Option #2: Using Data Analytics in a Hospital Setting

Scott Miner

Colorado State University – Global Campus

Option #2: Using Data Analytics in a Hospital Setting

A hospital is experiencing new incoming patients around the fall season of each year with a common virus. The hospital currently has no means to predict the amount of medicine needed on hand for the successive fall season to combat the illness (*Module 8: Portfolio Project*, 2020).

Definition of the Data Problem

The data problem the hospital is facing is the need to predict the amount of medicine to keep on hand each year. Additionally, the hospital needs to combine various disparate data sources, most likely spread across different systems, to find a solution to the data problem. Milovic and Milovic (2012) emphasize that one of the biggest challenges in healthcare data mining (DM) is data being voluminous and heterogeneous. Comprehensive data consists of numerous components, including interviews with patients, laboratory results, demographic information, and doctors' diagnoses. Koh and Tan (2011) point out these components are often lacking in healthcare settings since data frequently reside in different systems, and data warehouses (DWs) are often non-existent. Other challenges include data having missing, non-standardized, or corrupted values and modelers having to incorporate knowledge from several different staff within various departments (Koh and Tan, 2011).

The hospital, therefore, needs to utilize multiple team members, including domain experts (e.g., doctors and nurses), database experts, and predictive modeling experts. Domain experts may not always be familiar with the technical aspects of statistical modeling yet are needed to outline the data problem effectively. Likewise, database experts are necessary to identify the available data and explain how to access, join, and normalize the tables. Lastly, predictive modeling experts build predictive models and compare and evaluate the results to determine the best models for the solution (Abbott, 2014). Forming a cohesive unit comprised of these

members will allow the hospital to solve the data problem using predictive analytics successfully. Such a solution will allow for better inventory management year after year, leading to reduced costs annually, as well as better patient diagnoses and treatments, ultimately reducing patient expenses as well.

The amount of incoming data in healthcare has increased exponentially over the last few years. According to Fayyad et al. (1996), just 25 years ago, it was common to see medical diagnostic databases containing millions of rows with column counts in the thousands. Medical databases have only continued to grow. DesRoches et al. (2013), describe how in 2009, the Health Information Technology for Economic and Clinical Health (HITECH) Act authorized $30 billion for use in transitioning healthcare systems from paper-based to electronic health record (EHR) keeping. Menger et al. (2016) describe the benefits of a paperless EHR system, one such advantage being the capability to extract knowledge from datasets using various DM techniques.

Introduction to the CRISP-DM Methodology.

Fayyad et al. (1996) write that statisticians, data analysts, and others within the management information systems (MIS) communities often refer to DM as the process of extracting useful knowledge from datasets. Fayyad et al. present a method known as *Knowledge Discovery in Databases* (KDD) where DM is but a single step of five: (a) selection, (b) preprocessing, (c) transformation, (d) data mining, and (e) interpretation/evaluation. Figure 1 displays the five steps of the KDD process. Azevedo and Santos (2008) compare the KDD process to the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology developed by the SAS Institute. The researchers confirm the two methods are equivalent in the steps they employ, using only different names for the steps. Vleugel et al. (2010) compare these methodologies to the Cross-Industry Standard Process for Data Mining, otherwise known as

CRISP-DM. The CRISP-DM procedure, they write, was developed out of the need to consider DM from the business point-of-view. The CRISP-DM technique consists of six phases: (a) Business Understanding, (b) Data Understanding, (c) Data Preparation, (d) Modeling, (e) Evaluation, and (f) Deployment (Chapman et al., 2000). The difference between the CRISP-DM approach and the two techniques described above lies within steps (a) Business Understanding and (f) Deployment. The two steps exist only within the CRISP-DM methodology and are absent from both KDD and SEMMA methodologies (Vleugel et al., 2010).

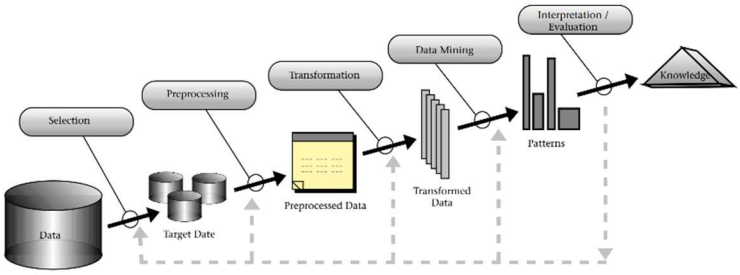

Figure 1. Five steps of the KDD process. Reprinted from "From Data Mining to Knowledge Discovery in Databases," by Fayyad et al., 1996, *AI Magazine*, 17(3), (p. 41).

