

HYBRID APPROACH FOR CLASSIFICATION OF MULTILABEL MEDICAL DATA USING CLASSIFICATION APPROACH (Machine Learning Based Approach)

Khushbu B.Patel

ABSTRACT

Medical data has been analyzed by medicos for better understanding of reports and information driven through various tests. There are chances to use machine learning to generate results based on certain criteria added with the help of human intelligence. The main aim of this research is to classify data for such further analysis. Exploratory data analysis (EDA) is performed for different set of data to focus on important features to get maximum insights from a data set. The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analyzing data. In this paper, the important factors are studies and the missing factors are predicted using K-means algorithm. The research proposes to use EDA along with machine learning techniques for classification of results. The result will generate categories that identifies words of medical terminology based on their relations.

INTRODUCTION

The main objective of this research to proposing an algorithm that is to classify data for further analysis medical . Exploratory data analysis (EDA) is performed for different set of data to focus on important features to get maximum insights from a data set. The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analyzing data. In this paper, the important factors are studies and the missing factors are predicted using K-means algorithm. The research proposes to use EDA along with machine learning techniques for classification of results. The result will generate categories that identifies words of medical terminology based on their relations. This research algorithm is works on important factors are studies and than the missing factors are we predicted using K-means algorithm. our research proposes to use EDA along with machine learning techniques and we are using here machine learning for classification of results generated by system. here result will generate categories that used to identifies words of medical terminology based on their relations

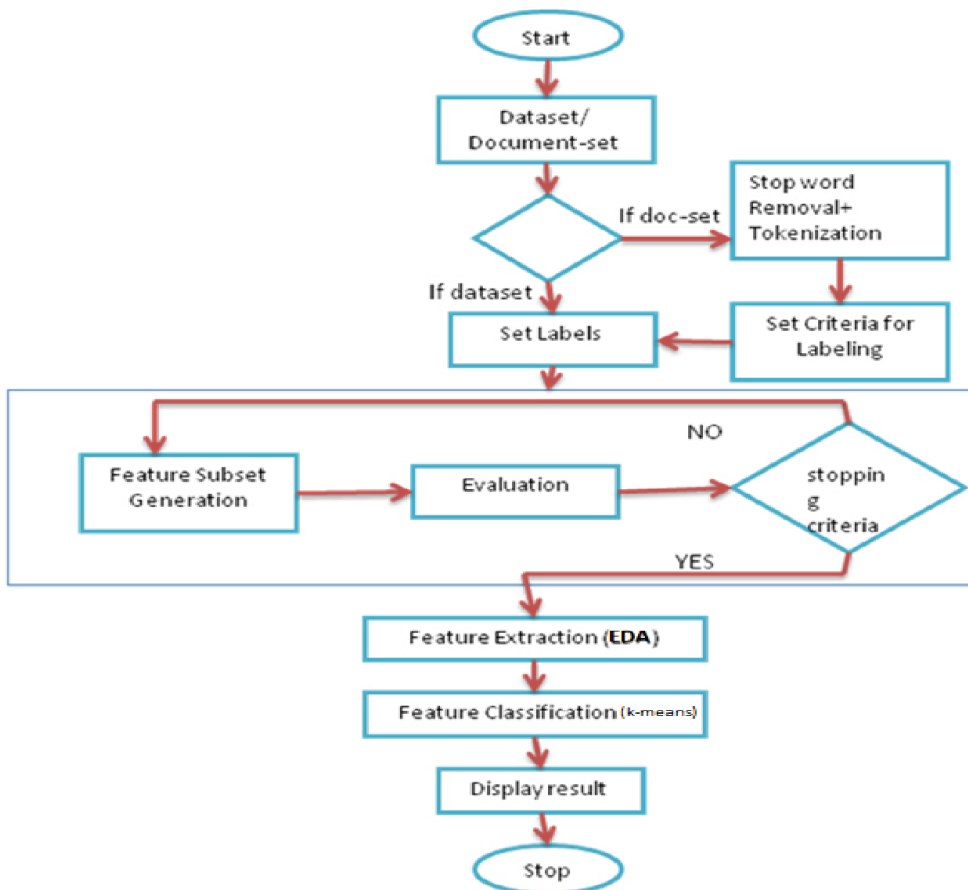
There are multiple types of documents which we use in our everyday work. The data is categorized in different types based on the values stored. In this research the categorization of Multi-label data between weakly labeled data and fully labelled data. The labels of training examples are incomplete, which commonly occurs in real applications in image classification are denoted as weakly labelled data. The classification accuracy need to be increased in order to check properly categorized data.

A motivation behind this research works we generating an algorithm for classification of multilabel medical data using Classification approach. and result will generate categories that identifies words of medical terminology based on their relations genatered according algorithm.

SYSTEM DESIGN AND IMPLEMENTATION

This diagram represent model of proposed algorithm. firstly dataset /documentation set is given if data set is ok than labeling is set accordingly based on criteria for labeling. After that feature subset is generated than we evaluate that subset. If it is based on criteria than feature extraction is done using EDA and if criteria is not on that it reprocess for feature generation and reevaluation .After successful feature extraction next we are going to done feature classification using k-means and after classification of medical labeling final result id display.

Architectural Overview



This diagram represent model of proposed algorithm. firstly dataset /documentation set is given if data set is ok than labeling is set accordingly based on criteria for labeling. After that feature subset is generated than we evaluate that subset. If it is based on criteria than feature extraction is done using EDA and if criteria is not on that it reprocess for feature generation and reevaluation .After successful feature extraction next we are going to done feature classification using k-means and after classification of medical labeling final result id display.

