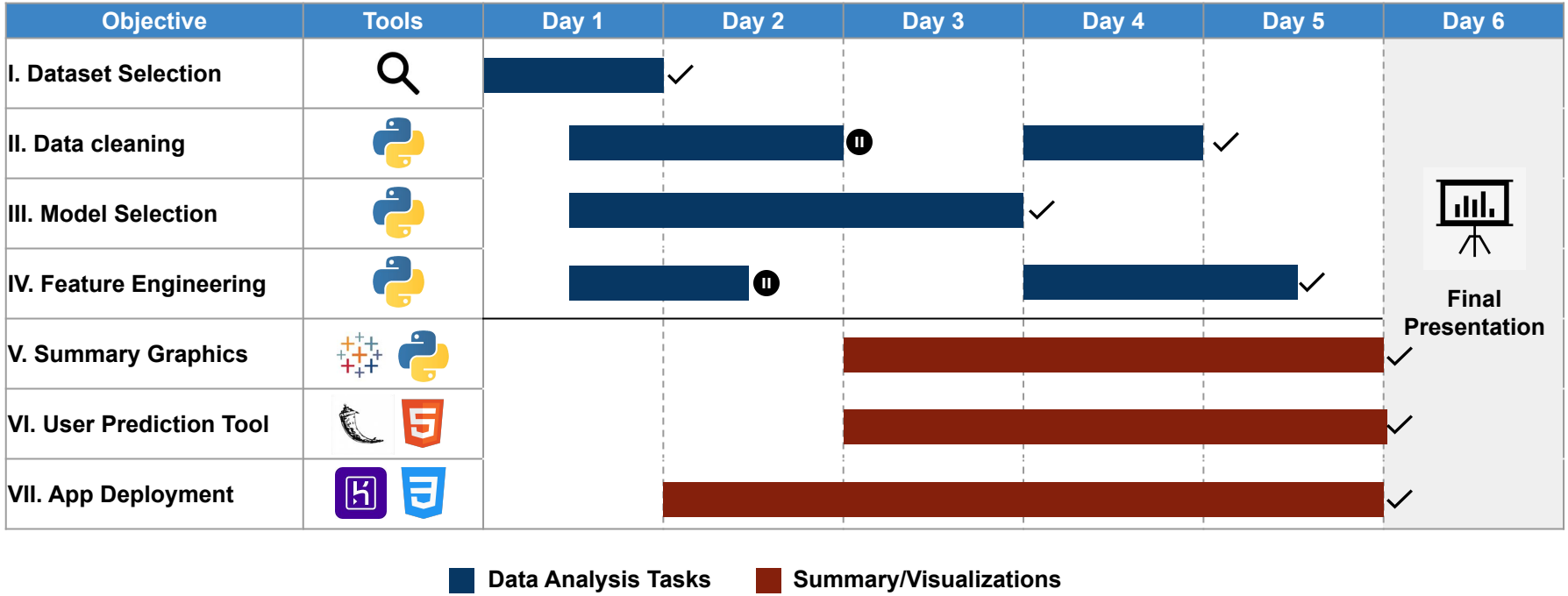
Rose or Jack?

Taylor Bohl
Harish Korrapati
Corey Lawson-Enos
Rhiana Schafer
Ishanjit Sidhu

PROJECT BACKGROUND & DESCRIPTION

We built a machine learning model that predicts whether you will survive a voyage on the Titanic based on your age, gender, passenger class, fare, and whether you travel solo or with family.



INITIAL DATA SELECTION & MANIPULATION

Raw Data
(n = 1309, cols = 14)

Feature	Format	Type	Definition	Pct. Null	
Passenger Class	int.	Cat.	Class (1st, 2nd, 3rd)	0.0%	
Survival	bool.	Cat.	Survived?	0.0%	1
Name	str.	pKey	Name	0.0%	3
Sex	str.	Cat.	Sex	0.0%	
Age	int.	Quant.	Age	20.1%	
Siblings/Spouses	int.	Quant.	# siblings/spouses aboard	0.0%	
Parent/Child	int.	Quant.	# parents/ children aboard	0.0%	
Ticket Number	str.	Cat.	Ticket ID	0.0%	3
Passenger Fare	float	Quant.	Ticket Cost	1.4%	
Cabin Number	str.	Cat.	Cabin+Deck alphanum.	77.5%	2
Port of Embarkation	str.	Cat.	Port of passenger origin	<0.1%	3
Lifeboat Boarded	str.	Cat.	Lifeboat boarded	62.9%	1
Body Number	int.	Quant.	Recovered body #	90.8%	1
Home Destination	str.	Cat.	Passenger's hometown	39.4%	3

