|  | **Member 1** | **Member 2** | **Member 3** | **Member 4** | **Member 5** |
| --- | --- | --- | --- | --- | --- |
| **SID:** | 520345357 | 540816688 | 540444920 |  |  |
| **Unikey:** | yyao5617 | open0051 | yeng0449 |  |  |
| **Project Name:** | OptiPaw: Data-driven Outcomes for Shelter Animals | | | | |
| **Mentor** | Lijia Yu | | | | |

## Abstract / introduction:

*This section should contain a brief overview of your proposed project and be 150 words max. For instance:*

* *Try to explain your research project to a non-expert, and argue why it is important*
* *What are the fields/domains your project investigates?*

Every year, 7.6 million companion animals enter US shelters. Many of these animals are surrendered by their owners, lost, or rescued from cruelty. Currently, there are no existing tools to help shelters to understand trends in animal outcomes.

In order to predict the outcome for shelter animals, we propose a tool called OptiPaw, it aims to predict whether shelter animals will be adopted, transferred, euthanized, or passed away by analyzing factors such as breed, age, and color. Additionally, OptiPaw examines popular pet names and how human preferences for specific animal characteristics influence adoption trends. The project leverages data science techniques, including data visualizations and machine learning models, to provide insights that help veterinarians in shelters allocate care more efficiently, prioritize resources for high-risk animals, and tailor adoption strategies based on community preferences.

Other than the exploration of data science, the project also investigates other fields: veterinary sciences by analyzing factors like breed, age, and color; probability and statistics by exploring how fate and likelihood influence outcomes; psychology by examining human preferences and behaviors in pet adoption; and ethical concerns such as genetic modification, animal cruelty, and adoption policies. Additionally, OptiPaw supports the "Adopt, don’t shop" movement by advocating for friendlier practices in pet adoption.

## Project Aims:

* *What do you hope your project will achieve/discover?*
* *What are the brief tools you will be using (i.e. coding languages, datasets if you already know what you will use, IDE, etc.)?*

Through Optipaw, we hope to achieve several key outcomes:

1. Predict the outcomes for shelter animals, including whether they will be adopted, transferred, euthanized, or pass away, based on factors such as breed, age, and color, using machine learning techniques. We can also predict the names of the animals based on attributes for fun!
2. Analyze trends in popular pet names through data visualization and how human preferences for specific animal characteristics affect adoption outcomes. This will offer adoption centers valuable insights, enabling them to develop more effective strategies to provide care and attention to animals that are less likely to be adopted.

Our final aim is to develop an interactive Shiny app that incorporates all our visualizations and predictions. Additionally, we will design an engaging questionnaire that gathers information about pets' attributes, enabling users to explore potential outcomes for their pets in an adoption scenario.

The datasets we will be using are:

1. Austin, US (2018-2022) - contains intakes.csv and outcomes.csv: <https://data.world/siyeh/austin-animal-center-live-data> (as test dataset)
2. Indiana, US (2017-2020): <https://data.world/city-of-bloomington/94d3f457-57b5-45be-bee0-a0106f59b7ed> (as training dataset)
3. California, US (2017-2024): <https://data.longbeach.gov/explore/dataset/animal-shelter-intakes-and-outcomes/export/?disjunctive.animal_type&disjunctive.primary_color&disjunctive.sex&disjunctive.intake_cond&disjunctive.intake_type&disjunctive.reason&disjunctive.outcome_type&disjunctive.outcome_subtype&disjunctive.intake_is_dead&disjunctive.outcome_is_dead> (as training dataset)

Specifically, for machine learning, we will be training our prediction models using datasets 2 and 3, and testing on dataset 1. Dataset 1 will be tested upon combining intakes.csv and outcomes.csv and stripping the OutcomeType attribute; the OutcomeType attribute will therefore be used to validate how accurate our models are - similar to a mark scheme. For data visualizations, we will combine and use all datasets 1, 2, and 3.

The tools we will be using are:

1. RStudio with pandas and visdat for data cleaning and checking data quality, and tidyverse, ggplot, gganimate, plotly, and visdat libraries for exploratory data analysis (EDA) and data visualization.
2. Python coded in Google Colab for machine learning predictions. Machine learning models including logistic regression, ridge regression, random forests, and neural networks, with evaluation metrics such as log loss, ROC curves, and F1 scores to identify the best-performing model.

