Titanic Survival Prediction using Supervised ML.

(Project Description)

Titanic Problem

Based on the sinking of the RMS Titanic, that ended up killing 1502 out of 2224 passengers and crew. One of the reasons for such loss was that there were not enough lifeboats for the passengers and crew. Some groups of people were more likely to survive than others. In this challenge you are requested to analyse data applying *machine learning* and predict which passenger survived the tragedy

Models Implemented

Right now I have implemented 9 *Machine Learning* models.

1. `Logistic Regression`
2. `Gaussian Naive Bayes`
3. `Support Vector Machines`
4. `Linear SVC`
5. `Perceptron`
6. `Decision Tree Classifier`
7. `Random Forest Classifier`
8. `KNN or k-Nearest Neighbors`
9. `Stochastic Gradient Descent`
10. `Gradient Boosting Classifier`

Platform Used:

Google Collab, VSCode, Pycharm

Language Used:

Python 3.6

Dataset Used:

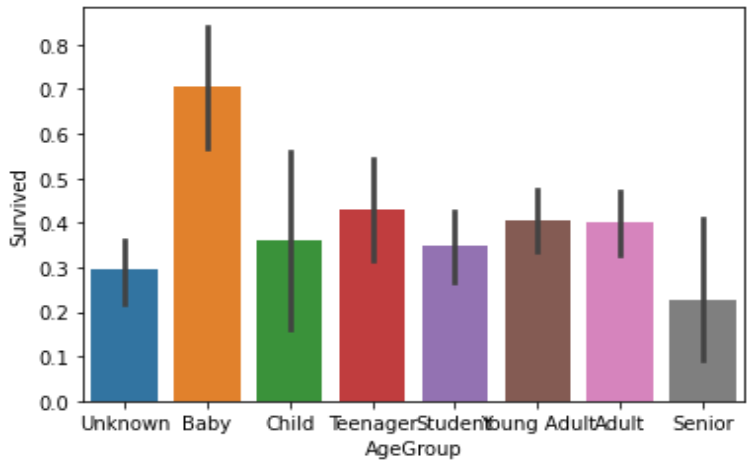
Training Data :
https://drive.google.com/file/d/1yWfF7Nq8FhJoabhWF59TulZv0iGKop0j/view?usp=sharing

Testing Data:
https://drive.google.com/file/d/1RnRSaIow9z0sDMrBt06tz73U5DM8uVfU/view?usp=sharing

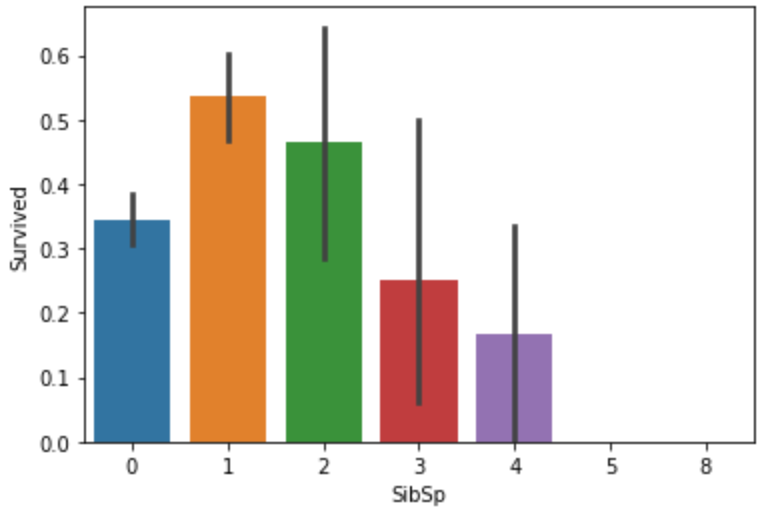
Tools Used:

1. **Numpy:** For faster array processing techniques.
2. **Pandas:** For data manipulation and analysis, opening csv files etc.
3. **Matplotlib:** For graph plotting purpose
4. **Seaborn:** For bar plotting.
5. **Sklearn:**Example: sklearn.ensemble, sklearn.linear_model, sklearn.neighbors, sklearn.tree, sklearn.svm, sklearn.naive_bayes

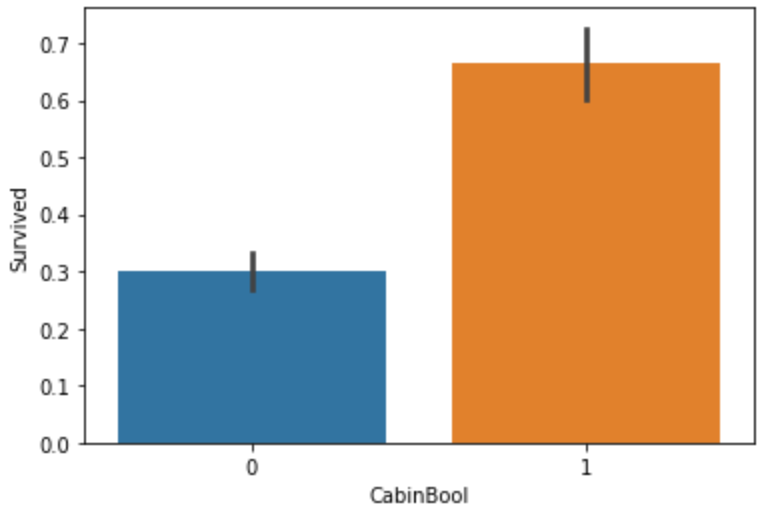
Visual Insights:



This visualizes how many categories of people was there.



This visualizes how many people with siblings numbers (0,1,2,3,4) survived.



This visualizes people surviving with cabin bool number 0 or 1. 1 is for elite people.

Accuracy of each model implemented:

	Model	Score
9	Gradient Boosting Classifier	81.01
8	Random Forest	78.21
7	Stochastic Gradient Descent	78.21
6	Logistic Regression	77.09
5	Gaussian Naive Bayes	77.09
4	Linear SVC	76.54
3	Support Vector Machines	75.98
2	KNN	75.42
1	Perceptron	74.86
0	Decision Tree	72.63

Output of the project:

To determine the amount of passengers survived and their details.

Output was exported to a csv file:
https://drive.google.com/file/d/1Qjeesv6KOr_tzlbgP40SEdvJi-STCTxA/view?usp=sharing

Data Insights:

Here are some insights I had on analysing the Data.

- Pclass, Sex, Cabin and Embarked are Categorical *features*.
- Comparing *Genders*, Females are way more likely to survive.
- The fares didn't contribute much to the model
- We decided to unite Age and Pclass due to the correlation with results
- Names are unique in the dataset, so they are useless without preprocessing
- Dividing age *feature* by groups is important to improve Machine Learning performance.

Conclusion:

Thanks to the teachers who guided me to complete this project. Online help was also taken from internet which helped me on every doubt i had during each step of the project. I mainly faced problem during the data analysis and data cleaning part as it has always been tedious task to determine the important features required and to clean unorganized data.

THANK YOU