TEACHING A CLASSIFICATION MODULE IN A DATA SCIENCE COURSE FOR UNDERGRADUATE NON-STEM STUDENTS USING PROJECT BASED LEARNING METHODOLOGY

Anna Khalemsky and Yelena Stukalin
Hadassah Academic College, Israel
Academic College of Tel Aviv–Yaffo, Israel
anaha@edu.hac.ac.il

The current paper presents a detailed plan for teaching the classification module as one of the most important parts of the full data science course for non-STEM undergraduate students. Classification is one of the most common data analytics tasks. It is employed in myriad disciplines including marketing, finance, sociology, psychology, education, medicine, and other non-STEM areas. It is, therefore, appropriate to extend carry off with data mining methods to students who will deal with such problems during their professional careers. The overall data science course combines theory and practice and is taught in a "hands-on" format. The main assignment that is run throughout the course is a practical project, which simulates comprehensive research, starting from research questions, through data mining, to interpretation and decision-making. The contents and highlights may vary between different students' majors. We recommend addressing the teaching of the data science course as an interactive and dynamic process.

INTRODUCTION

Most academic data science courses for non-STEM students include various types of data classification implementations. The coverage of topics often depends on the number of credit hours of the course, the background of the students, the specific types of analytical tasks that the students will face in the future. American Statistical Association (ASA) initiated the Guidelines for Assessment and Instruction in Statistics Education (Carver et al., 2016). The document provides recommendations for statistics educators regarding the requirements for the teaching of statistics, real-data integration, and the use of active learning methodologies, such as project-based learning (PBL) for undergraduate students. The exposure of students to real data significantly contributes both to course effectiveness and the interaction between the students and the lecturer (Nolan & Temple Lang, 2015). Preparing students for data analysis is a continuous and comprehensive process (Hawkins et al., 2014). It includes understanding the problem, selecting available and relevant analytical tools and methods, interpreting the results, and drawing conclusions (Garfield & Ben-Zvi, 2008). Familiarity with various alternative methods of data analysis, such as data-mining and visualization methods, significantly enriches the teaching and later also the quality of the research conducted by students (Wagaman, 2016).

Classification is one of the most useful implementations of data analysis, relatively easily interpretable, applicable, and very important for a wide range of fields (Gordon, 1999). Various classification algorithms and their modifications are available for supervised classification tasks and can be easily adopted and compared (Martignon & Laskey, 2019; Irimia-Dieguez et al., 2015). In the present article, we offer a work plan for classification module that is a part of full data science course that integrates theoretical and practical skills and provides a comprehensive methodological data analysis toolkit for undergraduate students with a non-mathematical background. Our primary aim as educators in statistics and data science is to equip non-STEM students with the ability to navigate an analytical process, spanning from defining problems through data analysis to decision-making. This course, with its accessible prerequisites, remarkably fosters motivation and effectiveness. Boasting an array of analytical techniques and user-friendly tools, it enables seamless operations. Students can employ techniques aligned with their majors, leveraging an extensive array of assessment metrics rooted in machine learning. The process encourages dynamic learning, encompassing critical thinking, concept comprehension, classification visualization, and result interpretation. This holistic approach empowers students, making the course a gateway to practical analytical prowess for diverse learners.

Project-based learning methodology is widely used in statistics and data science courses. Giving students practical projects during both basic and advanced courses can significantly enrich the learning process (Gary, 2015). The latter not only increases the students' engagement but also provides students

In: EM Jones (Ed.), *Fostering Learning of Statistics and Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
July 2023, Toronto, Canada. http://iase-web.org

with the opportunity to really participate in the process of implementing the acquired knowledge in practice. Students are involved in the whole process: from thinking about the narrative of the research idea and its potential importance, to processing the data analysis and further to interpreting the results (Saltz & Heckman, 2016). We see a great advantage in having the projects in small groups. Generating ideas, learning process, creating a careful work plan, and mutual help are just some of the benefits. We illustrate the practical projects' instruction by using medical data binary classification example.

Our data science basic course adopts a practical "hands-on" approach. Accessible statistical software is crucial for both lecturers and students, facilitating demonstration and practice. Our courses employ SPSS, STATA, and Weka for data mining. Familiarity with these tools streamlines data classification instruction. Weka's user-friendly interface enables rapid mastery of main modules. Alternatively, the course can center around R or Python, assuming students possess relevant experience. This ensures a dynamic learning experience in mastering essential data science skills.