Literature Review

Komenda et al. (2020) used the CRISP-DM methodology to analyze and visualize data via an interactive web-based dashboard to track the spread of the COVID-19 epidemic in the Czech Republic. After considering both the SEMMA and CRISP-DM models, the researchers ultimately decided upon the CRISP-DM methodology due to its adaptability. They describe the power of the approach as being demonstrated by its emphasis on the correct understanding of organizational processes and the proper implementation of all six phases. Moreover, CRISP-DM is an iterative process that allows analysts to return to previous steps should the need arise. Chapman et al. (2000) verify the dependent nature of the steps and explain that moving back and forth between phases is necessary. The outcome of one stage determines which comes next. Figure 2 displays the six stages of the CRISP-DM lifecycle and the cyclical nature of the

process. The arrows between steps indicate where dependencies are most likely to occur. Lastly, the CRISP-DM technique highlights the importance of checking one's discoveries before publication, a measure frequently omitted in other strategies (Komenda et al., 2020).

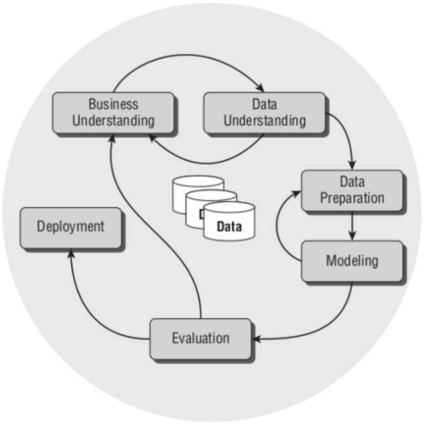

Figure 2. The six phases of the CRISP-DM process (Abbott, 2014, p. 21)

Likewise, Marcelino et al. (2015) used CRISP-DM to identify potential physiological pattern aberrations in senior populations to assess possible risk circumstances. They describe CRISP-DM as being the most widely used DM methodology. The eServices platform developed for the study potentiates constructing a considerable dataset, as the platform incorporates biosensor data, environmental data, and a log resultant from seniors' use of the platform. The platform provides various health and social benefits, including virtual doctor appointments, online games, and the ability to document and share experiences. Using the CRISP-DM methodology and C5.0 decision trees, Marcelino et al. found a solution that decreases false negatives, providing a 0.85 accuracy rate in assessing hazardous circumstances in vulnerable populations.

Menger et al. (2016) used the CRISP-DM methodology to analyze a case study conducted in the psychiatry department of the University Medical Center Utrecht. Menger et al. proposed a modification to the CRISP-DM approach termed CRISP-IDM: Cross-Industry Standard Process for *Interactive* Data Mining. The interactive approach utilizes a data visualization tool that incorporates direct feedback from the local workforce in *all* stages of the modeling process. As a result, the healthcare staff no longer felt alienated by the more technical

aspects of the modeling and evaluation phases, and two of the outcomes from the study were consequently implemented on the workplace floor and promoted by staff (Menger et al., 2016).

Lastly, Poucke et al. (2016) again used the CRISP-DM methodology, as well as interactive visual environments adapted for the medical community, to quantitatively assess platelet counts and Intensive Care Unit (ICU) survival rates. The researchers describe the goal of predictive analytics within healthcare as providing clinical staff with decision support that leads to "Predictive, Preventive, and Personalized Medicine (PPPM)" (p. 2). PPPM ultimately leads to reduced patient costs and increased quality of care. The challenge lies in integrating the multiple large-scale heterogeneous data sources that reside within the medical community: demographic data, administrative data, data on health risks and statuses, patient medical history data, current health management data, and outcome data (Lohr & Donaldson, 1994).

