ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

An Analysis of Data Science and its Applications

Mariadas Ronnie C.P
Asst. Professor,
Dept. of Computer Applications,
SCMS School of Technology and Management(SSTM),

Abstract — Data Science is a combination of multiple disciplines such as statistics, data analysis and machine learning that are used to perform data analysis and to extract knowledge from it. It is used to find patterns of data through data analysis and thereby make decisions. Through data science, organizations are able to make better decisions, predictive analysis and pattern discoveries. In this paper, some of the data science applications created through pycharm software are explained of about how each application would read data from csv, excel, json and mongodb to produce patterns of data and an explanation of several applications of data science has been provided.

Keywords—python, pandas, mongodb, json, pycharm

I. INTRODUCTION

Data Science is used in various organizations of the world today such as banking, consultancy, healthcare, and manufacturing. Data Science is needed for route planning so as to discover the best routes to ship, to foresee delays for flight/ship/train etc. by means of predictive analysis, to create promotional offers, to find the best suited time to deliver goods, to forecast the next years revenue for a company, to analyse health benefit of training, to predict who will win elections etc. Data Science is applied in every part of a business where data is available. Examples are Consumer goods, Stock markets, Industry, Politics, Logistic companies, E-commerce. Data Scientist jobs are one of the most demanding jobs in the present era. A Data Scientist requires expertise in several backgrounds such as Machine Learning, Statistics, Programming (Python or R), Mathematics, Databases. A Data Scientist must find patterns within the data. Before patterns are found, data must be organized in a standard format. The data scientist should be able to understand the business problem, explore and collect data from database, web logs, customer feedback, etc., extract the data and transform the data to a standardized format, clean the data by removing erroneous values from the data, find and replace missing values such as checking for missing values and replace them with a suitable value, normalize data by scaling the values in a practical range, analyse data, find patterns and make future predictions, represent the result by presenting the result with useful insights in a way the organization can understand.

This paper is organized as follows. First, how the various data science applications created through pycharm software read data from csv, excel, json and mongodb to produce patterns of data are explained. An introduction of how to install pycharm and what all libraries are needed (pandas and pymongo) are also explained. Later on, what are the application areas of data science are also explained.

II. ANALYSIS

Creating a data science application using pandas library in python

The pandas library

Pandas is the most popular python library used for data analysis. It provides highly optimized performance with back-end source code purely written in C or Python. It allows importing data from various file formats such as csv, JSON, NoSQL and relational databases, Microsoft Excel etc. Also, it allows various data manipulation operations such as selecting, merging, reshaping as well as features such as data cleaning and data wrangling. Pandas is built on the top of two Python libraries—matplotlib for data visualization and NumPy for mathematical operations. Pandas acts as a wrapper over these libraries that allows you to access many of matplotlib's and NumPy's methods with lesser amount of code.

The analyzation of data in pandas is done with series and Data frames data structures. These data structures are built on the top of the NumPy array which makes them fast and efficient. Series is a 1D array defined in pandas which can store any data type whereas Data frames is a 2D array defined in pandas consisting of rows and columns. Data frames are widely used and is the most important data structure.

Series

It is a 1D data structure based on homogeneous data.

10	36	72	84	97

The key points of Series are that it is homogeneous data, their size is immutable and their data values are mutable.

Data frames

It is a 2D data structure based on heterogeneous data.

Name	Address	Rating
Sarva	Kochi	5
Pandaris	Kochi	4
Ifthar	Kochi	4

The key points of Data frames are that it is heterogeneous data, their size is mutable and their data values are mutable.

How to install pandas library in pycharm python IDE

If you are using pycharm python IDE, follow the steps below to install *pandas* library;

- From your pycharm IDE Menu, open File > Settings.
- From the dialog box obtained, select your data science project that you have developed (**Project:dsproj**). *dsproj* is my data science project created in pycharm.
- When you select your project in python, it will display with two options such as Python Interpreter and Project Structure.
- Click on the option Python Interpreter.
- From the options displayed, click the small **+** symbol appearing at the top to add a new library to the project.
- We need the *pandas* library to be installed, so type pandas.
- When you type pandas, all the supporting libraries of *pandas* will be displayed. Now, click the button *Install Package*.
- After the successive installation, you will obtain the message such as *pandas* libraries have been successfully installed. Now terminate and close all the windows.

Python data processing

After installing *pandas* library in pycharm python editor, we could perform python data processing on csv, Excel and JSON. The data processing is done as follows;

Reading from csv file

In Data Science, reading data from CSV is a fundamental aspect. The data from various sources are exported to CSV format so that they could be used by other systems. The Pandas library used with python provides features using which we can read the CSV file completely as well as in sub sections, that is, for selected group of rows and columns.

