Predicting the Longevity of Dog Breeds Based on Various Attributes

Mr. Fernando	Mr. Hennayaka	Mr. Silva	Mr. Wanigasekara	Mrs. Begam
Faculty Of IT	*Faculty Of IT*	*Faculty Of IT*	*Faculty Of IT*	*Faculty Of IT*
Horizon Campus	*Horizon Campus*	*Horizon Campus*	*Horizon Campus*	*Horizon Campus*
Sri Lanka	Sri Lanka	Sri Lanka	Sri Lanka	Sri Lanka
ITBNM-2110-0016	ITBNM-2110-0021	ITBNM-2110-0052	ITBNM-2110-0088	ITBNM-2110-0006

***Abstract*—The research analyzes the attributes of score, size, suitability for children(score for kids,) and lifetime cost that influence the longevity of dog's life. Application of Linear Regression, Random Forest, Support Vector Regression (SVR), and XGBoost machine learning models were used to predict the span of life of dogs based on the above-mentioned attributes. Among these, the prepossessing techniques involved were scaling and feature engineering to upscale the performance of the models. Of these, the best model that emerged was the SVR that gave a concise R-squared value of 0.90 and MSE of 0.39. The results clearly indicate that genetics, financial investments, and care for the dogs influence the life span of dogs most dramatically. This information is very useful for any breeder or a pet owner. [1] [2] [3]**

A. Key Words

Dog Breeds, Longevity Prediction, Machine Learning, SVR, Random Forest, Data Science, Feature Engineering.

I. Introduction

A. Background Information

The study have indicated that the dog family depicts huge variation in life cycles dependent on the size of its species, inbreeding, and conditions of care [1] [2]. Normally, large dog breeds have shorter life spans, while dogs with mixed breeds mostly outlive purebred dogs since inbreeding does less harm [1]. This would help the veterinarians and pet owners also in being able to forecast and make better decisions about the care that is to be given to the dogs and breeds they have [3]. Traditionally, longevity studies have focused on genetic predisposition and congenital conditions. However, environmental factors such as a dog's size**,** dog's score**,** lifetime costs, and interactions with children may also play a significant role in determining lifespan. In the era of data science, machine learning models provide a powerful approach for predicting outcomes from complex, multi-dimensional datasets. By applying these techniques to dog breed data, this study seeks to provide a predictive framework that considers non-genetic factors.

II. Research Problems or Questions

This study addresses two main questions:

1) What are the most influential attributes for predicting a dog's lifespan?
2) Can machine learning models, such as Support Vector Regressor, Random Forest, and XGBoost, effectively predict the longevity of dog breeds

III. Significance of the Research

The mere fact that such insight into the longevity of dogs, and being able to predict it, may go a long way in helping a dog owner make informed decisions with respect to the care, diet, and health of the animal. This would also be a tool that the veterinarian can use to check for breed-related risks and build an emergent body of research in pet care analytics. The focus on non genetic factors extends the research on longevity beyond the level of biological explanations [1] [2].

IV. Literature Review

A. Overview of Relevant Literature

In fact, previous research has stipulated genetics as a fundamental factor in determining the dogs' life spans, especially about their sizes and inbreeding. Yordy et al. [1] have established a positive proper relation between body size and lifetime wherein the smaller dogs always outlived the larger ones. Similarly, in those variables that Urfer et al. [3] accounted for as specific risk factors in the lifespan of dogs within primary veterinary care were those dimensions wherein environmental causes alone were such that dietary and exercise-wise issues will also manifest themselves in the determination of longevity.

Socha et al. [2] extended that to investigate a broader set of attributes, including behavioral and financial lifetime cost and intelligence. They found that these factors are informative predictors of lifespan. The results are in agreement with the conclusions reached by Adams et al. [4], who investigated health issues in purebred dogs and came to the conclusion that economic and lifestyle factors might interact with genetic predisposition and influence lifespan.

Such uses in veterinary research also include machine learning for the prediction of disease outcomes, such as work done by Smith et al. [5] to predict Labrador retrievers that would develop hip dysplasia using machine learning. However, research incorporating non-genetic attributes for lifespan prediction remains scarce, a gap that this study seeks to address.

B. Key Theories or Concepts

The theories regarding animal longevity commonly focus on genetic and biological factors, but this research mainly focuses on economic and behavioral attributes like score, score for kids, size and lifetime cost. These attributes will act as better predictors of longevity when coupled with certain advanced models in machine learning. Applications of such models as Random Forest and Support Vector Regression are assumed to handle nonlinear relationships between attributes and lifespans more effectively [3].

C. Gaps or Controversies in the Literature

Important gap in the literature lies in the lack of studies that combine non-genetic factors with machine learning models to predict lifespan. While most of the studies are related to genetic predisposition, other non-genetic attributes, such as score, score for kids, size and lifetime cost, are usually ignored, though they likely have influential effects on the quality of life of a dog and its general longevity [2] [3].

V. Methodology

A. Research Design

This research involves a machine learning predictive modeling approach to estimate, using a number of attributes included, how long dog breeds may live in years. Some of the very important predictors include score, size, score for kids and lifetime cost.

