

# 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 terminolog 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 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.

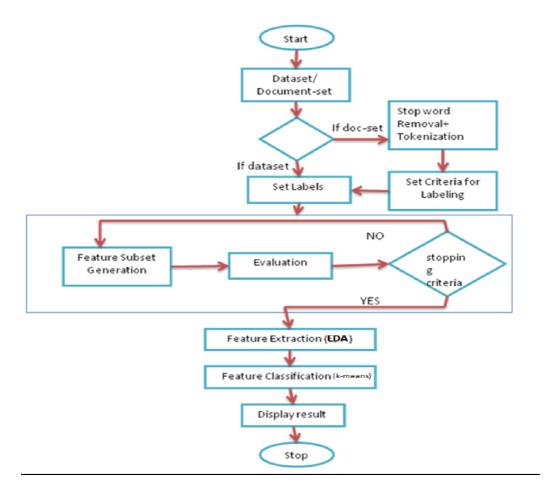
A motivation behind this research works we generating an algorithm for classification of multilabel medical data using Classification approach.and result will generate categories that identifies words of medical terminology based on their relations genatered according algorithm.



# 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 reevalution .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 reevalution. 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.

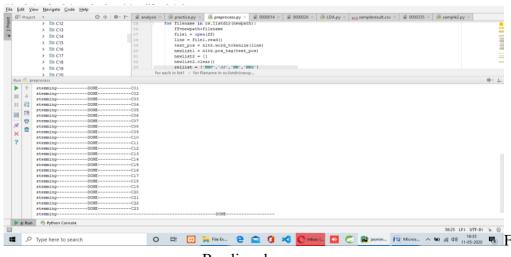
#### **PROPOSED STEPS:**

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 EVALUTION**

#### Implementation



Reading docs

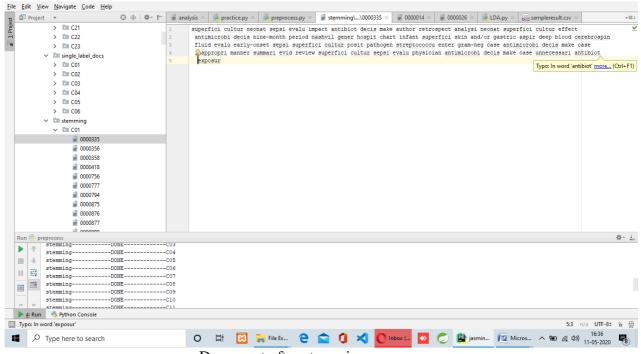


<u>E</u> dit <u>V</u> iew <u>N</u> avigate <u>C</u> ode <u>H</u> el	p									
🗗 Project 🔻	🕀 🕸 🕸 🗄 🏥 ana	alysis × 👘 practice.py ×	🐞 preprocess.py 🗵	i 0000014 ×	🗑 0000026 ×	ᡖ LDA.py 🗵	sampleresult.csv	🗑 🗑 🗑 🗑 🗑	ᡖ sample2.py ×	
> 🖿 C12	25		os.listdir(newpat)	h):						
> 🖿 C13	26	ff=newpath+:								
> 🖿 C14	27	filel = oper								
> C15	28	line = file								
> 🖿 C16	29		ltk.word_tokenize							
> D C17	30	newlist1 = 1 newlist2 =	ltk.pos_tag(text_)	pos)						
> = C18	32	newlist2.cle								
> C19	33		'NNP', 'JJ', 'NN', 'NN	NS'1						
		for each in list1 → for fil								
> C20 Run 🛑 preprocess										÷.
										- 94C
	ding package punkt to Users\sonal\AppData\Roam:	ing)nltk data								
	ge punkt is already up-to									
	ding package averaged pe:									
	Users\sonal\AppData\Roam									
🖸 🖪 [nltk_data] Unzipp	ping taggers\averaged pe:	rceptron tagger.zip.								
DONE	C01									
ONE     DONE     ONE     ONE										
DONE										
DONL										
?DONE										
DONE										
DONE										
DONE										
DONE										
DONE										
DONE	C12									
DONE	C13									
DONE	C14									
DONE										
DONE										
DONE										
DONE										
4: Run 😽 Python Console										
<u>, 1</u> , 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,									58:25 LF\$ UTF-	8≑ °a
P Type here to search		o 🛱 🔀 🕻	File Ex 🧲 🛾	🗅 🛈 🗙	Inbox	- 💿 🥏	😰 jasmi 🔮	resear ^ 🖸 🕯	■ <i>信</i> ( 句)) 16:31	20 🖣
		nrent	ocessing	the do	cumer	nte				
		prepr	occosing	, inc uc	cumer	113				

