Predicting Dog Breed Groups from Their Characteristics

Sian Campbell Thomas, Brown University
Github: https://github.com/campbellthomas/DataFinal.git

Introduction

Dogs are the most popular pets in the United States with over 65.1 million families owning one [1]. Before choosing a dog, potential owners evaluate breeds with respect to how well suited a breed is for their situation and needs. Do they want a dog for protection? Does their dog need to be good with kids? How much exercise will they be able to give their dog? These are all questions that must be answered before bringing a dog home. Despite their popularity as pets, approximately 3.1 million dogs enter animal shelters annually [2]. Ensuring that dog personalities align well with owner personalities could significantly reduce the rate of rehoming, which is emotionally challenging for the dog and a logistical burden for shelters. This project aims to address this problem by classifying dogs into different breed groups (mixed breed, working, hound, companion, herding, terrier, sporting, and hybrid) based on their characteristics, such as shedding, size, friendliness, and other factors potential owners might view as important. If successful, the project could improve the process of matching dog and owner personalities and reduce the need for rehoming.

The dataset used for this project was obtained from Kaggle and contains ordinal rankings of traits and continuous measurements of qualities for various dog breeds. It was originally created for a university research competition that focused on classifying dogs based on their suitability for apartment living. This project tackles a different problem: using the characteristics to predict a dog's breed group. Previous work with this dataset, limited to the apartment-living classification task, reported accuracy scores in the 40% range. This relatively low accuracyhighlights the inherent complexity of the dataset and provides a useful comparison for evaluating the performance of models in this study.

Exploratory Data Analysis

The dataset consists of 4,000 data points and 35 columns, each representing a specific dog characteristic. The dataset is notably imbalanced. Mixed-breed dogs are represented with much greater frequency than others, accounting for over 30% of the observations in the dataset. Other groups, such as working, companion, sporting, hound, and herding dogs, each contribute between 10-15%. Terrier and hybrid dogs are the least represented, together comprising roughly 10% of the dataset.

A key challenge was the presence of missing data, which primarily affected personality-related features such as sensitivity level, prey drive, exercise needs, intensity, and weight. It was important to note that the proportion of missing data aligned closely with the representation of each breed group, suggesting no significant bias caused by missing data.

Fig. 1: Distribution of Missing Data

Analyzing the averages of various features across breeds revealed substantial variability in characteristics such as shedding amount, general health, size, and energy level. These features appeared particularly promising for predicting breed group, given their distinctiveness across classes.

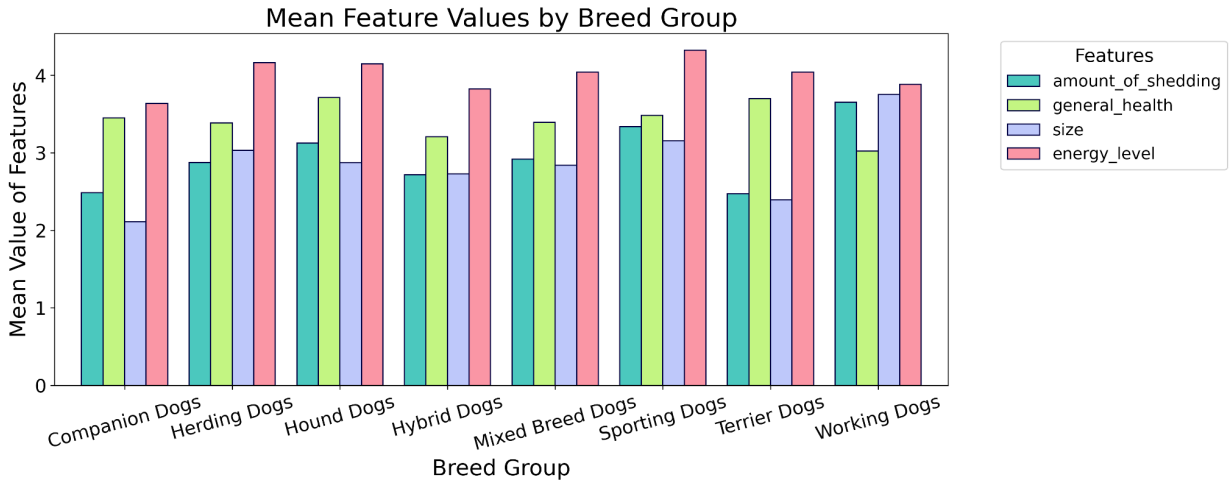

Fig. 2: Features with Heightened Variablility by Breed

Surprisingly, feature correlations were generally low. For example, traits like being kid-friendly and being friendly toward strangers, which intuitively seem related, showed weak correlation. Similarly, while smaller dogs are often considered healthier, no strong relationship between weight and general health was evident in the data. These findings informed the decision to retain all features except “Adapts well to apartment living.” This feature, originally the target variable of the dataset, lacked an ordinal ranking and was deemed irrelevant to this project’s goal.