DATA SCIENCE COURSE PLAN

In the present paper, we propose a data classification course for students with basic statistical knowledge, such as descriptive statistics, statistical inference, and linear regression. The complexity of the analysis depends on both the background of the students and their willingness to broaden their horizons and really reach the "knowledge discovery". The course's exact syllabus should be adjusted to the specific department. It is important to notice that despite the careful planning, we allow for dynamic changes that may accompany this applied course. For example, students may request additional instruction in the use of the software or the analysis of the results. Moreover, the unique characteristics of a database may require ongoing discussion and even additional unplanned explanations.

The data science course for non-STEM undergraduate students consists of 8 modules: 1) Introduction: knowledge as an intangible asset, SECI model of knowledge creation (Nonaka et al., 2000), knowledge discovery steps. Lessons 1-2. 2) Data preparation: data cleaning, data integration, data selection data transformation, scales of measurements. Lessons 3-4. 3) Data sources: data repositories. Lesson 5. 4) Data mining software. Lessons 6-7. 5) Evaluation measures: confusion matrix, accuracy, recall, specificity, precision, F-score, kappa statistic, ROC curve, AUC area. Lesson 8. 6) Data classification: binary classification, multiple groups classification, techniques, parameters, interpretation, implementations in different fields, unbalanced data, robustness. Lessons 9-11. 7) Cluster analysis: techniques, interpretation, latent class analysis. Lessons 12-13. 8) Visualization. Lesson 14.

PROJECT-BASED LEARNING: PRACTICAL PROJECT

The practical project consists of four assignments: approval of the project topic and the database, description of the knowledge discovery problem, classification analysis, oral presentation of main findings.

1. Proposal of a project topic and the database. The first assignment is given after mastering the theoretical material on data preparation and the introduction to data repositories. Students are required to search for a database in one of the existing repositories, such as Kaggle.com (Kaggle: Your Home for Data Science, n.d.), UCI data repository (UCI Machine Learning Repository: Data Sets, n.d.), Amazon datasets (Free Machine Learning Services: AWS, n.d.), etc. Students choose three potential databases and send the links of the databases to the lecturer or the teaching assistant for approval of one of them. At this stage, each student is assigned to a team of 2–3 students. The first assignment does not constitute any part of the final grade but can be returned to the team if none of the databases is not suitable for the project. It takes about 7 minutes for the experienced instructor to approve the database.
2. Knowledge discovery problem. The second task is assigned shortly after the instructor approves the database. Students must complete the following: (1) Define the "knowledge discovery problem" addressable through data analysis. (2) Present a concise literature review, citing relevant academic works on the database. (3) Analyze components of the "knowledge discovery problem," differentiating data, information, and knowledge, explicit and implicit knowledge, and potential innovative contributions. (4) Describe variable characteristics. This assignment contributes fifteen percent to the final grade and usually requires around 15 minutes for assessment and feedback.

3. Classification analysis. Following the supervised learning modules, the third assignment is assigned. Here, students outline data cleaning, feature selection, and analysis procedures. They must summarize pertinent results from diverse classification models, aligning with their chosen data type (binary or multi-class classification). This assignment contributes fifteen percent to the final grade and typically requires around 7 minutes for assessment.
4. Oral presentation. Concluding the classification course, the final assignment entails a ten-minute oral presentation. Students present project highlights, research queries, methodology, and key findings. Entire team participation in answering queries is mandatory. Presentation quality contributes ten percent to the final grade. The instructor records comments during the presentation.

STUDENTS' ASSESSMENT AND EVALUATION

Throughout the duration of the course, students undergo consistent and continuous evaluation and assessment. Assessment is the dynamic and continuous mutual estimation of the learning process (Chris & Duncan, 2013). We guide the students through the assessment process by providing them with theoretical material and practical tips, but students are expected to turn the information into knowledge themselves. Students have 6-8 Moodle quizzes during the semester, their average grade constitutes ten percent of the final course grade. We monitor students' homework accomplishments, their participation in active discussions during classes, and their work on practical projects. In some cases, we initiate personal meetings with students. In other cases, we meet teams to give them feedback and to get their feedback. If time permits, we invite some working groups to evaluate and give advice to other groups, because a practical project is not a competition. The data science course can be managed effectively for 40 students in each group.

