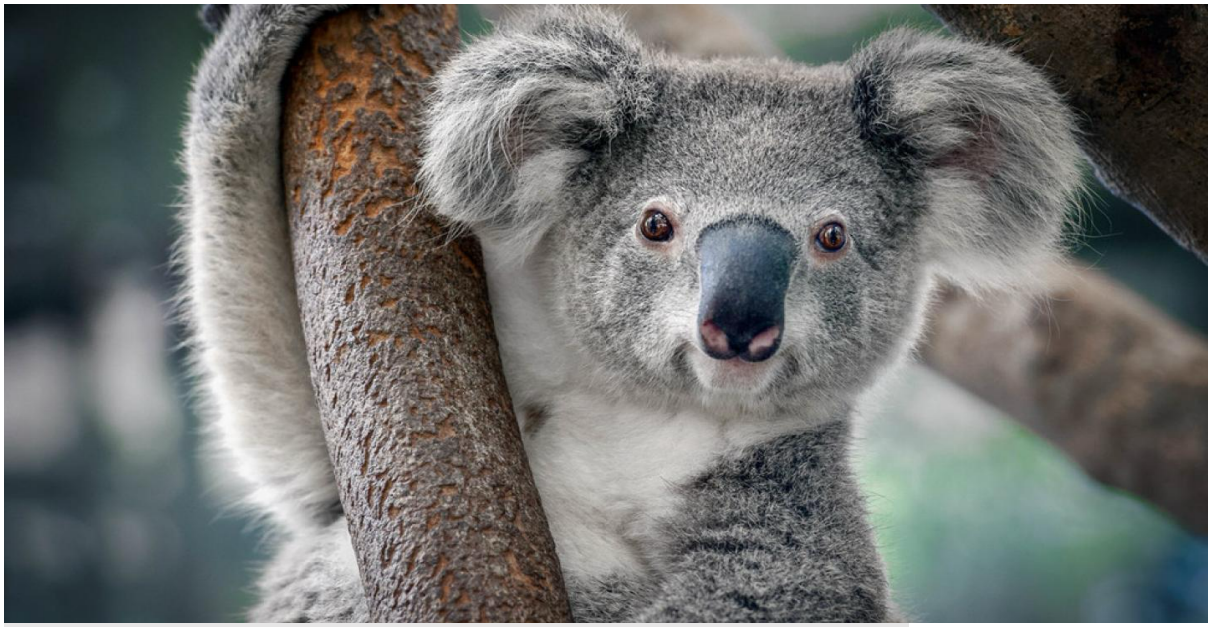
THE UNIVERSITY OF QUEENSLAND
AUSTRALIA



Koalas' hospital records analysis in Queensland

Data7001 Group I

Jie Zhang

We give consent for this report to be used as a teaching resource.

Table of Contents

1 Executive Summary

1.1 Research Aim and Motivation

The research is to investigate the many causes of the decline in the number of Koalas in Queensland.

This project's dataset was obtained from the Queensland Government Open Portal, which contains data from 1996 to 2022. Our study analysis revealed the primary causes of injured, sick and dead koalas in the wild. Also attempted to comprehend the many patterns of these reasons, which change by year and location.

Koalas are endemic to Australia, but their numbers have been declining. We hope that our initiative will act as a motivator for stakeholders such as the Animal Protection Organization, the State Government, and the local community in order to raise awareness about this problem. We believe that our study sheds light on the issue and aids to the preservation and prevention of Kolas in society.

1.2 Summary of Results

The primary insights from this study are that by successfully understanding these prediction models, we will be able to maintain Koloas in the future, particularly in Queensland suburbs, thereby helping to safeguard Australia's fauna.

The research has essentially accomplished its purpose, acquiring several insights via the use of many types of models and categorisation approaches. We were able to determine the primary causes of mortality in Koalas, which were strongly connected to the various variables of injury and disease. This research also helps to provide light on how these figures change between locations and years. In conclusion, we discussed several crucial techniques for preventing Koala population declines.

2 Introduction

2.1 Current issue: Koala population decline

The number of Koalas is decreasing in Australia at an alarming rate, this is due to many reasons such as habitat loss, accidents, bush fires, etc. “The Australian Koala Foundation has found that there are less than 57,920 Koalas left in the wild. In fact, the WWF-Australia projects that koalas will become extinct in the wild in eastern Australia by as early as 2050.” [1] With the help of our project, we want to spread awareness about these situations for the betterment of wildlife.

2.2 Research Analysis

To analyze and understand the dataset we first divided the data into two major factors that is injury and sickness. Then we classified different causes as mentioned below:

Injury: There were six factors classified under injury that are Vehicle Hit, Caused by Dog, Fall, Caught in Human Place, Attacked by Animals, Injury.

Sickness: There are ten different factors causing sickness in Koalas, that is Cystitis, Wasted, Conjunctivitis, Pneumonia, Blind, Chlamydiosis, Cancer, Bursitis, Nephritis and Sick.

By using the above factors, we build different models and studies how these trends over years and over different regions in Queensland.

2.3 Research Target

We mainly focus on the following questions and followed different approaches to answer this question. We then explained the results with the help of storytelling.

1. On what basis is the number of Koalas decreasing?
2. What different independent and dependent variables in this dataset?
3. Understanding how these factors are correlated to the death of these animals?

2.4 Key Stakeholders

The table 1 below examines important project stakeholders and provides a summary of their possible interests in this study.

Stakeholders	*Explanation*
Koalas	Avoid extinct species for a healthy growing environment.
Animal Protection Organization	Industry, researchers, veterinarians, non-governmental organizations, zoos and wildlife parks, community groups.
State Government	Knowing where koalas are frequently hurt can help to designate signage and give more organizational assistance. Implement new policies to reduce the number of injuries. Increasing budgets.
Local community	Informing people about koala conservation.

table 1 -Key stakeholders and explanation

3 Problem Solving with Data

3.1 Existing Problem

In Australia, there is presently no legislation in place to preserve koalas and their environment. The Koala's designation as "vulnerable" under the Environmental Protection and Biodiversity Conservation Act in 2012 made no difference and did not apply to Koalas in Victoria or South Australia. The Koala was designated as "Endangered" under the Environmental Protection and Biodiversity Conservation Act in 2022, and the law is no longer suitable for purpose. How come the listing got worse if it worked? This is meant to be Australia's leading environmental law. However, it has no teeth. The bulldozers may already be at work by the time you read this page, but it is not too late to intervene.

“In April 2012, the Australian Government declared the Koala as ‘VULNERABLE’ under the Federal EPBC Act in NSW, the ACT and QLD. Victoria and South Australia were excluded from the listing. In February 2022 the Koala was listed as ‘ENDANGERED’ in QLD, NSW, and ACT under the EPBC Act. The AKF believes that the Koala should have been listed in all States. Research conducted by the AKF strongly suggests the Koala’s conservation status should be upgraded to “CRITICALLY ENDANGERED” in the Southeast Queensland Bioregion as the Queensland Minister for the Environment has declared them to be ‘functionally extinct’.” [2]

Our approach focuses on the Queensland area, which has suffered a 53% drop. To further understand this, we used a dataset from the Queensland Data Source open site. Then, as described in the following parts, I studied various elements of it to get the necessary findings for stakeholders.

3.2 Main Focus

3.2.1 Outline of the Report

We structured our report in a simple manner which is easy to understand. At first, we started the report with the introduction, followed by the problem solving with data which outlines the main causes of decline of koalas and it's concerning issues.

