URL to Colab:

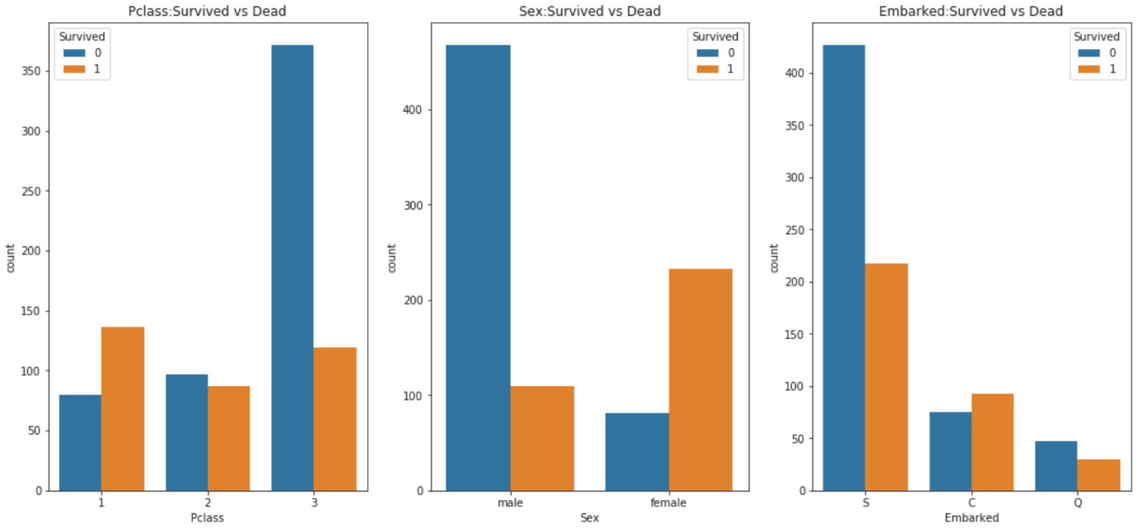
https://colab.research.google.com/drive/1NKlm-xC_MUfTGWpJ_YqoL8flLCvgAecR

Project Explanation:

This Titanic Machine Learning Project is to utilize the training dataset that includes the outcome for each passenger to predict what sorts of people were likely to survive.

Data Visualization

To have an overview of the whole dataset, I first visualized the relationship between the survival rate and three variables – Pclass, Sex and Embared by using the barplots.



From the above graph, I got the general idea that male passengers who were from the third class and embarked from Southampton had lower survival rates.

Data Processing

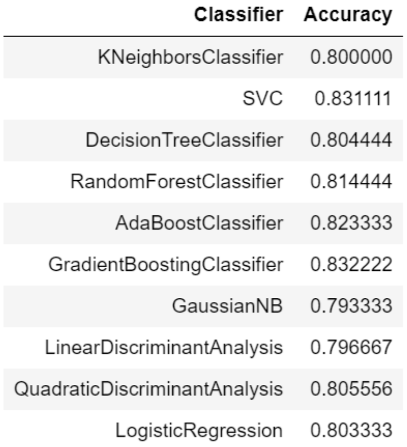
To translate collected data into usable information and feed my prediction model, I manipulated data with the following methods:

New Variable	**Original Variable**	**Manipulation Method**
Age	Age	• Fill null values with mean age value • Divide age values into five groups and find the cutting age value • Assign a group number to each age group • Map age values to age group values
Fare		• Fill null values with mean fare value • Divide fare values into four groups and find the cutting fare value • Assign a group number to each fare group • Map fare values to fare group values
Embarked	Embarked	Transfer categorical variable to numerical variable

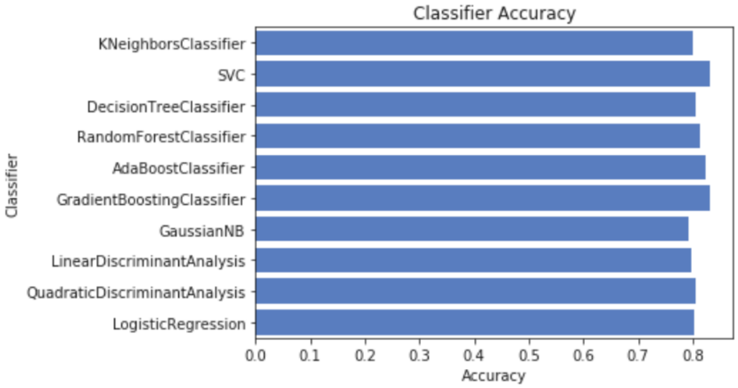
Title	Name	- Split titles and names and only keep titles - Replace uncommon titles with similar common titles - Transfer categorical variable to numerical variable
FamilySize	SibSp, Parch	Add the numbers of SibSp and Parch together to get the family size
Has_Cabin		- If the data type of the cell under Cabin column is a string, then assign "1" to this passenger to indicate that he has a cabin. - If the data type of the cell under Cabin column is a float, then assign "0" to this passenger to indicate that he doesn't have a cabin.
Pclass	Pclass	No Data Manipulation
Sex	Sex	No Data Manipulation
Parch	Parch	No Data Manipulation

Data Modeling

In the data modeling process, I utilized ten machine learning algorithms to make predictions of what type of people have higher survival rates. The accuracy comparisons of these ten machine learning models are listed below:



Classifier	Accuracy
KNeighborsClassifier	0.800000
SVC	0.831111
DecisionTreeClassifier	0.804444
RandomForestClassifier	0.814444
AdaBoostClassifier	0.823333
GradientBoostingClassifier	0.832222
GaussianNB	0.793333
LinearDiscriminantAnalysis	0.796667
QuadraticDiscriminantAnalysis	0.805556
LogisticRegression	0.803333



Recommendations to Management:

By comparing the accuracies of ten machine learning models for survival prediction, we can conclude that the Gradient Boosting Method is the best model for prediction based on current manipulated dataset. The accuracy rate of Gradient Boosting Model is 83.22%.

I also further evaluated the precision rate and recall rate of the model, which are 78.28% and 68.08% respectively.

Model Improvement:

There are also ways to improve the accuracy of the data analysis of this project. For example, we can conduct a more extensive feature engineering and remove the noisy features.