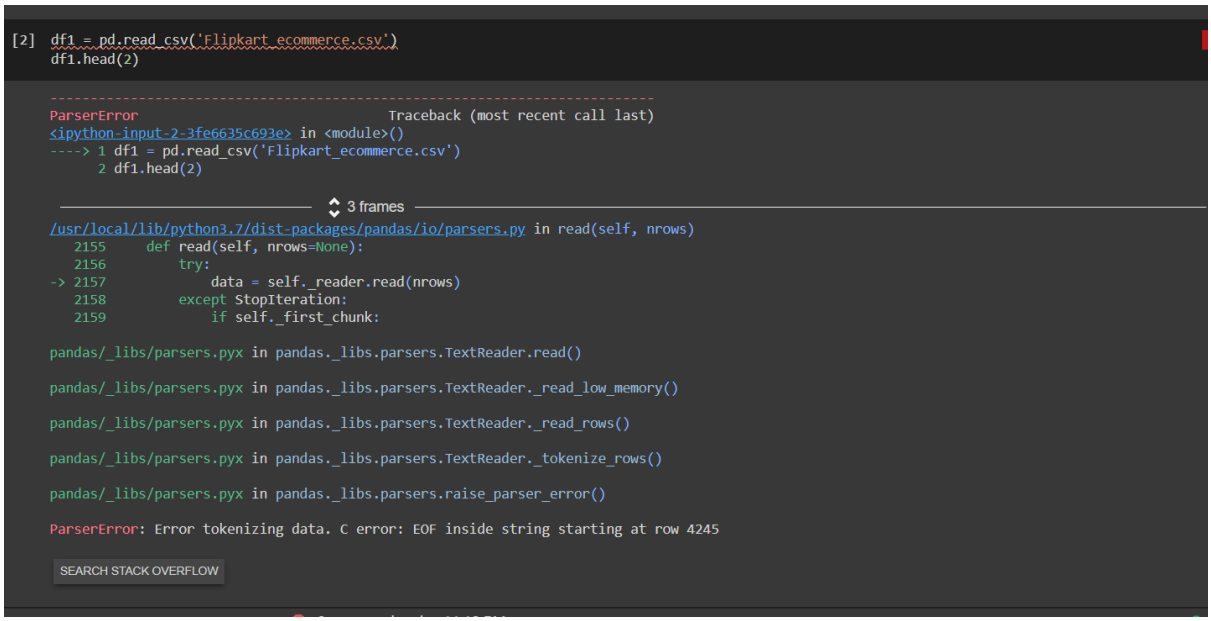
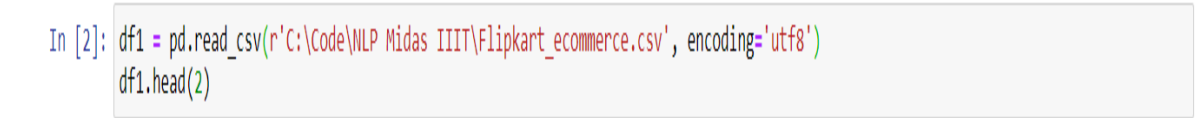
Errors

This error occurred sometimes and can be resolved by adding encoding = ‘utf8’