If you have a csv file named ***dsdata.csv*** as follows;

ID,Name of Employee,Salary,Joining date,Department
1,Mariadas Ronnie C P,35000,2016-06-13,CA
2,Ranjith S,35000,2019-09-23,CA
3,Sudha D,36000,2015-11-15,CA
4,Anjana S Chandran,45000,2009-05-11,CA
5,Praveen Kamath,65000,2007-03-27,CA
6,Jisha Liju Daniel,36000,2016-06-1,CA
7,Lakshmi Mahesh,45000,2007-07-30,CA
8,Jismy Joseph,36000,2015-09-17,CA
9,Shoby Sunny,36000,2016-06-01,CA
10,Vidya Gopinath,35000,2019-07-17,CA
11,Remya Raveendran,35000,2019-06-17,CA
12,Bindu John,36000,2015-06-17,CA
13,Anitha G Krishnan,36000,2017-04-17,CA
14,Arun Krishnan,36000,2017-05-17,BCom
15,Sujith,36000,2015-06-17,BCom
16,Arsha P B,36000,2014-06-17,BCom
17,Satheesh,36000,2015-07-17,BA
18,Rose Paul,36000,2016-06-01,BA
19,Binu V S,40000,2014-06-17,MBA
20,Vignesh Karthik,40000,2014-09-17,MBA

To run in pycharm, we have to save the csv file in your data science project folder. For example, in your ***PycharmProjects*** folder, where you are saving your python projects, save this file as ***dsdataex.csv*** . After this, we need to design a python code as follows;

```
import pandas as pd
data = pd.read_csv('dsdataex.csv')

#Reading a csv file

print (data)
print()

#Reading specific rows

print (data[0:5]['Salary'])
print()

#Reading specific columns

print (data.loc[:,['Salary','Name of Employee']])
print()

#Reading specific columns and rows

print (data.loc[[1,3,5],['Salary','Name of Employee']])
print()

#Reading specific columns for a range of rows

print (data.loc[2:6,['Salary','Name of Employee']])
```

The first statement imports the pandas library features into pd.

The read_csv function is used read the content of the CSV file as a pandas DataFrame into the python environment.

The python code uses the library pandas in order to execute the csv file in order to produce the outputs as follows;

#Output for Reading a csv file

	ID	Name of Employee	Salary	Joining date	Department
0	1	Mariadas Ronnie C P	35000	2016-06-13	CA
1	2	Ranjith S	35000	2019-09-23	CA
2	3	Sudha D	36000	2015-11-15	CA
3	4	Anjana S Chandran	45000	2009-05-11	CA
4	5	Praveen Kamath	65000	2007-03-27	CA
5	6	Jisha Liju Daniel	36000	2016-06-1	CA
6	7	Lakshmi Mahesh	45000	2007-07-30	CA
7	8	Jismy Joseph	36000	2015-09-17	CA
8	9	Shoby Sunny	36000	2016-06-01	CA
9	10	Vidya Gopinath	35000	2019-07-17	CA
10	11	Remya Raveendran	35000	2019-06-17	CA
11	12	Bindu John	36000	2015-06-17	CA
12	13	Anitha G Krishnan	36000	2017-04-17	CA
13	14	Arun Krishnan	36000	2017-05-17	BCom
14	15	Sujith	36000	2015-06-17	BCom
15	16	Arsha P B	36000	2014-06-17	BCom
16	17	Satheesh	36000	2015-07-17	BA

17 18 Rose Paul 36000 2016-06-01 BA
18 19 Binu V S 40000 2014-06-17 MBA
19 20 Vignesh Karthik 40000 2014-09-17 MBA

#Output for Reading specific rows

0 35000
1 35000
2 36000
3 45000
4 65000

#Output for Reading specific rows and columns

Salary Name of Employee
0 35000 Mariadas Ronnie C P
1 35000 Ranjith S
2 36000 Sudha D
3 45000 Anjana S Chandran
4 65000 Praveen Kamath
5 36000 Jisha Liju Daniel
6 45000 Lakshmi Mahesh
7 36000 Jismy Joseph
8 36000 Shoby Sunny
9 35000 Vidya Gopinath
10 35000 Remya Raveendran
11 36000 Bindu John
12 36000 Anitha G Krishnan
13 36000 Arun Krishnan
14 36000 Sujith
15 36000 Arsha P B
16 36000 Satheesh
17 36000 Rose Paul
18 40000 Binu V S
19 40000 Vignesh Karthik

#Output for Reading specific columns and rows

Salary Name of Employee
1 35000 Ranjith S
3 45000 Anjana S Chandran
5 36000 Jisha Liju Daniel

#Output for Reading specific columns for a range of rows

Salary Name of Employee
2 36000 Sudha D
3 45000 Anjana S Chandran
4 65000 Praveen Kamath
5 36000 Jisha Liju Daniel
6 45000 Lakshmi Mahesh

Reading from excel file

MS Excel is a widely used spread sheet program. Today, it is a widely used tool in Data Science. The Pandas library used with python provides features using which we can read the excel file completely as well as in sub sections, that is, for selected group of rows and columns. We could also read excel file with multiple sheets in it. To read excel file, the function read_excel is used.