Poucke et al. (2016) extracted 33 flat files for each of 11,994 patients admitted to one of five ICUs between 2001-2008 from the MIMIC-II (Multiparameter Intelligent Monitoring in Intensive Care) database. The MIMIC-II database is a comprehensive medical database containing patient demographic information, admission and discharge dates, ICD-9 codes, death dates, healthcare types and providers, and physiological information captured from hospital monitors (*MIMIC-II Databases, 2018).* The providers of the database have removed all protected health information (PHI) per HIPAA regulations and grant access only to legitimate researchers as authorized by Physionet.org (Goldberger et al., 2000).

Poucke et al. (2016) assimilated the flat files into a distributed Hadoop server. Barua and Mondal (2019) describe Hadoop as a framework that has, as its foundation, the Map-Reduce architecture, capable of utilizing parallelism. Figure 3 shows a logical architecture diagram of Map-Reduce. Figure 4 displays an overview of cloud computing services, which implement

structures like Map-Reduce and Hadoop. Barua and Mondal describe parallelism as distributed nodes completing algorithmic tasks on multiple partitions of a dataset simultaneously. Processing the data concurrently, as opposed to serially, allows the algorithms to execute much faster than in traditional computing environments (Barua & Mondal, 2019).

Poucke et al. (2016) used RapidMiner, a visual platform with a user-friendly GUI, to build predictive models on the dataset. Additionally, RapidMiner incorporates the ability to run R and Python scripts, if needed. The predictive algorithms that Poucke et al. used include Naïve Bayes, Decision Trees, Logistic Regression, Support Vector Machines, and Ensemble methods, including Random Forests, Boosting, Bagging, and Stacking. The results of the evaluation phase indicate that the AUPRC (Area Under the Precision Recall Curve) measure was highest (AUPRC = 0.764) for the Random Forest algorithm implementing the feature selection technique of Backward Elimination. AUPRC is typically used when data is unequal. The measure calculates the "fraction of negatives misclassified as positives" (p. 14). Other standard measures adopted in the evaluation phase include the AUC (Area Under the ROC Curve) and Accuracy. These measures, however, can be deceptive when data is unbalanced (Poucke et al., 2016).

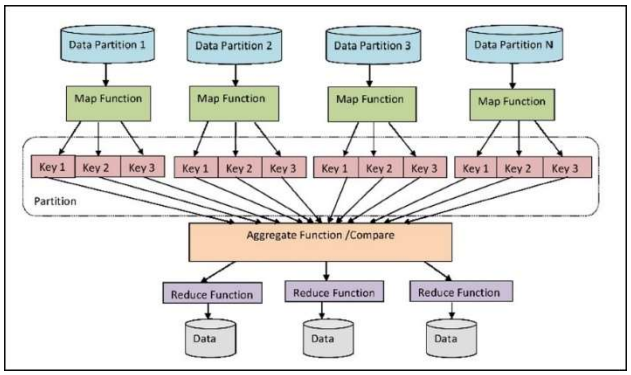

Figure 4. Overview of the Map-Reduce Framework. Reprinted from "A Comprehensive Survey on Cloud Data Mining (CDM) Frameworks and Algorithms," by Barua, H. B., & Mondal, K. C., 2019, *ACM Computing Surveys*, 52(5), (p. 8).

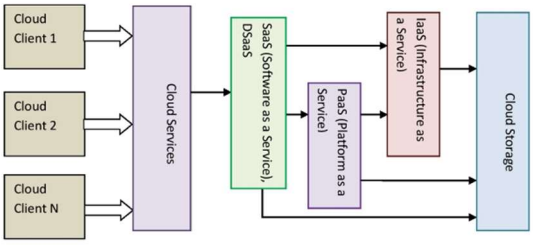

Figure 3. Overview of cloud computing services. Reprinted from "A Comprehensive Survey on Cloud Data Mining (CDM) Frameworks and Algorithms," by Barua, H. B., & Mondal, K. C., 2019, ACM Computing Surveys, 52(5), (p. 3).

A solution to the Hospital Problem using Predictive Analytics

The above studies provide a basis for solving the hospital problem using predictive analytics and the CRISP-DM methodology.

Business Understanding.

