International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Data Mining, Spidering and Analysis with Python

Tejas Kamble, Srujan Garde, Samruddhi Kadam
Department of Computer Engineering, Zeal College of Engineering and Research, Pune, Maharashtra, India

Article Info
Volume9, Issue 2
Page Number: 440-447

Publication Issue
March-April-2022

Article History
Accepted :03March2022
Published :10March2022

ABSTRACT

Data mining and analytics have played an important role in knowledge discovery and decision making/supports in the process industry over the past several decades. As a computational engine to data mining and analytics, machine learning serves as basic tools for information extraction, data pattern recognition and predictions. From the perspective of machine learning, this paper provides a review on existing data mining and analytics applications in the process industry over the past several decades. The state of the art of data mining and analytics are reviewed through eight unsupervised learning and ten supervised learning algorithms, as well as the application status of semi-supervised learning algorithms. Several perspectives are highlighted and discussed for future researches on data mining and analytics in the process industry.

Index theme: Data mining, Data spidering, Data wrangling, Data Analysis, Data Manipulation

I. INTRODUCTION

What is Data Mining?

Data Mining is a set of method that applies to large and complex databases. This is to eliminate the randomness and discover the hidden pattern. As these data mining methods are almost always computationally intensive.We use data mining tools, methodologies, and theories for revealing patterns in data. There are too many driving forces present. And, this is the reason why data mining has become such an important area of study.[1]

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of extracting information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.[1]

History of Data Mining

In 1960s statisticians used the terms “Data Fishing” or “Data Dredging”. That was to refer to what they

considered the bad practice of analyzing data. The term "Data Mining" appeared around 1990 in the database community.

Importance of data mining

As data mining is having spacious applications. Thus, it is the young and promising field for the present generation. It has attracted a great deal of attention in the information industry and in society.

Due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Thus, we use information and knowledge for applications ranging from market analysis. This is the reason why data mining, known as knowledge discovery from data.[2]

Data Mining Techniques

a. Artificial Neural Networks We use data mining in non-linear predictive models. As this learn through training and resemble biological neural networks in structure.[1]
b. Decision Trees As we use tree-shaped structures to represent sets of decisions. Also, these rules are generated for the classification of a dataset. These decisions generate rules for the classification of a dataset.[3]

   As there are specific decision tree methods that include Classification and Regression Trees and Chi-Square Automatic Interaction Detection (CHAID).
c. Genetic Algorithms There are the present genetic combination, mutation, and natural selection for optimization techniques. That is design based on the concepts of evolution.[2]
d. Nearest Neighbor Method A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) like. it in a historical dataset (where k ³ 1). Sometimes called the k-nearest neighbour technique.
e. Rule Induction The extraction of useful if-then

Below are 5 data mining techniques that can help you create optimal results.

1) Classification analysis

This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified. A classic example of classification analysis would be Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.[4]

2) Association rule learning

It refers to the method that can help you identify some interesting relations (dependency Modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset. Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design, and store layout. In IT, programmers use association rules to build programs capable of machine learning.[4]

3) Anomaly or outlier detection

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations, and exceptions. Often, they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the

ordinary has happened and requires additional attention. This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy.[4]

4) Clustering analysis

The cluster is a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different, or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise. A result of this analysis can be used to create customer profiling.[4]

5) Regression analysis

In statistical terms, a regression analysis is the process of Identifying and Analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting.

All of these data mining techniques can help analyze different data from different perspectives. Now you have the knowledge to decide the best technique to summarize data into useful information – information that can be used to solve a variety of business problems to increase revenue, customer satisfaction, or decrease unwanted cost.[4]