Methods

The dataset was split into 80% training and 20% testing sets before download, stratified to preserve breed distribution. To further refine the data, the training set was split into 75% training and 25% validation subsets using a stratified train_test_split to ensure balance across breeds.

This resulted in a final 60-20-20 split among training, validation, and testing sets. I chose these splits because they are the industry standard for smaller datasets, balancing the need for both a training/validation step and a testing step.

Preprocessing
Several preprocessing steps were undertaken to prepare the data for modeling. Continuous variables such as weight, height, and lifespan were initially stored as strings (e.g., “25-30 pounds”). These were converted into continuous variables by calculating the average between the given measurements. Continuous features were subsequently standardized using StandardScaler() to ensure equal weighting and to improve model performance and interpretability.

Breed group names were label-encoded to facilitate modeling. Ordinal variables were already ranked on a 1-5 scale and did not require additional encoding. Missing ordinal values were replaced with a rank of 0. Including missing data is necessary as not all data will always be available in the real world. Missing data is still meaningful to the model and should be accounted for during training. This approach retained all data without imputation, a crucial step given the dataset’s imbalance and limited size. After preprocessing, the dataset contained 33 features.

Machine Learning Pipeline
To ensure robust evaluation, a stratified 5-fold cross-validation pipeline was implemented. Stratification preserved the breed distribution across folds, addressing the dataset’s imbalance. The choice of 5 folds balanced computational efficiency with the need for sufficient training data in each fold and to account for the small dataset. Shuffling was incorporated to ensure representative sampling, further mitigating bias in the evaluation process.

The evaluation metric was the F1 score, which balances precision and recall. This metric was chosen because the dataset’s imbalance made an accuracy_score misleading eg. high accuracy could result from overpredicting the dominant class. The F1 score offered a more nuanced view of model performance, reflecting its ability to classify all breeds fairly. For this project, precision and recall were equally important, as false positives and false negatives had comparable weights. Because this is not a high stakes classification question, such as a medical diagnosis, I did consider using an accuracy_score to maximize overall correct classifications. However, due to imbalance and the model’s tendency to overpredict certain breeds, I made the decision to use an F1 score. This metric is more reflective of the model’s performance across all breeds.

Models and Hyperparameters
Four machine learning algorithms were evaluated:

1. Logistic Regression: This was chosen as the first model to try due to its simplicity and interpretability.

2. Random Forest: Chosen because it is well-suited for mixed data types and non-linear relationships. Because my dataset had ordinal and continuous along and was unlikely to have a linear relationship with so many variables, this was a good model to try.
3. Support Vector Machine (SVM): Ideal for datasets with high dimensionality and limited samples. I had 30+ variables and only 4000 samples. Computationally intensive and expensive.
4. XGBoost: Powerful and thorough algorithm designed for imbalanced data. Allowed for complex optimizations and utilized early stopping.

Below is a table of the hyperparameters I tuned, in bold are the best parameters.

Logistic Regression	C [.001**,** .01**,** .1, 1, 10, **100, 1000,** 10000]	Solver ['liblinear']	Penalty [**'l1'**, 'l2']
Random Forest	N estimators [100, 400, **500**, 600]	Max depth [None, **10,** 20, 30]	
Support Vector Machine	C [1e0, 1e1, 1e2, 1e3, **1e4,** 1e5]	Kernel ['linear', **'rbf'**, 'poly']	Gamma [1e-7, **1e-6,** 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1e0, 1e1]
XGBoost	N estimators [50, 100, 200, 500, **600,** 800]	Subsample [0.4, 0.6, **0.8**, 1.0]	Colsample by tree [.2, **0.4,** 0.6, 0.8, 1.0]

Table 1: Hyperparemeteres of models tested

Hyperparameters for each model were tuned using cross-validation. Four random states were looped through to assess variability, and the mean and standard deviation of F1 scores were recorded to quantify uncertainty.

