International Journal of Management, Finance and Accounting

Leveraging Business Data Analytics and Machine Learning Techniques for Competitive Advantage: Case Study Evidence from Small Businesses

Rathimala Kannan[1]*
Intan Soraya Binti Rosdi[1]
Kannan Ramakrishnan[2]
Haziq Riza Bin Abdul Rasid[1]
Mohamed Haryz Izzudin Bin Mohamed Rafy[1]
Sukinurlin Yusuf[1]
Siti Nurfara Alia binti Mohd Salamun[1]

*Corresponding author: rathimala.kannan@mmu.edu.my
[1]Faculty of Management, Multimedia University
[2]Faculty of Computing and Informatics, Multimedia University

Abstract

Data analytics is the essential component in deriving insights from data obtained from multiple sources. It represents the technology, methods and techniques used to obtain insights from massive datasets. As data increases, companies are looking for ways to gain relevant business insights underneath layers of data and information, to help them better understand new business ventures, opportunities, business trends and complex challenges. However, to date, while the extensive benefits of business data analytics to large organizations are widely published, micro, small, and medium sized organisations have not fully grasped the potential benefits to be gained from data analytics using machine learning techniques. This study is guided by the research question of how data analytics using machine learning techniques can benefit small businesses. Using the case study method, this paper outlines how small businesses in two different industries i.e. healthcare and retail can leverage data analytics and machine learning techniques to gain competitive advantage from the data. Details on the respective benefits gained by the small business owners featured in the two case studies provide important answers to the research question.

Keywords: Data Analytics, Machine learning Techniques, Prediction models, Market Basket Analysis.

Submitted on 15 November 2020; Accepted on 29 November 2020; Published on 25 February 2021.

https://doi.org/10.33093/ijomfa.2021.2.1.3
Published by MMU PRESS. URL https://journals.mmupress.com/index.php/ijomfa/index

1. Introduction

The world is transitioning to Industry 4.0, a visualisation of what the future of the manufacturing is like. Industry 4.0 embraces information technologies to boost competitiveness and efficiency of organisations by interconnecting data, people and machinery in its value chain (Gottge, Menzel, & Forslund, 2020). Many large organizations from various industries have recognised the benefits of connecting data, people and machinery by leveraging data analytics and machine learning techniques (Raguseo, 2018).

Data analytics is the essential component in deriving insights or value from data obtained from multiple sources. It represents the technology, methods and techniques used to obtain insights from massive datasets. The techniques automatically analyse data to deduce non-linear relationships and causal effects which is generally scattered in the datasets (Lu, 2020). However, to date, while the extensive benefits of business data analytics to large organizations are widely published such as data mining and predictive analytics applications for the delivery of healthcare services (Malik, Abdallah, & Ala'raj, 2018), Retailing and retailing research in the age of big data analytics (Dekimpe, 2020), it is found that micro, small, and medium sized organisations have not fully grasped the potential benefits to be gained from data analytics using machine learning techniques (Dwija Utama, Diryana Sudirman, & Alamsyah, 2020) .

Hence, the question to be answered in this study is: how can the use of data analytics using machine learning techniques benefit small businesses? Using the case study method, this paper's objective is to identify the specific benefits brought about by business data analytics to small businesses. The use of data analytics and machine learning techniques in two different industries is discussed in separate case studies featuring small businesses in the respective industries, i.e. healthcare and retail.

2. Analytical Processes

The two case studies featured in this paper are from coursework projects for a 'Business Data Analytics for Managers' course in a Master of Business Administration (MBA) program at a Malaysian higher education institution. The analysis for the two case studies were conducted using an open source software, KNIME analytics platform. Both case studies follow the Cross-Industry Standard Process for Data Mining (CRISP_DM), which is the most widely used analytics model using a set of guidelines to help plan, organize and

execute data analytics projects (Dwija Utama et al., 2020). The CRISP_DM consists of 6 phases, namely:

1. Business Understanding – Understanding the business problem and defining business requirements and objectives.
2. Data Understanding – Collecting data and assessing data quality.
3. Data Preparation – Handling of problems such as missing data, incorrect data etc. using data pre-processing methods.
4. Modelling – Using one or more machine learning techniques to build models.
5. Evaluation – Evaluating the models' effectiveness based on standard metrics and to determine whether the defined objectives are achieved.
6. Deployment – Deploying the model.

