

# DWDM

## Mini Project

---



Roll no	Name	ERP Number
PE03	Vishwajeet Shinde	1032190069
PE04	Aniruddha Shende	1032190079
PE06	Hrishikesh Vaze	1032190087
PE16	Prajwal Singh	1032190822

# Problem Statement

---

- Data mining with web scraping and retrieving data in correct format. Analyze data about research papers(IEEE). Making inference of given data using visualization and ml algorithm
- Research papers is best way to demonstrate any topic to all world and now many people are publishing their research papers in different websites such as IEEE etc. A research paper is an essay in which you explain what you have learned after exploring your subject in depth. In this you include information from resources such as books, articles, interviews, and websites. You also use your own ideas, knowledge, and opinions.
- In different colleges teachers and students are publishing their papers on the internet but it's huge in numbers and so it takes a huge amount of time to keep a record of each and every paper. For that automation is a good way which helps us to save a lot of time and our work. So this system is to automate this work i.e to keep record of every research paper with a single link of that paper. With help of link /url given this system uses selenium with python for web scraping which gets different attributes related to papers like author, data, ISSN, DOI title, publication etc. this data helps us to keep record of every paper and gives information about that papers in short. So by using xpath and other selenium html parser we get that data and save it in a dataframe. After that this system create a csv file of that data and preprocess it with different preprocessing techniques like checking empty values, data transformation, visualization and correlation analysis.
- After getting data in correct format by using different visualization tools like Tableau public and orange get different visualizations to understand the pattern and data relatively. Then we can apply machine learning model to get inference of data.

# Purpose

---

- Extract data of IEEE research papers from IEEE websites to collect data by using web scraping, preprocessing data and making inferences of it.
- The goal of this project is to retrieve data about IEEE papers from given links and analyze it.
- After getting data in the correct format, visualize data with different tools and make inferences of it by using ML algorithms.

# Literature Survey Continued

---

Literature Survey

Serial Number	Research Paper Info	Dataset	Technology/ Tools	Conclusion	Limitations/ Future Work
1	Web <u>scrapping</u> and storing data in a database, a case study of the used cars market	Cars Dataset	BeautifulSoup Pandas Numpy Python	it deals with the process of web scraping data from different locations on the Internet for the purpose of collecting and analyzing data of the used cars market.	It can be used in further analyzes.
2	WEB-BASED COLLEGE PLACEMENT ANALYSIS  Akhilesh Shinde, Aniket Shinde, Vivek Singh, Siddhesh Shivdikar, Sharvari Govilkar  Published 5 October 2021	College Placement dataset	Python Matplotlib BeautifulSoup Pandas Numpy	This project can be used in college for their placement activities. colleges can rectify the problems because of which companies are not coming for placement and use for the betterment of the college placement	1. The Dataset can be further extended by adding current and next year recurring placement data 2. Working for more accuracy

# Literature Survey

---

3	Data Analysis by Web Scraping using Python  David Mathews, Sandeep Mathur	Retail	Beautiful Soup Pandas Matplotlib	Thus categorized retail dataset.	Hidden web data need synthetic and semantic matching to fully achieve automatic integration.
4	Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice  Stephanie Lunn	Dataset from indeed.com	Beautiful Soup LXML WordCloud Pandas Matplotlib	It demonstrates techniques that could be utilized for numerous applications to further knowledge in CSE and analysis of it.	More detailed analysis
5	Comparison of E-commerce Products using web mining By Riya Shah	Dataset created for ecommerce price data	Django Mongodb	It will help users in decision making while buying products online. This website will facilitate users to analyze prices that are present on different e-commerce shopping websites	Can be combined together to get automatic suggestions from the website itself for searched products. Further use of AI/NLP

# Abstract

---

- This system extracts data from different IEEE paper's websites with help of automation via selenium using python. This system uses XPath with selenium for web scraping to get data that contain different attributes of IEEE papers from its website. Then it converts output data to a CSV file on which preprocessing is done with help of pandas and NumPy.
- It will help us to automate data mining and create excel files containing all types of attributes related to IEEE research papers. This describes how web scraping and natural language processing can be utilized to get correct data related to research papers from different websites. It demonstrates how web scraping can be useful for extrapolating large amounts of data from publicly available web pages to extract data from a wide array of sources and to further information in the field. We discuss how natural language processing can be used to reliably obtain salient information from textual data, and how it can complement qualitative analysis.

