Predicting Risk in Shelter Animals - Project Proposal

Kfir Diamond (209080134) and Daniel Tugendhaft (318465291)

1. Dataset Selection and Brief Description: The selected dataset is "Louisville Metro KY - Animal Service Intake and Outcome". It contains a list of animal cases admitted to Louisville County Animal Services. The data includes features such as animal type, sex, bites, pet size and breed. The dataset contains 64,000+ records and 15+ relevant features.

2. Motivation and Core Problem: The core problem we will address is predicting negative risk for animals admitted to the shelter. We define an animal as "At-Risk" if its final outcome type is Euthanasia, Died, or Disposal. This classification problem has high practical significance, as early predictions can help allocate resources to prevent negative outcomes.

3A. Supervised Learning – Classification:

1. **Logistic Regression:** A baseline model for binary classification, using the logistic function.
2. **Random Forest:** A fundamental supervised learning model for classification.
3. **XGBoost:** An ensemble learning method that combines several "weak classifiers," giving extra weight to points misclassified in previous rounds.
4. **MLP:** Building a basic neural network that will use appropriate activation functions, such as Softmax or Sigmoid.
5. **SVM:** A powerful classification model that finds the best possible way to identify the different categories in the data.

3B. Unsupervised Analysis: We will perform Clustering on the dataset to identify natural groupings of animals with similar characteristics. We plan to use K-Means and Hierarchical Clustering.

4A. Classification Experiments: We will train and test all models, comparing their performance after applying various feature engineering techniques (such as comparing performance with and without PCA).

4B. Clustering Experiments: We will compare the results of K-Means with Hierarchical Clustering. We will also compute the Silhouette Coefficient to evaluate cluster quality.

4C. Evaluation Metrics:

- For **Classification**, we will use metrics such as Accuracy, F1 Score, Precision, and Recall, which will help assess the model's ability to accurately identify "At-Risk" cases.
- For **Unsupervised Analysis**, we will use metrics such as Cohesion and Separation, in addition to the Silhouette Coefficient.