

**School of Computing Science and Engineering**

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

**Review III Report**

**Programme:** MTECH CSE with Specialisation in Big Data Analytics

**Course:** NoSQL Databases (CSE6006)

**Slot:** E1

**Faculty:** Dr. A. Bhuvaneswari

**Component:** J

**Title: Predicting IMDB Rating and Spoilers Detection of Movies using Machine Learning and Deep Learning**

**Team Member(s):** Kush Desai (20MCB1002)

Raginee Titar (20MCB1004)

Raviraj Patil (20MCB1016)

**Abstract:**

The film world isn't just an industry or a focal point of diversion, rather it is currently a focal point of worldwide business. Everywhere in the world is presently amped up for a's film industry achievement, ubiquity and so on A colossal information is accessible online about these films prosperity or prominence. We have utilized film dataset from mainstream site Kaggle, put away in NoSQL information base. Then different machine learning algorithms such as KNN, random forest are applied to the data set. The model gives good classification measures with the data set and we can also derive the spoilers with the help of natural language processing which was given in comments by users.

1. **Introduction**
   1. **Objective and goal of the project**

Study the different machine learning models and deep learning models for predicting Movies’ Imdb ratings and spoiler detection respectively.

* 1. **Problem Statement**

(a) Predicting IMDb Rating of Movies by Machine Learning Methods.

(b) Spoiler Detection using Deep Learning

* 1. **Motivation**

(a) Can you utilize the metadata to identify the imdb score?

(b) Apart from deriving imdb scores, the metadata available can be used for other tasks as well like spoiler detection etc.

* 1. **Challenges**

(a) Use of mongo db.

(b) predict that which movie is in which genre.

(c) Deal with the spoilers.

1. **Literature Survey**

|  |  |  |
| --- | --- | --- |
| Paper  No. | Methodology | Results |
| [1] | Word2vec + Linear SVM, Bag of Words + Random Forest, Word2vec+Random Forest, (Word2vec+Bag of Words) + Random Forest approaches used for feature vectors and introduced three unique classification strategies. | the (Word2vec+Bag of Words) + Random Forest model has given the maximum accuracy of 84.14 percent. |
| [2] | Author has used Bi-LSTM network and RNN-s. In this study, they perform movie genre classification from plot outlines using Bi-LSTM network. | Rather than utilizing entire plot rundown as info, partition it into its sentences and train the organization utilizing those sentences. |
| [3] | Author proposed a classification plan of pre-discharge film prevalence dependent on innate traits utilizing plan of pre-discharge film prominence dependent on natural credits utilizing C4.S and PART classifier calculation and characterized the connection between properties of post-delivery movies using correlation coefficient. | movie popularity classified with 77% accuracy. The classification accuracy of both the decision tree and rules showed the strength of chief position along with spending plan. |
| [4] | Author performed analysis of movie  reviews using a combination of deep learning techniques, LSTM and CNN. CNN is a commonly used deep learning method used for feature learning. LSTM is a unique sort of RNN, which is equipped for adapting long term conditions. Proposed model allowed remembering of information over a much longer period | LSTM–CNN provided better accuracy with an overall accuracy of 79% with compare to CNN-LSTM. |
| [5] | Author used Linear Regression, Factorization Machines techniques in order to predict movie success by anticipating IMDb appraisals for newly delivered films utilizing online media information and contrast it with current examinations. Fostered a system to assemble the film information from various sources including online media. | Factorization Machines approach is utilized to foresee film accomplishment by anticipating IMDb appraisals for recently delivered motion pictures by consolidating film metadata with web-based media information. |
| [6] | Data collected from two platforms, Wikipedia and IMDb. Author used various classification algorithms like Random forest, Naïve bayes etc. to predict movie ratings. To make the data set a balanced one, SMOTE is applied afterwards. Thus, the accuracies increase. | Random  forest, J48 and bagging gives same classification accuracy.  . |
| [7] | The text blob tool in python is used to correct misspelled words in user text. Author used RNN, Naïve bayes and LSTM for classification. | TF IDF Representation Of a review performed better  than Bag of words. NAVIE BAYES performed well in  machine learning algorithms and LSTM in deep learning  algorithms in the context of the text |
| [8] | Author proposed method to unveil the important factors influencing  the score of IMDb Movie Data.Author used SVM, Adaboost, Random forest classified to predict the movie score. | Random Forest addressed the film includes all the more precisely. The achievement rate for all models are better in contrast with the past investigations. The results obtained are better than that of some standard libraries and common studies. |
| [9] | Author designed and implemented  a tool to remove the IMDb dataset files and import them into a database. This apparatus utilizes archive situated information structures, and permits others to expand the code to change structures dependent on their necessities. At last, a web interface to perform inquiries against the import information to validate the import process was also developed. | delivered a website which performs  various queries on some of the data collections, which validates the imported data. |
| [10] | Author did a contextual investigation for the representation and examination of huge and complex worldly multivariate organizations determined from the Internet Movie Database (IMDB). Author incorporated network examination strategies with perception to address adaptability and intricacy issues. He applied island investigation for a particular time frame cut all together to distinguish significant and significant subgraphs. Further, a transient Kevin Bacon chart and a worldly two mode network are separated to give understanding and information on the development. | Shown some patterns in the evolution. Visualization with galaxy of movies is shown in the network of specific time slice. |