For this we collected the necessary datasets from a valid website. And performed data cleaning steps. This process of data imputation is performed under is my data fit for use. Next, we analyzed different aspects of information in this dataset.

Lastly, we portrayed our results and conclusions in the form on storytelling. This helps to understand how Koalas are decreasing in Queensland and how this emergency needs to be addressed by the stakeholders.

The above-mentioned steps are mentioned as a detailed analytical process in the following sections of the report.

3.2.2 Design Thinking

We gain from design thinking, which is a powerful technique, in our project.

We used agile and iterative processes at various phases of the projects, for example: we completed many adjustments from the trial presentation and made a lot of progress throughout the modest steps iterative process. A simple graphic may go a long way towards generating comments and encouraging stakeholders to consider what insights are important to them. The sooner and more frequently you do it, the better.

4 Getting the Data We Need

4.1 Data Source

The main dataset for this project is from the Queensland Government open data portal. https://www.data.qld.gov.au/dataset/koala-hospital-data

This main dataset is about koala which has records from the year 1996-2022 in hospital.

4.2 Data quality

Our data is legitimate and accurate because it comes directly from the government, so you can rely on it. Following a review of the original data, we discovered that they are also organized but there were some incomplete records or records which we didn't require in our project so we had to clean the dataset before taking that dataset into consideration.

Original data looks like table 2:

_id	Record ...	Koala N...	Call Dat...	LAT	LNG	Adult Fa...	Adult Size	Adult G.
1	A 55200	EOA	2016-01...	-27.5042...	153.241...	RSPCA	Adult (M...	
2	A 63587	EOA	2016-01...	-27.1303...	152.918...	RSPCA	Adult (M...	
3	A 62232	ELANISHA	2016-01...	-28.1259...	153.465...		Adult (M...	
4	A 60186	SAUCE	2016-01...	-28.2514...	152.891...	separate...	Sub-Adu...	
5	A 59732	EOA	2016-01...	-27.2903...	152.996...		Adult (M...	
6	A 60856	TOBY	2016-01...	-27.3621...	152.940...		Adult (M...	

table 2 - original data

Note: Koala database from the year 1996-2022 in hospital, Queensland. (https://www.data.qld.gov.au/dataset/koala-hospital-data)

4.3 Data Privacy

Koala's hospital records, including Koala Name, Latitude, Longitude, Adult Size are all public information, so there was no data privacy issue in our project.

5 Is the Data Fit for Use

5.1 Data preprocessing

We have gathered data from the Queensland government open data portal. The dataset consists of 56935 rows and 41 columns such as (Koala Name, Latitude, Longitude, Adult Size etc.)

5.2 Data Cleaning Process

Data is cleaned to increase productivity and allow for the highest quality information for decision making. We combined 3 original datasets together and using tableau and excel to clean data. When there is too much data, it might be difficult to see what is needed or appropriate, impacting productivity and efficiency.

5.2.1 Incorrect records

Variables were not recorded, but comments wrote down the details. Some of the incorrect records are mentioned below:

494 under threat but records were False.

116 orphan records were not recorded correctly, comments wrote mum dead.

86 records should be Cystitis.

57 ticks' records, only 40 were recorded as sick.

40 records were Hit by train.

Imputation Methods

Specific reasons: Caused By Dog/ Orphaned/ Under Threat/ Conjunctivitis/ Cystitis Wasted/ Vehicle Hit/ Fall using FALSE or TRUE to fill, according to the comments' details.

5.2.2 Variables we don't use

In this report we use limited variables to do analysis, some of the columns we can use in future study. Important variables involved but records don't make sense.

Delete

Record no/ koala name/ Post code/ LAT/ LNG/ Adult Fate/ Other Adult situation/ Other Young Fate/ Other Koala/ Found Address/ Koala Location/ Description Road/ Speed Limit/ Release

Date/ Release Location/ Release Suburb Release Post Code/ Release LAT/ Release LNG/ Field Comments sick are blank and injured are blank 88 records.

5.2.3 Important lost variables

301 records of attacked by animals (farm: mainly cows or horses): create a new column Attacked by animals 306 records of blind: create a column Blind create Cancer 75 records, create Bursitis 75 records, create 444 Caught in human place.

5.2.4 Making the data confess

Combine 3 main datasets. Final main dataset is: 56847 rows 28 columns.
Data Quality: Complete and clean.
Records in the table were in the form of True and False. So, to carry out the analysis process we had to convert all those record to numerical values i.e., in the form of 0 and 1.

6 Visual Exploratory Data Analysis

6.1 The significance and goal of exploratory data analysis

Exploratory data analysis is the attempt to explore the structure of a data set or a particular pattern of data with as few assumptions as possible. Common methods of
exploratory data analysis includes graphing, tabulating, and calculating feature sizes. Exploratory data analysis is a very important and useful step before attempting data modelling, a concept introduced and named by the famous American statistician John Tukey. Exploratory data analysis is valuable to data scientists because it ensures that the results, they generate are valid, can be correctly parsed and are applicable to the desired objectives. A more specific reason why exploratory data analysis is a necessary step to take before modelling is that data scientists need to have an in-depth understanding of the data and develop an intuition that can quickly identify where problems arise. Without exploratory data analysis, problems such as missing data, data outliers or errors in model selection can go unnoticed for long periods of time, which can lead to data scientists making decisions based on incorrect information or creating a host of other problems.

6.2 Visualization of data

Visualization of data is the most common approach to exploratory data analysis, using it to assess patterns and describe the data using several quantitative methods. "Exploratory data analysis places greater emphasis on intuition and data visualization, on methodological variety and flexibility, so that the analyst can see at a glance the valuable information implicit in the data, show the general patterns it follows and the distinctive and outstanding features that make it unique, facilitate the discovery of patterns and enlightenment, and meet the multifaceted requirements of the analyst, which is the main contribution of EDA to data analysis. the main contribution of EDA to data analysis." This presentation will use data visualization to explore the various causes of koala fatalities.

First, we used time series plots to visualize and analyze data on koala survival, and mortality data by year. The leftmost subplot in Figure 1 shows the change in the number of surviving koalas, the middle table shows the number of koala deaths by year, and the rightmost graph shows the total number of koalas seen in Queensland. The graph shows that the number of surviving koalas, the number of deaths and the total number of hospital records all peaked in 2006. At one point the number of koala deaths reached an unexpected 737 and the total number of visits reached an unexpected 4489.

