Data Analytics for HR Students: Using RapidMiner to Develop Systems Thinking Skills

Anna Czegledi and Ken Dafoe

Anna Czegledi is a Full Time Professor at Conestoga College School of Business, 299 Doon Valley Drive, Kitchener, Ontario N2G 4M4, Canada (Email: Aczegledi@conestogac.on.ca).

Ken Dafoe is a Partial-Load Faculty at Conestoga College School of Business, 299 Doon Valley Drive, Kitchener, Ontario N2G 4M4, Canada (Email: kdafoe@conestogac.on.ca).

Correspondence concerning this paper should be addressed to A. Czegledi.

Abstract

In this study, we investigated the value of the data mining tool RapidMiner for teaching data analytics to human resource (HR) students. In a way, the idea for another analysis tool contradicts the demand for demonstrable spreadsheet skills in business professions. Fluency in MS Excel is seen as a necessary skill for data analysis in business; however, more advanced analysis like predictive analytics (i.e., predicting employee churn, salary levels, survey text analysis, etc.) leaves a gap in MS Excel's skill set. Introducing
HR students to RapidMiner fills that gap. We wanted to provide students with a deeper learning experience beyond the spreadsheet to give them a competitive advantage in the job market. This study explained how faculty could use RapidMiner to teach data analytics to HR students. Based on their experience, students completed a survey that analyzed the impact of using RapidMiner on their engagement, learning satisfaction and understanding of Systems Thinking. We adapted active learning strategies and an HR data set from IBM to reduce the gap between business education and practice. We provided practical recommendations on using RapidMiner for teaching Data Analytics and as a method to develop fundamentals of systems thinking for business students. Our findings showed that using RapidMiner tool resulted in high student engagement, satisfaction, confidence and helped introduce the Systems Thinking approach.

**Keywords:** RapidMiner; data analytics; systems thinking; HR education; experiential learning

**DOI:** https://dx.doi.org/10.15239/j.brcacadjb.2023.13.01.ja02

**Data Availability:** Data used in this study are available upon request.

**INTRODUCTION**

Employers expect HR graduates to be able to analyze data, apply professional judgment and see "the big picture" to provide the relevant information at the right time for making the right decisions (KPMG, 2019). Human resource expenses could account for more than half of organizational expenditures, and it is critical that HR professionals provide value-added information (Society for Human Resource Management, 2017). As Human Resources continues to transform itself away from a purely administrative/managerial role and into a more strategic player within the organization, it faces a question: How to build the data skills it needs for a successful transformation? In this paper, we studied the
effects of introducing systems thinking through data analytic skills to HR students using the popular data mining tool RapidMiner. We selected RapidMiner for four reasons: cost, interface, repeatability, and its application to HR analytics.

First, like programming languages Python and R, RapidMiner has a free version of its Windows, Mac or Linux application which supports the analysis of up to 10,000 rows of data. RapidMiner also provides an academic version that students and educators can request (no fee) that allows loading up to 100,000 rows.

The second reason that sets RapidMiner apart as a tool for learning data mining skills is its interface. RapidMiner has options to drag and drop an integrated environment where students learn to create simple to complex data models by piping or connecting an exhaustive set of operators without entering a single line of code. Some students entering the Human Resource program lack data analytics and system thinking skills. It would be problematic to immerse them in the IDE (Integrated Development Environment) of a high-level programming language like Python or a language for statistical computing like R without any experience in data preparation, cleaning, structuring, or assembling. Teaching students about mapping an HR problem into a programming language and choosing the best analytic tool to solve the problem without any previous knowledge of working with data could be too challenging. We felt that actual learning would be supplanted by memorization of steps that would collapse when variations of the HR problem were introduced.

Thirdly is the issue of MS Excel. Many businesses use MS Excel as their go-to data analysis tool. However, ubiquity makes it less of choice and more of a policy decision. If MS Excel is the only data tool at your disposal, then every analysis must look like a spreadsheet. This is a problem. A spreadsheet is a visual environment for analyzing data, meaning that it becomes difficult to identify all the steps taken to convert the raw data into the final analyzed visualization. There is no clear path back to the raw data because the analysis often transforms the data as part
of the analysis process. If a student can only see the solution and not the iterative steps that produced the solution, then the learning process is affected. There becomes little chance for repeatability of a process because the data has been transformed without any clear path to its original state. In addition, MS Excel has other limitations. It is a good descriptive analytic tool, but falls short as a predictive tool and has no built-in text analytic capabilities without third-party add-ins, often by incurring associated costs.

Lastly, we needed to choose an HR problem with an applied solution. The integrated, modular approach to building a data model in RapidMiner makes it conducive to modelling the scenarios that we created for HR students. Using RapidMiner, our students could work with analytic operators in a graphical environment without memorizing more text-based parameters, which was an advantage to visualizing the process steps towards a solution.

RapidMiner also allowed us to focus on Systems Thinking by providing an integrated/ deep learning experience that included working with real business data. Our goal was to measure the impact of using RapidMiner on business students’ engagement, learning satisfaction, and understanding of interconnections (Systems Thinking).

**Prior Literature**

Data analytics is a critical element of human resource management as it helps to provide insights based on data-driven decision-making processes (Shet et al., 2021) and strategic organizational capability (Margherita, 2022; Fernandez & Gallardo-Gallardo, 2020). Human resource analytics (HRA) is an innovative practice in human resource management (HRM) (Huselid, 2018; Boudreau & Cascio, 2017; Levenson, 2017; Rasmussen & Ulrich, 2015). New technologies such as artificial intelligence, computational intelligence techniques, data mining, machine learning, and improved data-driven decision-making in HRM (Duan et al., 2019; Dwivedi et al.,
However, to get the full benefits of data analytics in HRA, it should be integrated with business requirements (Levenson, 2005, 2011; Angrave et al., 2016; Lawler et al., 2004; Boudreau & Ramstad, 2007).

