Presidency of Ministers Council
Ministry of Higher Education
and Scientific Research
Koya University
Faculty of Engineering
Software Engineering Department

هەرێمی کوردستانی عێراق
سەرۆکایەتی ئە نجومەنی وەزیران
وەزارەتی خوێندنی باڵا و توێژینەوەی زانستی
سەرۆکایەتی زانکۆی کۆیە
فاکەڵتی ئەندازیاری
بەشی ئەندازیاری پرۆگرامساز ی

Animal Condition

Group - B

Prepared by/

Zhya Rebwar
Mohammed Nabaz
Frishta Abdulsamad
Zmnako Karwan

Barez Azad
Aya Hassan
Husen Adnan
Huda Haval

Supervised by/ Dr. Abdulbasit Kamil Faeq

Contents

List of Tables

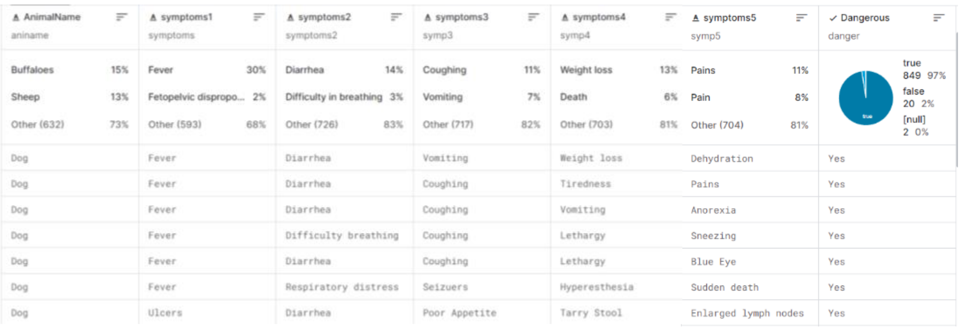
1. Introduction

Figure 1-1 As you go through the Animal Condition Classification

Diving into assessment of Animal Condition using Python and machine learning. Equipped with neural networks, KNN, Bayesian classifiers, and SVM, our goal is to predict animal danger from a rich dataset that was given in Kaggle. We work on a dataset which is a collection of data about animals having 5 types of different symptoms; The symptoms decide whether the animals are dangerous or not. [1]

2. Data Assessment

As you go through the Dataset, they are assured to confront challenges such as class imbalance and the need for feature engineering. Addressing these challenges will be crucial for achieving robust classification models. Thus, this dataset serves as a rich resource for those eager to make a meaningful impact in the field of animal health assessment. The dataset includes 871 records each record represents some important values, although the data is very imbalance between the ration of either the animal is dangerous or not. [2]



AnimalName		symptoms1		symptoms2		symptoms3		symptoms4		symptoms5		Dangerous
aniname		symptoms		symptoms2		symp3		symp4		symp5		danger
Buffaloes	15%	Fever	30%	Diarrhea	14%	Coughing	11%	Weight loss	13%	Pains	11%	true 849 97% false 20 2% [null] 2 0%
Sheep	13%	Fetopelvic dispropo...	2%	Difficulty in breathing	3%	Vomiting	7%	Death	6%	Pain	8%	
Other (632)	73%	Other (593)	68%	Other (726)	83%	Other (717)	82%	Other (703)	81%	Other (704)	81%	
Dog		Fever		Diarrhea		Vomiting		Weight loss		Dehydration		Yes
Dog		Fever		Diarrhea		Coughing		Tiredness		Pains		Yes
Dog		Fever		Diarrhea		Coughing		Vomiting		Anorexia		Yes
Dog		Fever		Difficulty breathing		Coughing		Lethargy		Sneezing		Yes
Dog		Fever		Diarrhea		Coughing		Lethargy		Blue Eye		Yes
Dog		Fever		Respiratory distress		Seizuers		Hyperesthesia		Sudden death		Yes
Dog		Ulcers		Diarrhea		Poor Appetite		Tarry Stool		Enlarged lymph nodes		Yes

Figure 2-1 The data characteristics and examples

For the imbalance data we need to do balance it so we make operations on it with equal (or near equal) relation between the data and there are multiple ways to deal with it, the way we used is the **Over-sampling** technique, with this we make sure to have the higher number of the lower ratio feature value as well as achieving the balance of the data.

