International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 14, Issue 12, 2020

Big Data Analytics in the perspective of Digital Businesses: A Case Study Approach

Maqsood Mahmud, Department of Management Information Systems (MIS), College of Business Administration, Imam Abdulrahman Bin Faisal University, P.O.Box 1982, Dammam, Saudi Arabia mMahmud@iau.edu.sa

The evolution of digital mining tools has resulted in the ease of access to massive quantities of information that can be used by digital businesses. Data analysis methodologies are utilised to scan enormous quantities of data for critical business guidance. The process of digging and drilling through data is used to obtain market insights, as well as to access obscure information in a wide range of data sources or even the existing real-time increasing web data ocean. Data analytics tools certainly extract concealed associations, forecast potential events, and further interpret and distribute business supplies. Such hidden knowledge seeks to achieve competitive advantages, strengthen client interactions, and even prevent fraudulent activities. In this study, quantitative analysis was performed with Exploratory Data Analysis (EDA) techniques by using three different case studies. These cases were analysed using secondary datasets for digital businesses by utilising data mining tools and secondary datasets from Kaggle. Our results showed that data mining tools like Rapidminer and/or Tableau can efficiently handle diverse kinds of data from various digital organisations, and hence, big data from diverse organisations with high volume, high velocity, and high veracity. The three case studies resulted in the conclusion that the extracted data can be tactfully transformed into valuable information using the market available data mining tools.

Keywords: *Big data, Knowledge extraction, Data archaeology, Pattern analysis, Digital businesses, Data mining tools, Digital organisations.*

INTRODUCTION

Data mining, forecasting relevant data, and extraction from vast repositories remain efficacious innovative technical knowledges with exceptional capabilities for predicting future behaviours and trends. Thus, they enable businesses to concentrate solely on the more crucial data in their

database systems, as well as succeed in making knowledge-based decision making. The remarkable progress that has been executed in data digging techniques has become an altered interest of information acquisition towards knowledge generation and aggregation. Today's efficient and comparatively affordable hardware technology, combined with sophisticated software, makes data mining a vital business tool. Apart from this, the Internet also portrays an influential role because networking and messaging have grown diverse and widespread, and data mining is performed globally through the adoption of integrated networks. The excessive amount of expert knowledge is not only accessed by the executive level, but rather includes all levels of the corporation (Alazemi, A. & Alazemi, A., 2016).

LITERATURE REVIEW

Business intelligence (BI) is often viewed as an infrastructure, method, application or program which captures and preserves data, and evaluates it by employing analysis techniques, as well as provides details or insights, promotes analysis, and questions, which fundamentally enables companies to enhance their decision-making process. In brief, BI could be regarded as a cycle that converts data into meaningful information, and subsequently, into knowledge. Becoming deeply embedded throughout the field of decision support systems (DSS), BI has endured substantial progression within the last few years and became an aspect of DSS, which mostly drives a great deal of interest from both the industrial sector, and analysts. Almost all BI applications are developed on top of relational databases. As a response, data mining (DM) incorporation with relational databases is an essential concern to be addressed while evaluating DM incorporation with BI (Azevedo, A. & Santos, M., 2013).

BACKGROUND

There have been several breakthroughs contributing to the BI processes which we see today. Such advances set foot on the glory days of mathematical concepts, as well as quantitative research employing correlation and Bayesian approaches throughout the mid seventeen hundreds. Upon the emergence of industrial digital equipment after the end of WWII, large amounts of data has become preserved in magnetic tapes to optimise industry procedures. Dating from the initial nineteen-sixties, data generated in early industrial machines allowed researchers to address basic analytical market issues (Alazemi, A. & Alazemi, A., 2016).

Today, BI technologies are much more sophisticated and have more than just monitoring functionality; they can reveal invisible trends, forecast potential events, and assess risks. Many of these technological services were first established throughout the nineteen-nineties, with exponential acceleration within the beginning of the new millenniums (Alazemi, A. & Alazemi, A., 2016).

