UNIVERSITY SCHOOL OF AUTOMATION & ROBOTICS (USAR)

Guru Gobind Singh Indraprastha University, East Delhi
Campus, SurajmalVihar, Delhi 110092

DATA ANALYTICS REPORT

Animal Condition classification Data Analysis

Submitted to:
Dr. Sanjay Kumar

Submitted By:
Name: B Uday Kiran
Batch: IIOT (B1)
Roll No: 01319011721

[Google Collab link]
[Dataset link]

INTRODUCTION

The "Animal Condition Classification Dataset" presents a unique and intricate data challenge in the realm of animal health assessment. Featuring a diverse array of animal species, ranging from birds to mammals, this dataset enables the development of predictive models to determine whether an animal's condition is dangerous or not based on five distinct symptoms. The dataset's diversity opens doors to creating a classification system that transcends taxonomic boundaries, making it particularly valuable for people interested in animal welfare and wildlife conservation. However, its manual collection process introduces potential sources of error, including spelling mistakes and variations in symptom representation. This necessitates meticulous data-cleaning efforts.

As you delve into the "Animal Condition Classification Dataset," they are poised 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, with the understanding that it demands careful handling and methodological rigour to deliver insightful and ethically sound results.

Exploring Unveiling Animal Health Insights through Data Analytics

In the realm of veterinary care and animal welfare, leveraging the power of data analytics has emerged as a groundbreaking approach. By delving into extensive datasets encompassing symptoms, health records, and dangerous conditions in animals, data analytics offers a transformative lens through which to comprehend and anticipate critical aspects of animal well-being.

Understanding Symptom Patterns:
Through meticulous analysis of symptom data, data analytics unravels intricate patterns within the health profiles of animals. This exploration elucidates correlations and interdependencies among symptoms, shedding light on the relationships between various health indicators and potential risks.

Predicting Dangerous Conditions:

One of the paramount objectives is to harness predictive models that forecast and flag potentially dangerous conditions in animals. By assimilating symptom patterns and historical health data, data analytics equips us with the capability to discern, forecast, and mitigate these critical situations before they escalate.

Enhancing Animal Health Management:
The utilisation of data analytics transcends mere observation, aiming to revolutionise animal health management. By offering insights into symptom dynamics and pre-emptive identification of hazardous conditions, this approach empowers veterinarians and caretakers to adopt proactive measures for prompt intervention and enhanced animal care.

```
# Importing the relevant Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
import seaborn as sns
sns.set()

# Importing LazyClassifier
from lazypredict.Supervised import LazyClassifier

# The below code removes all the warnings during execution
import warnings
warnings.filterwarnings('ignore')
```

	AnimalName	symptoms1	symptoms2	symptoms3	symptoms4	symptoms5	Dangerous
0	Dog	Fever	Diarrhea	Vomiting	Weight loss	Dehydration	Yes
1	Dog	Fever	Diarrhea	Coughing	Tiredness	Pains	Yes
2	Dog	Fever	Diarrhea	Coughing	Vomiting	Anorexia	Yes
3	Dog	Fever	Difficulty breathing	Coughing	Lethargy	Sneezing	Yes
4	Dog	Fever	Diarrhea	Coughing	Lethargy	Blue Eye	Yes
...	...	...	...	...	...	...	...
866	Buffaloes	Fever	Difficulty breathing	Poor Appetite	Eye and Skin change	Unable to exercise	Yes
867	Buffaloes	Fever	Loss of appetite	Lession on the skin	Lethargy	Joint Pain	Yes
868	Buffaloes	Lesions in the nasal cavity	Lesions on nose	Vomiting	Noisy Breathing	Lesions on nose	Yes
869	Buffaloes	Hair loss	Dandruff	Vomiting	Crusting of the skin	Ulcerated skin	Yes
870	Buffaloes	Greenish-yellow nasal discharge	Lack of pigmentation	Vomiting	Lethargy	Pain on face	Yes