# Methodologies

---

## Data mining :

1.selenium web scraping

## Data pre-processing:

1.checking empty values

2.data transformation

## Visualization

1.orange

## Machine learning model

1.Clustering - Kmean text

# WORKFLOW DIVIDED INTO 3 PARTS

---

## **Data mining**

- The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining

## **Data preprocessing**

- The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

- 

## **Machine learning model -Kmean**

- A machine learning algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data. Machine learning (ML) algorithms are broadly categorized as either supervised or unsupervised.
- K-Means clustering intends to partition  $n$  objects into  $k$  clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly  $k$  different clusters of greatest possible distinction.
- As we applied  $k$  mean clustering we can get clusters which are categorized journal wise and citations are clustered accordingly. So it will give us more understanding of data.



# Dataset

Unnamed: 0	authors	title of paper	journal name	citations	doi	page no	month of publication	year of publication	ISSN	url	Journal_encoded
0	Andrija Goranović, Marcus Meisel, Stefan Wilke...	Hyperledger Fabric Smart Grid Communication Te...	IEEE Conference Publication	5.0	10.1109/WFCS.2019.8758000	27-29	July	2019.0	1882-2161.0	<a href="https://ieeexplore.ieee.org/abstract/document/...">https://ieeexplore.ieee.org/abstract/document/...</a>	2
1	Obaidur Rahman, Kashem M. Muttaqi, Danny Sutanto	Three Phase Power Flow Analysis of Distributio...	IEEE Conference Publication	163.0	10.1109/AUPEC.2018.8758001	27-30	July	2019.0	2169-3536	<a href="https://ieeexplore.ieee.org/abstract/document/...">https://ieeexplore.ieee.org/abstract/document/...</a>	2
2	Shiva Azimi, Brejesh Lal, Tapan K. Gandhi	Performance Evaluation of 3D Keypoint Detectors...	IEEE Conference Publication	6.0	10.23919/MVA.2019.8758002	27-31	July	2019.0	1882-0927.0	<a href="https://ieeexplore.ieee.org/abstract/document/...">https://ieeexplore.ieee.org/abstract/document/...</a>	2
3	S.A Syed Mustafa, I. Musirin, M.M. Othman, M....	Location and Sizing of Distributed Generation ...	IEEE Conference Publication	42.0	10.1109/AUPEC.2018.8758003	27-30	July	2019.0	2169-3536	<a href="https://ieeexplore.ieee.org/abstract/document/...">https://ieeexplore.ieee.org/abstract/document/...</a>	2

colab.research.google.com/drive/1IcJ1QWMNVJbBzZW6-2EMfz3RuMU2dZ0?usp=sharing#scrollTo=g63Sdp34gju\_