Historically, BI is often conducted through conventional approaches. Those approaches were troublesome, unreliable, and essentially ineffective. To eliminate these complications,

electronic applications, including worksheets, and online analytical processing (OLAP) offered a modern means of gaining business insights. Such technology had its flaws, like restricted features and capabilities, as with OLAP, which cannot offer prediction assessment, but instead are applied as front-side analytical techniques through evaluation (Alazemi, A. & Alazemi, A., 2016).

DATA MINING FUNDAMENTALS

Data mining is acknowledged as the analysis of data utilising several metrics to derive undercover information or insight into the data. Data mining is commonly identified as the information exploration in the repositories or the knowledge discovery in databases (KDD) procedure stage. Data analysis methods entail presenting some categorisation and grouping, correlation of mining rules, anomaly recognition, summarisation, regression, and sequential patterns. Several of these technologies are quite beneficial for BI systems. Naturally, most DM algorithms and operating systems use BI technologies of this sort (Alazemi, A. & Alazemi, A., 2016).

Data mining is widely adopted in BI applications, and many instances of models can be presented. Business intelligence and data mining have different origins, and as a result, possess substantially distinctive features. Data mining evolved from a systematic background, and therefore, it is non-business oriented (Azevedo, Ana., 2014). Data mining applications still require a great deal of effort to produce the desired outcomes. Conversely, BI is embedded in the industrial sector and market. consequently, the methods of business intelligence are designed to be easy for an untrained user to use (Azevedo, A., 2014), (Saleem, F. & Malibari, A., 2011).

WHY IS DATA MINING IMPORTANT

The question here is that "why is data mining so crucial?". To break it down to you, the amount of information generated is multiplying every couple of years. Unorganised data only represents 90 per cent of the digital world. Increasingly, data does not necessarily equal greater knowledge (Gallant, D., et al., 2016), Chetan G. and Ahmed F. (2020). Data mining helps you to:

A. Search through all the messy and redundant activity in your results (*Li, A. & Zhang, L., 2009)*
B. Recognise what is important, and then make proper use of this knowledge to determine the possible options and/or cases (*Li, A. & Zhang, L.., 2009)*
C. Accelerate the amount of informed decision-making (*Gallant, D. et al., 2016).*

METHODOLOGY

The cycle of data mining is broken down into five stages. First, companies gather and transmit data to their data centres. Next, they save and control the data, whether on in-house servers or a cloud. Company examiners, administration groups, as well as information technology (IT) experts have exposure to the data and decide whether they want to arrange it. After that, the software program filters the data depending upon the outcomes of the user, and subsequently, the end-user displays the data in a handy-to-share way, including a chart or a list. Data mining systems examine data connections and habits depending on what consumers are looking for. For instance, a business can use data mining tools to build clusters of information. Try imagining, for example, that a diner chooses to use data mining to decide if certain specials should be served. It displays the information it has gathered and generates categories depending on when consumers come and what they tend to order most as depicted by Kemal, Mohammed. (2019) that how successful organizations use data visualization. This article will adopt a quantitative analysis based on exploratory data analysis (EDA) techniques using the Rapidminer and Tableau tools. Three cases will be selected for EDA analysis using secondary datasets for an organisation that is closely related to digital businesses using the Rapid Miner and/or Tableau (RapidMiner | Best Data Science & Machine Learning Platform), (Business Intelligence and Analytics Software (tableau.com) and (Li, A. & Zhang, L., 2009), Haroon, D. (2017), Mathur, P. (2019)..

CHALLENGES OF DATA MINING

Big Data:

Large data problems are extensive and common in several ranges that capture, preserve, and analyse data. Big data is distinguished by four main difficulties: quantity, range, accuracy, and speed. Data drilling helps to resolve these problems and discover the importance of data (Lee, R., 2020). (Najafabadi, M.M., et al., 2015).

Over-Fitting Models:

Over-fitting arises when the model illustrates some regular deviations inside a particular sample, preferably than the underlying group patterns. Over-fitted models are frequently too complicated and use an array of independent variables to make predictions. Consequently, the threat of over-fitting rises in volume and a variety of results (Lee, R., 2020).