Results

The baseline model, which predicted the most common breed group (mixed breed) for all samples, achieved an accuracy of 34.5% and an F1 score of 17.7%. This served as a benchmark for evaluating the predictive power of the machine learning models.

Model Performance

In terms of accuracy, XGBoost performed best, achieving 46.8% (±0.9%), followed closely by Random Forest at 46.0% (±0.3%). Logistic Regression achieved 40%, while SVM performed poorly with an accuracy of only 15.7%. Results are visualized in the graph below (fig. 3).

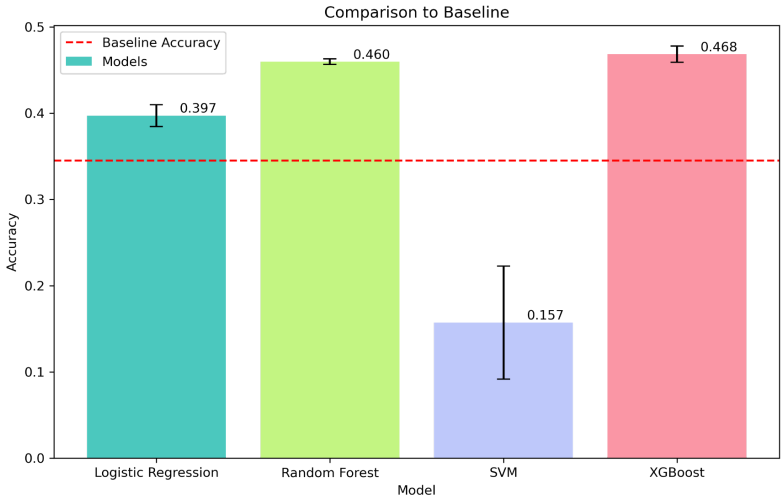

Fig 3. Comparing Models to the Baseline Accuracy Score

However, F1 scores provided a more meaningful comparison given the dataset's imbalance. Random Forest had the highest F1 score at 43.1%, followed by XGBoost at 42.8% and Logistic Regression at 40%. SVM had an extremely low F1 score of 10.9%. The results are visualized below.

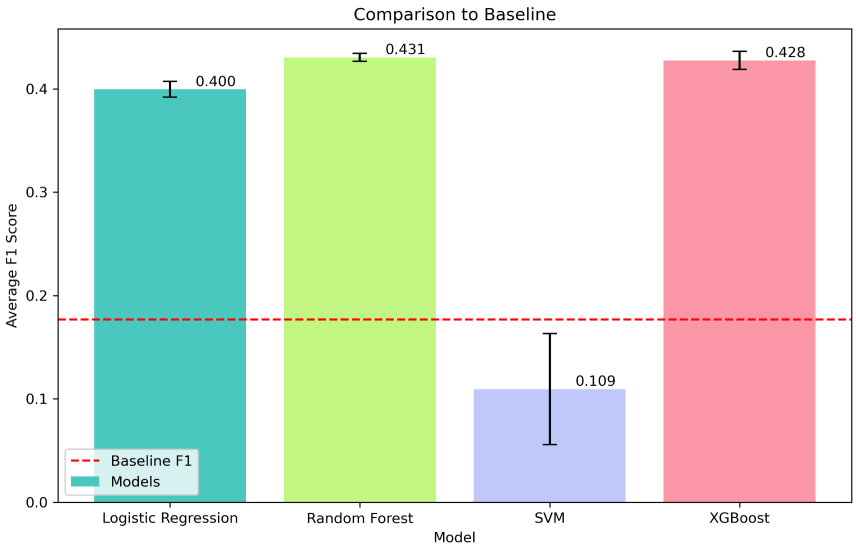

Fig 4. Comparing Models to the Baseline F1 Score

Both Random Forest and XGBoost outperformed the baseline, but did not perform well. Random Forest's F1 score exceeded the baseline by 18 standard deviations, while XGBoost exceeded it by 16. Logistic Regression also surpassed the baseline but by a narrower margin. SVM consistently underperformed, falling below the baseline by several standard deviations.

I compared the confusion matrices of both Random Forest and XGBoost to better understand the models. First, I looked at Random Forest when using accuracy to optimize. This resulted in significant overprediction of hound dogs and no classification of companion or terrier dogs.