If you have created a new project to design a data science program to read the contents of an excel file using pycharm, you have to install packages such as *pandas* and *openpyxl*.

If you have an excel file named ***inputdata.xlsx*** as follows;

#Data in Sheet 1

ID	Name of Employee	Salary	Joining date	Department
1,	Mariadas Ronnie C P,	35000,	13-06-2016,	CA
2,	Ranjith S,	35000,	23-09-2019,	CA
3,	Sudha D,	36000,	15-11-2015,	CA
4,	Anjana S Chandran,	45000,	11-05-2009,	CA
5,	Praveen Kamath,	65000,	27-03-2007,	CA
6,	Jisha Liju Daniel,	36000,	01-06-2016,	CA
7,	Lakshmi Mahesh,	45000,	30-07-2007,	CA
8,	Jismy Joseph,	36000,	17-09-2015,	CA
9,	Shoby Sunny,	36000,	01-06-2016,	CA
10,	Vidya Gopinath,	35000,	17-07-2019,	CA
11,	Remya Raveendran,	35000,	17-06-2019,	CA
12,	Bindu John,	36000,	17-06-2015,	CA
13,	Anitha G Krishnan,	36000,	17-04-2017,	CA
14,	Arun Krishnan,	36000,	17-05-2017,	BCom
15,	Sujith,	36000,	17-06-2015,	BCom
16,	Arsha P B,	36000,	17-06-2014,	BCom
17,	Satheesh,	36000,	17-07-2015,	BA
18,	Rose Paul,	36000,	01-06-2016,	BA
19,	Binu V S,	40000,	17-06-2014,	MBA
20,	Vignesh Karthik,	40000,	17-09-2014,	MBA

#Data in Sheet 2

ID	Name of Employee	Zipcode
1	Mariadas Ronnie C P	682002
2	Ranjith S	686796
3	Sudha D	694778
4	Anjana S Chandran	677277

5	Praveen Kamath	698987
6	Jisha Liju Daniel	665100
7	Lakshmi Mahesh	656743
8	Jismy Joseph	687457
9	Shoby Sunny	689799
10	Vidya Gopinath	672001
11	Remya Raveendran	683567
12	Bindu John	687567
13	Anitha G Krishnan	694534
14	Arun Krishnan	672008
15	Sujith	682015
16	Arsha P B	689142
17	Satheesh	702132
18	Rose Paul	702193
19	Binu V S	718348
20	Vignesh Karthik	729494

To run in pycharm, we have to save the excel file in your data science project folder. For example, in your ***PycharmProjects*** folder, where you are saving your python projects, save this file as ***inputdata.xlsx*** . After this, we need to design a python code as follows;

```
import pandas as pd
data = pd.read_excel('inputdata.xlsx')


#Reading an excel file

print (data)

#Reading specific rows and columns

print (data.loc[[1,3,5],['Salary','Name of Employee']])


#Reading multiple sheets

with pd.ExcelFile('inputdata.xlsx') as xls:
    data1 = pd.read_excel(xls, 'Sheet1')
    data2 = pd.read_excel(xls, 'Sheet2')

print("****Output Sheet 1****")
print (data1[0:5]['Name of Employee'])
print("")
print("***Output Sheet 2****")
print (data2[0:5]['Zipcode'])
```

The first statement imports the pandas library features into pd.

The read_excel function is used read the content of the excel file as a pandas DataFrame into the python environment. By default, the data is read from Sheet 1.

We can read specific rows and columns from the excel sheet as the same way as reading specific rows and columns from a csv file. Also, we can read data from multiple sheets in an excel file using the logic given in the program above. The *ExcelFile* function parses the file *inputdata.xlsx* into a Data frame.

The python code uses the library *pandas* and *openpyxl* in order to execute the excel file in order to produce the outputs as follows;

#Output for Reading an excel file

```
  ID,  Name of Employee, Salary, Joining date, Department
0  1, Mariadas Ronnie C P,35000,   13-06-2016,    CA
1  2, Ranjith S,          35000,   23-09-2019,    CA
2  3, Sudha D,            36000,   15-11-2015,    CA
3  4, Anjana S Chandran,  45000,   11-05-2009,    CA
4  5, Praveen Kamath,     65000,   27-03-2007,    CA
5  6, Jisha Liju Daniel,  36000,   01-06-2016,    CA
6  7, Lakshmi Mahesh,     45000,   30-07-2007,    CA
7  8, Jismy Joseph,       36000,   17-09-2015,    CA
8  9, Shoby Sunny,        36000,   01-06-2016,    CA
9  10,Vidya Gopinath,     35000,   17-07-2019,    CA
10 11,Remya Raveendran,   35000,   17-06-2019,    CA
11 12,Bindu John,         36000,   17-06-2015,    CA
12 13,Anitha G Krishnan,  36000,   17-04-2017,    CA
13 14,Arun Krishnan,      36000,   17-05-2017,    BCom
14 15,Sujith,             36000,   17-06-2015,    BCom
15 16,Arsha P B,          36000,   17-06-2014,    BCom
16 17,Satheesh,           36000,   17-07-2015,    BA
17 18,Rose Paul,          36000,   01-06-2016,    BA
18 19,Binu V S,           40000,   17-06-2014,    MBA
19 20,Vignesh Karthik,    40000,   17-09-2014,    MBA
```