Cost of Scale:

As data pace persists expanding data quantity and range, companies will scale and extend these models throughout the whole organisation. Unlocking the maximum benefits of data mining with these models demands substantial improvement in computing technology and processing

capability. Businesses should purchase and maintain strong devices, systems, and software equipped to accommodate the businesses' enormous volume and range of information (Lee, R., 2020).

Secrecy and Protection:

Enhanced data room demands have caused several businesses to switch to cloud computing and storehouse. Although the cloud has empowered countless technological progressions in the data mining field, the design of the service produces major challenges to secrecy and protection. Companies must secure their data against fraudulent activity to preserve the trust of their stakeholders and consumers (Lee, R., 2020).

TRENDS OF DATA MINING

Language Standardisation:

Closely related to how structured query language has emerged to be the next dominant language for repositories, clients are actively trying to strive for centralisation through data analysis. This drive helps users to communicate easily with several various mining platforms by just knowing one basic language (Lee, R., 2020).

Scientific Mining:

Data mining is being applied in science and academic study with its confirmed effectiveness in the business sector. Psychologists are increasingly employing correlation analysis to document and classify wider trends in human activity to help their studies. Similarly, economists use forecasting algorithms to estimate possible price conditions dependent upon existing variables (Lee, R., 2020).

Complex Data Objects:

While data mining is growing to affect many divisions and sectors, new approaches are being established to examine increasingly diverse and complex data (Lee, R., 2020).

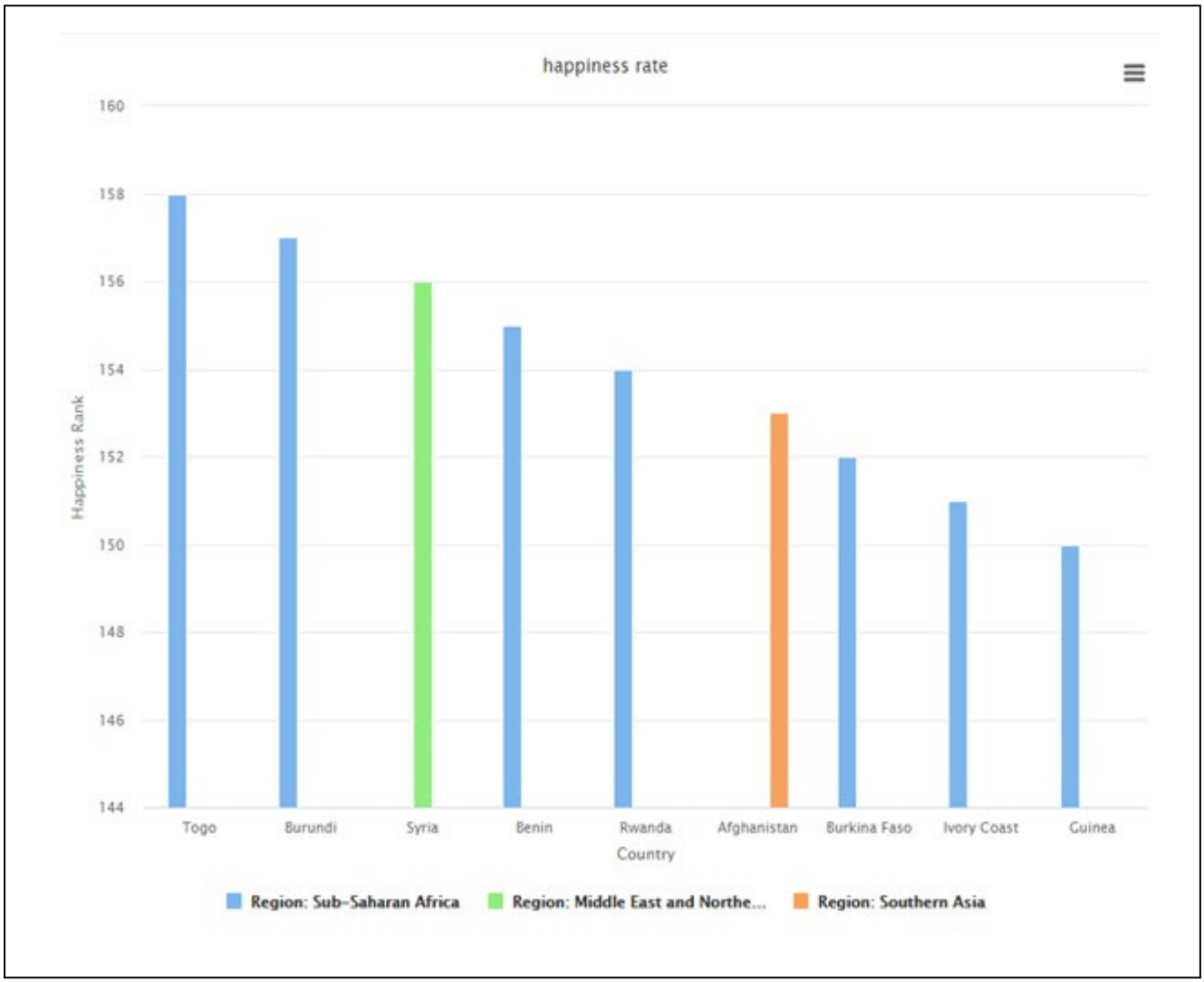
The Higher Pace of Computing:

As data size, intricacy, and scope rise, data digging technologies demand quicker and more reliable systems and far more powerful data analysis processes. Each new observation introduces an additional loop of computation to the study (Lee, R., 2020).

Web Mining:

Through the growth of the Internet, the exploration of habits and developments of usage is of considerable importance to organisations. Web mining follows relatively similar approaches to data analysis and employs them instantly on the Network. The three major categories of web mining are material mining, architecture mining, and utilising mining (Lee, R., 2020).

Figure 1. Happiness trend anlaysis around the world



VISUALISATION AND RESULTS

In this section, we discuss three different cases of big data (diverse data) analysis using the Rapidminer/Tableau tool. The first case is about the 'world's happiness rate analytics', while the second case is about the 'hotel booking business analytics'. The third case is about the 'World's Academia Ranking Analytics'. All these big data cases are analysed using a market available tool named, 'Rapidminer/Tableau'. Moreover, datasets are adopted from Kaggle,

which is the world's most popular online cloud-based programming and dataset community platform. Here, experts from all over the world contribute their experiences regarding the world's popular problems related to artificial intelligence, machine learning, and deep learning. It is a subsidiary of Google LLC, started in 2010, by offering machine learning competitions. Moreover, it permits users to discover, learn, and publish data sets. It also helps naive users to investigate and develop models in a web-based data-science environment. It helps researchers to contribute with other data scientists and machine learning engineers. Moreover, it holds various monetary and non-monetary awards competitions to solve data science challenges, including Microsoft and/or Google challenges, etc.