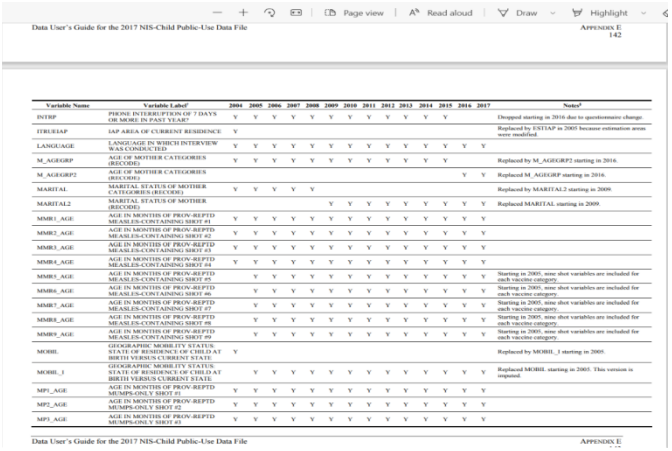
II. IMPLEMENTATION OVERVIEW

Data variables are the core objects of data mining. Data variable gives us the information about where to look at in Data set.

We have enough scrapped Data From National Immunization Survey-Child, now we are going to spider through the data and do analysis of it.

Experimental work: Data set-National immunization survey-child

- Programming language: Python
- Libraries/moduls used: Pandas,numpy,scipy And
- Compiler:Jupyternotebook,ipython.

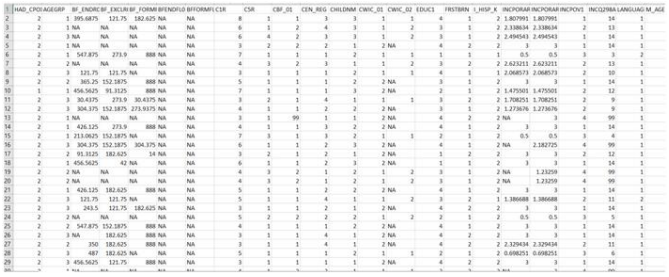


Data User's Guide for the 2017 NIS-Child Public-Use Data File — APPENDIX E, 142

Variable Name	Variable Label	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	Notes
INTRP	PHONE INTERRUPTION OF 7 DAYS OR MORE IN PAST YEAR?	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Dropped starting in 2016 due to questionnaire change.
IFRUEIAP	IAP AREA OF CURRENT RESIDENCE	Y														Replaced by ESTIAP in 2005 because estimation areas were modified.
LANGUAGE	LANGUAGE IN WHICH INTERVIEW WAS CONDUCTED	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
M_AGEGRP	AGE OF MOTHER CATEGORIES (RECODE)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Replaced by M_AGEGRP2 starting in 2016.
M_AGEGRP2	AGE OF MOTHER CATEGORIES (RECODE)													Y	Y	Replaced M_AGEGRP starting in 2016.
MARITAL	MARITAL STATUS OF MOTHER CATEGORIES (RECODE)	Y	Y	Y	Y	Y										Replaced by MARITAL2 starting in 2009.
MARITAL2	MARITAL STATUS OF MOTHER (RECODE)						Y	Y	Y	Y	Y	Y	Y	Y	Y	Replaced MARITAL starting in 2009.
MMR1_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #1	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			
MMR2_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			
MMR3_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #3	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			
MMR4_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #4	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			
MMR5_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #5		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Starting in 2005, nine shot variables are included for each vaccine category.
MMR6_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #6		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Starting in 2005, nine shot variables are included for each vaccine category.
MMR7_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #7		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Starting in 2005, nine shot variables are included for each vaccine category.
MMR8_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #8		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Starting in 2005, nine shot variables are included for each vaccine category.
MMR9_AGE	AGE IN MONTHS OF PROV-REPTD MEASLES-CONTAINING SHOT #9		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Starting in 2005, nine shot variables are included for each vaccine category.
MOBIL	GEOGRAPHIC MOBILITY STATUS: STATE OF RESIDENCE OF CHILD AT BIRTH VERSUS CURRENT STATE	Y														Replaced by MOBIL_1 starting in 2005.
MOBIL_1	GEOGRAPHIC MOBILITY STATUS: STATE OF RESIDENCE OF CHILD AT BIRTH VERSUS CURRENT STATE		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Replaced MOBIL starting in 2005. This version is imputed.
MP1_AGE	AGE IN MONTHS OF PROV-REPTD MUMPS-ONLY SHOT #1		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
MP2_AGE	AGE IN MONTHS OF PROV-REPTD MUMPS-ONLY SHOT #2		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	
MP3_AGE	AGE IN MONTHS OF PROV-REPTD MUMPS-ONLY SHOT #3		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	