Data Analytics could be defined as "the application of logical and statistical techniques to examine a dataset and to draw inferences from data, with the goal of data analytics to identify patterns or relationships within a dataset" (Teimourzadeh, et al., 2022, page 3). The conclusions from the data analytics would help with value-added decision-making. Analytics helps to identify meaningful patterns and insights and distribute the results to proper stakeholders (Lepenioti et al., 2019).

Data Analytics can be subdivided into three classifications: descriptive (What happened and why?), predictive (Will it happen again and why?), and prescriptive analytics (What should be done and why?). Descriptive analytics (also called business intelligence) are used most often in business and involve analyzing data through data aggregation plus data mining methods (Fahl, 2017). Although descriptive analytics are used in our assignment, it is really a predictive exercise. Using RapidMiner we introduce students to Decision Trees, Naive Bayes, Random Forests, and other predictive algorithms that are evaluated simultaneously in the AutoModel feature of RapidMiner. Predictive analytics aims to obtain information from data to predict behavior patterns and trends (Fahl, 2017). Students were required to analyze records about employee stay and churn to find patterns that predict the likelihood of current employees who may be contemplating leaving the company.

Effective application of Data Analytics skills requires a Systems Thinking approach from decision-makers. The term "systems thinking" refers to a management and operations approach where single business decisions are analyzed based on their systematic consequences (Vassallo, 2017). Systems Thinking believes that the world is systemic; as a result, a problem could be an emergent property of an interrelated whole (Flood, 2010). Monat and Gannon (2015) emphasized that System Thinking has
great power in solving complex problems that are not solvable using conventional reductionist thinking. Checkland (1981) identified that Systems Thinking focuses on the concept of wholeness (complete and not divided into parts) captured in the word (System). Senge (1990) made System Thinking more popular. His definition of Systems Thinking as a framework for seeing interrelationships rather than things, for seeing patterns of change rather than snapshots. It is consistent with Chapman (2004, p.14): "the core aspects of Systems Thinking are gaining a bigger picture (going up a level of abstraction) and appreciating other people’s perspectives. Systems Thinking could help HR students in developing a coherent understanding of business and make value-added decisions, which is a critical issue as HR function could become irrelevant if it doesn’t modernize its approach to understanding and planning for the future needs of the business (KPMG, 2019).

Education researchers work toward teaching and learning approaches that foster students’ Systems Thinking in various education fields (Assaraf & Orion, 2005; Molderez & Ceulemans, 2018; Hrin et al., 2017; Khanna et al., 2021; Gilissen et al., 2020); however, its integration in teaching business is still largely unexplored (Seiler & Kowalsky, 2011; Gregory & Miller, 2014). With relatively few exceptions (Atwater & Pittman, 2006; Atwater et al., 2008; Bui & Baruch, 2012; Waddock, 2010; Werhane, 2008; Turner & Baker, 2019), Systems Thinking does not seem to have a significant impact on business education. Therefore, it is critical to integrate Systems Thinking in business education in general and HR undergraduate and graduate curricula (Yawson, 2013).

Shaping decisions about people and the workforce using data insights is a central focus for future HR professionals (KPMG, 2019). HR professionals should maximize the power of data science to generate predictive and actionable insights for organizations. Using Artificial Intelligence (AI) and related technologies will be the biggest challenge for the HR function and require to enhance analytics capabilities. RapidMiner is AI as many of its processes support machine learning, for example, a deep learning
algorithm and advanced text analytics. In an HR environment, processes are interlinked and dependent on each other, which reinforces the importance of Systems Thinking for HR professionals. The power of the next generation of HR is in creating a holistic and mutually reinforcing "whole system" approach to building the workforce (and organization) of the future (KPMG, 2019). Systems Thinking increases stakeholder coordination across all parts of information systems and could help design reporting metrics that the HR professional can employ to assess future performance in an always-changing environment. HR professionals need several competencies: Data Driven, Business Acumen, Digital Integration, and People’s Advocate (van Vulpen, 2022, fig. 1). To be successful in the job market, HR graduates need robust data analytics and systems thinking skills to provide value-added decisions based on the most relevant and complete data.

Currently, the MS Excel skills for Human Resource students is a job requirement. Postings for entry-level HR positions list proficiency in Excel as one of their desired skills, but the depth of those skills are often undefined. Yet, human resource analytics as a discipline continues to expand in literature and practice. Big data, artificial intelligence, and machine learning are all technical disciplines adopted by companies with a desire to use data and algorithms to help solve complex problems—problems that often stretch beyond the ability of an MS Excel workbook to model. The disconnect here is something that was not lost on us; that companies are moving beyond MS Excel but Human Resources as a discipline is not. We recognized that to meet the demands of modern HR analytics is no longer a specialty skill but a graduating requirement for job competitiveness. As a first requirement to implement HR Analytics effectively is a number of employees with the knowledge and skills to collect the correct data, perform the right statistical analyses and then communicate the results in a meaningful and accessible way (Marler & Boudreau, 2017). Faculty could educate HR students about HR analytics, and students can enter the HR workforce with an already established working knowledge of HR analytics, effecting change immediately in HR
departments. Well-developed HR analytics skills could help graduates make significant human capital and strategic business decisions and gain a competitive advantage (Gurusinghe et al., 2021).

Based on the relevant literature review, we identified the critical importance of introducing the data mining tool RapidMiner for teaching data analytics to human resource (HR) students. In a way, the idea for another analysis tool contradicts the demand for demonstrable spreadsheet skills in business professions. Fluency in MS Excel is seen as a necessary skill for data analysis in business (Formby et al., 2017). However, more advanced analysis like predictive analytics (i.e., predicting employee churn, salary levels, survey text analysis, etc.) leaves a gap in MS Excel's skill set. Introducing HR students to RapidMiner fills that gap.