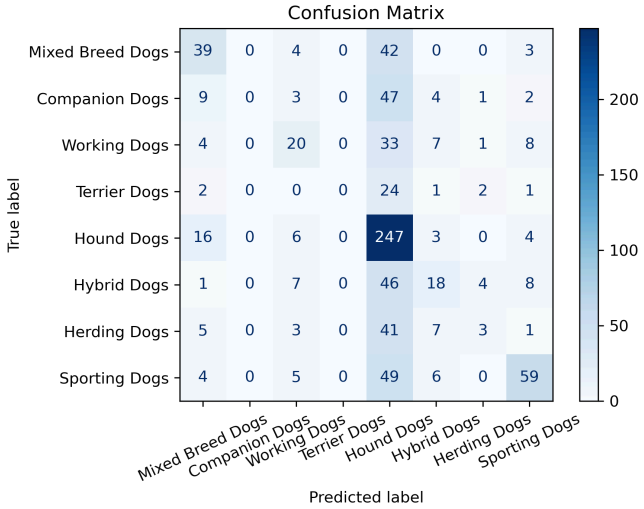

Fig. 5: Confusion Matrix of Random Forest Model Using Accuracy

I then compared this to the confusion matrix when using F1 during cross validation. This vastly improved the over prediction of hound dogs, but they were still prone to many false positives.

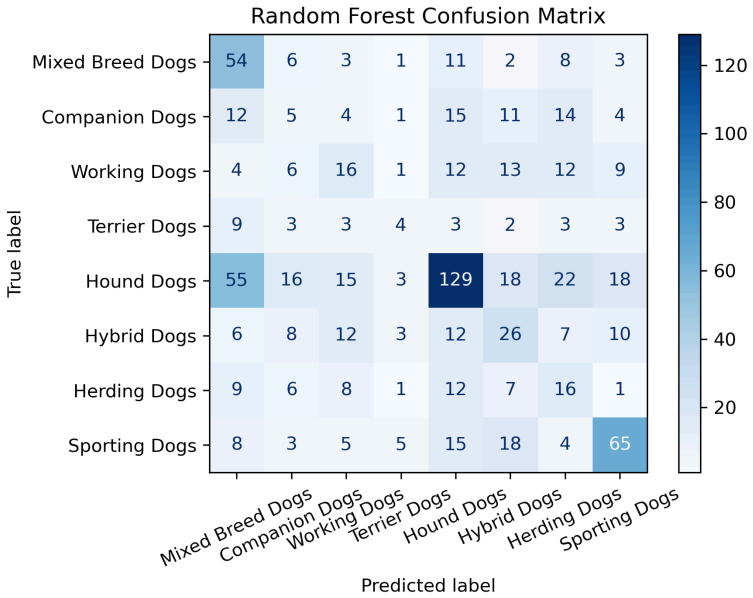

Fig. 6: Confusion Matrix Random Forest Model Using F1

Lastly, I compared the random forest matrices to a confusio matrix for XGBoost using F1 to help decide which model was better (since they performed very similarly). The matrix displayed that it also massively overpredicted hound dogs while rarely predicting companion and terrier dogs. This model's overprediction was similar to random forest using accuracy.

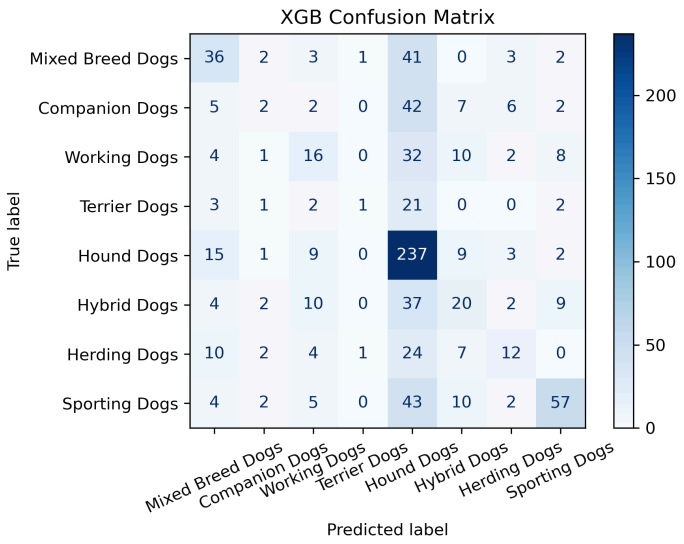

Fig. 7: Confusion Matrix of XGBoost Model Using F1

The confusion matrices lead me to believe that the Random Forest model using 500 estimators and a maximum depth of 10 was the bet model.