Data User's Guide for the 2017 NIS-Child Public-Use Data File — APPENDIX E

variable referencing column name



	HAD_CPOI	AGEGRP	BF_ENDR	BF_EXCLR	BF_FORMR	BFENDFLG	BFFORMFLG	CSR	CBF_01	CEN_REG	CHILDNM	CWIC_01	CWIC_02	EDUC1	FRSTBRN	I_HISP_K	INCPOVAR	INCPOVAR	INCPOVI	INCQ298A	LANGUAG	M_AGE
2	2	1	395.6875	121.75	182.625	NA	NA		8	1	1	3	3	1	1	4	1	2	1.807991	1.807991	1	14	1
3	2	1	NA	NA	NA	NA	NA		6	1	2	4	3	1	2	3	1	2	2.338634	2.338634	2	13	1
4	2	1	NA	NA	NA	NA	NA		6	4	2	3	3	1	1	2	3	1	2.494543	2.494543	1	14	1
5	2	1	NA	NA	NA	NA	NA		3	2	2	2	1	2	NA	4	2	2	3	3	1	14	1
6	2	1	547.875	273.9	888	NA	NA		7	1	1	1	2	1	1	1	1	1	0.5	0.5	3	3	2
7	2	2	NA	NA	NA	NA	NA		4	3	2	3	3	1	1	2	3	2	2.622321	2.622321	2	13	1
8	2	3	121.75	121.75	NA	NA	NA		3	1	1	1	2	1	1	4	1	1	2.068573	2.068573	2	10	1
9	2	2	365.25	152.1875	888	NA	NA		5	1	1	1	2	2	NA	3	1	2	3	3	1	14	1
10	1	1	456.5625	91.3125	888	NA	NA		7	1	1	1	3	2	NA	2	1	2	1.475501	1.475501	2	12	1
11	2	3	30.4375	273.9	30.4375	NA	NA		3	2	1	4	1	1	1	3	2	2	1.708251	1.708251	2	9	1
12	2	3	304.375	152.1875	273.9375	NA	NA		4	1	1	2	2	2	NA	3	1	2	1.273676	1.273676	2	9	1
13	2	1	NA	NA	NA	NA	NA		3	1	99	1	1	2	NA	4	2	2	NA	3	1	4	99
14	2	1	426.125	273.9	888	NA	NA		4	1	1	3	2	2	NA	4	1	2	3	3	1	14	1
15	2	1	213.0625	152.1875	NA	NA	NA		7	3	1	3	2	1	1	2	1	2	0.5	0.5	3	4	1
16	2	3	304.375	152.1875	304.375	NA	NA		6	1	1	2	3	2	NA	4	1	2	NA	2.182725	4	99	1
17	2	2	91.3125	182.625	14	NA	NA		3	2	1	2	1	2	NA	1	1	2	2	3	2	12	1
18	2	1	456.5625	42	NA	NA	NA		6	1	1	2	3	2	NA	1	1	2	2	3	2	14	1
19	2	2	NA	NA	NA	NA	NA		4	3	2	1	2	1	2	3	1	2	NA	1.23259	4	99	1
20	2	2	NA	NA	NA	NA	NA		4	3	2	1	2	1	2	3	1	2	NA	1.23259	4	99	1
21	2	1	426.125	182.625	888	NA	NA		5	1	1	2	2	2	NA	4	1	2	3	3	1	14	1
22	2	3	121.75	121.75	NA	NA	NA		5	1	1	4	1	1	2	3	2	1	1.386688	1.386688	2	11	2
23	2	3	243.5	121.75	182.625	NA	NA		3	1	1	2	1	2	NA	4	2	2	3	3	1	14	1
24	2	2	NA	NA	NA	NA	NA		5	2	2	2	2	1	2	2	1	2	0.5	0.5	3	5	1
25	2	2	547.875	152.1875	888	NA	NA		4	1	1	1	2	2	NA	4	1	2	3	3	1	14	1
26	2	3	NA	182.625	888	NA	NA		3	1	1	4	1	2	NA	4	2	2	3	3	1	14	1
27	2	2	350	182.625	888	NA	NA		3	1	1	4	1	2	NA	4	2	2	2.329434	2.329434	2	11	1
28	2	3	487	182.625	NA	NA	NA		5	1	1	1	2	1	1	2	1	2	0.698251	0.698251	3	6	1
29	2	3	456.5625	121.75	888	NA	NA		3	1	1	1	1	2	NA	4	2	2	3	3	1	14	1