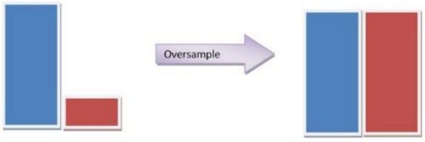
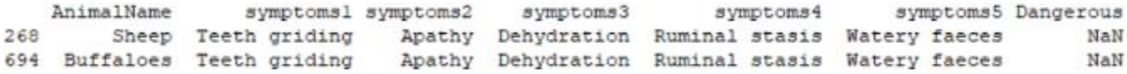


Figure 2-2 Oversampling Illustration

3. Missing Data

Missing data could result from a human factor, a problem in electrical sensors, or other factors. And when this happens, you can lose significant information. Missing Data can be handled



```
    AnimalName      symptoms1 symptoms2    symptoms3       symptoms4      symptoms5 Dangerous
268      Sheep  Teeth griding    Apathy  Dehydration  Ruminal stasis  Watery faeces       NaN
694  Buffaloes  Teeth griding    Apathy  Dehydration  Ruminal stasis  Watery faeces       NaN
```

Figure 3-1 Missing Target Records

through Deletion & Imputation. Searching through the records we found two target features are given without values to be precise 2 records.

Imputation

Dealing with missing data is to fill in the missing value with a substituted value. This method requires replacing the missing value with a specific value. To use it, have domain knowledge of the dataset is required. We have used mice imputation to solve this matter.

The required time (elapse time) for the imputation is: 0.030600786209106445 seconds using MICE Algorithm.

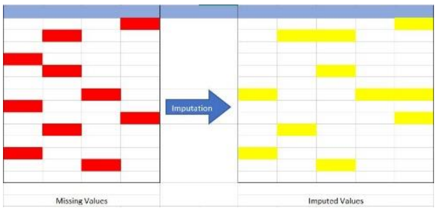


Figure 3-2 Missing Imputation Illustration

4. Feature Importance and Types

Determining the importance of each feature in a dataset when building a predictive model, these scores are calculated using a variety of techniques, such as decision trees, Permutation Importance & Correlation Criteria. To understand the characteristics of the dataset better. The feature that we have used is called decision tree Gini technique and we have used this library from sklearn.ensemble import RandomForestClassifier.

Decision Trees

Decision trees and random forests are models that use trees to make predictions. They provide a measure called feature importance, which shows how much each feature contributes to the accuracy of the model. One type of feature importance is Gini importance, which gauges how a feature reduces impurity in the tree. Gini Index is a metric that assesses the randomness or impurity in a dataset, aiming to decrease impurities from the top (root nodes) to the bottom (leaf nodes) of a decision tree. [3] For this the elapsed time was about: 0.28786420822143555 seconds; and the data were with highest importance were:

	Feature	Importance
274	symptoms3_Diarrhea	0.071861
101	symptoms1_Weakness	0.059620
150	symptoms2_Epistaxis	0.053465
160	symptoms2_Heavy Breathing	0.045763
138	symptoms2_Depression	0.045295

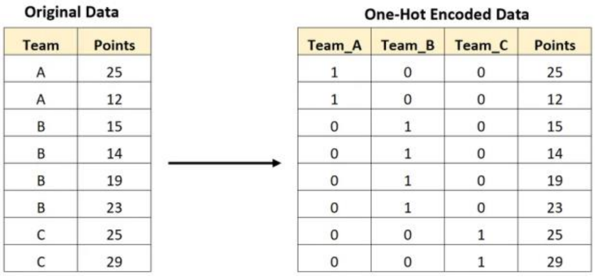
Although they (importance value) may change but these are the most common.

5. Conversion Of Categorical to Numerical

machines cannot interpret the categorical data directly. Therefore, the categorical data must be converted into numerical data for further processing. [4]

One-Hot Encoding: can be applied to the integer representation. [5] This is where the integer encoded variable is removed and a new binary variable is added for each unique integer value; Used for SVM, MLP (neural network), KNN.

For this the elapsed time was: 0.011002063751220703 seconds



Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Figure 5-1 One hot encoding Illustration

Label Encoder: assigns a unique number (starting from 0) to each class of data. [6]

Used in Bayesian Classifier as It's based on probabilities and counts. Therefore, using label encoding, where different categories are represented by different integers, does not introduce misleading ordinal relationships, as the actual values are not considered in a meaningful way.

The label encoder requires less time as its only conversion of the categories to numbers.