#Output for Reading specific rows and columns in an excel file

```
    Salary    Name of Employee
1  35000,   Ranjith S,
3  45000,   Anjana S Chandran,
5  36000,   Jisha Liju Daniel,
```

#Output for Reading multiple sheets in an excel file

```
****Output Sheet 1****
0   Mariadas Ronnie C P,
1   Ranjith S,
2   Sudha D,
3   Anjana S Chandran,
4   Praveen Kamath,

****Output Sheet 2****
0   682002
1   686796
2   694778
3   677277
4   698987
```

Reading from JSON file

JSON stands for Javascript object notation. It is used for storing and transporting data, such as, sending data from server to a web page, sending data between servers, web applications and web connected devices. Its format is text-only, so that the data could be easily transmitted between computers and could be used by any programming languages.

When there is a large dataset of JSON, working with it to extract data could be very difficult if they are large enough to fit in memory. In these situations, we need to use python and pandas library which contains supporting functions to analyse and explore the JSON data. To read JSON file using python, the pandas library function read_json is used.

If you have created a new project to design a data science program to read the contents of a JSON file using pycharm, you have to install packages such as *pandas in your project*.

If you have a JSON file named ***input.json*** as follows;

```
{
  "ID":["1","2","3","4","5","6","7","8" ],
  "Name":["Mariadas Ronnie C P","Ranjith S","Sudha D","Anjana S Chandran","Praveen Kamath","Sujith","Satheesh","Binu V S" ],

"Salary":["35000","35000","36000","45000","65000","36000","36000","40000" ],
  "Joining Date":[ "13-06-2016","23-09-2019","15-11-2015","11-05-2009","27-03-2007","17-06-2015", "17-07-2015","17-06-2014"],
  "Dept":[
"CA","CA","CA","CA","CA","BCom","BA","MBA"]
}
```

To run in pycharm, we have to save the JSON file in your data science project folder. For example, in your ***PycharmProjects*** folder, where you are saving your python projects, save this file as ***input.json***. After this, we need to design a python code as follows;

```
import pandas as pd
data = pd.read_json("input.json")

# Reading a JSON file
print(data)

# Reading specific rows and columns
print (data.loc[[1,3,5],['Salary','Name']])

# Reading JSON file as records
print(data.to_json(orient='records', lines=True))
```

The read_json function is used read the content of the JSONfile as a pandas DataFrame into the python environment.

We can read specific rows and columns from the JSON file as the same way as reading specific rows and columns from an excel and a csv file. The *to_json* function is used to display JSON file contents into individual records.

The python code uses the library *pandas* in order to execute the JSON file in order to produce the outputs as follows;

 # Output for Reading a JSON file

	ID	Name	Salary	Joining Date	Dept
0	1	Mariadas Ronnie C P	35000	13-06-2016	CA
1	2	Ranjith S	35000	23-09-2019	CA
2	3	Sudha D	36000	15-11-2015	CA
3	4	Anjana S Chandran	45000	11-05-2009	CA
4	5	Praveen Kamath	65000	27-03-2007	CA
5	6	Sujith	36000	17-06-2015	BCom
6	7	Satheesh	36000	17-07-2015	BA
7	8	Binu V S	40000	17-06-2014	MBA

 # Output for Reading specific rows and columns

	Salary	Name
1	35000	Ranjith S
3	45000	Anjana S Chandran
5	36000	Sujith

 # Output for Reading JSON file as records

{"ID":1,"Name":"Mariadas Ronnie C P","Salary":35000,"Joining Date":"13-06-2016","Dept":"CA"}
{"ID":2,"Name":"Ranjith S","Salary":35000,"Joining Date":"23-09-2019","Dept":"CA"}
{"ID":3,"Name":"Sudha D","Salary":36000,"Joining Date":"15-11-2015","Dept":"CA"}
{"ID":4,"Name":"Anjana S Chandran","Salary":45000,"Joining Date":"11-05-2009","Dept":"CA"}
{"ID":5,"Name":"Praveen Kamath","Salary":65000,"Joining Date":"27-03-2007","Dept":"CA"}
{"ID":6,"Name":"Sujith","Salary":36000,"Joining Date":"17-06-2015","Dept":"BCom"}
{"ID":7,"Name":"Satheesh","Salary":36000,"Joining Date":"17-07-2015","Dept":"BA"}
{"ID":8,"Name":"Binu V S","Salary":40000,"Joining Date":"17-06-2014","Dept":"MBA"}

Reading from Mongodb

Mongodb is an object oriented, simple, dynamic, NoSQL database. In Mongodb, data are stored as a collection instead of storing as rows and columns as in a traditional RDBMS. Mongodb database support dynamic database schema which makes them responsive to the change in the structure of data. That is why, it is so popular in the field of data science.