1. **Data Set Description:**

**Dataset source :**

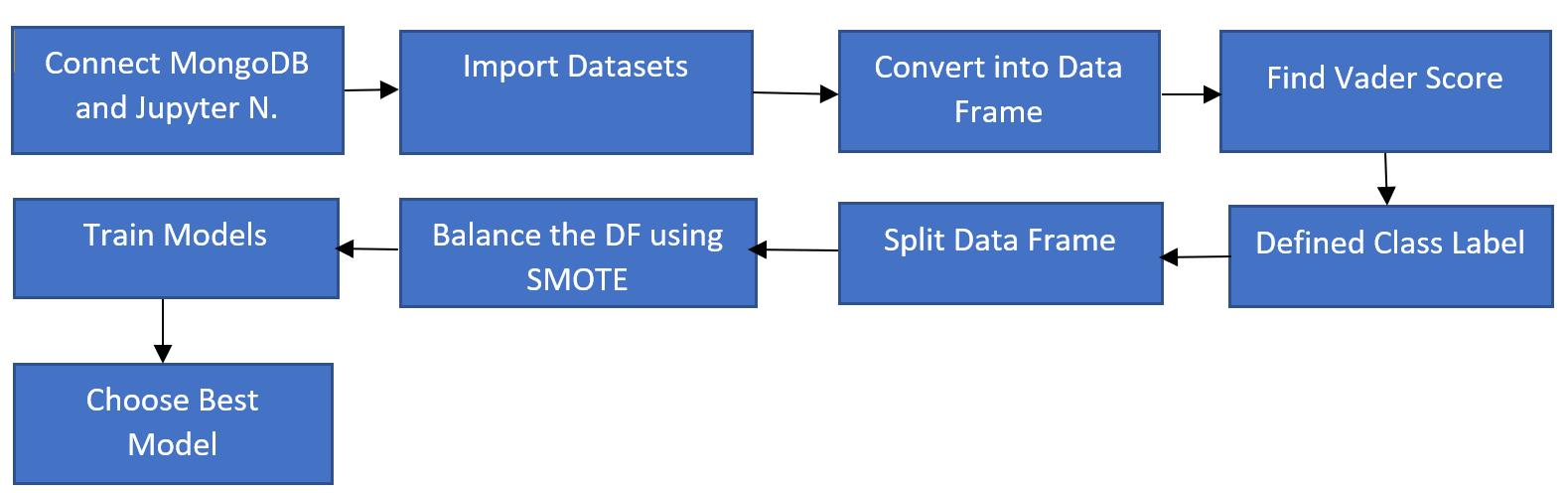
[**https://www.kaggle.com/rmisra/imdb-spoiler-dataset?select=IMDB\_reviews.json**](https://www.kaggle.com/rmisra/imdb-spoiler-dataset?select=IMDB_reviews.json)

**Dataset information :**

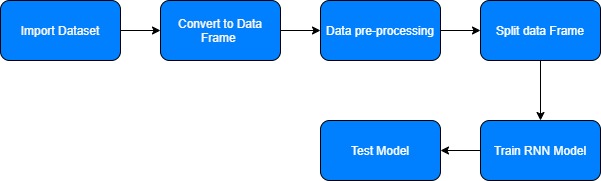
Dataset is generated from user review on movie and tv show. The dataset consists of 7 entity and around 570000 rows, The review consists of understanding of the review data is spoil or not. The user wants to watch a movie or tv show, the spoiler may reduce interest of show. Therefore, dataset divided into spoil or not to understand the spoil review given by the user.