Project	-	⊕ ÷ ₩- ⊮	= 🔡 ana	alysis × 🛛 į	👵 practice.py	🛛 🛛 👘 prep	ocess.py $ imes$	🗧 stemming	\\0000335	× 🗑 stopwo	ords\\00003	335 × 🗐 🗐 0000	014 × 🛛 🗑 0000	026 × 🛛 🎁	LDA.py ×	* 1
	> 🖿 C21		1	Superfi	cial cultur	res neonata	l sepsis e	valuations :	Impact ant:	ibiotic deci	ision makin	ng authors re	trospective a	nalysis n	eonatal	
	> 🖿 C22		2	superfi	cial cultur	res effect	antimicrob	ial decision	n nine-mont	th period Na	ashville G	eneral Hospit	al charts inf	ants supe:	rficial	
	> C23		3	skin an	d/or gastr:	ic aspirate	deep bloo	d cerebrosp:	nal fluid	evaluation	early-ons	et sepsis Sup	erficial cult	ures posi	tive	
	single_label_docs		4										propriate mann			
	> C01		5	review	superficia:	l cultures	sepsis eva	luation phys	sician ant:	imicrobial d	decision m	aking case u	nnecessary ant	ibiotic e:	xposure	
	> = C02															
	> CO2															
	> 🛅 C04															
	> 🖿 C05															
	> 🖿 C06															
	> istemming															
,	stopwords															
	✓ I C01															
	0000335															
	0000356															
	0000358															
	iii 0000338 iii 0000418															
	0000756															
	iii 0000777															
	iii 0000794															
	iii 0000876															
	4 000077		_													
Run 🤛 pre	eprocess															
	stemmingI															
+	stemmingI															
_	stemmingI															
- <b>5</b>	stemmingI															
	stemmingI	DONE	C08													
<u> </u>	stemmingI	DONE	C09													
	stemmingI	DONE	C10													
>> >>	stemmingI	DONF														
	🐴 Python Console															
▶ <u>4</u> : Run															5:1 n/a UTF-8	3¢ %
▶ <u>4</u> : Run																