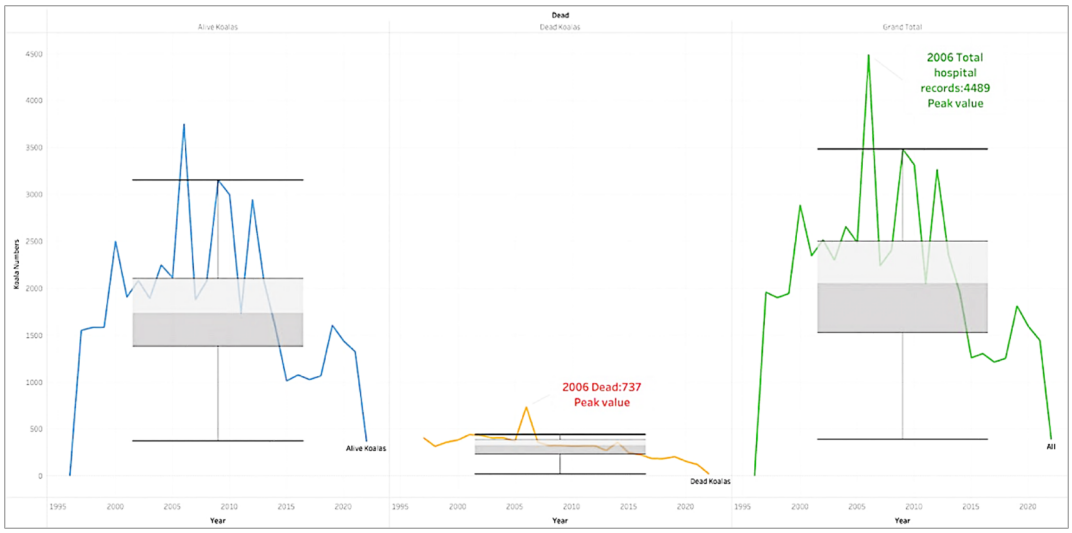


figure 1- Annual koala mortality trends

After we had identified the theme of studying the causes of koala mortality, we decided to explore the causes of koala mortality in two ways, namely injury and disease. Table 3 and table 4 show the common causes of injury and disease and their frequency and proportion,

respectively.

S. No	Injured Koalas	Frequency	Proportion
1	Vehicle Hit	11329	19.92%
2	Caused By Dog	4150	7.30%
3	Fall	1045	1.83%
4	Caught in Human Place	443	0.77%
5	Attacked by Animals	301	0.53%
6	Injured	11263	19.81%

table 3 - The most common causes of injury

From the proportion index in table 3 it can be concluded that the top three most frequent factors leading to injury were mainly vehicle impacts, caused by dogs and falls. The two factors that caused the lowest frequency of injuries to koalas were Attacked by Animals and Caught in Human Place.

S. No	Causes of Sickness	Frequency	Proportion
1	Cystitis	11289	19.85%
2	Wasted	10918	19.20%
3	Conjunctivitis	8365	14.71%
4	Pneumonia	435	0.76%
5	BLIND	306	0.53%
6	Chlamydiosis	152	0.26%
7	CANCER	83	0.14%
8	Bursitis	75	0.13%
9	Nephritis	59	0.10%
10	Sick	20418	35.91%

table 4 - The most common causes of sickness

Table 4 shows the common causes of illness in koalas. The top 3 leading causes of sickness are Cystitis, wasted and Conjunctivitis. Diseases such as blindness, cancer and Nephritis are

very rare.

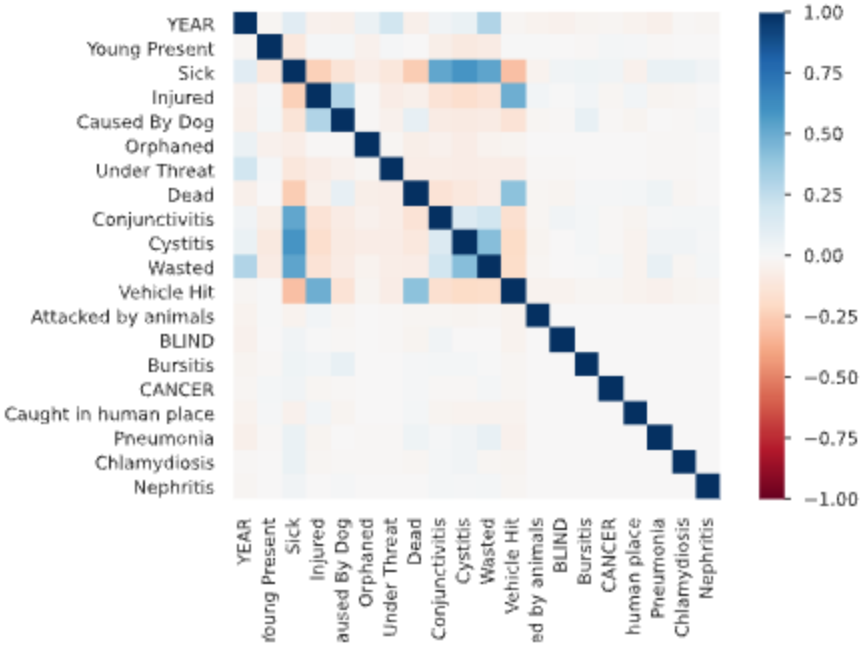


figure 2 - The correlation between all variables

The above descriptive analysis explores the relationship between the causes of injury and the relationship between the causes of illness, respectively. The two are now combined for analysis.

Figure 2 shows a correlation matrix that provides us with the following conclusions:

1. In the case of injury. Many dog attacks were in the context of injuries: many of the attacks were in the presence of young people, while most of the injuries to koalas were when they were attacked by animals.

2. Similarly, in the case of illness. Many diseases are directly linked to death. directly to death. In addition to disease, vehicle impacts are a cause of death on site. Impacts are also a cause of death on site.

7 Model Generation

7.1 Define Variables

Following data exploration and visualization of all aspects in the dataset, we settle on two sets of Independent Variables Sets:

Group1: Injury related variables sets are:

Caused by Dog, Vehicle Hit, Fall, attacked by animals, caught in human place, Under Threat and Injured itself, in total is 7.

Group2: Sickness related variables sets are:

They are Conjunctivitis, Cystitis, Wasted, Bursitis, cancer, Pneumonia, Chlamydiosis, BLIND, Nephritis and Sick itself, in total 10.

Dependent Variable: Death

Because we are going to use injury related variables sets and sickness related variables sets to predict death. These independent variables may be summarized in table 5.

Group	Independent Variables Sets	Explanations
Group 1	**Injury related variables**	Caused by Dog
		Vehicle Hit
		Fall
		Attacked by animals
		Caught in human place
Group 2	**Sickness related variables**	Under Threat
		Conjunctivitis
		Cystitis
		Wasted
		Bursitis
		cancer
		Pneumonia
		Chlamydiosis
		BLIND
		Nephritis
		Sick

table 5 – Variables Definition

7.2 Model Selection

The process of picking one final machine learning model from a group of candidate machine learning models for a training dataset is known as model selection.

Fitting models is simple but choosing among them is the fundamental problem of applied machine learning.

To begin, we must abandon the notion of a "best" model.

Given the statistical noise in the data, the incompleteness of the data sample, and the limits of each model type, all models have some prediction inaccuracy. As a result, the concept of a perfect or best model is useless. Instead, we must look for a "good enough" model.

In our study, we tested numerous categorization models (supervised and unsupervised) to produce predictions, as well as statistical and performance assessments.

In the next section, we will go through our assessment criteria in further detail.

7.3 Model Performances Evaluations

7.3.1 Key Assessment Criteria

The first assessment measure that anyone would employ is "Accuracy".

Accuracy is defined as the percentage of correct predictions made to total forecasts made.

Precision: a classification model's ability to identify just the relevant data points. As a result, it is also essential.

However, they are not enough to evaluate classification models. The ROC curve or AUC are primarily used to assess and compare various learning models, specifically shows in figure 3.

