Integrating RapidMiner in Business Analytics Education: An Instructional Approach for Skill Development

American Journal of Education and Learning
Vol. 10, No. 1, 104-116, 2025
e-ISSN:2518-6647

Anand Jeyaraj

Raj Soin College of Business, Wright State University, 3640 Colonel Glenn Highway, Dayton, OH 45435, USA.
Email: anand.jeyaraj@wright.edu

ABSTRACT

The continued growth of business analytics discipline raises the need for students as future professionals to be trained in business analytics concepts and applications to enable data-driven decision-making within organizations. As business analytics evolves to incorporate data mining and machine learning applications, students need to develop an overall understanding of the process of acquiring, preparing, and analyzing data. This paper describes a framework involving five stages—preparation, exploration, modeling, optimization, and validation—that can be used to instruct students on the business analytics process in the context of the RapidMiner software package. Further, this paper illustrates the application of the framework using a specific example that uses decision trees along with a discussion of the descriptive statistics, visual analysis using charts, identifying the variables for analysis, ways to optimize and validate models, and assess model performance. This includes training and testing (holdout) samples, unbalanced data, confusion matrix, precision and recall metrics, and AUC and F1 metrics. Student performance on individual assignments following in-class instruction and demonstration based on the framework shows that it was helpful in student learning. Although the framework was tested in the context of RapidMiner, it can be extended for business analytics instruction using other software tools.

Keywords: *Business analytics, Instructional framework, RapidMiner, Preparation, Exploration, Modeling, Optimization, Validation.*

DOI: 10.55284/ajel.v10i1.1458
Citation | Jeyaraj, A. (2025). Integrating RapidMiner in Business Analytics Education: An Instructional Approach for Skill Development. *American Journal of Education and Learning, 10*(1), 104–116.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The author confirms that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Competing Interests: The author declares that there are no conflicts of interests regarding the publication of this paper.
History: History: Received: 17 April 2025/ Revised: 26 May 2025/ Accepted: 6 June 2025/ Published: 26 June 2025
Publisher: Online Science Publishing

Highlights of this paper

- An instructional framework involving five stages—preparation, exploration, modeling, optimization, and validation—for business analytics education is described.
- The application of the instructional framework is shown in the context of the RapidMiner software using a dataset on customer churn available on the Kaggle website.
- Student performance on assignments showed that the instructional approach was effective in enabling the development of students' business analytics skills.

1. INTRODUCTION

As the business analytics discipline continues to mature and its applications find broader acceptance in enabling data-driven business decision-making within organizations, the need to train and equip current students, i.e., our future professionals, in business analytics and related software tools gains considerable importance (Jeyaraj, 2019; Nguyen, Gardner, & Sheridan, 2020; Provost & Fawcett, 2013). Such skills enable students to become more proficient in business analytics and also enhance analytical and critical thinking, problem solving, data-driven storytelling, and data-driven decision-making appropriate for operational and strategic levels of the organization.

Analytics has been applied to understand various business activities such as credit scoring, bankruptcy, customer churn, mortgage loans, and product sales (Boughaci, Alkhawaldeh, & Haddadi, 2021; Haddadi, Boughaci, & Alkhawaldeh, 2024). Depending on the goals, different techniques such as prediction, classification, clustering, and even a combination of these methods (i.e., ensemble modeling) have been used (Dhar & Bose, 2024; Sivri & Ustundag, 2024). These suggest that there is considerable potential for the application of business analytics for real-world impacts. Therefore, the need for students to be exposed to the principles and methods of business analytics gains prominence, which includes the hands-on use and training on relevant software tools for data analysis and visualization.

While several software tools[1] are available for data preparation as well as data mining and machine learning applications, RapidMiner was the software tool of choice for activities related to business analytics explained in this paper. RapidMiner can be viewed as a low-code/no-code software environment that provides a visual programming interface to create the necessary workflows to accomplish activities related to data preparation, data mining, predictive modeling, and machine learning. It offers a suite of operators grouped into various categories such as blending, cleansing, modeling, scoring, and validation that facilitate business analytics tasks[2]. The visual programming interface can be used to construct a workflow by simply dragging and dropping the operators into the workspace, adjusting the hyperparameters for the operators, and establishing connections between the operators. The RapidMiner visual interface is particularly helpful for business students—especially if they are not immersed in information technologies, data and programming environments, or data preparation and analysis activities—to navigate the complexities of data preparation and analysis without having to be overwhelmed by the intricacies of managing data using specialized information technologies.

This paper describes an instructional approach used to help business students learn the business analytics concepts and apply them using RapidMiner. This approach is particularly relevant for business education since there are opportunities to apply business analytics methods in a variety of settings such as sales, marketing, finance, supply chain management, and human resources for various purposes such as predicting sales, segmenting

[1] Examples include Python (and its numerous libraries such as numpy, pandas, and matplotlib), R (with its libraries such as Tidyverse and Plotly), and SAS, which are largely code-based environments.

[2] See: https://docs.rapidminer.com/2025.0/studio/operators/index.html for a full list of RapidMiner operators.

consumers, optimizing supply chains, enhancing customer service, and enhancing operational efficiencies. Business students may gain skills that may be of significant value in their professional careers through their engagement with business analytics using the instructional approach. The ensuing sections of the paper describe the organizing framework for learning business analytics and the application of the framework in a business analytics course, followed by a discussion and conclusion.

2. FRAMEWORK

Adopting the notion that students should have an overall understanding of the considerations underlying business analytics for more effective engagement, the organizing framework depicted in Figure 1 is used for instruction[3]. The framework identifies five broad stages[4]—preparation, exploration, modeling, optimization, and validation—and can be used for theoretical discussions (which enable students to understand the concepts, principles, and methods of business analytics) and practical applications (which enable students to obtain hands-on experience with business analytics, in this particular case, through the use of RapidMiner).