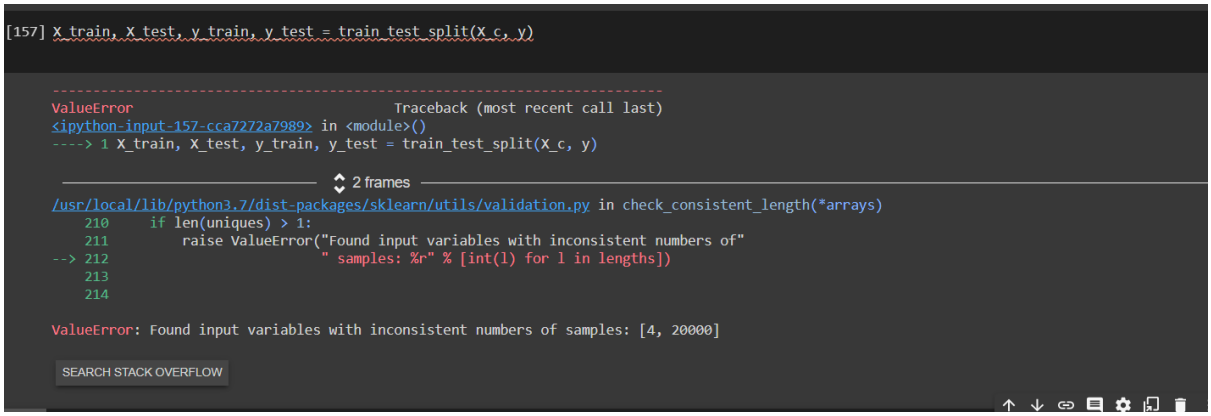


```
[2] df1 = pd.read_csv('Flipkart_ecommerce.csv')
    df1.head(2)

---------------------------------------------------------------------------
ParserError                               Traceback (most recent call last)
<ipython-input-2-3fe6635c693e> in <module>()
----> 1 df1 = pd.read_csv('Flipkart_ecommerce.csv')
      2 df1.head(2)

                                  3 frames
/usr/local/lib/python3.7/dist-packages/pandas/io/parsers.py in read(self, nrows)
   2155     def read(self, nrows=None):
   2156         try:
-> 2157             data = self._reader.read(nrows)
   2158         except StopIteration:
   2159             if self._first_chunk:

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader.read()

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_low_memory()

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._read_rows()

pandas/_libs/parsers.pyx in pandas._libs.parsers.TextReader._tokenize_rows()

pandas/_libs/parsers.pyx in pandas._libs.parsers.raise_parser_error()

ParserError: Error tokenizing data. C error: EOF inside string starting at row 4245

SEARCH STACK OVERFLOW
```



```
In [2]: df1 = pd.read_csv(r'C:\Code\NLP Midas IIIT\Flipkart_ecommerce.csv', encoding='utf8')
        df1.head(2)
```

By using Tfidf the dimension of X was changing therefore while implementing train_test_split it gave an error.

