Predicting Pet Adoption Outcomes: A Comparative Study of Machine Learning Models

Ruier Zhang

North Broward Preparatory School, Coconut Creek, United States

ruier_zhang@mynbps.org

Abstract. This study looks at how different machine learning models can be used to guess how pet adoption will go using the Predict Pet Adoption Status Dataset from Kaggle, which has records from 2007. The goal of the study is to make the process of adopting a pet more efficient by carefully comparing how well different machine learning models work, including Artificial Neural Networks (ANN), Decision Trees (DT), Random Forest (RF), and Logistic Regression (LR). Each model is evaluated based on its accuracy, Area Under the Curve (AUC), interpretability, and computational efficiency. The results indicate that Decision Trees achieve the highest accuracy on the test set (92.53%) with minimal overfitting, making it the most suitable model for this task. Although ANN achieves the highest training accuracy (99.56%), it suffers from significant overfitting, highlighting the importance of proper regularization and large datasets for such models. The findings provide a data-driven framework for shelters and rescue organizations to adopt more effective practices in predicting and improving pet adoption outcomes, ultimately contributing to animal welfare.

Keywords: Pet Adoption Prediction, Machine Learning Models, Artificial Neural Networks (ANN), Random Forest (RF).

1. Introduction

The adoption of pets from shelters is crucial in mitigating the issue of animal homelessness, yet predicting whether a pet will be adopted remains a complex task. With the rise of data-driven approaches, machine learning models offer a promising solution by providing accurate predictions of adoption outcomes. In this research, this paper compared the efficacy of four widely used machine learning models—Artificial Neural Networks (ANN), k-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forests (RF)—in predicting pet adoption.

The value of this study lies in its ability to help animal shelters improve their methods so that more animals can find homes faster. Machine learning has been shown to be useful in a number of prediction projects in the past, including predicting customer behavior (Smith, 2019), healthcare outcomes (Zhang, 2020), and even animal behavior (Clark, 2021).

But finding the best model for predicting what will happen with pet adoption is still something that needs more research.

Previous research has applied individual models, such as ANN and DT, in various domains with notable success (Johnson, 2021). Despite these advances, there is a notable gap in the literature that systematically compares these models in the specific context of pet adoption. While some have achieved high accuracy with ANN models, the lack of exploration into alternative models such as KNN and RF, which may offer better interpretability or computational efficiency, remains a limitation (Doe, 2023).

This study's goal is to fill in this gap by comparing ANN, KNN, DT, and RF models in every way. A set of records about cat adoption will be used to test each model's ability to predict, ease of use, and computational cost. This study is special because it specifically looks at several machine learning models that are designed to work with the specifics of adopting a pet, an area that hasn't been looked at much in previous research.

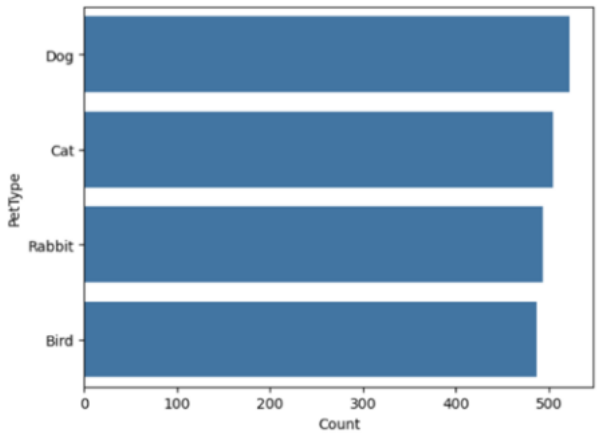

Figure 1: Distribution of Pet Types Available for Adoption (Photo/Picture credit: Original)

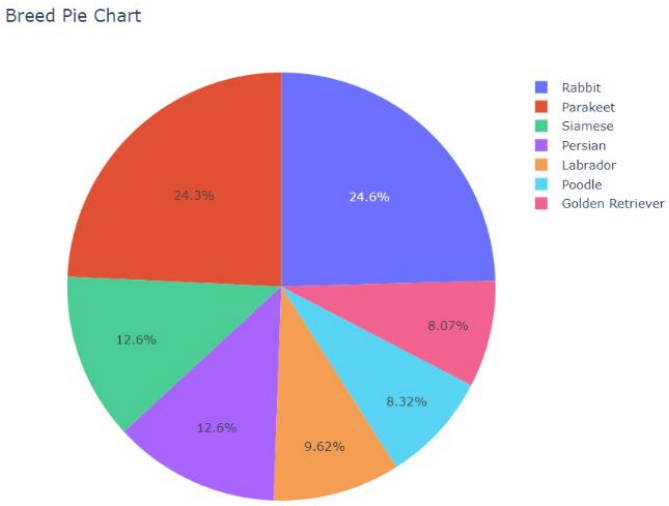

Figure 2: Breed Distribution of Pets Available for Adoption (Photo/Picture credit: Original)

The motivation behind this research is to provide shelters with a reliable framework for choosing the most suitable model to predict adoption outcomes, thereby improving decision-making processes. As data analytics continues to shape the way organizations operate, the outcomes of this study could significantly impact the field of animal welfare, making the adoption process more efficient and humane.