To work Mongodb with pycharm, you need to install mongodb driver package in pycharm which is *pymongo*. On your new project, select settings->your project name->python interpreter. From there, install the package

pymongo. After successful installation of *pymongo*, use the statement in the pycharm project code editor as follows to check whether the python package works successfully;

Import pymongo

If your project runs successfully without bugs, it means that your pymongo package is installed successfully.

A python code is designed as follows to demonstrate inserting and finding records in a Mongodb database;

```
import pymongo
myclient =
pymongo.MongoClient("mongodb://localhost:27017/")
#Creating database
mydb = myclient["employee"]
#Creating collection
mycol = mydb["faculty"]
#Creating records
mylist = [
{ "_id": 1,"name": "Mariadas Ronnie C P", "address":
"Mattancherry"},
{ "_id": 2,"name": "Ranjith S", "address":
"Chottanikkara"},
{ "_id": 3,"name": "Sudha D", "address": "Aluva"},
{ "_id": 4,"name": "Anjana S Chandran", "address":
"Palarivattom"},
{ "_id": 5,"name": "Praveen Kamath", "address":
"Thammanam"},
{ "_id": 6,"name": "Jisha Liju Daniel", "address":
"Kakkanad"},
{ "_id": 7,"name": "Lakshmi Mahesh", "address":
"Ernakulam"},
{ "_id": 8,"name": "Jismy Joseph", "address":
"Perumbavoor"},
{ "_id": 9,"name": "Shoby Sunny", "address":
"Kakkanad"},
{ "_id": 10,"name": "Vidya Gopinath", "address":
"Ernakulam"},
{ "_id": 11,"name": "Remya Raveendran", "address":
"Kaladi"},
{ "_id": 12,"name": "Bindu John", "address": "Mala"},
{ "_id": 13,"name": "Anitha G Krishnan", "address":
"Ernakulam"},
{ "_id": 14,"name": "Arun Krishnan", "address":
"Ernakulam"},
{ "_id": 15,"name": "Sujith", "address": "Edakochi"},
{ "_id": 16,"name": "Arsha P B", "address": "Ernakulam"},
{ "_id": 17,"name": "Satheesh", "address": "Pambakuda"},
{ "_id": 18,"name": "Rose Paul", "address": "Ernakulam"},
{ "_id": 19,"name": "Binu V S", "address": "Trivandrum"},
{ "_id": 20,"name": "Vignesh Karthik", "address":
"Tirunelveli"}
]
#Insert records
x = mycol.insert_many(mylist)
#Find specific records
res = mycol.find({"address": "Ernakulam"})
for x in res:
print(x)
#Updating records
oldvalue = { "address": "Mattancherry" }
newvalue = { "$set": { "address": "Fort Kochi" } }
mycol.update_one(oldvalue, newvalue)
#print "faculty" after the update:
for x in mycol.find():
print(x)
#Deleting records
deldata = { "address": "Tirunelveli" }
mycol.delete_one(deldata)
#print "faculty" after the delete:
for x in mycol.find():
print(x)
```

```
#Output for inserting and finding records
{'_id': 7, 'name': 'Lakshmi Mahesh', 'address': 'Ernakulam'}
{'_id': 10, 'name': 'Vidya Gopinath', 'address': 'Ernakulam'}
{'_id': 13, 'name': 'Anitha G Krishnan', 'address':
'Ernakulam'}
{'_id': 14, 'name': 'Arun Krishnan', 'address': 'Ernakulam'}
{'_id': 16, 'name': 'Arsha P B', 'address': 'Ernakulam'}
{'_id': 18, 'name': 'Rose Paul', 'address': 'Ernakulam'}

#Output for updating records
{'_id': 1, 'name': 'Mariadas Ronnie C P', 'address': 'Fort
Kochi'}
{'_id': 2, 'name': 'Ranjith S', 'address': 'Chottanikkara'}
{'_id': 3, 'name': 'Sudha D', 'address': 'Aluva'}
{'_id': 4, 'name': 'Anjana S Chandran', 'address':
'Palarivattom'}
{'_id': 5, 'name': 'Praveen Kamath', 'address': 'Thammanam'}
{'_id': 6, 'name': 'Jisha Liju Daniel', 'address': 'Kakkanad'}
{'_id': 7, 'name': 'Lakshmi Mahesh', 'address': 'Ernakulam'}
{'_id': 8, 'name': 'Jismy Joseph', 'address': 'Perumbavoor'}
{'_id': 9, 'name': 'Shoby Sunny', 'address': 'Kakkanad'}
{'_id': 10, 'name': 'Vidya Gopinath', 'address': 'Ernakulam'}
{'_id': 11, 'name': 'Remya Raveendran', 'address': 'Kaladi'}
{'_id': 12, 'name': 'Bindu John', 'address': 'Mala'}
{'_id': 13, 'name': 'Anitha G Krishnan', 'address':
'Ernakulam'}
{'_id': 14, 'name': 'Arun Krishnan', 'address': 'Ernakulam'}
{'_id': 15, 'name': 'Sujith', 'address': 'Edakochi'}
{'_id': 16, 'name': 'Arsha P B', 'address': 'Ernakulam'}
{'_id': 17, 'name': 'Satheesh', 'address': 'Pambakuda'}
{'_id': 18, 'name': 'Rose Paul', 'address': 'Ernakulam'}
{'_id': 19, 'name': 'Binu V S', 'address': 'Trivandrum'}
{'_id': 20, 'name': 'Vignesh Karthik', 'address': 'Tirunelveli'}

#Output for deleting records
{'_id': 1, 'name': 'Mariadas Ronnie C P', 'address': 'Fort
Kochi'}
{'_id': 2, 'name': 'Ranjith S', 'address': 'Chottanikkara'}
{'_id': 3, 'name': 'Sudha D', 'address': 'Aluva'}
{'_id': 4, 'name': 'Anjana S Chandran', 'address':
'Palarivattom'}
{'_id': 5, 'name': 'Praveen Kamath', 'address': 'Thammanam'}
{'_id': 6, 'name': 'Jisha Liju Daniel', 'address': 'Kakkanad'}
{'_id': 7, 'name': 'Lakshmi Mahesh', 'address': 'Ernakulam'}
{'_id': 8, 'name': 'Jismy Joseph', 'address': 'Perumbavoor'}
{'_id': 9, 'name': 'Shoby Sunny', 'address': 'Kakkanad'}
{'_id': 10, 'name': 'Vidya Gopinath', 'address': 'Ernakulam'}
{'_id': 11, 'name': 'Remya Raveendran', 'address': 'Kaladi'}
{'_id': 12, 'name': 'Bindu John', 'address': 'Mala'}
```