Document after stopping process





Document after stemming process

p P	roject	Ŧ	⊕ ÷ ∳ • I+	👘 practice	.py 🛛 🛛 🎼	preprocess	.py 🛛 👔	sel_cl.py	× 🗒	0000014 >	i 🗐 0000	026 ×	👍 LDA.py	× 📰 san	npleresult.cs	v × 🛛 🗑 0000	335 ×			
	jasmi	n_ME E:\M.E\jasmin	ME	Q. naive			< ⊗	<b>+</b> J		+n -n	2. <b>T</b>		atch Case	Words	Reger	One match				
>	a a	I		67		new=Selle					orm(X tra:				- neges	onematch				
>	d 🖿	iabetes-data		68		k+=500			,					,						
>	DI D	reprocessed		69	x	train X	test t	train	v test	= train	test enl	it/X net	a med ta	rdet tes	t sizes0	3 random st	(Omenter			
~	Di n	unit			for i in rang	ge(0,6)														
tun:		el_cl 🔜 sel_cl																		-
•	+	Target names are	printed below:																	
_		['C01', 'C02', '	C03', 'C04', 'C05', '	C06', 'C07	', 'COS',	'C09', '	C10', 'C	:11', 'CI	12', '0	:13', 'C1	4', 'C15'	, 'C16'	, 'C17',	'C18', 'C	19', 'C20	', 'C21', 'C	222', 'C23	3']		
	+ ;	Length of data i	s printed below:																	
11	9 <b>-</b> \$	18302																		
- 0		Count Vectorizer	shape:																	
9		(18302, 41371)																		
		data labeling:																		
e   .		(18302, 41371)																		
×			0.09390263675731336																	
			0.45503013519180013																	
?			0.09186109249677796																	
			0.10634524136211702																	
			0.06349689060625388																	
			0.2613876769310571																	
			0.06425121332533683																	
			0.07491365952348601 0.05246758068346656																	
			0.07785069190134751																	
			0.049933174034411006																	
			0.09982647314985896																	
			0.07036177652908228																	
			0.06523938094240538																	
			0.07553410861036915																	
			0.1646860192319083																	
			0.05835766886741211																	
			0.13639636020100798																	
		(0, 23268)	0.07558292915802403																	
		(0, 21454)	0.07992724774040373																	
		(0, 21305)	0.3115201005705003																	
4:	Run	😽 Python Console																		
-	_																	74:56 L	F¢ UTF-8¢	2
	0.	Type here to sear		C			Eile C	-	-	-	10	Inhau (			i	Micros	A 000	( 1))	16:48	
	$\mathcal{P}$	type nere to sear	cn	C			File E	···· C			$\sim$ $\cup$	mbox (.			jasrnin	Wilcros		((7. 47))	11-05-2020	Ę

Results after labeling and training process

I



. P	Project 🔻 😳 崇 🕸	🗜 🎼 practice.py × 🞼 preproces.py × 📫 sel_cl.py × 🚆 0000014 × 🔮 0000026 × 👘 LDA.py × 🔛 sampleresult.csv × 😫 0000335 ×	
	jasmin_ME E:\M.E\jasmin_ME	Q* naive $\leftrightarrow$ 😒 🛧 💺 👷 $+_{\Pi}{\Pi} \boxtimes_{\Pi}$ 🏋 🗌 Match Case 🗌 Words 🗌 Regex One match	
>	Car all     diabetes-data     To preprocessed     To runit	67     X_new=SelectKBest(ch12, k=kK).fit_transform(X_train_thidr,med.target)       68     kk+=500       64     X_train_X_treat_u_train_u_text = train_text snlt/(X_new_med_target_text size=0 %_random state=0)       67     X_train_X_text_u_train_u_text = train_text snlt/(X_new_med_target_text size=0 %_random state=0)	
Run:		(or mininglos)	¢.
	<ul> <li>shape of training set: (12811, 500)</li> <li>shape of test set: (5491, 500)</li> <li>labelled records: 0.4332544163174</li> <li>Number of labels</li> <li>shape of training set: (12811, 1000)</li> <li>shape of test set: (5491, 1000)</li> <li>labelled records: 0.4866144600254</li> </ul>	1000	
	Number of labels= shape of training set: (1281, 2000) shape of test set: (5491, 2000) labelled records: 0.5212165361500	£37	
	<pre>shape of test set: (5491, 1500) labelled records: 0.5077399380804 Number of labels</pre>	2000	
4	<pre>shape of test set: (5491, 1500) labelled records: 0.5077399380804 Number of labels shape of training set: (12811, 2000) shape of test set: (5491, 2000) labelled records: 0.5212165361500</pre>	637	36 LF= UTF-8= %