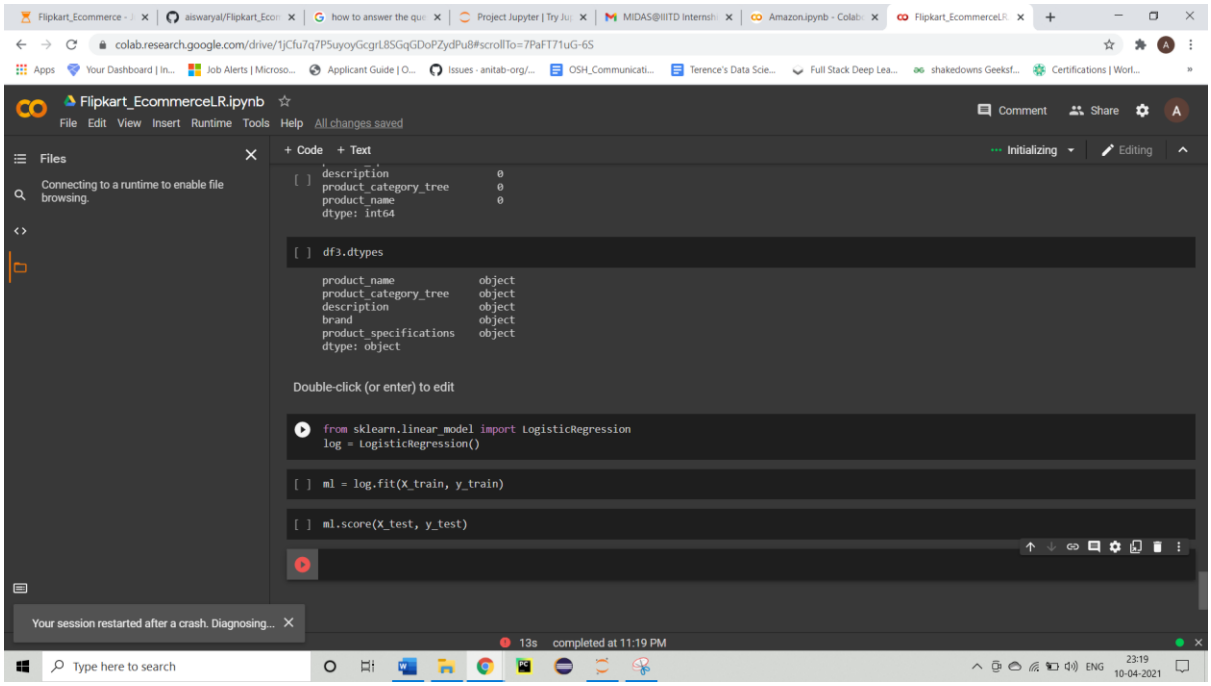


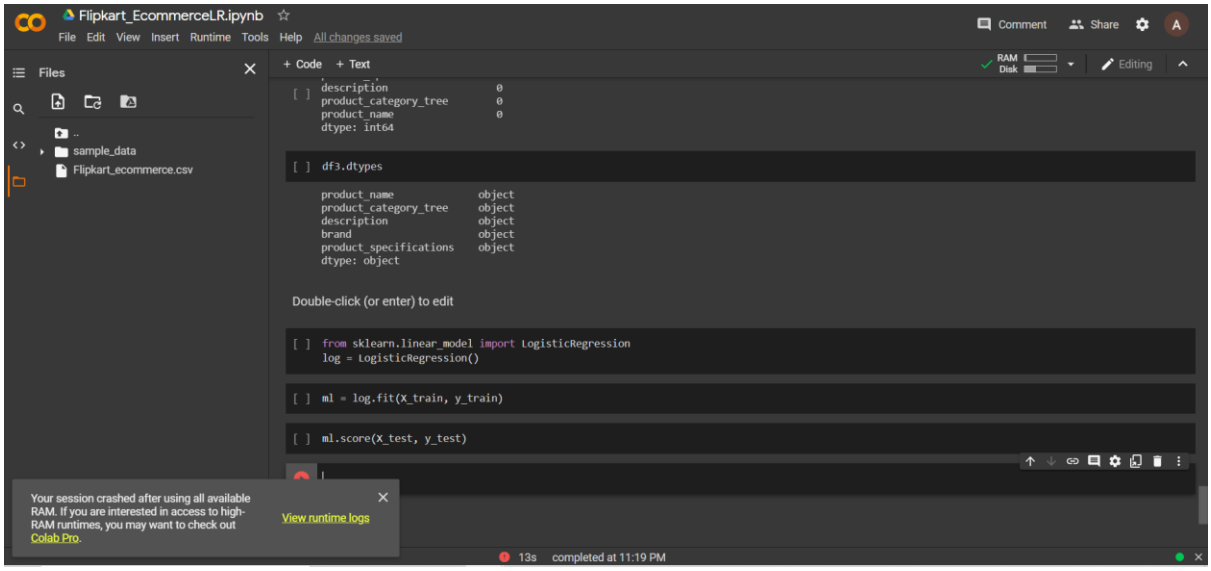
```
[157] X_train, X_test, y_train, y_test = train_test_split(X_c, y)

---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-157-cca7272a7989> in <module>()
----> 1 X_train, X_test, y_train, y_test = train_test_split(X_c, y)

                                  2 frames
/usr/local/lib/python3.7/dist-packages/sklearn/utils/validation.py in check_consistent_length(*arrays)
    210     if len(uniques) > 1:
    211         raise ValueError("Found input variables with inconsistent numbers of"
--> 212                          " samples: %r" % [int(l) for l in lengths])
    213
    214

ValueError: Found input variables with inconsistent numbers of samples: [4, 20000]

SEARCH STACK OVERFLOW
```

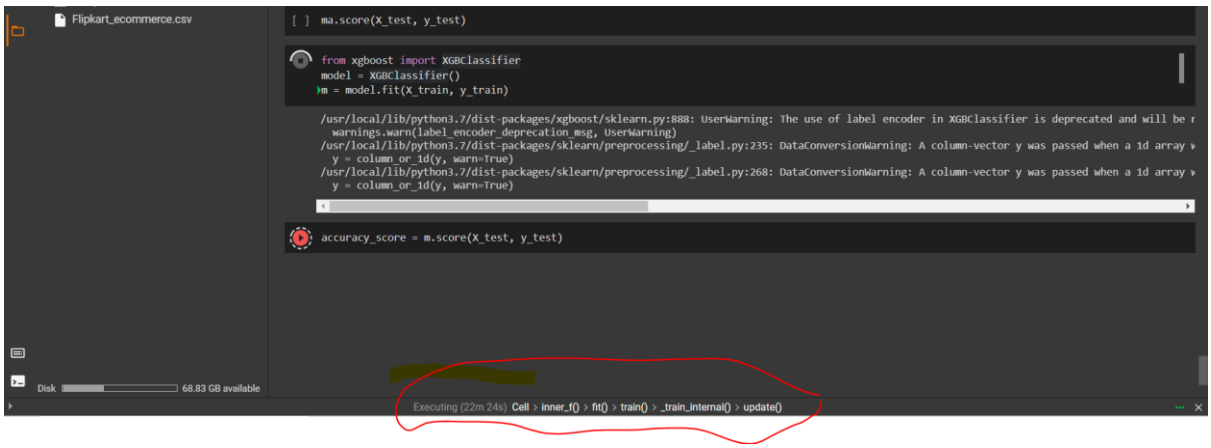
```
could not convert string to float: '["clothing >> men\'s clothing >> inner wear & sleep wear >> vests >> l\'appel du vide vests >> l\'appel du vide men\'s vest (pack of 2)"]'
```