CASE 1: WORLD'S HAPPINESS RATE ANALYTICS

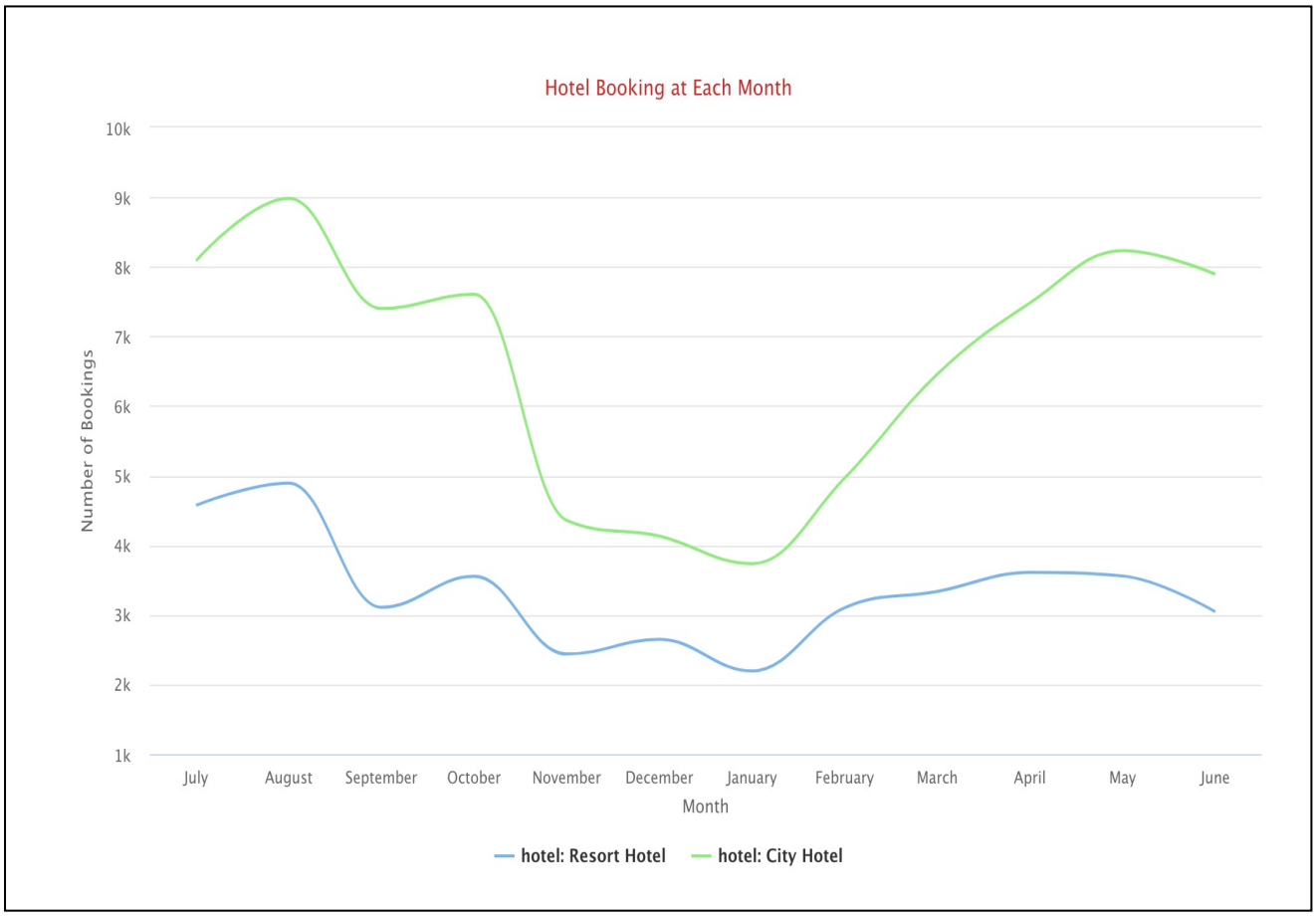
The visualisation of raw data simplifies the process of extraction of knowledge. The inference for this case study is taken from the existing counter case studies in literature, such as displeased people that were inclined to suicide attempts. This case study discussed a unique data mining approach for the uncovering and prediction of a suicidal case within social media networks in Singapore (Seah, J. & Shim, K., 2018). Similar displeased cases are also discussed in (Amirkhanyan, A. & Meinel, C., 2017) and (Deng, Z. & Fang, & Guo-Dong., 2007) as inference. The software tool Rapidminer/Tableau is one of the data mining tools which can extract the needed data from a substantial amount of raw data and represent it as whole information. The dataset for the big data analytics process case was adopted from Kaggle (J. F. R. L. J. S. J.-E. D. N. Helliwell, "world happiness report 2020,"). In our case study, the first visual is a graph about the happiness rank in some of the countries, and the happiness rate visual shows the least happy countries around the world. The country Togo is at the top with a happiness ranking of 158 points in Sub-Saharan Africa, while in contrast, Guinea is at the lowest ranking with 150 points as per Figure 1. The only middle eastern country is Syria, representing a happiness rank of 156. The happiness rank in every country around the world, and specifically to the dataset used, is based on several factors, such as the standard error, market (gross domestic product GDP per capita), family, well-being, and independence.