We wanted to provide students a deeper learning experience beyond the spreadsheet to give them a competitive advantage on the job market. We focused on three goals for this study. First, to investigate how RapidMiner could be used to teach data analytics to HR students. Second, how using RapidMiner could help to enhance systems thinking skills. Third, the benefits and drawbacks of using RapidMiner for teaching data analytics to HR students.

What is RapidMiner?

RapidMiner is a cross-platform data-mining application that supports data/text mining and advanced analytics (Wibowo et al., 2021). RapidMiner has unique characteristics that focus on Systems Thinking by providing an integrated/ deep learning experience for working with real business data. Students could provide cogent recommendations while solving real business problems by creating data models in its drag-and-drop process environment. RapidMiner promotes itself as "AI for Everyone" because there is no programming requirement to use the application. The company began as a data science project in 2001 at the University of Dortmund in Germany and has since grown to support over 150,000 users with offices in the USA, UK, Germany and Hungary (RapidMiner, n.d.). As an application, RapidMiner is a data processing
system where users import data, and then select a series of connected operators to perform functions on the data, providing modelled results that can be saved and recalled at will. RapidMiner studio is the downloadable application from RapidMiner.com, and, once installed, presents users with the following layout of five areas (fig. 2):

- **Repository.** The repository is the top left box on the screen and is where data that is to be used in RapidMiner is imported and stored. Data can be in a variety of formats, including CSV, XLSX, and authorized connections to database applications. The Local Repository contains all imported data sets, easily reused as needed.

- **Operators.** The operators in the bottom left box on the screen is where all RapidMiner functions exist. Similar to MS Excel, RapidMiner also uses functions to perform tasks. The Operator area categorizes operators based on tasks, and indented within each category are folders and sub-folders of operators that include data access (read and write), cleaning (missing and normalizing functions), modelling and validation functions. Additional operators beyond the base RapidMiner functions can be downloaded (mostly for free) from the marketplace - available from the main menu. Learning the most common operators and those that are good for particular models is the most difficult part of RapidMiner.

- **Parameters.** The parameters area in the top right of the screen is where a user can modify the processing of a particular operator. A selected operator will display its default options in the parameter box, allowing users to modify settings that affect the model output.

- **Help.** Help is in the bottom right of the screen, where descriptions of the selected operators and their parameters are provided to users. Often there are also examples of how a selected operator can be used.

- **Process.** The process area is the large canvas in the center of the RapidMiner screen. This is where operators are dragged in, and connected (piped or strung together), to create a data model that will run a particular process. Figure 3 includes an example of a data
model (process) that retrieves employee salary data and performs outlier detection.

Figure 3 lists four operators: Retrieve data (Employee Salary Data), Select Attributes (Salary), Detect Outliers (Univariate) in the salary column, and Generate Outlier Flag. Using these four operators we can produce the figure 4 subset of results, with outlier salaries clearly labelled under the Outlier Flag. If required, we can use multiple methods of detecting outliers including quartiles, histogram, or z-scores by simply selecting a different parameter during the model runtime (whereas in MS Excel these detection methods must be individually calculated by manual processes).

After the outliers have been detected, RapidMiner has a built-in visualization/charting feature where we can easily plot the results in a scatter plot and save the plot in a variety of graphic formats (Fig. 5). In figure 5, we can see the shape of the outlier results, with the actual salary outliers clearly in green dots in the top right of the scatter plot.

If an analyst needs to extract this outlier data set for use in another application, RapidMiner has write operators that include: CSV, MS Excel, MS-Access, Database, and many cloud-based providers (Salesforce, Amazon S3, Dropbox, MS-Azure and Google). There is also a free Tableau Marketplace extension that writes a RapidMiner example set to Tableau - should Tableau be the visualization application of choice for a company.

Using RapidMiner, students get hands-on experience creating data models that produce repeatable output. And if RapidMiner Server is installed, data and processes can be shared within an organization. HR students get practical experience with an advanced machine learning application and an understanding of how data flows from source to output - a demonstrated example of how an information system works. This is an invaluable experience. Students learn that data is a shared resource within a company and that results often extend in importance beyond their initial goal. For example, in our outlier example shown in figures 4-5: What are the positions whose salaries are outliers? What
are the employee attributes - age, gender, years in company, years with manager, bonus percentages, etc.? And how do these affect or correlate to their outlier status? System Thinking helps our students become better problem solvers by seeing modeled results as more than a singularly departmental concern.

**Research Questions**

Our research goal was to describe the RapidMiner application in teaching data analytics and how this approach could help to develop Systems Thinking and provide a deep learning experience, high engagement, confidence and student satisfaction. Based on our literature review, although business educators have made significant progress in implementing and sharing Data Analytics methods into the curriculum (Wymbs, 2016; Teimourzadeh, et al., 2022; Clayton & Clopton, 2019; Paul & MacDonald, 2020), we are not aware of any research that specifically investigates the impact of using the RapidMiner application on the following four constructs: Systems Thinking, Student Engagement, Student Satisfaction, and Student Confidence. Our review of related research indicates the importance of understanding of those four constructs. The enabler of these constructs is the application of the RapidMiner tool. Hence, our research questions are as follows:

**Q1:** What effect does the application of RapidMiner in an assignment have on HR students' perception of learning engagement?

Our assumption was that including RapidMiner could increase students learning engagement, which could be defined as high interest in course material, a personal connection with the subject and an understanding of how this assignment helps with a future career or job search after graduation (Czegledi & Smiderle, 2020). Engagement is also linked to interest and enjoyment in a task (Strati et al., 2017), which could lead to deeper learning for students.
Q2: What effect does the application of RapidMiner in an assignment have on HR students' perception of learning satisfaction?

Our assumption was that including RapidMiner could increase students' learning satisfaction, which could be defined as feeling confident with their ability to complete assigned tasks with quality and efficiency (Czegledi & Smiderle, 2020). In the RapidMiner assignment, to increase students' satisfaction and engagement real business data was used, which linked to more deep learning compared to surface learning and provided an environment for active learning (Smart & Csapo, 2007) as students have a more active role in the learning process.