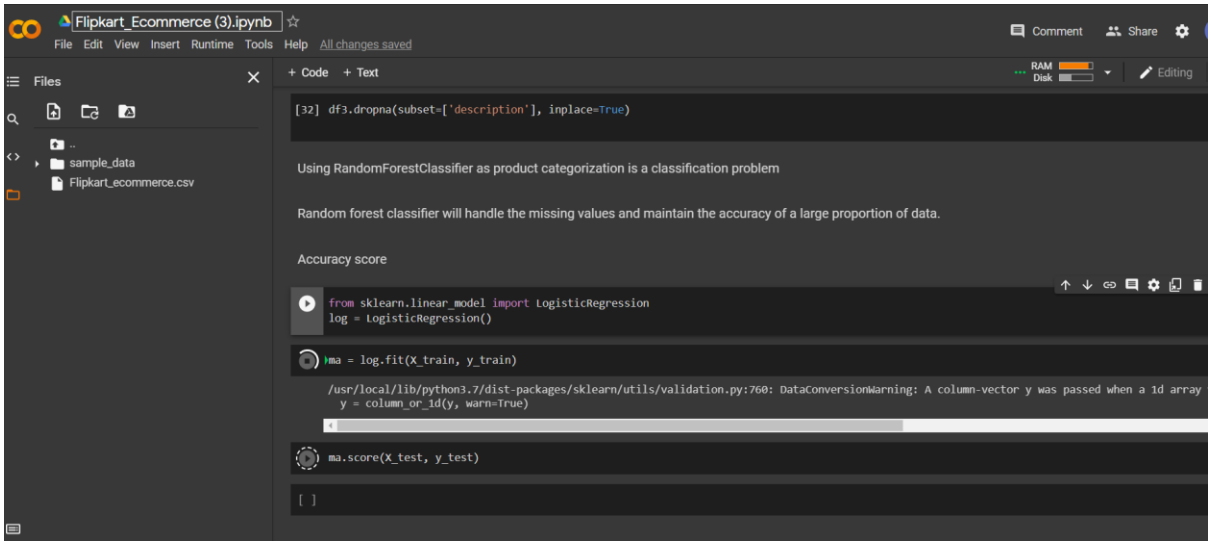
```
[42] from sklearn.ensemble import RandomForestRegressor
     RF_clf = RandomForestRegressor(n_estimators=10)

     ml = RF_clf.fit(X_train, y_train)

     /usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:4: DataConversionWarning: A column-vector y was passed when a 1d array was expected.
       after removing the cwd from sys.path.
     ---------------------------------------------------------------------------
     ValueError                                Traceback (most recent call last)
     <ipython-input-42-a41ea83fb0b5> in <module>()
           2 RF_clf = RandomForestRegressor(n_estimators=10)
           3
     ----> 4 ml = RF_clf.fit(X_train, y_train)

                                    1 frames
     /usr/local/lib/python3.7/dist-packages/numpy/core/_asarray.py in ascontiguousarray(a, dtype)
         175
         176     """
     --> 177     return array(a, dtype, copy=False, order='C', ndmin=1)
         178
         179

     ValueError: could not convert string to float: '["jewellery >> necklaces & chains >> necklaces"]'

     SEARCH STACK OVERFLOW
```

Taking very long to respond and many times it says “Your session crashed after using all available RAM.” On Google colab









Even on Jupyter notebook it takes a very long time and the code is not executed.

```
In [*]: !pip install xgboost

In [*]: !pip install --upgrade xgboost

In [*]: import re

        regex = re.compile(r"\[|\]|<", re.IGNORECASE)

        X_train.columns = [regex.sub("_", col) if any(x in str(col) for x in set(('[', ']', '<'))) else col for col in X_train.columns.va

In [*]: from xgboost import XGBClassifier
        model = XGBClassifier(n_estimators = 10)
        ml = model.fit(X_train, y_train)

In [*]: accuracy_score = ml.score(X_test, y_test)

In [*]: from sklearn.linear_model import LogisticRegression
        log = LogisticRegression()

In [*]: ml = log.fit(X_train, y_train)

In [*]: ml.score(X_test, y_test)

In [ ]:
```

I was only able to implement RandomForestClassifier(). Other Machine learning Algorithms could not be implemented because RAM was exhausted in the middle of execution of code.

Image can also be used during prediction. I did not use it because I do not have enough knowledge in that domain and I am learning about it.