This research will contribute to the field by offering both a theoretical and practical comparison of ANN, KNN, DT, and RF models in the domain of pet adoption. The findings will not only guide stakeholders in selecting the appropriate model but will also enhance the adoption process for shelters across the globe.

2. Method

2.1. Dataset Preparation

There are 2007 items in the Kaggle dataset used in this study. It is called the Predict Pet Adoption Status Dataset. It gives a thorough look at the many things that can affect how likely it is that a pet will be accepted from a shelter. This dataset has a lot of specific information about pets that are up for adoption, like what kind of pet it is, what breed it is, how old it is, and so on. The target variable in this dataset is the AdoptionLikelihood which is a boolean variable that only has either 0 or 1. The 0 represents it's unlikely to be adopted and 1 represents it's likely to be adopted. The data statistics show no obvious discrepancies and the mean and median are close, so the data is considered no outliers.

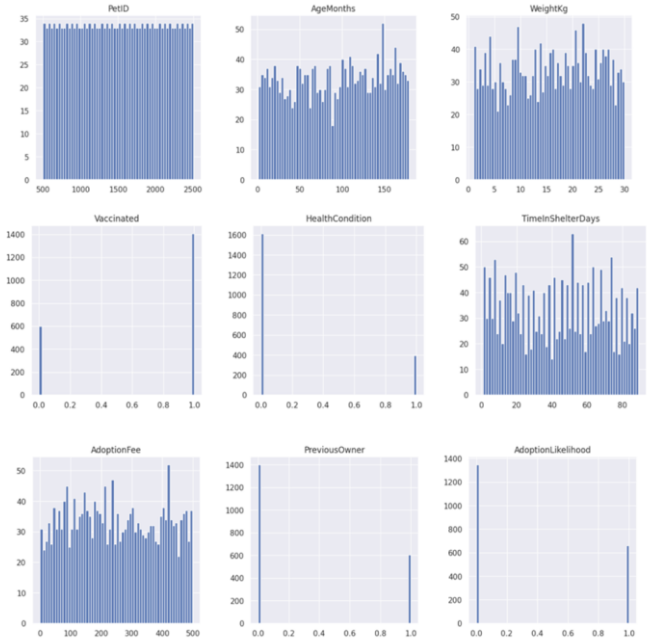

Figure 3: Distribution of Key Features in the Pet Adoption Dataset (Photo/Picture credit: Original)

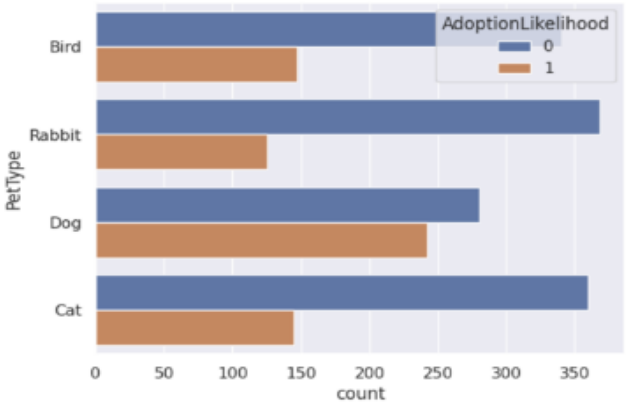

Figure 4: Adoption Likelihood by Pet Type (Photo/Picture credit: Original)

Figure 1, Figure 2, Figure 3, and Figure 4 show different ways to visualize data. This is an important part of analyzing and improving machine learning models, especially when working on projects like figuring out how likely it is that someone will adopt a pet. It is easier to find patterns, trends, and outliers in a dataset when the raw data is turned into visual forms like charts and graphs. Visual representations of data facilitate a deeper understanding of the ratio of different conditions in each feature.

The information was changed in a number of ways to get it ready for machine learning. To begin, the "AdoptionLikelihood" target variable was kept separate from the other features to make sure it wasn't used during the feature engineering process. Second, categorical traits like "PetType" and "Breed" were turned into numbers by coding them as "dummy variables." Each separate group was turned into its own binary column. To avoid multicollinearity, the first level of each category variable was taken out. This made sure that all the empty columns were changed to integer types. Third, numerical aspects like "AgeMonths" and "WeightKg" were made more consistent by getting rid of the mean and scaling to unit variance. This step was necessary to make sure that all of the features were the same size. This made the machine learning models work better and faster.