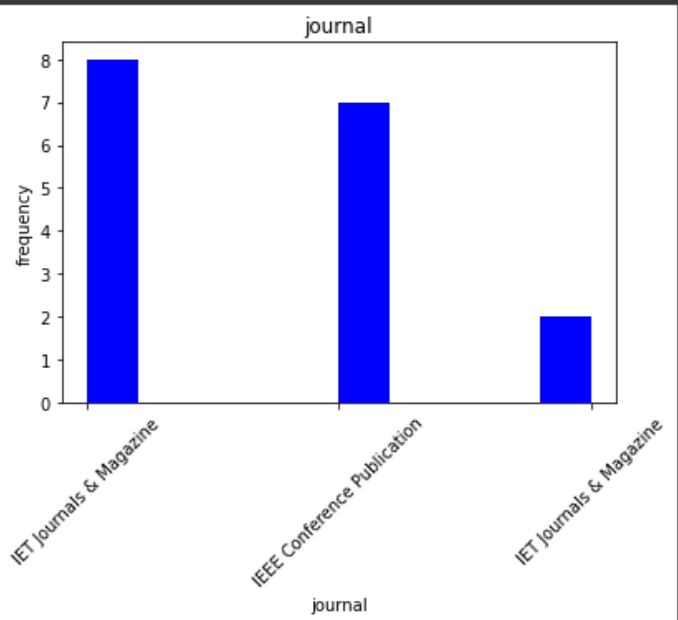
### Preprocessing.ipynb

File Edit View Insert Runtime Tools Help Last saved at 09:18

+ Code + Text

```
index=np.arange(len(journal))
plt.xticks(index,journal,rotation=45)
plt.show()
```


3



journal	frequency
IET Journals & Magazine	8
IEEE Conference Publication	7
IET Journals & Magazine	2

```
[ ] df['journal name'].value_counts()[ :20].plot(kind='pie')
```

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f34ce8e5810>




output.xls

colab.research.google.com/drive/1IcJ1QWMNVJbBzZW6-2EMfz3RuMU2dZ0?usp=sharing#scrollTo=g63Sdp34gju\_

### Preprocessing.ipynb

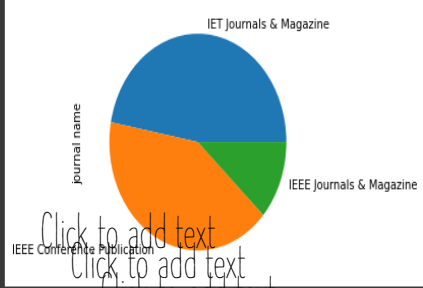
File Edit View Insert Runtime Tools Help Last saved at 09:18

+ Code + Text



```
[ ] df['journal name'].value_counts()[ :20].plot(kind='pie')
```

<matplotlib.axes.\_subplots.AxesSubplot at 0x7f34ce8e5810>



Click to add text

Click to add text

Click to add text

Click to add text

Click to add text

Click to add text

output.xls

colab.research.google.com/drive/1IcJ1QWMNVJbBzZW6-2EMfz3RuMU2dZ0?usp=sharing#scrollTo=g63Sdp34gju\_

