**Sornsawan Permpipat 300386177**

**Project: Forecasting Dog Adoption**

**Business Domain: Sonoma County Animal Shelter**

The Sonoma County Animal Shelter plays a pivotal role in managing the intake, care, and placement of animals, including dogs, in the community. The shelter's core mission is to increase adoption rates and optimize the allocation of resources to improve animal welfare. With a focus on efficiently matching animals with potential adopters, the shelter aims to create a better environment for its animals and the people who choose to adopt them.

The primary goal of this analysis is to identify the factors that influence the likelihood of a dog being adopted or experiencing other outcomes such as return to owner, or transfer. Addressing these questions will provide valuable insights that can enhance the shelter’s operations and decision-making process. Based on existing knowledge, the following hypotheses were developed to guide this analysis:

**Hypothesis 1:** Younger dogs and certain breeds are more likely to be adopted compared to older dogs or less popular breeds.

**Hypothesis 2:** Dogs with healthy or treatable conditions have a higher chance of being adopted compared to those with severe health issues or behavioral problems.

**Hypothesis 3:** Size and color of the dog might influence adoption rates, with smaller or more visually appealing dogs potentially being adopted faster than larger or less typical-colored dogs.

# **Data Preparation**

## **Data Inventory**

This dataset, sourced from the Sonoma County government portal, includes information on animal intakes and outcomes, such as adoption, transfer, and other status types [[Links](https://data.sonomacounty.ca.gov/Government/Animal-Shelter-Intake-and-Outcome/924a-vesw/about_data)]. For the analysis, I filtered the data to include only animals of the dog type, ensuring that the focus is on predicting dog adoptions specifically. The dataset contains 13,640 records and 23 columns, each representing a unique animal intake. As shown in Figure 1, most of the features are of object type, with some missing values.

A screenshot of a computer

Description automatically generatedThe dataset contains several columns that provide detailed information about the animals in the shelter. These include the animal's Name, Type, Breed, and Color, which describe the basic characteristics of the animal. The Sex column indicates whether the animal is male, female, neutered male, or spayed female, and the Size column specifies whether the animal is large, medium, small, or toy sized. The Date of Birth provides an approximate birth date, and the Impound Number and Kennel Number correspond to the animal’s specific shelter location. Additional columns track the intake and outcome of the animal, including the Animal ID, a unique identifier for each animal, and the Intake Date and Outcome Date, which mark when the animal entered and left the shelter. The Days in Shelter column calculates the duration the animal spent in the shelter. The Intake Type and Outcome Type columns explain the reasons for the animal's entry and release, with corresponding subtypes giving more specific details. The dataset also includes columns about the animal's condition at intake and release (Intake Condition and Outcome Condition), as well as the jurisdictions involved in the intake and release processes (Intake Jurisdiction and Outcome Jurisdiction). Finally, the Outcome Zip Code indicates the zip code of the area where the animal was released, and the Location column contains the latitude and longitude coordinates for the outcome jurisdiction.

Figure Dataset Features

**Data Processing**

Figure 2 provides a summary of the raw data, indicating that, on average, dogs stay in shelters for 21 days. In addition, Figure 3 displays a preview of the dataset.

A table with numbers and text

Description automatically generated

Figure 2 Summary Statistic on Raw Data

A screenshot of a computer

Description automatically generated

Figure 3 Preview of Dataset

I performed several steps to prepare the data for analysis. First, I cleaned the data by removing rows with missing values and dropping columns containing unique identifiers, as these did not provide meaningful insights into the dataset’s patterns. For feature engineering, I introduced a new "Age" feature by calculating the difference between the Date of Birth and Outcome Date. Since the Age variable was skewed, I applied a log transformation to make its distribution more suitable for analysis. Similarly, for the Days in Shelter variable, I applied a square root transformation, as this provided a better fit, as shown in Figure 4.

During encoding, I applied one-hot encoding to categorical variables and used frequency encoding for the Breed column to assign numerical values based on breed popularity. For the Sex column, I created two new features: "IsSterilization" to capture sterilization status and "Group\_Sex" to categorize animals by sex. Additionally, in the Color column, which originally had 205 distinct values, I grouped these into four main categories: Monochrome, Brown, Brindle/Patterned/Tricolor, and Other, to simplify the data and highlight key patterns.

