ADVANCED CANINE ANALYTICS - *OPTIMIZING DOG HEALTH AND SERVICE SUITABILITY WITH MACHINE LEARNING*

[Project Overview and Methodology 2](#_Toc166103380)

[Synthetic Data Generation and Data Analysis 3](#_Toc166103381)

[USE CASE 1 (Regression): Regression Models Evaluations and Weight Predictions 6](#_Toc166103382)

[USE CASE 1 (Regression): Best Fit Model Analysis and Predictions 7](#_Toc166103383)

[Use Case 2 (Classification): Dog Service Rating classifier 10](#_Toc166103384)

[Closing Remarks 12](#_Toc166103385)

ADVANCED CANINE ANALYTICS

*…OPTIMIZING DOG HEALTH AND SERVICE SUITABILITY WITH MACHINE LEARNING*

# Project Overview and Methodology

introduction

The intersection of machine learning and canine care provides me with a unique opportunity to address real-world problems that affect pet owners and professionals in the veterinary and service sectors.

My project focuses on two primary challenges: predicting a dog’s weight based on various features and classifying dogs as suitable service animals. The rationale behind choosing two distinct use cases—**one focusing on regression and the other on classification**—is to demonstrate the comprehensive capability of machine learning. This approach highlights how ML can handle both quantitative data for precise measurements and categorical data for decision-making, effectively showcasing the versatility of ML in solving varied and complex problems in canine care and service optimization.

key machine learning prototypes

To address the challenges in canine care, I’ve crafted two machine learning prototypes:

* **Regression Models Evaluations (DogWeightRegressionEvaluator):** This evaluates several regression models (linear, gradient boosting, logistic and Random Forest) to determine the best fit model and further extends the model to predict dog weight from attributes such as breed.
* **Random Forest Classifier (DogServiceRatingClassifier)**: This classifier determines a dog’s suitability as a service animal such as Seeing Eye, Police K9, Search and Rescue, Military, Assistance, Emotional Support,

synthetic data generation

Given the scarcity of specific datasets for this nuanced analysis, I’ve turned to synthetic data generation. Real-world data from dog owners, shelters, veterinary records, and training centers is essential for refining these prototypes and better reflecting practical scenarios. However, while awaiting such data, I’ve used synthetic data to test and validate my prototype for my college project submission. This approach gives me full control over feature distribution, enabling the modeling of rare but crucial scenarios and ensuring robustness.

* **Synthetic Data Generation (DogSynDataGenerator):** I have designed this tool to generate synthetic data that closely mimics real-world canine datasets.

# Synthetic Data Generation and Data Analysis

This tool generates synthetic data that closely mimics real-world canine datasets.

Data generation rules

The following details the high-level data generation model simulation

1. Generate Dog Breeds
   1. Small Breeds (5): B-S1, B-S2, B-S3, B-S4, B-S5
   2. Medium Breeds (5): B-M1, B-M2, B-M3, B-M4, B-M5
   3. Large Breeds (9): B-L1, B-L2, B-L3, B-L4, B-L5, B-L6, B-L7, B-L8, B-L9
2. Generate age within the specified range.
   1. All Breeds: 1 to 10 years
3. Calculate weight as a function of age, where weight increases upto age of 3 years and after that the fluctuations are minimal. The weights will be in these ranges:
   1. Small Breed: 10 – 30 lbs
   2. Medium Breed: 31 – 60 lbs
   3. Large Breed: 61 – 100 lbs
4. Calculate height as a function of age with little to no fluctuations after 2 years. The heights will be in these ranges:
   1. Small Breed: 5 – 10 inches
   2. Medium Breed: 11 – 20 inches
   3. Large Breed: 20 – 30 inches
5. Assign service suitability service ratings based on breed size:
   1. Small Breed: Set high ratings for 'Emotional Support' and 'Family'. Set very low ratings for other services.
   2. Medium Breed: Set moderate to high ratings for 'Emotional Support' and 'Family' and average for the rest.
   3. Large Breed: Distribute ratings more evenly across all services
6. Adjust ratings randomly to add variability and simulate real-life anomalies.
7. Variable data volumes are generated for each breed to ensure a diverse and representative sample.

output

The Output is a csv file: Synthetic\_dog\_data.csv

data analysis and visualizations

The following details the high-level data generation model simulation

**D O G S Y N T H E T I C D A T A**

A table with numbers and numbers

Description automatically generated

A graph of a graph

Description automatically generated

A group of different types of service rating with Salt Lake Temple in the background

Description automatically generated with medium confidence

A group of scatter plots

Description automatically generated

synthetic data analysis

This section provides a series of reports on the synthetic data characteristics, with each plot and statistic offering insights into the overall health and service distribution of the dog population.

