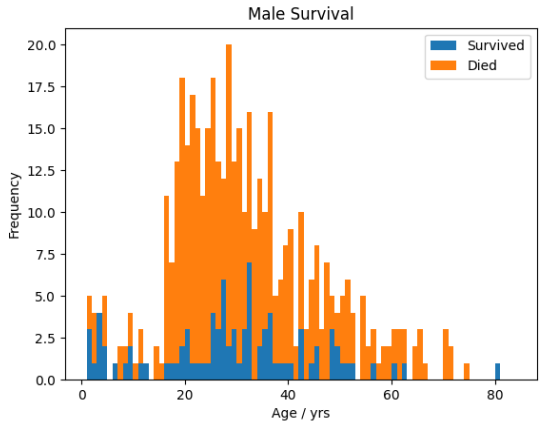
Titanic Survival

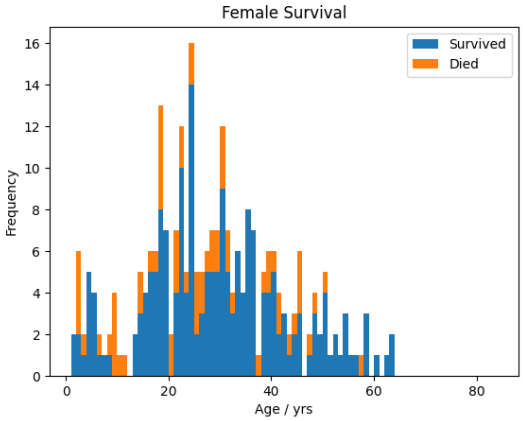
Predict whether a passenger will survive the Titanic disaster.

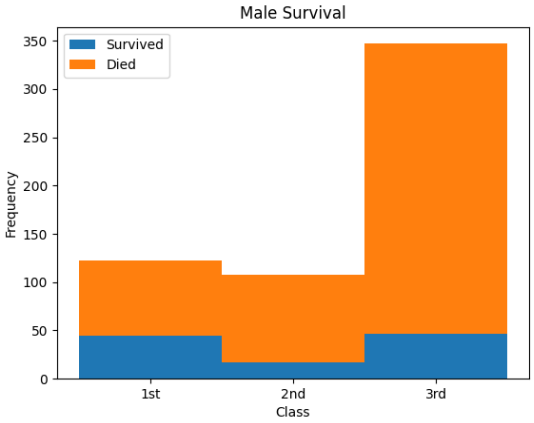
Learning Objectives

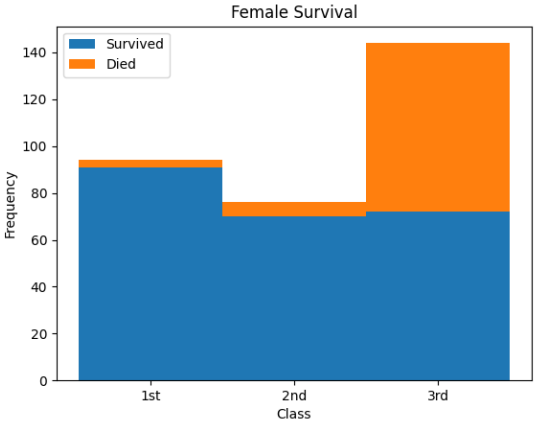
- Cleaning data
- Handling missing data
- Feature importance
- Feature engineering

