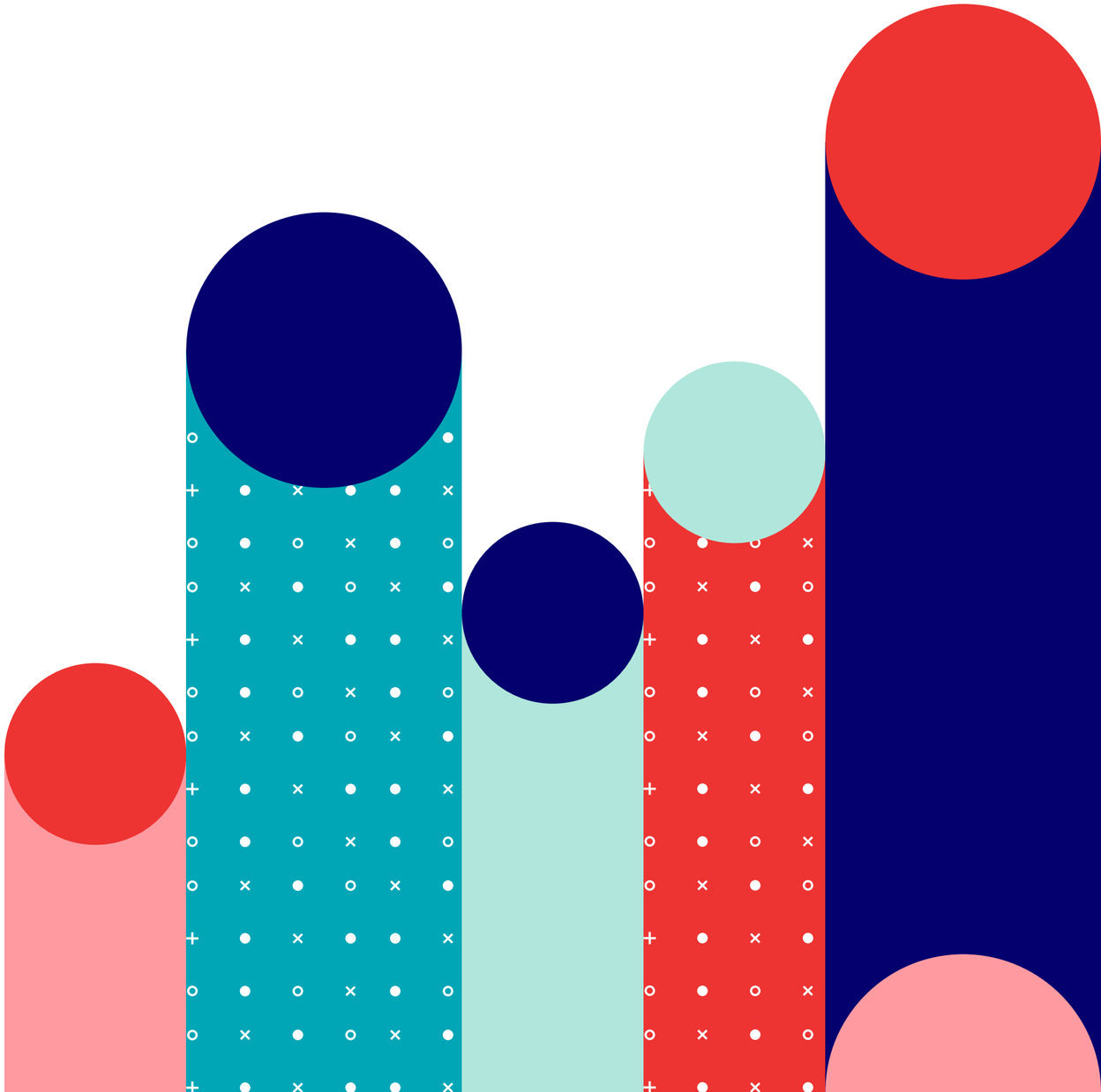
Analyzing Titanic Survival Data

What factors influenced survival of passengers?

