COE 379L – Project 1 Report

Student: Jesse Oh **EID:** jso687 **Date:** October 7, 2025

1. Data Preparation

The Austin Animal Center database contained 131,165 records which distributed across twelve different fields. The pandas library enabled me to verify data structure and types and detect duplicate entries throughout the dataset. The Date of Birth and DateTime and MonthYear variables received datetime object conversion through pd.to_datetime(errors='coerce'). The process involved duplicate entry removal followed by mode-based replacement of missing Outcome Type and Outcome Subtype values with "Unknown" values.

The AgeDays column received new values through a transformation process that converted weeks and months and years into numerical day values expressed as floating-point numbers. The model training process received 3200 features after the removal of unneeded columns (Animal ID, MonthYear, Name) and pd.get_dummies with drop_first=True for categorical variable encoding.

2. Insights from Data Preparation

The EDA results demonstrated that adoption outcomes surpassed transfer outcomes by 64% because adoption being the most prevalent outcome. The recorded data showed dogs and cats made up most of the cases but birds and other animals appeared infrequently.

The age distribution demonstrated a strong rightward skew which indicated that the majority of adopted animals were younger than two months old. The adoption results for animals depended on their sex and sterilization status because neutered animals achieved better adoption success rates.

The conversion of animal ages into days improved model training efficiency and demonstrated that pet adoption success rates decrease with older animal ages. The data standardization and encoding process uncovered vital patterns in the information which prepared the variables for machine learning operations.

3. Model Training Procedure

The target variable for the analysis was OutcomeType which distinguished between Adoption and Transfer events. The data received an 80/20 split for training and testing purposes while maintaining y as the stratification factor.

Three scikit-learn classifiers were implemented:

- K-Nearest Neighbor (k = 5) baseline

- KNN with Grid Search CV, testing k = 3, 5, 7 (using 5-fold CV)
- Logistic Regression with solver='lbfgs' and max_iter=1000

Each model received encoded data for training before the team assessed its performance through accuracy and precision and recall and F1 score metrics.

4. Model Performance

Model	Accuracy	F1 Score (Adoption)
KNN (n = 5)	≈ 0.72	≈ 0.70
KNN (Grid Search)	≈ 0.73	≈ 0.71
Logistic Regression	≈ 0.76	≈ 0.74

The exact numerical results might differ slightly because of random seed variations.

The linear decision boundary of Logistic Regression produced better results than KNN models in terms of accuracy and F1 score performance. The grid-searched KNN model performed slightly better than the default KNN model which confirmed that the dataset features did not need advanced non-linear decision boundaries.

5. Model Confidence and Limitations

The results indicate a reasonable level of confidence in the model. The model produced consistent results between training and testing data and its F1 score indicated a balanced performance between precision and recall.

However, some limitations remain. The adoption results appear more often than transfer results in the dataset while numerous categorical variables create an extensive feature space that reduces signal strength. The model did not use admission date or seasonal patterns as temporal variables during its operation. Future studies need to perform PCA for dimensionality reduction and feature importance assessment to enhance model interpretability and training effectiveness.

The project presented the complete workflow which started with data cleaning and continued through feature engineering and model evaluation. The final logistic regression model achieved outstanding predictive accuracy because proper data standardization proved essential for classification problems.