Q3: What effect does the application of RapidMiner in an assignment have on HR students' perception of confidence related to working with business information?

Our assumption was that including RapidMiner positively impacts confidence, which we defined as feeling comfortable to complete assigned tasks and not being afraid to try new tasks as students were able to build transferrable and Systems Thinking skills to work with complex business data and People analytics. People analytics is an important element of HRM; however, few organizations are strong in this area (Deloitte, 2015). Strong People analytics could provide a competitive advantage to graduates prior to entering the job market.

Q4: What effect does the application RapidMiner in an assignment have on HR students' perception of the interconnection between parts of the business (systems thinking)?

Our assumption was that including RapidMiner could help students develop an understanding of the interconnection between elements of the business (which is one element of Systems Thinking). Our primary focus is developing a strategy for teaching elements of Systems Thinking for HR students and supporting our goal, real business data was introduced. This approach could positively impact students' engagement, satisfaction, and confidence, which leads to an easier introduction and application
of Systems Thinking concepts in teaching HR students. For this study we focused on the practical approach of Systems Thinking (Monat et al., 2020), such as understanding interrelationships between different elements of a complex situation or business units of an organization.

**Method**

**RapidMiner Assignment Design**

The RapidMiner assignment was designed for final semester post-graduate HR certificate students. Open-source data was used for the assignment (IBM's Employee Attrition and Performance dataset (https://zdataset.com/free-dataset/ibm-hr-analytics-employee-attrition-performance/), which included 1,470 data points. This data set provided students with the means to complete several analytics tasks such as basic descriptive statistics on the data, creating a decision tree to understand the patterns of exiting employees, and the ability to accurately predict the leave/stay status of an employee. This assignment helped students to understand the factors in an organization that affect an employee's company/work experience related to their leaving. After completing all analytics tasks, students provided analysis and recommendations on employee retention.

Two strategies aided students in their understanding of how to approach the topic of the assignment: Why do employees leave our company? The first strategy was introducing students to data mining instead of simply descriptive analytics. A descriptive approach only gave them basic insight into what happened with employees but no predictive capabilities around explaining how it happened or whether it will continue. Moving quickly into RapidMiner provided more options for how to classify this business problem and how to model solutions beyond the descriptive capabilities of spreadsheets. Our impetus was that the ability to work with HR data beyond simple analysis can make graduate students more attractive to employers.
Our second strategy was to increase student thinking of problems beyond the domain of HR, and to see how HR solutions could affect departments across the entire business (Systems Thinking). Figure 6 flowcharts the steps we used to employ the two strategies just discussed.

RapidMiner assignment required significant time effort for students; therefore, 20% of their total course mark was allocated. Marks were assigned to evaluate the following:

- Data wrangling and Exploratory Data Analysis (EDA) – cleaning, gaps, removals (Using MS Excel and RapidMiner Turboprep)
- Detect outliers and correlations – summarize findings (Using RapidMiner Automodel).
- Create a prediction model of the attrition dataset. Summarize findings, and evaluate the model's effectiveness by testing it on five untested employee records (using RapidMiner Studio & Automodel).
- Answer assignment questions about the use case of the model in HR (posed by senior management) – successes, problems, pitfalls, ethics, etc. (Writeup using MS Word).

Data Collection and Participants
HR graduate certificate students participated in this study and provided self-reported information based on their experience with the RapidMiner application on their level of engagement, satisfaction, confidence and understanding of interconnections between parts of the business (Systems Thinking). Upon completion of the assignment, students were asked to complete a survey based on their experience, as the quality of learning can be improved by adapting the teaching process as a result of students' feedback (Marsh, 1982). Ninety students participated in the survey. Data was collected through the Qualtrics application. In developing the survey questions, a literature review was completed with a focus on these constructs. Several studies were considered in designing the survey (Li et al., 2007; Topala & Tomozii, 2014; Chang & Chang, 2012; Czegledi & Smiderle, 2020; Lovelace et al., 2016). In designing survey questions, we
considered both cognitive and attitudinal questions (Cooper & Schindler, 2014), which often required a statement of direction (whether positive or negative) regarding a particular object (in this study, it would be the RapidMiner experience and its impact). The survey included two types of questions: 5-point Likert scale, which allowed students to indicate the extent to which they agreed or disagreed with statements related to their work with RapidMiner and several open-ended questions to add more depth to participants’ answers. The 5-point Likert-type balanced scaling approach was applied to ensure that participants choose a positive, negative or neutral response equally. Students indicated the extent to which they agreed or disagreed with statements regarding the RapidMiner impact on their level of engagement, satisfaction, confidence and understanding of business interconnections. Survey questions were pre-tested with other faculty to find potential issues, such as ambiguous language or if the survey completion time was significantly longer than anticipated. Table 10 includes survey questions for which students had provided self-reflective qualitative comments about their experience working with the RapidMiner in addition to 5-point Likert scale questions. This study received ethical approval from the institution’s Research Ethics Board, where the study was completed.

Results and Discussion
For our Likert-type questions on the survey, we used the standard five-point scale: strongly agree, agree, neutral, disagree, and strongly disagree. We asked students twenty-four questions that we grouped into three categories for analysis. These divisions allowed us to measure responses across three domains of interest and gauge students’ perceptions as identified in the research questions. Below are our groupings:

**Category 1:** Students' perceptions of engagement, satisfaction, and confidence (10 questions).

**Category 2:** Student’s perceptions of Systems Thinking (4 questions).
Category 3: Student’s Perceptions of RapidMiner as an analytics tool (10 questions).