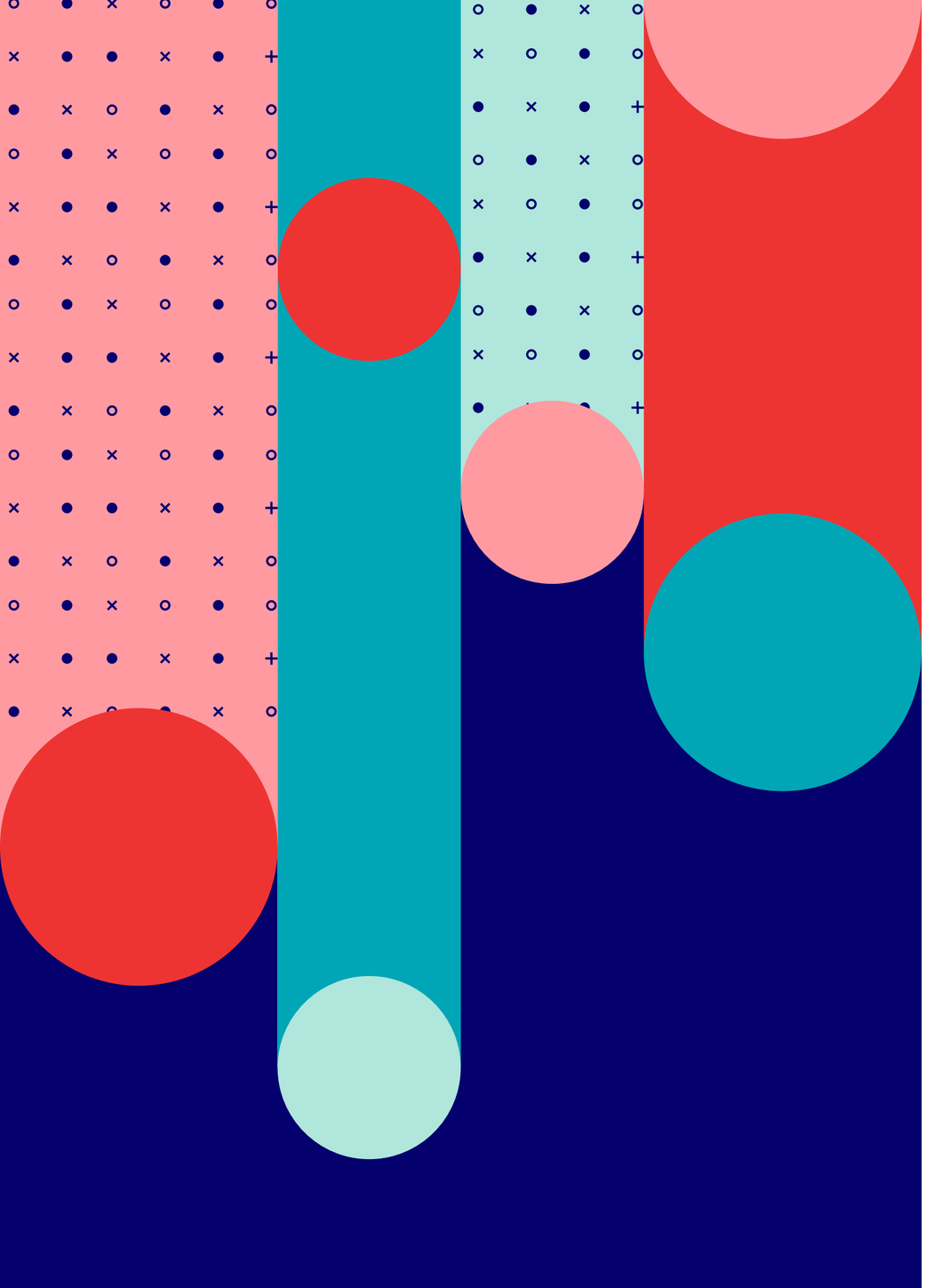
Contents

- Exploratory Data Analysis
  - Introduction to the dataset
  - Data Exploration
  - Data Visualization
- Machine Learning
  - Predictive analysis of survival using classifiers.
  - Feature Importance



Introduction to the Titanic Survival Dataset

The dataset contains a wealth of information about Titanic passengers, including details such as their age, gender, ticket class, fare, and most crucially, whether they survived or not. It serves as a valuable resource for exploring the factors that influenced survival rates among the diverse group of individuals aboard the ship.