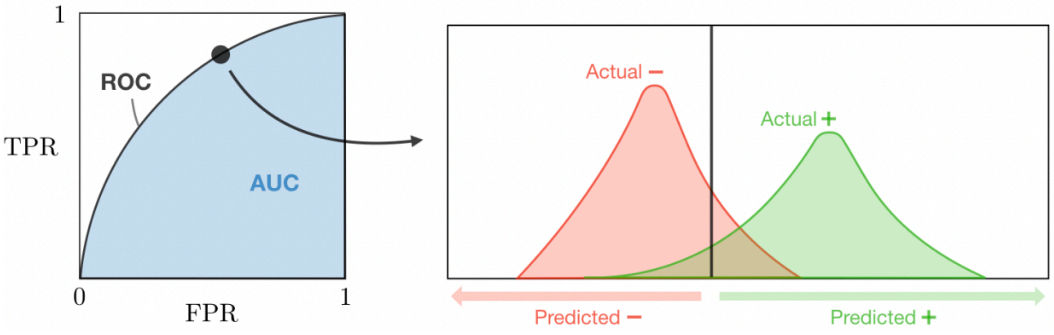


figure 3 - ROC and AUC definition

Note:

TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

y_i is true value, $\hat{y}_i$ is prediction, N is the number of samples.

Evaluation Criteria	Mathematics Explanations
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
F1 score	$\frac{2TP}{2TP + FP + FN}$
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
TPR	$\frac{TP}{TP + FN}$
FPR	$\frac{FP}{TN + FP}$
ROC/AUC	*The area under the receiving operating curve, also noted AUC or AUROC, is the area below the ROC as shown in the following figure:*

table 6 - Assessment Criteria and mathematics explanations

7.4 Model performances

7.4.1 Logistic Regression

Despite its name, logistic regression is a classification model rather than a regression model. For binary and linear classification issues, logistic regression is a simpler and more efficient technique [3].

logistic function $p(x) = \frac{1}{1+e^{-(x-\mu)/s}}$

note: where μ is a location parameter (the midpoint of the curve, where p(μ)=1/2 and is a scale parameter [4].

7.4.1.1 Model Training - Logistic Regression

We split data into 70% training data and 30% testing data. Use logistic regression algorithm to fit the training data, and to predict the test data.

Following that, we may assess the model using a matrix.

Group1: Injury related variables to predict Death

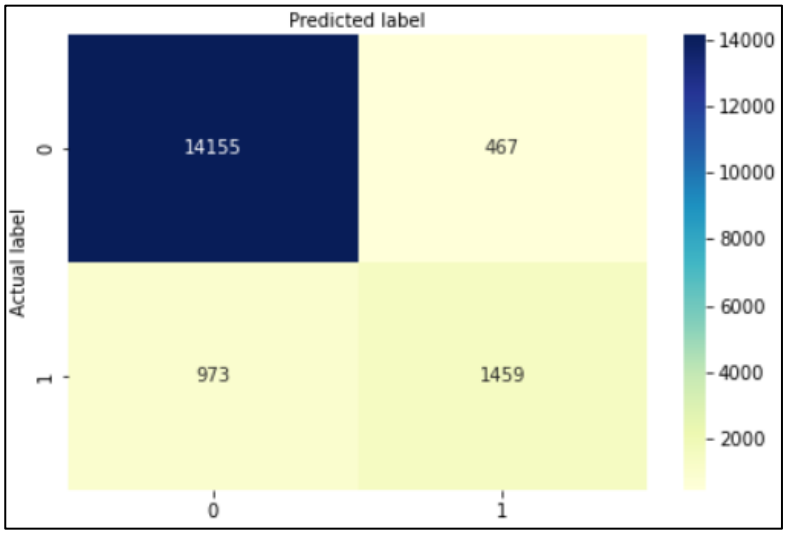


figure 4 -Confusion matrix of logistic regression (Injury model)

	precision	recall	f1-score	support
Alive	0.94	0.97	0.95	14622
Dead	0.76	0.60	0.67	2432
accuracy			0.92	17054
macro avg	0.85	0.78	0.81	17054
weighted avg	0.91	0.92	0.91	17054

table 7 - Evaluation index of logistic regression (Injury model)

The Injury model's conclusion may be drawn from the table 7, which is Logistic regression predicts Koala death by injury with a 92% accuracy. Recall and f1-score are good for predicting “Alive”.

However, the model has limitations, such as a precision of 76% for predicting "Dead," as well as recall (60%) and f1-score (67%) respectively.

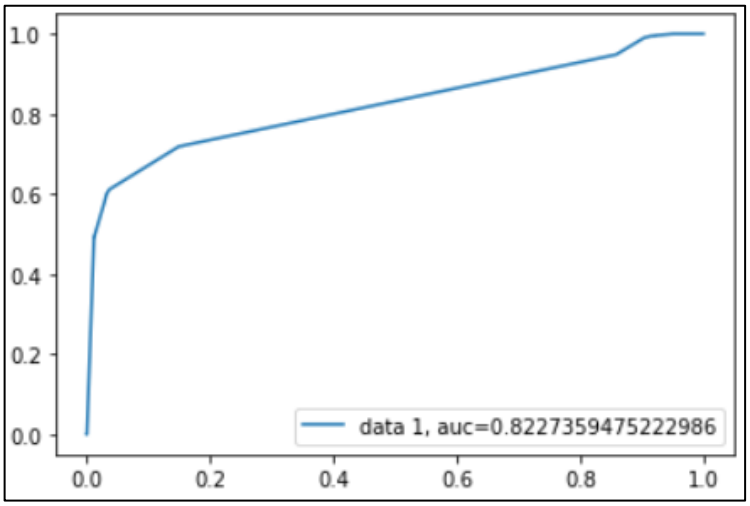


figure 5 - ROC/AUC of logistic regression (Injury model)

When we plot the ROC/AUC to validate the performance of the logistic regression model, the AUC of the Injury model is 82%, indicating that the model performs well.
According to this, let's look at the Sick model in the subsequent research.

Group2: Sick related variables to predict Death

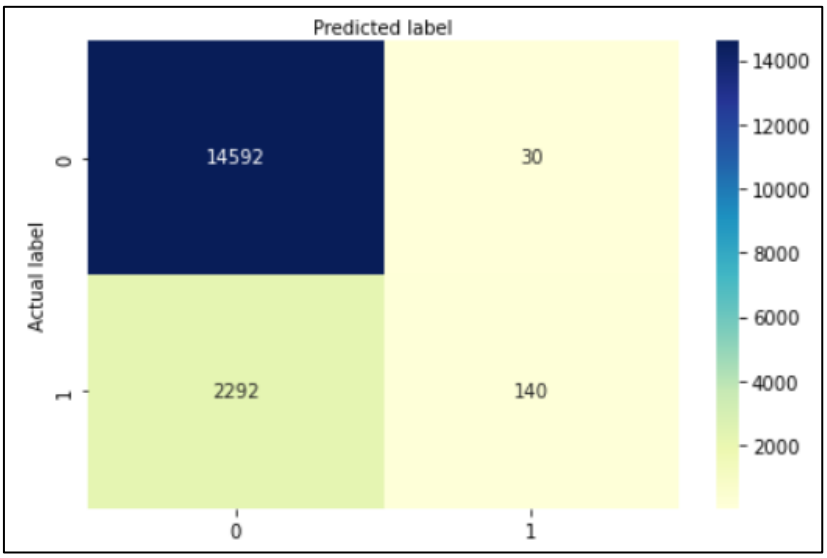


figure 6 -Confusion matrix of logistic regression (Sick model)

	precision	recall	f1-score	support
Alive	0.86	1.00	0.93	14622
Dead	0.82	0.06	0.11	2432
accuracy			0.86	17054
macro avg	0.84	0.53	0.52	17054
weighted avg	0.86	0.86	0.81	17054

table 8 - Evaluation index of logistic regression (Sick model)

The image illustrates the Sick model's conclusion: Logistic regression predicts Koala mortality by injury with an 86% accuracy. While the model except recall is only 6%, the other performance indices are satisfactory. Predicting "Alive" with precision (86%), recall (100%), and f1-score (91%).