Evaluation is the construction of the final grade in the data classification course, and it is definitely a simpler task. Homework (Moodle quizzes) constitutes 10% of the final grade. Specifically, students must submit at least 80% of all quizzes. If students submit more than that, we consider the best grades. Compliance with the requirements for submitting the quizzes is a prerequisite for taking the final exam and, therefore, for passing the course. The practical project constitutes 40% of the final grade. The final exam constitutes 50% of the grade. Students must obtain a passing grade in the final exam in order for the other components to be counted toward the final grade.

CLASSIFICATION MODULE AND 3-RD ASSIGNMENT ILLUSTRATION IN THE PROJECT

Our classification module spans three lessons, each lasting 3 academic hours. Beginning with supervised learning concepts and classification problem definitions, we delve into software practice, parameter tuning's impact on outcomes, and crucial result interpretation. Figure 1 outlines this detailed module plan. Lesson 10 introduces a multivariate dataset from Kaggle, tailored to students' majors. Here, medical data is showcased using the "Cardiovascular Study Dataset." With 15 independent variables and a dichotomous 0-1 dependent variable, the dataset of 3,390 patients contributes to an ongoing Framingham, Massachusetts cardiovascular study, projecting a decade-long coronary heart disease risk prediction (Cardiovascular Study Dataset, n.d.).

We use Weka software to run the algorithms and produce their visual representation and SPSS software (version 27) to analyze the logistic regression output (Table 1). We prepare with students a table of six performance metrics for each model. Table 2 presents the results of the comparison of the four binary classification methods: Logistic regression (model 1); J48 classification tree: unpruned (model 2); Neural network (model 3); Naïve Bayes (model 4). In the F-score, Precision, and Recall columns, the first row in each cell presents the overall accuracy (weighted average of disease and non-disease classification), the second row reflects the disease classification, and the third row shows the non-disease classification. Those teams that choose to work with binary classification, use logistic regression output and are expected to know how to interpret the results.

Table 1: Cardiovascular Study database. Stepwise logistic regression results.

Variable	Estimate B	SE	Wald $\chi 2$	p-value	exp(B)	CI 95% (exp B)
Age	0.066	0.007	84.211	0.000	1.069	1.054,1.084
Sex (male)	-0.489	0.120	16.480	0.000	0.614	0.485,0.777
Cigarettes per day	0.022	0.005	22.643	0.000	1.023	1.013,1.032
Total cholesterol	0.003	0.001	6.472	0.011	1.003	1.001,1.006
SysBP	0.016	0.002	45.081	0.000	1.016	1.011,1.021
Glucose	0.009	0.002	21.526	0.000	1.009	1.005,1.013
Constant	-8.804	0.520	286.956	0.000	0.000	

Figure 1: Classification module

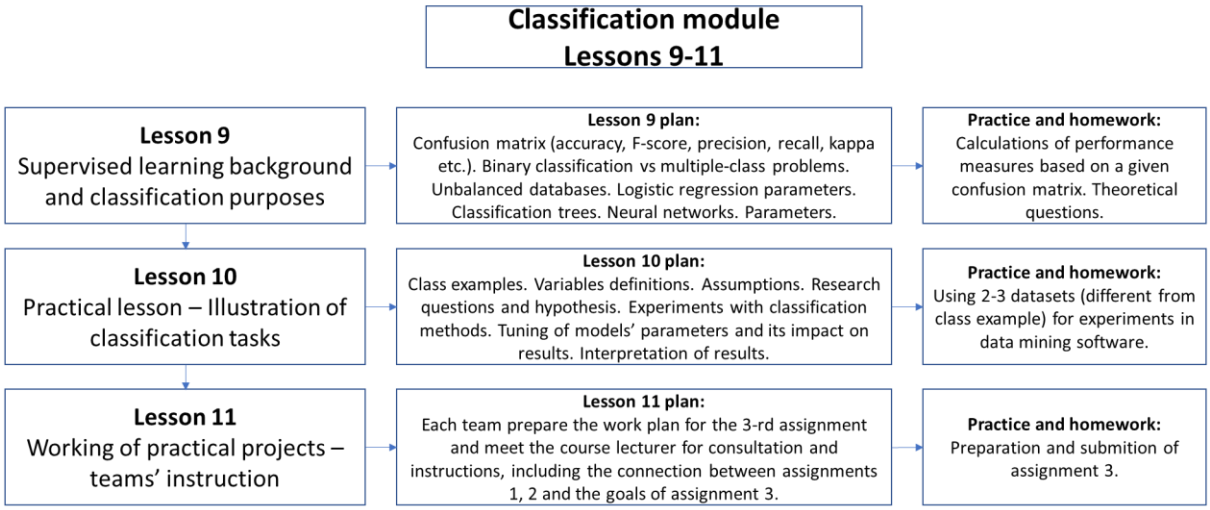

Table 2: Cardiovascular Study dataset. A comparison of the results of the logistic regression, J48 classification tree, neural network, and naïve Bayes binary classification methods.