The first stage in the CRISP-DM life cycle is business understanding. In this phase, it is crucial to determine business objectives, success criteria, available resources, project requirements, and data mining goals (Chapman et al., 2000). Additionally, in this stage, modelers discover the modeling methods and how to deploy the models (Abbott, 2014). Modelers may choose to interview domain experts, as outlined in the study by Menger et al. (2016). When determining useful tools to aid in the analysis, analysts should recall the studies performed by Menger et al. and Poucke et al. (2016). Each group of researchers utilized visual analytic tools with user-friendly interfaces to aid practitioners and clinical end-users in their abilities to understand the modeling processes. SAS Enterprise Miner is like RapidMiner in its visual presentation. For instance, each software allows modelers to process big data in a code-free environment via drag-and-drop nodes. Either the functions of SAS Enterprise Miner or RapidMiner would be useful in aiding analysts in their abilities to solve the hospital data problem. Each software offers a GUI and the ability to build predictive algorithms in code-free environments (Poucke et al., 2016; Georges & Andersson, 2017).

Additionally, each software allows modelers the capability to expand their settings using R or Python scripting techniques, should the need arise (Poucke et al., 2016; Georges & Andersson, 2017). Moreover, like RapidMiner, SAS Enterprise Miner offers the capability to implement high-performance computing nodes that operate on distributed datasets. Examples of predictive algorithms with high-performance computing nodes in SAS Enterprise Miner include

Bayesian networks, *k*-means clustering, forests, neural networks, and support vector machines (Baxter & Huddleston, 2016).

The next step in the Business Understanding phase is to determine the unit of analysis and target variables. The unit of analysis in the hospital problem is a hospital patient. In determining target variables, it is often useful to predict both categorical and numeric outputs. In the hospital problem, the categorical target variable determines whether a patient was diagnosed with the virus. A value of 1 reflects a concrete diagnosis, while the value 0 represents the absence of a diagnosis. The numerical, continuous-valued output variable, on the other hand, is the amount of medication needed in milliliters (ML). Using these two variables, modelers can then calculate a score for each record as such: $Score = P(Diagnosed = 1) \times Estimated\ Medication\ Amount\ in\ mL$ (Abbot, 2014).

Abbott (2014) writes that to assess model accuracy in classification problems, analysts use the Percent Correct Classification (PCC) measure along with confusion matrices. For continuous-valued estimation problems, analysts use the following metrics: R^2, average error, Mean Squared Error (MSE), median error, average absolute error, and the median absolute error. These values constitute the success criteria. To determine the model selection and evaluation criteria, modelers use the multiplicative method described above to determine the predicted cumulative pharmaceutical amount. Modelers subtract this cumulative value from the aggregated actual pharmaceutical amount. The model that provides the smallest difference between these two accumulated values, without resulting in a negative outcome, allows the hospital to purchase the most cost-efficient amount of medicine and still treat all infected patients as needed (Abbott, 2014).

Data Understanding.

In the data understanding phase, analysts begin with an initial data collection and analyze the data to gain familiarity with the dataset. Then, modelers describe, explore, and verify the quality of the data. This phase allows analysts to identify data problems, including quality issues or missing values, uncover initial understandings, and identify data subsets that warrant further analysis. This phase is a modeler's first encounter with the dataset, where the practitioner performs the initial preparatory work needed for further analysis (Chapman et al., 2000).

Data Preparation.

The Data Preparation phase is often the most prolonged phase of the project, where modelers build the final dataset used for analysis from the raw data. Tasks in this phase include analyzing and selecting essential variables. In the hospital problem, examples of significant features include patients' age, weight, height, demographic information, and other details pertinent to a healthcare dataset. Additional tasks in this phase include outlier detection, as well as attribute removal, addition, and discretization. Here, analysts remove noisy or redundant data identified in the exploration process of the Data Understanding phase. Further steps include cleaning, integrating, and reformatting the data to build the final dataset delivered to the modeling tool (Chapman et al., 2000).

Modeling.

The modeling phase is where practitioners select modeling techniques, generate test designs, build the predictive models, and assess what they have created. In this stage, it is often necessary to return to the Data Preparation stage, because specific algorithms sometimes require particular inputs. In the hospital problem, analysts build models to predict patient diagnoses, as well as the amount of medicine needed to combat the virus. Examples of predictive classifier

algorithms include Decision Trees (DTs), Artificial Neural Networks (ANNs), Logistic Regression, k-Nearest Neighbors (k-NN), Naïve Bayes (NB), and Support Vector Machines (SVMs). To predict patient diagnoses, modelers can use these classification algorithms. Furthermore, ensembles typically produce more accurate results than single models and are formed by combining predictions from two or more single models. Examples of model ensembles include bagging, boosting, random forests, and heterogeneous ensembles, which are composed of varying algorithms (Abbott, 2014).