Because the target variable (Outcome\_Type) had four popular values and several rare ones, I grouped the rare outcome types into a single category called "Other" to make the analysis more focused and improve model performance. After these steps, the processed dataset contained 13,505 entries and 20 columns, each consisting of either integer or float data types. Figure 5 provides an overview of the cleaned dataset.

|  |  |
| --- | --- |
| A group of graphs showing different sizes of data  Description automatically generated with medium confidence  Figure 4 Comparison of Log and Square Root-Transformed Age and Days in Shelters | A screenshot of a computer program  Description automatically generated  Figure 5 Cleaned Dataset |

**Model Implementation**

For feature selection, I employed three methods to optimize the model’s performance. The first method, All Features (passthrough), keeps all available features without any selection, providing a baseline for comparison. The second method, SelectKBest, selects the top 10 features based on their relevance to the target variable, measured using the f\_classif statistical test, which is effective for identifying features that have a strong relationship with the outcome. Lastly, the VarianceThreshold method removes features with low variance, specifically those with a variance below 0.2, as these are unlikely to contribute meaningful information to the model.

For this analysis, I propose using a combination of classification models to predict the likelihood of dog adoption, return to owner, transfer, or other outcomes. The rationale for selecting these models is based on the nature of the problem (classification) and the type of data I have. Given that the dataset includes both numerical and categorical features, and that the target variable is categorical (outcome status). Below is an overview of the models and their suitability for this task:

1. Logistic Regression: It is a foundational model for classification tasks, extended here to multinomial logistic regression for multi-class prediction. It was selected for its interpretability and simplicity, making it ideal for understanding how dog characteristics relate to adoption outcomes.
2. K-Nearest Neighbors (KNN): Can captures non-linear relationships by comparing instances to their nearest neighbors. This makes it suitable for datasets where adoption outcomes depend on interacting factors like breed, age, and health condition. Feature scaling was applied to ensure optimal performance, given KNN's sensitivity to feature magnitude.
3. Support Vector Machine (SVM): Both linear and RBF kernels of SVM were implemented. SVM's ability to work effectively in high-dimensional spaces was leveraged, particularly for capturing complex adoption patterns.
4. Decision Tree: Due to its ability to model non-linear relationships and provide interpretable decision rules. With a constrained depth, it avoided overfitting while effectively capturing key feature interactions.
5. Naive Bayes: Due to its simplicity and efficiency. It assumes feature independence, making it computationally efficient and a useful baseline for comparison.
6. Random Forest: Due to its robustness to overfitting and ability to process both categorical and numerical data make it a strong contender for this dataset.
7. XGBoost: it is well-suited for handling complex datasets that contain both numerical and categorical features. Its advanced gradient boosting approach optimizes predictive performance more effectively than many other algorithms, making it a powerful choice for this dataset.

**Model Workflow**

Since the dataset has both numerical and categories features, using several techniques instead of relying on just one model is a better approach. First, I split the training and testing as 80:20, then, feature selection using three methods: 'All Features', 'SelectKBest', and 'VarianceThreshold'. Next, eight machine learning models are tested: Logistic Regression, K-Nearest Neighbors, Linear SVM, RBF SVM, Decision Tree, Naive Bayes, Random Forest, and XGBoost. For each model, a pipeline is created that includes the selected features, data scaling for consistency, and the model itself. Each model is trained and tested, and their results are recorded to find the best one.

After identifying the best-performing model, hyperparameter tuning is done using GridSearchCV. This process fine-tunes the model by testing different settings, such as tree depth, learning speed, and data sampling rate, to improve its accuracy. The tuned model is then tested again to confirm its performance.

Finally, the best model is evaluated using measures accuracy and detailed metrics such as precision, recall, and F1 score.

