Pet Adoption Outcome Prediction Using Machine Learning

Author: Ivy Sun

Course: CSCI 39542 - Data Science

1. Introduction

This project investigates how pet characteristics influence adoption outcomes using real-world data from **PetFinder.my**, Malaysia's largest pet adoption platform.
Each record describes a unique pet and includes attributes such as:

- Species
- Breed
- Age
- Color
- Gender
- Health
- Number of photos

Along with the adoption speed (0–4), where lower values indicate faster adoption.

The goal is to predict how quickly a pet will be adopted and identify which characteristics most strongly influence adoption speed. This can help shelters allocate resources earlier, prioritize animals needing more support, and potentially reduce stress and euthanasia rates.

This topic is personally meaningful because I have adopted two cats myself, and I hope these insights can support better outcomes for animals in shelters.

2. Data Overview

This dataset contains over 15,000 adoption records with the following types of feature:

Numerical

- Age (in months)
- Adoption fee
- PhotoAmt (number of photos uploaded)

Categorical

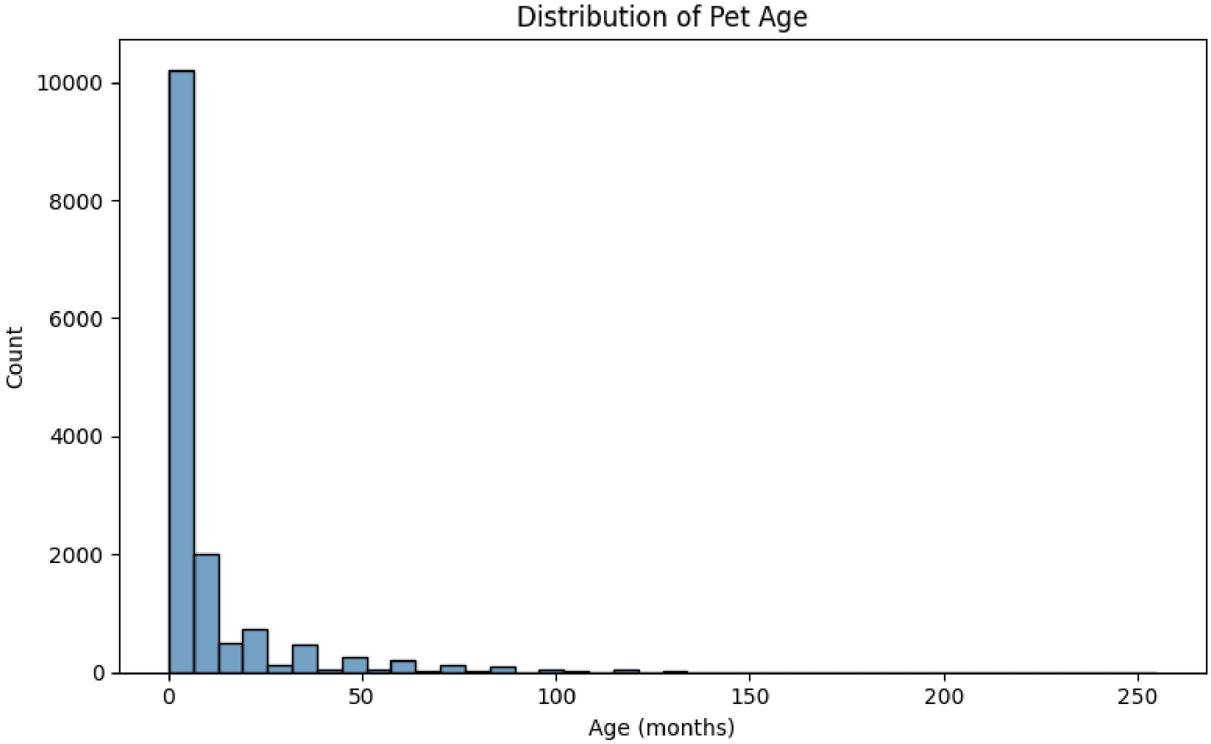
- Type

- Breed
- Gender
- Color
- Health Condition
- State (regional code in Malaysia)

