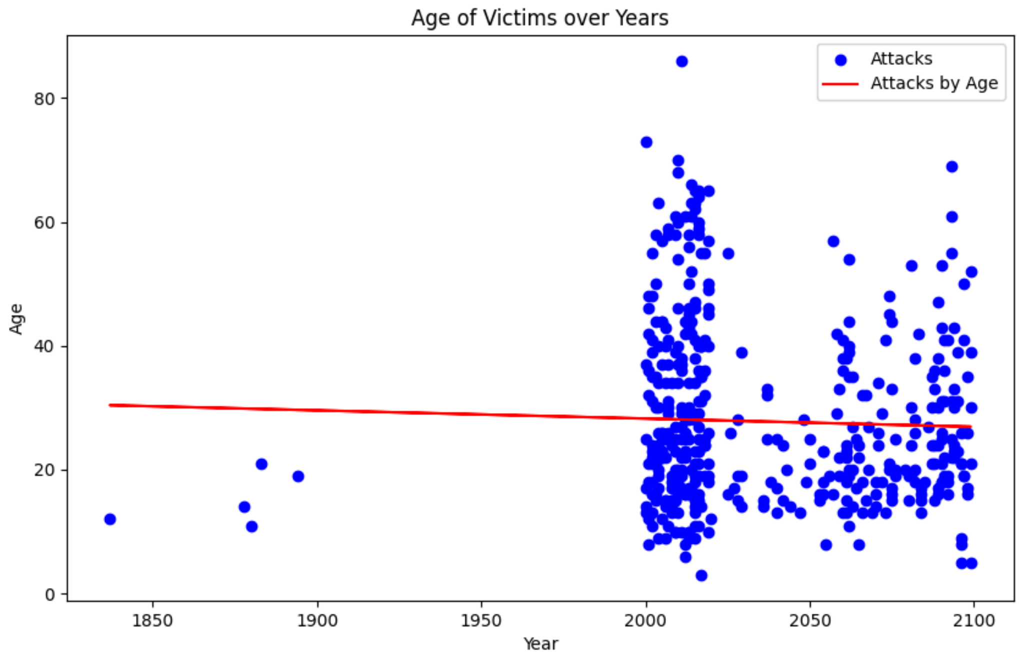
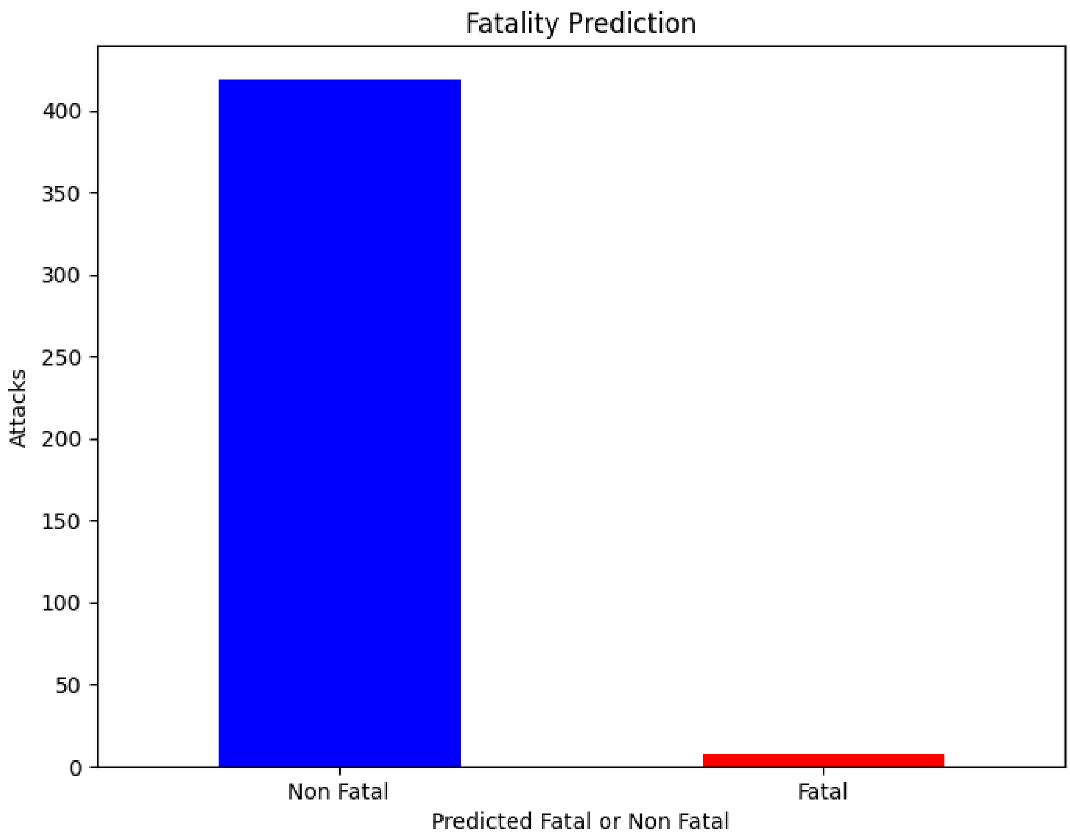
Linear Regression:



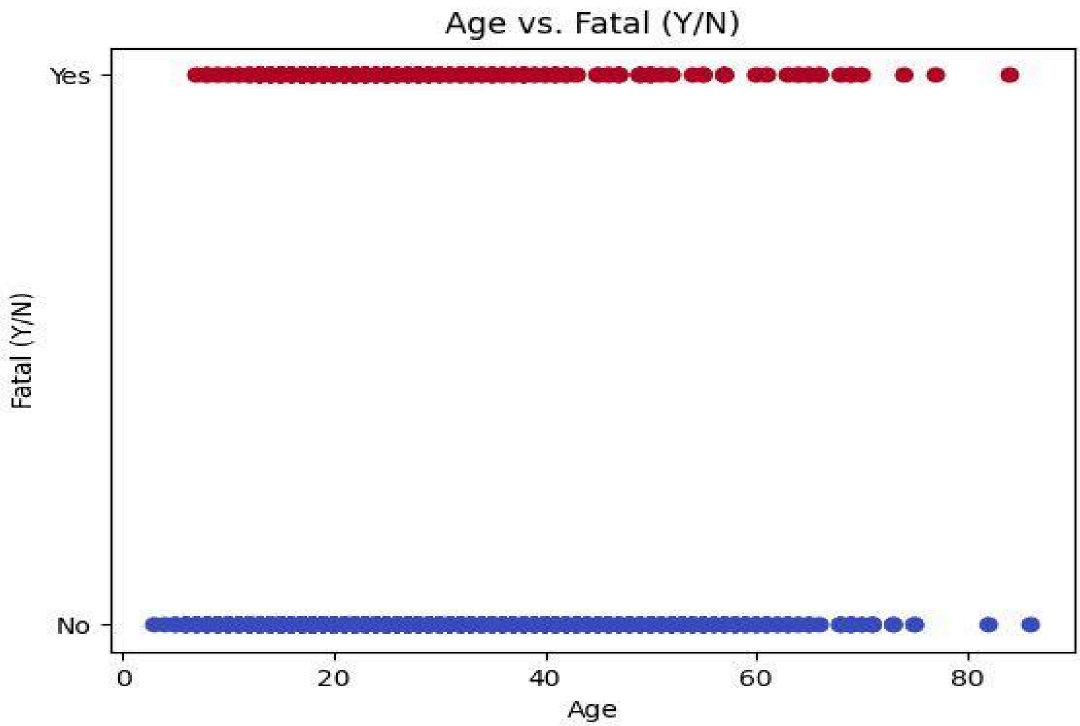
The figure depicts a trend in shark attack victim ages over time, indicating a slight decrease in predicted victim age from around the late 1850s to the year 2100. Linear Regression was selected due to its suitability for modeling and predicting relationships between numerical variables in the shark data, which could potentially provide insights into factors influencing shark attacks over time and with respect to age or other features, allowing for easy identification of key predictors.

