**Big Data Assignment – II**

**Titanic: Machine Learning from Disaster, Predict Survival on the Titanic**

**Model’s and Accuracy’s**

|  |  |
| --- | --- |
| **Model** | **Accuracy** |
| *Random Forest Classifier* | *0.7870036101083032* |
| *Random Forest with Tuning (k = 5)* | *0.8194945848375451* |
| *Decision Tree Classifier* | *0.8194945848375451* |
| *Decision Tree Classifier with Tuning (k = 5)* | *0.8267148014440433* |
| *Logistic Regression Classifier* | *0.7870036101083032* |
| *Logistic Regression Classifier with Tuning (k = 3)* | *0.8267148014440433* |
| *Multilayer Perceptron Classifier* | *0.75 – 0.78* |

**Steps**

**1.Preprocessing**

I loaded Train.csv and removed the ‘titles’ from each name by using two splits and then I categorized these names into four categories as ‘Boy’, ‘Girl’, ‘Lady’, ‘Sir’ based on conditions using regex\_replace and when function. Next, I found family size for each name by using a UDF function that calculates family size = Sibsp + Parch + 1. I handled null values in the ‘Fare’ column and ‘Age’ column by filling with their respective average values. I used String Indexer and One Hot Encoder on the title column and sex column to get vector values of my categorical columns then I used vector assembler to put together my features that contains my numerical column’s and newly generated vectorised columns as input.

Features = ["titleVec", "SexVec", "Pclass", "Fare", "familysize", "Age"]

**2.Model Building**

I split the pre-processed training data into ‘train’ and ‘test’ with percentage splits of 70% for my train and 30% for my test for my first three models (random forest, decision tree and logistic regression) but, for my multilayer perceptron I used a 60:40 split for the ‘train’ and ‘test’ and this improved my accuracy score. I fitted the models with features as the vector assembled features and prediction as ‘Survived’ on my ‘train’ and predicted the values for ‘Survived’ in my ‘test’. I evaluated the accuracy of the model using Multiclass Classification Evaluator. I tuned my first three models (random forest, decision tree, logistic regression) by passing in parameters and by using cross-validation with different values of k = 3,5 and selected k for the model that gave me the highest accuracy.

**3.Model Selection**

Among the four models, I would choose either choose Logistic Regression Classifier with k = 3 or Decision Tree Classifier with k = 5 because they are giving me the highest accuracy of **‘*0.8267148014440433’***. If I have a personal choice between the two I would choose the decision tree classifier as it will show a tree structure in the order of importance of my features that are useful in predicting the survival of the people on Titanic and this will help in understanding and interpreting more which features matter the most if they are to survive from the disaster. Basically, it would help me create rules using my features to predict survival.

Finally, I will pre-process the actual test data by following the same steps I followed on my train data and use the selected model (Decision Tree Classifier with k = 5) to predict the values of the column ‘Survived’ and calculate accuracy.