**CS5560 Knowledge Discovery Management**

**In Class Assignment # 12**

**23rd April 2020**

Name: Srichakradhar Reddy Nagireddy

ID: 11

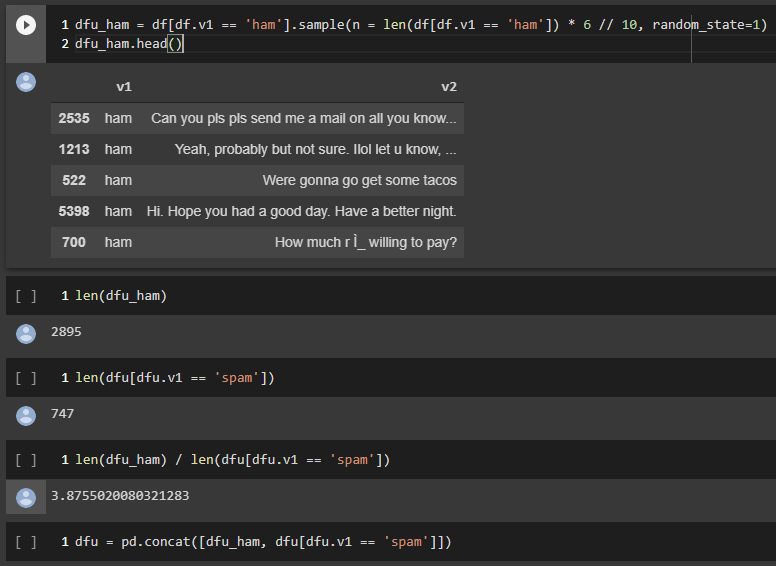
Perform the following:

1. The data in spam.csv file has a class imbalance issue. Usually there are various techniques to handle class imbalance in the data, but the two widely used one are:
   1. Random Oversampling: Randomly duplicate examples in the minority class.
   2. Random Undersampling: Randomly delete examples in the majority class.
2. ICP Tasks:

* Solve the data imbalance issue with either random Undersampling or random oversampling.
* Report on new class ratio

# Random Under sampling:

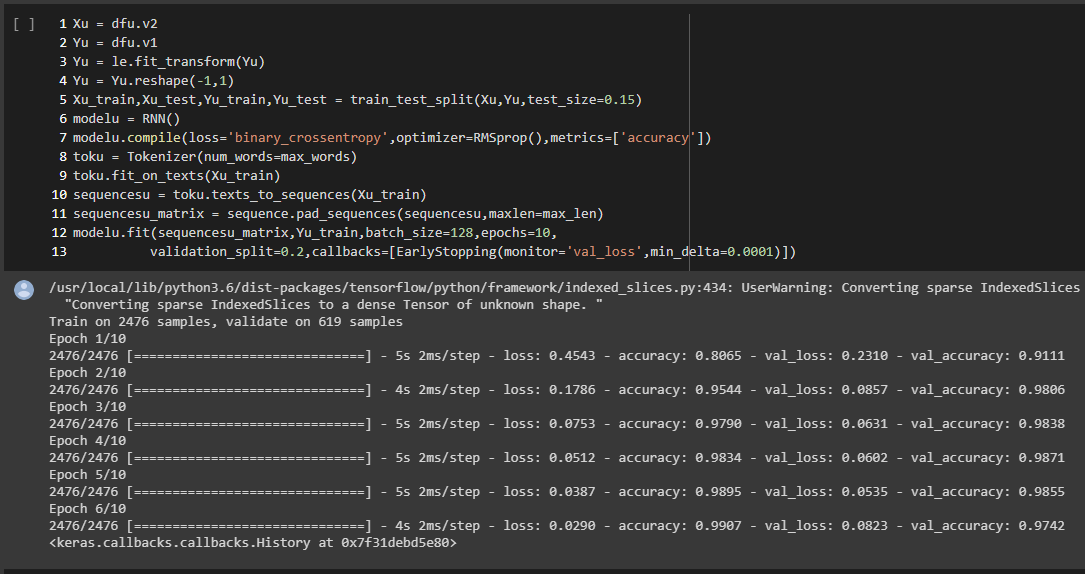
## Create the dataset



New class **Ratio** = 3.87 : 1

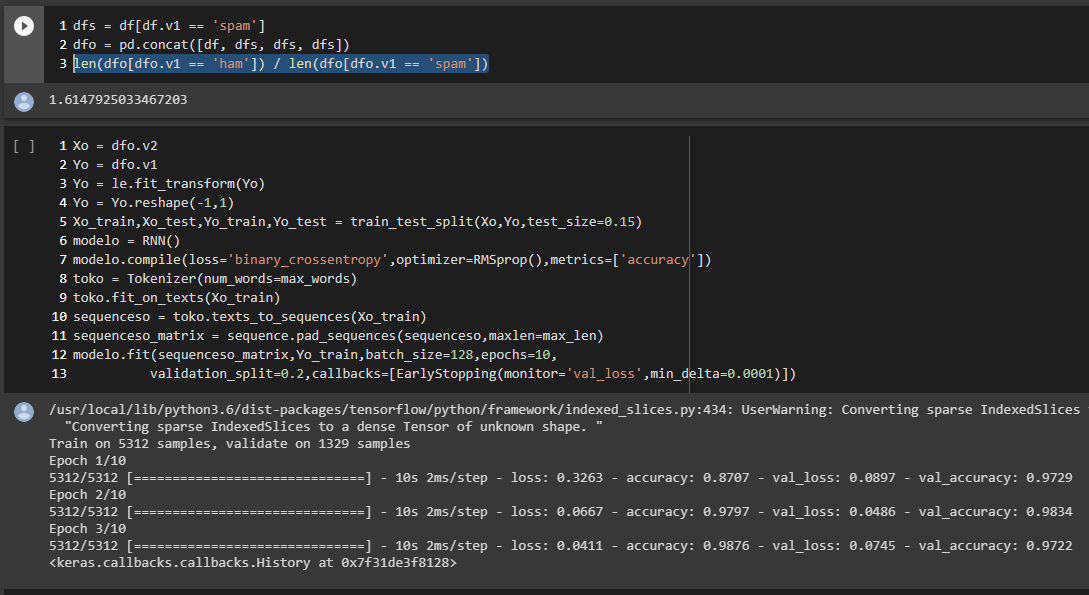
**Accuracy**: **97%** vs 98% for default model (reduced – may be because of underfitting)

## Compare the model performance on this modified data.



# Random Over sampling:

## Create the dataset and compare the model performance on this modified data



New class **Ratio** = 1.61 : 1

**Accuracy**: **97%** vs 98% for default model (reduced – may be because of overfitting)