Naive Bias:



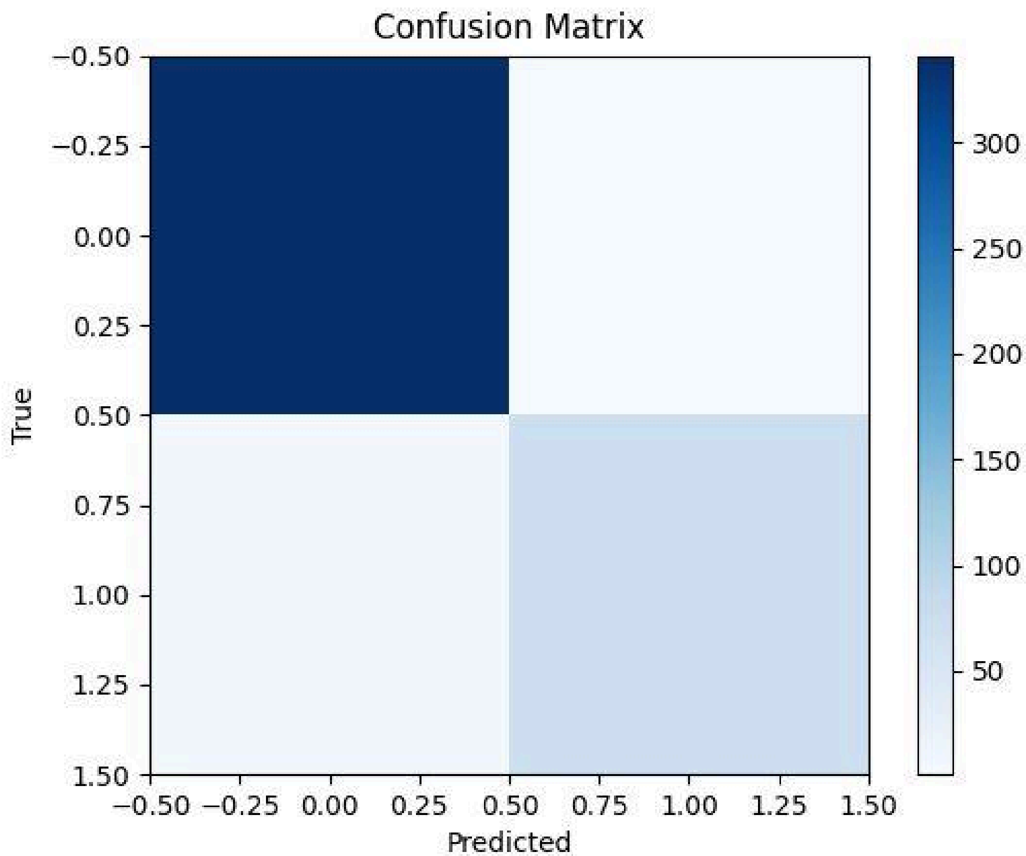
The figure illustrates the distribution of predicted fatalities and non-fatal incidents in shark attack data, with approximately 8 instances predicted as fatal and around 410 instances predicted as non-fatal. Naive Bayes could classify incidents based on their likelihood of resulting in a fatal outcome, providing a straightforward and interpretable approach to classify and compare shark attack attacks with various features. However, Naive Bayes assumes independence among features, which may not hold true in some real-world scenarios.

Logistic Regression:



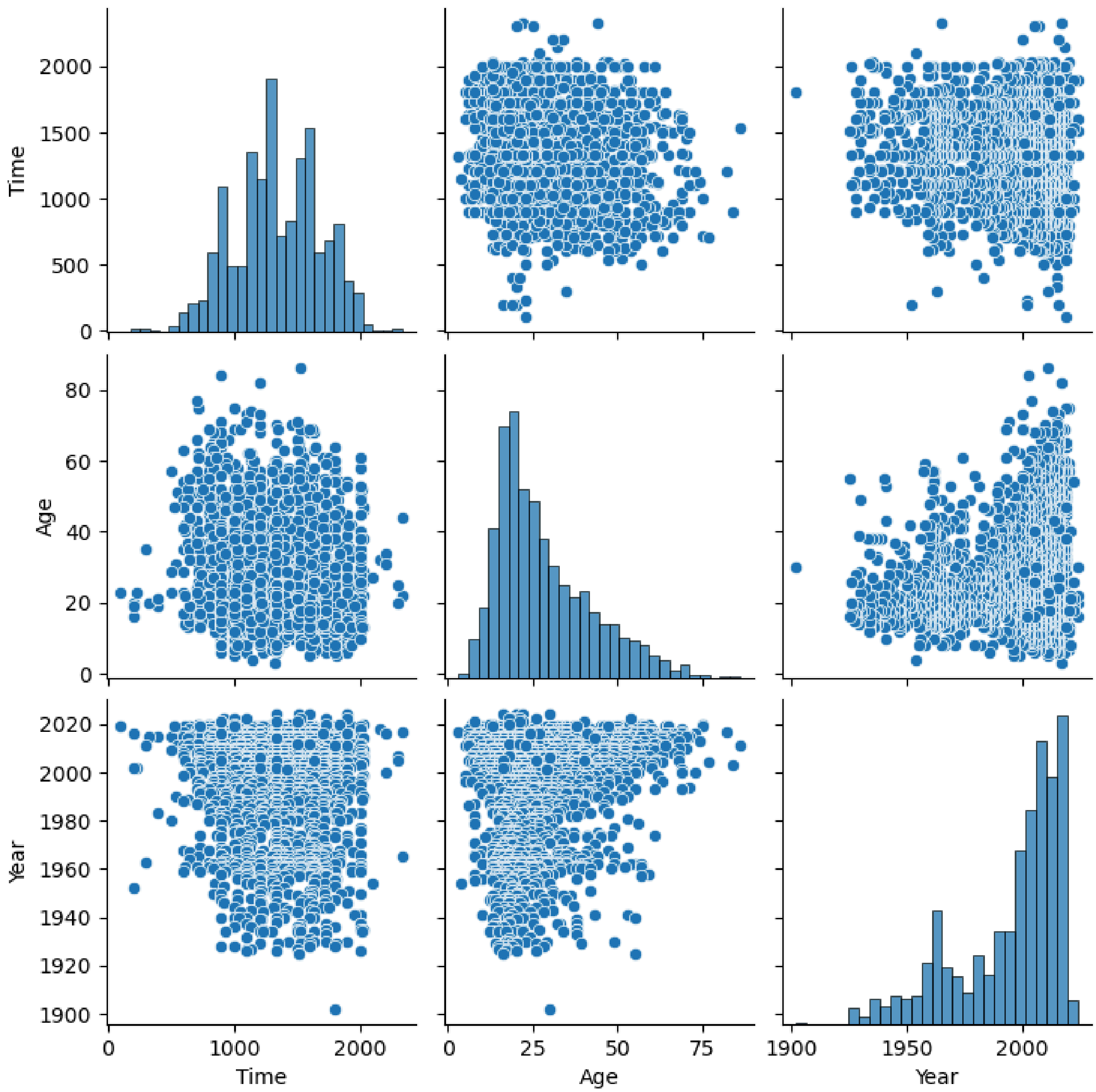
Logistic regression was chosen to look at the relationship between victim ages and fatality due to the fatality variable being binary. The interpretability of logistic regreen allowed us to understand the impact of the predictor variables on the likelihood of a fatal outcome, and what age ranges have fallen victim to a fatal attack most. Logistic regression also offers computational efficiency and assumes linearity, making it very practical for the shark attack dataset. By using logistic regression, we have gained information about the possible relationship between age and fatality rate of shark attacks. This information can be used to aid the development of shark attack prevention strategies and safety measures to mitigate the amount of shark attack fatalities in the waters.

