מבט אל מאחורי הקלעים בכריית נתונים בחינוך (מאמר קצר)

יעל פלדמן-מגור
מכון ויצמן למדע,
האוניברסיטה הפתוחה
yael.feldman-maggor@weizmann.ac.il

שגיב ברהום
האוניברסיטה הפתוחה
sagivba@openu.ac.il

רון בלונדר
מכון ויצמן למדע
ron.blonder@weizmann.ac.il

ענבל טובי-ערד
האוניברסיטה הפתוחה
inbaltu@openu.ac.il

Behind the Scenes of Educational Data Mining (Short Paper)

Yael Feldman-Maggor
Weizmann Institute of Science,
The Open University of Israel
yael.feldman-maggor@weizmann.ac.il

Sagiv Barhoom
The Open University of Israel
sagivba@openu.ac.il

Ron Blonder
Weizmann Institute of Science
ron.blonder@weizmann.ac.il

Inbal Tuvi-Arad
The Open University of Israel
inbaltu@openu.ac.il

Abstract

Research based on educational data mining conducted at academic institutions is often limited by the institutional policy with regards to the type of learning management system and the detail level of its activity reports. In many cases, only raw data is provided to the researchers. Such data normally contain numerous fictitious user activities that can create a bias to the activity trends and lead to inaccurate conclusions unless careful strategies for data cleaning, filtering and indexing are applied. Nevertheless, pre-processing stages are not always reported in detail in the scientific literature. As educational data mining and learning analytics methodologies become increasingly popular in educational research, it is important to promote the awareness of researchers and educational policymakers, especially those new to the field, to several pre-processing stages that are essential to create a reliable database prior to any analysis. To address this goal, we suggest here a working process based on four pre-analysis stages: data gathering, data interpretation, database creation, and data organization. These stages were applied to educational data collected from several chemistry courses conducted at two academic institutions. Our results show that adequate pre-processing of the data can prevent major inaccuracies in the research findings, and significantly increase the authenticity and reliability of the conclusions.

Keywords: Learning analytics, Educational Data mining, Learning management system (LMS), Moodle, Higher Education.

תקציר

אחד האתגרים המשפיעים על איכות מחקרים בחינוך המבוססים על כריית נתונים הוא קבלת מסד נתונים מהימן. מחקרים המבוססים על כריית נתונים חינוכיים הנערכים במוסדות אקדמיים מושפעים לרוב ממדיניות המוסד בו מתקיים המחקר, מסוג מערכת ניהול הלמידה ורמת הפירוט של דוחות הפעילות שלה. הנתונים הגולמיים המתקבלים מדוחות אלו לרוב אינם מתאימים לניתוח ישיר ומצריכים עיבוד ובדיקות מקדימות. לדוגמה, הדוחות לרוב כוללים נתונים על משתמשים שאינם לומדים בקורסים ועלולים לכלול אי דיוקים בהקשר לזמני כניסה של משתמשים בשל תקלות טכניות. לפיכך לפני שלב ניתוח הנתונים על החוקרים לבצע ניקוי, מיון וסינון קפדני של הנתונים. עם זאת, השלבים הראשונים של מחקר מבוסס נתונים נשארים לרב "מאחורי הקלעים" בהצגת המחקר. ככל שכריית נתונים חינוכיים הופכת פופולרית יותר ויותר במחקר חינוכי, חשוב לקדם את המודעות של החוקרים וקובעי המדיניות, במיוחד אלה מתוכם החדשים בתחום, למספר שלבי עיבוד מקדים החיוניים ליצירת בסיס נתונים מהימן לפני ניתוח הנתונים. במאמר זה מוצע תהליך עבודה המבוסס על ארבעה שלבים עיקריים: איסוף הנתונים, פרשנות הנתונים, בניית מסד הנתונים וארגון הנתונים כאשר כל שלב מורכב ממספר תתי שלבים. תהליך זה פותח על בסיס כריית נתונים מכמה קורסים בכימיה שנערכו בשני מוסדות אקדמיים. התוצאות מראות כי עיבוד מוקדם של הנתונים יכול למנוע אי-דיוקים גדולים בממצאי המחקר, ולהגדיל משמעותית את מהימנות המסקנות.

מילות מפתח: כריית נתונים, ניקוי נתונים, מערכת לניהול למידה, השכלה גבוהה, הוראת הכימיה.