Feature Importance

To interpret the models, global feature importances were calculated using permutation importance, SHAP values, and Gini importance. Permutation importance identified average lifespan as the most significant feature, nearly double the importance of others. Weight, height, and shedding amount also featured prominently. Interestingly, size did not appear in the top 10, perhaps because weight and height indirectly captured this characteristic.

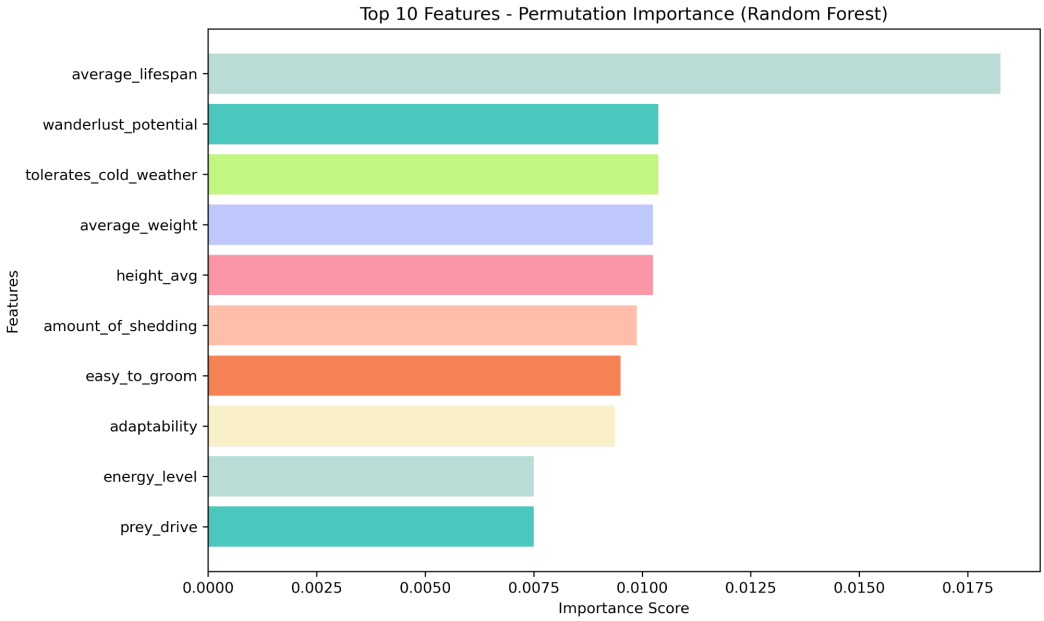

Fig 8: Permutation Importance of RF Model

SHAP values corroborated these findings, highlighting similar features but with varied rankings. SHAP provided granular insights into the marginal impact of each feature, showing that traits like tolerance to cold weather and wanderlust potential were consistently important.

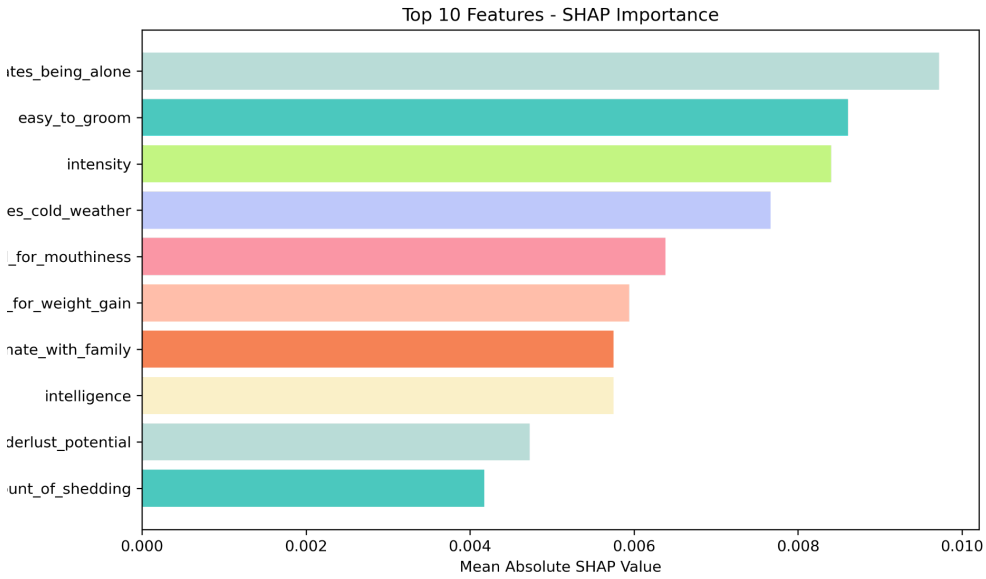

Fig 9: SHAP Importance of RF Model

Gini importance added another perspective, emphasizing features like adaptability and energy level in addition to tolerance to cold weather. All three rated tolerance to cold weather, indicating that was continuously a predictive feature.