Features

#	Column	Non-Null Count	Dtype
0	PassengerId	891 non-null	int64
1	Survived	891 non-null	int64
2	Pclass	891 non-null	int64
3	Name	891 non-null	object
4	Sex	891 non-null	object
5	Age	714 non-null	float64
6	SibSp	891 non-null	int64
7	Parch	891 non-null	int64
8	Ticket	891 non-null	object
9	Fare	891 non-null	float64
10	Cabin	204 non-null	object
11	Embarked	889 non-null	object

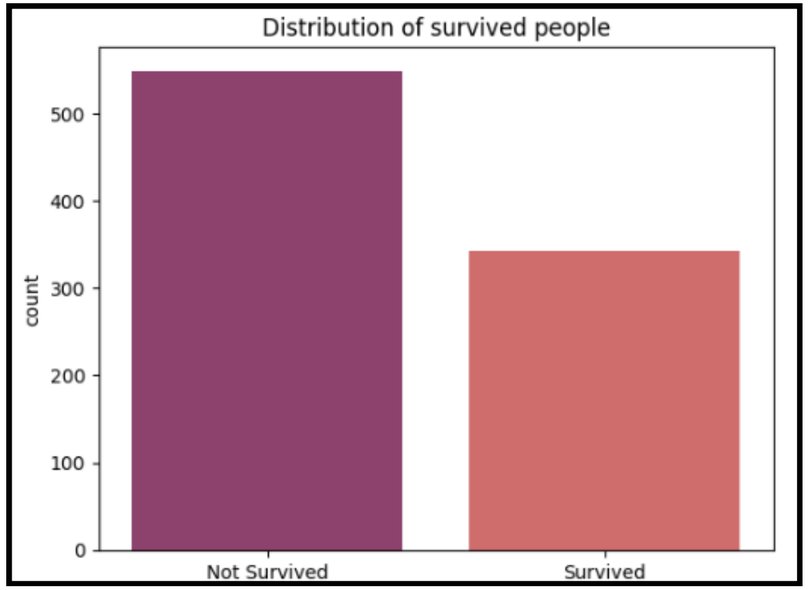
Missing Values

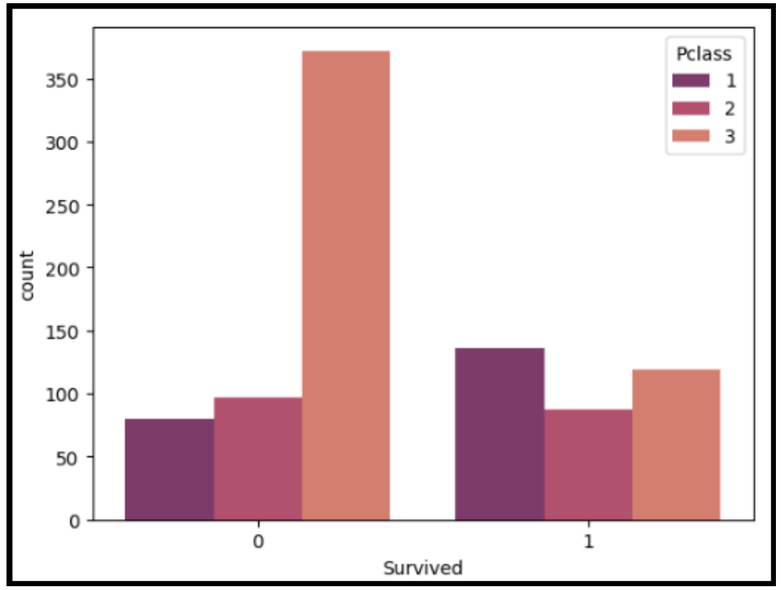
How to handle missing values?