Many classification algorithms possess regression counterparts that predict continuous-valued outcome variables, including DTs, ANNs, k-NN, and SVMs. The only difference between classification ANNs and regression ANNs, for instance, is the output layer nodes. In classification ANNs, the output layer nodes implement a sigmoidal activation function, while in regression ANNs, a linear activation function is used in the output layer nodes. Both versions of the algorithm allow for an extremely flexible, powerful model capable of predicting nonlinear outputs from inputs. ANNs, like other models requiring numeric explanatory variables, including linear regression, logistic regression, k-NN, and SVMs, necessitate the conversion of categorical variables to 1/0 dummy variables and the imputation of missing values (Abbott, 2014).

DTs are one of the most popular DM algorithms and are capable of handling missing values automatically, as well as ordinal, nominal, and continuous predictor variables. Naïve Bayes, on the other hand, requires all inputs to be categorical. Additionally, in the modeling stage, analysts fine-tune the parameter settings. Examples of DT parameters include the maximum depth, the minimum number of records allowed in leaf nodes, and the minimum number of records permitted in parent nodes. Neural Network settings, on the other hand,

include the learning rate, the momentum, the number of epochs, the number of hidden layers, the number of neurons in each hidden layer, and the stopping criteria (Abbott, 2014).

Evaluation.

In the evaluation phase, practitioners assess the results to determine the best model to meet the business objective. In the hospital example, the business objective is to determine the model that best predicts the amount of medication needed in the successive fall season to ensure the effective treatment of all patients and to reduce inventory overstock and costs where possible. To calculate the predicted score for each record, modelers use the formula $Score = P(Diagnosed = 1) \times Estimated\ Medication\ Amount\ in\ mL.$ Practitioners sum and subtract the predicted scores from the accumulated actual values. After reviewing the steps needed to construct the model and ensuring it achieves the business objectives, practitioners choose the model that most closely matches the set threshold when the cumulative predicted score is subtracted from the aggregated actual amount of medicine (Abbott, 2014).

Deployment.

In the final phase, the deployment phase, analysts present their findings to key stakeholders. The stakeholders can then decide to purchase the amount of medication the model recommends. In subsequent analyses, the hospital needs only to compute model scores and compare them to thresholds defined in the evaluation phase to determine the amount of medicine necessary for the following fall season (Abbott, 2014). This phase additionally includes monitoring and maintaining the deployment plan, producing the final report, and documenting the experience (Chapman et al., 2000).

Summary and Conclusion

The hospital data problem presents a challenging problem that nevertheless can be solved effectively by implementing the CRISP-DM methodology along with a competent team comprising domain experts, database experts, and predictive modeling experts. The most critical stage of the CRISP-DM approach is the Business Understanding stage, where the business objectives are defined, along with the unit of analysis, target variable(s), success criteria, modeling objectives, and modeling selection and evaluation measures (Abbott, 2014). By working with domain experts up-front to gain a thorough understanding of these crucial components, modelers can successfully solve the hospital problem despite the challenges inherent to the industry. These challenges include numerous heterogeneous data sources, absent DWs, and data sources that contain many missing or invalid values (Koh and Tan, 2011).

In this example, we built predictive models within a visual analytics tool like SAS Enterprise Miner or RapidMiner to solve the hospital data problem. We implemented multiple predictive algorithms, including DTs, Linear and Logistic Regression, ANNs, k-NNs, NBs, and SVMs, to predict target variables including whether a patient was diagnosed with the infection, as well as the amount of medicine the hospital needs to keep on hand for the following fall season. By computing a score based on these values, we created a cumulative predictive score, which we then subtracted from the aggregated actual amount of medicine. The result of this subtraction, calculated in the evaluation phase, informed us which model was the best predictor to solve the hospital data problem. The best predictor is the model nearest the threshold set in the evaluation phase (Abbott, 2014). The threshold amount should reflect the business objective, namely, to provide treatment to all infected patients and reduce the overstock costs associated with the pharmaceutical supply. Finally, in the deployment phase, we presented our results to key decision-makers and stakeholders.