CASE 2: HOTEL BOOKING BUSINESS ANALYTICS

The second chosen case for big data analytics is related to the '*hotel booking business Analytics*' case. The inference for this case study is taken from the existing case study of the customers' mining case study discussed by Bodendorf, F. & Kaiser, C. (2010). In Athanasopoulos, G. & Hyndman, R. (2011)., the authors articulated challenges in the traditional design and the value of feedback in forecasting competitions. Similarly, the dataset for the 'hotel booking business analytics' case is adopted from Kaggle. In this case, the dataset consists of a hotel booking history of more than 100,000 hotel stay between 2015 and 2017, and from 178 different countries, which was adopted from Kaggle (A. A. L. N. Nuno Antonio, "Hotel booking demand," 2019. Additionally, the dataset shows the booking of two hotel types: city hotels, and resort hotels. The result shows that August had the highest number of bookings

for both cities, and resort hotels. In second place was the month of July. It can be seen from Figure 2 below that January would be the best month of the year to book a hotel, if someone is seeking a less crowded time to spend their vacation, especially during the Coronavirus pandemic of 2020.

Figure 2. Trend analysis of hotel bookings for each month



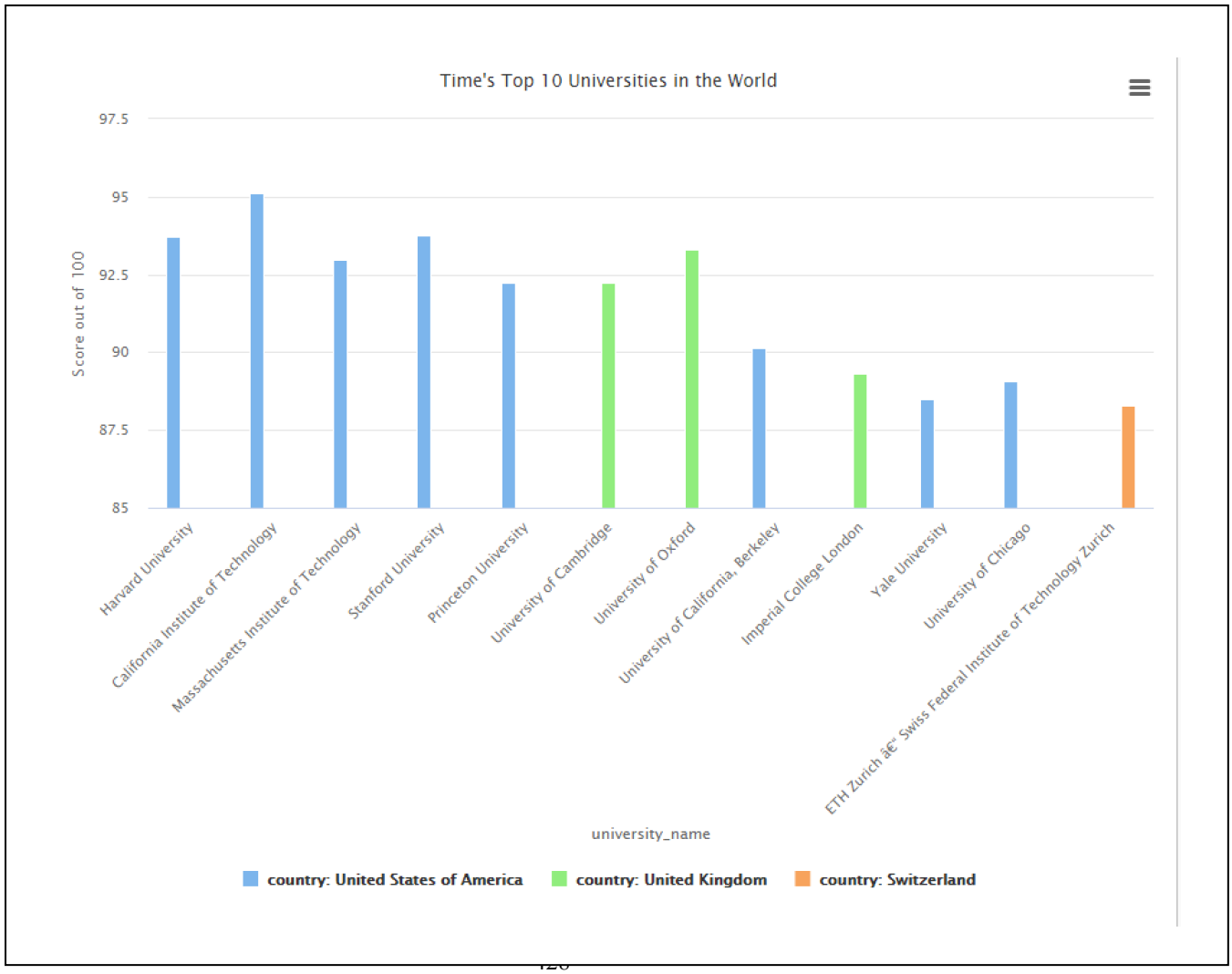
CASE 3: WORLD'S ACADEMIA RANKING

The third case that we selected for big data analytics is related to 'world academia ranking'. A dataset from Kaggle (T. H. E. W. U. Ranking, "World University Rankings," 2010.) was obtained. The inference for this case study is taken from the existing case study of a log mining case study of Qafqaz University, as discussed by Adamov, A. (2014) and a research design for a data analysis system for student education improvement. This case study focused deeply on the students' progression system in the university students' ranking (Singh, K., 2016). The newly updated science-wide author databases of standardised citation indicators published in the journal of Public Library of Science (PLOS) is also a correlated case study for inference cited in (Ioannidis J., Boyack K., & Baas J., 2020). In our case of the 'world's academia ranking' case study, the Kaggle dataset consists of more than 1,000 data rows and 14 columns containing university ranking criteria for more than 500 universities around the world, based

