International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Data Analyzer Using the Concept of Machine Learning

Minu M.S., Prangad Khanna, Sachit Jain, Ashutosh Saxena

Abstract: ***As the workplace is becoming more tech-driven and fast-moving, data analysis has become an imperative part of any industry. Analyzing a data helps in simplification of complex information. This complex information is extremely difficult to interpret and a data analyzer shows only the most necessary outputs from the stored data. It requires detailed procedures and methodology. The biggest advantage of this interpretation of information is that it gives all the technical insights which is extremely important for any organization. Objective behind making a data analyzer is to provide accurate information to the respective organization. Accurate, significant and timely documentation is important for an effective decision-making process. The data presented can be in the form of figures, graphs, mathematical modelling etc. The key algorithm used for basic integration in this project is shunting algorithm.***

Keywords: shunt yard algorithm, fragmentation, integration, machine learning.

I. INTRODUCTION

The history and future of any organization is depicted through its data. The information stored in the database helps in understanding the statistics of a company at a much deeper level. This information stored, can be very difficult to read and interpret. Thus, a medium is needed to project this information in a much-simplified manner, which is solved using a data analyzer. Considering the field of national security, if a person in doubt needs to be identified through a course of data and information, it becomes extremely difficult to track that particular person down due to immense information stored. A data analyzer will help in fastening this process by analyzing all the data and finally providing common links based on which, one can find the right match. Similarly, for other fields such as business, stock markets, engineering etc [1][2]. The analyzer will provide all the important data which is needful for the user and be extremely precise in giving a clear picture about the situation. Data analysis refers to fragmenting a complete into its unconnected components. The process is all about obtaining unanalyzed data and converting it into information that is useful for decision-making by users. Basic application of machine learning has been applied in this project.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Minu M.S., Department of Computer Science SRM Institute of Science & Technology Chennai, India msminu1990@gmail.com
Prangad Khanna, Department of Computer Science SRM Institute of Science & Technology Chennai, India khanna.prangad1998@gmail.com
Sachit Jain, Department of Computer Science SRM Institute of Science & Technology Chennai, India sachitj4@gmail.com
Ashutosh Saxnea, Department of Computer Science SRM Institute of Science & Technology Chennai, India ashutoshsaxena20150@gmail.com

The concept of machine learning is about studying and understanding algorithms and statistical models. It helps in filtering out data based on certain criteria given by the user. Using the concept of machine learning it is possible to swiftly and automatically construct models that can analyze prodigious, more multiplex data and deliver at a fast rate. By constructing these models, we can avoid unknown risks. Key algorithm followed of integrating the data is “Shunt yard Algorithm” which is a simple technique for parsing infix expressions containing binary operators of varying precedence [3].

Involvement of data integration is the most important aspect in this proposed project. It is divided in to several components and each of it are a crucial part of the entire working mechanism of the system. Each step and section for data integration follows systematic approach and doesn't overlap unnecessarily with other frame of work layers. Here are the following details about each step-in data integration for general analysis:

Data Modeling

Data modeling is done on various different types of fields and projects. These models are highly ongoing and it doesn't promote final data model for any organization or application. They are developed in a manner that they are subjected to as a living document so that they can adapt to any changes made at a later stage of a project. This very factor helps in making the work flow efficient. A data model provides information about the base data that can be used by database developers for creating a physical database. This is a visual representation of all the rules, objects, compliances and government stated policies hence maintaining a good quality of data throughout [3][4].

Data Profiling

Data processing and its analysis can’t take place without the involvement of data profiling, reviewing any source data for content and quality is done through it. As the data gets bigger and infrastructure moves towards the cloud storage, data profiling becomes exceedingly important. It can easily identify the issues in the quality of data, that one can handle via scripts and integration tools of the data for carrying data from source to destination respectively. The effectiveness comes when it highlights the data that has been suffering from multiple issues in terms of quality and content. This mechanism carries out assessment based on which final data is presented.

Data cleansing

The process involves correction of all the incorrect data and identification of inaccurate records, therefore replacing them with what is needed at that time.

The step provides the modified data which is later used in the final stages of the module and is the most precise ensuring that no waste data lives in the analysis procedure.

Data Merging

This mechanism allows you to create various amount of variations while projecting a document. The merged document contains boilerplate information from the target data [4]. The variation created helps in using the information in multiple ways hence helping in the analysis.

Data warehousing

Once the data has been modulated and modifications are performed, it needs to get stored somewhere. Due to immense work on it, the data gets securely saved in the data warehouse. The warehouse can accommodate multiple formats of data and can allow segments for easy recognition of information. The structure of the warehouse created can vary in accordance to how big the data is. Any malfunction in this segment can cause a failure therefore impacting the analysis of the subject.