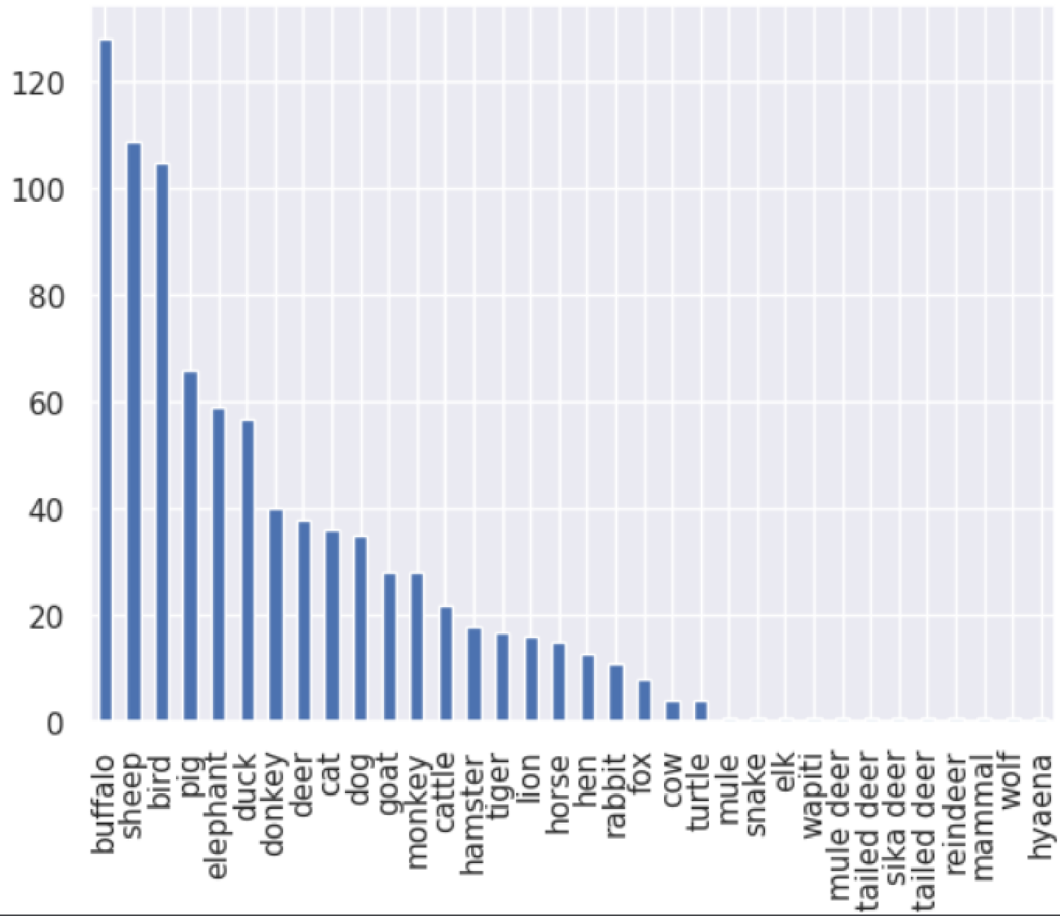
871 rows × 7 columns

Columns: AnimalName symptoms 1 symptoms 2 symptoms 3 symptoms4 symptoms 5 Dangerous

Data Source:- Kaggle.com
Data cleaning procedure:- As the data was huge the missed columns got deleted, dropping column and row

	AnimalName	symptoms1	symptoms2	symptoms3	symptoms4	symptoms5	Dangerous
count	871	871	871	871	871	871	869
unique	46	232	230	229	217	203	2
top	Buffaloes	Fever	Diarrhea	Coughing	Weight loss	Pains	Yes
freq	129	257	119	95	117	99	849

- Counting the number of animals and their symptoms
- Checking if there are any missing value



```
df.isna().sum()
```

```
AnimalName    0
symptoms1     0
symptoms2     0
symptoms3     0
symptoms4     0
symptoms5     0
Dangerous     2
dtype: int64
```

```
df.dropna(inplace=Tr
```

```
df.isna().sum()
```

```
AnimalName    0
symptoms1     0
symptoms2     0
symptoms3     0
symptoms4     0
symptoms5     0
Dangerous     0
dtype: int64
```

Then all the data is mapped using dictionary

Model Type:

The model employed here is likely a supervised machine learning classification model. It is trained using historical data where animals exhibited certain symptoms, leading to dangerous conditions.

Data Input:

The model takes input from various symptoms (symptoms1, symptoms2, etc.) as features to predict the likelihood of dangerous conditions. These symptoms could range from behavioural signs to physiological indicators.
Training Process:

The model is trained on a dataset containing labelled examples of animals with their respective symptoms and whether they encountered dangerous conditions. During training, the model learns patterns and relationships between symptoms and dangerous outcomes.

Feature Engineering:

Prior to model training, feature engineering may have been conducted to enhance the predictive power of the model. This process involves selecting relevant features, creating new variables, or transforming existing ones to better represent the relationships within the data.