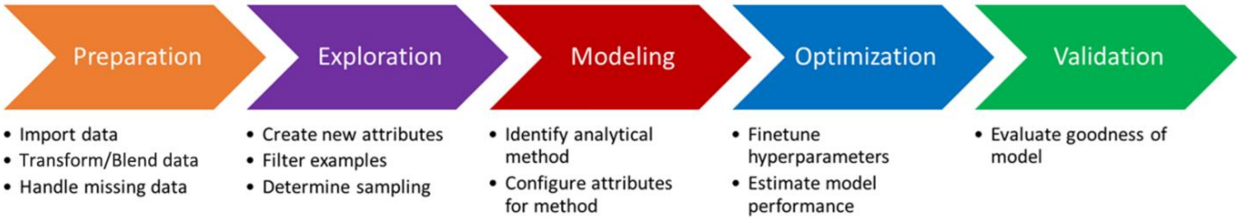


Figure 1. Framework for RapidMiner instruction.

The preparation stage deals with acquiring data from various internal or external sources, consolidating or blending data in different containers or formats (e.g., CSV, XML), transforming data to ensure consistency, renaming attributes, and handling missing data. The culmination of activities in the preparation stage results in a usable dataset that has been acquired, transformed, coded, corrected, and verified for use in the next stage.

The exploration stage represents the preliminary examination of the data prior to applying specific methods. Activities include creating new attributes for better representation, examining associations or correlations to identify attributes that cannot be used together, identifying other attributes that may not have much variation, and identifying necessary roles for attributes (e.g., unique identifier such as ID cannot be used in modeling). Moreover, visualizations such as box plots and scatter plots can be used to examine distributions and relationships.

The modeling stage deals with various analytical methods available for analyzing data and the selection of the methods that may be appropriate for the business problem at hand. In this study, three broad classes of methods are

[3] Several frameworks such as Cross Industry Standard Process for Data Mining (CRISP-DM), Sample-Explore-Modify-Model-Assess (SEMMA), Analytics Solutions Unified Method for Data Mining (ASUM-DM), Knowledge Discovery in Databases (KDD), and Team Data Science Process (TDSP) have been proposed and extended to the business analytics contexts over time (Ahmad, Yaacob, Ibrahim, & Wan Fakhruddin, 2022; Shafique & Qaiser, 2014). Despite variations in the number of stages seen in such frameworks, they generally include stages that broadly deal with data acquisition, data preparation, behavioral modeling, and model evaluation. Further, custom frameworks have been adopted for business analytics with stages that are more specific to the learning contexts and overall goals (Jeyaraj, 2019; Khan, Nadeem, & Ali, 2019; Zhang, Chen, & Wei, 2020). We acknowledge that one of these frameworks could be used for business analytics if students are already knowledgeable, but the framework in Figure 1 is presented as an instructional approach to help them learn business analytics. That is, students will have to first receive instruction on the various activities involved in business analytics before they can begin to apply their learning.

[4] These stages could be iterative depending on the dataset and the necessary activities. However, to facilitate ease of instruction and discussion, the stages are presented as sequential. It is possible that students may have to revisit an earlier stage in the process to reevaluate, respecify, or rerun the models.

considered—regression, classification, and segmentation. Regression is appropriate for problems that require numeric values to be predicted (e.g., amount of insurance premium), classification is relevant for the prediction of classes (e.g., risk rating of customer as Low, Medium, or High), and segmentation is helpful for problems that require the identification of clusters of cases. Ensemble modeling that allows the use of multiple methods in conjunction is also possible. It may be necessary to apply sampling methods to balance the data for more accurate modeling especially if data for the primary variable of interest is skewed.

The optimization stage allows for the identification of the optimal values for hyperparameters for the analytical method chosen for analysis (e.g., tree depth for decision trees). This finetuning of hyperparameters impacts the model training and its performance. Hyperparameters may be initialized before the training and can be systematically evaluated using a range of values. While a manual random search for optimal values is possible (i.e., text one hyperparameter value at a time), a systematic and exhaustive search for optimal values from a range of predefined values is also available although it could be computationally intensive. The optimization process involves an internal validation in that the model parameters are assessed for every combination of relevant values for hyperparameters. The internal validation process could be based on k-fold cross-validation in which the dataset is split into k subsets and model performance is assessed by using k-1 subsets for training and 1 subset for evaluation.

The validation stage encompasses the assessment and evaluation activities through which the goodness of the optimized models is determined. This is typically accomplished by applying the optimized model on the testing dataset. If testing data is not available, then one of the validation methods such as split validation or cross validation to assess model performance. The confusion matrix and ROC (receiver operating characteristic) curve may be examined to validate overall model performance.

Table 1 provides a summary of the major activities in each stage specified in the instructional framework and the related RapidMiner operators that may be appropriate in each stage[5]. Each of these operators can be activated by placing it on the RapidMiner workflow, adjusting its settings, and connecting it with other operators through the relevant input or output ports. For instance, the settings for the Read CSV operator can be adjusted to point to a specific CSV file available on a secondary storage device and the "out" port can be connected to the "res" port (in the RapidMiner workflow environment). Depending on the type of operator, different ports may be available for use. The Decision Tree operator, for instance, includes the "tra" (to provide the training dataset as input) port as well as the "mod" (to obtain the decision tree model) and "exa" (to obtain the dataset) ports for subsequent steps. Some operators can be applied for different problems. For instance, the Support Vector Machine operator can be used for both classification (i.e., prediction of a class) and regression (i.e., prediction of a numeric value) depending on the type of data represented in the variable with the "label" role.