II. SURVEY

A survey was conducted to have a detailed knowledge and deep understanding of the concepts used in data analyzers. Various models were studied in an efficient way to have an understanding of all the basic mechanisms. Here are some of the techniques, models and data analyzers that have been used in recent times for a useful output of information. The survey includes all the models and techniques which has its application in Data analyzing field. The survey helped in giving a clear picture as to what all has not been implemented.

First was the machine learning for an efficient judgement and prediction of human performance in collaborative learning environments. This Proposes a machine learning based architecture to understand behaviors, dynamics and interaction in the CLE [5].

Second, A new semi-supervised support vector machine using active learning. This paper presents a semi supervised support vector ML algorithm based on active learning. Methods were discussed to improvise the current machine [6].

Third, Energy information analysis using data algorithms based on big data platform. It finds out about the most advantageous algorithm that can efficiently manage the energy data based on big data in particular and by drawing a comparison based on data which is analyzed using certain algorithms such as the support vector machine algorithm and ridge regression algorithm [7].

Fourth, Financial data mining based on support vector machines and ensemble learning. The paper basically compares the support vector machine learning and ensemble learning in terms of financial data. The ensemble learning achieves improvement in the performance over the SVM learning [8].

Fifth, Analysis model of technical and economic data of mining enterprise based on big data analytics. This paper analyses the fluctuation pattern and influencing factors of mineral products by using prediction models: the casual prediction model and extrapolation model [9].

Sixth, Machine learning and its applications review. Paper presented work done by various authors in the field of machine learning. Described application areas using the SVM, clustering, decision trees, logistic regression, etc. [10].

Seventh, the most detailed and closely studied paper by the team. Forecasting Nike's sale using Facebook data. This paper tests if it is possible to forecast the accurate sales for Nike from Facebook data and how events related to Nike affect the activity on Nike's Facebook page. The paper considers various data sets like social data, sales & financial data and google trends. The paper covers huge amount of aspects in machine learning [11].

III. RELATED WORK

Data analysis is a process of examining, cleansing, altering, and creating data with the agenda uncovering all the useful information, giving precise conclusions, and support the basic decision-making mechanism**.** Data analysis has multiple sides to it and approaches, diverse techniques that are used in different domains. Various industries make use of this analysis to work in a systematic way and also a secure environment is maintained. There have been many models and techniques that have been proposed in the recent years. The following involves data analysis:

A. Marketing field:

In the field of Marketing, data analyzers play a key role in making decisions, suggesting practices which are beneficial to make any vital choices. These involve scientific methodology helping any business achieve an effective output or goal. The most common approach for making such suggestions is Data Mining [4]. Data mining is a technique whose ultimate focus is constructing and knowledge discovery used for entirely predictive motive and not just illustrative justification. [10]

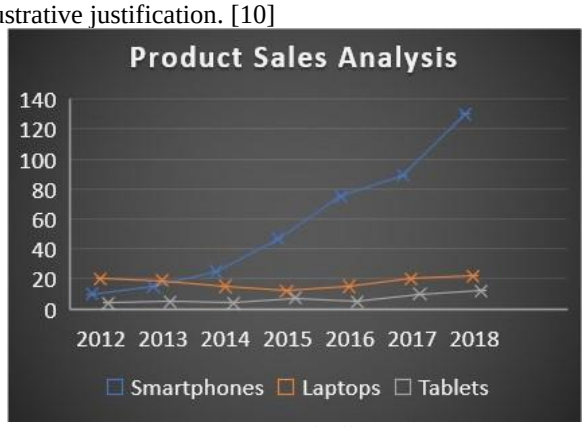

Fig 1. Sales Analysis of different products

B. Analysis of Stock Market:

Data analysis is being used in the industry of studying stock market and making predictions based on it. This involves statistical data analyzers which can further be divided into descriptive statistics, exploratory data analysis (EDA), and confirmatory data analysis (CDA). Each of these perform different actions and study the data in a different manner [5]. The stock market is an area where it's extremely important that the data read or interpreted is presented in the most correct manner. It must be error free and this is achieved via the modes of analysis mentioned above.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