1. **Methodology and Algorithm used:**
   1. **System Architecture diagram**



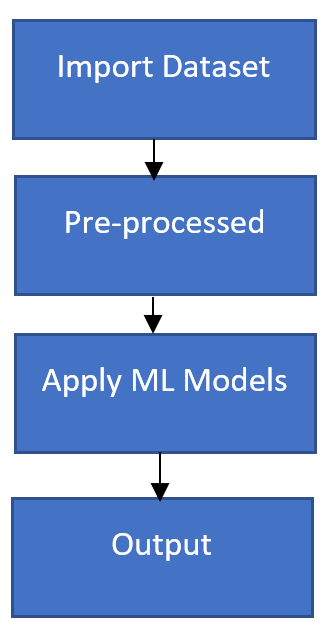
Predicting IMDB Rating



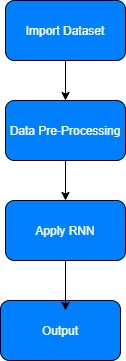
Spoilers Detection

* 1. **Flowchart**

A. Predicting IMDB Rating



B. Spoiler Detection



* 1. **Algorithm design**

**XGB boost:**

XGBoost is a choice tree-based gathering Machine Learning calculation that uses an inclination boosting system. In expectation issues including unstructured information (pictures, text, and so on) counterfeit neural organizations will in general beat any remaining calculations or systems. In any case, with regards to too little to-medium organized/even information, choice tree-based calculations are viewed as top tier at this moment. If it's not too much trouble, see the graph beneath for the advancement of tree-based calculations throughout the long term.

**Random Forest:**

Random forest, similar to its name suggests, comprises of countless decision trees that work as a troupe. Every individual tree in the random forest lets out a class forecast and the class with the most votes turns into our model's expectation.

The principal idea driving random forest is a straightforward however amazing one the astuteness of groups.

**K Nearest Neighbour:**

K-nearest neighbors (KNN) calculation is a sort of directed ML calculation that can be utilized for both classifications just as regression prescient issues. Be that as it may, it is principally utilized for characterization prescient issues in the industry. K-nearest neighbors (KNN) calculation utilizes 'highlight closeness' to expect the potential gains of new data centers which further infers that the new data point will be given out a value reliant upon how eagerly it arranges with the concentrations in the planning set.

1. **Experimental setup**

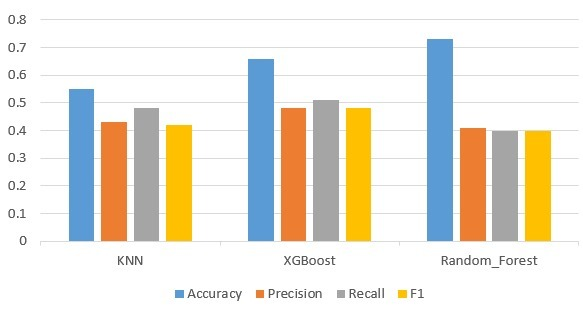
5.1 **Hardware Requirements**

Laptop with windows 10 OS

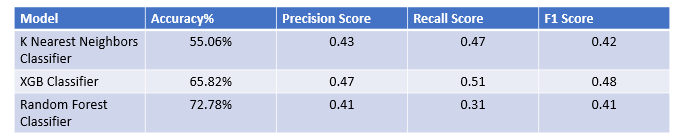
5.2 **Software Requirements**

mongo db, Jupiter Notebook.

