**Milestone Report – 2**

**Capstone Project: News Article Classification**

* **Dayakar & Krupasindhu (18th Oct 2021)**

1. **Project Architecture Design Completed** (Attached .PPT document for reference)
2. **ML Approach Documentation completed** (Attached .PPT document for reference)
3. **Developed Python script for Data preparation**
4. **Install python mongodb connectors for connecting with Pyspark**

**\* !pip3 install pymongo[srv]**

**\* !pip3 install Pyspark**

**\* !pip3 install mlflow**

**\* Load the data from mongodb**

**\* Perform EDA to study the data**

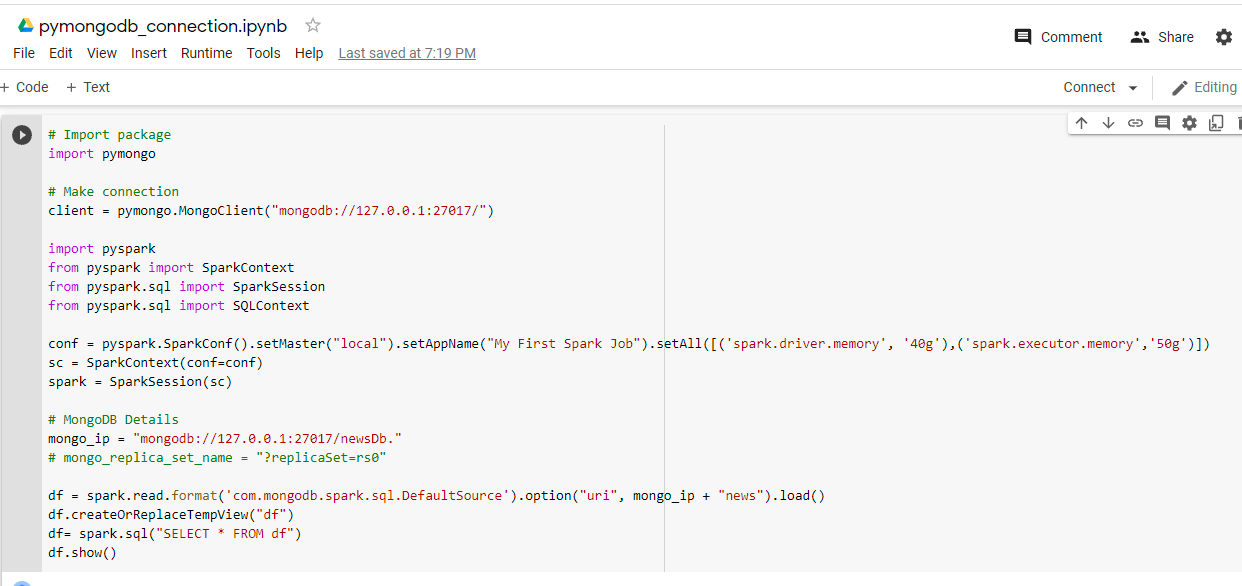
**\* Select the relevant columns for modeling**

**\* Split the data for train and test**

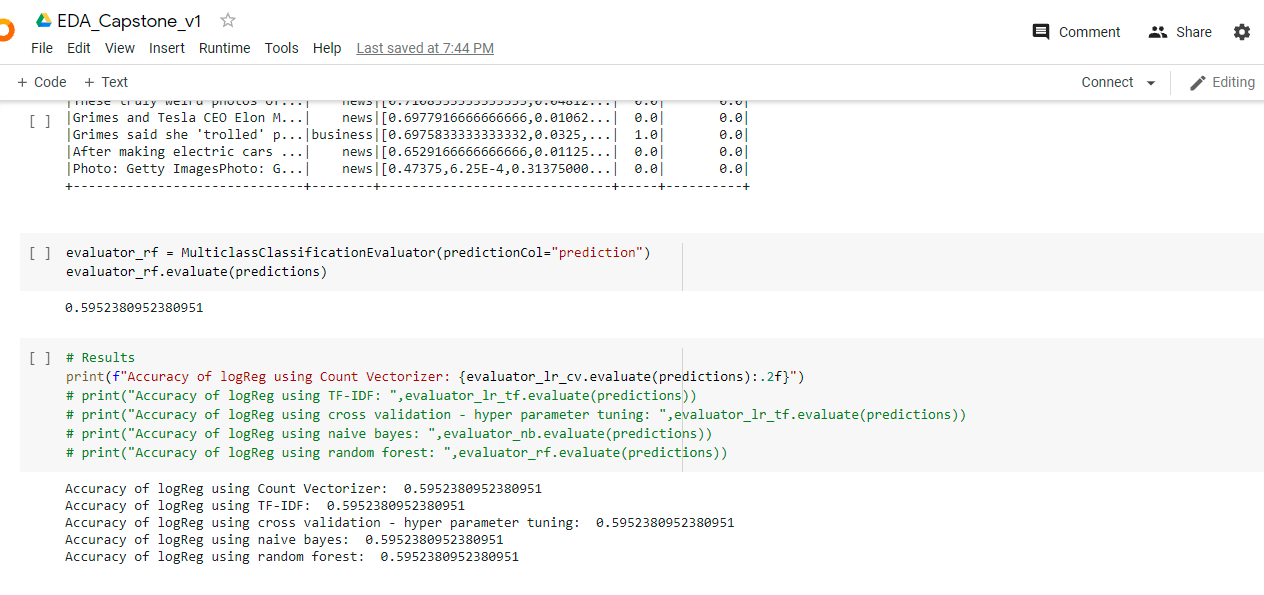
**\* Model the data with different algorithms**

**\* Evaluate the models and select the best performing and optimized model**

**\* Save the model using pyfunc/mlflow**



1. **Results after the EDA done on the sample data:**



1. **GitHub Link / URL:**

<https://github.com/bksindhu/Capstone---News-articles-classifier/tree/main/Milestone2%20-%20Data%20Preparation>

<https://github.com/DayakarKodirekka/Capstone---News-articles-classifier>

1. **Challenges:**

Connecting to mongodb using pyspark