6. 4 Algorithms Used

We used different machine learning tools to classify animal conditions. Neural Networks learned intricate patterns, K-Nearest Neighbors compared features with nearby data points, Bayesian Classifiers calculated condition probabilities, and Support Vector Machines separated conditions effectively. This mix of methods gave our system the flexibility to handle various aspects of animal condition assessment.

1. KNN: for the KNN we try to find the best k value(optimal-k) then train the set of data and then do the test and prediction. The accuracy of this algorithm is between 95 to 99, although it has a probability of being a 100 but it is very little because of the randomness of data selection.

The total estimated time for this algorithm is about: 7.353438138961792 seconds

2. SVM: we used the kernel as it’s non-linear data with the sigmoid function which is:

$K(xi,xj)=\tanh(\alpha xiTxj+c)$ this type is usually used for binary classification in certain types of medical which is our case. The accuracy of this algorithm is between 85 to 95 may be higher in some cases depending on the randomness of selection of data.

The total estimated time for this algorithm is about: 0.37119483947753906 seconds

3. Bayesian Classifier: we use Multinomial bayes classifier which is the "Multinomial" part indicates that the features are assumed to be discrete, typically representing counts. The accuracy of this algorithm is between 89 to 97 maybe higher or lower depending on the amount of randomness for the data selection.

The total estimated time for this algorithm is about: 0.27504634857177734 seconds

4. Neural Networks: for the neural network we used 2-hidden layers each with the ReLU activation function and the first hidden layer being 64 neurons and second being 32 neurons the output layer is 1 neuron with activation function of sigmoid. ReLU is used in hidden layers to introduce non-linearity and enable the network to learn complex patterns, Sigmoid is used in the output layer for binary classification problems, as it produces probabilities that can be thresholder to make class predictions. The epochs are equal to 50 and the batch-size is 30. It's success rate is between 97 to 99 maybe it gives answer of 100 but its actually 99 with a very high point that it will be rounded to 100.

And the estimation time is 8.17303466796875 seconds which is higher by far than others.

7. Data Visualization

There can be a lot of graphs and charts to represent the data. Each used in a particular time. Line Chart, Bar Chart & Area Chart are most common, and among all the types we have used pie chart & Confusion Matrix. Finally represented the accuracy in array form. [7]

8. Conclusion

Choosing the best algorithm depends on the data and time, according to our dataset it is obvious that the neural network gives the best outcome but the times that the algorithm alone requires is very high according to other algorithms so that is a disadvantage.

While the other algorithms are not so far behind in terms of accuracy, second best being KNN but also with a high time, coming to the 3rd which is the Bayesian Classifier can be reliable for the dataset in terms of time and accuracy of the data.

So, it may be the best choice for the current evaluation; There are also different types of imbalancing techniques effects the algorithm which is a part of the algorithm indirectly, when using under-sampler techniques, the accuracy increases by each of the algorithms, the Neural Network being still the best after that Bayesian Classifier comes in that case.

References

[1] www.kaggle.com. (n.d.). *Animal Condition*. [online] Available at: https://www.kaggle.com/datasets/willianoliveiragibin/animal-condition [Accessed 1 Jan. 2024].

[2] freeCodeCamp.org. (2022). *How to Handle Missing Data in a Dataset*. [online] Available at: https://www.freecodecamp.org/news/how-to-handle-missing-data-in-a-dataset/.

[3] Aporia. (n.d.). *Feature Importance: 7 Methods and a Quick Tutorial*. [online] Available at: https://www.aporia.com/learn/feature-importance/feature-importance-7-methods-and-a-quick-tutorial/#:~:text=What%20Is%20Feature%20Importance%3F.

[4] GeeksforGeeks. (2021). *How to convert categorical string data into numeric in Python?* [online] Available at: https://geeksforgeeks.org/how-to-convert-categorical-string-data-into-numeric-in-python/ [Accessed 25 Dec. 2023].

[5] Brownlee, J. (2017). *Why One-Hot Encode Data in Machine Learning?* [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/.

[6] chugh, aakarsha (2018). *ML | Label Encoding of datasets in Python*. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/.

[7] Durgapal, A. (2023). *Data Visualization using Python*. [online] Medium. Available at: https://medium.com/@ayushmandurgapal/data-visualization-using-python-1f0b032ff2db#:~:text=There%20can%20be%20many%20useful [Accessed 28 Jan. 2023].