III. DATA ANALYSIS

Analysis 1

- Average number of influenza vaccines for those children we know received breast-milk as a child and those who know did not.
- Mining the data related to requirement ,we need cells of data containing no of vaccine and child who received breast-milk,are follows

”CBF_01”, ”P_NUMFLU ”

```
def average_influenza_doses() :
ave_milkfed_dose =
np.mean(milk_fed[”P_NUMFLU ”])
ave_milknotfed_dose =
```

np.mean(milk_notfed["P_NUMFLU "]) ave_up = (ave_milkfed_dose, ave_milknotfed_dose)
return ave_tup
average_influenza_doses()
(1.8799187420058687, 1.5963945918878317)

Analysis 2

- . Calculate the ratio of the number of children who contracted chickenpox but were vaccinated against it (at least one varicella dose) versus those who were vaccinated but did not contract chicken pox. Return results by sex.$

def chickenpox_by_sex() :
CPox ={}
CPox["male"] = male
CPox["female"]=female
return CPox
chickenpox_by_sex()
′ male ′ : 0.009675583380762664, ′ female ′ : 0.0077918259335489565

Analysis 3

we might look at the correlation between the use of the vaccine and whether it results in prevention of the infection or disease.

- Some notes on interpreting the answer. The had-chickenpox-column is either 1 (for yes) or 2 (for no), and the num-chickenpox-vaccine-column is the number of doses a child has been given of the varicella vaccine. A positive correlation (e.g., corr ¿ 0) means that an increase in had-chickenpox-column (which means more no's) would also increase the values of num-chickenpox-vaccine-column (which means more doses of vaccine). If there is a negative correlation (e.g., corr ¡ 0), it indicates that having had chickenpox is related to an increase in the number of vaccine doses. $

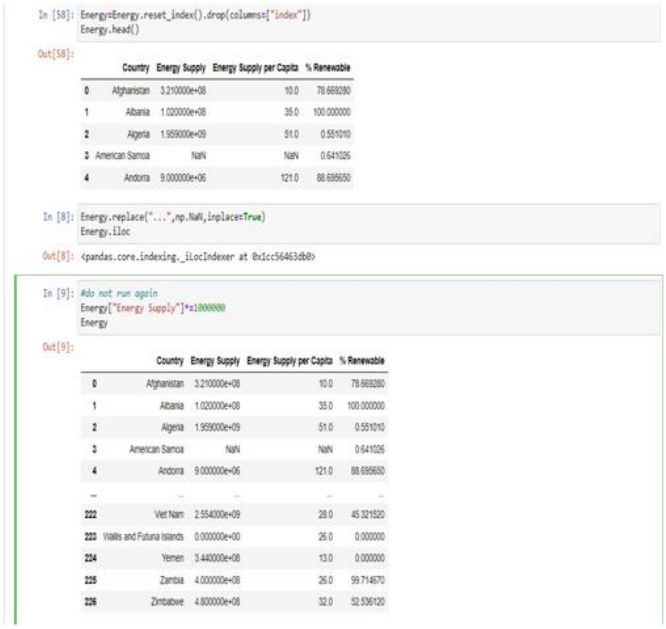
req=["HAD_CPOX", "P_NUMV RC"] $
def corr_chickenpox() :
corr, pval = stats.pearsonr(had_cpox["HAD_CPOX"], had_cpox["P_NUMVRC"])
return corr
corr_chickenpox()
07044873460147986

