Random Forest Model Report

# 1. Model Architecture

## 1.1 Dataset Description

The dataset used in this report is for predicting pet adoption speed, which is a multiclass classification problem.  
The target variable is 'AdoptionSpeed' with 5 categories (0 to 4).  
The features include categorical features such as 'Breed1', 'Breed2', 'Color1', 'Color2', 'Color3', and others.  
There are numerical features such as 'Age', 'Fee', 'PhotoAmt', etc.  
Target encoding was used for categorical features such as 'Breed1\_encoded', 'Color1\_encoded', etc.

## 1.2 Random Forest Model

The Random Forest Classifier is an ensemble method that builds multiple decision trees and merges them to get a more  
accurate and stable prediction. It is robust against overfitting and can model complex non-linear relationships.  
The model uses 'n\_estimators' (number of trees) and 'max\_depth' (maximum depth of each tree) hyperparameters to control  
the complexity of the model. Each tree in the forest is trained on a random subset of the data, and predictions are made  
by aggregating the predictions of all trees.

# 2. Evaluation Metrics

（和决策树一样，应该也和所有的一样）

## 2.1 Accuracy

Accuracy is the proportion of correctly predicted samples out of the total number of samples. It is a commonly used metric  
for classification problems, but can be misleading in the case of imbalanced datasets.

## 2.2 Confusion Matrix

The confusion matrix is a tool used to evaluate the performance of a classification model by comparing the true  
labels with the predicted labels. It provides a more detailed view of the model’s accuracy for each class.

## 2.3 Precision, Recall, F1-Score

Precision is the proportion of positive predictions that are correct. Recall is the proportion of actual positives  
that are correctly identified by the model. F1-score is the harmonic mean of precision and recall, providing a single metric  
to evaluate both aspects. These are especially important when dealing with imbalanced data.

## 2.4 Macro and Weighted Average

Macro average computes the metric independently for each class and then takes the average, treating all classes equally.  
Weighted average takes the support (the number of true instances for each class) into account when averaging the metrics.

# 3. Results and Analysis

## 3.1 Confusion Matrix

The confusion matrix for the Random Forest model is as follows:  
  
[[ 6 20 20 3 18]  
 [ 3 163 166 70 74]  
 [ 2 118 231 86 131]  
 [ 2 72 154 157 111]  
 [ 2 51 101 61 353]]  
  
- Class 0 has low recognition performance, often misclassified as classes 1 and 2.  
- Class 4, the most frequent class, is predicted more accurately with a majority of correct predictions.  
- Class 2 and 3 show moderate predictive stability, but confusion among classes remains noticeable.

## 3.2 Classification Report

The classification report is as follows:  
  
precision recall f1-score support  
  
0.0 0.40 0.09 0.15 67  
1.0 0.38 0.34 0.36 476  
2.0 0.34 0.41 0.37 568  
3.0 0.42 0.32 0.36 496  
4.0 0.51 0.62 0.56 568  
  
accuracy 0.42 2175  
macro avg 0.41 0.36 0.36 2175  
weighted avg 0.42 0.42 0.41 2175  
  
- Class 0 shows poor precision and recall, indicating difficulty in recognizing rare classes.  
- Class 4 achieves the highest precision and recall, consistent with its majority presence.  
- Overall accuracy is 42%, showing a clear improvement over the decision tree baseline (35%).

## 3.3 Results Analysis

The Random Forest model demonstrates improved predictive performance compared to a basic decision tree.  
However, class imbalance remains a major issue, especially for class 0, where the model struggles.  
Further techniques like SMOTE oversampling, undersampling, or using class\_weight adjustments are recommended to balance the model's predictive ability across classes.

## 3.4 Conclusion

The Random Forest model offers a strong baseline with an accuracy of 42%.   
It shows better handling of frequent classes but continues to suffer from misclassifications among minority classes.  
Further hyperparameter tuning and class balancing strategies are expected to enhance model performance.

# 4. Feature Importance

Below is the feature importance chart for the Random Forest model, which highlights the most important features  
in predicting pet adoption speed. 'Age' and 'PhotoAmt' are found to be the most influential features.