[5] Certain operators may span multiple stages depending on how they could be used. For instance, the Cross Validation operator is generally used in the Optimization stage to assess model performance even as the search for optimal values for the hyperparameters is being conducted. However, it can also be used in the Validation stage especially if testing data is not available to assess the performance of the optimized model.

Table 1. Stages in RapidMiner operators.

Stage	Goals	RapidMiner operators
Preparation	▪ Acquire data from various sources ▪ Combine or blend data in different containers ▪ Apply transformations for data consistency ▪ Rename attributes as needed ▪ Handle missing data	Read CSV, read excel, read XML Nominal to binominal, nominal to numerical, numerical to polynominal, date to nominal Union, join Trim, split Rename Replace missing values Normalize
Exploration	▪ Create new attributes as needed ▪ Exclude unnecessary attributes from analysis ▪ Filter examples as needed	Generate attributes, map Remove correlated attributes Remove useless attributes Select attributes Remove duplicates Filter examples Set role
Modeling	▪ Identify analytical method ▪ Apply sampling methods to balance data ▪ Configure roles for necessary attributes ▪ Determine training and testing data ▪ Identify model parameters	Sample Set role Multiply Split data *Regression*: Linear regression *Classification*: k-NN, decision tree, Support vector machine, Neural net *Segmentation*: Clustering (k-means) *Ensemble*: Random forest, bagging, AdaBoost, vote, stacking
Optimization	▪ Apply analytical method on training data ▪ Evaluate model for hyperparameter combinations ▪ Identify optimal values for hyperparameters	Optimize parameters (Grid) Loop parameters Apply model, performance Split validation, cross validation
Validation	▪ Apply optimized model on testing data ▪ Evaluate goodness of optimized model	Apply model Performance Split validation, cross validation

3. APPLICATION

3.1. Context

The instructional approach was implemented in an introductory undergraduate course on business analytics in the business school at a large university in Midwestern USA. The course is open to students in different business disciplines such as accountancy, marketing, finance, and information systems. Students are typically in their junior or senior years of study when entering the course. It may be the first introduction to business analytics for the majority of students although it is possible that some students have taken an introductory course on business data that introduces the fundamentals of data acquisition, preparation, and visualization. The course introduces students to the concepts and principles of business analytics including data mining and machine learning.

3.2. Instruction

The illustration below is based on the dataset on credit card customers available at: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers and is aimed at capturing customer churn. The dataset contained 10000 examples (cases) and 21 usable attributes describing the customer (e.g., Customer_Age, Gender, Education_Level, Marital_Status, Income_Category, Dependent_count), their relationship with the bank (e.g., Months_on_book, Credit_Limit, Total_Relationship_Count, Total_Revolving_Bal, Total_Trans_Amount, Total_Trans_Count), and their status (i.e., Existing customer or Attrited customer). Since the target attribute was binominal, the dataset was deemed appropriate for classification methods, which implies that methods such as Decision Trees, k- Nearest Neighbor (k-NN), Artificial Neural Network (ANN), and Support Vector Machine (SVM) may be appropriate.

Figure 2 shows the RapidMiner process related to the preparation and exploration stages of the instructional framework. The data downloaded from the web site was in CSV format and was imported into RapidMiner using the Read CSV operator. The operators for these two stages are placed in a subprocess named "Prep-Explore" in the overall RapidMiner process.

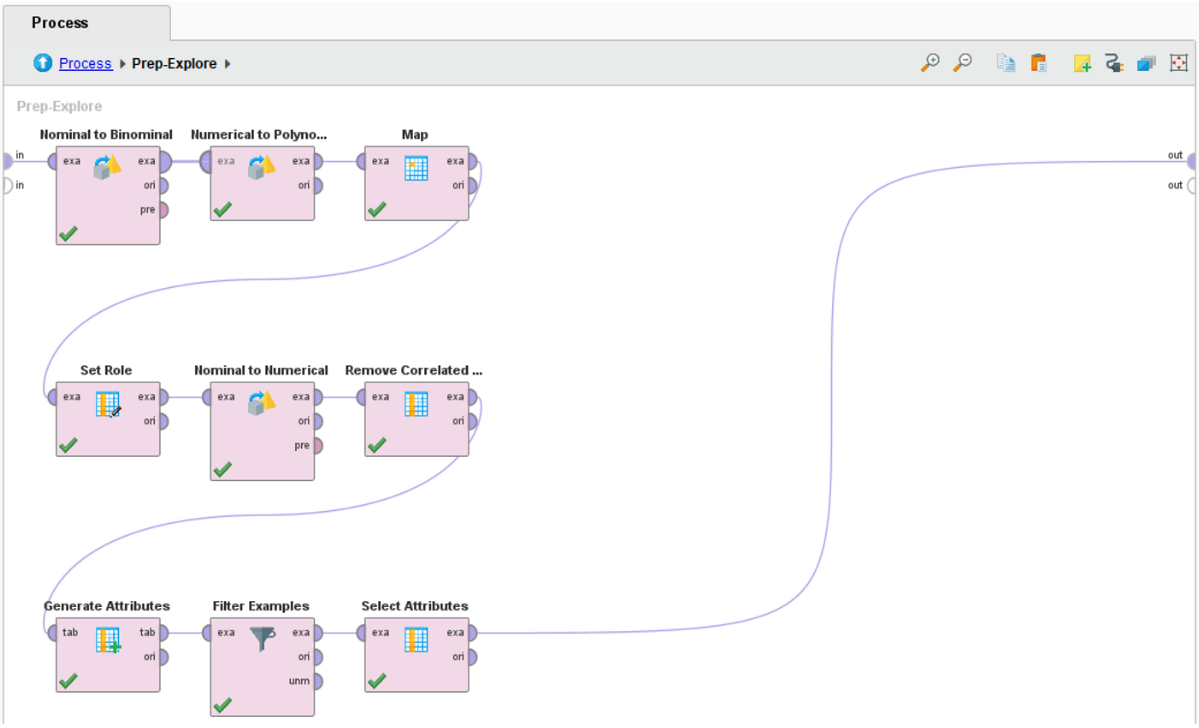


Figure 2. Preparation and exploration.