References

Abbott, D. (2014). Applied predictive analytics: principles and techniques for the professional data analyst. Indianapolis, IN: Wiley.

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. 6.

Barua, H. B., & Mondal, K. C. (2019). A Comprehensive Survey on Cloud Data Mining (CDM) Frameworks and Algorithms. *ACM Computing Surveys,* 52(5), 1–62. https://doi.org/10.1145/3349265

Baxter, A., & Huddleston, E. (2016). *SAS Help Center: Overview of the SAS Enterprise Miner High-Performance Procedures*. Retrieved May 28, 2020, from https://documentation.sas.com/?docsetId=emhpprcref&docsetTarget=emhpprcref_intro_sect001.htm&docsetVersion=14.2&locale=en

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide, USA: SPSS Inc. *CRISPWP*-0800.

DesRoches, C. M., Charles, D., Furukawa, M. F., Joshi, M. S., Kralovec, P., Mostashari, F., Worzala, C., & Jha, A. K. (2013). Adoption of Electronic Health Records Grows Rapidly, But Fewer Than Half of US Hospitals Had At Least A Basic System In 2012. *Health Affairs*, 8.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* 17(3), 37—54.

Georges, J., & Andersson, C. (2017). *Advanced Predictive Modeling Using SAS® Enterprise Miner™ Course Notes*. Cary, NC: SAS Institute. Retrieved June 6, 2020, from https://coursecms.csuglobal.edu/file/af7b20c0-6082-49ee-8997-

9bb3f672ea82/2/production/MIS530_ModuleLectures/media/MIS530_Advanced_Predictive_Modeling_Using_SAS_Enterprise_Miner_Course_Notes.pdf

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *circulation*, 101(23), e215-e220.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, 19(2), 65.

Komenda, M., Bulhart, V., Karolyi, M., Jarkovský, J., Mužík, J., Májek, O., Šnajdrová, L., Růžičková, P., Rážová, J., Prymula, R., Macková, B., Březovský, P., Marounek, J., Černý, V., & Dušek, L. (2020). Complex Reporting of the COVID-19 Epidemic in the Czech Republic: Use of an Interactive Web-Based App in Practice. *Journal of Medical Internet Research*, 22(5), N.PAG. https://doi.org/10.2196/19367

Lohr, K. N., & Donaldson, M. S. (Eds.). (1994). *Health data in the information age: Use, disclosure, and privacy.* National Academies Press.

Marcelino, I., Lopes, D., Reis, M., Silva, F., Laza, R., & Pereira, A. (2015). Using the eServices Platform for Detecting Behavior Patterns Deviation in the Elderly Assisted Living: A Case Study. *BioMed Research International*, 2015, 1–10. https://doi.org/10.1155/2015/530828

Menger, V., Spruit, M., Hagoort, K., & Scheepers, F. (2016). Transitioning to a Data Driven Mental Health Practice: Collaborative Expert Sessions for Knowledge and Hypothesis Finding. *Computational & Mathematical Methods in Medicine*, 1–11. https://doi.org/10.1155/2016/9089321

Milovic, B., & Milovic, M. (2012). Prediction and decision making in health care using data mining. *Kuwait Chapter of the Arabian Journal of Business and Management Review*, 1(12), 126.

MIMIC-II Databases. (2018, September 18). MIMIC-II Databases. https://archive.physionet.org/mimic2/

Module 8: Portfolio Project. (2020). Retrieved June 6, 2020, from https://csuglobal.instructure.com/courses/20979/assignments/427772

Poucke, S. V., Zhang, Z., Schmitz, M., Vukicevic, M., Laenen, M. V., Celi, L. A., & Deyne, C. D. (2016). Scalable Predictive Analysis in Critically Ill Patients Using a Visual Open Data Analysis Platform. *PLOS ONE*, 22.

Vleugel, A., Spruit, M., & van Daal, A. (2010). Historical Data Analysis through Data Mining from an Outsourcing Perspective: The Three-Phases Model. International Journal of Business Intelligence Research (IJBIR), 1(3), 42-65. https://doi.org/10.4018/jbir.2010070104