The 6 phases of CRISP_DM are illustrated in Figure 1 below.

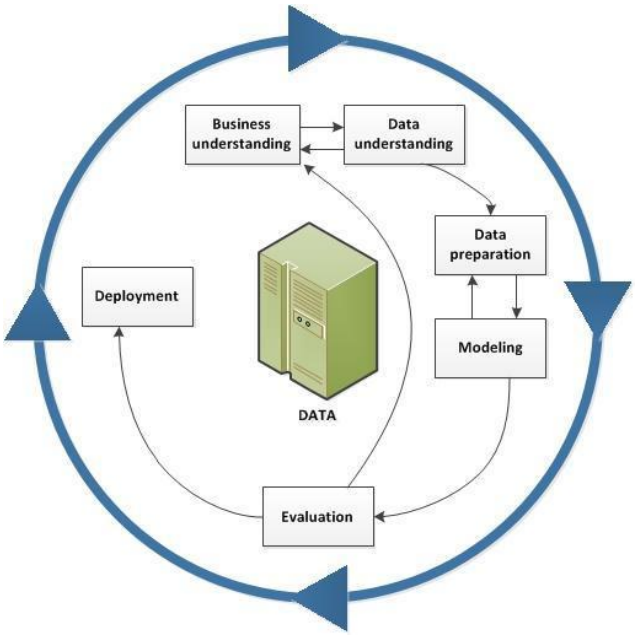


Figure 1: Cross-Industry Standard Process- Data Mining (CRISP-DM)

3. Case Study 1 – Corporate Wellness Program Provider

The first case study is about a company in the healthcare industry in Malaysia, which collects health-related data from corporate company employees to predict the probability of

obesity-related issues for the employees. There have been studies on analytics in the healthcare industry, which establish that benefits of using data analytics for healthcare are in the form of prediction of diseases (Chen, Hao, Hwang, Wang, & Wang, 2017; Kaur & Kumari, 2019); optimisation of pharmaceutical outcomes (Hernandez & Zhang, 2017), prediction of surgery outcomes such as bariatric surgery (Johnston et al., 2019), and prediction of whole-body fat percentage and visceral adipose tissue mass (Swainson, Batterham, Tsakirides, Rutherford, & Hind, 2017).

3.1 Phase 1: Business Understanding

The company specializes in providing services such as corporate wellness program since its establishment in 1993. One of its core offerings is its corporate wellness program that promises to enhance employee health and wellness and to help build a healthy lifestyle culture at the workplace. The three main causes of chronic diseases in Malaysia is poor nutrition, physically inactivity, and unhealthy lifestyle. The company engages their own in-house professionals i.e. dietitians and nutritionists educating clients on healthy eating, fitness instructors guiding them on exercising, and panel doctors teaching healthy lifestyle choices. In general, the program is to educate and promote healthy living among employees. The conventional measurement of obesity utilizes the body mass index (BMI) criterion. Although there are benefits to this method, there is a concern that using BMI, not all individuals at risk of obesity-associated medical conditions are being accurately identified. By calculating a person's BMI, the system can detect if a person is either underweight, normal, overweight, or obese. The BMI calculation is commonly used by many experts including doctors and researchers as the calculation is relatively easy. However, the accuracy of the BMI in determining a person's obesity level has been questioned since it only measures one's weight and height. Yet, there are no alternative formulae in calculating the obesity of a person.