Random Forest:

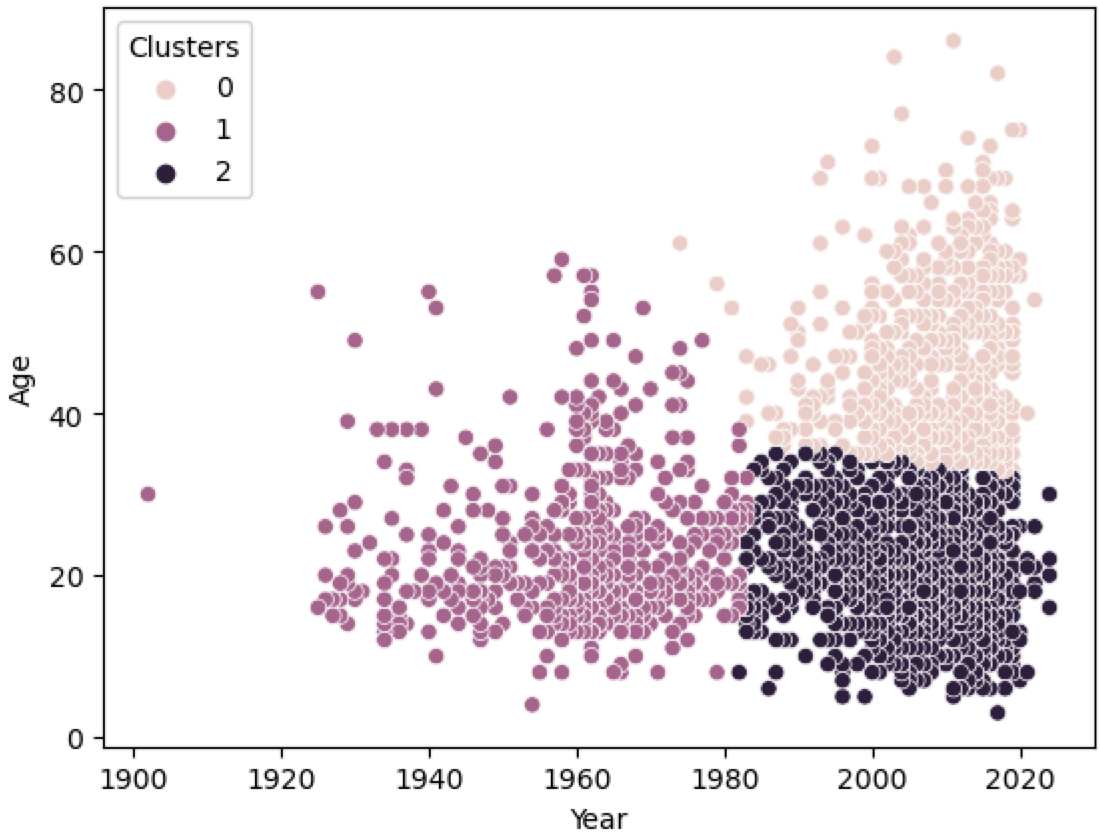


The Random Forest algorithm is beneficial for analyzing our shark attack data because it can handle both numerical and categorical data, suiting the different varieties of information being held by the dataset. By analyzing the shark attack data with Random Forests, insights were gained into the factors that influence shark attacks, like location, activity and time of day. These insights can give us valuable ideas for how to prevent future shark attacks to ensure the safety of beach goers. The above graph shows how accurate the prediction model is, the top left shows how many times the model correctly predicted a non-fatal attack as non-fatal. The bottom right shows the amount of times the model correctly predicted a fatal attack as fatal. The top right and bottom left show incorrectly predicted attacks.

k-Means:



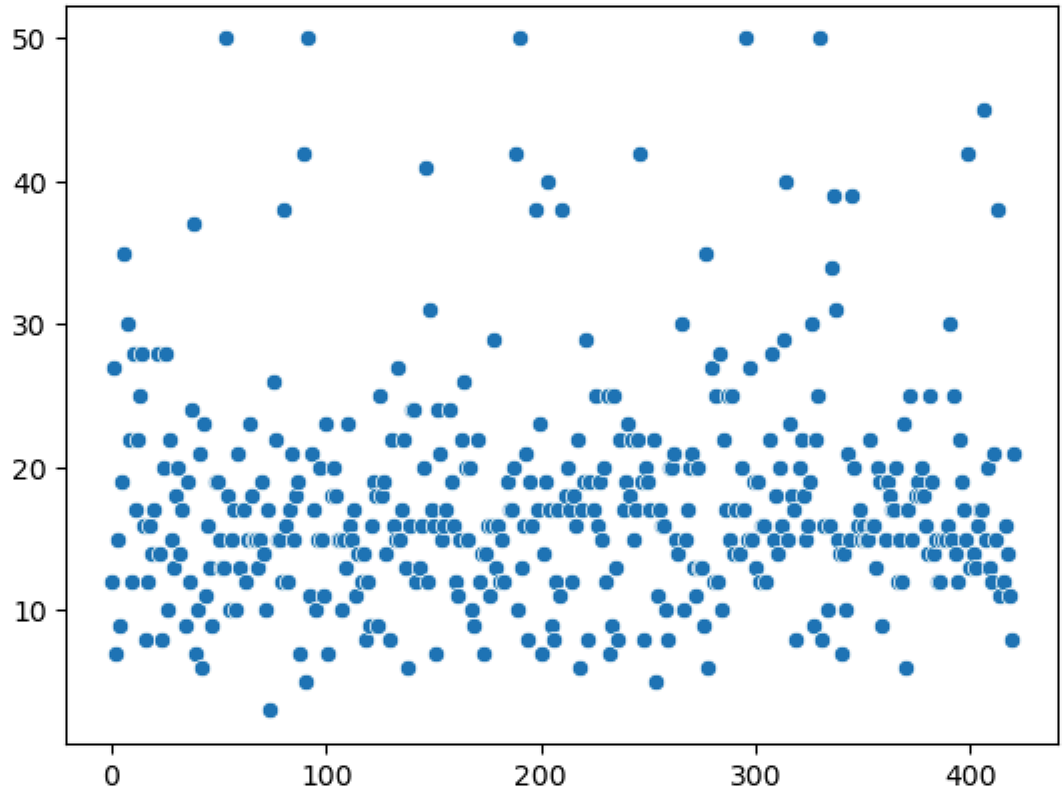
Each of the numerical features can be compared in order to determine which may fit a k-Means clustering model. In this case, the year of the attack (from 1900 until 2024), the age of the victim, and the time of day the attack had occurred (from 00:00 to 24:00) are compared. However, there is very little natural clustering that occurs, which shows that k-Means is not necessarily suitable for analysis. That being said, age and year can have a slight cluster group, as there is a gap of older victims prior to the 2000s.



This shows the age and year scatterplot being segmented into 3 clusters, where the outliers in the gap are handled in part by cluster 0 and by cluster 1. However, trend data is not usefully determined by this relationship using this model, as the desired trend would be for future data to better determine victim demographic trends, while these clusters would be for categorizing newly-incoming data about attacks that occurred in the past.

k-NN:

Similarly to k-Means, a clustering prediction model is less useful than other models for predictions. The goal would again be to determine the demographics of future attacks, while the numerical features provided would not provide a useful prediction pairing.



For instance, given training data with the year and age, the training prediction data (pictured above) has an accuracy of 2.6%, which is not useful even if were more accurate.

Peer Evaluation Form for Final Group Work

CSE 487B

- Please write the names of your group members.

Group member 1 : Matt Kreuzer

Group member 2 : Ben Dykes

Group member 3 : Dylan Vasapollo

- Rate each groupmate on a scale of 5 on the following points, with 5 being HIGHEST and 1 being LOWEST.

Evaluation Criteria	**Group member 1**	**Group member 2**	**Group member 3**
How effectively did your group mate work with you?	5	5	5
Contribution in writing the report	5	5	5
Demonstrates a cooperative and supporting attitude.	5	5	5
Contributes significantly to the success of the project.	5	5	5
TOTAL	20	20	20

Also please state the overall contribution of your teammate in percentage below, with total of all the three members accounting for 100% (33.33+33.33+33.33 ~100%) :

Group member 1 : 33.33%

Group member 2 : 33.33%

Group member 3 : 33.33%