Category 1 - Students’ Perceptions of Learning Engagement, Satisfaction, Confidence

Before we began the analysis of the Likert-type question responses in the first category of 10 questions, we ran a Cronbach’s alpha test in R using the ltm (Latent Trait Models) package. Chronbach’s alpha tests for internal consistency across the grouped question responses (SPSS FAQ, n.d.), in other words, are the questions closely related so that the category 1 questions reliably hold together as a unit? A coefficient from the test typically ranges between 0 and 1, whereas a coefficient closer to 1 indicates that the set of items performs strongly as a group. Our Cronbach’s score of .881 assured us that our category one questions have high inter-relatedness as a unit and that it was acceptable to analyze them as such (fig. 7).

Following the positivity of the Cronbach's coefficient, we grouped the category 1 student responses into Agree (Strongly Agree/Agree), Neutral, and Disagree (Strongly Disagree/Disagree). Table 1, "Students' perceptions of learning engagement, satisfaction, confidence" provides a breakdown of the survey results of students' perceptions using the Likert-scale format. 76% agreed/strongly agreed that Using RapidMiner increased learning engagement in this course (E1). Students commented, "At first it was confusing and overwhelming, but after using it while reviewing class recordings to accomplish the assignments, it was not that difficult. It was then exciting to see how entering data leads to new information to keep advancing to the next part of the assignment". Using real company data as a part of the RapidMiner was one of key elements in increasing learning engagement (E2) 77% agreed/strongly agreed (higher compared to fictitious data (E3). Also, using real company information as a part of the RapidMiner experience increased learning satisfaction (S1) 79% agreed/strongly agreed. Using RapidMiner could be beneficial for students’ future HR career, which students recognized and led to strong
scores on S2 (74% agreed/strongly agreed that the skills and knowledge acquired using RapidMiner will benefit their future career).

Using RapidMiner increased students’ confidence in working with Data Analytics requirements (69%, C1) and large volume of information (72%, C2). Students' comments support this observation: "I have a great experience with RapidMiner and learned how easy analyze business data with RapidMiner. It enhanced my data analyzing skills". "Overall, it was a good experience. I have learned new skills and techniques which will help me in future as HR". And as a result, it helped increased confidence in moving from school to work/industry (C3) 63% agreed/strongly agreed. Students' comments support this observation: "The experience has been new, and there is a lot to learn. Analytics is the future of HR, and learning this will just prepare us for the future". Students also identified that RapidMiner increased their confidence in working with real business data (70%, C4); however, they need more practice as only 61% felt confident using RapidMiner to help solve HR problems (C5).

An average of 70% of students' answers to the ten questions agreed with positive scores. 6% on average disagreed, and 24% were neutral. Noticeable standouts were the two double-digit disagree scores of questions Q2_2 and Q3_1. We interpret these two higher-than-average negatives and lower-than-average positive scores as indicators of confidence around the students' ability to use RapidMiner in a data model outside the assignment. In other words, confident with the data presented in a structured assignment but unsure in an unstructured setting. In our qualitative analysis section, we look for text patterns that give more definition and explanation to this uncertainty. However, we wanted to understand the neutral scores better because these represented an average of 24% of all answers. In Likert-type questionnaires, the question of how to treat the neutral scores is always an issue. Knowing whether neutral scores should be left alone or attributed to the agree or disagree column in a question is vital for the evaluations.
To determine how to use the neutral scores, we used the calculation parameters outlined in a 2015 paper by Abhijit Gosavi, *Analyzing Responses from Likert Surveys and Risk-Adjusted Ranking: A Data Analytics Perspective*, where we grouped the strongly agree/agree, and the strongly disagree/disagree answers each to their categories, resulting in three categories: agree, disagree and neutral with 90 responses per question. Treating the agree/disagree categorized results as a binomial distribution, we calculated the standard error (SE) and margin of error (MOE) for each question using the SE equation. We then calculated the confidence intervals (Stage 1: (agree + neutral) ± MOE or Stage 2: (disagree + neutral) ± MOE) across the two groups to see if we had overlaps in the intervals (Gosavi, 2015).

\[
SE = (n \times P1) \times (1 - P1)
\]

The margin of error with 0.05 alpha

\[
MOE = 1.96 \times SE
\]

Table 2 shows the results of the Excel analysis for the first category of ten questions. In our data, all cases of the agree number plus the neutral number exceeded the total of the disagree value (as evidenced by the bolded false formatting). In other words, all the individual question responses passed stage 1 of the equation. In stage 2, we added the neutral values to the disagree and passed the equation with no confidence interval overlaps (the second level of the bolded false conditions). For the series of questions in category one, we could reject the null hypothesis: The Agree value was not <= the Disagree value in each question’s response, with an alpha value of 0.05. Therefore, we could accept the alternative hypothesis (the Agree value was >= the Disagree value) and claim with 95% confidence that the number of those who chose to agree to our questions statistically exceeded those who chose to disagree (Gosavi, 2015) in the category one responses (Table 2).
Knowing that our questions in category 1 were highly inter-related and that student responses were statistically in agreement with the questions, next we wanted to test for a relationship between responses and time spent learning RapidMiner. We formulated two questions: 1) Were students who spent time outside of class on RapidMiner more likely to select agree to the questions, and 2) Were they more inclined to agree based on how much time they spent? In table 3, we compared our agree, neutral and disagree counts across all category one questions to the number of hours spent on the RapidMiner assignment outside class time. Students were given two hours per week of class time out of a 3-hour class to work on their assignments. They had 2.5 weeks to complete the assignment, so 4 hours of class time in total with instructor access. The time category counts results are included in table 3.

A Pearson's chi-square test (fig. 8) was run against the table 3 numbers using the stats package in R. With a test statistic of 38.002 between the observed and expected values, the resulting p-value of 0.0004998 informed us that there is dependency among the categories (agree, neutral, disagree) with only a 5 in 10,000 probability that the numbers are due to chance. With this assurance, we continued our analysis.