1. **Results and Discussion** 
   1. **Discuss on performance metrics.**



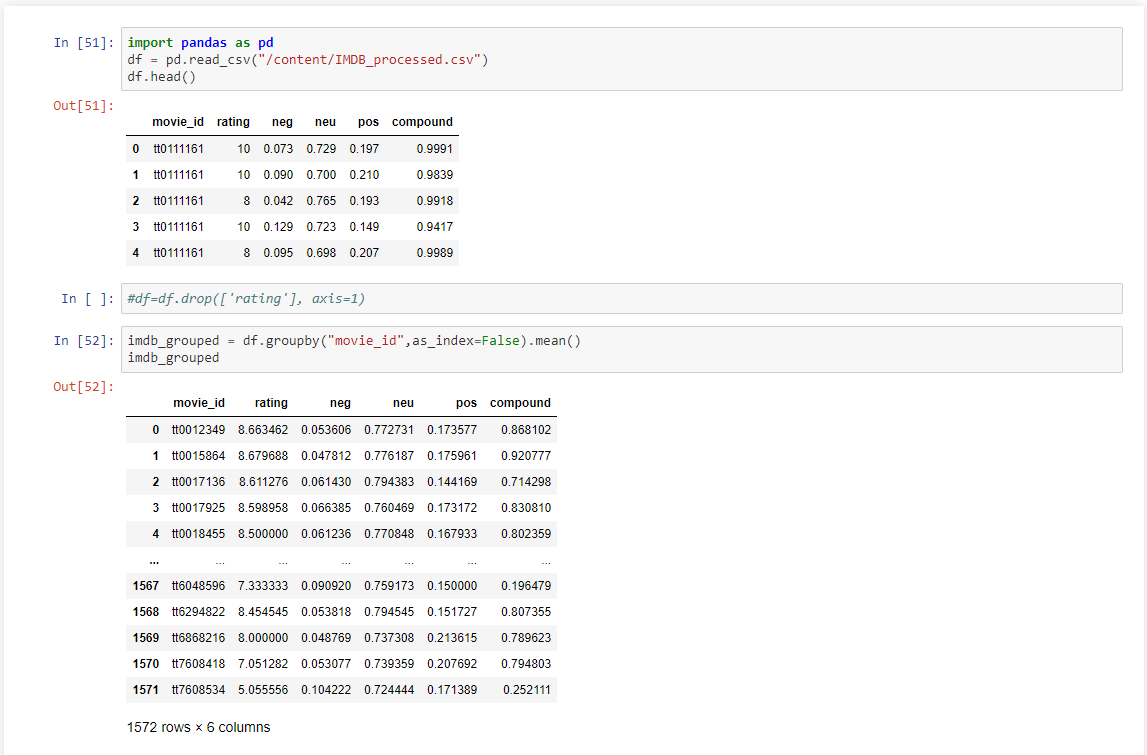
* 1. **Tables**

 As we can see in the above diagram Random Forest give better accuracy as compared to XGBoost and KNN algorithm, Random Forest gives 72% accuracy with Precision Score 41%, Recall Score 39%, and F1 Score 40% but XGBoost provide better F1 Score, Precision, Recall Score compare to Random Forest. As shown in the above diagram, XGBoost gives 65% accuracy and Precision Score, Recall Score, and F1 Score as 0.47, 0.51, and 0.48 respectively. KNN provides the lowest accuracy for a given dataset.