### Preprocessing.ipynb

File Edit View Insert Runtime Tools Help Last saved at 09:18

+ Code + Text

For Missing Values

```
df.isnull().sum()
```

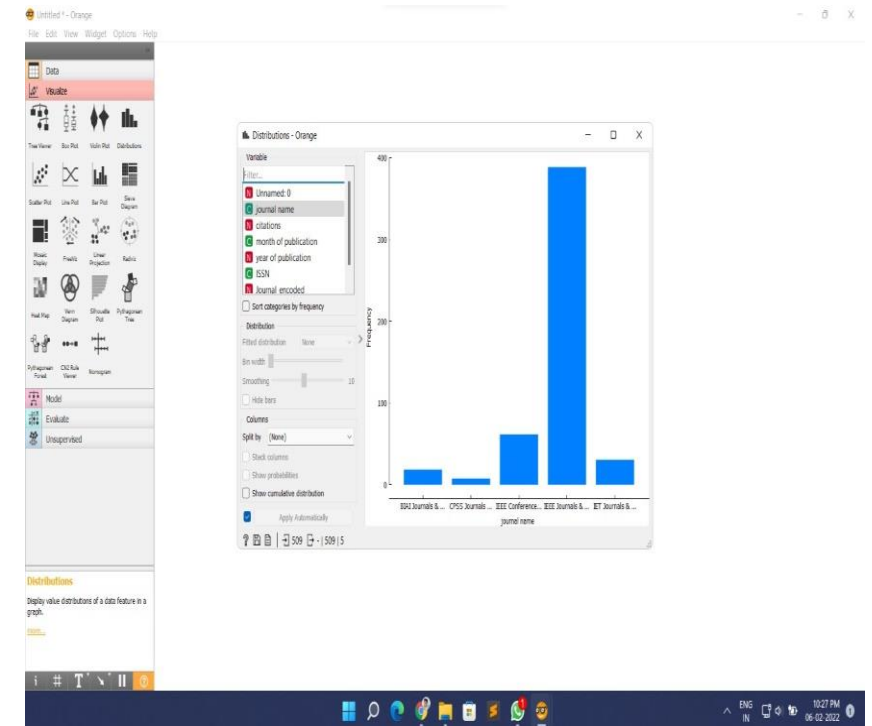
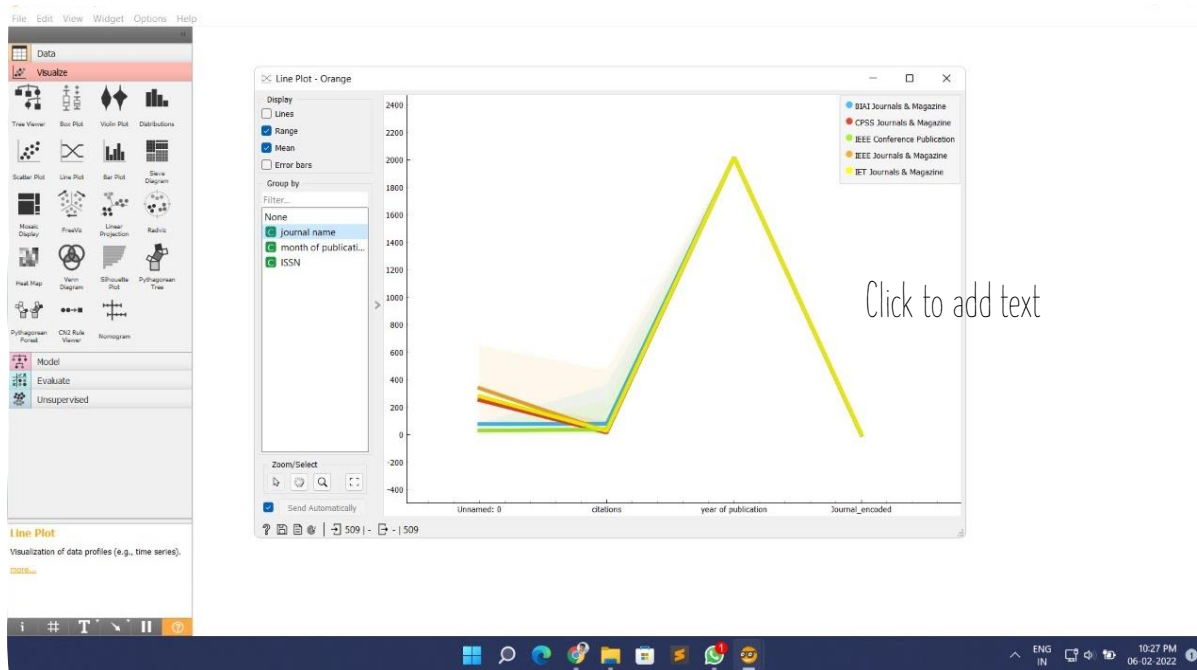
ors	0
e of paper	0
nal name	0
tions	0
	0
no	0
h of publication	0
of publication	0
	0
	0
me	17
e: int64	

```
df.drop('volume',axis=1) # dropping column volume
```

oding on Journal Name

```
sklearn import preprocessing
l_encoder = preprocessing.LabelEncoder()
df['Journal_encoded']= l_encoder.fit_transform(df['journal name'])
```