In order to add more depth to our analysis for table 3 data, the results were re-calculated by percentages, split into two categories: students who spent one hour or less on RapidMiner outside of class time and those who spent more than one hour (Table 4).

Table 4 shows that students who spent >1 hour on RapidMiner outside of class time were 11% more positive in their assessment of RapidMiner than those who spent less than one hour. We also found reductions in disagreement by 45% but an increase in neutral reviews by 16% for those spending greater than one hour. To understand the likelihood of achieving an agree response over a disagree response, we created a 2x2 matrix (table 5) and calculated the Relative Risk (RR) and the Odds Ratio (OR).

Relative Risk and Odds Ratios measure the likelihood or chance of an event occurring between two groups (Relative Risk, 2013). In our
case, the possibility of an agree response vs. a disagree response under students’ invested time of \( \leq \) or \( > \) one hour. We created our matrix using the values already determined in table 3.

Using table 5, we determined the likelihood of an agree response to the category one questions from students based on the time spent on RapidMiner outside of class (Andrade, 2015). Table 6 shows our calculations and interpretations.

Based on our analysis, we were able to derive the following conclusions from our analysis of the category one student responses to questions about student perceptions of engagement, satisfaction and confidence:

- Chronbach’s Alpha coefficient of .881 (figure 7) assured us that our category one questions had high inter-relatedness as a unit. Proceeding with the analysis of the category one questions was supported.
- An average of 70% of students’ answers to the ten questions in category one was in agreement with positive scores, 6% on average disagreed, and 24% were neutral (table 1).
- For the series of questions in category one, we could claim with 95% confidence that the number of students who agreed to our questions statistically exceeded those who chose to disagree (table 2).
- A chi-squared test statistic of 38.002 between the observed and expected values, with the resulting p-value of 0.0004998, informed us that there is dependency among the categories (agree, neutral, disagree) with only a 5 in 10,000 probability that the numbers were due to chance (figure 8).
- Students who spent \( >1 \) hour on RapidMiner outside of class time were 11% more positive in their assessment of RapidMiner than those who spent less than one hour. We also found reductions in disagreement by 45% and a reduction in neutral reviews by 16% for those spending greater than one hour (table 6).
- The odds ratio informed us that students who spend \( >1 \) of study on RapidMiner outside of the designated 4-hour class time are 3.33
times more likely to give an agreed response to the category one questions (table 6).

The following two tables (7 and 8) provide summary analysis of the category two and three sets of Likert responses. For a detailed comparisons of the results see table 9.

**Category 2 Responses - Students' Perceptions of Systems Thinking**

Table 7, "Student perceptions of Systems Thinking" provides a breakdown of the survey results of students' perceptions using the Likert-scale format. 77% agreed/strongly agreed that RapidMiner helped to adopt a "big picture" approach to solving business problems with data (ST1), and 78% agreed/strongly agreed that RapidMiner helped to understand the importance of measuring the performance of a data model (ST2). The majority of students (79%) realized that with RapidMiner, they could do deeper data analysis compared to MS Excel (ST3); however, only 62% agreed that with RapidMiner they could analyze complex business problems without programming (ST4).

**Category 3 Responses - Students' Perceptions of RapidMiner Functionality as an Analytics Tool**

Table 8, "Students Perceptions of RapidMiner Functionality as an Analytics Tool provided a breakdown of the survey results of students' perceptions using the Likert-scale format. 69% recommended using RapidMiner to other students (RM1). 59% suggested using RapidMiner in more courses (RM2), lower number could be related to not clear how RapidMiner could be applied in other HR courses. 57% of students identified that RapidMiner was difficult to learn (RM3). Only 52% found that building a data/process model with RapidMiner was straightforward (RM4). However once students learned the basic functionality, RapidMiner was easier to use (76%, RM5), as was supported by students' comments: "Although I found it a bit tough in the beginning, it gradually became easy. It was a fascinating experience to analyze the data using RapidMiner". One of
the potential challenges with using RapidMiner could be understanding basic descriptive statistics (73%, RM6) and more data analysis skills before students could use RapidMiner effectively (64%, RM7). Students agreed that provided instructions were sufficient to complete the RapidMiner work (78%, RM8), and RapidMiner increased their interest in the HR Analytics course (68%, RM9), as supported by student’s comments: ”RapidMiner gave me a big insight about data analytics and helped me in understanding the importance of data analysis in an organization”. However, as it was their first experience with RapidMiner 58% agreed that they would need technical support to use RapidMiner in their future HR career (RM10).

After completing the detailed analysis of the category 1 student responses (tables 1-6), we performed the same analysis for categories 2 and 3 responses. For this paper, we have not included the details of these analyses; however, table 9 summarizes and compares the findings across all three Likert question categories.

Based on the summaries in table 9, students were more positive in their perceptions of RapidMiner (B) across all three response categories above neutral or disagree responses. Both Cronbach’s Alpha (A) and the chi-squared tests (D) confirmed relatedness in the Likert-type question categories with good Cronbach scores and statistically significant p-values in the chi-squared tests within each category of questions. We found grouping similarities across the three categories when studying the agree, neutral, and disagree responses percentages (E) to a comparison of >1 hour time spent outside of class on RapidMiner gave us standard time spent increases/decreases over <= hour. However, when we compared the percentage changes of category three (Student’s Perceptions of RapidMiner as an analytics tool), it scored noticeably lower than the more positive responses in categories one and two. In category 3, agree scores were at their lowest with an 8% increase, neutral scores were up 17%, and disagreements were only reduced by 6%. The ratio calculations also showed less positive risk and odds ratios. From these
comparisons, we inferred that student responses expressed trepidation when applying their skills using the data mining tool RapidMiner outside of the framed assignment (where data and goal are provided to them) versus an unframed real-world experience where data and the goal may not be so clearly defined. This inference is also supported by the confidence intervals expressed through the margin of error calculations (C), where only category three could not express an across-the-board declaration of 95% confidence of agree scores over disagree because of CI overlap in 4 of the ten questions.