{'_id': 13, 'name': 'Anitha G Krishnan', 'address': 'Ernakulam'}
{'_id': 14, 'name': 'Arun Krishnan', 'address': 'Ernakulam'}
{'_id': 15, 'name': 'Sujith', 'address': 'Edakochi'}
{'_id': 16, 'name': 'Arsha P B', 'address': 'Ernakulam'}
{'_id': 17, 'name': 'Satheesh', 'address': 'Pambakuda'}
{'_id': 18, 'name': 'Rose Paul', 'address': 'Ernakulam'}
{'_id': 19, 'name': 'Binu V S', 'address': 'Trivandrum'}

Applications of Data science

1. Fraud and risk detection

We are witnessing the larger amount in the quantity of data, both at global and economic level entities where the data can be audio, video, text, pictures etc. Thus, we are in a need to implement the data analytics software to prevent and detect fraud. Data are available from various sources like call centers, social media, phone, email, fax etc. We need to use these data for making strategy decisions, detect and prevent fraud etc. The usage of data analysis software provides in-depth analysis of the phenomena of fraud and corruption since information and communication technology has become an instrument of the economy.

There are a variety of tools available in the market for supporting anti-fraud activities. An example is Forensic Data Analytics (FDA) tools. An example of these tools are Microsoft Excel, M S Access and SQL Server. Even though they are important FDA tools, they are focused to be used primarily on matching, grouping, ordering, joining or filtering data. To improve the data analytics process for preventing and detecting fraud data is to provide an in-depth examination of the meaning and features of the data using some specific methods and techniques. These techniques make careful examination of the data and identifies its strength, weakness, dysfunction, vulnerability and other risk factors which may constitute threats and finally suggest guidelines for removing threats.

There are different types of data analysis. The difference depends on the nature of data, practical usability and applicability, scope etc. The two classical types of analysis are namely *operational analysis* and *strategic analysis [1]*. Operational analysis is used to exploit data and present information to comply with currently running activities in order to detect fraud thereby providing maximum efficiency. The role of operational analysis is to detect and defend illegal activities such as examination of links and their characteristics, the movement of money and other variables, communication through email, social networking etc. The strategic analysis on the other hand, offers a macro view of fraud. Digital statistical tools are used to explore data and its variations. It has a powerful interface and a statistical processing engine. In the early days, statistics and exploration of data were developed independently of the visual techniques. With the usage of new generation software, we could access large data sets and could find patterns and threats within a short span of time which would take hours or days in conventional data mining process.