PROPOSED STEPS:

STEP1: Start the process

STEP 2: That We need to input one the document we can process

STEP 3: Data set medical term that has all the possible medical words and meaning.

STEP 4: That Document-set will be process first NLP(natural language process)

STEP 5: First process we be done for removing all the common english word.

STEP 6: At a same time the process of data labeling will be done over dataset

STEP 7: The token generated and the label of medical dataset will be put combine merge matrix.

STEP 8:Each feature match with label evolution up to all the label are process

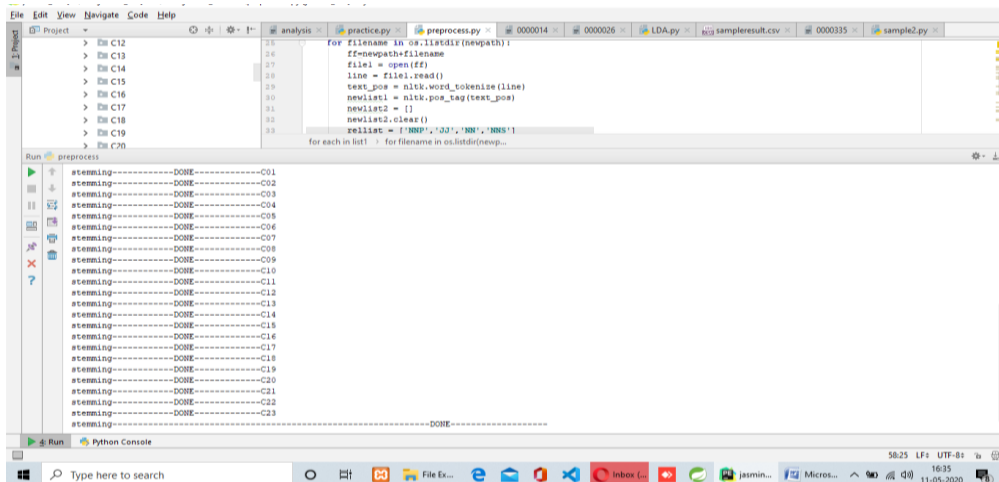
STEP 9: Feature Extraction process each for matching criteria for the document word.