output.xls

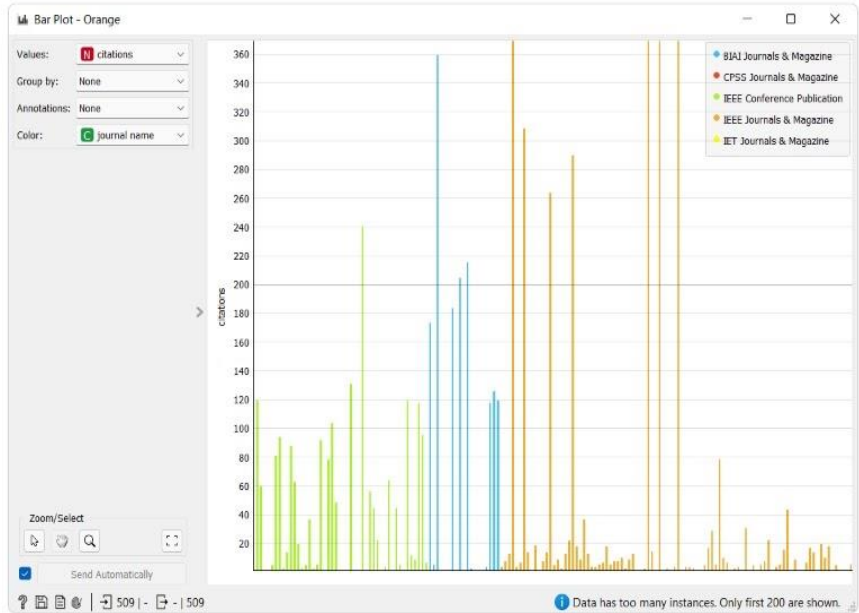


**Data**  
**Visualize**

Tree Viewer Box Plot Violin Plot Distributions  
Scatter Plot Line Plot Bar Plot Sieve Diagram  
Mosaic Display FreeViz Linear Projection Radviz  
Heat Map Venn Diagram Silhouette Plot Pythagorean Tree  
Pythagorean Forest Chi2 Rule Viewer Nomogram

**Model**  
**Evaluate**  
**Unsupervised**

**Bar Plot**  
Visualizes comparisons among categorical variables.  
[More...](#)

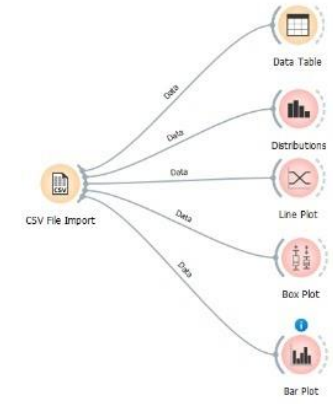


**Data**  
**Visualize**

Tree Viewer Box Plot Violin Plot Distributions  
Scatter Plot Line Plot Bar Plot Sieve Diagram  
Mosaic Display FreeViz Linear Projection Radviz  
Heat Map Venn Diagram Silhouette Plot Pythagorean Tree  
Pythagorean Forest Chi2 Rule Viewer Nomogram