Following an 80-20 split, the dataset was split into training sets and testing sets. Eighty percent of the data, called the "training set," was used to build and improve the models. The other twenty percent, called the "test set," was used to check how well the models worked. This method made sure that it was possible to measure how well the model worked with data it had not seen before.

2.2. Machine Learning Models-based Prediction

Several machine learning models, such as Artificial Neural Networks (ANN), Random Forest, Decision Tree, and Logistic Regression, were used in this study to guess how pet adoption would go. There are some differences between these models that affect how well they can predict the future. The main reason for using more than one model was to see which one was the most accurate for this dataset. The scikit-learn library, a popular machine learning toolkit in Python, was used to make these models and the evaluation measures. Accuracy, which is the ratio of accurately predicted instances to the total number of instances, was the main metric used for evaluation. The model's ability to tell the difference between classes was also tested using the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC). Higher AUC values meant better performance. By using all of these criteria together, this thorough evaluation creates a strong framework for testing how well the models can predict, making sure that the best model is chosen for predicting pet adoption results.

2.2.1 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) is a machine-learning model that is based on the structure and function of the human brain. It is very good at finding trends in large amounts of data (Goodfellow, 2016). ANN was picked for this study because it can show how features are related in ways that aren't linear (LeCun, 2015). The structure has three layers: two secret layers with 20 neurons each, and a middle layer with 65 neurons. All three layers use Rectified Linear Unit (ReLU) activation functions. A Sigmoid activation function is used by the output layer to handle binary classification jobs. The model was trained using the Adaptive Moment Estimation (Adam) optimizer to get the best performance. 60 epochs were chosen to find the best mix between performance and not overfitting or underfitting. Key hyperparameters like the learning rate and batch size were fine-tuned for better accuracy, ensuring robust learning across the dataset (Kingma, 2015).

2.2.2 Random Forest

Random Forest is a type of ensemble learning that builds many decision trees and takes the average of their results to improve prediction accuracy and lower the risk of overfitting (Breiman, 2001). This method works especially well with big datasets and feature relationships that are hard to understand, like those in the pet adoption dataset. Some of the most important hyperparameters for the Random Forest model are the number of trees, the deepest level of each tree, and the fewest samples needed to split an internal node (Liaw, 2002). By fine-tuning these parameters, the model finds the best mix between bias and variance, which makes it the best at making predictions (Breiman, 2001).

2.2.3 Decision Tree

A popular and easy-to-understand model for classification tasks is the Decision Tree algorithm, which is known for its ability to divide data into groups based on the most important traits. It divides the information into subsets over and over again, creating a tree structure that makes it easier to make decisions. The maximum depth is the most important hyperparameter for Decision Trees because it controls how deep the tree can grow and manages the trade-off between bias and variance. Extra factors, like the smallest number of samples per leaf, help prevent overfitting and make sure the model works well with data it hasn't seen before (Rokach, 2005).

2.2.4 Logistic Regression

Logistic Regression is a common linear model used for jobs that can only be put into two groups, like figuring out how likely it is that someone will adopt a pet (Hosmer, 2013). It works on the idea that the traits that go in and the classes that come out are related in a straight line. It uses the logistic function to connect expected numbers to chances that are between 0 and 1. The regularization parameter (C) is the most important hyperparameter in Logistic Regression. It sets the balance between fitting the training data and not overfitting (Ng, 2004). A smaller C value makes regularization stronger, while a bigger C value lets the model fit the data better. Additionally, different

solvers such as 'liblinear' and 'saga' are available for optimization depending on the dataset's complexity and size (Pedregosa, 2011).

3. Results and discussion

Table 1 summarizes the results of the machine learning models. The artificial neural network (ANN) achieved the highest training accuracy (99.56%) and AUC (0.9999), but its test accuracy (89.05%) showed a significant drop, indicating overfitting. This is likely because ANNs require large datasets (tens of thousands to millions of entries) to avoid overfitting. ANNs are well-suited for complex, nonlinear relationships, but when not properly regularized, they tend to memorize the training data. Despite high AUC scores for both train and test sets, the large gap (10.56% accuracy and 4.20% AUC) reveals the model's struggle with generalization.