STEP10: After this process will used in K-means algorithm

STEP11:Display Result

RESULT AND EVALUATION

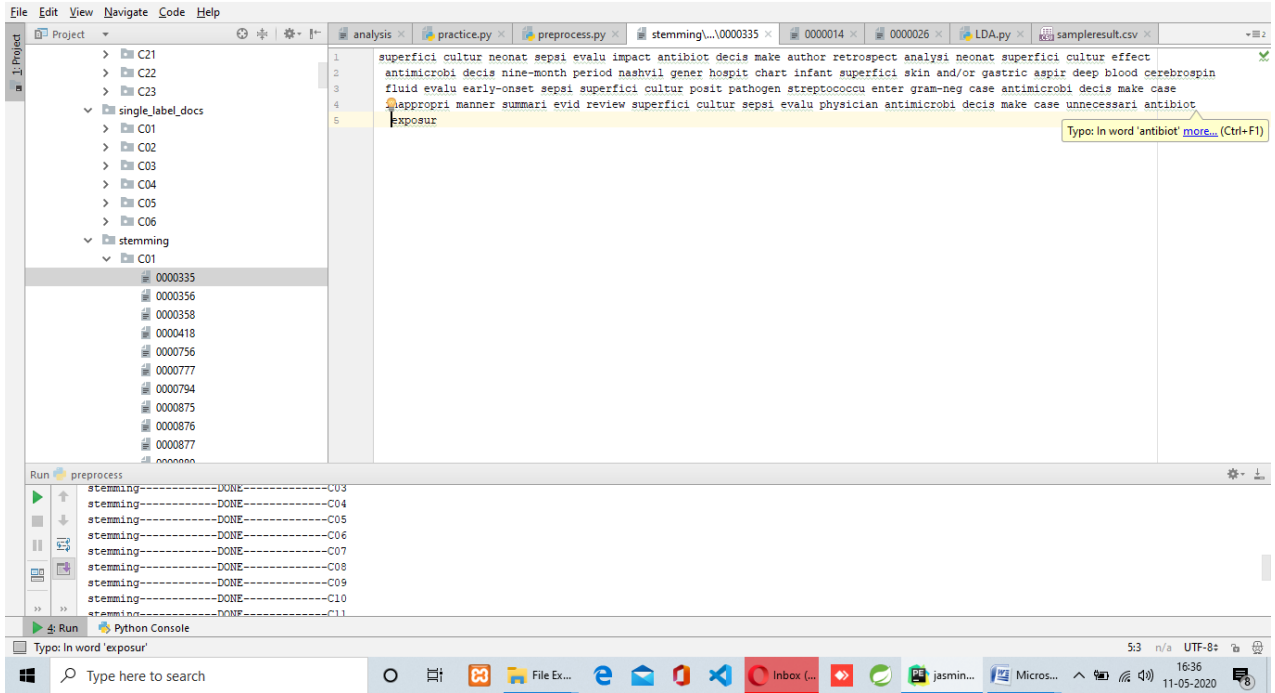
Implementation



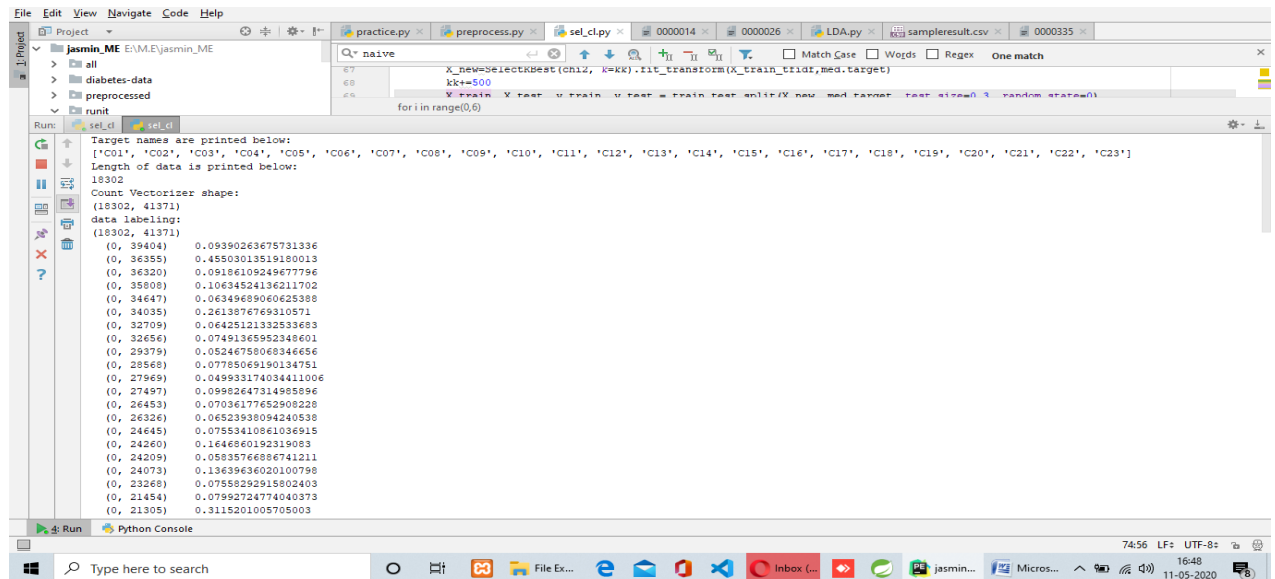
```
for filename in os.listdir(newpath):
    ff=newpath+filename
    file = open(ff)
    line = file.readline()
    text_pos = nltk.word_tokenize(line)
    newlist1 = nltk.pos_tag(text_pos)
    newlist2 = []
    newlist1.clear()
    rellist = ['NNP', 'JJ', 'NN', 'NNS']
    for each in list1:
        for filename in os.listdir(newp...
```

```
stemming-----DONE-----C01
stemming-----DONE-----C02
stemming-----DONE-----C03
stemming-----DONE-----C04
stemming-----DONE-----C05
stemming-----DONE-----C06
stemming-----DONE-----C07
stemming-----DONE-----C08
stemming-----DONE-----C09
stemming-----DONE-----C10
stemming-----DONE-----C11
stemming-----DONE-----C12
stemming-----DONE-----C13
stemming-----DONE-----C14
stemming-----DONE-----C15
stemming-----DONE-----C16
stemming-----DONE-----C17
stemming-----DONE-----C18
stemming-----DONE-----C19
stemming-----DONE-----C20
stemming-----DONE-----C21
stemming-----DONE-----C22
stemming-----DONE-----C23
stemming-----DONE-----
```