IV. ENERGY SUPPLY DATA ANALYSIS

1) Manipulation

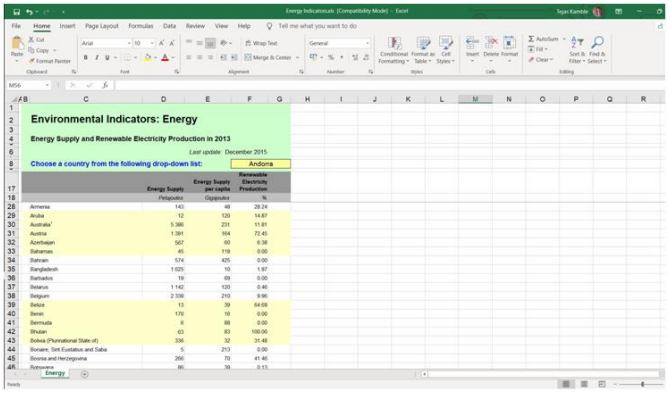
Convert Energy Supply to gigajoules (Note: there are 1,000,000 gigajoules in a petajoule). For all countries which have missing data (e.g. data with "...") make sure this is reflected as np.NaN values.

Energy["Energy Supply"]*=1000000 Energy
Easy manipulation of data.



	Country	Energy Supply	Energy Supply per Capita	% Renewable
0	Afghanistan	3.210000e+08	10.0	78.669280
1	Albania	1.020000e+08	35.0	100.000000
2	Algeria	1.959000e+09	51.0	0.551010
3	American Samoa	NaN	NaN	0.641026
4	Andorra	9.000000e+06	121.0	88.695650

	Country	Energy Supply	Energy Supply per Capita	% Renewable
0	Afghanistan	3.210000e+08	10.0	78.669280
1	Albania	1.020000e+08	35.0	100.000000
2	Algeria	1.959000e+09	51.0	0.551010
3	American Samoa	NaN	NaN	0.641026
4	Andorra	9.000000e+06	121.0	88.695650
...	...	...	...	...
222	Viet Nam	2.554000e+09	28.0	45.321520
223	Wallis and Futuna Islands	0.000000e+00	26.0	0.000000
224	Yemen	3.440000e+08	13.0	0.000000
225	Zambia	4.000000e+08	26.0	99.714670
226	Zimbabwe	4.800000e+08	32.0	52.536120



Data set-Energy supply and renewable electricity production

2) Spidering and Grabbing

Top 15 countries for average GDP over the last 10 years. This function should return a Series named avgGDP with 15 countries and their average GDP sorted in descending order.

def gdp_avg(row):

```
  data=row[['2006', '2007', '2008','2009', '2010', '2011', '2012', '2013', '2014', '2015']]
  return pd.Series({"avgGDP":np.nanmean(data)})
avgGDP=GDP_avg.apply(gdp_avg,axis="columns")
avgGDP=avgGDP.sort_values(by="avgGDP",ascending=False)
avgGDP
```

Top 15 countries for average GDP over the last 10 years Country
United States 1.536434e+13
China 6.348609e+12 Japan 5.542208e+12
Germany 3.493025e+12 France 2.681725e+12
United Kingdom 2.487907e+12 Brazil 2.189794e+12
Italy 2.120175e+12 India 1.769297e+12
Canada 1.660647e+12
Russian Federation 1.565459e+12 Spain 1.418078e+12
Australia 1.164043e+12 South Korea 1.106715e+12
Iran 4.441558e+11
Name: avgGDP, dtype: float64