B. Data Collection Methods

For this study a data set of 87 dog breeds have been used from the source of 'Kaggle' with features related to type, score, popularity ranking, size, intelligence, score for kids, etc. Data prepossessing techniques such as data cleaning, data transformation and data reduction were implied to the dataset to prepare it for building a predictive modeling approach [2].

C. Sample Selection

The study sample included multiple dog breeds, ensuring a broad range of sizes and care-related variables. In the study, there were some mainly focused variables which were:

- longevity(years) as the target variable.
- score, score for kids, size and lifetime cost as predictor variables.

D. Data Analysis Techniques

The data was prepossessed by normalizing the numerical features using Min-Max scaling, and even created polynomial features along with interaction terms to capture non-linear relationships. Various machine learning models were then applied to these including:

1) Linear Regression
2) Decision Tree Regressor
3) Random Forest Regressor
4) Support Vector Regressor
5) XGBoost Regressor

All these models were then evaluated by Mean Squared Error and R-squared values. The results of cross-validation showed that indeed these models had performed well and also Hyper-parameter tuning was performed using GridSearchCV. SVR was selected as the best-performing model, capturing complex relationships between the attributes [3].

VI. Results

A. Presentation of Findings

The following table summarizes the final performance of each model on the test set.

TABLE I
PERFORMANCE OF MODELS

Model	Mean Squared Error(MSE)	R-Squared
Random Forest	0.8298	0.7926
SVR	0.3928	0.9018
XGBoost	1.1355	0.7161

B. Data Analysis and Interpretaion

The SVR model performed the best, with an MSE of 0.3928 and an R-squared of 0.9018. This suggests that the model can explain 92% of the variance in the predicted longevity of dog breeds, outperforming other models such as Random Forest and XGBoost. The analysis indicates that size, lifetime cost and the overall score were the most significant predictors of longevity.

C. Support for Research Questions or Hypothesis

The results support the research questions that machine learning models, such as Support Vector can effectively predict the longevity of dog breeds and size, lifetime cost and overall score are the most influential attributes that helped to predict the dog's lifespan according to this study.

VII. Discussion

A. Interpretation of Results

The best-performing model, Support Vector Regression (SVR), effectively handled the non-linear relationship between breed attributes and lifespan. These findings confirm the results from Urfer et al. [3], who highlighted that environmental factors like lifetime cost significantly affect a dog's lifespan.

B. Comparison with Existing Literature

The findings corroborate previous research, such as Yordy et al. [1], which established the relationship between size and longevity, while also expanding on this by incorporating economic and behavioral factors. This study also complements the work of Socha et al. [2], who noted the importance of lifetime cost and intelligence scores in lifespan prediction.

C. Implications and Limitations of the study

While the study shows that non-genetic factors can significantly predict dog breed longevity, the dataset's size and the exclusion of genetic factors limit the model's predicting performance. Further research incorporating genetic data alongside environmental and economic factors would likely yield more robust predictions.

VIII. Conclusion

A. Summary of key findings

This study demonstrated that dog breed longevity can be predicted using machine learning models based on non-genetic factors. The SVR model emerged as the most accurate, achieving an MSE of 0.3928 and an R-squared of 0.9018.

B. Contributions to the filed

This research contributes to the growing field of pet care analytics by introducing a framework for predicting the lifespan of dog breeds based on non-genetic attributes. It provides a valuable tool for veterinarians, pet owners, and breeders to make informed decisions.

C. Recommendations for Future Research

Future work should focus on expanding the dataset and incorporating additional factors, such as diet, exercise, genetic ailments and congenital ailments to improve model accuracy and applicability.

References

[1] J. Yordy, C. Kraus, J. J. Hayward, M. E. White, L. M. Shannon, K. E. Creevy, D. E. Promislow, and A. R. Boyko, "Body size, inbreeding, and lifespan in domestic dogs," *Conservation genetics*, vol. 21, pp. 137–148, 2020.

[2] S. Socha, M. MiroĹska, and D. KoĹodziejczyk, "Analysis of factors affecting the quality and length of life of dogs," *Acta Sci. Pol. Zootechnica*, vol. 21, no. 3, pp. 3–12, 2022.

[3] S. R. Urfer, M. Wang, M. Yang, E. M. Lund, and S. L. Lefebvre, "Risk factors associated with lifespan in pet dogs evaluated in primary care veterinary hospitals," *Journal of the American Animal Hospital Association*, vol. 55, no. 3, pp. 130–137, 2019.

[4] V. Adams, K. Evans, J. Sampson, and J. Wood, "Methods and mortality results of a health survey of purebred dogs in the uk," *The Journal of small animal practice*, vol. 51, pp. 512–24, 10 2010.

[5] A. L. Smith, R. M. A. Packer, and S. M. Caney, "Predicting the development of canine hip dysplasia in a labrador retriever population using machine learning," *Journal of Veterinary Internal Medicine*, vol. 29, no. 6, pp. 1387–1394, 2015.