1. Proxy for outcome of interest (Survival)

2. Excessive null values

3. Initial analysis deemed irrelevant

Rationale for Elimination

Objectives (Dataset 1.1)

- Take an initial pass by pulling a set of data with clean and reliable data to minimize noise
- Use simple, explainable data to understand drivers of survival rate
- Conduct basic feature engineering to combine similar columns into more predictive drivers

Reduced Data
(n = 1045, cols = 5)

Feature	Type	Unique Vals.
Passenger Class	Cat.	3
Sex	Cat.	2
Age	Quant.	-
Family size	Quant.	-
Passenger Fare	Quant.	-

One-hot encoding

Final Dataset
n = 1045
cols = 8

Source

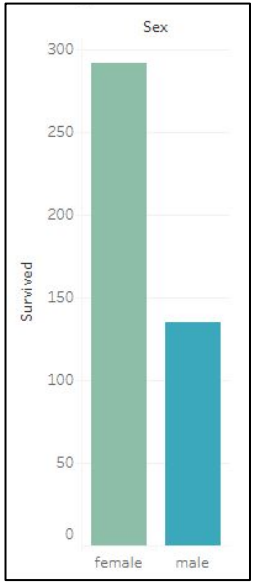
Kaggle
Titanic - Machine Learning from Disaster
Source URL: https://www.kaggle.com/competitions/titanic/data
Full File: https://www.kaggle.com/datasets/vinicius150987/titanic3?resource=download

PRELIMINARY HYPOTHESES

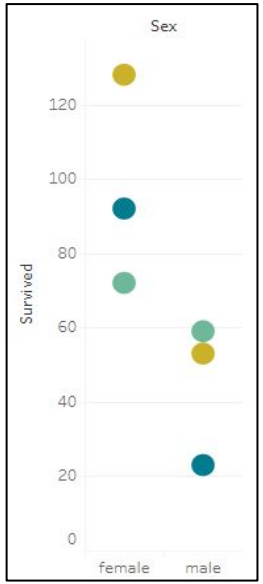
Hypotheses

I. Sex has a significant effect on survival likelihood
II. Class has a significant effect on survival likelihood
III. Family size has a significant effect on survival likelihood

Count of Survivors by Sex

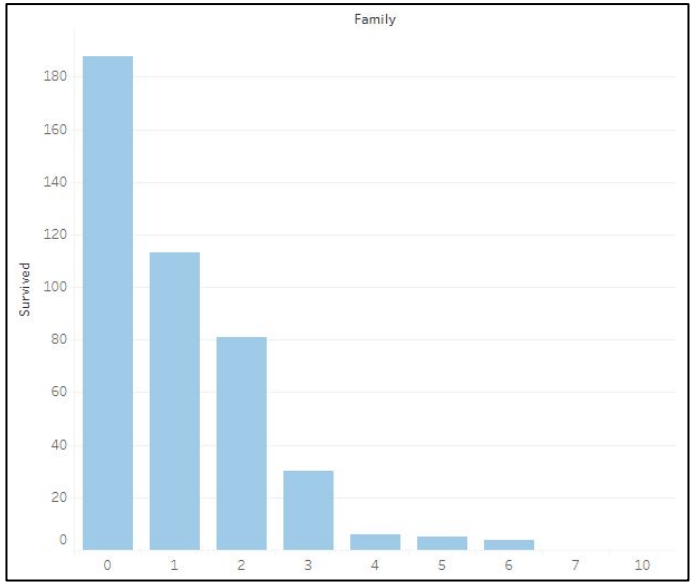


Count of Survivors by Sex
Grouped by Class



1st **2nd** **3rd**

Count of Survivors by Number of Family Members



Important questions to consider for future model refinement:

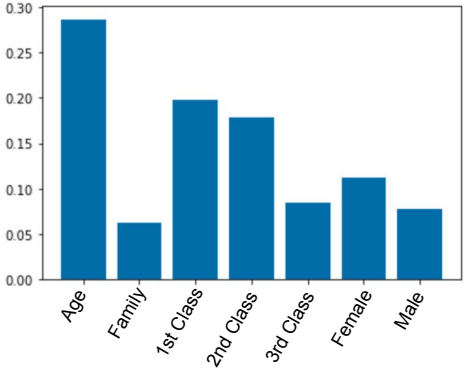
- Are you more likely to survive if you are alone or female?
- Are there far more women and solo travelers than not?
- Is it possible that people with missing family info are coded as 0 in data?
- What are potential reasons first-class passengers appear more likely to survive?
- What are the viable paths to survival?