Model	Accuracy	Kappa	F-score	Precision	Recall	ROC area
Model 1: Logistic Regression	85.6%	0.1214	80.5% 15.0% 92.2%	83.4% 69.4% 85.9%	85.6% 8.4% 99.3%	72.4%
Model 2: J48	83.7%	0.1054	79.8% 16.8% 90.9%	78.4% 36.1% 85.9%	83.7% 11.0% 96.6%	63.9%
Model 3: NN	84.0%	0.1339	80.4% 19.8% 91.1%	79.3% 40.4% 46.2%	84.0% 13.1% 96.6%	66.7%
Model 4: Naïve Bayes	81.6%	0.1883	80.3% 29.1% 89.4%	79.4% 34.7% 87.3%	81.6% 25.0% 91.6%	71.2%

STUDENTS' FEEDBACK

We designed a survey to capture data science students' practical project feedback. Presenting a summary from 43 responses, focusing solely on project-related questions. Our ethics committee endorsed the questionnaire. Table 3 showcases descriptive statistics for ten questions, scored on a 1-10 scale.

Table 3: Descriptive statistics of students' responses

Question #	Average (std)	Median (IQR)	Skewness
1) Do you think that the combination of the practical project that simulates real research enriches the course?	7.49 (2.11)	8 (6,9)	-0.766
2) Do you feel more engaged when there is a practical project in the course?	7.44 (2.12)	8 (7,9)	-1.120
3) Do you prefer a more structured data mining course?	6.7 (2.7)	7 (5,10)	-0.44
4) Do you like the "freedom" of choosing your own topic and data?	6.7 (2.96)	7 (5,9)	-0.616
5) Do you like the "freedom" in decision-making?	6.44 (2.95)	7 (5,9)	-0.568
6) How much do you like teamwork during practical projects?	8 (2.09)	8 (7,10)	-0.752
7) Do you feel free to contact the lecturer for help?	7.6 (2.95)	9 (5,10)	-1.011
8) Do you feel free to ask other students for help?	6.84 (2.53)	7 (5,9)	-0.682
9) Are you using external information sources for learning goals?	7.26 (2.54)	8 (5,10)	-0.825
10) Do you think that learning materials will be useful for you in the future?	5.39 (2.38)	6(4,7)	-0.407

DATABASES FOR DIFFERENT ILLUSTRATION GOALS

In this section we present the list of six different databases from the Kaggle repository that can be used for different teaching purposes, such as providing class examples for students that study in various faculties or for additional exercises. All databases are suitable for classification analysis.

Education: Student performance prediction (*Student Performance Prediction*, n.d.) (can be transformed to binary / multiple class classification). Medicine: Diabetes dataset (*Diabetes Dataset*, n.d.) (binary classification). Business: Company's Ideal Customers / Marketing Strategy (*Company's Ideal Customers | Marketing Strategy*, n.d.) (binary classification). Finance: Company Bankruptcy Prediction (*Company Bankruptcy Prediction*, n.d.) (binary classification). Political sciences: Voters and Non-Voters (*Voters and Non-Voters*, n.d.) (multiple classes). Psychology: Personality classification Data: 16 Personalities (*Personality Classification Data*, n.d.) (multiple classes).

DISCUSSION AND CONCLUSIONS

Analytical software that forms the basis for data mining and knowledge discovery are available and relatively easy to use. The statistics instructor's primary objectives in teaching non-STEM students are to analyze the "knowledge problem", to identify the most suitable methods for a given task, to differentiate within the range of methods, and to teach how to interpret the results in the best way. These objectives correspond to those in the Guidelines for Assessment and Instruction in Statistics Education.