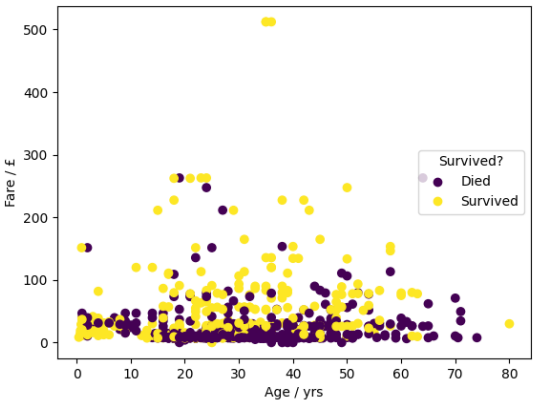
Visualising the Data



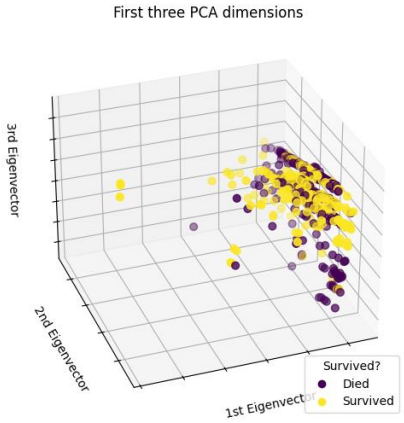








Principal Component Analysis (PCA) was applied to the iris dataset, creating three new features which are a linear combination of the original features. This transformation was chosen to maximise the variance. However, this didn't provide any useful information.



Feature Importance & Creation

It is clear from the histograms that sex, age, and passenger class (or fare) play a significant role in determining survival. Using this information, naïve decision trees were constructed and tested on the dataset.

Naïve Decision Tree	Score
Predict all die	0.6220
Predict survive iff female	0.7655
Predict survive iff female or under 16 years old	0.6603
Predict survive iff female or (under 16 yrs and 1st class)	0.7679

The importance of the remaining features was considered:

Feature	Decision
Name	Too difficult to encode. Used to extract other information
Siblings/Spouse	Combined with Parents/Children to form FamilySize feature
Parents/Children	Combined with Parents/Children to form FamilySize feature
Ticket	Unlikely to provide useful information, ignored
Fare	Provides similar class information to PassengerClass, ignored
Cabin	Most passengers do not have a cabin, ignored
Embarked	Does not provide useful information, ignored

From these features, new features were created: IsChild, Surname, FamilySize, FamilySurvivalRate. IsChild was a binary flag that was true if the passenger was younger than 16 years old; the surnames were extracted from the names; FamilySize was calculated from the number of parents, children, and siblings; FamilySurvivalRate was calculated from the passengers who had the same Surname.

Some passengers had missing ages. If these passengers had the title Miss or Master, they were assumed to be a child. Otherwise, they were assumed to be an adult.

Support Vector Machines

SVMs with different kernels were trained on the training dataset, considering only Sex, PassengerClass, IsChild, FamilySize, and FamilySurvivalRate. Their scores on the testing dataset are summarised here:

SVM Kernel	Score
Linear	0.6324
RBF	0.7679
Sigmoid	0.5754
Polynomial	0.7655

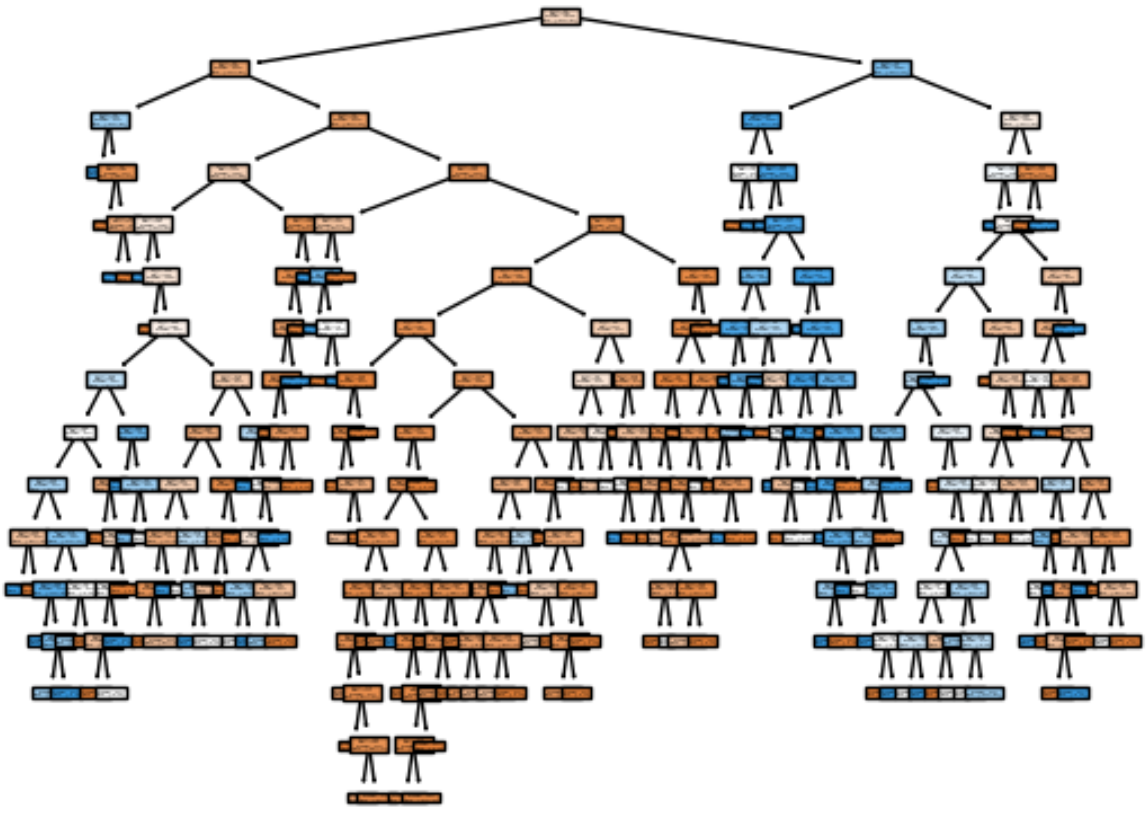
All the SVMs achieved worst scores than the best naïve decision tree – another approach was required.

Machine Learning Decision Tree

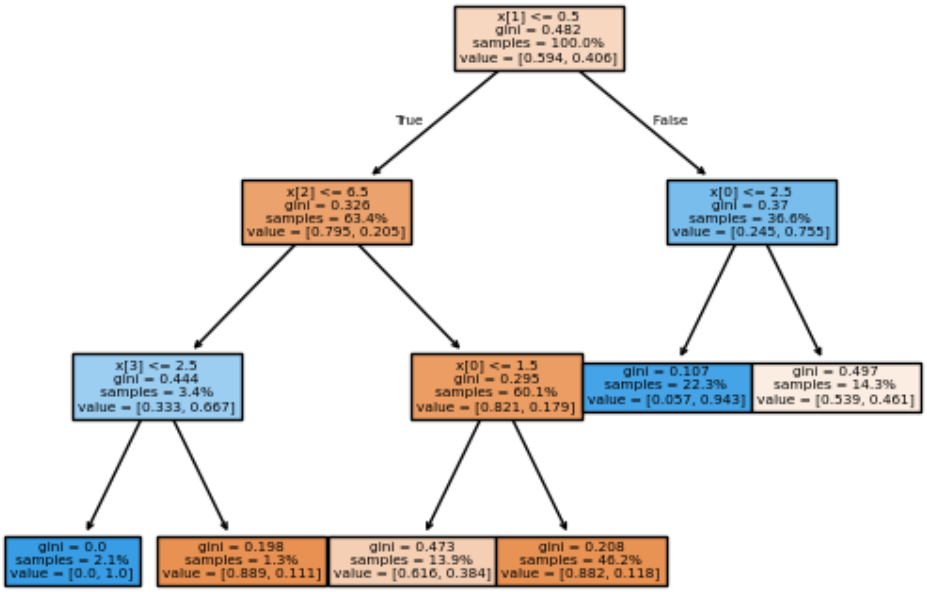
A Decision Tree Classifier was trained on the same data points as the SVMs. It was then pruned to a maximum depth of 4. The resulting trees and their scores are given below:

Full Decision Tree	0.6172
Pruned Decision Tree	0.6770

Full Tree



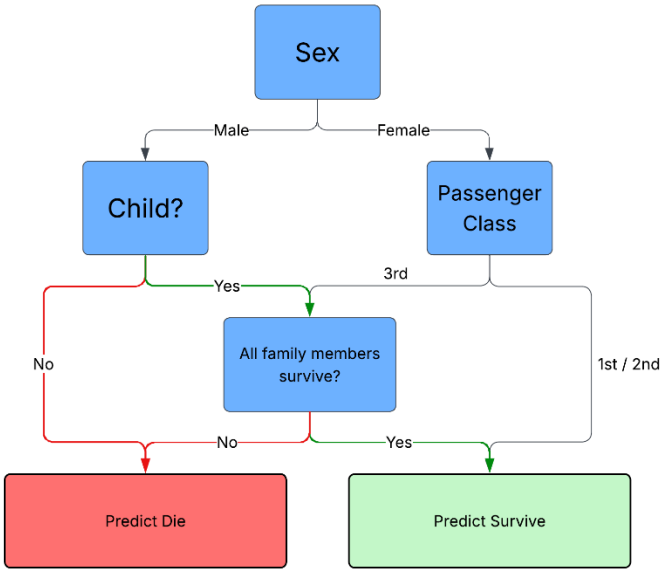
Pruned Tree



The Decision Trees performed worse than the SVMs. It's time to go back to the data.

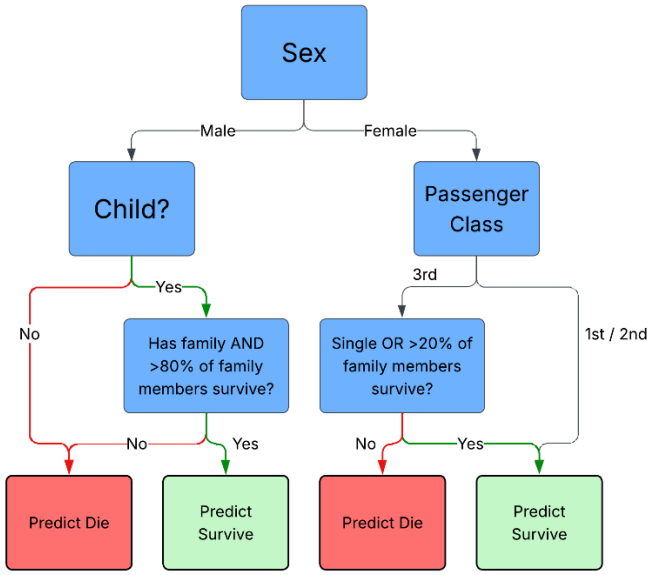
Manual Decision Tree

First & Second Class women survive at a much higher rate than Third Class women. Further investigation of the data showed that family survival rates are usually either very high (>75%) or very low (<25%). This means that families tend to survive or die together. A more informed decision tree was constructed from this information, and is shown below:



This achieved a score of 0.7919 on the test dataset.

But what if the women and children are travelling alone? Single women tend to survive, whilst single boys tend to die. As seen above, not every family *actually* survives and dies together, and within each family, women still have a higher survival rate than boys. This information was incorporated into an updated decision tree, shown below:



This achieved a score of 0.8134 on the test dataset.

Conclusion

This dataset has shown that feature importance and engineering is critical for developing accurate models. All the machine-learning models that were tested performed worse than the feature-informed manual decision trees.

The final highest score of 0.8134 could potentially be improved by incorporating some Machine Learning in the decision process. Currently, all adult men are assumed to die, which is clearly not the case in the training dataset. A model could be fitted only on the adult men and used to predict their survival. A similar strategy could be employed for the single and $1^{st}/2^{nd}$ class women.