Training and testing set segregation

~ 1	1.14	asmin_ME_E:\M.E\jasmin_ME			anne denne Terre Perce Tehedding Courth Merberton	
					Layers import Input, Dense, Embedding, Conv2D, MaxPool2D Layers import Reshape, Flatten, Dropout, Concatenate	
	_				callbacks import ModelCheckpoint	
		diabetes-data	3		optimizers import Adam	
		preprocessed			nodels import Model	
	/ D	💷 runit	•	I I OM KE	adels import nodel	
	~	iasmin medical				
Run	<b>1</b>	model				÷-
¢	+	3600/8529 [=====>	] =	ETA: 2:1	loss: 0.6936 - accuracy: 0.5144	
	12	3630/8529 [====>		ETA: 2:1	loss: 0.6935 - accuracy: 0.5146	
	+	3660/8529 [>				
n.	<u>5</u> =3	3690/8529 [>				
÷.,	2.5	3720/8529 [=====>		ETA: 2:1	loss: 0.6929 - accuracy: 0.5175	
		3750/8529 [>				
	<b>D</b>	3780/8529 [>	1 -	ETA: 2:0	loss: 0.6929 - accuracy: 0.5167	
S.	.0.	3810/8529 [=====>				
	龠	3840/8529 [=====>				
×		3870/8529 [>		ETA: 2:0	loss: 0.6927 - accuracy: 0.5168	
2		3900/8529 [>	1 -	ETA: 2:0	loss: 0.6923 - accuracy: 0.5177	
		3930/8529 [====>				
		3960/8529 [=====>		ETA: 2:0	loss: 0.6916 - accuracy: 0.5199	
		3990/8529 [>		ETA: 2:0	loss: 0.6918 - accuracy: 0.5195	
		4020/8529 [>	1 -	ETA: 2:0	loss: 0.6918 - accuracy: 0.5192	
		4050/8529 [>				
		4080/8529 [========>	] =	ETA: 1:5	loss: 0.6917 - accuracy: 0.5189	
		4110/8529 [>		ETA: 1:5	loss: 0.6918 - accuracy: 0.5185	
		4140/8529 [>	1 -	ETA: 1:5	loss: 0.6920 - accuracy: 0.5184	
		4170/8529 [====>		ETA: 1:5	loss: 0.6920 - accuracy: 0.5182	
		4200/8529 [=====>	] =	ETA: 1:5	loss: 0.6922 - accuracy: 0.5171	
		4230/8529 [>	1 -	ETA: 1:5	loss: 0.6919 - accuracy: 0.5180	
		4260/8529 [>	1 -	ETA: 1:5	loss: 0.6919 - accuracy: 0.5181	
		4290/8529 [=====>				
		4320/8529 [===========>	] =	ETA: 1:5	loss: 0.6918 - accuracy: 0.5185	
_		4350/8529 [>		ETA: 1:5	loss: 0.6919 - accuracy: 0.5182	
▶.	: Ru	un 📑 Python Console				
-						136:89 CRLF¢ UTF-8¢ %
ms	_					8