More recent research has established that in order to determine whether a person is actually obese, overweight, or normal, other body composition analysis such as muscle mass, lean body mass, and fat mass should be taken into consideration. This is because different people are weighted differently (muscle-weight or fat-weight). It is found that whole-body fat percentage (PBF), and visceral adipose tissue (VFL), are correlated with

obesity-related disease trajectories, which are not fully accounted through BMI evaluation (Swainson et al., 2017).

In this case study, data analytics and machine leaning techniques are used to predict visceral fat level (VFL) based on body mass index (BMI), skeletal muscle mass (SMM), body fat mass (BFM), percentage body fat (PBF) and a few anthropometric measurements. The stakeholders involved are the corporate wellness team, the one responsible in conducting the program and also the employees of a client company. The company benefited from the solution presented in this project in which they are able to predict the obesity level of the employees by measuring not only the weight and height, but also the other body composition analyses such as body mass, skeletal muscle, body fat, and also visceral fat. As for benefits to the employees, the data had enabled them to receive more precise advice on their lifestyles so that their health is better maintained (Swainson et al., 2017; Wedell-Neergaard et al., 2019).

3.2 Data Understanding

The original dataset contains 1,079 rows and 83 columns with client demographic information and body composition analysis such as weight, height, level of protein, mineral and water in body, body mass index, body fat mass, skeletal muscle mass, percentage body fat and visceral fat level. Among all the variables and items, the target variable in this study is the visceral fat level (VFL). The data distribution consists of 377 males and 702 females with a percentage of 35% male and 65% female. An overview of the data is provided in Table 1 below.

Table 1: Overview of data

1. Name 2. ID 3. Height 4. Date of Birth 5. Gender 6. Age 7. Mobile Number 8. Phone Number 9. Zip Code	31. Lower Limit (SMM Normal Range) 32. Upper Limit (SMM Normal Range) 33. BMI (Body Mass Index) 34. Lower Limit (BMI Normal Range) 35. Upper Limit (BMI Normal Range)	61. Weight Control 62. BFM Control 63. FFM Control 64. BMR (Basal Metabolic Rate) 65. WHR (Waist-Hip Ratio) 66. Lower Limit (WHR Normal Range)

10. Address	36. PBF (Percent Body Fat)	67. Upper Limit (WHR Normal Range)
11. E-mail	37. Lower Limit (PBF Normal Range)	68. VFL (Visceral Fat Level)
12. Date of Registration	38. Upper Limit (PBF Normal Range)	69. Obesity Degree
13. Memo	39. FFM of Right Arm (Muscles segmentation)	70. Lower Limit (Obesity Degree Normal Range)
14. Test Date / Time	40. FFM% of Right Arm	71. Upper Limit (Obesity Degree Normal Range)
15. Weight	41. FFM of Left Arm	72. 20kHz-RA Impedance (Electrical flow points)
16. Lower Limit (Weight Normal Range)	42. FFM% of Left Arm	73. 20kHz-LA Impedance
17. Upper Limit (Weight Normal Range)	43. FFM of Trunk	74. 20kHz-TR Impedance
18. TBW (Total Body Water)	44. FFM% of Trunk	75. 20kHz-RL Impedance
19. Lower Limit (TBW Normal Range)	45. FFM of Right Leg	76. 20kHz-LL Impedance
20. Upper Limit (TBW Normal Range)	46. FFM% of Right Leg	77. 100kHz-RA Impedance
21. Protein	47. FFM of Left Leg	78. 100kHz-LA Impedance
22. Lower Limit (Protein Normal Range)	48. FFM% of Left Leg	79. 100kHz-TR Impedance
23. Upper Limit (Protein Normal Range)	49. BFM of Right Arm (Fats segmentation)	80. 100kHz-RL Impedance
24. Minerals	50. BFM% of Right Arm	81. 100kHz-LL Impedance
25. Lower Limit (Minerals Normal Range)	51. BFM of Left Arm	82. InBody Type
26. Upper Limit (Minerals Normal Range)	52. BFM% of Left Arm	83. Local ID
27. BFM (Body Fat Mass)	53. BFM of Trunk	
28. Lower Limit (BFM Normal Range)	54. BFM% of Trunk	
29. Upper Limit (BFM Normal Range)	55. BFM of Right Leg	
30. SMM (Skeletal Muscle Mass)	56. BFM% of Right Leg	
	57. BFM of Left Leg	
	58. BFM% of Left Leg	
	59. InBody Score	
	60. Target Weight	