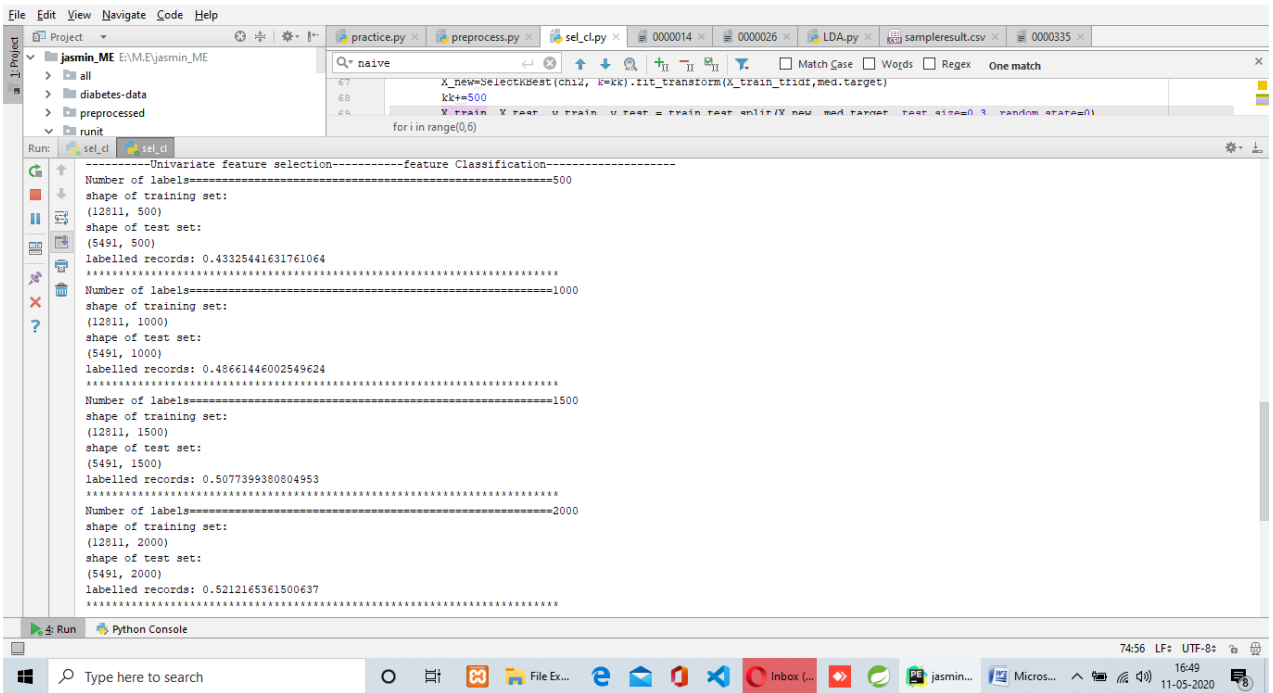
Reading docs



Document after stemming process



Results after labeling and training process



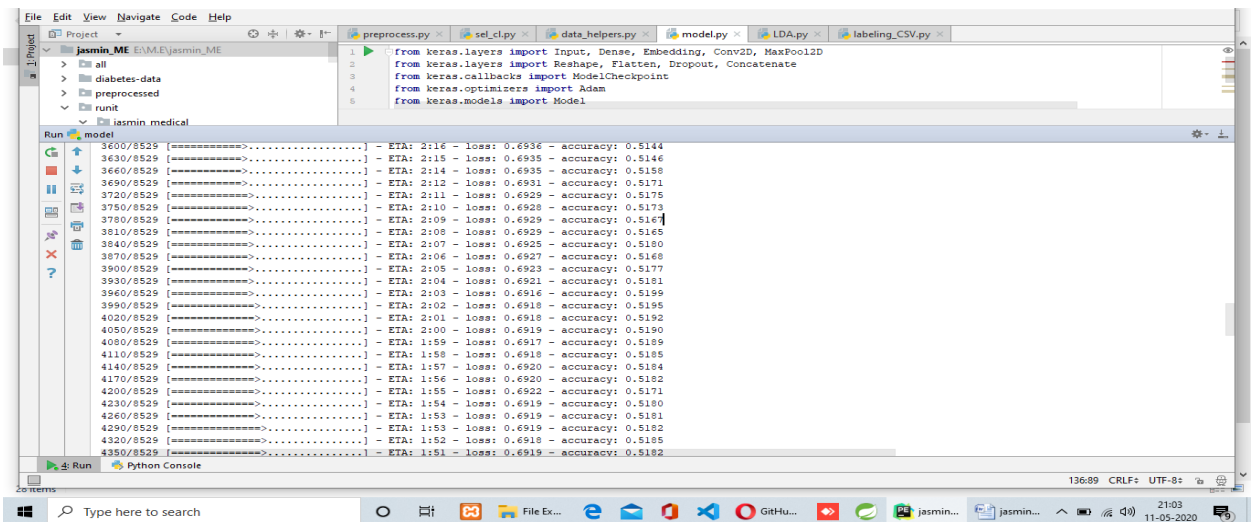
```

naive
X_new=SelectKBest(chi2, k=k).fit_transform(X_train_tfidf,med.target)
k=500
X_train X_test u_train u_test v_train v_test enlin(X_new, med.target, test_size=0.9, random_state=0)
for i in range(0,6)

-----Univariate feature selection-----feature Classification-----
Number of labels=====500
shape of training set:
(12811, 500)
shape of test set:
(5491, 500)
labelled records: 0.43325441631761064
*****
Number of labels=====1000
shape of training set:
(12811, 1000)
shape of test set:
(5491, 1000)
labelled records: 0.48661446002549624
*****
Number of labels=====1500
shape of training set:
(12811, 1500)
shape of test set:
(5491, 1500)
labelled records: 0.5077399380804953
*****
Number of labels=====2000
shape of training set:
(12811, 2000)
shape of test set:
(5491, 2000)
labelled records: 0.5212165361500637
*****