Data Summary and Statistics

* Synthetic dog data covers weight, height, and age metrics, exhibiting a normal distribution, indicating a healthy dog population.
* Average weight aligns with expectations for domestic dogs, with low standard deviation implying weight consistency.

Distribution Insights

* Histograms illustrate normal distribution of age, weight, and height, ensuring data quality and reliability.
* Most dogs fall within medium to large size categories, with a minority in very small or extra-large categories.

Service Type Analysis

* Bar charts visualize service type ratings, revealing varied performance across different dog services.
* Some breeds excel in companionship roles like "Emotional Support" and "Family" services.

3D Scatter Plots of Service Type Rating

* The 3D scatter plots showcase data distribution by service types, aiding in identifying clusters and outliers.
* Service types exhibit clustering, with no significant correlation between size and service rating.

In conclusion, synthetic data illuminates the healthy diversity of the dog population, revealing its versatile capabilities and unraveling the complexities of service suitability.

# USE CASE 1 (Regression): Regression Models Evaluations and Weight Predictions

Objective

Evaluate dog weight using regression models based on features like breed. Uncover insights into factors influencing dog weight. Provide recommendations for weight prediction.

Data preparation

The dataset containing synthetic dog data is loaded and divided into training and testing sets. This dataset comprises information on dog breeds, age, and weight, essential for model training.

regression model evaluation and best fit selection

Four regression models, namely Linear Regression, Decision Tree, Gradient Boosting, and Random Forest, are assessed for their predictive performance. These models utilize features such as dog breed and age to predict dog weight. Evaluation metrics such as Mean Squared Error (MSE), R-squared (R2), Mean Absolute Error (MAE), and Explained Variance Score (EVS) are utilized for model assessment. Sklearn library is utilized for model implementation.

The model exhibiting the highest performance across evaluation metrics is selected as the best fit.

regression models evaluation metrics

1. MSE (Mean Squared Error): High MSE indicates that the model predictions are generally far from the actual values. High MSE suggests either model inadequacy, presence of outliers, or excessive variance in the data.
2. R2 (R-squared): A high R-squared value close to 1 indicates that a large proportion of the variance in the dependent variable has been explained by the independent variables in the model.
3. MAE (Mean Absolute Error): Like MSE, but less sensitive to large errors (outliers). High values suggest that the model may have systematic errors.
4. EVS (Explained Variance Score): If EVS is significantly lower than R², it could indicate an issue with the model’s bias or an occurrence of residuals that are not uniformly distributed.

**R E G R E S S I O N M O D E L S C O M P A R I S I O N S & B E S T F I T**

The recommended model based on the composite score is: **Gradient Boosting**

A green and grey bar with numbers

Description automatically generated

A screenshot of a graph

Description automatically generated

# 

# USE CASE 1 (Regression): Best Fit Model Analysis and Predictions

analysis of best fit model – gradient boosting

Feature Importance:

An analysis of feature importance is conducted to discern the relative significance of features such as dog breed and age in predicting dog weight. This analysis aids in identifying pivotal predictors for weight estimation. This indicates that both the age of the dog and its breed size play critical roles in determining its weight, with age being the most significant predictor.

Correlation Analysis:

A correlation heatmap is generated to visualize relationships between features and the target variable (weight). This analysis assists in identifying potential correlations and understanding underlying patterns in the data. Certain breed categories (e.g., B\_L4, B\_L5) exhibit high correlations not only with weight but also with each other, it could signal multicollinearity. This would imply that these breed variables might contribute redundant information, which could skew the model’s performance and interpretability.

Scatter Plots:

Scatter plots are employed to visualize relationships between individual features and dog weight. These plots provide insights into the distribution and correlation of features with the target variable.

V I S U A L S F O R B E S T-F I T M O D E L - *G R A D I E N T B O O S T I N G*

**FEATURE IMPORTANCE**

A graph of different colored bars

Description automatically generated with medium confidence

**CORRELATION HEATMAP**

A graph of heatmap

Description automatically generated

**SCATTER PLOTS - FEATURES VS WEIGHTS (LBS)**

A screenshot of a graph

Description automatically generated

model predictions - weight progression by breed

Weight progression of dogs is visualized by breed, showcasing variations in weight with age for different breeds. This analysis aids in understanding growth patterns of various dog breeds over time.

