In the Data processing part, our team dropped the cabin column because of lack of too much data and it doesn’t relate with survival. Furthermore, we replaced some missing age data with average values. According to the article “ 7 Ways to Handle Missing Values in Machine Learning” by Satyam Kumar, we learned that replacing missing values with mean is a statistical approach. For categorical data “embared”, we replace the missing part with “S” which is the most frequent one appearing in our data. Satyam points out the advantage of filling the missing data, he said it can prevent data loss which results in deletion of rows and columns.

Before we get to the model development, here is some brief description of the model. The dataset has numerical data: age of passengers, family member number and some categorical attributes:sex, ticekt class, fareness of passenger, port of embarkation.Then, our team encoded the categories attribute and finally we are ready to build our model.

We utilized LogisticRegression, both Naive Bayes, and Linear SVM models to find out the relationship between those attributes of passengers and if they survived. We got 77.01% accuracy with the logistic regression model. In this model, we examined the relationship between numerical and categorical attributes with survival. With 10 fold validation, we finally get 80.81% accuracy so it adds more validity to our result. Also, we also found that iterations does’t change the accuracy at a range of 100 to 900. Later, we apply the Categorical Naive Bayes Classifier model and linear SVM model. After converting the numbercial attributes to categorical data, the model had 79.32% accuracy and we got 81.16% accuracy for the 10 fold validation test. The accuracy of this model is slightly higher than our logistic regression model. However, the Linear SVM model only got 73.74% accuracy which has the lowest accuracy between these ML models. Finally, according to the visualization plot for correlation, we can clearly see that Survival is correlated to age, fare, class, and sex.

Works Cited

Kumar, Satyam. “7 Ways to Handle Missing Values in Machine Learning.” *Medium*, Towards Data Science, 2 Aug. 2020, towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.