Classification Write-up

Ridley Arp

Metis Classifications: Jan-Feb ‘22

**Abstract**:

This project was designed to assist in giving early warning of the likelihood of multiple tumours existing in any given presentation in a canine or feline. It also was designed to try and highlight traits or aspects of the presentation which might indicate a heightened risk of multiple tumours in that case.

**Design:**

This project arose from a lack of available information on cancers in pets and provides a prediction of whether the attending vet should expect to see only a singular tumourous lesion, or if there are likely to be more, indicated by a target of Multiple Instances. This also helps to highlight potential risk factors that might lead to more numerous or more aggressive tumourous growths.

**Data:**

The dataset contains 109895 prior cases, with 10 features per case, all of which are categorical. Feature examples include Species, Breed, Gender, Tumour type, and Neuter status. Extensive feature engineering helped to segment Breeds and Cancer type into manageable bins to build a consistent signal, resulting in a total of 15 features used to create the final model

**Algorithms:**

*Feature Engineering*

1.Seperated Gender from neuter status to provide separate signals for each

2.Segmented breeds according to historic stressors and training of each given breed. This determined by AKC classification(Sporting, Toy, Working, Herding, Non-Sporting)

3.Created independent features for each of the six major types of cancer, with rarer types being included in their own feature.

4. isolated tumour location by overall body area(Head, Pelvic, Trunk Internal, Trunk External, Extremities)

*Models*

1.Random forest, Decision Tree, Logistic Regression, and AdaBoost were all tested, with Decision Tree providing the best Negative accuracy and Logistic Regression providing the most accurate Positive accuracy. A final Voting Classifier was created from a weighted ensemble of the Decision Tree and Logistic Regression.

*Model Evaluation/Selection*

Data was separated into a 60/20/20 Train/Test/Hold split with accuracy and confusion matrixes informing the performance of each model. Ensembling of models based on individual strengths with hyperparameter tuning provided the final model accuracy of .76

**Communication:**

Presentation slides containing metrics and additional features which show increased chance of higher risk cancers.