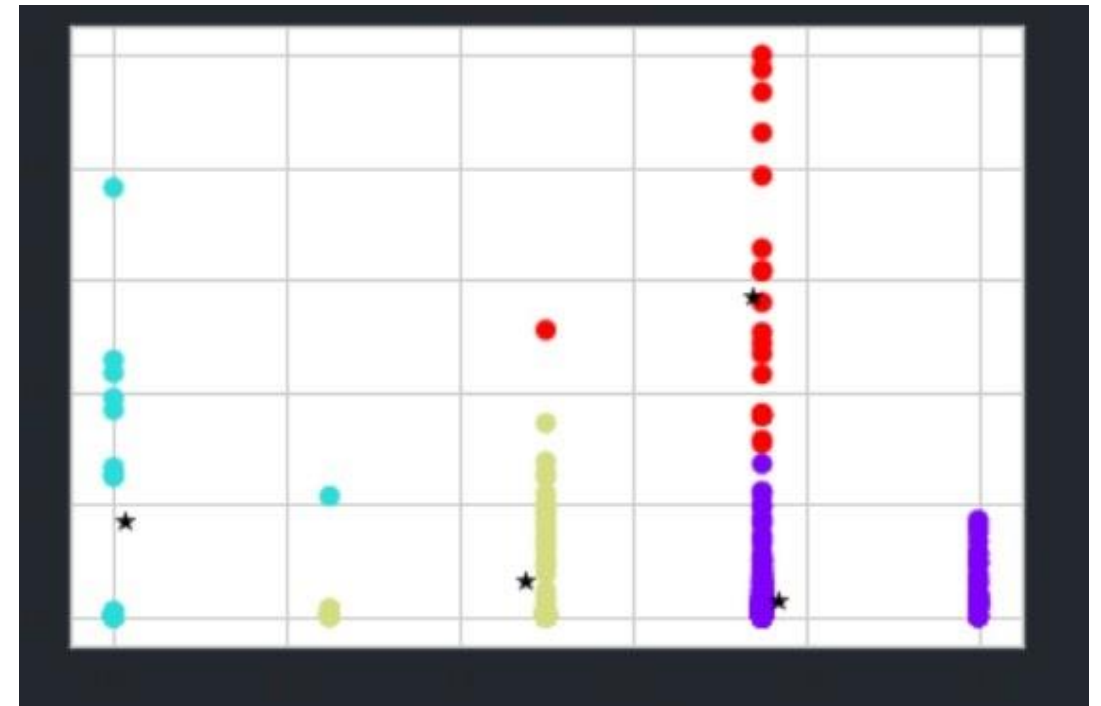
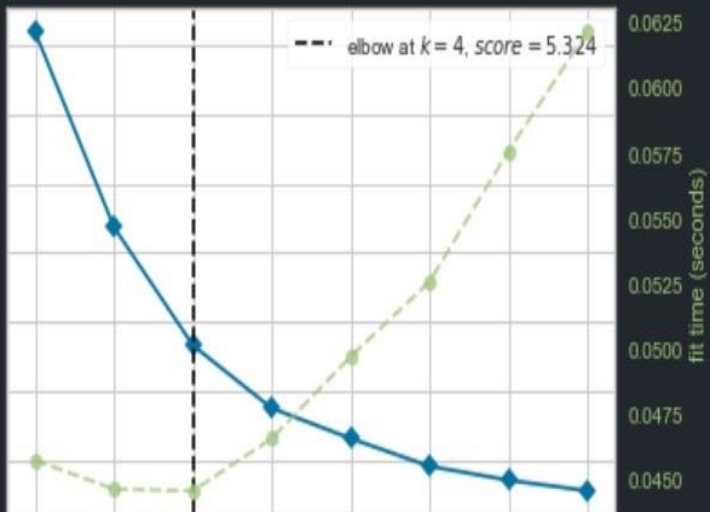
**Model**  
**Evaluate**  
**Unsupervised**

Select a widget to show its description.  
See [workflow examples](#), [YouTube tutorials](#), or per the [welcome screen](#).



```
model = model = KMeans(random_state=40)
visualizer = KElbowVisualizer(model, k=(2,10))
visualizer.fit(df[['Journal_encoded', 'citations']])
visualizer.show()
```

Click to add text  
Click to add text



# Conclusion

---

- According to the requirements of the project this system automatically extracted data in raw format using web scraping (selenium) and preprocessed it using different preprocessing techniques. After that for better understanding of data we used visualization from different tools like orange etc.
- After understanding and getting data in correct format we applied a machine learning model like K means for clustering the given papers according to citations.
- So thus automated data extraction and getting data in correct format for making inference of it is implemented.

# Future Scope

---

- 
- This system has done web scraping to extract data from IEEE websites and transform raw data to correct format by using preprocessing and for further understanding of data it uses an ML model like kmeans.
- But for the future we can make a system that can extract data from not only websites but also from given information from that research paper which will require NLP text mining.
-

# References

---

- <https://ieeexplore.ieee.org/Xplore/home.jsp>
- <https://ieeexplore.ieee.org/document/8822022>
- <https://ieeexplore.ieee.org/document/9274270>
- <https://github.com/siddheshshivdikar/college-placement-scraping>
- <https://ieeexplore.ieee.org/browse/conferences/title>