3) Spidering info

.GDP changed over the 10 year span for the country with the 6th largest average GDP $

```
def answer_four():
change_GDP=main_ds[main_ds["Rank"]==4]["2015"]-main_ds[main_ds["Rank"]==4]["2006"]
  return pd.to_numeric(change_GDP)[0]
answer_four()
246702696075.3999
mean energy supply per capita
def answer_five(main_ds):
  return np.mean(main_ds["Energy Supply per Capita"])
answer_five(main_ds)
157.6
```

4) 5Learning data

What country has the maximum percentage Renewable and what is the percentage $

```
def answer_six():
max_renew=main_ds["% Renewable"].idxmax(),np.max(main_ds["% Renewable"])
  return max_renew
answer_six()
('Brazil', 69.64803).
```

New column that is the ratio of Self-Citations to Total Citations. What is the maximum value for this new column, and what country has the highest ratio.

```
def answer_seven():
  main_ds["Citation_ratio"]=main_ds["Self-citations"]/main_ds["Citations"]
max_cit_ration=main_ds["Citation_ratio"].idxmax(),np.max(main_ds["Citation_ratio"])
  return max_cit_ration
answer_seven()
('China', 0.6893126179389422)
```

5) Data sorting

. Create a column that estimates the population using Energy Supply and Energy Supply per capita. What is the third most populous country according to this estimate

```
def answer_eight():
est_pop=main_ds
est_pop["PopEst"]=est_pop["Energy Supply"]/est_pop["Energy Supply per Capita"]
max_pop=est_pop.sort_values(by="PopEst",ascending=False)
  return max_pop.index[2]
answer_eight()
'United States'
```

6) Statastical Info

. Create a column that estimates the number of citable documents per person. What is the correlation between the number of citable documents per capita and the energy supply per capita? Use the .corr() method, (Pearson's correlation).

```
def answer_nine():
cit_doc=main_ds
```

```
cit_doc["Citabledocs per Capita"]=cit_doc["Citabledocuments"]/cit_doc["PopEst"]
cit_doc["Citabledocs per Capita"]=cit_doc["Citable docs per Capita"].astype(float)
cit_doc["Energy Supply per Capita"]=cit_doc["Energy Supply per Capita"].astype(float)
  return cit_doc["Citable docs per Capita"].corr(cit_doc["Energy Supply per Capita"])
answer_nine()
7940010435442946
```

def plot9(): import matplotlib as plt %matplotlib inline

Top15 = answer_one() Top15['PopEst'] = Top15['Energy Supply'] / Top15['Energy Supply per Capita'] Top15['Citable docs per Capita'] = Top15['Citable documents'] / Top15['PopEst'] Top15.plot(x='Citable docs per Capita', y='Energy Supply per Capita', kind='scatter', xlim=[0, 0.0006])

7) Mapping

```
def answer_eleven():
grp_cont=main_dsContinentDict = {'China':'Asia', 'United States':'North America', 'Japan':'Asia', 'United Kingdom':'Europe', 'Russian Federation':'Europe', 'Canada':'North America', 'Germany':'Europe', 'India':'Asia', 'France':'Europe', 'South Korea':'Asia', 'Italy':'Europe', 'Spain':'Europe', 'Iran':'Asia', Australia':'Australia', 'Brazil':'South America'}
grp_cont['Continent'] = grp_cont.index.to_series().map(ContinentDict)
grp_cont=grp_cont.groupby("Continent")["PopEst"].agg(["size","sum","mean","std"])
  return grp_cont
answer_eleven()
```

Continent	size	sum	mean	std
Asia	5	2.898666e+09	5.797333e+08	6.790979e+08
Australia	1	2.331602e+07	2.331602e+07	NaN
Europe	6	4.579297e+08	7.632161e+07	3.464767e+07
North America	2	3.528552e+08	1.764276e+08	1.996696e+08
South America	1	2.059153e+08	2.059153e+08	NaN

V. APPLICATIONS

Market Basket Analysis

Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behaviour of a buyer. This information may help the retailer to know the buyer's needs and change the store's layout accordingly. Using differential analysis comparison of results between different stores, between customers in different demographic groups can be done.

