Preparation of Dataset:

To prepare the dataset, I first identified and removed 17 duplicate rows. For missing values in the Name column, I filled them with the most common name, “Max,” since this feature is not critical for training. In the Outcome Type column, 40 rows had missing values; given the dataset size of over 131,000 entries, I chose to drop these rows as their proportion was insignificant. For the Outcome Subtype column, 65,346 values were missing, all corresponding to adoption outcomes. I imputed these with “Foster,”since this is the most common type of outcome subtype for the adoption outcomes. Next, I converted the Age upon Outcome column into a numerical format, expressing all values in days for consistency. For categorical variables such as Animal Type, Sex upon Outcome, and Color, I transformed them into suitable forms, including grouping the many distinct colors into three broader categories: “Light,” “Dark,” and “Other.” Finally, I applied one-hot encoding to all categorical features so they could be effectively used as model inputs. Finally, for training the classification model, I used the following features: Age_years, Animal Type, Sex upon Outcome, and ColorGroup, while “Outcome Type” was treated as the target variable. These features were selected because they are directly related to an animal’s likelihood of adoption or transfer. Age is important since younger animals are generally adopted more quickly. Animal Type (dog, cat, etc.) captures species-specific adoption patterns. Sex upon Outcome reflects whether the animal is spayed, neutered, or intact, which can influence adoptability. Finally, ColorGroup was included because visual traits such as coat color may affect how appealing animals are to potential adopters. Together, these features provide both biological and perceptual factors that are useful for predicting the outcome.

Insights get from the dataset:

From the data preparation and exploratory analysis, several key insights emerged. The dataset is dominated by young animals, with the majority under two years old and very few older than ten, indicating that most outcomes occur early in an animal’s life. Cats and dogs make up the overwhelming majority of cases, while birds and livestock appear only rarely, suggesting that the model will generalize better for the common species. In terms of outcomes, adoption is more frequent than transfer, but transfer still represents a significant portion, making this a moderately imbalanced binary classification problem. Additionally, most animals have outcomes are neutered males or spayed females, with fewer intact animals and a notable “unknown” category that could add noise but still provides useful information. And the color categories showed that darker colors were more common among the animals, while light colors were less frequent.

Training the model:

To train the model, I first split the dataset into training and testing sets using a 70/30 ratio while keeping the class distribution consistent through stratification. I then trained three classification models — a K-Nearest Neighbors (KNN) classifier(testing with the k = 10), a KNN with Grid Search Cross-Validation to find the optimal number of neighbors, and a Linear Classification (SGDClassifer)

model. Each model was trained on the processed dataset that included one-hot encoded categorical features and a numerical age feature. After training, I evaluated the models on the test set using accuracy, precision, recall, and F1-score to compare their performance and understand how well each model predicted animal outcomes.

How does the model perform to predict the class&How confident are you in the model?:

For this classification problem, True recall is the most important metric. (True is transfer and false is adoption). In the context of predicting animal outcomes (such as whether an animal will be adopted or transferred), recall measures how well the model correctly identifies all actual positive cases—in this case, all transfers. A high recall means the shelter can accurately identify most animals that are likely to be transferred and take action to improve their adoption chances. While precision and accuracy are also valuable, missing potential transfer cases (low recall) could mean fewer opportunities to increase adoption rates, which is more critical in a real-world shelter setting. Therefore, maximizing recall provides more meaningful insight and better supports decision-making for animal welfare management.

I am fairly confident in the performance of all three models, as their accuracies are consistently around 0.85–0.86, indicating strong generalization. The KNN model (k=10) achieved balanced performance but showed slightly lower recall for the “True” class (Transfer), meaning it sometimes failed to identify transfers correctly. The Grid Search CV model, which tested 1-100 k-values and selected k=42 as the optimal parameter, achieved a slightly better overall performance compared to other two methods with accuracy 0.86 and higher F1-score, suggesting that tuning helped improve stability and reduce overfitting. The Linear Classification model produced similar accuracy , showing that a linear decision boundary works reasonably well for this dataset.

For all three methods, they all achieve high F1-scores for the “False”(Adpotion) class across all experiments. However, the models struggle more with predicting transfers, as indicated by the relatively lower recall for the “True” class. This means some transfers are misclassified as adoptions. Based on the report, for all the models, the recall for the Transfer class is noticeably lower than that of Adoption, likely due to class imbalance and overlapping feature patterns. Since adoptions make up a larger portion of the dataset, the model tends to favor predicting the majority class to minimize overall error, resulting in more missed transfer cases. Additionally, features such as color, age, and sex may not strongly distinguish between adoption and transfer outcomes, making it harder for the model to correctly identify transfers. As a result, the model achieves high recall for adoption but struggles to capture all true transfer cases, leading to a lower recall value.

Use of AI

[1]. Tool: ChatGPT
Prompt: Change the Age upon outcome which is the object datatype, in the form of like 2 weeks 6 months, 4 days and 4 years to float
Output:

```
unit_to_days = {"day":1, "week":7, "month":30, "year":365}

def parse_age(s):
    if pd.isna(s):
        return np.nan
    s = str(s).lower().strip()
    match = re.match(r"(\d+)\s*(day|week|month|year)s?", s)
    if match:
        value, unit = int(match.group(1)), match.group(2)
        return value * unit_to_days[unit]
    return np.nan
```