1. **Screenshots of the project**

**7.1. Predicting IMDB Rating**

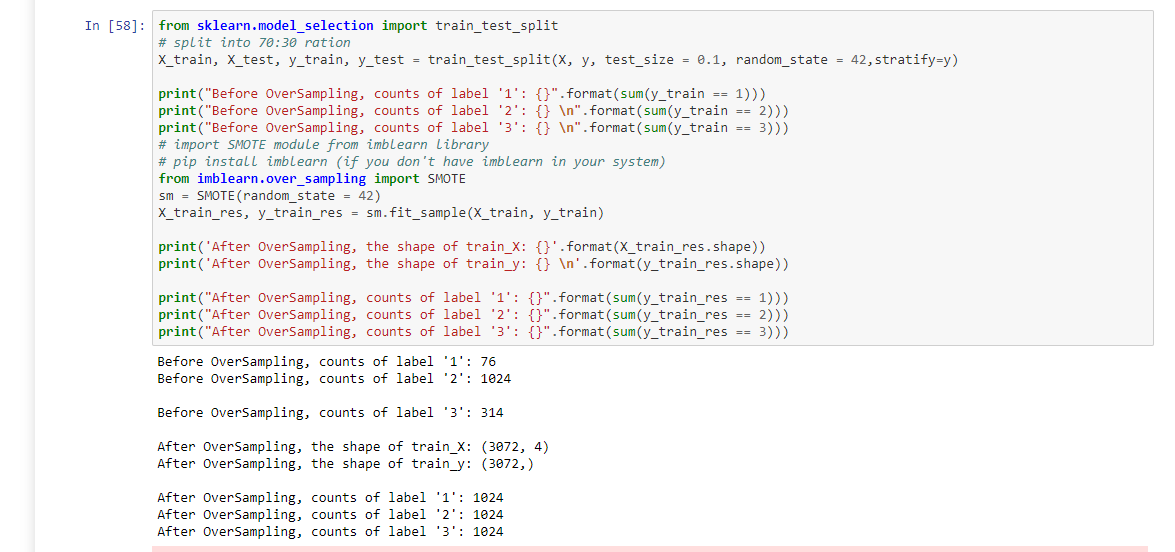
* **Import dataset**



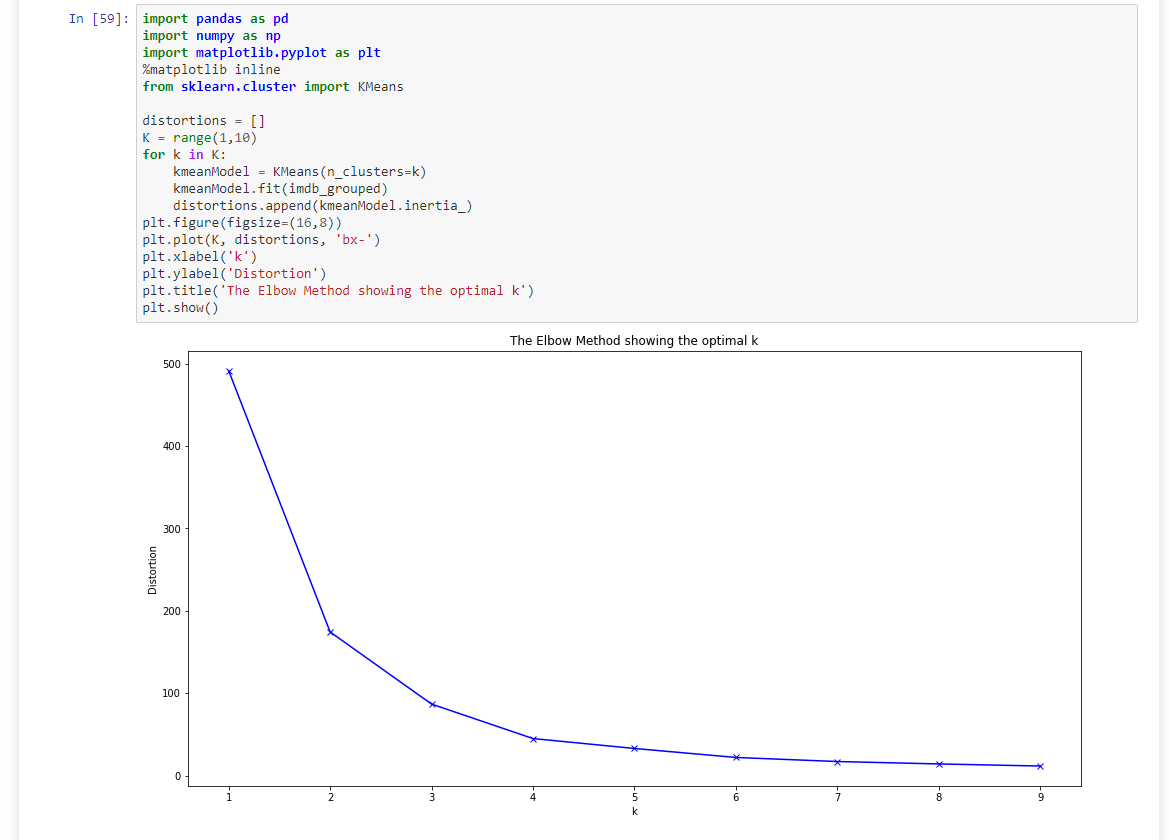
* **Create class label from rating table.**