The Nominal to Binominal operator was used to recode Gender as binominal since it represented only two levels (M and F). The Numerical to Nominal operator was used to recode the ClientNum attribute since the ID values cannot be used for numeric computations. The Map operator was used to recode the Attrition_Flag attribute such that the two classes were worded as Existing and Attrited. The "Label" role for the Attrition_Flag attribute to enable classification and the "ID" role for the ClientNum attribute was given using the Set Role operator. The Nominal to Numerical operator was used to transform the Gender attribute with M and F codes into numerical attributes (i.e., one for each level). The Remove Correlated Attributes operator was used to assess associations and exclude attributes that shared high correlations with other attributes. This operator is helpful since the Nominal to Numerical operator created two separate binary variables based on Gender (i.e., Gender=M and Gender=F) and only one of the two attributes can be used in classification models to eliminate multicollinearity problems. The Generate Attributes operator was used to construct new attributes based on the Education_Level, Marital_Status, and Income_Category attributes. For instance, the new attribute Married was coded as 1 representing Marital_Status = Married and 0 representing other statuses such as Single and Divorced. The Filter Examples operator was to exclude examples that represent card categories other than "Blue" since 93% of the examples dealt with the "Blue" card category. The Select Attributes operator was used to include/exclude attributes for the classification method.

Figure 3 shows the RapidMiner process related to the modeling and optimization stages of the instructional framework. For modeling, the Decision Tree operator was placed within the “training panel” of the Cross Validation operator, which automatically splits the dataset into the training and testing datasets typically using the k (=10 by default) parameter and bootstrapping the examples in the original dataset into 90% on the training dataset and 10% on the testing dataset. Thus, the Decision Tree operator served to first train a decision tree model with the examples in the training dataset based on the patterns between the Attrition_Flag label and the other attributes. The Apply Model operator was used to score the prediction of the Attrition_Flag label for examples in the testing dataset using the trained decision tree model. The Performance operator was used to assess the extent to which the predicted classes matched with the observed classes for the testing dataset. Both the Apply Model and Performance operators were placed within the “testing panel” of the Cross Validation operator.

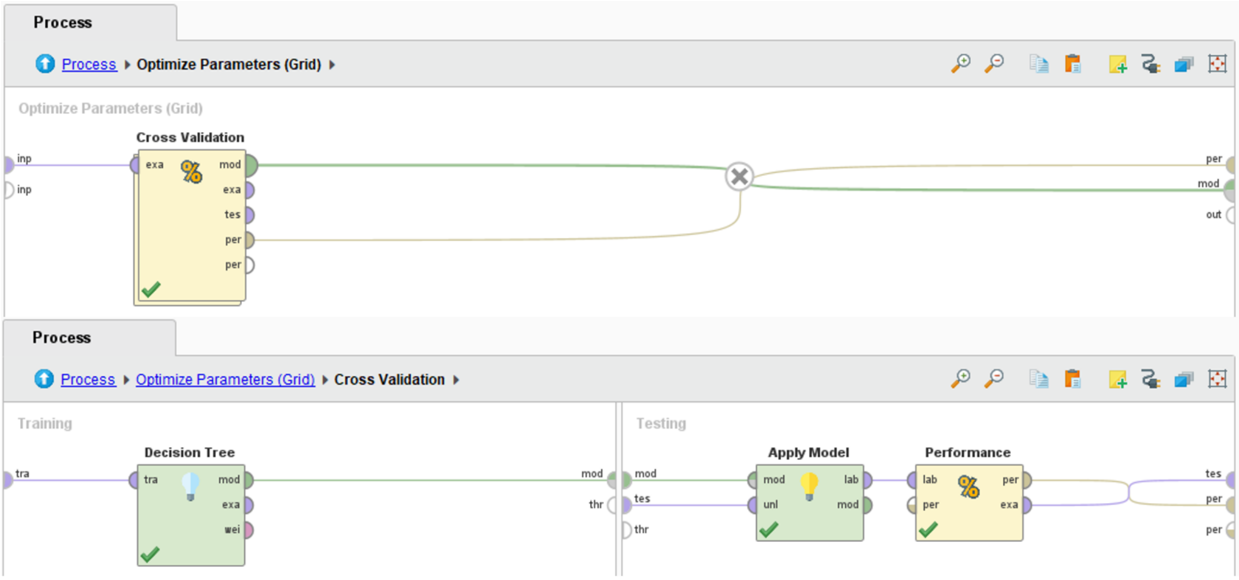


Figure 3. Modeling and optimization.

The Cross Validation operator was placed within the Optimize Parameters (Grid) operator to handle optimization activities. The Optimize Parameters operator enables the hyperparameters of the operators to be examined with a set of values. For instance, the maximal_depth parameter of the Decision Tree operator was allowed to range from 1 to 10 in steps of 1.

Figure 4 shows the overall RapidMiner process across all stages of the instructional framework to complete the decision tree analysis. As noted earlier, the Prep-Explore subprocess contains all operators shown in Figure 2 and the Optimize Parameters (Grid) operators contains all operators shown in Figure 3. The Split Data operator was used to identify the training and testing (holdout) samples from the dataset. The Sample operator was used to balance the examples for the classes represented in the Attrition_Flag variable.