```

Training and testing set segregation



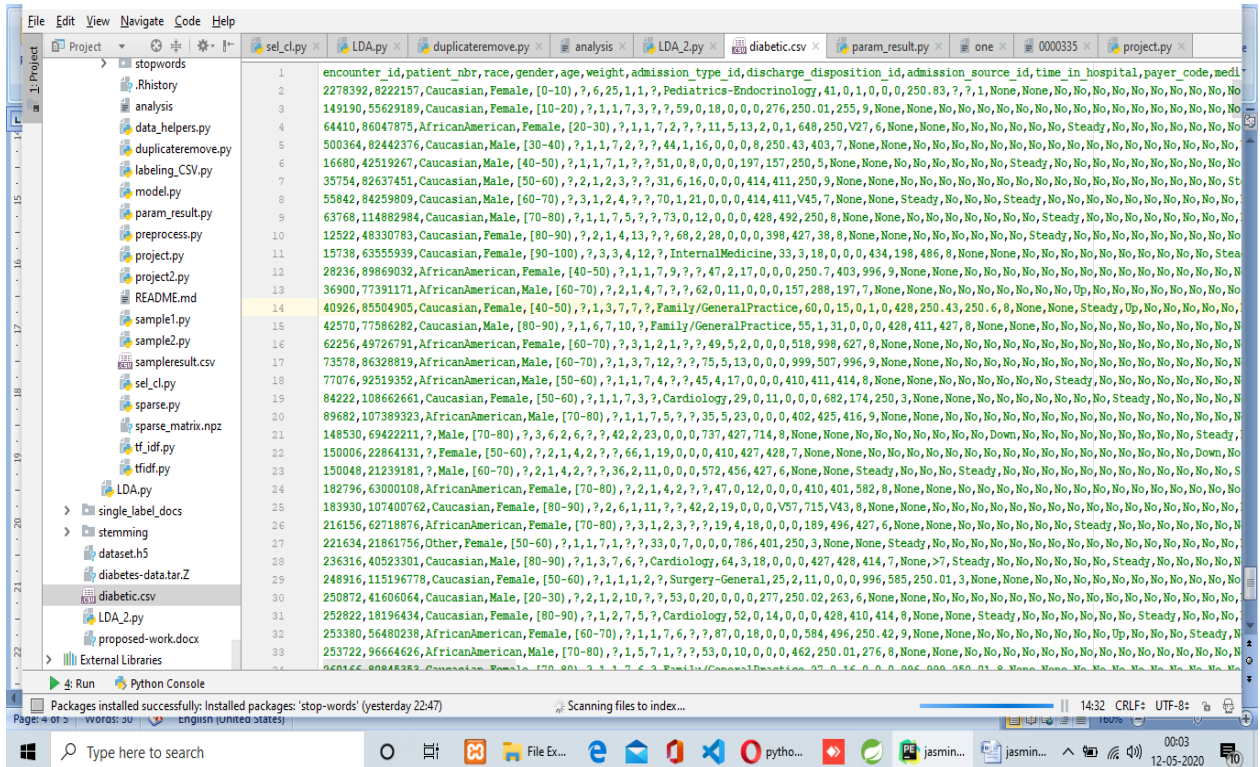
```

from keras.layers import Input, Dense, Embedding, Conv2D, MaxPool2D
from keras.layers import Reshape, Flatten, Dropout, Concatenate
from keras.callbacks import ModelCheckpoint
from keras.optimizers import Adam
from keras.models import Model

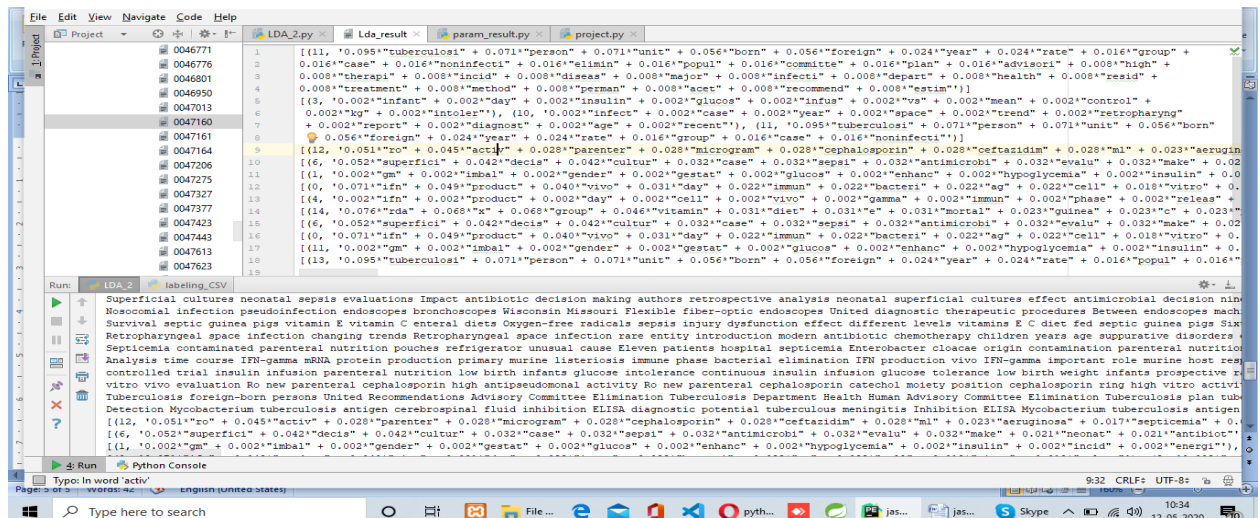
3600/8529 [=====] - ETA: 2:16 - loss: 0.6936 - accuracy: 0.5144
3630/8529 [=====] - ETA: 2:15 - loss: 0.6935 - accuracy: 0.5146
3660/8529 [=====] - ETA: 2:14 - loss: 0.6935 - accuracy: 0.5158
3690/8529 [=====] - ETA: 2:12 - loss: 0.6931 - accuracy: 0.5171
3720/8529 [=====] - ETA: 2:11 - loss: 0.6929 - accuracy: 0.5175
3750/8529 [=====] - ETA: 2:10 - loss: 0.6928 - accuracy: 0.5173
3780/8529 [=====] - ETA: 2:09 - loss: 0.6929 - accuracy: 0.5167
3810/8529 [=====] - ETA: 2:08 - loss: 0.6929 - accuracy: 0.5165
3840/8529 [=====] - ETA: 2:07 - loss: 0.6925 - accuracy: 0.5180
3870/8529 [=====] - ETA: 2:06 - loss: 0.6927 - accuracy: 0.5168
3900/8529 [=====] - ETA: 2:05 - loss: 0.6923 - accuracy: 0.5177
3930/8529 [=====] - ETA: 2:04 - loss: 0.6921 - accuracy: 0.5161
3960/8529 [=====] - ETA: 2:03 - loss: 0.6916 - accuracy: 0.5189
3990/8529 [=====] - ETA: 2:02 - loss: 0.6918 - accuracy: 0.5195
4020/8529 [=====] - ETA: 2:01 - loss: 0.6918 - accuracy: 0.5192
4050/8529 [=====] - ETA: 2:00 - loss: 0.6919 - accuracy: 0.5190
4080/8529 [=====] - ETA: 1:59 - loss: 0.6917 - accuracy: 0.5189
4110/8529 [=====] - ETA: 1:58 - loss: 0.6918 - accuracy: 0.5185
4140/8529 [=====] - ETA: 1:57 - loss: 0.6920 - accuracy: 0.5184
4170/8529 [=====] - ETA: 1:56 - loss: 0.6920 - accuracy: 0.5182
4200/8529 [=====] - ETA: 1:55 - loss: 0.6922 - accuracy: 0.5171
4230/8529 [=====] - ETA: 1:54 - loss: 0.6915 - accuracy: 0.5180
4260/8529 [=====] - ETA: 1:53 - loss: 0.6919 - accuracy: 0.5181
4290/8529 [=====] - ETA: 1:53 - loss: 0.6919 - accuracy: 0.5182
4320/8529 [=====] - ETA: 1:52 - loss: 0.6918 - accuracy: 0.5185
4350/8529 [=====] - ETA: 1:51 - loss: 0.6919 - accuracy: 0.5182