מבוא

התפתחות האינטרנט הובילה מוסדות אקדמיים רבים ברחבי העולם להציע מספר הולך וגדל של קורסים מתוקשבים. מרבית הקורסים, מעוצבים ונבנים באמצעות מערכת לניהול למידה המהווה את אתר הלמידה האינטרנטי של הקורס. בזמן שהלומד נמצא באינטראקציה עם המערכת נאספים נתונים אודות פעילותו המתוקשבת (Baker & Inventado, 2014) כגון: זמן הפעלת וידאו, זמן כניסה לקבצים באתר הקורס ועוד. נתונים אלו מאפשרים לנתח סטטיסטית התנהגות של מדגם גדול של סטודנטים, אך מדובר בנתונים שדורשים ניתוח מורכב (Bergner, et at 2012). בתחום החינוך, התפתחו שתי דיסציפלינות המתמחות בניתוח נתונים מסוג זה – כריית נתונים חינוכיים – EDM וניתוח למידה-LA, (Baker& Inventado, 2014). במאמר זה לא נתייחס להבדלים ואלגוריתמים הנבנים בשיטות אלו אלא לשלבים המקדימים של איסוף וניקוי הנתונים. היכולת להסיק מסקנות אמינות בעלות משמעות במדעי הלמידה ולבנות מודלים לחיזוי התנהגות לומדים על סמך כריית נתונים תלויה באיכותם ובמהימנותם שכן פרמטרים אלה משפיעים באופן מהותי על פרשנות הנתונים וממצאי המחקר. על כן נחוץ לתעד את אופן הטיפול בנתונים טרם ניתוחם (Pelánek, Rihák, & Papoušek, 2016).

לאחר קבלת הנתונים השלב המכריע בתהליך כריית המידע כרוך בניקוי הנתונים, זיהוי אי דיוקים ונתונים לא עקביים (Romero, C., Romero, J. R., & Ventura, 2014). בנוסף, יש צורך לעבד את הנתונים הגולמיים הזמינים לפורמט מתאים לניתוח (Liñán & Pérez, 2015). במחקרים רבים המפורסמים בספרות המדעית החוקרים מסתפקים בתיאור המשתנים בהם התמקדו זאת, למרות ששלב הניקוי יכול להוות יותר מ-60% מהזמן המוקדש לעיבוד ולניתוח הנתונים (Romero et al, 2014).

מטרות המחקר

במאמר זה מוצגים שלבי עבודה לעיבוד מוקדם של נתונים חינוכיים מקוונים שנאספו מקורסים מתוקשבים לכימיה שנלמדו באוניברסיטה הפתוחה ובמכון ויצמן למדע בין השנים 2016-2019 ומאפייניהם מופיעים בטבלה 1. הנתונים נאספו במסגרת מחקר רחב יותר העוסק בלמידה מקוונת בכימיה. מחקרים קודמים שעסקו בנושא התמקדו בהיבט הטכני (Romero et al, 2014) או במהימנותם (Pelánek et al,2016). ייחודיות המאמר היא התייחסות לשני ההיבטים הללו יחד. בעוד ששלב העיבוד המקדים של נתונים עשוי להיראות מובן מאליו עבור חוקרים בעלי ניסיון קודם בכריית נתונים מטרתנו היא לפנות לחוקרים המתחילים את דרכם בכריית נתונים שיתכן ואינם בקיאים במורכבות התהליך.

טבלה 1. מאפייני הקורסים מהם הופקו הנתונים

שם הקורס	מוסד	מספר סטודנטים ממוצע בסמסטר	מספר הפעלות בשנה	חומרי למידה	מרכיבי הערכה
עולם הכימיה	האוניברסיטה הפתוחה	237	3	ספר לימוד, 6-13 מפגשי הנחייה מתוקשבים או פנים מול פנים בהתאם לקבוצה אליה נרשם הסטודנט, אתר קורס הכולל מצגות, קישורים, הקלטות מפגשי ההנחיה ותרגילים.	הגשת 2-5 מטלות, מבחן.
כימיה כללית א	האוניברסיטה הפתוחה	134	2	ספר לימוד, 6-13 מפגשי הנחייה מתוקשבים או פנים מול פנים בהתאם לקבוצה אליה נרשם הסטודנט, אתר קורס הכולל מצגות, קישורים, הקלטות מפגשי ההנחיה ותרגילים, מפגש מעבדה של 4 שעות.	הגשת 2-8 מטלות, מבחן, השתתפות במפגש מעבדה.
כימיה כללית	האוניברסיטה הפתוחה	144	2	ספר לימוד, 6-13 מפגשי הנחייה מתוקשבים או פנים מול פנים בהתאם לקבוצה אליה נרשם הסטודנט, אתר קורס הכולל מצגות, קישורים, הקלטות מפגשי ההנחיה ותרגילים, מפגש מעבדה של 4 שעות.	הגשת 2-8 מטלות, מבחן, השתתפות בשני מפגשי מעבדה.
מבוא לחומרים וננוטכנולוגיה	מכון ויצמן למדע	40	1	אתר הקורס הכולל בתוכו: 13 שיעורים מתוקשבים, בחנים, משימות על לוחות פדלט, פורום לדיון ושאילת שאלות ומפגש מעבדה של ארבע שעות.	מענה על בחני הקורס, השתתפות פעילה בפורום ובלוח הפדלט בקורס, הגשת משימה מסכמת.