**Results Interpretation and Implications**

Figure 6 shows the best model pipeline found using GridSearchCV, the pipeline has three main parts: passthrough (All features), RobustScaler, and XGBClassifier. The passthrough step means no extra changes were made to the data before scaling, as this was the best choice for the model. Next, the RobustScaler adjusts the dataset, this helps reduce the effect of extreme values (outliers) on the model. Finally, the last step is the XGBClassifier. Figure 7 shows the F1-score for the model, which balances precision and recall, also highlights good performance. For example, the F1-scores for class 2 were high at 92% for precision and 93% for recall, meaning the model performed very well in identifying most of the instances for this class. However, for class 3, the F1-score is lower due to the low recall of 41%, meaning the model missed many of the instances of this class. Additionally, the model shows an F1-score of 0.83, which is a good indication of the model’s balance between precision and recall across all classes. while the model is generally performing well, there is room for improvement in handling underrepresented classes, particularly class 3. This could include adding more data, especially for the underperforming class to address class imbalance. Overall, the model meets the goal of achieving a decent accuracy (83%), but further fine-tuning, additional data, and possibly different techniques are needed to improve performance, especially for the less frequent classes.

|  |  |
| --- | --- |
| A screenshot of a computer  Description automatically generated  Figure 6 Best Model Pipeline | Figure 7 Accuracy, Precision, Recall, and F1 score |

Figure 8 shows that the model performs well overall, with high accuracy in predicting ADOPTION and RETURN TO OWNER. However, it struggles with the TRANSFER class, which has the most misclassifications, especially predicted as ADOPTION or OTHER. While RETURN TO OWNER also has some misclassifications, TRANSFER and OTHER is the most challenging for the model.

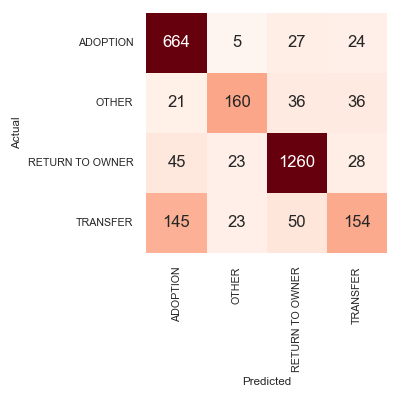


Figure 8 Confusion Matrix

The report initially hypothesized three key factors influencing dog adoption rates. Hypothesis 1 proposed that younger dogs and specific breeds are more likely to be adopted. Figure 9 shows the data partially supports this, showing that age (Age\_log feature importance: 0.022526) and breed popularity (Breed\_Frequency feature importance: 0.017340) play minor roles in adoption. Hypothesis 2 posited that healthier or treatable conditions would increase adoption likelihood. This is strongly supported by the data shown in Figure 9, with Intake\_Condition\_UNTREATABLE (feature importance: 0.360463) being the most influential factor, while features like HEALTHY (0.056963) and TREATABLE/MANAGEABLE (0.022941) reinforce the importance of health. Lastly, Hypothesis 3 suggested that size and color might influence adoption rates, though the data shows limited support. Figure 9 shows Size\_PUPPY (0.061436) indicates some preference for smaller dogs, but color features have minimal impact. Overall, health condition emerges as the most significant factor, while other attributes, including age, breed, size, and color, play lesser roles.