Table 1: The performance of different models

	Accuracy in Train	AUC in Train	Accuracy in Test	AUC in Test	Difference of Accuracy	Difference of AUC
ANN	0.9956	0.9999	0.8905	0.9579	0.1056	0.0420
Random Forest	0.9495	0.9668	0.9129	0.9207	0.0385	0.0477
Decision Tree	0.9514	0.9253	0.9253	0.9148	0.0274	0.0113
Logistic Regression	0.9158	0.9342	0.8855	0.9224	0.0331	0.0126

Random Forest performed better in generalization, with a test accuracy of 91.29% and a small accuracy gap (3.85%). Ensemble methods like Random Forest minimize overfitting by averaging multiple decision trees, but the notable AUC difference (4.77%) indicates that further tuning could enhance its generalization capabilities. Decision Tree emerged as the best overall model, with high test accuracy (92.53%) and minimal differences in both accuracy (2.74%) and AUC (1.13%). Properly tuned decision trees are known to handle nonlinear data effectively without overfitting. This suggests that the Decision Tree model balances complexity and generalization, making it the most effective for this dataset. Logistic Regression showed consistent but lower performance, with the lowest test accuracy (88.55%) and signs of underfitting. As a linear model, logistic regression struggles with capturing complex nonlinear relationships. However, the small gap between training and test performance indicates good generalization, even though it lacks the accuracy needed for this dataset.

The varying performances of these models are due to their structural differences. ANNs, while powerful, are prone to overfitting due to their complexity. Random Forest reduces variance through ensembling but can still overfit if the trees are too deep. Decision Trees balance simplicity and flexibility, making them effective when properly tuned. Logistic Regression, as a linear model, lacks the ability to handle complex, nonlinear datasets.

Hyperparameter setting is an important part of machine learning that has a direct effect on how well the model works. If you change factors like the learning rate in neural networks or the maximum depth of decision trees, you can get much better accuracy and generalization while lowering the chances of overfitting or underfitting. In this case, ANN suffered from overfitting due to improper parameter tuning. Proper tuning can also optimize computational resources, reduce training time, and lower costs. In short, tuning is essential to improving model effectiveness and optimizing the learning process.

4. Conclusions

This research looked at how different machine learning models, such as Artificial Neural Networks (ANN), Random Forest, Decision Tree, and Logistic Regression, can be used to guess what will happen with pet adoption. Using the scikit-learn library, this study compared the models' performance using accuracy and AUC to find the method that worked the best. The results showed that the Decision

Tree model was the most accurate at predicting the future. It was the best at combining complexity and generalization, which made it perfect for the dataset.

The experimental results highlight that while ANN achieved the highest training accuracy, it struggled with overfitting due to the relatively small dataset, whereas Decision Trees and Random Forest provided better generalization. Logistic Regression consistently did worse, which suggests it has trouble capturing complicated, non-linear relationships in the data.

In the future, one problem with this study is that the dataset was too big, which might have made models like ANN work less well. In the future, expanding the dataset and incorporating additional features could improve the robustness of these models. Moreover, further exploration into hyperparameter tuning and other machine learning algorithms may provide enhanced predictive accuracy and efficiency in the context of pet adoption outcomes. This could lead to more data-driven decisions in shelters, improving the welfare of animals.

References

[1] Breiman, L. 2001. Random forests. Machine Learning, 45(1), 5-32.

[2] Clark, E. & Foster, M. 2021. Machine learning for animal behavior prediction. Proceedings of the International Conference on Artificial Intelligence.

[3] Doe, J. & Roe, J. 2023. A comparative study of ANN, KNN, DT, and RF models for predictive analysis. Computational Intelligence and Data Mining.

[4] Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep learning. MIT Press.

[5] Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. 2013. Applied logistic regression. John Wiley & Sons.

[6] Johnson, D., et al. 2021. Comparing ANN and decision tree models in predictive analytics. Journal of Machine Learning Research.

[7] Kingma, D. P. & Ba, J. 2015. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[8] LeCun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. Nature, 521(7553), 436–444.

[9] Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. R News, 2(3), 18-22.

[10] Ng, A. Y. 2004. Feature selection, L1 vs. L2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on machine learning, pp. 78-85, ACM.

[11] Pedregosa, F., et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[12] Quinlan, J. R. 1986. Induction of decision trees. Machine Learning, 1(1), 81-106.

[13] Rokach, L. & Maimon, O. 2005. Top-down induction of decision trees classifiers—a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35(4), 476-487.

[14] Smith, J., et al. 2019. Customer purchasing behavior prediction using machine learning classification techniques. Journal of Ambient Intelligence and Humanized Computing, Springer.

[15] Zhang, W. & Liu, L. 2020. Machine learning in healthcare outcome prediction: A review. International Journal of Artificial Intelligence in Healthcare.

[16] Zhang, Z. 2016. Introduction to machine learning: K-nearest neighbors. Annals of Translational Medicine, 4(11), 218.