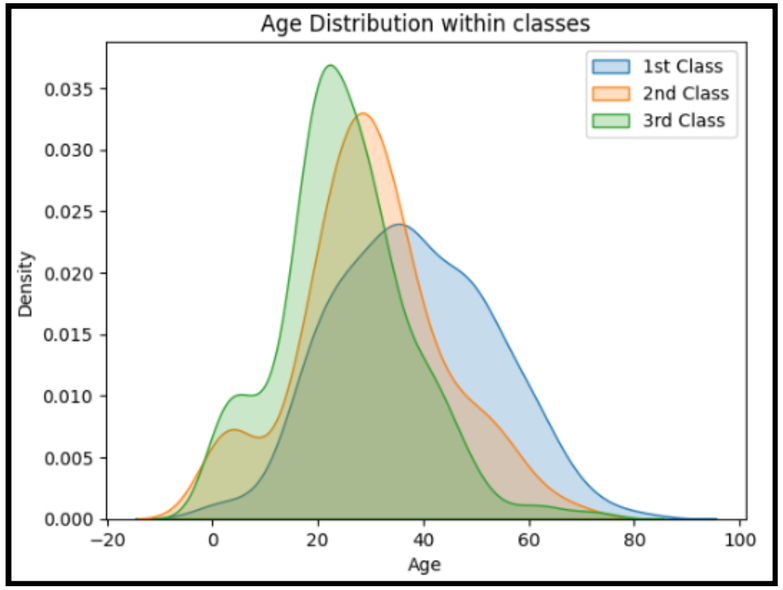
If the number of missing values is significantly large, and there is not enough information to reproduce the data, in such cases we drop the column.
Eg: We drop the cabin column as it had over 95% data missing

Reproduce the missing values using statistical methods like mean, median and mode.
Eg: We fill the missing values in Age column by the mean of the pclass that passenger belongs.