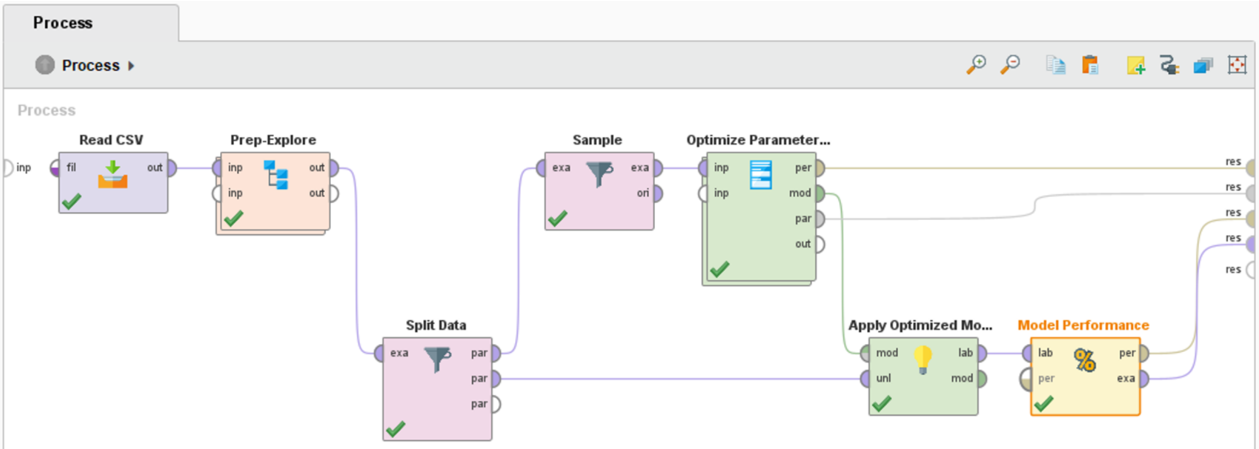


Figure 4. Overall rapidMiner process.

For the Validation stage, the optimized decision tree model was used to score the examples in the training (holdout) sample using the Apply Model operator (renamed as Apply Optimized Model for clarity) in the RapidMiner process. The Performance (Binomial Classification) operator (renamed as Model Performance for clarity) was used to assess the extent to which the optimized model was successful in scoring the examples in the testing (holdout) sample.

3.3. Extension

Once students have been introduced to the application of the framework using one of the analytical methods as described in section 3.2, they can be invited to apply the same approach in implementing other analytical methods. Further, it is possible to apply multiple modeling methods (e.g., classification can be accomplished using k-NN, decision tree, and support vector machine as shown in Table 1) and compare the performance of the different models for the same dataset. Similarly, the hyperparameters across the different analytical methods can be adjusted to finetune the models and the results.

4. DISCUSSION

4.1. Results

Table 2 shows the descriptive statistics of the variables used in the decision tree analysis. The statistics shows that 84% of the examples described Existing customers while the remaining 16% referred to Attrited customers.

Table 2. Descriptive statistics of variables.

Variable	Mean (SD)	Range	Frequency
Attrition_flag			Existing = 7847, Attrited = 1488
Customer_Age	46.38 (8.07)	26 – 73	
Dependent_count	2.34 (1.29)	0 – 5	
Months_on_book	35.97 (8.03)	13 – 56	
Total_relationship_count	3.85 (1.54)	1 – 6	
Credit_limit	7340.70 (7640.35)	1438.30 – 34516.00	
Total_revolving_balance	1157.47 (815.78)	0 – 2517	
Total_trans_amount	4125.44 (3045.50)	510 – 18484	
Total_trans_count	63.66 (22.69)	10 – 139	
Gender			F = 5063, M = 4272
Married			Yes = 4390, No = 4945
Low_Income			Yes = 3374, No = 5961
Graduate			Yes = 3751, No = 5584

Note: N = 9335; SD: Standard deviation.

Figure 5 illustrates examples of the visual analysis that was conducted on the variables. Both charts show the distribution (i.e, frequency) of Existing and Attrited customers on the Y-axis by different variables on the X-axis (gender on the first panel and age on the second panel) also color coded by the Attrition_Flag variable (i.e., blue for Existing customers and green for Attrited customers). The gender chart shows that attrition is slightly higher for females (i.e., 877 cases, also 877 / 5063 = 17.3% of females) relative to males (i.e., 611 cases, also 611 / 4272 = 14.3%). The age chart shows that attrition is generally higher in the age groups from 40 to 50 years relative to the other age groups represented in the data (i.e., 770 of 1488 cases = 51.7%), which represents more than half of all attritions represented in the data.