****

* **Split dataset into train and test**

****

* **Apply elbow method to find k value**

****

* **Train KNN classifier model**

****

* **Train XGBoost model**

****

* **Train Random Forest Classifier model**

****

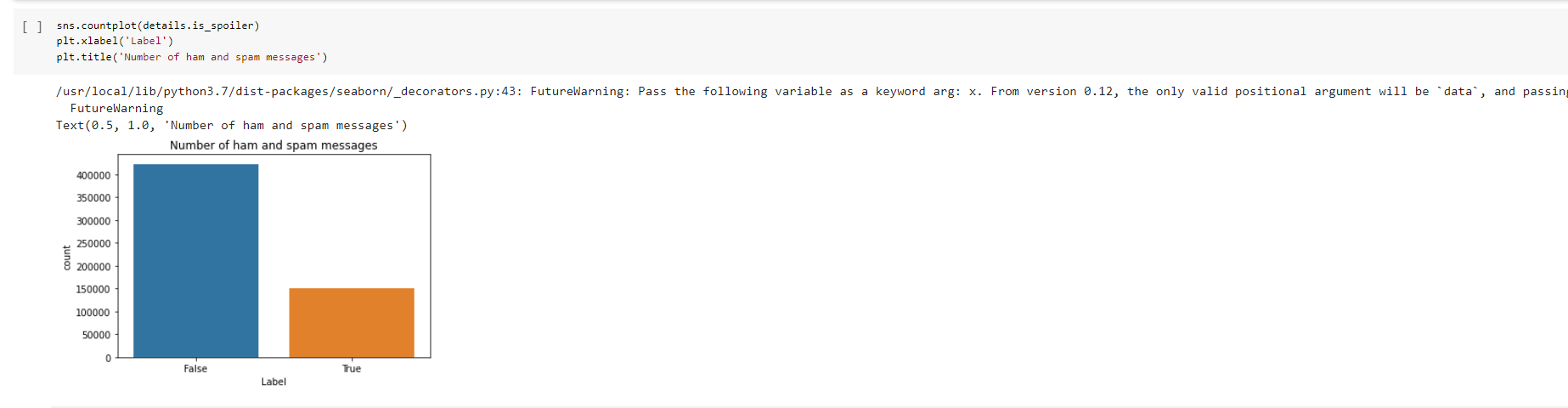
**7.2 Spoilers Detection**

* **Import dataset**

****

****

**Plot rainfall tomorrow True VS False**

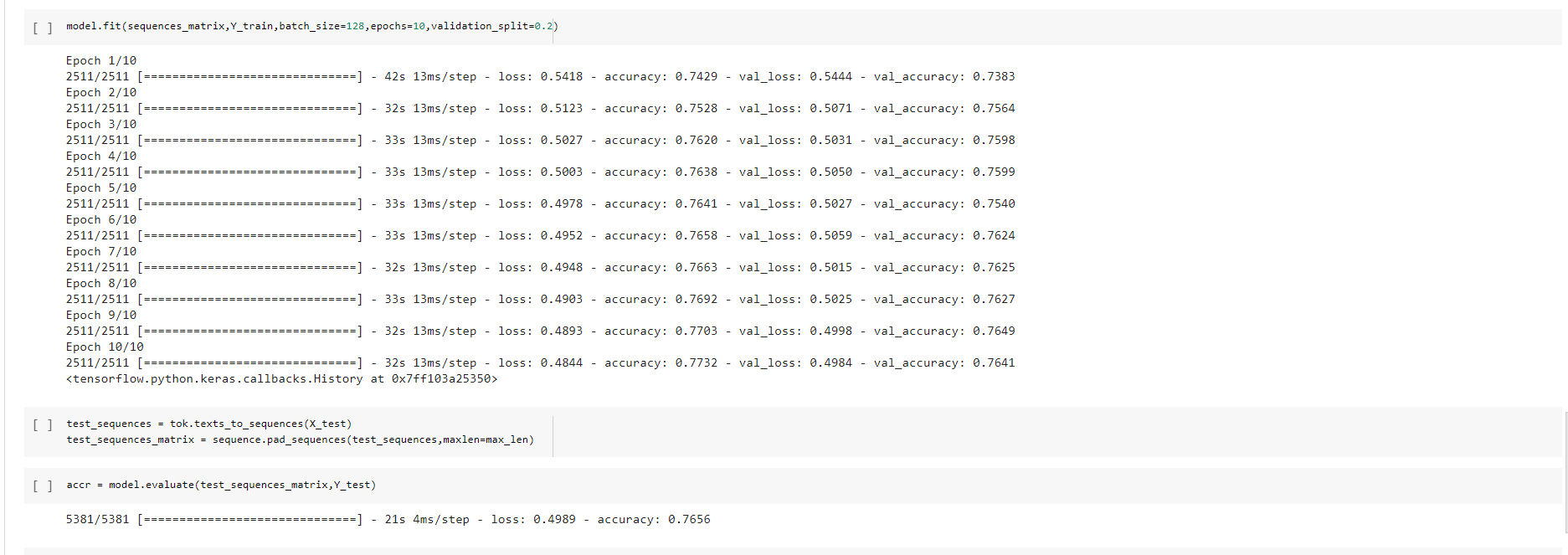
****

* **Split dataset in train and test and create RNN model**

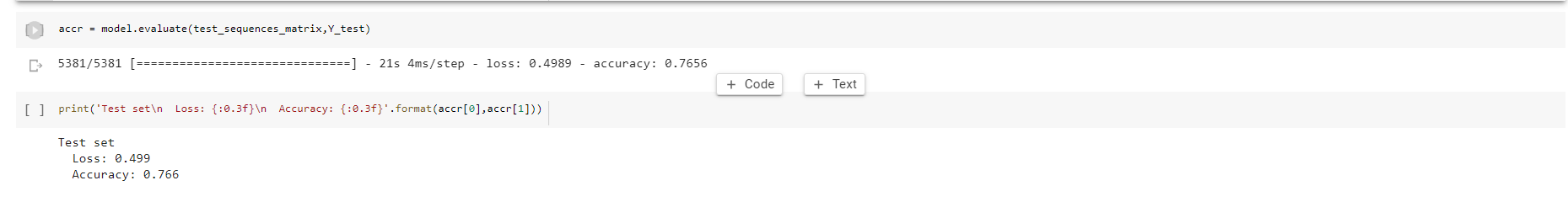
****

****

* **Train RNN model**

****

* **Test model**

****

1. **Conclusion**

Random Forest Classifier gave good accuracy (73%) compared to KNN and XGBoost. XGB classifier achieves the best performance metrics other than accuracy in rating prediction. Spoiler detection using RNN gives an accuracy of 76.6%.

1. **References**
2. Rachiraju, S.C. and Revanth, M., 2020, May. Feature Extraction and Classification of Movie Reviews using Advanced Machine Learning Models. In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 814-817). IEEE.
3. Ertugrul, A.M. and Karagoz, P., 2018, January. Movie genre classification from plot summaries using bidirectional lstm. In 2018 IEEE 12th International Conference on Semantic Computing (ICSC) (pp. 248-251). IEEE.
4. Asad, K.I., Ahmed, T. and Rahman, M.S., 2012, May. Movie popularity classification based on inherent movie attributes using C4. 5, PART and correlation coefficient. In 2012 International Conference on Informatics, Electronics & Vision (ICIEV) (pp. 747-752). IEEE.