3.3 Data Preparation

The dataset is explored to detect missing and incorrect values. In addition, other variables that are not relevant in obtaining the desired outcome were also eliminated, such as demographic information, the mineral, protein, total body water, waist-hip ratio (WHR), and the basal metabolic ratios (BMR). After eliminating the unused variables and missing data, the dataset was only left with 1,077 rows and 9 columns.

3.4 Modelling

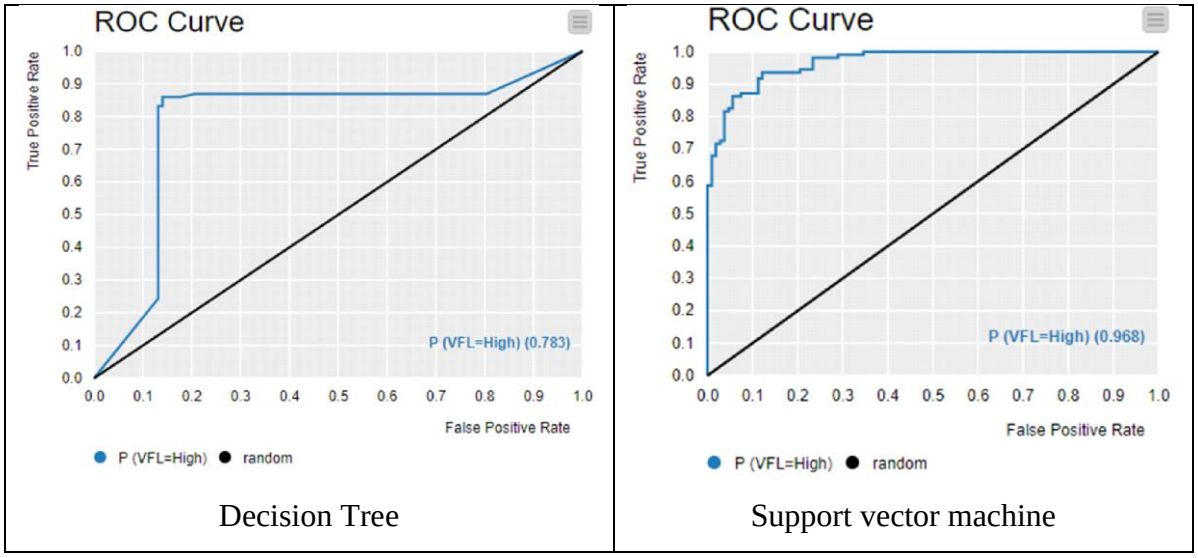
Supervised learning method is used to build the prediction model. Two machine learning techniques, i.e. decision tree and support vector machine (SVM) were applied to build the predictive models.

3.5 Evaluation

The evaluation phase involves comparing the performance measure for the decision tree and the SVM, and selecting the best method to accurately predict visceral fat level (VFL). The performance evaluation comparison metrics were precision, sensitivity, accuracy and the ROC curve. Table 2 and Figure 2 below illustrate the comparison of the two predictive models showing the SVM outperforming the decision tree.

Table 2: Evaluation of predictive models

Performance Evaluation Measure	Decision Tree	Support Vector Machine
Precision	0.857	0.922
Sensitivity	0.841	0.872
Accuracy	0.85	0.898



Decision Tree | Support vector machine

Figure 2: Evaluation of predictive models using the ROC curve

3.6 Deployment

Based on the performance evaluation metrics, the machine learning technique SVM performs better prediction of VFL, and hence is recommended for use in the deployment phase. In this case study, the stakeholders that involved is the corporate wellness team, the one responsible in conducting the program, and also the bank employees. The company would benefit from the solution presented in the project in which they are able to predict the obesity level of the bank employees by measuring not only the weight and height, but as well as the other body composition analysis such as body mass, skeletal muscle, body fat, and also the visceral fat. As for the bank employees, they will receive a more precise advice on the lifestyle they are supposed to perform so that their health is maintained.