Education

There is a new emerging field, called Educational Data Mining, concerns with de- veloping methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learn- ing behaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the student.

Fraud Detection

Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

Intrusion Detection

Any action that will compromise the integrity and confidentiality of a resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and

information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

Lie Detection

Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This file includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

Financial Banking

With computerised banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately ap- parent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting,targeting, acquiring, retaining and maintaining a profitable customer

Research Analysis

History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualisation and visual data mining provide us with a clear view of the data.

Criminal Investigation

Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complex- ity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. These information can be used to perform crime matching process.

VI. PRO'S AND CON'S

❖ **Pro's of Data Mining**

Better Customer Relationship Management

Being able to ensure good customer relationship management is one of the key advantages of data mining. It helps businesses know what type of customers to approach with different kinds of products. This guarantees the sale of the product and not the pitching of the product.

Forecasting Market Trends

Marketing and retailing depend on the current market trends that are followed by customers. Data mining allows these industries to find the correct trends through market research which, in turn, helps them in choosing their marketing strategies.

Helps Stay Ahead of Competition

With so much new data, well analyzed, your RD department will be at the fore- front of trends and will be able to think about the next product. Data mining offers many advantages over the possibilities of personalization, consistency with the current and future needs of consumers.

Anomaly Detection with more accurate Analysis

The analysis is much more accurate with data mining since it is possible to classify all the information according to the priorities that you previously identified. It is capable of analyzing databases with a huge amount of data. Data mining can become very

useful for various financial institutions. Banks and credit card companies can obtain information on loans and know the creditworthiness of customers.

❖ **Con's of data mining**

Expensive in the Initial Stage:

With a large amount of data getting generated every day, it is pretty much ev- ident that it will draw a lot of expenses associated with its storage as well as maintenance. This is one of the main disadvantages of data mining.

In order to successfully operate data mining, your company needs the appropriate specialists.Dependingon the type of data you want to collect, a lot of work may be required, or sometimes the initial investment to obtain the technologies needed for data collection can be very expensive.

Security of the Critical Data:

Companies hold a lot of critical information on their customers and employees as well. There's always a risk of being hacked, as a massive amount of valuable data gets stored in the data mining systems. Security issues during data mining

Non-Verified data updation,Security architect evaluation,Dataanonymization,Filtering validating external sources,Data storage location,Distributed frameworks for data.

Data Mining Violates User Privacy:

It is comprehended that data mining uses market-based techniques to gather data on people. Most of the time, private information that companies hold is traded to others or leaked.

Organizations gather information on their consumers in several ways to under- stand their purchasing behaviour and much more.

Lack of Precision or Incorrect Information

The data mining tools analyze data without actually knowing its meaning. Theypresent the results in the form of various visualizations. However, these patterns are not meaningful by themselves, but only after the user has assessed them. eg If incorrect information is applied for decision-making, it can cause severe out- comes.

VII. FUTURE SCOPE

- In the future, data mining will include more complex data types. In addition, for any model that has been designed, further refinement is possible by examining other variables and their relationships.
- Research in data mining will result in new methods to determine the most in- teresting characteristics in the data. As models are developed and implemented, they can be used as a tool in enrollment management.

VIII. CONCLUSION

Data mining, along with traditional data analysis, is a valuable tool that that is being used in Strategic Enrollment Management to achieve desired enrollment targets in colleges and universities. In situations where it has been applied, it has been proven to successfully predict enrollment, at least to a degree. More research is needed to fully take advantage of the data mining processes and technologies

IX. REFERENCES

[1]. G. S. L. M. J. A. Berry, Mastering Data Mining, New York: Wiley, 2000.

[2]. H. M. a. P. S. D. Hand, Principles of Data Mining., Cambridge, : MA:MIT Press., 2001.

[3]. J. F. R. O. a. C. S. L. Breiman, Classification and Regression Trees., Wadsworth, 1984.

[4]. G. S. L. M. J. A. Berry, Data Mining Techniques, New York: Wiley, 1997.