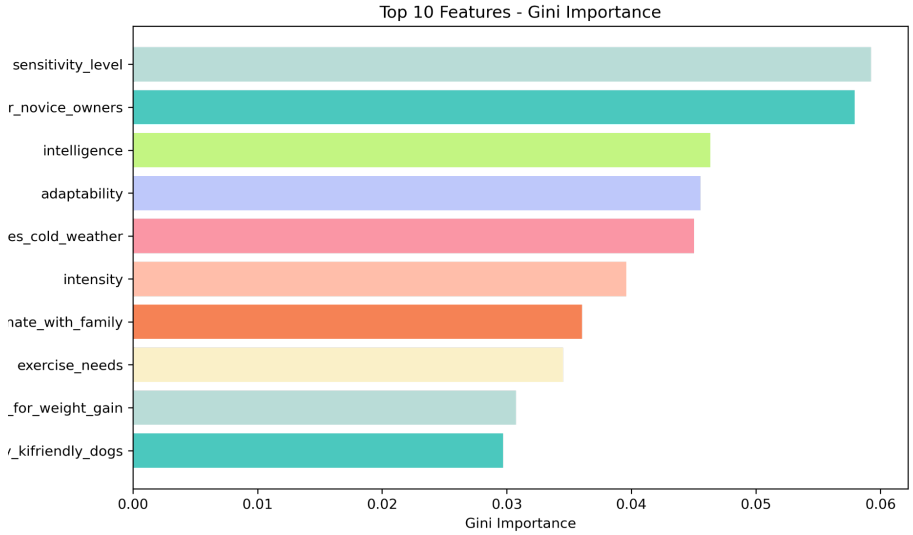

Fig 10: GINI Importance of RF Model

Local Interpretability

SHAP values were analyzed for misclassified and correctly classified samples. Regardless of the prediction outcome, features such as intensity, exercise needs, and tolerance to cold weather consistently influenced predictions. This uniformity suggests that the models relied heavily on a core feature subset, limiting their ability to differentiate between similar breeds. The shap force diagrams for Hounds (most commonly predicted, with many false positives), Companion dogs (least commonly predicted, many false negatives), and sporting dogs (also commonly predicted)

Fig. 11 Hound SHAP Local Importance

Fig. 12 Companion Dog SHAP Local Importance

Fig. 13 Sporting Dog SHAP Local Importance

Outlook

Improving the Model

Several strategies could improve model performance. First, removing strongly correlated features could simplify the model and reduce noise. Expanding the search range in GridSearchCV would also allow for more comprehensive hyperparameter tuning. Finally, obtaining additional data, particularly for underrepresented breeds, would enhance model generalizability and reduce imbalance. I believe additional data and applying a more intensive model would be the most effective measures.

Improving Interpretability

Analyzing performance metrics for each breed could reveal nuances in the model's strengths and weaknesses. Advanced tools like LIME or SHAP can provide deeper insights into predictions for underrepresented breeds, making the model more trustworthy and actionable.

Conclusion

This project demonstrated the challenges of predicting dog breeds based on their characteristics. While Random Forest and XGBoost showed the most promise, their performance highlights the dataset's complexity and limitations. Insights from feature importance and interpretability analyses suggest that refining the feature set and collecting more balanced data are critical next steps.

Citations

[1.] Megna, M., & Bailie, K. (2024, October 22). Pet ownership statistics 2024. Forbes. https://www.forbes.com/advisor/pet-insurance/pet-ownership-statistics/#:~:text=66%25%20of%20U.S.%20households%20(86.9,fish%20(11.1%20million%20households).

[2.] ASPCA. (2019). Pet Statistics. https://www.aspca.org/helping-people-pets/shelter-intake-and-surrender/pet-statistics

[3] Kaggle Dataset: https://www.kaggle.com/datasets/agarwalyashhh/dog-adaptability/data