Target Variable

- AdoptionSpeed (0-4)

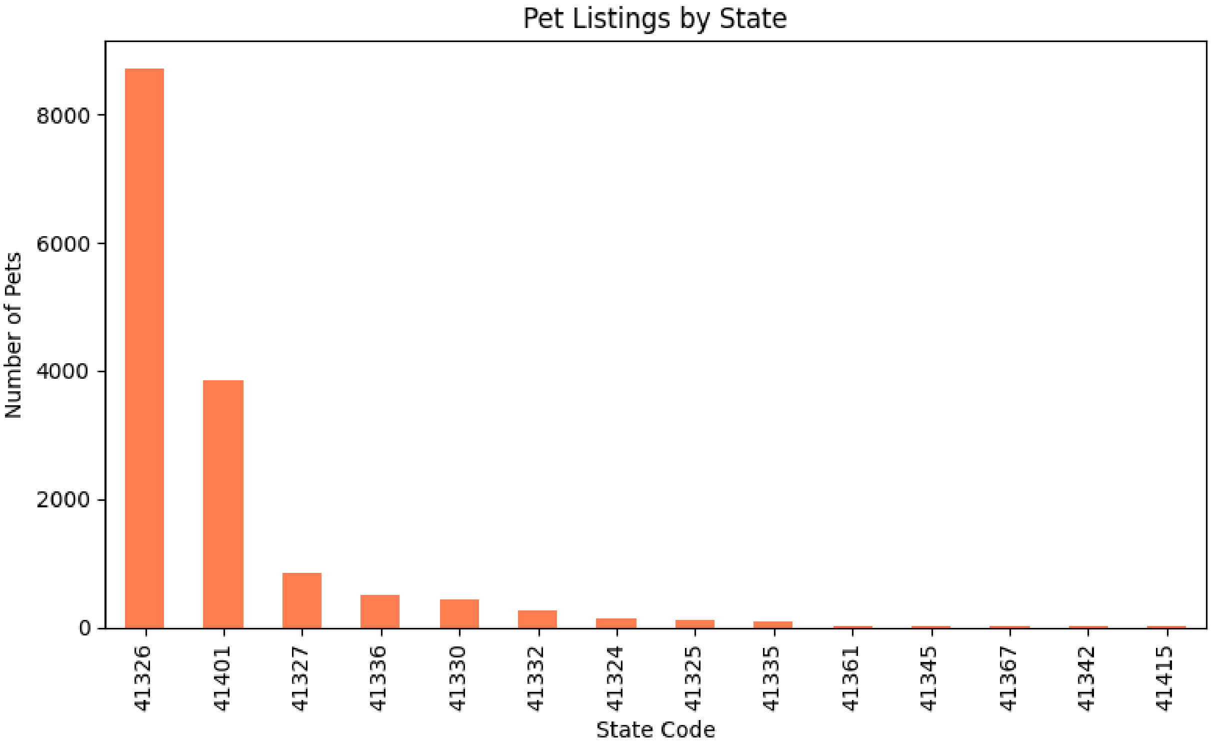
Age Distribution Histogram:



This age distribution is heavily skewed toward younger pets, especially under 6 months. Therefore, younger pets may overshadow older pets in adoption outcomes. This imbalance influences both visualization and modeling results.

3. Geographic Distribution (State)

Pet Listing by State:



This dataset shows that most adoption listings come from only a few Malaysian state codes, primarily 41326 and 41401.
This regional concentration of listings suggests uneven adoption activity and may affect model generalization.