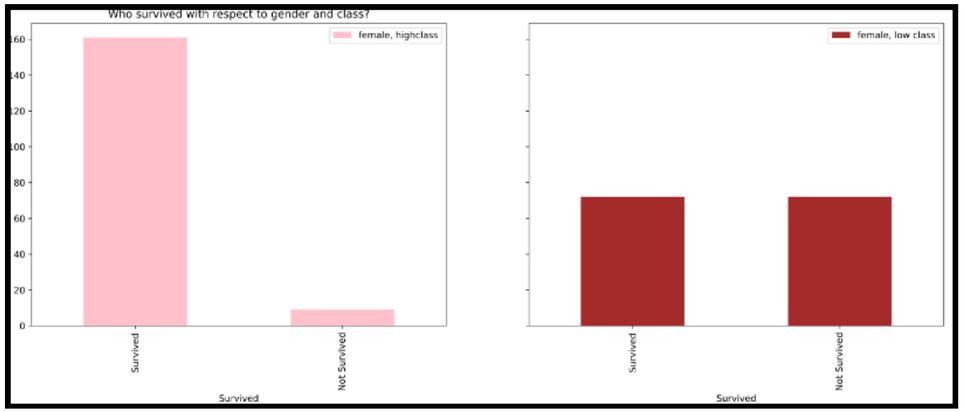
Relation between Age and Pclass

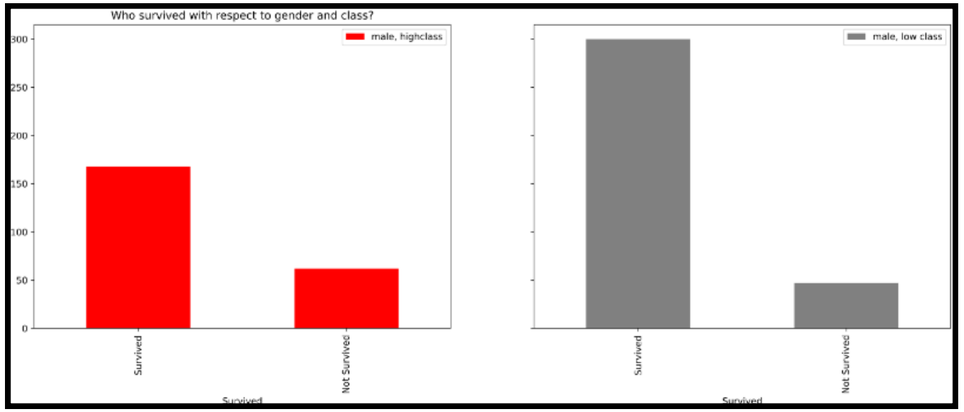






Exploring Relations with Survived column





Correlation Analysis

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	male	Q	S
PassengerId	1.000000	-0.005028	-0.035330	0.032625	-0.057686	-0.001657	0.012703	0.043136	-0.033694	0.022269
Survived	-0.005028	1.000000	-0.335549	-0.055958	-0.034040	0.083151	0.255290	-0.541585	0.004536	-0.151777
Pclass	-0.035330	-0.335549	1.000000	-0.394125	0.081656	0.016824	-0.548193	0.127741	0.220558	0.076466
Age	0.032625	-0.055958	-0.394125	1.000000	-0.241624	-0.172144	0.116083	0.084944	-0.069496	0.007387
SibSp	-0.057686	-0.034040	0.081656	-0.241624	1.000000	0.414542	0.160887	-0.116348	-0.026692	0.069438
Parch	-0.001657	0.083151	0.016824	-0.172144	0.414542	1.000000	0.217532	-0.247508	-0.081585	0.061512
Fare	0.012703	0.255290	-0.548193	0.116083	0.160887	0.217532	1.000000	-0.179958	-0.116684	-0.163758
male	0.043136	-0.541585	0.127741	0.084944	-0.116348	-0.247508	-0.179958	1.000000	-0.075217	0.121405
Q	-0.033694	0.004536	0.220558	-0.069496	-0.026692	-0.081585	-0.116684	-0.075217	1.000000	-0.499261
S	0.022269	-0.151777	0.076466	0.007387	0.069438	0.061512	-0.163758	0.121405	-0.499261	1.000000

Predictive Analysis

We train the following classifiers and compare their accuracy

- Logistic Regression
- XGBOOST Classifier
- Random Forest Classifier

Train-Test Split and Logistic Regression

1. For binary classification, we divide the data into training features (X) and target variable (y). We further split both X and y using train_test_split for cross validation. This helps us diagnose and rectify problems like overfitting and underfitting.

2. Logistic regression is a statistical method used for binary classification, modeling the probability of an observation belonging to one of two classes based on input features.

Logistic Regression vs XGBoost

- Linear Algorithm
- Binary Classification
- Provides Coefficients
- Limited ability to cover non-linear relations
- Generally faster to train
- Suitable for large datasets. with small number of features

- Ensemble Learning
- Versatile for both regression and classification
- Provides feature importance but is difficult to interpret
- Excellent at capturing non-linear relations
- Can be computationally more expensive
- Suitable for large dataset with large number of features

Random Forest

What are they?

Random Forest is an ensemble learning method that belongs to the family of decision tree algorithms. It operates by constructing a multitude of decision trees during training and outputs the mode of the classes (classification) or the average prediction (regression) of the individual trees. The "random" in Random Forest comes from the use of two sources of randomness: random feature selection and bootstrapped sampling of the training data.