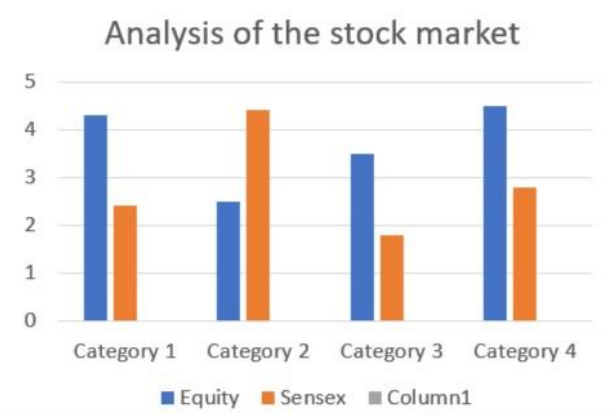

Fig 2. Analysis of Stock Market

C. Weather Predictions:

Predictions are made about the weather based on certain analysis. The weather department ensures that the most accurate data is presented to ensure safety and prevent disastrous encounters. Data analyzers are used, which read the weather map and other set of logical data to make decisions. Here the application of machine learning comes in account and with analysis a final report is produced. The below graph analyses the weather pattern for the next 30 days [7].

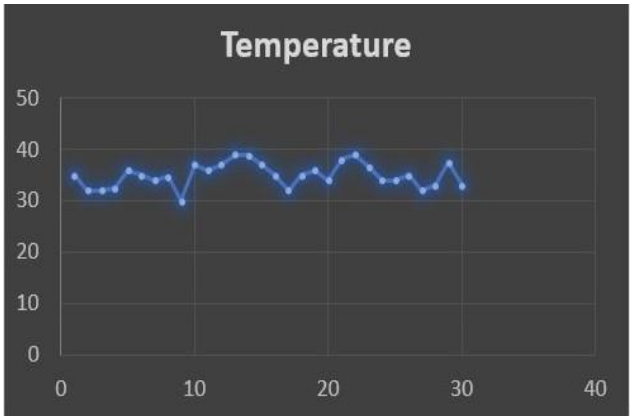

Fig 3. Analysis of the weather for the next 30 days

D. Security and Defence:

This is one of the most important field for any country. Data analyzers are used in bulk in the security and Defense of country. They make decisions based on movements, radiations, sounds and satellite images. The data is studied and through an analyzer the final prediction is made through which actions are taken [7]. Analyzing of data helps in taking actions rapidly and arrangements are made to ensure that people are secure from any damage.

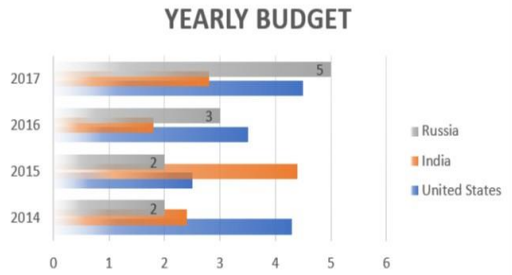

Fig 4. Analysis of the yearly budget spent by three countries

E. Healthcare ingression:

Ensuring that an individual's health who is admitted is monitored properly and to know the root cause of the health hazard, data analyzers are used. They study the patients' medical history and current problem and draw conclusions based on that. One of the most advanced data analyzers are used in this industry. They are also used to monitor a larger area which is affected by a certain virus and presents suggestions as to how one can overcome and fight it out.

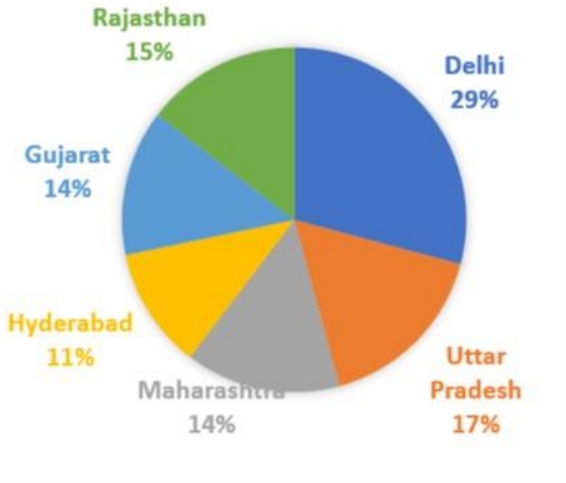

Fig 5. Analysis of number of fatal diseases incidents in 6 states

F. Cyber security:

Big data comes with threats and hostile actions, thus making it mandatory to mitigate all the cyber security risks. Analyzers recommend methods to make this environment secure and increases efficiency. The key advantage of using data analyzers here is that it gives fast solutions which prevents unethical leakage of data from any source [12].