Classification tasks underpin crucial decisions across diverse domains including finance, quality control, environment, epidemiology, and customer relations. Medical students, for instance, analyze patient conditions, leveraging medical history, data, and condition specifics for treatment decisions. Equipping future healthcare professionals with statistical and computational skills is vital. However, instructing data mining to non-STEM students is a nuanced endeavor. While various scientific fields embrace advanced statistical and data mining techniques, students may lack comprehensive backgrounds. Balancing these factors presents a challenge in achieving effective and dynamic learning experiences for non-STEM learners.

In summary, we've designed an innovative data science course tailored for non-STEM undergraduates. Comprising eight modules, each module is structured with a balanced mix: one-third theoretical study, one-third lecturer's demonstrations, and one-third student practice. With access to personal computers during lessons, a hands-on approach is maintained, integrating theory, practical projects, open discussions, and occasional analysis competitions. Notably, the practical project stands as a cornerstone. It empowers students to independently shape their projects, from database selection to methodologies. This facet serves as both active learning and essential training for aspiring professionals.

A common challenge lies in students approaching projects in a rigid, structured manner, contrary to the simulated analysis and decision-making process encouraged. Analysis of 43 questionnaires, part of ongoing research, highlights students' recognition of the benefits of Problem-Based Learning (PBL) and the latitude in dataset and topic selection. However, there remains room for enhancing their understanding of future utility in applying acquired learning outcomes.

REFERENCES

Cardiovascular Study Dataset. (n.d.). Retrieved June 17, 2023 from www.kaggle.com

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G., Velleman, P., Witmer, J., & Wood, B. (2016). Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016. *Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report 2016*.

Chris, B., & Duncan, H. (2013). *Evaluating and Assessing for Learning*. Routledge.

Company Bankruptcy Prediction. (n.d.). Retrieved June 17, 2023 from www.kaggle.com

Company's Ideal Customers | Marketing Strategy. (n.d.). Retrieved June 17, 2023 from www.kaggle.com

Diabetes Dataset. (n.d.). Retrieved June 17, 2023 from www.kaggle.com

Donoghue, T., Voytek, B., & Ellis, S. E. (2021). Teaching Creative and Practical Data Science at Scale. *Journal of Statistics and Data Science Education*, *29*(sup1), S27–S39.

Free Machine Learning Services—AWS. (n.d.). Retrieved May 4, 2022, from https://aws.amazon.com/free/machine-learning

Garfield, J., & Ben-Zvi, D. (2008). *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*. Springer Science & Business Media.

Gary, K. (2015). Project-Based Learning. *Computer*, *48*(9), 98–100.

Gordon, A. D. (1999). *Classification, 2nd Edition*. CRC Press.

Hawkins, A., Jolliffe, F., & Glickman, L. (2014). *Teaching Statistical Concepts*. Routledge.

Irimia-Dieguez, A. I., Blanco-Oliver, A., & Vazquez-Cueto, M. J. (2015). A Comparison of Classification/Regression Trees and Logistic Regression in Failure Models. *Procedia Economics and Finance*, *23*, 9–14.

Kaggle: Your Home for Data Science. (n.d.). Retrieved February 25, 2019 from https://www.kaggle.com/

Martignon, L., & Laskey, K. (2019). Statistical literacy for classification under risk: An educational perspective. *AStA Wirtschafts- Und Sozialstatistisches Archiv*, *13*(3), 269–278.

Nolan, D., & Temple Lang, D. (2015). Explorations in Statistics Research: An Approach to Expose Undergraduates to Authentic Data Analysis. *The American Statistician*, *69*(4), 292–299.

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, Ba and Leadership: A Unified Model of Dynamic Knowledge Creation. *Long Range Planning*, *33*(1), 5–34.

Personality classification Data: 16 Personalities. (n.d.). Retrieved June 17, 2023 from www.kaggle.com

Saltz, J., & Heckman, R. (2016). Big Data science education: A case study of a project-focused introductory course. *Themes in Science and Technology Education*, *8*(2), 85–94.

Student performance prediction. (n.d.). Retrieved June 17, 2023 from www.kaggle.com

UCI Machine Learning Repository: Data Sets. (n.d.). Retrieved December 10, 2016, from https://archive.ics.uci.edu/ml/datasets.html

Voters and Non-Voters. (n.d.). Retrieved June 17, 2023 from https://www.kaggle.com

Wagaman, A. (2016). Meeting Student Needs for Multivariate Data Analysis: A Case Study in Teaching an Undergraduate Multivariate Data Analysis Course. *The American Statistician*, *70*(4), 405–412.