```

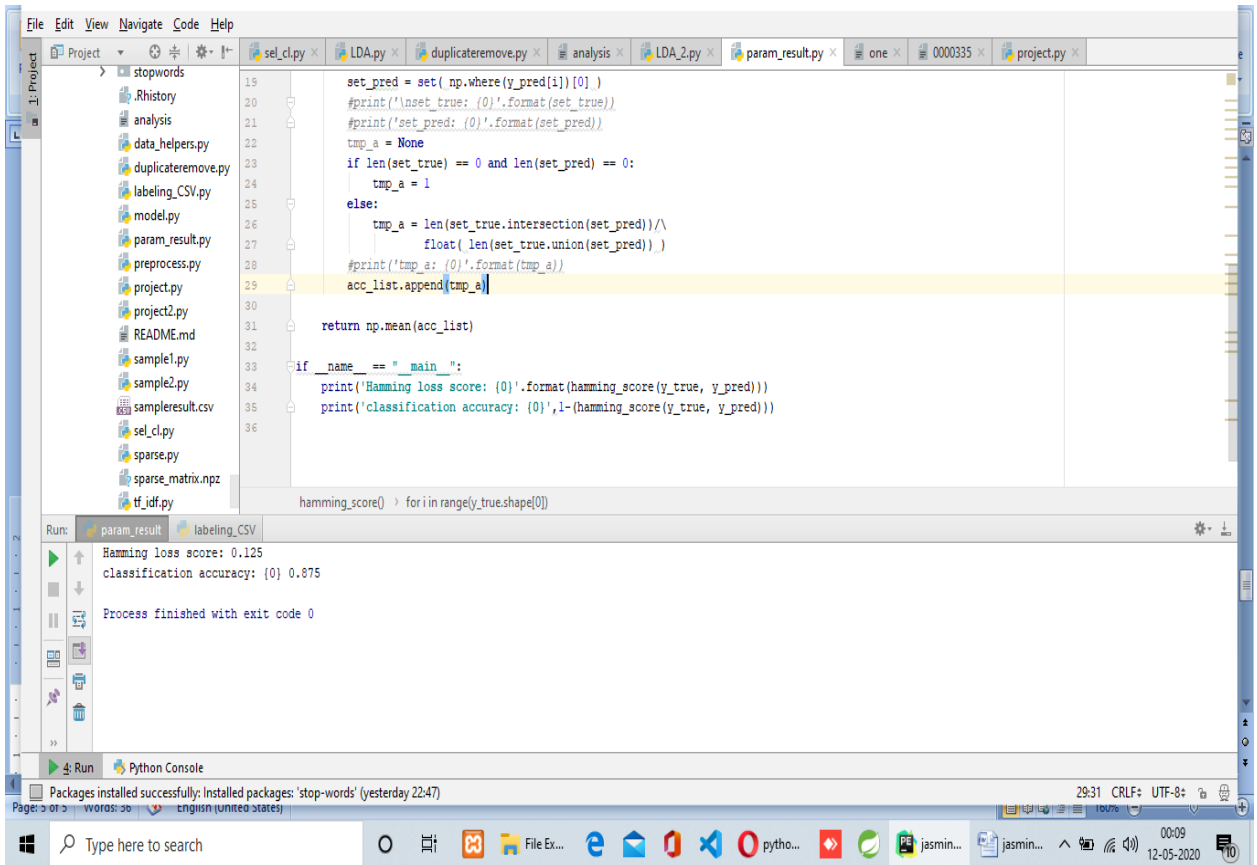
Training of csv data for labeling



Csv file for diabetic patient records



Results after LDA



```

19 set_pred = set( np.where(y_pred[i]) [0] )
20 #print('\nset_true: {0}'.format(set_true))
21 #print('set_pred: {0}'.format(set_pred))
22 tmp_a = None
23 if len(set_true) == 0 and len(set_pred) == 0:
24     tmp_a = 1
25 else:
26     tmp_a = len(set_true.intersection(set_pred)) /
27         float( len(set_true.union(set_pred)) )
28 #print('tmp_a: {0}'.format(tmp_a))
29 acc_list.append(tmp_a)
30
31 return np.mean(acc_list)
32
33 if __name__ == "__main__":
34     print('Hamming loss score: {0}'.format(hamming_score(y_true, y_pred)))
35     print('classification accuracy: {0},1-(hamming_score(y_true, y_pred))')
36
hamming_score() > for i in range(y_true.shape[0])
    
```