To understand if there was a relationship between the findings in the Likert-type questions and the student comments, we next analyzed the text of the student responses across four text domains: Top 4 ngrams by question using both the tidytext and gt packages in R, word synonym pairs using Word2Vec, negative/positive word contributions using the tidytext package in R, and Syuzhet word emotion categories in R.

**Qualitative Analysis - Student Perceptions of the RapidMiner Experience Expressed in Open-Ended Questions**

Students were asked four open-ended questions, each meant to elicit a text response. We wanted to see if these answers would correlate to the positive responses we received with the Likert-type questions. The four questions asked are included in table 10:

We began with a simple ngram analysis where n=2, but we found that a bigram was not sufficient enough to understand the context of how the student’s words were used. We tried higher word frequencies, but that did not translate into a better comprehension of student comments. The trade-off between the number of words in the ngram and the times used was n=4, providing us with the most interpretable visualization across all questions. The RStudio package **gt** (Grammar of Tables) was used to construct the following table based on our analysis. Each question is categorized with its own words (4 words for each line) and the count, n, of how many times that 4-word combination appeared (fig 9).
An analysis of Q7 ngrams gives us the sense that difficulty and time were factors, but were overcome with usage. Q8 is too ambiguous; the n=4 was not enough to provide an adequate analysis. Nor did we find that reducing the ngrams to n=2, or increasing them to n=6, improved our understanding of the comments. Q9 gives us importing data and fast as two possibilities for common word patterns. With Q10, more videos and hands on experience were the two most common factors that students would change.

The ngram analysis, although interesting, did not give us a clear understanding of what was most uniformly expressed by the students. The student comments needed more than word frequencies; a necessary analysis but not sufficient to clearly comprehend similarities across the comments for a given question. In other words, frequencies were not going to help us understand the student comments, but identifying word proximities (words that appeared close to each other across comments and could be mapped into vectors so that mathematically distances between the words could be computed), should provide us more textual insight. The goal of this next text analysis segment being that words in a closer vector space with one another probably have a similar meaning or could be synonymous.

We used the Word2Vec operator in RapidMiner across all the student comments (360 in total (90x4), to discover synonymous word pairs with a threshold cosine distance of <50. This was the opportunity to evaluate the generated stem rules produced by the Word2Vec algorithm to see if they help further explained or supported the ngrams. The RapidMiner process provided five rules across all the comments, as demonstrated in figure 10.

The proximity of the words experience and videos stemmed into the rule 1, experience:videos, reinforced the request that students associate their experience with Rapidminer with videos. In other words, a desire for more supporting videos. This rule is also enforced with rule 5, videos:helpful|data, stressing again the students desire for more videos that would help them when working with data in RapidMiner. Rule 2 finds the
word *helpful* synonymous with *experience|tool* so that students equated their experience using the tool RapidMiner as helpful (reading random comments seemed to equate helpful with a valuable skill). The words *good* and *time* were concatenated by Word2Vec but were not meant to convey the student’s attitude towards learning RapidMiner, but that they found it good and wished there was more time to work with understanding the tool. Not exactly opposite meanings, but through Word2Vec they are considered synonymous. The concatenation of *understand|difficult* with *time* in rule 4 are synonymous and associated with the *time* element - adding more clarity to rule 3.

The next text analysis we performed on the student’s comments was a sentiment analysis. We wanted to categorize the overall sentiment of their comment word choices into positive or negative - which would give us a sense of their emotional response to learning RapidMiner for the assignment. For this we used the tidyverse and tidytext packages in RStudio to identify by frequency, the top 10 positive and negative words used across all comments, questions Q7-Q10 inclusive. We then used ggplot2 to visualize the results (fig. 11).

We see that the frequency of positive words is greater than the identified negative words. Closer examination of the most frequently used negative words, *difficult* and *confusing*, is more qualified, i.e., is meant to convey the initial state of the students which was mitigated through time and experience as displayed in the stemming rules in figure 9. *Helpful*, *good*, and *support*, are the top positive words, but like the negative words, need further qualification as indicators of time and videos - as expressed in the stemming rules. Note that the words *time* and *video*, although important in the stemrules of Word2Vec, do not appear in the sentiment analysis because they are both sentiment neutral words.

As a final analysis, we grouped the words by emotion using the syuzhet package in RStudio. Syuzhet categorizes sentiment words across eight predefined emotions, which we visualized a simple bar plot using ggplot2 by percentage in each category. Positive emotions (trust, anticipation, joy)
scored with 76%, whereas more negative associated emotions (surprise, fear, sadness, anger, disgust) scored with 24%, which gives us a 3:1 ratio of positive emotions words to negative (fig 12) in the comments of the students.

Following the results of our text analysis, we can state that students were positive about their experience with RapidMiner, as expressed through their word choices. Even the synonymous word pairs, as determined using Word2Vec through RapidMiner, provide more positive associations across the student comments in context. Using these text mining models were able to better understand our student’s experience using RapidMiner in the context of an HR data-based assignment. Support between the ngrams and the Word2Vec stem rules, and the emotion in student comments seemed to reinforce three findings:

1. More time is needed with RapidMiner to gain a better understanding of a difficult tool. The text mining reinforces this finding as does the Likert analysis where the odds of students who spent more than an hour outside of class learning RapidMiner are 23.4x more likely to provide an agree response to category 1 questions.

2. More access to videos to support student’s understanding of RapidMiner. This is evident in the text mining where 4/10 responses cite videos in the Q10 answers, and the number may have been greater had we extracted ngrams of 5 terms. Also, Word2Vec synonyms included video in 2 of its 5 stemrules.

3. The experience of working with RapidMiner was helpful. Although some students found the application difficult, we found that time spent outside of class had an influence on this feeling with an odds ratio of 3.33x. We also found a higher frequency of positive word counts over negative, and that trust, anticipation and joy were the three larger percentages of feeling as expressed using the Syuzhet package in R.