However, the model has limitations: recall is only 6% and f1-score is only 11% for predicting "Dead." Other approaches for assessing model performance are still required.

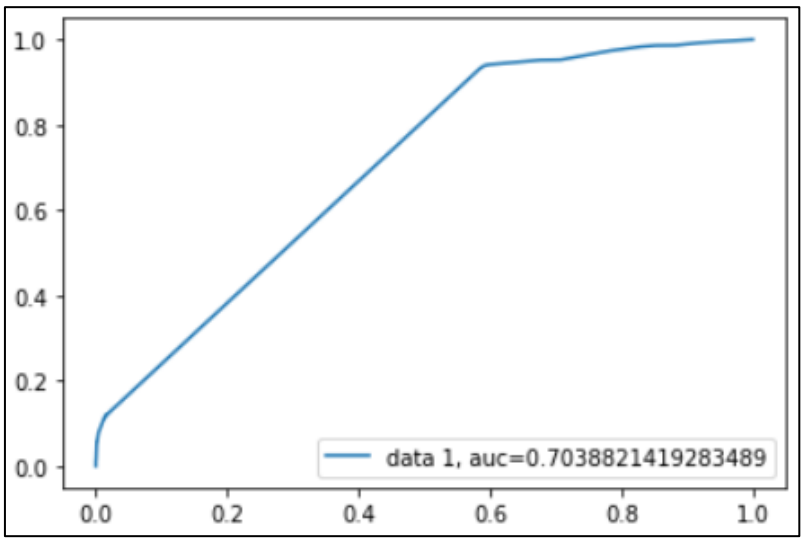


figure 7 - ROC/AUC of logistic regression (Sick model)

When we plot the ROC/AUC to assess the logistic regression model's performance, the AUC of the ill model is 70%, suggesting that the model works well.

Finally, the logistic regression technique performs well in both the Injury and Sick models. Furthermore, according to the confusion matrix and ROC/AUC studies, the Injury model outperforms the Sick model.

7.4.2 K Nearest Neighbors (KNN)

Figure 8 explains how k-Nearest Neighbors operate. Given a k value, what will be the prediction?

Because green is the most common colour in the k=3 circle, new data points will be projected to be green.

Because blue is the most common colour in the k=6 circle, new data points will be projected to be blue [5].

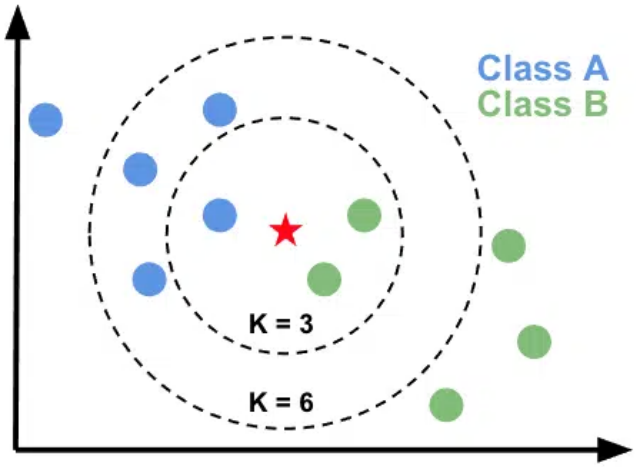


figure 8 – KNN algorithm

How to choose k

In the picture, we plot several ks with accuracy; k=2 is the best choice since the accuracy is the greatest, which is 85%; after that, the accuracy does not vary. Consequently, k=2 is the most suitable alternative in Sick model.

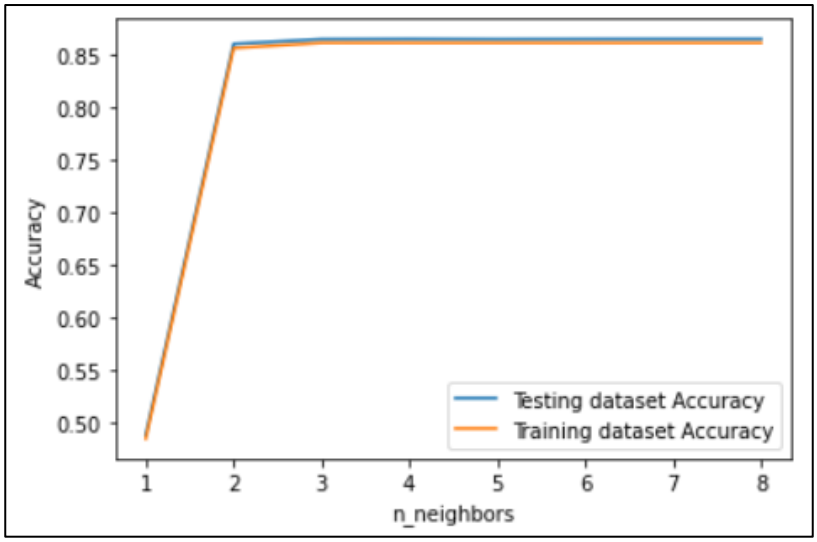


figure 9 -k values in KNN (Sick model)

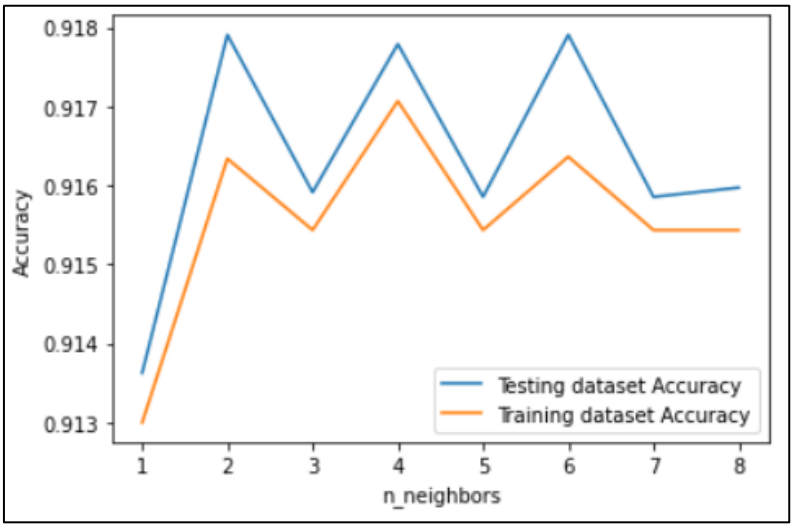


figure 10 -k values in KNN (Injury model)

We may also pick the optimum k from the figure 10 ; similarly, when k=2, the accuracy approached 90%, but we discovered that k=4 and k=6 had equal accuracy.
However, because of the model's complexity, we shall select the least k, which is k=2.

7.4.2.1 Model Training - KNN

In the same approach, we divided the data into 70% training data and 30% testing data. Use the KNN algorithm to fit the training data and predict the test data.
After fitting, the Sick model's testing data prediction accuracy is 86%, while the Injury model's testing data prediction accuracy is 92%.
This suggests that, in general, KNN is another acceptable model for classification in our study.