Training of csv data for labeling



🗊 Project 🔻 🙂 🖶 🕸 🗠	🐌 sel_cl.py 🔾	< 👘 LDA.py 🛛	👼 duplicateremove.py 🛛	🗑 analysis 🛛	🐌 LDA_2.py 🛛	📖 diabetic.csv 🗵	🕌 param_result.py 🛛	🗑 one 🛛	🗑 0000335 ×	🐌 project.py 🛛
> stopwords	1	encounter id,	patient nbr, race, gende	er,age,weight,	admission type	id, discharge di	sposition id, admissi	on source	id, time in ho	spital, payer code
2.Rhistory	2		57, Caucasian, Female, [0							
i analysis	3	149190, 556291	39, Caucasian, Female, [1	10-20),?,1,1,7	7,3,?,?,59,0,18	,0,0,0,276,250.0	1,255,9,None,None,No	, No, No, No	, No, No, No, No, N	io, No, No, No, No, No
👼 data_helpers.py	4	64410,8604787	5,AfricanAmerican,Fema	ale,[20-30),?,	1,1,7,2,?,?,11	,5,13,2,0,1,648,	250,V27,6,None,None,	No, No, No,	No, No, No, Stead	ty, No, No, No, No, No
duplicateremove.py	5	500364,824423	76,Caucasian,Male,[30-	40),?,1,1,7,2	2,?,?,44,1,16,0	,0,0,8,250.43,40	3,7,None,None,No,No,	No, No, No,	No, No, No, No, No	, No, No, No, No, No, No, I
labeling_CSV.py	6		7,Caucasian,Male,[40-5							
👘 model.py	7		L,Caucasian,Male,[50-6							
param_result.py	8		9,Caucasian,Male,[60-7							
	9		34, Caucasian, Male, [70-							
i preprocess.py	10		3, Caucasian, Female, [80							
👘 project.py	11		9, Caucasian, Female, [90							
🐌 project2.py	12		2,AfricanAmerican,Fema							
EREADME.md	13		L,AfricanAmerican,Male							
📥 sample1.py	14		5, Caucasian, Female, [40							
asample2.py	16		2,Caucasian,Male,[80-9 L,AfricanAmerican,Fema							
ampleresult.csv	17		9,AfricanAmerican,Male							
sel_cl.py	18		2,AfricanAmerican,Male							
	19		51, Caucasian, Female, [5							
i sparse.py	20		23, AfricanAmerican, Mal							
in sparse_matrix.npz	21		L1,?,Male,[70-80),?,3,							
👘 tf_idf.py	22		31,?,Female,[50-60),?,							
🐌 tfidf.py	23		31,?,Male,[60-70),?,2,							
LDA.py	24		08,AfricanAmerican,Fem							
> ingle label docs	25		762, Caucasian, Female, [							
> is stemming	26	216156,627188	76, AfricanAmerican, Fem	nale,[70-80),1	?,3,1,2,3,?,?,1	9,4,18,0,0,0,189	,496,427,6,None,None	, No, No, No	, No, No, No, Stea	dy, No, No, No, No, No, No
	27	221634,218617	56,Other,Female,[50-60	),?,1,1,7,1,3	?,?,33,0,7,0,0,	0,786,401,250,3,1	None, None, Steady, No,	No, No, No,	No, No, No, No, No	, No, No, No, No, No, No, I
👘 dataset.h5	28	236316,405233	01,Caucasian,Male,[80-	-90),?,1,3,7,6	6,?,Cardiology,	64,3,18,0,0,0,42	7,428,414,7,None,>7,	Steady, No	, No, No, No, No, N	io, Steady, No, No, No
🚽 diabetes-data.tar.Z	29	248916,115196	778, Caucasian, Female, [	[50-60),?,1,1,	1,2,?,Surgery-	General,25,2,11,	0,0,0,996,585,250.01	,3,None,N	one, No, No, No, N	io, No, No, No, No, No
diabetic.csv	30	250872,416060	64,Caucasian,Male,[20-	-30),?,2,1,2,1	10,?,?,53,0,20,	0,0,0,277,250.02	,263,6,None,None,No,	No,No,No,	No, No, No, No, No	, No, No, No, No, No, No, I
👍 LDA_2.py	31	252822,181964	34, Caucasian, Female, [8	30-90),?,1,2,7	7,5,?,Cardiolog	y,52,0,14,0,0,0,	428,410,414,8,None,N	Ione, Stead	y,No,No,No,No,	No, No, Steady, No, I
proposed-work.docx	32	253380,564802	38,AfricanAmerican,Fem	nale,[60-70),3	?,1,1,7,6,?,?,8	7,0,18,0,0,0,584	,496,250.42,9,None,N	None, No, No	, No, No, No, No, No, N	lo,Up,No,No,No,Sto
External Libraries	33	253722,966646	26,AfricanAmerican,Mal	Le,[70-80),?,1	1,5,7,1,?,?,53,	0,10,0,0,0,462,2	50.01,276,8,None,Nor	ne, No, No, N	o, No, No, No, No,	No, No, No, No, No, No
🕨 4: Run 🛛 🤞 Python Console	24	020122 0004E3	2 Caucagian Pomalo 12		1 6 3 Pamilu/Ca	nowal Departing 07	0 16 0 0 0 006 000	980 01 0	None Nene No. V	Io No No No No No
Packages installed successfully: Installe	d nackagori 'rto	n-words' (vesterda	, 22, 47)	: Scanning file	s to index				11.142	2 CRLF¢ UTF-8¢
or 5 words: 30 💓 English (Unit		p-words (yesterda)	22,41)	* Scanning me	es to muex					
		0	🛱 🔀 🐂 File	Ex 🤁 🛛	🗅 🛈 🗙	O pytho	🔊 💋 😰 jasmi	in 🔮 ji	asmin 🔨 🛍	00:03 ((c, \$)) 12-05-2020
										12 03-2020
			Csv file for	r diabe	tic patie	ent recor	ds			
					1					