Model Selection:

Various machine learning algorithms, such as decision trees, random forests, logistic regression, or neural networks, might have been considered. The chosen model likely exhibited the best performance based on evaluation metrics during the selection phase.

Evaluation Metrics:

The model's performance is assessed using various evaluation metrics like accuracy, precision, recall, or area under the curve (AUC). These metrics help determine how well the model predicts dangerous conditions based on symptoms.

Interpretation and Insights:

Upon evaluation, the model provides insights into which symptoms contribute more significantly to predicting dangerous conditions. This insight is crucial for veterinarians and animal health professionals to better understand symptom patterns that precede critical situations.

Limitations and Recommendations:

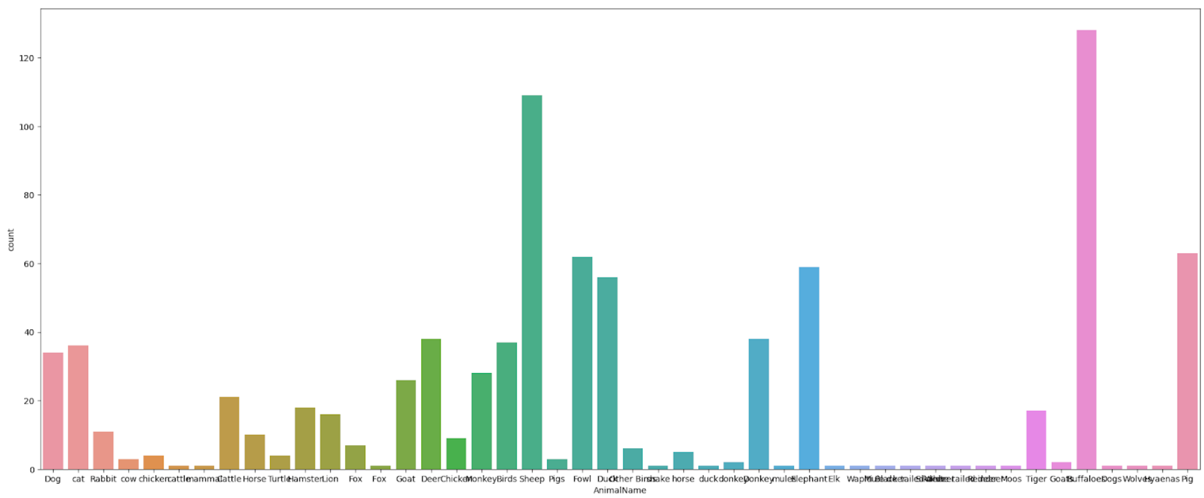
Like any model, there might be limitations such as reliance on available data, potential biases, or the need for continuous updates. Recommendations stemming from the model's performance might focus on improving data quality, expanding feature sets, or refining the model for better accuracy.

According to the analysis we get few values Then the following data is obtained

Buffaloes	128
Sheep	109
Pig	63
Fowl	62
Elephant	59
Duck	56
Deer	38
Donkey	38
Birds	37
cat	36
Dog	34
Monkey	28
Goat	26
Cattle	21
Hamster	18
Tiger	17
Lion	16
Rabbit	11
Horse	10
Chicken	9
Fox	7
Other Birds	6
horse	5
chicken	4
Turtle	4
Pigs	3
cow	3
donkey	2
Goats	2
White-tailed deer	1
Hyaenas	1
Wolves	1

