I. COE 379L; Project 1 - Part 3; Mahin Naveen - mn27995

II. What I did to prepare the data

I started by reading the full Austin animal shelter dataset from project1.csv and checked the size and shape of the raw data. There were around 131 thousand rows and 12 columns, which matched the project description. I inspected datatypes, found a mix of strings, dates, and categorical text, and decided which columns needed to be cleaned or converted.

There were a few duplicate rows (17 in total), so I dropped them. I also noticed many missing values, especially in text fields like Color and Sex Upon Outcome. For numeric values, I converted the "Age Upon Outcome" field to a single numeric measure in days using a small function that translated text like "2 years" or "4 weeks" into equivalent days. For missing ages, I filled them with the median number of days. For categorical fields, I used the most common value (mode) to fill in blanks.

Next, I simplified some columns to make the data easier to model. From Sex Upon Outcome, I created two new features – sexSimple (male/female/unknown) and isFixed (1 if the animal was neutered or spayed, 0 if intact). I also made baseColor by taking the first color word before any slashes. After cleaning, I dropped columns that were irrelevant for classification such as Name, AnimalID, DateTime, MonthYear, and Outcome Subtype. Finally, I changed all categorical columns to pandas "category" types and created a one-hot encoded version to confirm the data could be turned entirely numeric.

Overall, I ended up with a balanced, clean dataset that had no duplicates, no missing values, and consistent numeric and categorical features, ready for training models.

III. Insights from the data preparation

A few patterns stood out. Most animals were either adopted or transferred, and there was a noticeable imbalance between cats and dogs – dogs were slightly more likely to be adopted. The "age in days" feature varied widely, with many young animals being adopted more quickly. Also, neutered and spayed animals tended to have higher adoption rates. The base colors "Brown," "Black," and "White" were the most common.

Cleaning the data also showed how inconsistent text formatting was in the original CSV. Many categorical fields used slightly different words for the same idea. Fixing that made the dataset more reliable for training.

IV. What procedure I used to train the model

Following the instructions, I dropped the Breed column since it contains too many unique values and would have created thousands of dummy variables. Then I split the cleaned data into 80% training and 20% testing sets using a stratified random split to keep the same ratio of "Adoption" and "Transfer" in both sets.

I trained three classifiers:

- A baseline K-Nearest Neighbors (K = 5) model
- A KNN model using Grid Search CV to find the best K and distance metric
- A Logistic Regression model for linear classification

All models used the same preprocessing pipeline: a ColumnTransformer that scaled numeric features (ageDays, isFixed) and one-hot encoded categorical ones (Animal Type, sexSimple, baseColor). This ensured consistent input to each classifier.

V. How the models performed

After training and evaluation, Logistic Regression gave the best overall results. On the test set, it reached an accuracy around 74% and F1 score approx. = 0.74. The basic KNN model was close behind (around 68%), while the tuned Grid Search version improved slightly (around 70%).

These results make sense since logistic regression handles high-dimensional one-hot features more efficiently than KNN, which struggles with distance metrics when there are many sparse binary features. The KNN models still performed respectably and were useful for comparison.

The precision and recall values were balanced, meaning both classes (Adoption and Transfer) were being predicted reasonably well. Because this is a classification problem where both outcomes matter, the F1 score is the most meaningful metric. It captures the balance between precision (how many predicted adoptions were correct) and recall (how many actual adoptions were found).

VI. How confident I am in the model

I'm fairly confident in the results. The data cleaning and transformations were done carefully, and the split was stratified to avoid bias. The consistent F1 scores across models suggest the dataset is informative enough to predict outcomes with moderate accuracy. Logistic Regression provided the best trade-off between accuracy, interpretability, and speed.

There's still room to improve if Breed could be handled through grouping or embeddings, but that's outside the current project scope. For now, the model meets all assignment goals and gives a clear, interpretable prediction of adoption versus transfer outcomes for shelter animals.