Advantages in our analysis

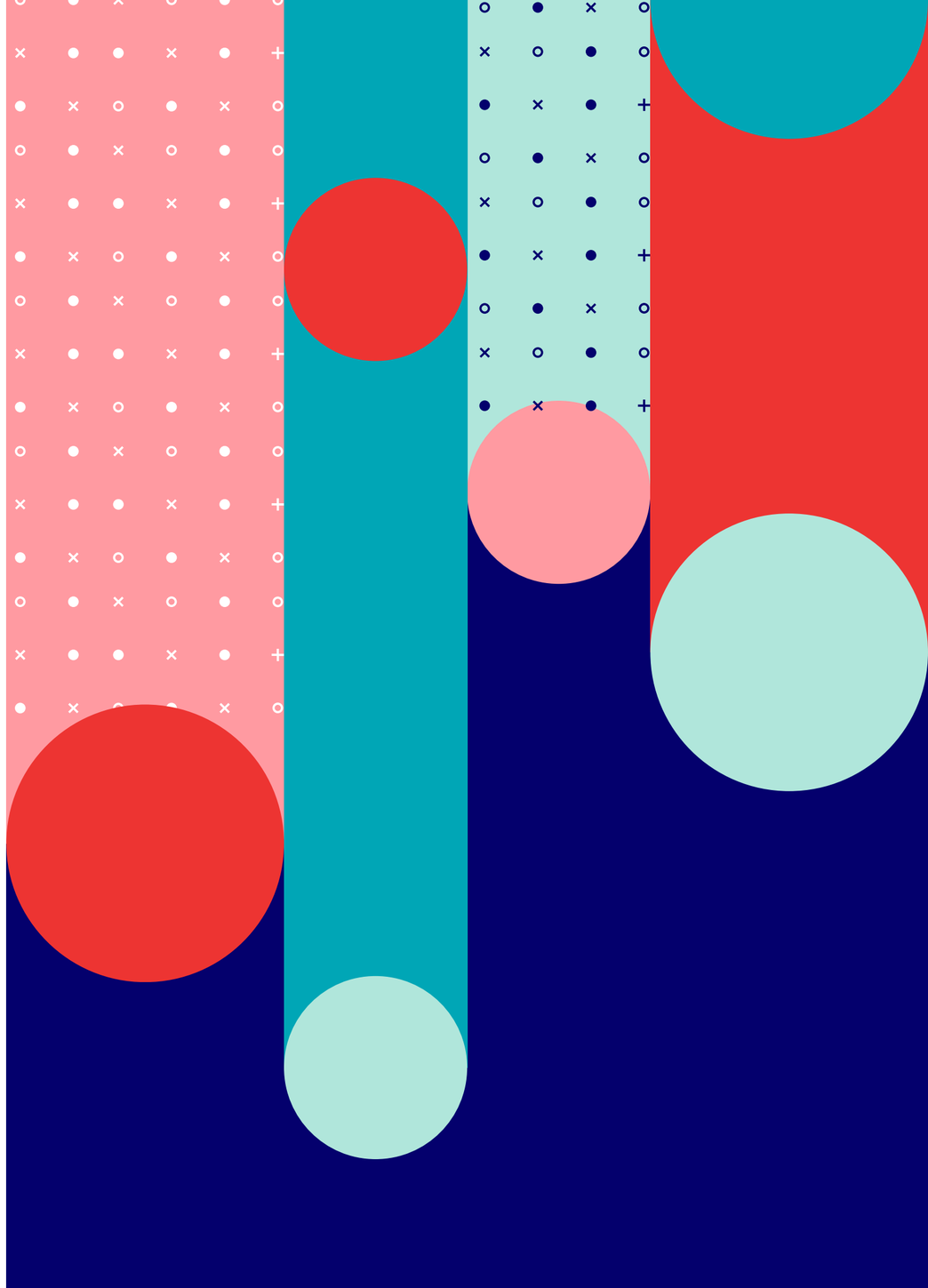
- Handling non-linear relations
- Provides feature importance
- Robustness to overfitting
- Handling categorical features
- Versatile nature
- Out-of-the-box performance