4. Modeling Approach

Models Used

1. Logistic Regression
2. Random Forest Classifier

Train-Test Split

- 80% training
- 20% testing

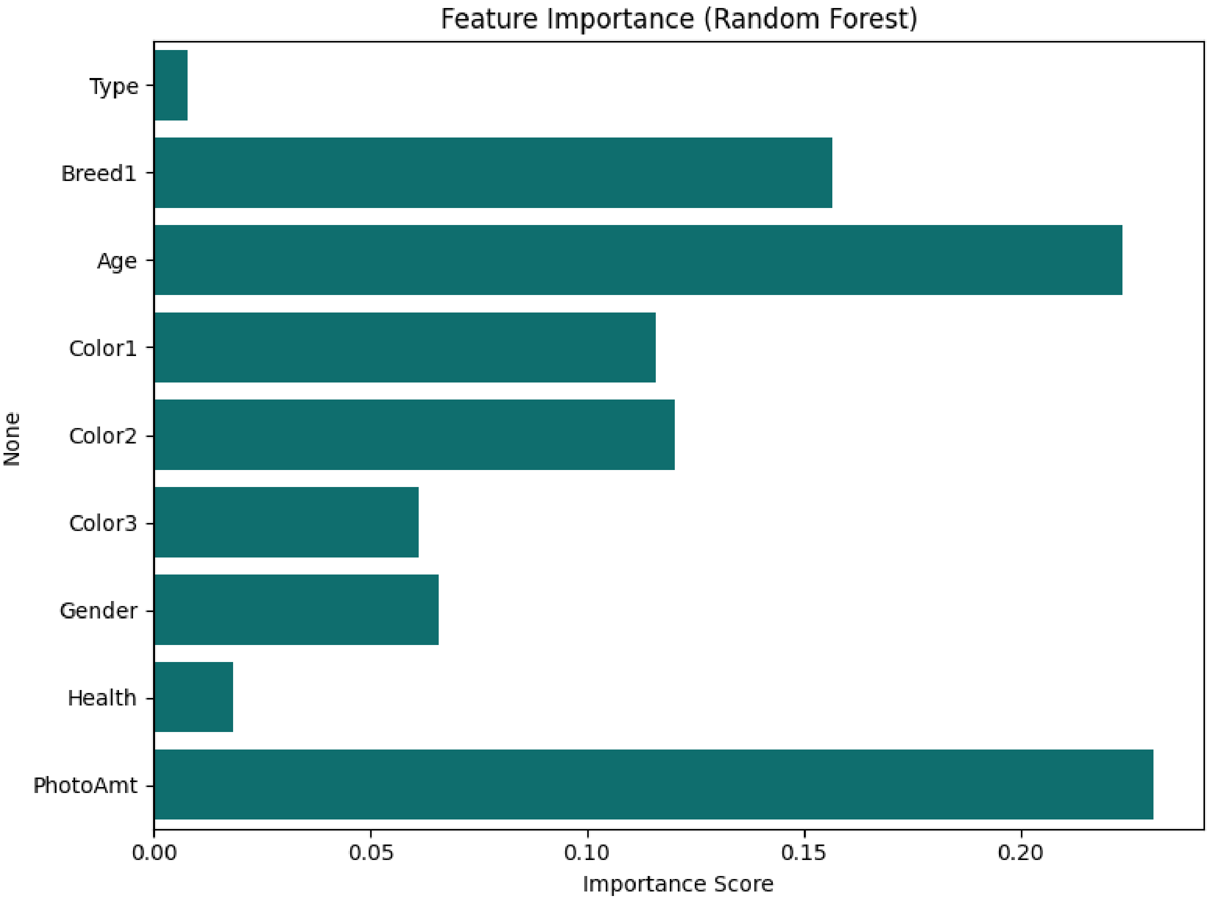
Results Summary

Model	**Accuracy**
Logistic Regression	~32%
Random Forest	~34% (best)

~34% accuracy has 5 classes target, the classes are imbalanced.
Random Forest performed better due to its ability to handle nonlinear relationships and mixed type data.