4. Case study 2 – Market Basket Analysis of a Convenience Store

The second case study is about a convenience store whereby the owner had initiated plans on having personalised product recommendation to his customers in the effort towards customer satisfaction and retention. With regard to the highly competitive retail industry in which the convenience store is operating, there are a noteworthy number of data analytics applications. One of the more popular applications is Market Basket Analysis, which uncover the useful patterns that are hidden in retail transaction data (Griva, Bardaki, Pramatari, & Papakiriakopoulos, 2018). In the retail business, customer satisfaction in their

shopping experience is vital to attract and retain customers for the longer term. Retailers have long recognized that data driven decision-making have risen in the agenda of many businesses. Retailers need to find innovative ways to enhance the customer shopping experience, and one approach is by gaining customer insights from customer data and shopping behaviour patterns. Hence, retailers can customize their offerings and services suit to the customer shopping needs and preferences (Halim, Halim, & Felecia, 2019; Musalem, Aburto, & Bosch, 2018).

4.1 Business Understanding

The convenience store provides retail sales in food and beverages, household items, and local products from small medium enterprise (SME) companies. The store was in need of business strategies to stay competitive in the market and it had turned to data analytics to better understand its customer needs and preferences. The main aim was to find relationships between products and identify customers' purchasing patterns. The project had involved studying customer behaviour by looking at products purchased by shoppers. It also examines the association between product or item purchases. Through the project, the company had managed to determine and predict customer behaviour through their purchasing patterns, especially the association between products purchased. Findings from the data analytics had served as the basis for decision making on marketing activities, promotional pricing, product placements, and inventory management.

4.2 Data Understanding

For the project, data is sourced from two tables consisting of (i) transaction data and (ii) product name and category data. The transaction data has 57,106 transaction item entries with variables such as method of payment, date and time of transaction, transaction identification, quantity of transaction, item purchased and total item price. The product category data set has 2068 product entries belonging to 96 types of product categories. Some examples of categories available in this data set are groceries, biscuits, frozen, food services, dry items, drinks, phone accessories and cigarettes. The initial exploratory data analysis had detected several data problems such as:

1. Multivalued entries: a single transaction is recorded as multiple entries due to each entry containing one item with its quantity purchased.

2. Negative price values: certain entries record prices as having negative values due to the possibility of refunds for certain products such as unfilled fuels.
3. Missing values: 2827 entries contain empty/missing item names.
4. Item name truncation: item names in the transaction data set are truncated, which makes it harder to match items with its categories from the product category data set.
5. Trailing whitespaces: some values contain hidden trailing whitespaces that may interfere with the grouping of similar values.

4.3 Data Preparation

A significant portion of data preparation work is conducted prior to performing descriptive analysis of the data. First, the initial transaction data set is modified to contain descriptive column names as the original data did not have any names or descriptions for each column. Then, columns containing data that are not useful or have an unknown purpose are removed from the data set. Meanwhile, a string manipulation node is set up to match items to categories from the item category data set based on the first 15 letters of the item name. This approach is used due to the truncation of item names in the transaction data set. Following that, the processed transaction and item category data sets are passed into a rule engine dictionary node that performs the matching and appends a new column for the item's category. Then, additional string manipulation is performed to remove trailing whitespaces from item names and entries with missing item names are removed from the data set.

In its current state, the data set has a single entry for each specific product purchased regardless of quantity, with a numeric quantity attribute being present in the same row. In order to correctly supply the correct number of items purchased per transaction to the machine learning technique association rule learner in the modelling stage, the number of item entries must be duplicated to match the purchased quantity within the specified transaction. This is achieved by using nested loops; the outer loop iterates over each row and extracts the quantity attribute from the row as a variable, and the inner counting loop duplicates the row a certain number of times based on the quantity variable. Following the completion of the outer loop's iterations, the quantity attribute is removed from the data set as it is no longer needed.