5. Sajeevan, A. and Lakshmi, K.S., 2019, July. An enhanced approach for movie review analysis using deep learning techniques. In 2019 International Conference on Communication and Electronics Systems (ICCES) (pp. 1788-1794). IEEE.
6. Cizmeci, B. and Ögüdücü, Ş.G., 2018, September. Predicting IMDb ratings of pre-release movies with factorization machines using social media. In 2018 3rd International Conference on Computer Science and Engineering (UBMK) (pp. 173-178). IEEE.
7. Bristi, W.R., Zaman, Z. and Sultana, N., 2019, July. Predicting IMDb Rating of Movies by Machine Learning Techniques. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.
8. Gogineni, S. and Pimpalshende, A., 2020, June. Predicting IMDB Movie Rating Using Deep Learning. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) (pp. 1139-1144). IEEE.
9. Dhir, R. and Raj, A., 2018, December. Movie Success Prediction using Machine Learning Algorithms and their Comparison. In 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC) (pp. 385-390). IEEE.
10. Butler, M. and Robila, S., 2016, April. Interface for querying and data mining for the IMDb dataset. In 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT) (pp. 1-6). IEEE.
11. Ahmed, A., Batagelj, V., Fu, X., Hong, S.H., Merrick, D. and Mrvar, A., 2007, February. Visualisation and analysis of the Internet movie database. In 2007 6th International Asia-Pacific Symposium on Visualization (pp. 17-24). IEEE.