5. Feature Importance Analysis

Feature Importance (Random Forest):



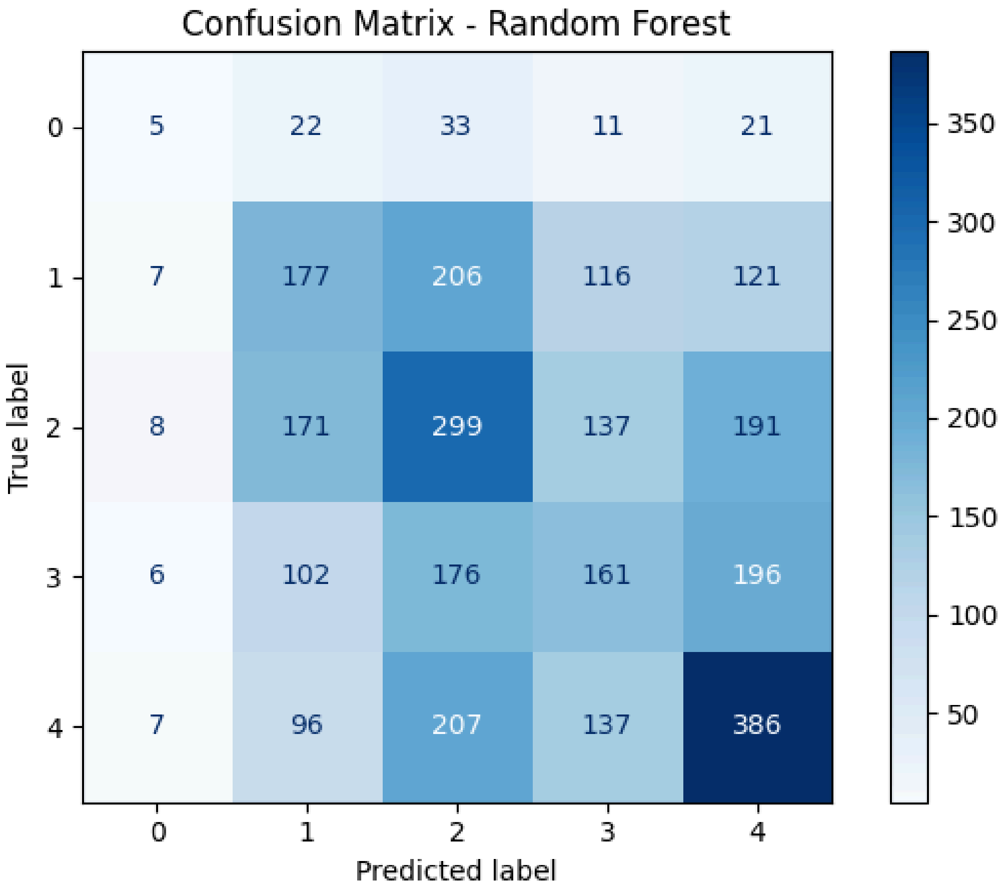
The most important predictors were:

- Breed
- Age
- PhotoAmt (Number of photos)

As we can see, certain breeds are adopted more quickly. Younger pets continue to be strongly favored. And pets with more photos tend to attract adopters faster, which is very important for shelters.

6. Confusion Matrix Analysis

Confusion Matrix - Random Forest:



Key Observations

- The model struggles with middle adoption categories (1, 2, 3).
- Class 4 has the lowest adoption and is predicted more accurately.
- Misclassifications occur because pets in middle categories often share similar characteristics.

This aligns with expectations, as adoption speed is influenced by many small factors not fully captured in numerical columns.

7. Conclusion

This project successfully explored pet adoption patterns and built predictive models based on key characteristics.

Breed, age and number of photos have the strongest influence on adoption speed. Regional activity is highly uneven, concentrated in major Malaysian states.

Shelter could increase adoption rates by improving photo quality, paying extra attention to older or less popular breeds, and focusing outreach in high volume regions.

References

- PetFind.my (https://www.kaggle.com/competitions/petfinder-adoption-prediction/data)
- Scikit-learn Documentation
- Seaborn & matplotlib Documentation