|  | **Member 1** | **Member 2** | **Member 3** | **Member 4** | **Member 5** |
| --- | --- | --- | --- | --- | --- |
| **SID:** | 520345357 | 540444920 | 540816688 |  |  |
| **Unikey:** | yyao5617 | yeng0449 | open0051 |  |  |
| **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, passed away, etc. by analyzing factors such as breed, sex, age, color, and more. 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 animals traits 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 are achieving 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 are also predicting 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.

Additionally, we are designing 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 have been 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 are training our prediction models using dataset 1, splitting it by outcome date where data before or during 2023 is used for training and data after 2023 is used for testing. Dataset 1 will be tested upon combining intakes.csv and outcomes.csv and stripping the OutcomeType attribute; the OutcomeType attribute is then used to validate how accurate our models are - similar to a mark scheme. For data visualizations, we have integrated and used all datasets 1, 2, and 3.

The tools we are using are:

1. Python pandas for data cleaning and checking data quality. RStudio tidyverse, ggplot, gganimate, and plotly libraries for exploratory data analysis (EDA) and data visualizations.
2. Machine learning models are coded with Python in Google Colab. Machine learning models including logistic regression, SVM, random forests, XGBoost, kNN and neural networks. Evaluation metrics such as accuracy, precision, recall, log loss, 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. For instance:*

| 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 | * Re-cleaning of data set * Check and confirm final visualizations * Attempt prediction models studied |
| 9 | * Prediction model executions and evaluations |
| 10 | * Editing of prediction model (changing scope of data set, method of splitting) * Updated prediction model evaluations using cross validation * Build recommendation system and audience questionnaire |
| 11 | * Prediction model comparisons * Update recommendation system (debug) * 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 explored three datasets by analyzing and visualizing the data to understand its key characteristics, uncover patterns, detect outliers, and identify relationships between variables. Methods used include clustering and dimensionality reduction techniques, which help create graphical representations of high-dimensional data with many variables. For visualization, we produced:  
  
(1) Heat map: This visualization presents the percentage of unique combinations between intake type and outcome type to analyse the relationship and discover any patterns between the two.

(2) Time-Series Stacked Bar Plot: This plot examines the popularity of pet adoption (by percentage of outcome type) 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.

(3) Comparative Box plot: This visualization compares 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.

Additionally, we’ve computed statistics and “top 10” trends for popular pet names to prime for the recommendation system in work package 3.

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

Our approach included:

(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) Support Vector Machine (SVM) for classifying the data by finding an optimal line or hyperplane.

(4) k Nearest Neighbor (kNN) for multi-classification tasks, predicting outcome type based on its nearest data points.

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

(6) XGBoost features regularization to reduce overfitting, making it ideal for large datasets.

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

To assess the performance of these models, we performed repeated (10 times) 5-fold cross validation to produce the following key evaluation metrics:

(1) Accuracy to detect the proportion of true matches in relation to all true and false positives and negatives classes.

(2) Precision to detect the proportion of classified matches that have been classified as true matches.

(3) Recall to detect the proportion of true matches that have been classified correctly.

(4) Log Loss to quantify prediction error by penalising incorrect predictions.

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

In this section, we explored all the tools that we have learned on modeling the relationships between pet characteristics and the target variable - adoption outcome.

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

In this work package, we built a collaborative filtering recommender system that predicts pet names based on the pet's characteristics, such as animal type, breed, sex, color, and age. The system uses techniques like cosine similarity and TF-IDF to analyse the similarity of the input data with training data. By analyzing patterns in existing pet names associated with their characteristics, the system will generate a set of suggested names for new pets.

For any user input, the model provides a personalized selection of names that best match the pet’s unique traits. This approach creates 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 similarity algorithms, we offer more personalized and relevant pet name recommendations, enhancing the overall adoption experience.

**Work Package 4 - Shiny App workflow**

Finally, we are demonstrating the workflow to develop a Shiny app with visualisations and 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 tables (DT package) and plots (plotly package), 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. Additionally, the recommendation system will also be part of this panel. 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.