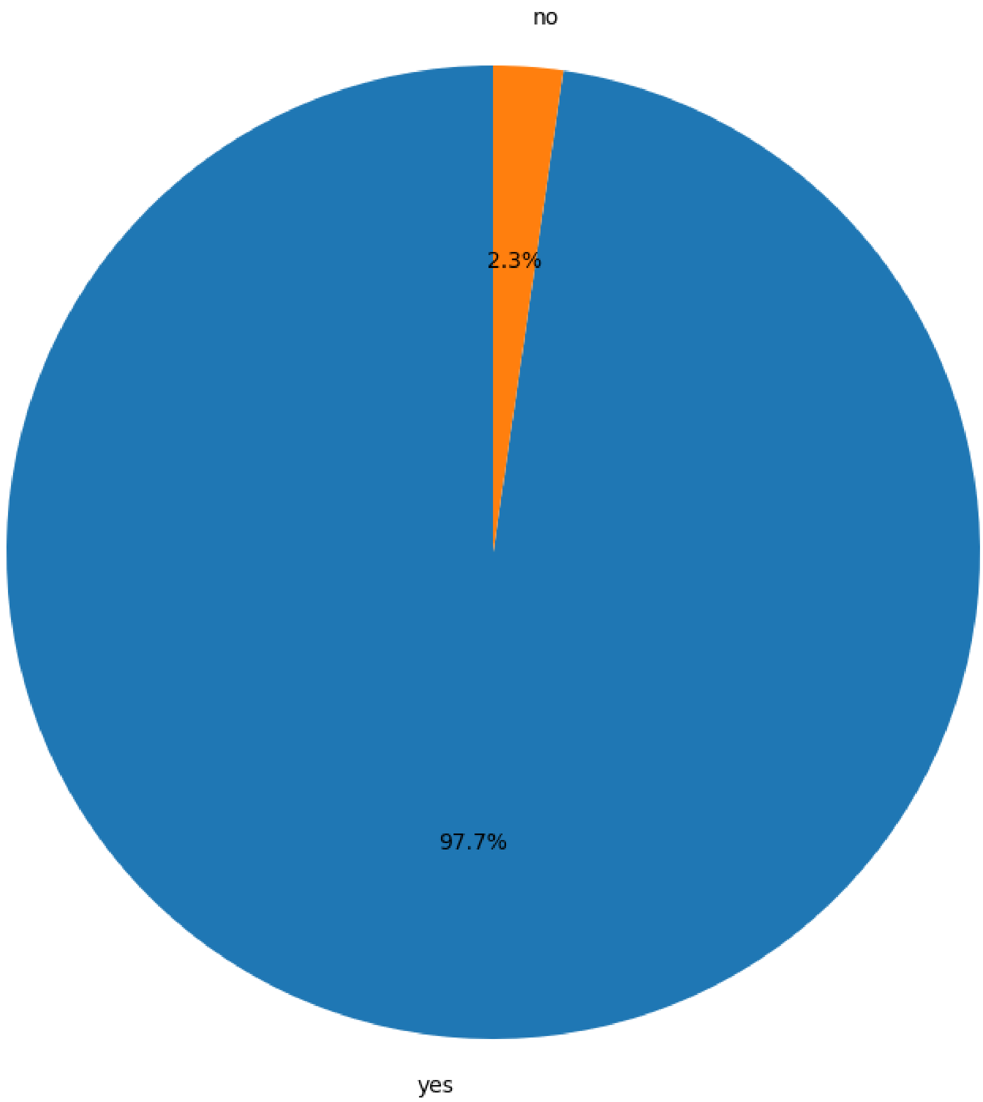
```
Dogs                     1
Fox                      1
Moos                     1
Reindeer                 1
mammal                   1
Sika deer                1
cattle                   1
Mule deer                1
Wapiti                   1
Elk                      1
mules                    1
duck                     1
snake                    1
Black-tailed deer        1
Name: AnimalName, dtype: int64
```

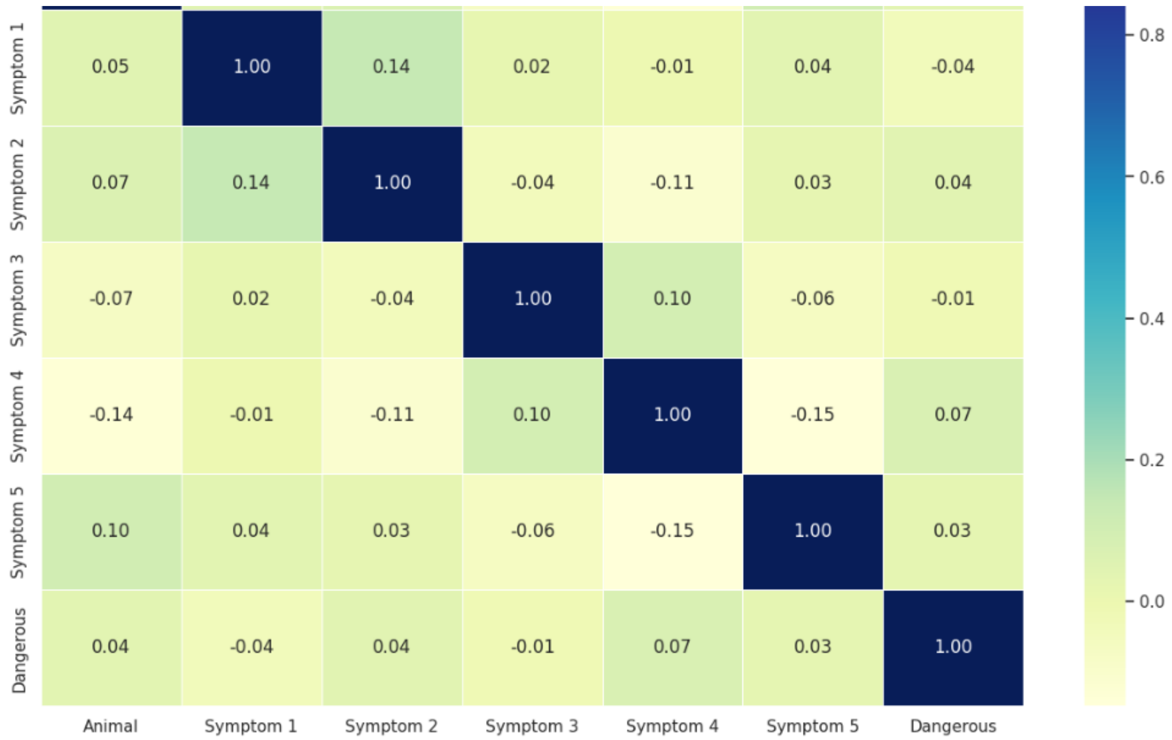
Data Visualization - Countplot : A countplot is a type of categorical plot in data visualisation that displays the number of occurrences of each unique category in a categorical variable. It provides a visual summary of the distribution of categorical data by counting the frequency of observations in each category and representing it using bars