## Project Timeframe

*This section should contain a brief overview of the timeline for your proposed project. Please come up with this together with your mentor.*

| Week | Agenda |
| --- | --- |
| 5 | * Teammates familiarise with potential datasets from kaggle competition * Teammates each learn and present meaning of code assigned * Lijia introduces suitable deep learning model |
| 6 | * Run assigned code * Teammates find usable data set and topic * Explored a few options (Mode of action, mushrooms - kaggle competitions) * Conduct EDA to justify any new directions |
| 7 | * Confirm data set and topic - shelter animal outcome project * Produce Progress Report 1 & Project Proposal * Learn new code assigned * Data cleaning/checking for data quality issues, finalize EDA |
| 8 | * Check and confirm final visualizations * Attempt prediction models studied |
| 9 | * Prediction model comparisons * Choose best model |
| 10 | * Progress Report 2 * Build audience questionnaire and shiny app |
| 11 | * Build and test shiny app * Make slides for presentation |
| 12 | * Presentation Night |

## Proposed Methodologies

*Include tools, skills, programs and packages that you will expect to use. Also include sources of information / data.*

**Work Package 1 - Exploratory Data Analysis and Data Visualization**

In this section, we aim to explore three datasets by analyzing and visualizing the data to understand its key characteristics, uncover patterns, detect outliers, and identify relationships between variables. Methods used may include clustering and dimensionality reduction techniques, which help create graphical representations of high-dimensional data with many variables. For visualization, we will include:  
  
(1) Radar Chart: This visualization will provide an overview of how strongly each independent variable (such as breed, age, and intake condition) influences the outcome type (adoption, transfer, or euthanasia) of shelter animals. This chart helps identify which factors play the most significant role in determining an animal's outcome.

(2) Bubble Plot: This plot will explore how various intake conditions (such as stray, owner surrender, or medical intake) affect the outcome, categorized by animal type and size. The insights from this plot will help shelters refine their adoption strategies and allocate resources more efficiently based on animal type and condition.

(3) Time-Series Scatter Plot: This plot will examine the popularity of pet adoption over time, focusing on how monthly adoption rates are influenced by recurring celebratory events. By identifying trends, shelters can better plan for periods of increased or decreased adoption activity.  
  
(4) Scatter Plot: This visualization will compare the length of time an animal spends in the shelter with its eventual outcome. By understanding this relationship, shelters can refine their practices to improve overall animal care and optimize the time it takes for pets to find permanent homes.

**Work packages 2 - Predictive Modeling for Shelter Animal Outcomes**

We will propose a series of models that capture relationships between adoption outcome and pet’s characteristics. During Week 4, we studied Python machine learning models under our mentor's guidance, focusing on various algorithms and evaluation metrics to find the best model for predicting shelter animal outcomes.

Our approach will include:

(1) Principal Component Analysis (PCA) for dimensionality reduction, helping us simplify complex datasets by retaining the most significant variables.

(2) Logistic Regression (LR) for binary classification tasks, where outcomes can be categorized into two types.

(3) Ridge Regression to address issues of overfitting by penalizing model complexity.

(4) Random Forest for enhanced accuracy and stability in predictions by aggregating multiple decision trees.

(5) Neural Networks (NN) for recognizing deeper patterns and nonlinear relationships in the data.

To assess the performance of these models, we will use the following key evaluation metrics:

(1) Area Under the ROC Curve (AUC) to measure the overall accuracy of the model.

(2) Log Loss to quantify prediction error.

(3) F1 Score to balance precision and recall, ensuring that our model performs well on both positive and negative classes.

In this section, we will explore all tools that we have learned on modeling the relationships between adoption outcome and pet’s characteristics.

**Work Package 3 - Data Driven Pet Name Suggestions**

In this work package, we aim to build a model that predicts potential pet names based on the pet's characteristics, such as breed, color, and age. The model will use clustering techniques like k-means or hierarchical clustering to group pets with similar attributes, identifying common names within each cluster. By analyzing patterns in existing pet names associated with specific breeds, colors, and ages, we will generate a set of suggested names for new pets.

For any user input (such as the pet’s breed, color, and age), the model will provide a personalized selection of names that best match the pet’s unique traits. This approach will create a data-driven, user-friendly tool that makes it easier for new pet owners to select an appropriate name based on their pet’s characteristics. By applying these clustering algorithms, we can offer more personalized and relevant pet name recommendations, enhancing the overall adoption experience.

**Work Package 4 - Building a Shiny App**

Finally, we will develop a Shiny app to visualize pet data and display prediction outcomes. The app will include multiple panels, each focusing on different aspects of the data. One panel will allow users to explore pet characteristics such as breed, age, and color through interactive graphs and tables, making it easy to analyze the distribution of these traits.

Another panel will be dedicated to predicting outcomes, such as the likelihood of adoption, transfer, or euthanasia, with dynamic visualizations that display the results. Users will have the ability to interact with the data using widgets such as dropdown menus, sliders, and checkboxes. These widgets will allow users to filter data by variables such as pet type, age, or shelter intake conditions.

By integrating these interactive elements, the Shiny app will make it easier for users to navigate through the data, explore relationships between pet characteristics and outcomes, and understand model predictions in an engaging and intuitive way. This will not only enhance data analysis but also offer valuable insights to help improve shelter management practices.