7.4.3 Other models

For fitting, we also attempted **Support Vector Machines (SVM)** and **K Means Clustering**.
However, SVM findings show that precision is only 61% and recall is just 8.1%.
And K means' performance is significantly worse, with only 23.5% accuracy. As a result, we did not choose to employ these two algorithms.

7.5 Models summary

Models	*Evaluations*	*Results*
Logistic Regression	Performance good both in Injury model and Sick model.	√
KNN	Both the Injury model and the Sick model have high prediction accuracy.	√
SVM	Precision and recall levels do not match expectations.	×
K means clustering	Prediction accuracy is just 23.5%, which is lower than that of guessing.	×

table 9 -Models' comparison

Our research indicates that the models we can use are logistic regression and KNN. In reference to table 9.

8 Statistical Analysis

In this section, we look at how the cause of injury affects death and how the cause of sickness affects death. We selected to fit the data using a classification model because the data are all type 0-1.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	33889.993[a]	.329	.522

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

table 10 - Logistic regression model summary (injury)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CausedByDog	3.932	.044	8115.959	1	.000	51.013
	Orphaned	.111	.096	1.339	1	.247	1.118
	UnderThreat	-1.648	.280	34.559	1	.000	.192
	VehicleHit	3.806	.035	11668.766	1	.000	44.991
	Fall	3.640	.070	2738.859	1	.000	38.088
	Attackedbyanimals	2.775	.127	480.306	1	.000	16.037
	BLIND	1.865	.156	142.820	1	.000	6.453
	Caughtinhumanplace	2.556	.110	536.763	1	.000	12.885
	Constant	-3.449	.030	13462.559	1	.000	.032

a. Variable(s) entered on step 1: CausedByDog, Orphaned, UnderThreat, VehicleHit, Fall, Attackedbyanimals, BLIND, Caughtinhumanplace.

table 11 - Logistic regression about injured.

Table 10 and table 11 explores the relationship between injuries and the various causes of injury. One of the indicators of how well a logistic regression model fits is the -2log likelihood. the larger the value, the better the model fit. the -2log likelihood value for injured and its various causes of injury is 33889.993, which demonstrates that the model fits well.
The p-value of Orphaned in this model is greater than 0.05, so it is not significant and needs to be removed. The three factors that can most influence injury are caused by dogs, vehicle impact, and fall.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	29198.068[a]	.547	.751

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

table 12 – Logistic regression model summary (sick)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Conjunctivitis	4.955	.058	7231.718	1	.000	141.832
	Cystitis	4.153	.044	8901.489	1	.000	63.649
	Wasted	3.123	.040	6105.004	1	.000	22.705
	BLIND	1.513	.173	76.377	1	.000	4.541
	Bursitis	4.207	.492	73.056	1	.000	67.130
	CANCER	3.132	.298	110.738	1	.000	22.925
	Pneumonia	1.826	.158	133.928	1	.000	6.212
	Chlamydiosis	5.155	.520	98.239	1	.000	173.230
	Nephritis	3.859	.509	57.557	1	.000	47.423
	Constant	-2.661	.021	16003.872	1	.000	.070

a. Variable(s) entered on step 1: Conjunctivitis, Cystitis, Wasted, BLIND, Bursitis, CANCER, Pneumonia, Chlamydiosis, Nephritis.

table 13 - Logistic regression about sick.

The two R-squares in the table 12 represent the two interval values of the proportion of variation in the model that can be explained by the dependent variable. For example, in this logistics model, sick is the dependent variable, and the cause of the sick koala is the independent variable. The interval of variation explained by sickness in the logistics model is [54.7% to 75.1%]. The three factors that can most influence sickness are chlamydia, conjunctivitis, and Bursitis.

9 Storytelling and Recommendations

9.1 Storytelling

We now have a better grasp of the elements that lead to koala fatalities as a result of our past research.

Natural catastrophes or forest fires may have contributed to an increase in koala injuries and illnesses in Queensland in 2006. Vehicle Hit is the most prevalent sort of koala injury, because to the sluggish movement of koalas, their inability to evade cars in time, and their own tiny size, which makes them less apparent to vehicle drivers. Because of the spread of the chlamydia virus among wild animals, cystitis has become the most frequent ailment in koalas. According to the logistic regression results, dog assaults on koalas are more likely to cause harm, and chlamydia is the most frequent cause of koala sicks, which would explain why cystitis is the most common disease in koalas.

Because the logistic regression model discussed previously only captures information from about half of the data, I aim to add some more factors to the variables in the future to study the data from as many perspectives as feasible. Deep-learning models for predicting independent variables might also be developed.

9.2 Recommendations

Queensland lost 700 hectares of tree cover between 2001 and 2021, accounting for 6.5% of the total loss since 2000. The state government's decision to reduce tree cover in urban development plans has resulted in changes in the regional climate, which has resulted in severe weather. To promote the state's greening, the government must grow its revenue and play its full role as the backbone of Queensland's important ecological initiatives. The Department of

Transport might impose speed limits in places frequented by koalas to help cars avoid injuring them. Because domestic dogs are softer in nature, the state authorities may even prohibit hounds in places favored by koalas.

To restrict the spread of the Chlamydia virus, animal protection organisations should arrange volunteer activities such as spraying the forests. Increase funding for studies into preventing koala diseases in order to lower the risk of koala illnesses. In addition, veterinarians should make regular trips to the forest or other areas where koalas reside to assist them.

Local communities might offer talks on the importance of koala conservation to boost awareness, or perhaps add a subject on koala protection to the curriculum of local schools.

10 Conclusion

From the analysis and models, we made we were able to draw the following conclusions:

1. The main causes of decrease in number of Koalas are due to Vehicle Hit which causes death at the spot and Sick.
2. There are two dependent variables Sickness and Injury whereas Death is an independent variable.
3. The year 2006 saw a peak as injury and sickness rate was very high this year. However, the yearly hospitalization records were decreasing over the years which is a good sign.

For obtaining the above findings we built different models like KNN, SVM, K-means, Logistic regression. The Logistic Regression and KNN stood out the best.

As a word of advice for our stakeholder, the animal protection organization should start research on koala sickness to decrease the rates of serious sickness among these animals. The state government can also take responsible actions to preserve the habitat of Koalas, for example by posting sign boards where Koalas live can prevent road accidents significantly. Lastly the local community plays an important role in helping the preserve lives of these native animals.

We believe that there is still possibilities of future work in our topic as this topic is very closely related to wildlife in Queensland Australia. For further research we would like to work on more variables(factors) and perform location clustering. Also, we would want to generate deep learning models to predict independent variables.

11 Improvements and Management of change

Teachers and classmates provided suggestions on our trail presentation, so we altered our model generation and improved other aspects of our final presentation. More information about our change may be seen in the table 14 below:

Feedbacks	Improvements and changes
How does your project solve a problem? Don't be general. Talk about specifics.	Analyze the main causes leading koalas to death. Using different algorithm models to predict death.
Data collection: you didn't explain.	Describe how we get our main datasets
Data confess: predict death from injury and sickness. Why not use the type of injury and sickness? You can use a one-hot encoding of categorical variables (dependent) to predict death.	Define variables
What are the most frequent causes of death? What types of injury/sickness?	General data analysis - Injury/Sick Vehicle Hit caused Death/Sick also caused Death
Storytelling: The slides on deforestation have nothing to do with your analysis. Your conclusions should be taken from your analysis.	We abandon the deforestation data

table 14 - Feedbacks and improvements

References

[1][2] Australian Koala Foundation. “The Koala - Endangered or Not?” Australian Koala Foundation, 2020. https://www.savethekoala.com/about-koalas/the-koala-endangered-or-not/.