■ 0046771     1 [(11, *0     ■ 0046776 2 0.016***     ■ 0046801 3 0.008***	Lda_result × 🎼 param_result.py × 🚺 project.py × 🛛 095*"tuberculosi" + 0.071*"person" + 0.071*"unit" + 0.056*"born" + 0.056*"foreign" + 0.024*"year" + 0.024*"rate" + 0.016*"group" + 🛛 💥
0046776 2 0.016*** 0046801 3 0.008***	
■ 0049950 5 [13, *0, 002=* ■ 0047161 7 + 0,002 ■ 0047161 7 + 0,002 ■ 0047161 9 [12, *0, 003 ■ 0047161 9 [12, *0, 003 ■ 0047161 1 [12, *0, 004 ■ 0047265 11 [11, *0, *0, 004 ■ 0047327 11 [11, *0, *0, 004737 ■ 0047327 12 [15, *0, *0, 004737 ■ 0047327 13 [15, *0, *0, 004737 ■ 004733 14 [15, *0, *0, 004738] ■ 004743 14 [15, *0, *0, 004738] ■ 004743 14 [15, *0, *0, 004748] ■ 004743 14 [15, *0, *0, 004748] ■ 004743 14 [11, *0, *0, 004748] ■ 004748 14 [11, *0, *0, *0, *0, *0, *0, *0, *0, *0, *0	<pre>ser + 0.016* Toolinfecti + 0.008* "elimin + 0.016* "popul" + 0.016* "committe" + 0.016* "plan" + 0.016* "advisori + 0.008* "meid" + exatinet. 0.008* "incid" + 0.008* "meid" + 0.008* "perman" + 0.008* "meid" + 0.002* "gender + 0.008* "meid" + 0.002* "gender + 0.002* "gender + 0.008* "meid" + 0.002* "gender + 0.008* "meid" + 0.002* "gender + 0.002*</pre>
Run: LDA 2 Jabeling CSV	
<ul> <li>Nosoccial infection pseudoinfection e Survival septic guinea pigs vitamin E Survival septic guinea pigs vitamin E Retropharyngeal space infection changi Controlled trial insulin infusion pare Controlled trial insulin infusion pare Controled trial insulin infusion pare Controled trial insul</li></ul>	Palastons Impact antibiotic decision making authors retrospective analysis mecnatal superficial cultures effect antimicrobial decision mi descopes bronchoscopes Microsnin Hissonir Harshell fiber-optic endoscopes United diagnostic therapeutic procedures Between endoscopes mach itamin C enteral disto Gxygen-free radicals aspais injury dysfunction effect different levels vitamins E C diet fed septic guines page 3 g transf Retropharyngeal page infection rare entry introduction modern antibiotic chemotherapy children years age supportant disorders ition pouches refrigerator unusual cause Tieven patients hospital septicemia Enterobacter cloaces origin contamination parenteral nutritic icin production primary mine lisericius immuse phase bacterial islamination TNP production VVO IRV-gemma important cole marine hospital icin production primary mine lisericius immuse phase bacterial islamination INP production VVO IRV-gemma important cole marine hospital cephalosportin high antipeedomonal activity Bo new parenteral cephalosportin catechol molety position cephalosportin ring high vitro activit descomendations Advisory Committee Elimination Dusculous perinterus Henlihotion ELISA Mycobacterium tuberculosis antiger "gender" + 0.028" "cephalosportin" + 0.028" "engravita" + 0.023" "merind" + 0.023"
🕨 4: Run 🛛 🚔 Python Console	
Typo: In word 'activ' or 5 vvoras: 42 😻 English (United States)	9-32 CRLF= UTF-81 % @ @
ho  Type here to search	O 🛱 🔀 🛼 File 😂 😭 🚺 式 🕐 pyth 🔯 ⊘ 🕼 jas 🔄 Skype ^ 📼 🔅 40) 10.34