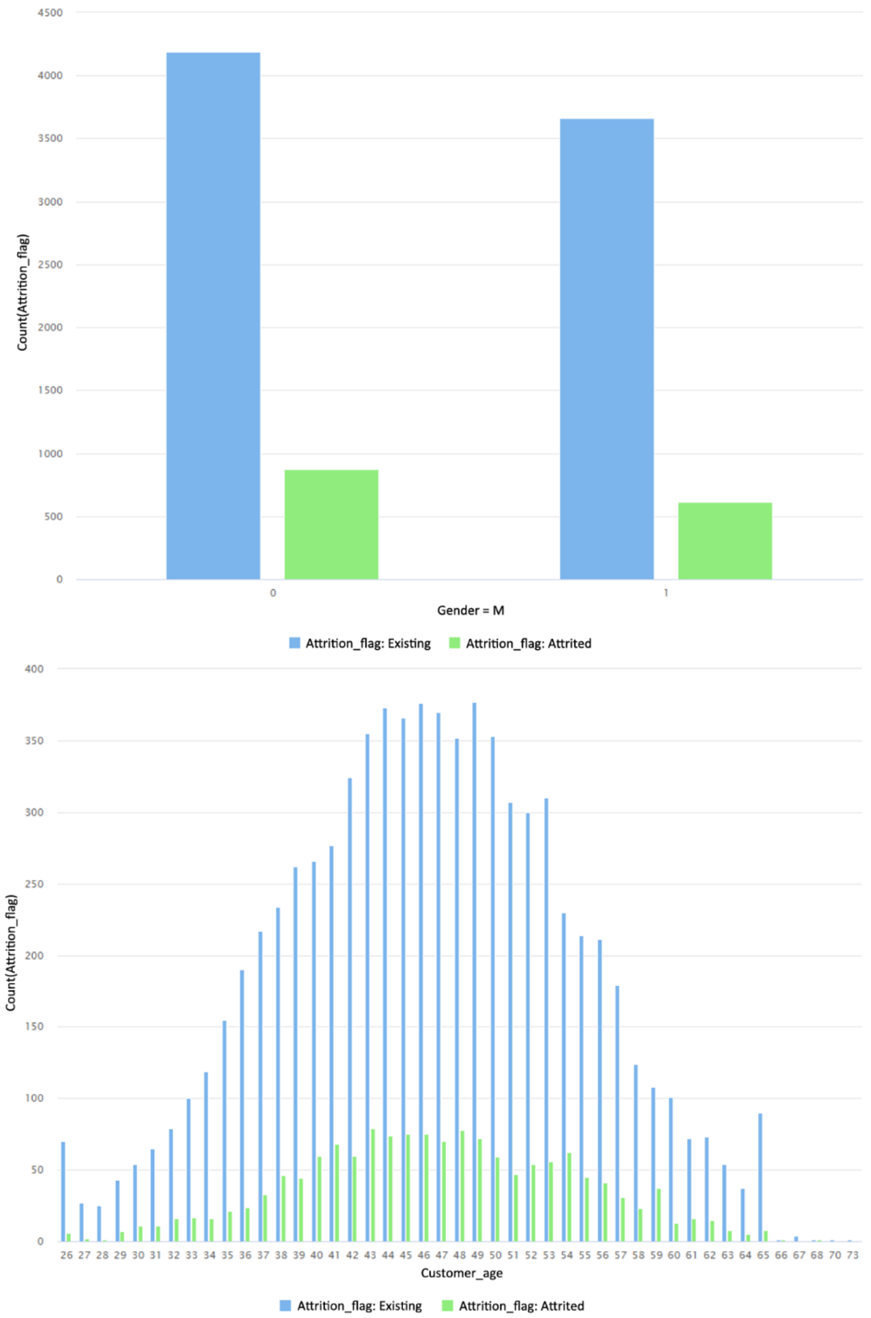
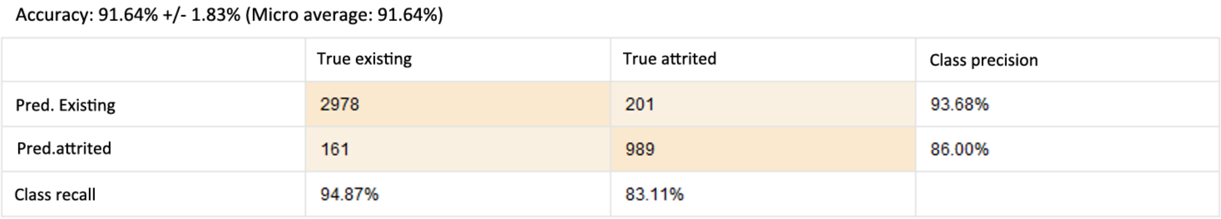


Figure 5. Visual analysis.

For the decision tree analysis, the Split Data operator was used to select choose 80% of the examples for training while the remaining 20% was retained for testing (holdout). The training sample had 7468 examples while the testing (holdout) sample had 1867 examples. The Sample operator was used to select all the Attrited cases in the training sample (i.e., 1190 examples) along with 50% of the Existing cases (i.e., (7468 − 1190) * 0.5 = 3139 examples) to somewhat balance the representation, such that the decision tree model was trained using 4329 examples. The k-fold cross-validation was executed with 90% of the sample (i.e., 4329 * 0.9 = 3896 examples) for training and 10% of the sample (i.e., 4329 * 0.1 = 433 examples) in each iteration. The optimized decision tree model was applied on the testing (holdout) sample, which included 1569 Existing cases and 298 Attrited cases.

The optimized decision tree model had an average accuracy of 91.64% (standard deviation = 1.83%) across the k folds used for cross-validation. Figure 6 shows the confusion matrix. The precision (i.e., the number of actual positive cases out of all predicted positive cases, computed as $Precision = \frac{TP}{TP+FP}$) for the Attrited class was 86% whereas the recall (i.e., the number of positive cases correctly identified by the model, computed as $Recall = \frac{TP}{TP+FN}$) for the Attrited class was 83.11%, in which TP, FP, and FN represent true positives, false positives, and false negatives respectively.



Accuracy: 91.64% +/- 1.83% (Micro average: 91.64%)

	True existing	True attrited	Class precision
Pred. Existing	2978	201	93.68%
Pred.attrited	161	989	86.00%
Class recall	94.87%	83.11%	

Figure 6. Confusion matrix for optimized decision tree.

Figure 7 shows the ROC curve for the cross-validation process used in optimizing the decision tree model. The AUC (Area under curve), i.e., the probability that the model will rank a random positive example higher than a random negative example, was 0.951, which is an indicator of an excellent model.

