Pranav Krishnan COE 379 Project 1 Report (part 3)

# 1. Data Preparation

The raw dataset contained 131,165 rows and 12 columns, which described various characteristics of animals in Austin shelters.

Initially, I noticed all the columns were stored as object types, so I had to do quite a few preprocessing steps to format the data properly:

1. **Date Conversion:**
    - Converted Date of Birth, DateTime, and MonthYear to datetime64 to allow for chronological analysis based on the actual date.

2. **Duplicate Removal:**
    - Found 17 duplicate rows using animals.duplicated().sum() and removed them with drop_duplicates().

3. **Missing Values:**
    - Replaced NaN values in *Name* and *Outcome Subtype* with "Unknown".
    - Filled Outcome Type missing entries using the mode, which was (‘Adoption’).

4. **Age Normalization:**
    - Extracted numeric values and units from *Age upon Outcome*, converting all ages into days to have consistent numeric comparisons

5. **Irrelevant Columns:**
    - Dropped *Animal ID, Name,* and *MonthYear* as they added nothing to help me predict.

6. **Categorical Conversion & Encoding:**
    - Converted Outcome *Type, Outcome Subtype, Animal Type, Sex upon Outcome*, *Breed,* and *Color* all to categorical types from objects, then applied one-hot encoding to create numeric dummy variables for machine learning in part 2.

After reformatting and cleaning, the dataset had 131,148 rows and 3,150 columns (~400 MB). Finally, per instructions, I dropped the Breed column before modeling.

## 2. Exploratory Insights

- **Animal Type:** Dogs (~68 K) and Cats (~63 K) dominate the dataset while birds and livestock are much rarer.

- **Outcome Type:** Most records are **Adoptions** (~85 K), with Transfers (~47 K) having the second most.

- **Sex upon Outcome:** Neutered Male and Spayed Female animals form the majority, and intact animals are uncommon.

- **Age Distribution:** Most animals are under 2 years old, mainly “2 months”, “1 year”, and “2 years”.

These distributions show the dataset is slightly skewed toward adopted dogs and cats that are already neutered/spayed with the rest being much more uncommon.

## 3. Model Training Procedure

To predict **Outcome Type (Adoption or Transfer)**:

1. **Train/Test Split:** Used *train_test_split* with 30% test data and *stratify*=*y* to preserve class ratios.

2. **Feature Selection:** Included *Age upon Outcome (days)* and a subset of dummy variables *(Animal Type_, Sex upon Outcome_)* to avoid memory overload where this wasn't done.

3. **Models Trained:**

    - **Baseline KNN (k = 3)**

    - **Optimized KNN using GridSearchCV** (*k ∈ [1 … 29], 3-fold CV*)

    - **Linear Classification (Logistic Regression)**

All models used *accuracy, precision, recall,* and *f1-score* as metrics, computed with *classification_report().*

# 4. Model Performance

| Model | Accuracy | Precision (macro) | Recall (macro) | F1 (macro) |
|---|---|---|---|---|
| KNN (k = 3) | 0.8167 | 0.8025 | 0.7986 | 0.8005 |
| KNN (best k = 28) | 0.8546 | 0.8614 | 0.8205 | 0.8343 |
| Logistic Regression | 0.8603 | 0.8730 | 0.8236 | 0.8395 |

**Observations:**

- Baseline KNN performed well (≈ 81% accuracy).

- GridSearchCV improved KNN to ≈ 85%, confirming that optimal neighbor selection (k = 28) does help.

- **Logistic Regression achieved the best results**, with ~86% accuracy and F1 ≈ 0.84, slightly outperforming the KNN.

# 5. Model Confidence & Interpretation

- The F1-score was the most important metric because it balances precision and recall, and we want to minimize false classifications.
- Logistic Regression's stable performance across classes suggests that it is a good generalization.
- The high accuracy (~86%) and strong F1 values make it so that we know the model is pretty reliable and confident in predicting animal outcomes based on features like age, type, and sex status.