[3] www.sciencedirect.com. “Logistic Regression - an Overview | ScienceDirect Topics,” n.d. https://www.sciencedirect.com/topics/computer-science/logistic-regression.

[4]Wikipedia Contributors. “Logistic Regression.” Wikipedia. Wikimedia Foundation, April 12, 2019. https://en.wikipedia.org/wiki/Logistic_regression.

[5] Chouinard, Jean-Christophe. “K-Nearest Neighbors (KNN) in Python.” JC Chouinard, n.d. https://www.jcchouinard.com/k-nearest-neighbors/.

Appendix

Appendix A: Modeling - Injury model (python code)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_excel(r'/Users/jiezhang/Documents/7001 introduction to data science/group project/koala data/project data /koalabase combined_clean.xlsx',
sheet_name='Sheet1')
df.replace({False: 0, True: 1}, inplace=True)
```

```
feature_cols = ['Caused By Dog', 'Orphaned', 'Under Threat', 'Vehicle Hit','Fall','Attacked by animals','BLIND','Caught in human place','Injured']
X = df[feature_cols] # Features
y = df.Dead # Target variable
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=14)
```

```
# import the class
from sklearn.linear_model import LogisticRegression
# instantiate the model (using the default parameters)
logreg = LogisticRegression(random_state=14)
# fit the model with data
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
```

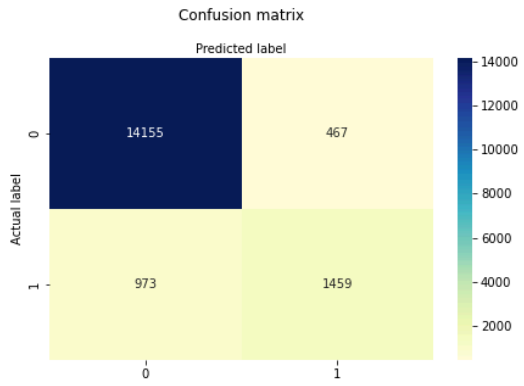
```
# import the metrics class
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

```
array([[14155,   467],
       [  973,  1459]])
```

```
# import required modules
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

```
Text(0.5, 257.44, 'Predicted label')
```



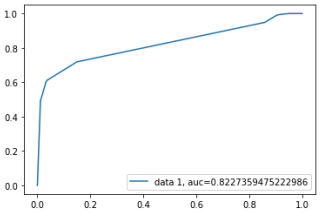

```
from sklearn.metrics import classification_report
target_names = ['Alive', 'Dead']
print(classification_report(y_test, y_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

       Alive       0.94      0.97      0.95     14622
        Dead       0.76      0.60      0.67      2432

    accuracy                           0.92     17054
   macro avg       0.85      0.78      0.81     17054
weighted avg       0.91      0.92      0.91     17054
```

```
y_pred_proba = logreg.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
plt.plot(fpr,tpr,label="data 1, auc="+str(auc))
plt.legend(loc=4)
plt.show()
```



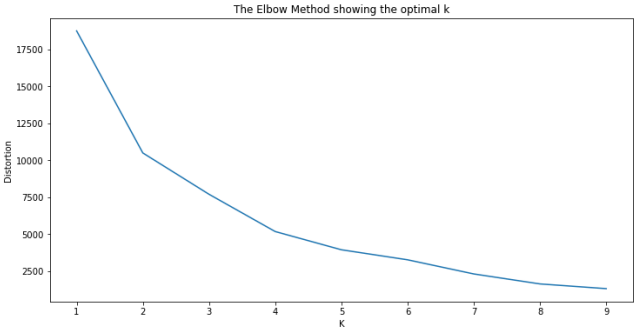

```
from sklearn.cluster import KMeans
kmeans = KMeans(2)
kmeans.fit(X_train)
identified_clusters = kmeans.predict(X_test)
identified_clusters
centers = kmeans.cluster_centers_
print(centers)
```

```
[[ 3.83568337e-02  2.52338406e-02  3.46572665e-02  7.88258347e-15
   1.20061427e-02  5.23523663e-03  5.79366187e-03  7.22462655e-03
  -1.06303855e-14]
 [ 1.60412926e-01  1.93895871e-02  1.79533214e-03  7.14811490e-01
   3.76122083e-02  6.28366248e-03  3.50089767e-03  9.15619390e-03
   7.08527828e-01]]
```

```
distortions = []
K=range(1,10)
for k in K:
    kmeanModel = KMeans(k)
    kmeanModel.fit(X_train)
    distortions.append(kmeanModel.inertia_)
print(distortions)
```

```
[18755.289832127037, 10496.208628384871, 7700.234787780818, 5178.255173036786, 3944.905915348315, 3266.060986081672, 2309.2586333473314, 1634.2101714479134, 1315.1429180443943]
```

```
# Plot the distortions of K means
plt.figure(figsize=(12,6))
plt.plot(K, distortions)
plt.xlabel('K')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```




```
score = metrics.accuracy_score(y_test,kmeans.predict(X_test))
print(score)
```

```
0.7645127242875571
```

```
#Import svm model
from sklearn import svm
#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel
rbf_svc = svm.SVC(kernel='rbf')
rbf_svc.kernel
#Train the model using the training sets
clf.fit(X_train, y_train)
#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy: how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.9155036941480005
```

```
# Model Precision: what percentage of positive tuples are labeled as such?
print("Precision:",metrics.precision_score(y_test, y_pred))
# Model Recall: what percentage of positive tuples are labelled as such?
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Precision: 0.7568688439606014
Recall: 0.600328947368421
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, y_train)
# Predict on dataset which model has not seen before
print(knn.predict(X_test))
train_accuracy = knn.score(X_train, y_train)
print(train_accuracy)
test_accuracy= knn.score(X_test, y_test)
print(test_accuracy)
```

```
[0 0 0 ... 0 1 0]
0.9163399678327302
0.9179078222117978
```

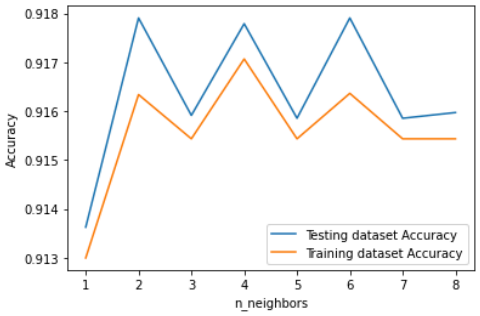
```
neighbors = np.arange(1, 9)
train_accuracy = np.empty(len(neighbors))
test_accuracy = np.empty(len(neighbors))

# Loop over K values
for i, k in enumerate(neighbors):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)

    # Compute training and test data accuracy
    train_accuracy[i] = knn.score(X_train, y_train)
    test_accuracy[i] = knn.score(X_test, y_test)

# Generate plot
plt.plot(neighbors, test_accuracy, label = 'Testing dataset Accuracy')
plt.plot(neighbors, train_accuracy, label = 'Training dataset Accuracy')

plt.legend()
plt.xlabel('n_neighbors')
plt.ylabel('Accuracy')
plt.show()
```