M O D E L P R E D I C T I O N S - W E I G H T P R O G R E S S I O N B Y B R E E D

A graph of different colored lines

Description automatically generated

conclusion

The Gradient Boosting model showed strong predictive capabilities, supported by insights from feature importance, scatter plots, and correlation heatmaps. The results, including weight progression by breed and age, present a promising commercial opportunity. However, refinement opportunities exist:

* Explore strategies to reduce Mean Squared Error.
* Utilize techniques to address high multicollinearity.
* Adjust model complexity and consider additional features for improvement.
* Fine-tune parameters for optimal performance.

# Use Case 2 (Classification): Dog Service Rating classifier

objective

This prototype is designed to predict the suitability of different dog breeds for specific service roles based on various features. This report presents the evaluation and visualization of the classifier's performance using regression models and other analytical techniques.

Model Comparison and Best Fit:

The Dog Service Rating Classifier utilizes a Random Forest Classifier to predict the suitability of dog breeds for specific service roles.

Visual Insights for Best Fit Model:

Feature Importance:

The importance of each feature in predicting service suitability is visualized through a bar plot, providing insights into the factors influencing the classifier's decisions. The model's feature importance graph indicates that the most influential features for predicting suitability for family settings are primarily identifiers for large breeds (e.g., B\_L2, B\_L4, B\_L6, B\_L5, B\_L9), with importance values decreasing notably as features shift from large to medium and small breed identifiers. This trend suggests a potential bias in the model toward characteristics more common in larger breeds.

Confusion Matrix:

A confusion matrix is presented to analyze the classifier's performance in terms of true positives, false positives, true negatives, and false negatives. This aids in understanding the model's predictive accuracy and potential errors. The model demonstrates high accuracy and effectiveness in its predictive capability, evidenced by a substantial number of true positives and true negatives, alongside minimal false positives and false negatives. This low error rate provides a strong basis for further refinement, potentially through tuning threshold values or additional feature engineering to further reduce misclassification.

ROC Curve:

The ROC curve illustrates the trade-off between sensitivity and specificity of the classifier, providing a comprehensive view of its performance across different thresholds. The ROC curve for the model, with an AUC of 0.82, showcases good predictive performance and a balanced trade-off between sensitivity and specificity, effectively discriminating between suitable and not suitable cases for family settings. The shape of the ROC curve indicates potential for further optimization by adjusting the classification threshold to enhance either sensitivity or specificity based on specific requirements.

D O G S E R V I C E C L A S S I F I C A T I O N - R A N D O M F O R E S T C L A S S I F I E R

A graph of different colored bars

Description automatically generated

A close-up of a guide

Description automatically generatedA diagram of a diagram

Description automatically generated with medium confidence

A graph with a green line

Description automatically generated

service suitability ranking table

**Suitability Ranking Table:** A heatmap visualizes the suitability rankings of different dog breeds for the specific service role, offering a clear comparison and ranking based on their predicted ratings.

A chart of dog breeds

Description automatically generated

The Dog Service Rating Classifier demonstrates strong predictive capabilities in determining the suitability of dog breeds for specific service roles. Through model comparison and visual analysis, it provides valuable insights for decision-making in service selection. Further refinement and optimization can enhance its performance and applicability in diverse service scenarios.

# Closing Remarks

lessons learned

Reflecting on this journey, I've realized the essence of thorough model evaluation. It wasn't just about accuracy; delving into feature importance, correlations, heatmaps strengthened my understanding of various regression models.

1. I began by exploring various project options, which led me to study neural network models, gaining substantial insights into their structures and applications.
2. I learned to effectively manage both quantitative and categorical data, enhancing my appreciation for machine learning's adaptability across different data types.
3. Generating synthetic data that closely resembles real-world scenarios proved challenging. I encountered multiple obstacles, some of which I couldn't overcome, as detailed in the specific use cases.
4. My experience with iterative model development, especially the setbacks from misrepresented synthetic data, taught me that continuous improvement in machine learning is not always a linear process.
5. I discovered that prototyping is crucial for identifying and understanding the key drivers within models, underscoring its importance in the early stages of model development.

future expansion and commercial viability

Further exploration and potential expansions, these prototypes could be adapted for real-world applications.

* Future Expansions:
  + Behavioral Insights: Incorporate behavioral data to refine suitability assessments.
  + Lifestyle Matching: Customize models to account for owner lifestyle, optimizing dog-owner pairing. Extended Applications:
* Extended Applications:
  + Veterinary Systems Integration: Employ regression models for proactive dog weight management.
  + Service Dog Certification: Utilize classifiers to streamline the certification of service dogs.