Based on the preliminary descriptive analysis conducted on the data sets, the top ten categories of products purchased at the convenience store are Primax 95 fuel, bread/buns/cakes, diesel fuel, Primax 97 fuel, prepared food, water, cigarettes, mineral water and bottled or canned drinks. Meanwhile, we also explored the data to find product categories that were least purchased in the store. The top ten least purchased products categories are basic groceries, other groceries, snacks, household accessories, hair care, instant food, lubricants, ointments/plasters, household fragrances and canned food.

4.4 Modelling

In order to uncover patterns in the transaction data, the machine learning technique for association rule mining was applied. The application is also referred to as the Market Basket Analysis, which is a data mining method that examines large transactional databases to determine items that are most frequently purchased together. Association rule mining is an important component of the analytical system in the retail sector to determine the placement of goods and the design of sales promotions for different segments of customers to improve customer satisfaction and hence the profits of the store (Dwija Utama et al., 2020).

Using the prepared data, an association rule learner node is used within KNIME in order to mine for and discover association rules within the transaction data. The association rule learner utilises the *a priori* algorithm. After experimenting with several settings, the learner is configured to have a minimum set size of 1, a minimum support of 0.1% and a minimum rule confidence of 10%.

4.5 Evaluation

From the obtained results, rule support values range between 2 to 43 transactions, which is approximately equivalent to the range of 0.016% and 0.349% of transactions. Meanwhile, rule confidence percentages range from the pre-set minimum of 10% to a maximum of 36.4%. The low amount of rule confidence may be due to the fact that the minimum support setting needs to be set at a higher value as a result of the relatively small number of transactions being processed; results may differ if the modelling was performed on a larger data set spanning longer time periods. Hence, to provide a better measure for the impact of each rule on the convenience store's sales figures, the rule revenue is used as a metric. Rule revenue values range between RM1.80 to RM38.70 depending on the value of the identified

antecedent products. In summary, the model mostly discovers rules that have antecedent-consequent item relations that can be easily described due to the nature of the purchased items. However, the model is also found to be capable of uncovering rules with pairs of items that do not necessarily make sense at a glance, but have the data to support the existence of the rule.

4.6 Deployment

The discovered association rules can be utilised in a number of ways by the store's management team. First, data on the associations between products can be used to drive merchandising strategies. One such strategy that can be performed based on the information is the rearrangement of the store layout to position products with strong associations closer together in order to influence customers to pick all of the associated products while shopping in the store. This may cause customers to include the consequent products in their purchases based on association alone, even though they did not initially plan to purchase the product. This will then translate to an increase in convenience store sales and profits that may not have been obtained if the products were placed in separate, distant locations.

Second, the discovered association rules can also be used to select products to be bundled together for in-store offers and promotions. For instance, the bread and coffee product combination mentioned in the "Evaluation" section can be marketed as a breakfast bundle, prompting the customer to purchase both at the same time. Furthermore, additional value can be added to the bundle by offering the bundle at a cheaper price compared to the usual total price of the two items. This will further incentivise customers to purchase both antecedent and consequent products at the same time due to the perceived increase in value. Alternatively, if the store owner has limited control over the promotions that can be designed, store employees can instead offer the consequent item as an additional purchase option to customers who are checking out with the antecedent item. This upselling strategy may also have the same effect of influencing customers to purchase the consequent item, thereby netting more sales for the store.

Third, association rules can also be beneficial for demand planning. By having the knowledge of items that are related to each other, store owners can make decisions to order more of a consequent product if the demand for the antecedent products are higher.

Additionally, the improvements made in demand planning can be leveraged to create promotions and drive merchandising strategies without worrying about a supply shortage.