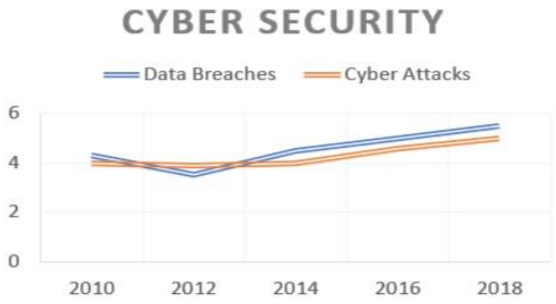

Fig 6. Analysis showing security threats over the years

G. Biosystems data analysis:

The biological systems generate a huge amount of data. The data stored needs to be converted in to information (which is summarized with full specifications) and finally made available to the user visually. Models are constructed to take the analysis forward which helps in the integration of bioinformatics and biostatistics that are entered in the storage. Here analyzers can only be used once models are created [9].

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

H. Transportation :

It is very important to control traffic and manage transportation in a way that least amount of time is taken to cover a distance. With the help of manual inputs and data analysis using a GPS monitoring system a balanced plan can be given in order to travel efficiently and in time. This mechanism also suggests and makes prediction based on the data as to where one can find congestion. Google maps analyzes your daily commute for work and hence gives the exact time it will for you to reach to your work office. This data is given on the basis of how many traffic signals are there on your way and identifying the density of cars travelling in that direction. That is how it calculates the approximate time taken to reach your destination from a particular way.

I. Delivery Logistics

Timely delivery for a product is a must and of utmost importance. Data analysis in the field of delivery logistics help in finding suitable means of transportation, show progress and give statistical growth recommendations. A large amount of companies make use of it in order to grow effectively without incurring any damages or losses due to malfunctioning at production end [9][12].

IV. SYSTEM ARCHITECTURE

The system Architecture of the Data analyzer proposed follows simple working mechanism. At first the data is stored in the data warehouse and segmented in to different fields for a better processing and to avoid data overload and confusion between different sets of data [14].

Since one has to find useful data for an active decision-making process, there is a need to find it in a measurable way to know if the particular organization is advancing towards its ultimate goal. Key points or indicators must be identified at an early stage to make the work flow easier.

In accordance to the requirement, data from the warehouse is produced which is monitored and reduced before presenting it to the next level [15]. At this stage of reduction, all the unwanted information is removed keeping the filtration in mind. The data that has to be analyzed could be textual, numerical or even pictures and graphs. The main problem arises with the textual data since it can be stored in multiple formats and can be collected from various sources. This particular issue is solved using segmentation of data. The database holding up the data stores information according to fields and internally it is again divided in to parts which can segregate different formats of similar data. After reduction, data cleansing is done to improve data quality. It is considered to be the most important step in the entire data value chain [15]. With the help of this method, all the junk data is removed and information produced is made to be accurate, precise and error free. After the cleansing is done, the information is passed for processing and made available at user level.

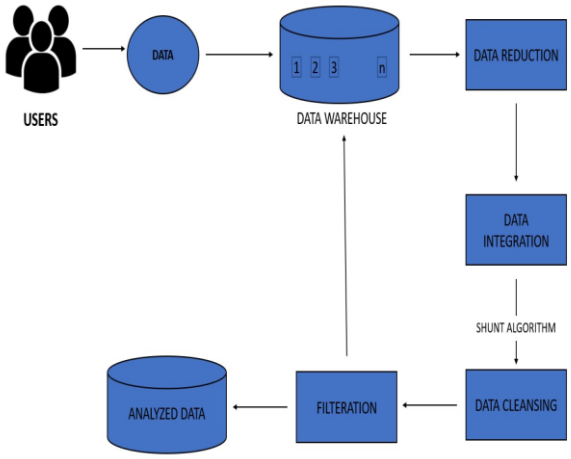

Fig 7. System Architecture of the proposed Data Analyzer

The biggest advantage of such analysis is, that it reduces time and complexity to find a particular information or record. The framework produced is different from the other proposed modules in many ways. At first it is extremely cost efficient since it can accept all sorts of inputs (Textual, numerical, pictures and graphs). Secondly, the information produced is filtered with sequence of steps and integrated to provide accurate information to its users. Third, the data gets stored in segments which makes it useful for different types of data to be stored and retrieved [12]. Hence not causing data overload. Fourth, the analyzer proposed can be used in multiple fields at a small-scale level which makes work load less and saves time (Ex:A student's data is stored and one needs to find all the needful information and history backdrop of the student, then this analyzer with the help of a few manual inputs, shall produce the needful data by connecting it with other sources attached to the warehouse.