Visual analysis is an important technique to understand the location in which the fraud has happened and discover the patterns based on fraud behaviors. Analytical tools are used to identify, explore, index and process data. The challenge is about how to implement automated methods and tools in order to detect and prevent fraud and to make decisions. The system should be capable in such a way that it should be able to retrieve data from different sources of different formats. Even though the data is in different formats, they should be still treated using the same methods so that the database creation would be homogeneous. The actual mode of fraud detection is the combination of human and technical support. Human intervention brings the exploitation of results no matter how the technologies have been improved.

The integration of data analytics with the fraud detection system brings out certain benefits and limitations. The benefits are such that answers would be obtained in real time to a series of questions regarding fraud issues. The other benefits are faster access to data, elimination of duplicate records, high productivity, increased rate of fraud detection, fast detection and recovery from fraud activity, statistical analysis with greater accuracy, reducing claims related to fraudulency, improving the analytical product quality. The limitation is such that the analytical tools used are not cheap. The other limitations are, not all the data are available in databases and the usage of analytical tools does not save your time.

2. Healthcare

Medical imaging is one of the applications in image processing. The scanned images of patients are used in data science to find the patients defects and to decide which treatment should be suggested for patients in order to recover from the disease. The usage of deep learning algorithms [2] helps to determine the difference in resolution, dimension between medical images obtained from x-ray etc. The proper analyzation of these images helps doctors to make proper decision on which treatment could be given for patients, improving medical accuracy, accurately detect various diseases etc.

An important application of data science in health care is drug discovery. To find a new drug, it requires a number of procedures and tests and it would take lot of time and resources. By the usage of data analysis, it would be able to perform millions of calculations within a minimum amount of time and resources. By analyzing data and by using various algorithms, it is now possible to detect the effect of the drug in human body and its probability of success. Data such as the patient's response to the drug and its side effects are analyzed and is used to improve the efficiency of the drug. Due to these advantages, a drug could be released within a minimum span of time.

Predictive analytics is used in healthcare to develop plans to decide the most effective treatment for the patient. By using predictive analytical tools, doctors could diagnose disease in a patient at an earlier level. So, it helps the doctors to suggest patients about the various prevention measures that

is needed to be taken if diseases are identified at an earlier stage.

Another application of data science in healthcare is to provide virtual assistance. There are certain applications in data science that would predict the type of disease depending on the inputs provided by the patients as symptoms of the disease. Other type of applications would notify the patients on a daily basis about the medical intake they should follow for avoiding the skipping of medicinal intakes.

3. IoT

Data are generated as huge amount from IoT devices such as sensors, satellites, RFID's, actuators, consumer appliances, social media etc. The data generated from sensors and objects are transferred through networks and are stored in the cloud for big data processing. Applications that support data science are used by scientists to analyze structured and unstructured data generated from IoT devices. These applications extract information to identify trends, patterns and make effective decisions. The IoT data is mostly collected from sensors. So, regarding external influences such as noise, heterogeneity, these data will be different from normal big data.

Smart devices that use data and connectivity are integrated with each other and the IoT integrates these collected data from the devices through intelligence applications. The integrated data produce a huge amount of data so that data science could play an important role in IoT for pattern recognition, decision making etc.

An example of business analytics technologies that are integrated with IoT devices are wearable health monitoring sensors. The health data collected through sensors report the change in the normal activities of its members so that healthcare professionals could collect the data and analyze, monitor their patients efficiently. Another example is by monitoring the environment conditions, the level of energy consumption and the performance of equipment. Thus, it requires IoT to collect the data and data science to extract useful information and thereby monitoring the performance and changes of the object. An example of IoT monitoring and control systems is Smart Home Technology [3] which is used to save energy and to protect family and property. In terms of protecting family and property, users can control the IoT system using their smartphone or laptop. They could adjust the lights, control locking and unlocking of doors and could also manage the security systems. Another application is the usage of IoT in smart cars where the various parts of the car could be controlled and monitored.

4. Targeted Advertising

Customer prospecting [4] is a challenging segment of the online display advertising market. It deals with the delivery of advertisements to customers who have no previous interactions with any particular brand, but likely to become customers of that particular brand when a proper advertisement has been shown. Real time bidding exchanges normally auction off web sites for placing online display ads on it through which a customer using a website gets interested in. Thus, through website auctions, advertisers could target customers through their ads. Auctions are done in real time when the customer navigates through the browser. At the time of the auction, the advertiser could decide whether to bid for the space on the website, how much bidding for the space should be done and what to put as ad, if the advertiser wins the auction. The ad would be put by the advertiser depending on the previously collected data depending upon the consumer and the website. Real time auction happens in websites in billions and advertisers would require large scale data to come to a decision within a short span of time. Thus, machine learning comes in optimizing ads because of the massive amount of data based on consumer behavior, data based on the actions of consumers related to brand behavior and based on the ability to make advertising decisions and delivering advertisements in real time.