Results after LDA



Image: - 0 + 0 - 1       Image: - 0 + 0 - 0 - 0 + 0 + 0 + 0       Image: - 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0	e <u>E</u> dit <u>V</u> iew	<u>N</u> avigate <u>C</u> ode <u>H</u> elp															
<pre>andpoint file and person file and file for the file</pre>	Project		👍 sel_cl.py	y 🛛 🐞 LDA	A.py 🛛 👘 dı	uplicateremove.p	y 🛛 🗐 analys	is 🛛 👘 LDA_2.	py 🗵 🚺 para	m_result.py ×	🗒 one 🖯	0000335	i 🗵 👘 projec	t.py ×			
# sparse_matrix.npz       hamming_score() > for i in range(y_true.shape(0))         Run:       param_tesut       habeling_CSV         Hamming loss score:       0.125         classification accuracy:       (0) 0.875         Process finished with exit code 0         Image: Sparse_matrix.npz		stopwords Rhistory analysis data_helpers.py duplicateremove.py babeling_CSV.py model.py param_result.py project.py README.md README.md sample1.py sample2.py samplers.put.csv	19         20           21         ←           22         23           23         24           25         ⊖           27         ←           28         29           30         31           32         ⊖           33         ⊖11           34         35	set, fpr: fpr: tmp, if : els: fpr: acc return n if pr: acc print(')	<pre>pred = set int('inset_int('set pro- a = None len(set_true tmp_a = len tmp_a = len flo int('tmp_a; list.append np.mean(acc_ == " main Hamming loss</pre>	<pre>( np.where(y_j true: (0)'.form ed: (0)'.form e) == 0 and le n(set_true.int h(set_true.int h(set_true.int (0)'.format(' (1(tmp_a)) ]ist) ": s score: (0)'.</pre>	<pre>interf(1)(0); mat(set_true tf(set_pred)) en(set_pred)) en(set_pred) ersection(se crue.union(se crue.union(se mp_s))</pre>	))_ == 0: t_pred))_(\ t_pred))_) ng_score(y_tr	e, y_pred)))		i one >	k ₩ 0000335	j × 👫 projec	t.py ×			
Image: State of the state		👍 sparse.py	36														
Run: param_result habeling_CSV ** 1 Hamming loss score: 0.125 classification accuracy: {0} 0.875 Frocess finished with exit code 0 ** 1 ** 1 Packages installed successfully: Installed packages: 'stop-words' (yesterday 2247) 2931 CRLF: UTF-8: 0 0 2931 CRLF: UTF-8: 0 0 0000			ł	hamming sco	re() → foriin	range(v true.shar	pe[0])										
classification accuracy: {0} 0.875 Process finished with exit code 0				-	-	5.0- 1										<u>4</u> -	+
▶ ± Run       ♣ Python Console         Packages installed successfully: Installed packages: 'stop-words' (yesterday 22:47)       29:31 CRLF: UTF-8: 6         ≥ or > words: so @ registion (united states)		lassification accura	су: {0} 0.8														
	▶ <u>4</u> : Run Packages insta	talled successfully: Installe	d packages: 's	stop-words' (y	resterday 22:47)											F-8≑ îa	tig
			in succes		0 Ħ	8	File Ex 🤁		× 0 P	ytho 🔷		🤨 jasmin	-		= 11) 0	00:09 05-2020	F

Hamming loss and classification accuracy results

		]	BR (Bi	nary Rele	evance) 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

# RESULTS

# Binary Relevance Classifier output

-



Volume: 06 Issue: 05 | May - 2022

Impact Factor: 7.185

ISSN: 2582-3930

			ML (n	nulti-labe	el) KNN Classifier
	ML(multi label)			F1-	
Dataset	Accuracy	Precision	Recall	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