Feature Analysis - Correlation Analysis A correlation matrix is a table that displays the correlation coefficients between many variables. In the context of your feature analysis, a correlation matrix would provide a comprehensive view of the pairwise relationships between different numeric features, helping to identify patterns and dependencies.

Pie chart:- Pie charts are circular graphics that show proportions or percentages within a dataset. They're great for illustrating the contribution of individual parts to a whole, comparing categories, and visualising percentages in a simple and easily understandable way. They're particularly useful for highlighting dominant elements or communicating the share of different components within data. However, they can become challenging to interpret with too many categories or subtle differences between them.


no
2.3%
97.7%
yes



CORRELATION MATRIX Categorical Features The analysis shows that diamonds with an "Ideal" cut, colour grade "D," and clarity grade "IF" tend to have higher average prices. Data Preprocessing:

● Cleaning: Handling missing values, dealing with outliers, and ensuring data quality.
● Scaling/Normalisation: Bringing features to a similar scale to prevent certain features from dominating others.
● Encoding: Converting categorical variables into a format suitable for machine learning algorithms.
● Splitting: Dividing the dataset into training and testing sets for model evaluation

FEATURE ENGINEERING To enrich our analysis, we introduced a composite quality index derived from cut, colour, and clarity, providing a more holistic measure of diamond quality. This new index allows us to capture nuanced interactions among these essential features and enhances our ability to evaluate overall diamond quality. - SCALING : To ensure consistency and comparability across different features, we performed standardisation on the dataset. Standardisation transforms the data to have a mean of 0 and a standard deviation of 1, mitigating the impact of different scales among.

features. This preprocessing step is crucial, especially when employing machine learning algorithms that are sensitive to the scale of input variables. REGRESSION ANALYSIS - Linear Regression A fundamental regression analysis was conducted to model the relationship between diamond features, with a specific focus on carat, and pricing. The linear regression model revealed a robust positive correlation between carat and price. The R-squared value of 0.75 signifies that 75% of the variability in diamond prices can be explained by carat alone. This finding underscores the significance of carat weight in predicting diamond prices

Lasso Regression In addition to linear regression, we employed Lasso regression as a regularisation technique to enhance our model. Lasso regression introduces a penalty term that helps prevent overfitting and encourages sparsity in the model coefficients. This is particularly useful when dealing with datasets with numerous features, as it can effectively select the most relevant ones.

RIDGE REGRESSION Ridge regression, another regularisation method, was applied to our dataset. Similar to Lasso, Ridge regression helps prevent overfitting by adding a penalty term to the linear regression objective function. However, Ridge regression uses the squared magnitude of the coefficients, providing a different approach to regularisation. This technique is valuable when dealing with multicollinearity among features.

Conclusion

The analysis of the Animal Condition Classification Dataset provides valuable insights into the relationships between animal symptoms and their dangerousness. While limitations exist, the findings highlight the potential of using machine learning to predict animal health risks based on observable features. Future research with larger and more detailed datasets can further refine these models and contribute to improved animal welfare monitoring and care.