Appendix B: Modeling - Sick model (python code)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_excel(r'/Users/jiezhang/Documents/7001 introduction to data science/group project/koala data/project data /koalabase combined_clean.xlsx',
sheet_name='Sheet1')
df.replace({False: 0, True: 1}, inplace=True)
```

```
feature_cols = ['Conjunctivitis', 'Cystitis', 'Wasted', 'Bursitis','CANCER','Pneumonia','Chlamydiosis','BLIND','Nephritis','Sick']
X = df[feature_cols] # Features
y = df.Dead # Target variable
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=14)
```

```
# import the class
from sklearn.linear_model import LogisticRegression
# instantiate the model (using the default parameters)
logreg = LogisticRegression(random_state=14)
# fit the model with data
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
```

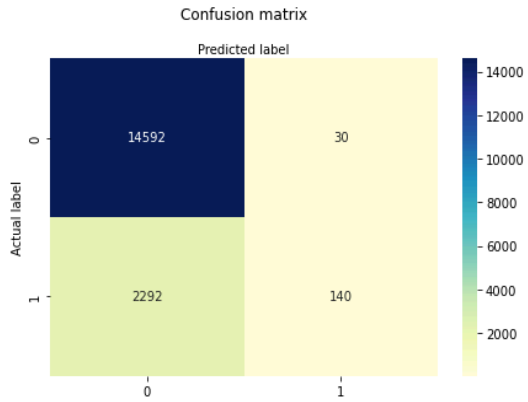
```
# import the metrics class
from sklearn import metrics
cnf_matrix = metrics.confusion_matrix(y_test, y_pred)
cnf_matrix
```

```
array([[14592,    30],
       [ 2292,   140]])
```

```
# import required modules
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

```
Text(0.5, 257.44, 'Predicted label')
```




```
from sklearn.metrics import classification_report
target_names = ['Alive', 'Dead']
print(classification_report(y_test, y_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

       Alive       0.86      1.00      0.93     14622
        Dead       0.82      0.06      0.11      2432

    accuracy                           0.86     17054
   macro avg       0.84      0.53      0.52     17054
weighted avg       0.86      0.86      0.81     17054
```

```
y_pred_proba = logreg.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
plt.plot(fpr,tpr,label="data 1, auc="+str(auc))
plt.legend(loc=4)
plt.show()
```

```
score = metrics.accuracy_score(y_test,kmeans.predict(X_test))
print(score)
```

```
0.4815292599976545
```

```
#Import svm model
from sklearn import svm
#Create a svm Classifier
clf = svm.SVC(kernel='linear') # Linear Kernel
rbf_svc = svm.SVC(kernel='rbf')
rbf_svc.kernel
#Train the model using the training sets
clf.fit(X_train, y_train)
#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy: how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

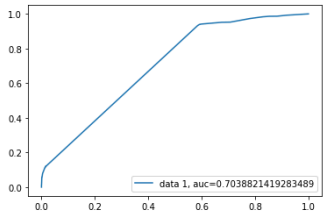
```
Accuracy: 0.8614987686173332
```

```
# Model Precision: what percentage of positive tuples are labeled as such?
print("Precision:",metrics.precision_score(y_test, y_pred))
# Model Recall: what percentage of positive tuples are labelled as such?
print("Recall:",metrics.recall_score(y_test, y_pred))
```

```
Precision: 0.6073619631901841
Recall: 0.08141447368421052
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
knn = KNeighborsClassifier(n_neighbors=2)
knn.fit(X_train, y_train)
# Predict on dataset which model has not seen before
print(knn.predict(X_test))
train_accuracy = knn.score(X_train, y_train)
print(train_accuracy)
test_accuracy= knn.score(X_test, y_test)
print(test_accuracy)
```

```
[0 0 0 ... 0 0 0]
0.8563530357860877
0.8603846604902076
```



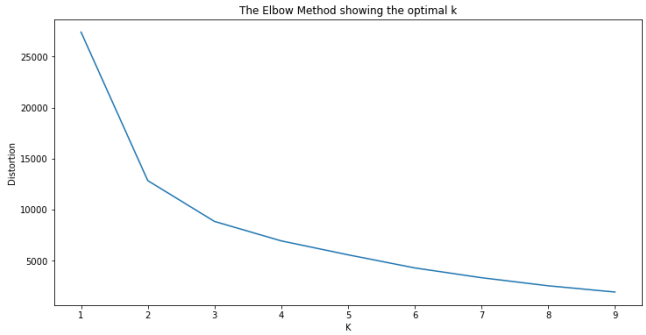

```
from sklearn.cluster import KMeans
kmeans = KMeans(2)
kmeans.fit(X_train)
identified_clusters = kmeans.predict(X_test)
identified_clusters
centers = kmeans.cluster_centers_
print(centers)
```

```
[[3.87794198e-01 5.23877942e-01 4.86247947e-01 3.21565408e-03
  2.94198139e-03 1.53256705e-02 7.25232622e-03 8.21018062e-03
  2.66830870e-03 9.74753695e-01]
 [6.03749603e-03 9.25484588e-03 2.40705434e-02 7.94407372e-05
  5.16364792e-04 2.38322212e-03 7.94407372e-05 3.37623133e-03
  7.94407372e-05 1.52655666e-14]]
```

```
distortions = []
K=range(1,10)
for k in K:
    kmeanModel = KMeans(k)
    kmeanModel.fit(X_train)
    distortions.append(kmeanModel.inertia_)
print(distortions)
```

```
[27397.784856251066, 12836.902420359227, 8832.608771236452, 6927.348677696677, 5563.901127942194, 4275.3940849028, 3315.7994201081115, 2522.677966013719, 1910.5887488927997]
```

```
# Plot the distortions of K means
plt.figure(figsize=(12,6))
plt.plot(K, distortions)
plt.xlabel('K')
plt.ylabel('Distortion')
plt.title('The Elbow Method showing the optimal k')
plt.show()
```



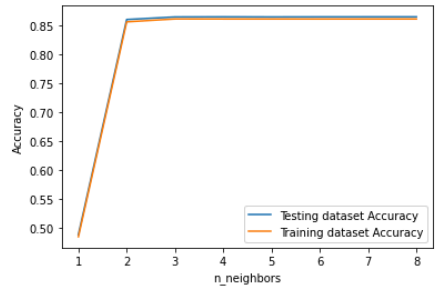

```
neighbors = np.arange(1, 9)
train_accuracy = np.empty(len(neighbors))
test_accuracy = np.empty(len(neighbors))

# Loop over K values
for i, k in enumerate(neighbors):
    knn = KNeighborsClassifier(n_neighbors=k)
    knn.fit(X_train, y_train)

    # Compute training and test data accuracy
    train_accuracy[i] = knn.score(X_train, y_train)
    test_accuracy[i] = knn.score(X_test, y_test)

# Generate plot
plt.plot(neighbors, test_accuracy, label = 'Testing dataset Accuracy')
plt.plot(neighbors, train_accuracy, label = 'Training dataset Accuracy')

plt.legend()
plt.xlabel('n_neighbors')
plt.ylabel('Accuracy')
plt.show()
```



Appendix C: Variables Correlation Analysis

The codes are shared at a public URL; to access the code, click here:
https://colab.research.google.com/drive/1FC-E1hOJ6Fwx3YVWUI3Yvul3gJiICXAT?usp=sharing#scrollTo=ytaCL7B4QP7M