Ryane Li
COE 379L

COE 379L Project 1 Part 3 Report:

What did you do to prepare the data?
To prepare the data, I performed a comprehensive cleaning and transformation process to make the raw dataset of over 130,000 animal records suitable for machine learning. First, I addressed data quality issues. This involved removing **17 duplicate rows** and **9,907 duplicate Animal IDs** to ensure each record was unique. I also dropped records with missing Outcome Type values, as this was our target variable for prediction. The initial dataset had all columns stored as text, so I converted them to appropriate data types like **numeric, categorical, and datetime formats**. A key step was feature engineering. The original Age column contained text strings like "2 weeks" or "3 years". I wrote a custom function to convert these into a single, numeric unit: age in days. This created a new Ageindays feature that the models could easily interpret. Finally, to handle categorical data (like Animal Type and Sex upon Outcome), I used one-hot encoding. This process converts text categories into a binary format (0s and 1s) across multiple new columns, creating a numerical representation of the data. Initially, this expanded the dataset to over 3,000 features, but I strategically removed the 2,527 features related to Breed to simplify the model and prevent overfitting. The final, clean dataset contained **121,258 records** and **620 features.**

What insights did you get from your data preparation?
The data preparation phase revealed several important insights about the dataset and the animal shelter's operations.

The most significant insight was the distribution of outcomes. I found that **62.3%** of the animals were **adopted**, while **37.7%** were **transferred** to partner organizations. This slight class imbalance is important to account for during model training, which I handled using stratified sampling to ensure the training and testing sets reflected this real-world distribution.

The data also showed that the shelter handles a diverse population of animals, primarily cats and dogs, across a very wide age range. The Ageindays feature I created confirmed that age is a critical factor, spanning from just a few days old to many years. This reinforced the importance of including age as a key predictor in the model. Finally, the sheer volume of high-quality data (over 121,000 clean records) confirmed that the dataset was robust enough to train a reliable and powerful predictive model.

What procedure did you use to train the model?
I used a structured and comparative procedure to train and evaluate three different classification models.

First, I split the prepared dataset into a training set and a testing set using an **80/20 ratio**. I used a stratified split to ensure that the proportion of adoptions and transfers (our two classes) was the same in both the training **(97,006 records)** and testing **(24,252 records)** sets. This prevents the model from being biased towards the majority class.

I then trained three different models to compare their performance:

Ryane Li
COE 379L

Basic K-Nearest Neighbors (KNN): A straightforward implementation with k=5 to establish a baseline performance.
Optimized K-Nearest Neighbors (KNN): I used GridSearchCV to systematically test different hyperparameters (like the number of neighbors and the distance metric) to find the best-performing KNN model. The best parameters found were k=3 with a Manhattan distance metric.
Logistic Regression: A linear model that is well-suited for binary classification tasks like this one. By training multiple models, I could confidently identify the one that was not only the most accurate but also the best fit for the specific goals of this project.

How does the model perform to predict the class?
The model performance was excellent, with **Logistic Regression** emerging as the clear winner. It achieved near-perfect accuracy in predicting whether an animal would be adopted or transferred.

Summary Metrics:

Model	Accuracy	Precision	Recall	F1-Score
KNN (basic)	0.9677	0.9609	0.9883	0.9744
KNN (grid search)	0.9713	0.9657	0.9891	0.9773
Logistic Regression	0.9995	0.9992	1.00	0.9996

The recall of 100% means that the model successfully identified every single animal that was ultimately adopted, ensuring no adoption opportunities were missed **(zero false negatives)**.
The precision value of 99.92% means that when the model predicted an adoption, it was correct 99.92% of the time, resulting in very **few false positives**.

How confident are you in the model?
I am confident in the Logistic Regression model's ability to reliably predict animal shelter outcomes. This is due to the following key factors. An accuracy of 99.95% is important, but the 100% recall is what matters most. For an animal shelter, the cost of a false negative (failing to identify an adoptable animal) is very high. Since this model had the highest recall score, it would have the lowest risk of raising a false negative. The model converged quickly, taking only **389 iterations**, which means that the optimization process was relatively stable and did not struggle to find a solution. This suggests that the model has learned the patterns in the data effectively without overfitting. While not as perfect, the strong performance of the two KNN models (including one optimized with 3-fold cross-validation) serves as further validation. The fact that different algorithms could also achieve high accuracy (>97%) on the same data confirms that the underlying patterns are strong and predictable.
In summary, the combination of near-perfect predictive accuracy, a 100% recall score, and a stable training process makes this model a trustworthy and valuable tool. I am confident it can be deployed to help a shelter optimize its resources and, most importantly, maximize positive outcomes for its animals.