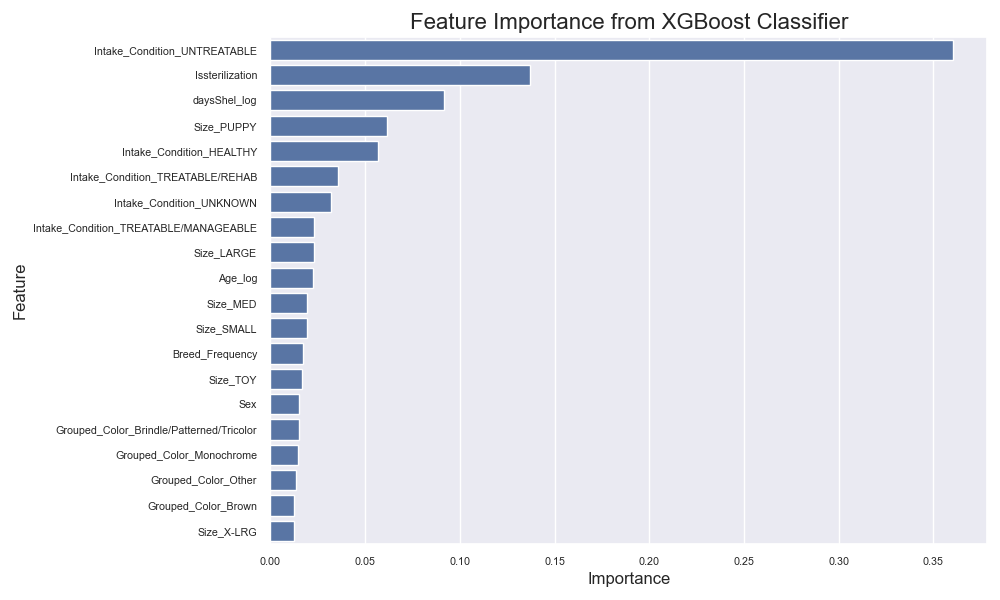
****

Figure 9 Features Important XGBoost Classifier

**Out-of-Sample Predictions**

First, I used X\_test to create a summarized description (using 50% values) that represented a typical test sample, capturing the central tendencies for relevant dog features. This summarized row formed the basis of an out-of-sample prediction by creating a new data frame containing these descriptive values.

After running the model on the new data frame, the model predicted that the representative dog profile belonged to the “ADOPTION” class (class label [0]) with a probability of 0.76. This probability means that, according to the model, this specific dog (a Parson Russell Terrier mix) has a 76% likelihood of being adopted. This insight can guide shelter staff, helping them understand which characteristics are associated with a higher probability of adoption. It also suggests that dogs with profiles similar to this one may have better adoption outcomes, allowing shelters to focus on such attributes to increase adoption rates.

**Conclusion**

In this analysis, I explored the process of using machine learning to predict the outcome of dog adoptions in a shelter. The workflow involved several key steps, including data preprocessing, feature selection, model training, and evaluation. Initially, I applied feature selection methods to reduce dimensionality and improve the model’s performance. I tested eight machine learning models: Logistic Regression, K-Nearest Neighbors (KNN), Linear Support Vector Machine (SVM), Radial Basis Function (RBF) SVM, Decision Tree, Naive Bayes, Random Forest, and XGBoost. To optimize the models, I fine-tuned their hyperparameters using GridSearchCV, which helps in selecting the best combination of parameters by evaluating different options based on accuracy. This model tuning process was crucial in enhancing performance and improving prediction accuracy. After training, the model was evaluated using cross-validation and test data to assess its ability to predict various adoption outcomes, such as ADOPTION, TRANSFER, RETURN TO OWNER, and OTHER.

The main insights from the model show that the animal's condition when it first enters the shelter is the most important factor. For example, whether the animal is healthy or untreatable can have a big impact on the model's prediction. The sterilization status, or whether the animal has been sterilized, is also an important factor. The number of days the animal has spent in the shelter (daysShel\_log) is another key feature, suggesting that the longer the animal stays in the shelter, the more it affects the prediction. Lastly, the animal's size and age, like whether it’s a puppy or a large size, also matter but are not as important as the condition at intake or sterilization status.

From a business or managerial perspective, the key implication of these findings is that shelters can prioritize efforts to improve adoption rates by focusing on factors that influence adoption outcomes. For example, by reducing the shelter stay duration (e.g., through targeted adoption campaigns or partnerships), promoting sterilization programs, and addressing health issues early, shelters can increase the likelihood of adoption and reduce the need for euthanasia. Additionally, understanding the importance of certain breeds and age of the dogs can help to allocate resources more effectively, ensuring that dogs with higher adoption potential are highlighted in adoption events. This predictive modeling approach offers actionable insights that can drive operational decisions in animal shelters, ultimately improving adoption rates and outcomes for the dogs.

**Video presentation**

Here are two links you can choose from; either one will work. I included both to ensure the content always remains accessible.

YouTube: <https://www.youtube.com/watch?v=iyMTn71Ky10>

Google Drive: <https://drive.google.com/file/d/1lh3oFKsWyBHlQ2gyqj_F7DpC89fPCPG7/view?usp=sharing>