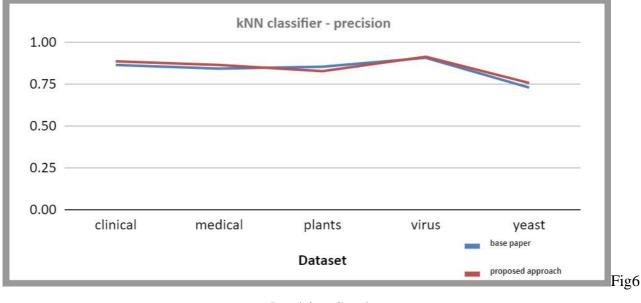
		Proposed Approach	
	Hamming		
Dataset	Loss	Accuracy	
clinical	0.114	0.8053	
medical	0.125	0.875	
Plants	0.148	0.7248	
Virus	0.085	0.7698	
Yeast	0.035	0.5477	

Accuracy Outcomes with Hamming Loss



Comparison of Base Paper kNNClassifier Precision Results WithBase 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



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

https://www.google.com/search?q=lda+machine+learning&sxsrf=APq- 2.
 WBvoZKrg1ykODIJIcJJRuTbOEXcG8g%3A1643876186790&ei=Wo\_7YYfcL4ejoATvy4ngDQ&o
 q
 =lda&gs\_lcp=Cgdnd3Mtd2l6EAEYADIECCMQJzIECAAQQzIECAQQZAQQSQM6EAguELEDEIMBEMcBE
 N
 EDEEM6BQgAELEDOg4ILhCABBCxAxDHARCjAjoLCC4QgAQQxwEQ0QM6DQguELEDEM
 C
 BEKMCEEM6BQguEIAESgUIPBIBMUoECEEYAEoECEYYAFCilQNY56sDYNTPA2gCcAJ4BI
 A
 BgQKIAbgLkgEFMC43LjGYAQCgAQGwAQrIAQnAAQE&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 Ver- beek, 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,<sup>1,2</sup> Qingyu Chen,<sup>1</sup> Yifan Peng,<sup>1</sup> Yang Xiang,<sup>2</sup> Cui Tao,<sup>2</sup> and Zhiyong Lu<sup>1</sup> <sup>1</sup>National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland, USA, and <sup>2</sup>The University of Texas School of Biomedical Informatics, Houston, Texas, USA

Corresponding Author: Zhiyong Lu, PhD, National Center for Biotechnology Information, National Library of Medicine, Na- tional 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<sup>a,1</sup>, Bruno CADONNA<sup>b</sup>, Ivo RAKOVAC<sup>a</sup>, Christian GÜTL<sup>c</sup>, Hubert LEITNER<sup>d</sup>, Günther STARK<sup>d</sup>, Thomas R. PIEBER<sup>a,e</sup>, Peter BECK<sup>a a</sup>JOANNEUM RESEARCH Forschungsges.m.b.H., Institute for Biomedicine and Health Sciences, Graz, Austria
<sup>b</sup>Free University of Bozen – Bolzano, Faculty of Computer Science, Bolzano, Italy <sup>c</sup>Graz, University of Technology, Institute for Information Systems and Computer Media, Graz, Austria
<sup>d</sup>Steiermärkische Krankenanstaltenges.m.b.H., Graz, Austria
<sup>e</sup>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.<u>https://www.javatpoint.com/machine-learning</u>



18 https://vitalflux.com/machine-learning-list-of-35-free-online-books/

19. https://towardsdatascience.com/natural-language-processing-nlp-for-machine-learning-

<u>d44498845d5b#:~:text=NLP%20is%20a%20field%20in,and%20potentially%20generate%20human</u> %20language.&text=Information%20Retrieval(Google%20finds%20relevant,Gmail%20structures% 20events%20from%20emails).

20. https://www.javatpoint.com/python-tutorial

L