V. FUTURE FRAMEWORKS

The data analyzer proposed can be improvised in many ways in terms of storage capacity. Since the current one can analyze all types of data but can hold up only limited set of information due to low data storage therefore causing burdensome. This can be used in the future in a much more coherent way, by providing abundance of storage so that this issue doesn't emerge. It has been particularly made to study data and analyze it at a small scale. Thus, changing it for a large-scale analysis is one of the key future works. Next is the inclusion of a methodology which can test accuracy in parameter estimation. This is important to make sure that the analyzed data is the most precise and has no bugs in it. It is the most self-commissioning techniques and completely based on statistical analysis. Thirdly, including a Statistical Power Analysis technique, which is an analysis directly related to tests of hypotheses. While conducting tests of hypotheses, one can commit two types of errors namely, Type I error and Type II error. The technique deals with type II error only. Which is the most tedious one to solve.

Main agenda is to determine the compact sample size that can be used for detection while testing at the desired level of significance.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Fourth, adding user defined rules for generating extensive range of additional information according to ones needs. This further can be used for drilldown and analysis. Advantage of a drilldown function is that it can enable us to extract any data and can be viewed at press of a key. The rules basically provide supportive supplementary information for the dataset.

VI. RESULTS AND ANALYSIS

Data analyzer proposed will provide exact knowledge instead of unimportant assumptions. It is considerably fast and highly precise than manual analyzing techniques. Interface of the analyzer is user friendly, simple and intuitive. The advantage is that it can reduce effort and time for analyzing a data. Segmentation and integration allow the data to be stored using various criteria. Reduction process is rapid and useful data is presented after filtering.

REFERENCES

1. Sheikh Shaugat Abdullah, Mohammad Saiedur Rahaman and Mohammad Saidur Rahman, “Analysis of stock market using text mining and natural language processing”.
2. Sheena Angra, Sachin Ahuja, “Machine learning and its applications”, International Conference On Big Data Analytics and computational Intelligence (ICBDACI), 2017.
3. Chenn-Jung Huang, Ming-Chou Liu, San-Shine Chu, Chin-Lun Cheng, “Application of Machine Learning Techniques to Web-Based Intelligent Learning Diagnosis System”, Fourth International Conference on Hybrid Intelligent Systems (HIS’04), 2004.
4. Linda Camilla Boldt, Vinothan Vinayagamoorthy, Florian Winder, Melanie Schnittger , Mats Ekran, Raghava Rao Mukkamala1, Niels Buus Lassen, Benjamin Flesch, Abid Hussain and Ravi Vatrapu, “Forecasting Nike’s Sales using Facebook Data”, IEEE International Conference on Big Data (Big Data), 2016.
5. A. ,Sen, “The us fashion industry: a supply chain review,” International Journal of Production Economics, vol. 114, no. 2, pp. 571–593, 2008.
6. S. Thomassey, “Sales forecasts in clothing industry: The key success factor of the supply chain management,” International Journal of Production Economics, vol. 128, no. 2, pp. 470– 483, 2010.
7. H. Choi and H. Varian, “Predicting the present with google trends,” Economic Record, vol. 88, no. s1, pp. 2–9, 2012.
8. M. Coakley and D. Song, “Using google trends to predict retail sales,” http://www.pwc. com/us/en/retail-consumer/publications/assets/ pwc-using-google-trends-to-predict-retail-sales.pdf, price waterhouse Coopers.
9. S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, “Predicting consumer behavior with web search,” Proceedings of the National academy of sciences, vol. 107, no. 41, pp. 17 486–17 490, 2010.
10. Nike Inc, “Nike, inc. reports fiscal 2014 fourth quarter and full year results,”http://s3.amazonaws.com/nikeinc/assets/31010/NIKE_Inc_Q414_Press_Release_-_6-25-2014_6P_CLEAN_1_.pdf?1403806476, June 2014.
11. A. Hussain and R. Vatrapu, “Social data analytics tool (sodato),” in DESRIST-2014 Conference (in press), ser. Lecture Notes in Computer Science (LNCS). Springer, 2014.
12. G. Shmueli and O. R. Koppius, “Predictive analytics in information systems research,” MIS Q., vol. 35, no. 3, pp. 553–572, Sep. 2011. [Online].Available:http://dl.acm.org/citation.cfm?id=2208923.2208926
13. B. Flesch, R. Vatrapu, R. R. Mukkamala, and A. Hussain, “Social set visualizer: A set theoretical approach to big social data analytics of real-world events,” in Big Data (Big Data), 2015 IEEE International Conference on. IEEE, 2015, pp. 2418–2427.
14. W. S. Woon, “Introduction to the event study methodology,” Singapore Management University, 2004.
15. D. Cram, “The event study webpage,” http://web.mit.edu/doncram/www/eventstudy.html.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)