MACHINE LEARNING ALGORITHMS

Project Summary:

The objective of the project is to create the best fit model which can predict whether a person will survive or not utilizing Titanic Dataset
The Models used to compare are figure out the best models are: Logistic Regression, Support Vector Machine, Multilayer Perceptron, Random Forest & Boosting

Developed platform: Jupyter Notebook – Python Version 3.7.4   **Date:** 07/06/2020

Developed by:
Jeyamaruthi Jayakumar
ID: 1001757737
Department of Computer Science & Engineering
The University of Texas at Arlington
Portfolio: https://jeyamaruthi.uta.cloud/blog/
Email ID: Jeyamaruthi.jayakumar@mavs.uta.edu

1   Project Description

The objective of the project is to create the best fit model which can predict whether a person will survive or not utilizing Titanic Dataset

The Models used to compare are figure out the best models are: Logistic Regression, Support Vector Machine, Multilayer Perceptron, Random Forest & Boosting.

URL: https://jeyamaruthi.uta.cloud/blog/

2   Detailed Description

This dataset contains information about 891 people who were on board the ship when departed on April 15th, 1912. As noted in the description on Kaggle's website, some people aboard the ship were more likely to survive the wreck than others. There were not enough lifeboats for everybody so women, children, and the upper-class were prioritized. Using the information about these 891 passengers, the challenge is to build a model to predict which people would survive based on the following fields:

- **Name** (str) - Name of the passenger
- **Pclass** (int) - Ticket class
- **Sex** (str) - Sex of the passenger
- **Age** (float) - Age in years
- **SibSp** (int) - Number of siblings and spouses aboard
- **Parch** (int) - Number of parents and children aboard
- **Ticket** (str) - Ticket number
- **Fare** (float) - Passenger fare
- **Cabin** (str) - Cabin number
- **Embarked** (str) - Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Firstly, we cleanse and clean the data, to make it compatible to train the model. You can see that in '**01_Foundations**'.

'**02_LR**' => 1.) Here we had created a Logistic Regression model utilizing Sklearn. Then we train the model with GridSearchCV with iteration = 5.

2.) Then we run the model with different parameter types to check the best possible prediction score for Logistic Regression. Then, we utilize the library joblib & the function best_estimator_ to get the finest parameters for Logistic Regression.

3.) Finally, we export the model to LR_Model.pkl

Similarly, we do the above **three** steps for the other 4 models i.e. we follow the same for Support Vector Machine, Multilayer Perceptron, Random Forest & Boosting.

Also, we export the model as {Name}_Model.pkl (where Name = SVM, MLP, RF, GB)

With utilizing the models and dataset we generated through above steps, we evaluate each model and compares its Accuracy, Precision, Recall & Latency.

After doing so, we figure out that the Random Forest gives an Accuracy of 81% & Precision of 87.5% making it has the best model to predict utilizing the given data.

3   Conclusion

Hence, we were able to find the best fitting model for Titanic Dataset by comparing the Accuracy, Precision, Recall & Latency of each model.