This knowledge will help us improve HR student experience with the data mining tool RapidMiner through its paced introduction, supplemented with teacher explanations and existing linked videos to content,
and by reinforcing the skill development benefits of such a tool for the
HR job market.

**Using RapidMiner to Develop Data Analytics and Systems Thinking Skills: Potential Implementation Strategy Issues and Possible Resolutions**

Utilizing RapidMiner to develop Data Analytics and Systems Thinking skills for HR students could have some challenges. These challenges will need to be addressed before the RapidMiner application’s benefits can be fully realized. Here we identify several critical issues with incorporating the application of the RapidMiner when teaching HR students data analytics:

1. Faculty require working knowledge of RapidMiner and statistics
2. Students need basic knowledge of statistics
3. Access to business data for faculty and students
4. Students require extra support (video, instructor access, etc.)
5. Systems Thinking and DA in teaching HR students

Several possible resolutions are identified for each issue for practical application of this experiential learning approach:

1. **Faculty require working knowledge of RapidMiner and statistics**
   - Faculty need to invest time in learning RapidMiner. To help faculty gain required knowledge and skills RapidMiner tutorials with videos could be used, which are available for free at RapidMiner resource center (https://rapidminer.com/resource/).
   - There is also an active community of Q&A responses (https://community.rapidminer.com/).

2. **Students need basic knowledge in statistics**
   - Statistical literacy plays an important role in data interpretation (Teimourzadeh, et al., 2022; Schield, 1998, 1999) and some students may have limited skills in this area. Therefore, it is recommended
to review basic statistical concepts with students prior to using RapidMiner. Several trainings are available on LinkedIn Training.

- For data modeling it is suggested to help students to learn concepts with a focus on predictive models

3. Access to business data for faculty and students

- Students potentially limited prior experience with real business data (extra support will be required from faculty)
- Basic data cleaning/wrangling concepts and methods will need to be taught to HR students

4. Students require extra support and time commitment

- Students need to invest the time to learn how to use RapidMiner to its full extent. However, with the AutoModel feature students can quickly gain data mining results without a major time commitment to understanding all criterion in the model. AutoModel is an important feature that combines several predictive algorithms in one process so that users can gauge the effectiveness of each model through a base comparison of classification error scores. This is as opposed to modeling each algorithm independently.
- Some students may have mental barriers to RapidMiner learning (anything related to "programming", or "Math"). It is recommended to start with more basic concepts related to everyday life, which could be easier to understand. For example, analysis of the readily available "Titanic" (https://www.kaggle.com/c/titanic/data) data set, which was used in RapidMiner as a primer to this assignment to help teach analytic concepts.
- Significant effort and time for instructor preparation and student evaluations are required. Faculty could ask students to work in groups or use data from the same company (like the IBM attrition data set) for several groups. Faculty can select a company/industry with which they are familiar or have prior professional experience.
- Faculty could create instructional Videos to help with challenging tasks. Several students identified a positive impact of instructional videos on their learning process, for example: "In the beginning, I
was struggling to understand the process like how to attach files and import them. I watched the class recording videos again and again. Then I was able to understand what I need to do”.

5. Systems Thinking and Data Analytics in teaching HR students

- To help students understand the importance of ST and DA, faculty could include a discussion of current trends in the HR profession and requirements for HR graduates to be successful in job markets. ST and DA are critical elements (KPMG, 2019) for successful HR careers. A deep understanding of the knowledge hidden in HR data is a critical element of organizational decision making (Ranjan et al., 2008). DA could help find patterns helpful for value-added analysis (Wardhani & Gata, 2022).

Limitations and Future Research

The study has some limitations and opportunities for further research. In the next research stage, Kirkpatrick’s training evaluation model could be applied (Kirkpatrick and Kirkpatrick, 2005, 2006). Kirkpatrick’s four levels model is outcome-based and objective oriented and focuses on determining the results of an educational strategy (Reio et al., 2017) and will provide additional depth to the future study and enhance the practical application for the business environment (Kirkpatrick, 2018). ………………..

Conclusion

In this study, we investigated and described RapidMiner for teaching Data Analytics to human resource students. MS Excel is the standard tool for data analysis in business; however, it does not allow for predictive analytics. We wanted to provide students with a deeper learning experience beyond the spreadsheet by using the open-source data mining tool RapidMiner to give them a competitive advantage in the job market. Based on students' survey results, using RapidMiner resulted in high engagement, learning satisfaction and understanding of Systems Thinking.
We adapted active learning strategies and an HR data set from IBM to reduce the gap between business education and practice. We also provided practical recommendations on using RapidMiner for teaching Data Analytics as a method to develop fundamentals of Systems Thinking for business students. We recommend that each institution evaluate the analytic capabilities of their students as a prerequisite before embarking on a similar assignment offering. Ensuring the students are equipped with basic statistical and data knowledge before the assignment helps make the analytical requirements of the assignment more understandable and appealing as a job skill.

References


Molderez, I., & Ceulemans, K. (2018). The power of art to foster systems thinking, one of the key competencies of education for sustainable development. Journal of Cleaner Production, 186, 758–770. https://doi.org/10.1016/j.jclepro.2018.03.120


RapidMiner - https://rapidminer.com/


Vassallo, S. (2017). Design Thinking Needs To Think Bigger. Design thinking, as it was conceived 15 years ago, has outlived its usefulness. Enter systems thinking. https://www.fastcompany.com/90112320/design-thinking-needs-to-think-bigger

Waddock, S. (2010). Finding wisdom within—The role of seeing and reflective practice in developing moral imagination, aesthetic sensibil-


**Web Appendix**

A web appendix for this paper is available at https://dx.doi.org/10.15239/j.brcacadjb.2023.13.01.wa02