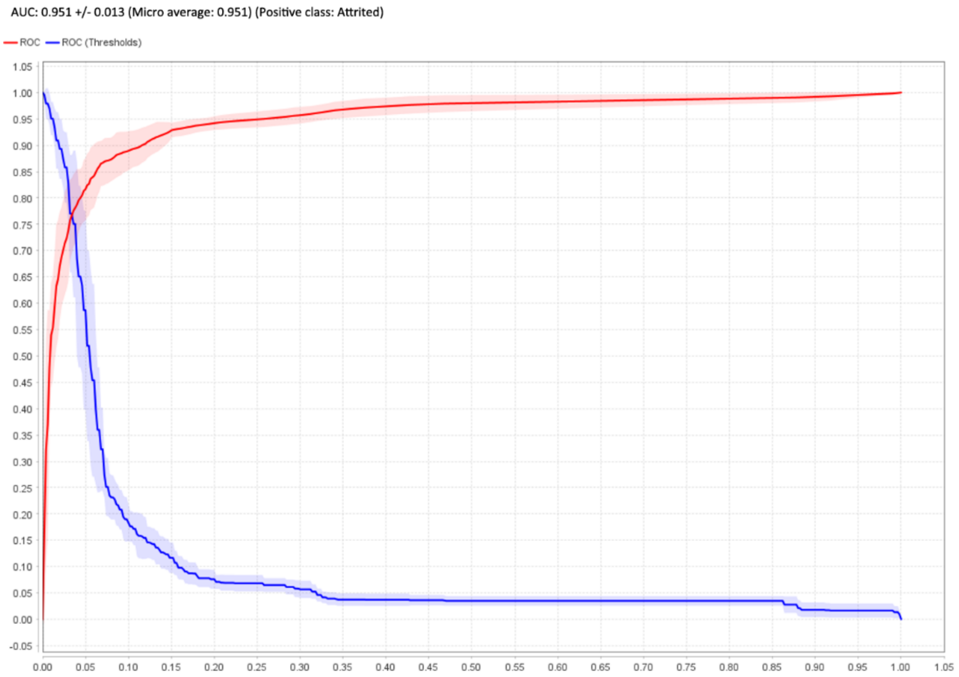
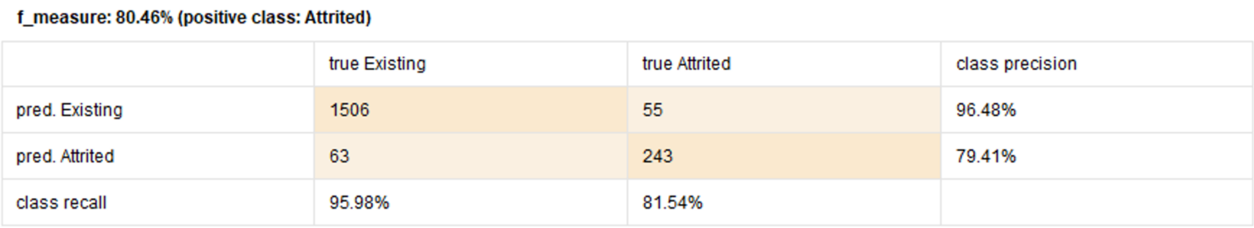


Figure 7. ROC curve for optimized decision tree.

Figure 8 shows the confusion matrix in validating the optimized decision tree on the testing (holdout) sample. The precision was 79.41% whereas the recall was 81.54% for the Attrited class. Since the dataset was imbalanced on the Existing vs. Attrited classes for the label, the F1 score was used to assess performance since the accuracy can be high even if the model correctly guesses the Existing class. The F1 score computed as: $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$ was 80.46% indicating that the model was reasonably good.



f_measure: 80.46% (positive class: Attrited)

	true Existing	true Attrited	class precision
pred. Existing	1506	55	96.48%
pred. Attrited	63	243	79.41%
class recall	95.98%	81.54%	

Figure 8. Confusion matrix for validation.

Figure 9 shows the ROC curve in validating the optimized decision tree. The AUC was 0.950, which is an indicator of an excellent model.

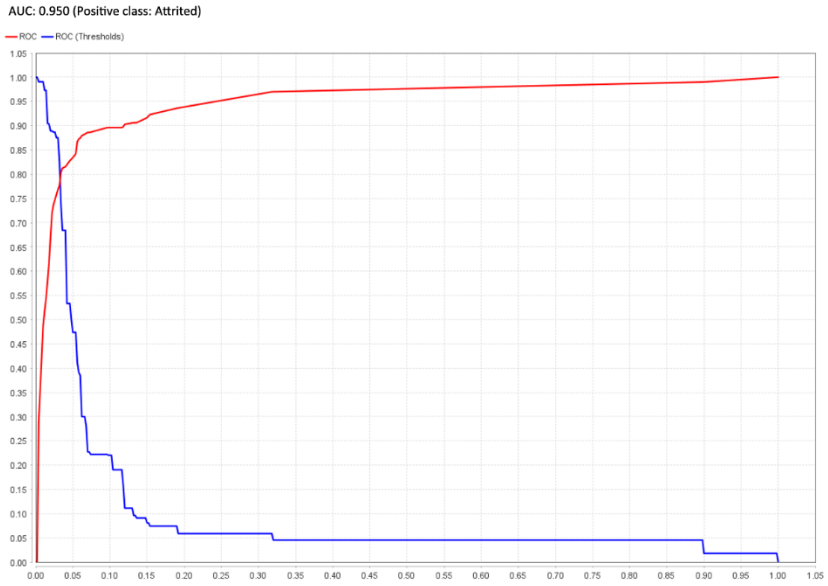


Figure 9. ROC curve for validation.

4.2. Performance

Students were required to complete three assignments for the course using RapidMiner. The requirements were designed such that students will have to begin with the preparation stage of the framework for each assignment. The first assignment dealt with only the descriptive statistics and additionally required use of the exploration stage of the framework. The second assignment required students to apply data mining methods (such as regression or classification) and the use of the modeling and validation stages of the framework as well. The third assignment involved machine learning methods (such as ANN and SVM) and also the use of the optimization stage of the framework. In doing so, students had the flexibility to use any dataset from public domain web sites such as the UCI Machine Learning Repository and Kaggle. This presented each student the opportunity to choose datasets representing topics that are of interest to them or appropriate for their primary disciplines such as marketing, insurance, and finance.