5. Website recommendations

Recommendation systems are an important component in the tourism industry. The most important components are online booking and reservation systems. Recommendation systems are used by tourists in order to find the best place for stay when they tour a place. The system would provide tourists with a list of hotels out of which the users could choose some. The data that would be collected by the recommendation systems would get limited if it collects data from homogeneous systems. As the data collected could be residing in heterogeneous systems too, collecting data from them would lead to complexity. Thus, developers should deal with the heterogeneity of the data collected from different types of sources. The recommendation systems should be able to recommend a hotel or list of hotels to users based on numerical and textual data by means of machine learning to achieve accurate recommendations. It should be able to mine user reviews from other travelers by means of ranks/votes made by them and thus obtaining an accuracy in recommendation. An efficient big data solution is by means of using Hadoop since it deals with data heterogeneity.

Since sentimental analysis [5] is the process of extracting opinions from reviews, it is used by many online recommendation systems to make travelers rate hotels based on location, rooms, cleanliness, service, staff and safety. Along with sentimental analysis, machine learning is used to categorize the reviews based on positive, negative and neutral values. Afterwards, training dataset is used to train and validate the performance.

There are two types of datasets namely training dataset and test dataset. The classification factors are learned from the training dataset and the classification accuracy is evaluated based on the test dataset. Another approach is by using the cluster-based approach along with collaborative filtering thereby reducing the computation time. Since the available web data are enormous, the problem that should be tackled is how to manage this data and how to improve time efficiency and performance. The solution is by using Hadoop and NoSQL which efficiently handles the time

efficiency and performance when dealing with these huge amounts of data.

6. Advanced Image Recognition

Image processing plays a vital role in various organizations, industries, healthcare, defense etc. It means that an image will be analyzed as input and afterwards, image processing techniques will be applied to that image to obtain an output which can be a different form of the image or the part of an image. The resultant image produced by image processing techniques could be huge resulting the categorization of big data. The huge amount of information is stored as structured or unstructured data. Big data analytics for data mining on the data formed through image processing has a profound application in the field of education, research, govt. organizations, healthcare, defense, business etc. Visualization techniques [6] are used for message communication in the form of image, animation or video. There are different visualization techniques such as abstract visualization and scientific visualization. Images can be subject to noise. Many algorithms could be used for de-noising an image. To overcome resolution loss, image restoration techniques are used. For example, Adobe Photoshop as a software is used for this purpose. Image retrieval is the process of retrieving images from a huge database system. The different techniques of retrieving images are content based, document-based image retrieval. Algorithms are mainly used for feature extraction, smoothing, reconstruction and enhancing the quality of images. In future, image and big data may be combined into a hybrid system. Research could be done in detecting satellite images, detect health care problems such as tumor, problems related to diseases in fruits or vegetables which may of big help for farmers.

III. Conclusion

As companies need data for their data driven decision models, the companies will focus on specific areas such as marketing, customer acquisition, innovation to improve customer experience. Data Scientists are the backbone of the huge amount of data processing companies or organizations. Data scientists provide the essential data that the companies need. The purpose of each Data Scientist is to extract, preprocess and analyze data for the companies. Through this, the companies could be able to make better decisions. Various companies have their own requirements and use data accordingly. Thus, the goal of the Data Scientist is to make businesses grow better. With the decisions and insights provided, the companies would be able to adopt appropriate strategies and customize themselves for enhanced customer experience. Thus, this paper was providing an insight of about how the data science applications produce patterns of data from huge amount of csv, excel, json, mongodb data to produce patterns of data.

REFERENCES

[1] Adrian Bănărescua, Detecting and Preventing Fraud with Data Analytics, Procedia Economics and Finance 32 (2015) 1827 – 1836

[2] Krutika H. Churi, Dept. of I.T, Sonopant Dandekar College, Palghar, India, Data Science in Healthcare, SAMRIDDHI Volume 13, Special Issue 1, 2021

[3] Sarfraz Nawaz Brohi, Mohsen Marjani, Ibrahim Abaker Targio Hashem, Thulasyammal Ramiah Pillai, Sukhminder Kaur, and Sagaya Sabestinal Amalathas, A Data Science Methodology for Internet-of-Things, n International Conference for Emerging Technologies in Computing (iCETiC 2019)

[4] C. Perlich · B. Dalessandro · T. Raeder · O. Stitelman, F. Provost, Machine learning for targeted display advertising: transfer learning in action, Mach Learn (2014) 95:103–127

[5] Bushra Ramzan, Imran Sarwar Bajwa, Noreen Jamil, Riaz Ul Amin, Shabana Ramzan, Farhan Mirza, and Nadeem Sarwar, An Intelligent Data Analysis for Recommendation Systems Using Machine Learning, Hindawi Scientific Programming Volume 2019, Article ID 5941096

[6] Ezhilraman, S & Srinivasan, Sujatha. (2018). State of the art in image processing & big data analytics: Issues and challenges. International Journal of Engineering and Technology (UAE).