5. Discussion and Conclusion

This paper shows how individual entrepreneurs and small and medium sized organizations can leverage data analytics and machine learning techniques to gain competitive advantage from the data. The respective benefits gained by the small business owners featured in the two case studies provide answers to the research question on how the use of data analytics using machine learning techniques can benefit small businesses.

As data increases, companies are looking for ways to gain relevant business insights underneath layers of data and information. Data analytics using machine learning techniques provide the methods and tools that enable companies to retrieve hidden gems to better understand new business ventures, opportunities, business trends and complex challenges. Data mining solutions support intelligent decision-making through intuitive interfaces for users to learn patterns in data, potential business ventures, insight into their customers' buying habits, detection of discrepancies and problems that might otherwise remain obscured.

Acknowledgement

The authors would like to thank the **Centre of Business Excellence, Faculty of Management** for their initiatives to write case studies.

References

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction by Machine Learning over Big Data from Healthcare Communities. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.2694446

Dekimpe, M. G. (2020). Retailing and retailing research in the age of big data analytics. *International Journal of Research in Marketing*, *37*(1), 3–14. https://doi.org/10.1016/j.ijresmar.2019.09.001

Dwija Utama, I., Diryana Sudirman, I., & Alamsyah, D. P. (2020). Optimizing the Products Combination by Using Data Mining Market Basket Analysis Approach. *Journal of Critical Reviews*, *7*(19), 5085–5093. Retrieved from

http://www.jcreview.com/fulltext/197-1598174408.pdf?1602220537

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2018.01.029

Halim, K. K., Halim, S., & Felecia. (2019). Business intelligence for designing restaurant marketing strategy: A case study. In *Procedia Computer Science* (Vol. 161, pp. 615–622). Elsevier B.V. https://doi.org/10.1016/j.procs.2019.11.164

Hernandez, I., & Zhang, Y. (2017). Using predictive analytics and big data to optimize pharmaceutical outcomes. *American Journal of Health-System Pharmacy*. https://doi.org/10.2146/ajhp161011

Johnston, S. S., Morton, J. M., Kalsekar, I., Ammann, E. M., Hsiao, C. W., & Reps, J. (2019). Using Machine Learning Applied to Real-World Healthcare Data for Predictive Analytics: An Applied Example in Bariatric Surgery. *Value in Health*, *22*(5), 580–586. https://doi.org/10.1016/j.jval.2019.01.011

Kaur, H., & Kumari, V. (2019). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2018.12.004

Lu, J. (2020). Data Analytics Research-Informed Teaching in a Digital Technologies Curriculum. *INFORMS Transactions on Education*, *20*(2), 57–72. https://doi.org/10.1287/ited.2019.0215

Malik, M. M., Abdallah, S., & Ala'raj, M. (2018, November 1). Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Annals of Operations Research*. Springer New York LLC. https://doi.org/10.1007/s10479-016-2393-z

Musalem, A., Aburto, L., & Bosch, M. (2018). Market basket analysis insights to support category management. *European Journal of Marketing*. https://doi.org/10.1108/EJM-06-2017-0367

Raguseo, E. (2018). Big data technologies: An empirical investigation on their adoption, benefits and risks for companies. *International Journal of Information Management*, *38*(1), 187–195. https://doi.org/10.1016/j.ijinfomgt.2017.07.008

Swainson, M. G., Batterham, A. M., Tsakirides, C., Rutherford, Z. H., & Hind, K. (2017). Prediction of whole-body fat percentage and visceral adipose tissue mass from five

anthropometric variables. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0177175

Wedell-Neergaard, A. S., Lang Lehrskov, L., Christensen, R. H., Legaard, G. E., Dorph, E., Larsen, M. K., … Krogh-Madsen, R. (2019). Exercise-Induced Changes in Visceral Adipose Tissue Mass Are Regulated by IL-6 Signaling: A Randomized Controlled Trial. *Cell Metabolism*. https://doi.org/10.1016/j.cmet.2018.12.007