The overall performance of students was reasonably good for all assignments. The descriptive statistics were as follows: first assignment (N = 19, mean = 80.73, standard deviation = 7.57, high = 96, low = 70), second assignment (N = 21, mean = 84.71, standard deviation = 6.99, high = 99, low = 72), and third assignment (N = 20, mean = 86.95, standard deviation = 7.40, high = 100, low = 75).

Students had largely accessed the UCI Machine Learning Repository to obtain datasets describing adult income[6], bank marketing[7], purchase intention[8], and bankruptcy prediction[9]. For the first assignment, students prepared reports based on the descriptive statistics and the visualizations produced by RapidMiner. For the second assignment, students had explored the data using descriptive statistics and visualizations followed by classification methods (i.e., decision trees) and regression methods (i.e., linear regression). They had primarily used the split sample estimation method and examined the classification accuracy for the classification methods and the R-squared statistic for the regression methods for the validation stage. For the third assignment, students used descriptive statistics and visualizations followed by machine learning methods including ANN, k-NN, and ensemble modeling. Students used cross-validation methods for the validation stage and also adjusted parameters such as number of nodes for ANN, number of neighbors for k-NN, and number of levels for decision trees.

4.3. Reflection

The instructional framework was effective in helping students learn the concepts of business analytics and apply them using RapidMiner. The stages embedded in the framework were useful in student learning as they outlined specific goals and tasks that students can focus on in dealing with datasets. Students became adept at building complex workflows as they spent time with the dataset and guided by the framework. Although the specific workflows used by students are not shown here to keep the overall presentation clear, it must be noted that they employed advanced operators such as Bagging (with Random Forest) and Boosting (with AdaBoost) with the Cross Validation operator. They had also applied the Multiply operator to direct different copies of the same dataset to be simultaneously fed into different operators such as Bagging, Boosting, and Voting. Students had also independently learned and applied other operators not shown in Table 1 in their RapidMiner workflows. Examples of such operators include subprocess operator (to streamline the workflow visual by organizing multiple operators into a subprocess), Read C4.5 and Read ARFF, Real to Integer (for data type conversion), Correlation Matrix, Deep Learning (i.e., deep neural network), Logistic Regression, and Weight by Information Gain.

Based on student performance, it is possible to conclude that the framework was useful in imparting business analytics knowledge and skills to students. One student stated: “I have learned to analyze data identifying the correct operators to apply to datasets, the purpose of classification and regression methods for data analysis, and how to analyze the data.” Another student stated: “The idea of cleansing and preparing data before running analysis on it was foreign to me, and very interesting to learn about. The data analysis itself was the most difficult part, and also the most rewarding part. Taking the time to set up the models and go through the descriptive statistics took the most time but are essential for proper and effective analysis.” Another student stated: “I learned how important it is to pay attention to the performance metrics of the model you have created, and being willing to try different

[6] https://archive.ics.uci.edu/dataset/2/adult

[7] https://archive.ics.uci.edu/dataset/222/bank+marketing

[8] https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

[9] https://archive.ics.uci.edu/dataset/365/polish+companies+bankruptcy+data

variations of variables and techniques to produce a better model. It is also helpful to try different models on the same dataset to see which one tends to perform better than the others."

5. CONCLUSION

An instructional framework to train business students in business analytics concepts and applications was developed and applied in a business analytics course. The framework includes five stages that takes students through the journey of preparing and exploring data, modeling, optimizing hyperparameters, and validating models. The decision tree (for classification) has been used as the illustration in this paper; however, any analytical method identified in Table 1 can be taught using the framework since the data analysis will have to proceed through the same five stages introduced in the framework. While it was applied with RapidMiner as the software package, the framework has the potential for application in business analytics contexts that may use other software packages.

REFERENCES

Ahmad, Z., Yaacob, S., Ibrahim, R., & Wan Fakhruddin, W. F. (2022). *The review for visual analytics methodology.* Paper presented at the International Congress on Human-Computer Interaction, Ankara, Turkey.

Boughaci, D., Alkhawaldeh, A. A. K., & Haddadi, D. (2021). Appropriate machine learning techniques for credit scoring and bankruptcy prediction in banking and finance: A comparative study. *Risk and Decision Analysis, 8*(1–2), 1–20.

Dhar, V., & Bose, I. (2024). Business analytics: A review of the state-of-the-art. *Journal of Business Analytics,* 7(1), 1–20.

Haddadi, D., Boughaci, D., & Alkhawaldeh, A. A. K. (2024). A hybrid machine learning model for market clustering. *Engineering, Technology & Applied Science Research, 14*(6), 18824–18828.

Jeyaraj, A. (2019). Pedagogy for business analytics courses. *Journal of Information Systems Education, 30*(2), 67-83.

Khan, R., Nadeem, A., & Ali, A. (2019). Business analytics: A framework. *International Journal of Computer Technology & Applications, 10*(2), 102-108.

Nguyen, A., Gardner, L., & Sheridan, D. (2020). Data analytics in higher education: An integrated view. *Journal of Information Systems Education, 31*(1), 61-71.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51-59. https://doi.org/10.1089/big.2013.1508

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217-222.

Sivri, M., & Ustundag, A. (2024). Applications of business analytics in industry: A review. *Journal of Manufacturing Science and Engineering, 146*(3), 031006.

Zhang, L., Chen, F., & Wei, W. (2020). A foundation course in business analytics: Design and implementation at two universities. *Journal of Information Systems Education, 31*(4), 244-259.