upon the ranking of 'Times University', which is considered one of the most specialised in university rankings. The total ranking score was given based on the publications and research by the university, teaching environment, number of citations, university income, and number of students. Figure 3 shows the top ten universities analysed by feeding the secondary data to the Rapidminer/Tableau data mining tool. The results showed that six out of the ten universities are based in the United States, three are in the United Kingdom, and one is in Switzerland. The California Institute of Technology came in at the top of the list with a total score of 95 per cent, while Stanford University and Harvard University came in second, and third place, respectively (T. H. E. W. U. Ranking, "World University Rankings,", 2010).

Thus, the above three cases show that market available BI and/or data mining tools, such as the Rapid miner, can suffice the needs of the business community in a variety of big data analytics issues or problems, whether it is from an academic perspective, pure customers oriented business needs or governmental and/or private sector organisational survey needs related to public well-being (Cao, L., 2008).

Figure 3. Top 10 universities trend analysis around the world



CONCLUSION

It is concluded that market-oriented data mining tools like Rapidminer/Tableau have the capabilities to analyse big data problems from various perspectives of digital business and diverse organisations. It is an essential tool to analyse, categorise, and understand big data and enhance the business intelligence processes of a digital organisation. These types of tools have evolved substantially over the past few years, and have now become significant tools for data grouping, anomaly identification, regression, pattern finding, and correlations. Furthermore, data mining technologies still require some effort to produce the desired outcomes, and they face several challenges, including legal and moral challenges, as part of functional limitations.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia. The author is grateful for this support.

REFERENCES

Adamov, A. (2014). Data Mining and Analysis in Depth. Case Study of Qafqaz University HTTP Server Log Analysis. 10.1109/ICAICT.2014.7035947.

Alazemi, A. & Alazemi, A. (2016). Overview of Business Intelligence through Data Mining. 10.4018/978-1-4666-9562-7.ch003.

Amirkhanyan, A. & Meinel, C. (2017). The framework for spatiotemporal sequential rule mining: Crime data case study. 34-38. 10.1109/ICKEA.2017.8169898.