Model Selection & Creation

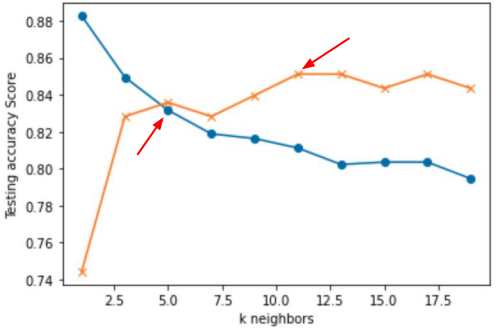
We selected four classification models to predict the survival of Titanic passengers....

Model Results			
Model	**Training Score**	**Testing Score**	**Notes**
Logistic Regression	76.7%	83.2%	❖ Needed to provide a higher value for max iterations
Random Forest	100.0%	84.0%	❖ 8 features ❖ 5 informative
Deep Neural Network	-	80.5%	**Total parameters: 1,621** ❖ **1st layer:** 40 units; activation = relu ❖ **2nd layer:** 20 units; activation = tanh ❖ **3rd layer:** 20 units; activation = relu ❖ **Output layer:** 1 unit; activation = sigmoid
K-Nearest Neighbors	82.6%	84.0%	❖ k: 5 provides the best accuracy where the classifier starts to stabilize ❖ model has highest performance at k: 11

I. RF Feature Importance



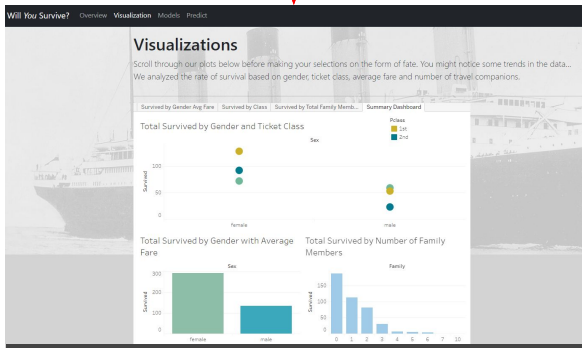
II. KNN



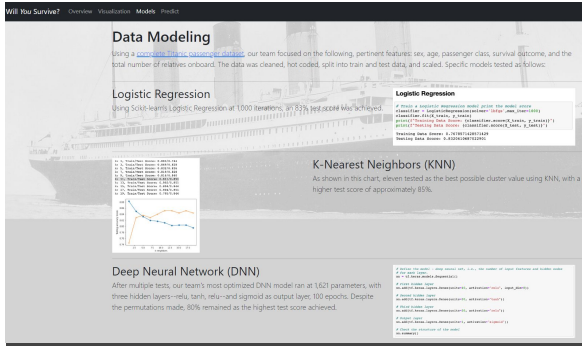
WEBPAGE PATHS

- **00 – Overview:** Our websites main page, featuring an overview of our project along with musical accompaniment, and links to the other routes.
- **01 – Visualizations:** Several tableau visualizations showing trends in the data
- **02 – Models:** Descriptions of the various machine learning models attempted on the data, with their relative levels of success
- **03 – Predict:** An interactive form that allows the you to enter in your information, and our machine learning model will predict whether you live or die!

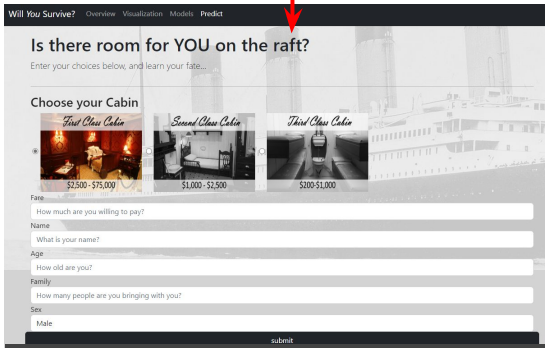
00 – Overview (Homepage)



01 – Visualizations



02 – Models



03 – Predict

INTERACTIVE DEMONSTRATION

LEARNINGS, PAIN POINTS & FUTURE ENHANCEMENTS

Learnings/Pain Points

Learnings:

- Picking the right data saves you a ton of time in the beginning
- Hard to improve beyond 84% accuracy - feature enhancements do not necessarily improve accuracy

Pain Points:

- Did not realize we needed to export the standardscaler for flask model to work
- Constant adjustments of CSS code
- Heroku deployment caused boat icon to vanish

Future Enhancements

Data Analysis:

- Find additional datasets with new features (e.g. socioeconomic features) to add and improve scores
- Create a database to hold input data to analyze participant trends
- Project what lifeboat passengers attempted to get on
- Understand if families stuck together or separated (reluctantly or willingly)
- Delve further into Random Forest documentation for parameter enhancements that might improve scores

Web Application/Visualizations:

- On form page, restrict max/min values based on class selection

Just Graduated!