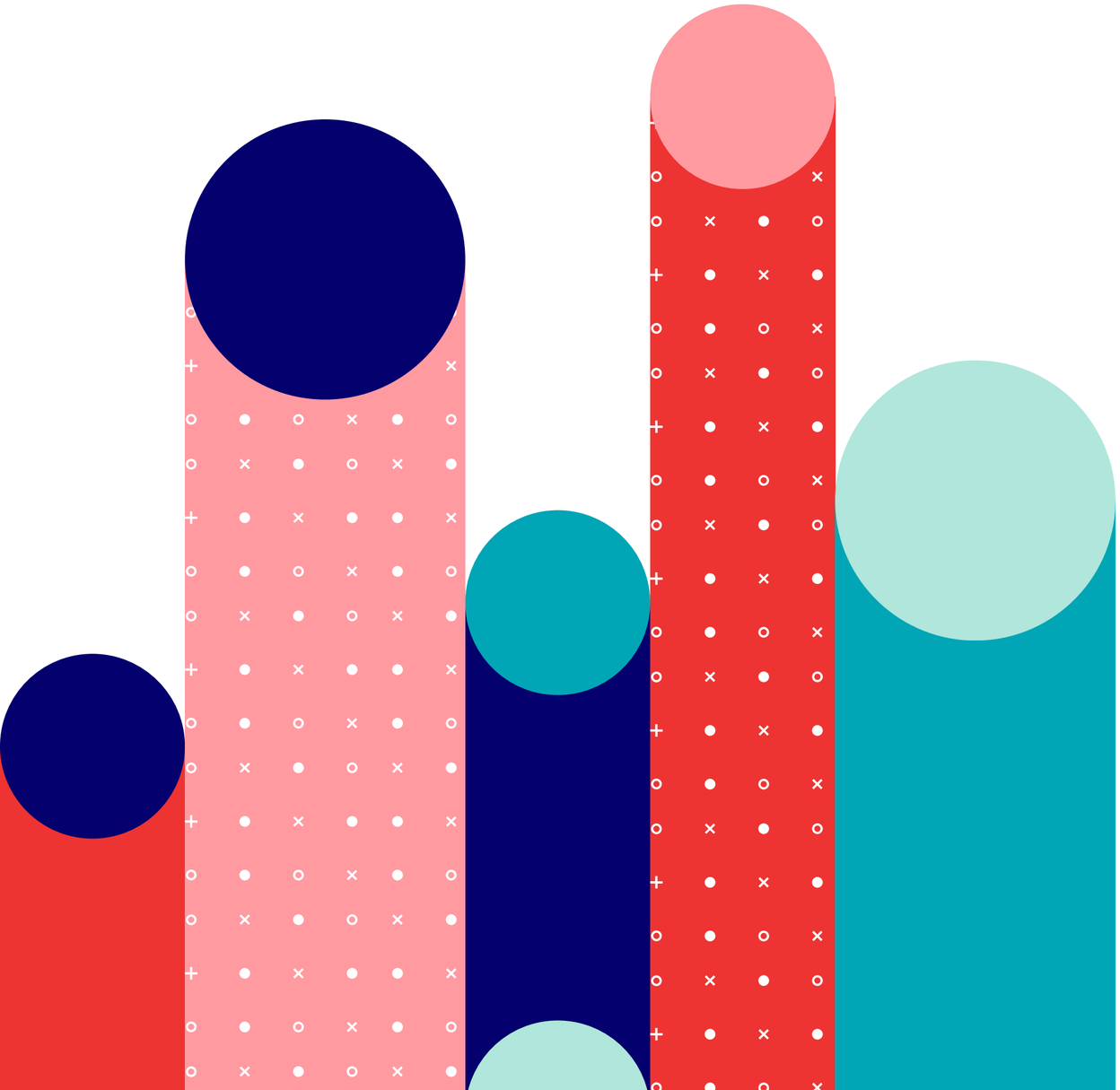
Comparing Model Performance

1. Both Logistic Regression and XGBoost Classifier provide the same accuracy ~ 79.77% This indicates a simple linear model is able to capture the relationship between the train and test data successfully without the threat of overfitting.

2. In cases where interpretability is not the main concern, we may prefer xgboost as it provides feature importance.

3. Random forest outperforms ~80% accuracy with n_estimator=100 and max_depth=3





Feature Importance and Inferences

1. Women and children were the first to board the titanic which means they are more likely to survive than men
2. The younger you are the more likely to survive
3. You have a higher chance of surviving if you have a first class ticket than having a second or third
4. You are more likely to survive if you are travels with 1 to 3 people and if you have 0 or more than three you have a less chance.

Thank you

Analysis by:
Sidhved Warik
sw6071
N10824439