Athanasopoulos, G. & Hyndman, R. (2011). The value of feedback in forecasting competitions. International Journal of Forecasting. 27. 845-849. 10.1016/j.ijforecast.2011.03.002.

Azevedo, A. & Santos, M. (2013). A Perspective on Data Mining Integration with Business Intelligence. 10.4018/978-1-4666-2455-9.ch097.

Azevedo, A. (2014). Data Mining and Business Intelligence: A Comparative, Historical Perspective. 1-11. 10.4018/978-1-4666-6477-7.ch001.

A. A. L. N. Nuno Antonio, "Hotel booking demand," 02 2019. [Online]. Available: https://www.kaggle.com/jessemostipak/hotel-booking-demand. [Accessed 10 07 2020].

Bodendorf, F. & Kaiser, C. (2010). Mining Customer Opinions on the Internet A Case Study in the Automotive Industry. 24-27. 10.1109/WKDD.2010.129.

Business Intelligence and Analytics Software (www.tableau.com)

Cao, L. (2008). Domain Driven Data Mining (D3M). 74-76. 10.1109/ICDMW.2008.98.

Chetan G. and Ahmed F. (2020). Deep Learning for Industrial AI: Challenges, New Methods and Best Practices. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). Association for Computing Machinery, New York, NY, USA, 3571–3572.

Deng, Z. & Fang, & Guo-Dong. (2007). Mining Top-Rank-K Frequent Patterns. Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007. 2. 851 - 856. 10.1109/ICMLC.2007.4370261.

Gallant, D. & Gauvin, L. & Berteaux, D. & Lecomte, N. (2016). The importance of data mining for conservation science: a case study on the wolverine. Biodiversity and Conservation. 25. 2629–2639. 10.1007/s10531-016-1188-5.

Haroon, D. (2017). Python Machine Learning Case Studies: Five Case Studies for the Data Scientist. 10.1007/978-1-4842-2823-4.

Ioannidis J., Boyack K., Baas J., (2020). Updated science-wide author databases of standardized citation indicators. PLoS Biol 18(10).: e3000918. https://doi.org/10.1371/journal.pbio.3000918

J. F. R. L. J. S. J.-E. D. N. Helliwell, "World Happiness Report 2020," 2020. [Online]. Available: https://www.kaggle.com/mathurinache/world-happiness-report. [Accessed 10 07 2020].

Kemal, M. (2019). Successful Company Using Data visualization. Report published.

Lardinois, Frederic; Mannes, John; Lynley, Matthew (March 8, 2017). "Google is acquiring data science community Kaggle". Techcrunch. Archived from the original on March 9, 2017. Retrieved March 9, 2017.

Lee, R. (2020). Data Mining. 10.1007/978-981-15-7695-9_4.

Li, A. & Zhang, L. (2009). A Study of the Gap from Data Mining to Its Application with Cases. 2009 International Conference on Business Intelligence and Financial Engineering, BIFE 2009. 464 - 467. 10.1109/BIFE.2009.111.

Mathur, P. (2019). Monetizing Healthcare Machine Learning: Cases Studies from Healthcare, Retail, and Finance. 10.1007/978-1-4842-3787-8_6.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. et al. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data 2, 1

RapidMiner | Best Data Science & Machine Learning Platform

Saleem, F. & Malibari, A. (2011). Data mining course in information system department- case study of King Abdulaziz University. 10.1109/ICEED.2011.6235378.

Seah, J. & Shim, K. (2018). Data Mining Approach to the Detection of Suicide in Social Media: A Case Study of Singapore. 5442-5444. 10.1109/BigData.2018.8622528.

Singh, K. (2016). Design and Research of Data Analysis System for Student Education Improvement (Case Study: Student Progression System in University). 508-512. 10.1109/ICMETE.2016.80.

T. H. E. W. U. Ranking, "World University Rankings," 2010. [Online]. Available: https://www.kaggle.com/mylesoneill/world-university-rankings?select=timesData.csv. [Accessed 10 07 2020].