Run: param_result labeling_CSV

Hamming loss score: 0.125
classification accuracy: {0} 0.875

Process finished with exit code 0

Hamming loss and classification accuracy results

RESULTS

BR (Binary Relevance) Classifier					
Dataset	ML(multi label) Accuracy	Precision	Recall	F1-Measure	Avg Accuracy
Clinical	0.8325	0.8691	0.8868	0.8758	0.7586
medical	0.7511	0.819	0.8405	0.8278	0.6942
Plants	0.7553	0.8756	0.759	0.8098	0.7203
Virus	0.8583	0.9012	0.8857	0.8928	0.8114
Yeast	0.5079	0.7091	0.5896	0.6417	0.1593

Binary Relevance Classifier output

ML (multi-label) KNN Classifier					
Dataset	ML(multi label) Accuracy	Precision	Recall	F1-Measure	Avg Accuracy
clinical	0.7389	0.8632	0.7383	0.7952	0.6647
medical	0.7157	0.8422	0.7379	0.7837	0.6738
plants	0.7492	0.853	0.7442	0.794	0.7079
virus	0.7853	0.9058	0.7978	0.8367	0.7336
yeast	0.5186	0.7285	0.5889	0.6488	0.192

KNN Classifier Output

Proposed Approach				
Dataset	Precision	Recall	F1-measure	Accuracy
Clinical	0.885	0.7456	0.801	0.8053
medical	0.864	0.7843	0.7965	0.875
Plants	0.8263	0.7866	0.801	0.7248
Virus	0.9124	0.8279	0.8475	0.7698
Yeast	0.756	0.6024	0.7088	0.5477

Accuracy with Precision, Recall F1-measure

Proposed Approach		
Dataset	Hamming Loss	Accuracy
clinical	0.114	0.8053
medical	0.125	0.875
Plants	0.148	0.7248
Virus	0.085	0.7698
Yeast	0.035	0.5477

Accuracy Outcomes with Hamming Loss

Comparison of Base Paper kNN Classifier Precision Results With Base Paper		
Dataset	Base paper	Proposed Approach
clinical	0.8632	0.885
medical	0.8422	0.864
Plants	0.853	0.8263
Virus	0.9058	0.9124
Yeast	0.7285	0.756

Comparison of Base Paper kNN Classifier Precision Results With Base Paper

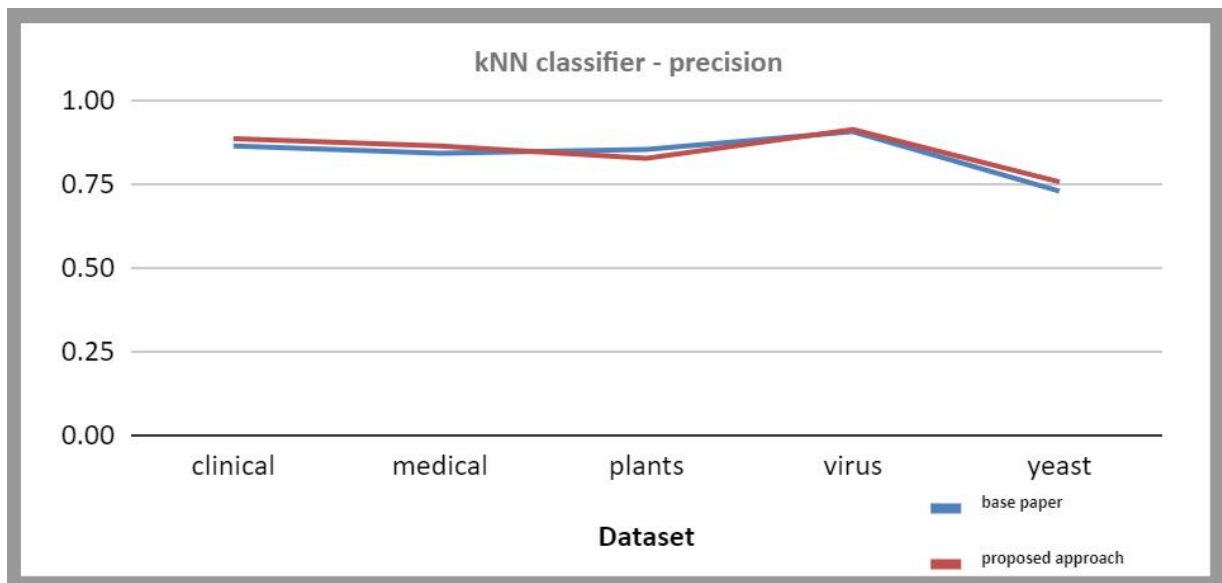


Fig6

Precision Graph

CONCLUSION

Here firstly we done classification of data for analysis. Then Exploratory data analysis (EDA) is used to performed for different set of data to focus on important features to get maximum insights from a data set. here The use of analytics in healthcare improves care by facilitating preventive care and EDA is a vital step while analyzing data. than the important factors are studies and the missing factors are predicted by using K-means algorithm. here, our research proposes to use EDA along with machine learning techniques for classification of results. The result will used to generate categories that contains identifies words of medical terminology based on their relations.

REFERENCES

1. <https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning>
2. [https://www.google.com/search?q=lda+machine+learning&sxsrf=APq-WBvoZKrg1ykODIJcJJRuTbOEXcG8g%3A1643876186790&ei=Wo_7YYfcL4ejoATvy4ngDQ&oeq=lda&gs_lcp=Cgdnd3Mtd2l6EAEYADIECCMQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIFCAAQgAQyBAGAEEMyBAGAEEMyBQgAEIAEOgcIABBHELADoggIABCABBCwAzoHCCMQ6gIQJzoNCC4QxwEQ0QM6gIQJzoICAAQgAQQsQM6EAgUELEDEIMBEMcBE N EDEEM6BQgAELEDOg4ILhCABBCxAxDHARCjAjoLCC4QgAQQxwEQ0QM6DQguELEDEM c BEKMCEEM6BQguEIAESgUIPBIBMUoECEEYAEoECEYYAFCilQNY56sDYNTPA2gCcAJ4BI A BgQKIAbgLkgEFMC43LjGYAQCgAQGwAQRlAQnAAQE&sclient=gws-wiz](https://www.google.com/search?q=lda+machine+learning&sxsrf=APq-WBvoZKrg1ykODIJcJJRuTbOEXcG8g%3A1643876186790&ei=Wo_7YYfcL4ejoATvy4ngDQ&oeq=lda&gs_lcp=Cgdnd3Mtd2l6EAEYADIECCMQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIFCAAQgAQyBAGAEEMyBAGAEEMyBQgAEIAEOgcIABBHELADoggIABCABBCwAzoHCCMQ6gIQJzoNCC4QxwEQ0QM6gIQJzoICAAQgAQQsQM6EAgUELEDEIMBEMcBE N EDEEM6BQgAELEDOg4ILhCABBCxAxDHARCjAjoLCC4QgAQQxwEQ0QM6DQguELEDEM c BEKMCEEM6BQguEIAESgUIPBIBMUoECEEYAEoECEYYAFCilQNY56sDYNTPA2gCcAJ4BI A BgQKIAbgLkgEFMC43LjGYAQCgAQGwAQRlAQnAAQE&sclient=gws-wiz)
3. <https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning>
4. [Guillaumin *et al.*, 2010] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 902–909, San Francisco, CA, 2010.
5. Kartik Dhiwar, PG Scholar, Department of Computer Science and Engineering, SSGI, SSTC, Bhilai (CG), India. Abhishek Kumar Dewangan Professor, Department of Computer Science and Engineering, SSGI, SSTC, Bhilai (CG), India.
6. Isabelle Augenstein, Diana Maynard and Fabio Ciravegna Department of Computer Science, The University of Sheffield, UK {i.augenstein,d.maynard,f.ciravegna}@dcs.shef.ac.uk
7. To cite this article: Yang Luo *et al* 2021 *J. Phys.: Conf. Ser.* 2010 012038
8. Jingcheng Du,^{1,2} Qingyu Chen,¹ Yifan Peng,¹ Yang Xiang,² Cui Tao,² and Zhiyong Lu¹
¹National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland, USA, and ²The University of Texas School of Biomedical Informatics, Houston, Texas, USA
Corresponding Author: Zhiyong Lu, PhD, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA (zhiyong.lu@nih.gov)
Received 10 December 2018; Revised 21 April 2019; Editorial Decision 6 May 2019; Accepted 8

May 2019

9. Alina Nesen Computer Science Purdue University
West Lafayette, IN, USA anesen@purdue.edu

10. Bharat Bhargava Computer Science Purdue University
West Lafayette, IN, USA bbshail@purdue.edu

11. Stephan SPAT^{a,1}, Bruno CADONNA^b, Ivo RAKOVAC^a, Christian GÜTL^c, Hubert LEITNER^d, Günther STARK^d, Thomas R. PIEBER^{a,e}, Peter BECK^a *JOANNEUM RESEARCH Forschungsges.m.b.H., Institute for Biomedicine and Health Sciences, Graz, Austria*

^b*Free University of Bozen – Bolzano, Faculty of Computer Science, Bolzano, Italy* ^c*Graz University of Technology, Institute for Information Systems and Computer Media, Graz, Austria*

^d*Steiermärkische Krankenanstaltenges.m.b.H., Graz, Austria*

^e*Medical University of Graz, Department of Internal Medicine, Division of Endocrinology and Nuclear Medicine, Graz, Austria*

12. Kazuteru Miyazaki
National Institution for Academic Degrees and Quality Enhancement of Higher Education 1-29-1,
Gakuennichimachi, Kodaira,

13. Dongfang Ma
School of Micro-Nanoelectronics, Zhejiang University, Hangzhou 310027, China
Xiaoqian Yan
Tongde Hospital of Zhejiang Province
, Hangzhou 310012, China

14. Ximin Li Second Clinical Medical
College Zhejiang Chinese Medical University, Hangzhou 310053, China
Shenghong Mou School of Micro-Nanoelectronics, Zhejiang University, Hangzhou
310027, China
Ying Lu
Tongde Hospital of Zhejiang Province
, Hangzhou 310012, China

15. Zhiyuan CHENG School of Micro-Nanoelectronics, Zhejiang University, Hangzhou
310027, China

16. Ruijian Yan Department of Orthopedic, Surgery Second Affiliated Hospital, Zhejiang
University School of Medicine, Hangzhou, China

17. <https://www.javatpoint.com/machine-learning>

- 18 <https://vitalflux.com/machine-learning-list-of-35-free-online-books/>
19. [https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b#:~:text=NLP%20is%20a%20field%20in,and%20potentially%20generate%20human%20language.&text=Information%20Retrieval\(Google%20finds%20relevant,Gmail%20structures%20events%20from%20emails\).](https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-d44498845d5b#:~:text=NLP%20is%20a%20field%20in,and%20potentially%20generate%20human%20language.&text=Information%20Retrieval(Google%20finds%20relevant,Gmail%20structures%20events%20from%20emails).)
20. <https://www.javatpoint.com/python-tutorial>