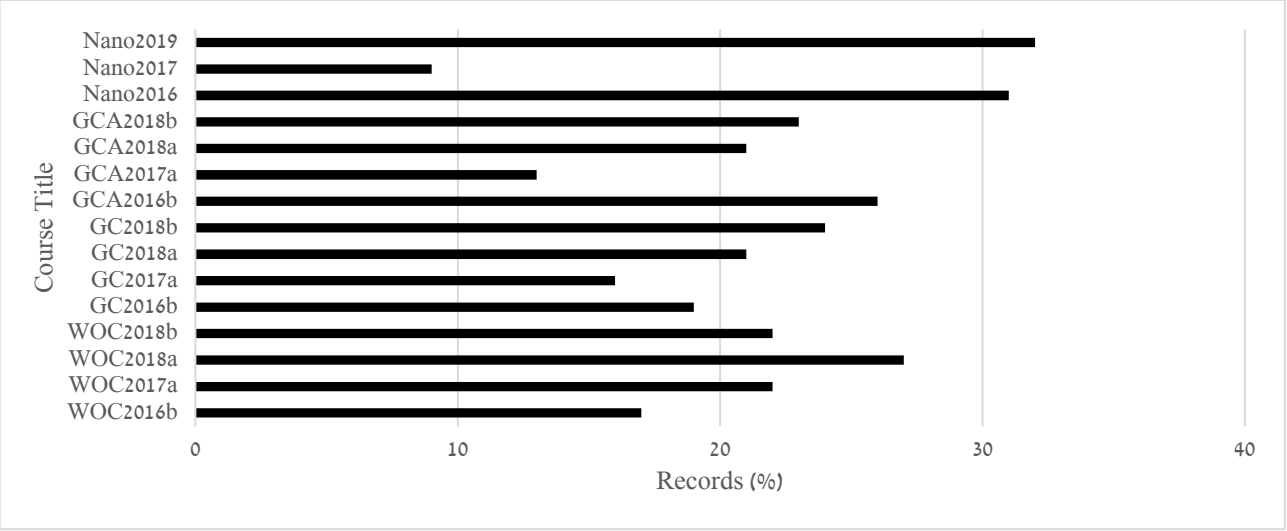
שלבי העבודה

1. **איסוף הנתונים:** הגדרת המקורות מהם ייאספו קבצי הנתונים, ותכנון לוח הזמנים לאיסופם תוך התחשבות בעדכוני תוכנה ומדיניות המוסד האקדמי שעלולים להגביל את הגישה לנתונים. כחוקרים לא הייתה לנו גישה לאחזור הנתונים או שליטה על רמת פירוטם. איסוף הנתונים לאורך מספר שנים דרש שיתוף פעולה עם מחלקות המחשוב ומערכות המידע בכל מוסד אקדמי.

2. **פרשנות הנתונים:** אתרי הקורסים שנחקרו מבוססים על מערכת ניהול למידה מסוג Moodle. הפרשנות של כל פרמטר מקובץ היומן של Moodle המפרט את פעילות המשתמשים אינה תמיד חד משמעית. על מנת ליצור קובץ המפרט את ההסבר לכל פעילות באתר התחברנו לכל קורס כמשתמש אורח וביצענו פעולות שונות במערכת. לאחר מכן בדקנו מיד את האופן בו פעולות אלה נרשמו בקובץ יומן Moodle. בהתבסס על פרשנות זו בנינו קובץ קונפיגורציה ששימש לאחר מכן כתבנית ארגון עבור כלל הנתונים במסד הנתונים של המחקר.

2.1 מדדים לקביעת מהימנות המידע

2.1.1 **סוג המשתמשים:** מערכות שונות לא תמיד מפרידות בין לומדים, מרצים וצוות טכני כך שסטטיסטיקת השימוש עשויה להיות לא מדויקת. איור 1 מציג את אחוז הרשומות של משתמשים שאינם סטודנטים מתוך כלל הרשומות בקורסים שהשתתפו במחקר. כפי שניתן לראות מאיור 1, השגיאה היחסית בין מספר הרשומות יכולה להיות משמעותית ולהגיע עד 32% מסך הפעולות הרשומות. במחקר זה, האבחנה בפעולות הסטודנטים בלבד באמצעות קובץ היומן של Moodle לא הספיקה וההפרדה הושגה על ידי שילובם עם נתוני ההישגים הלימודיים של הסטודנטים.

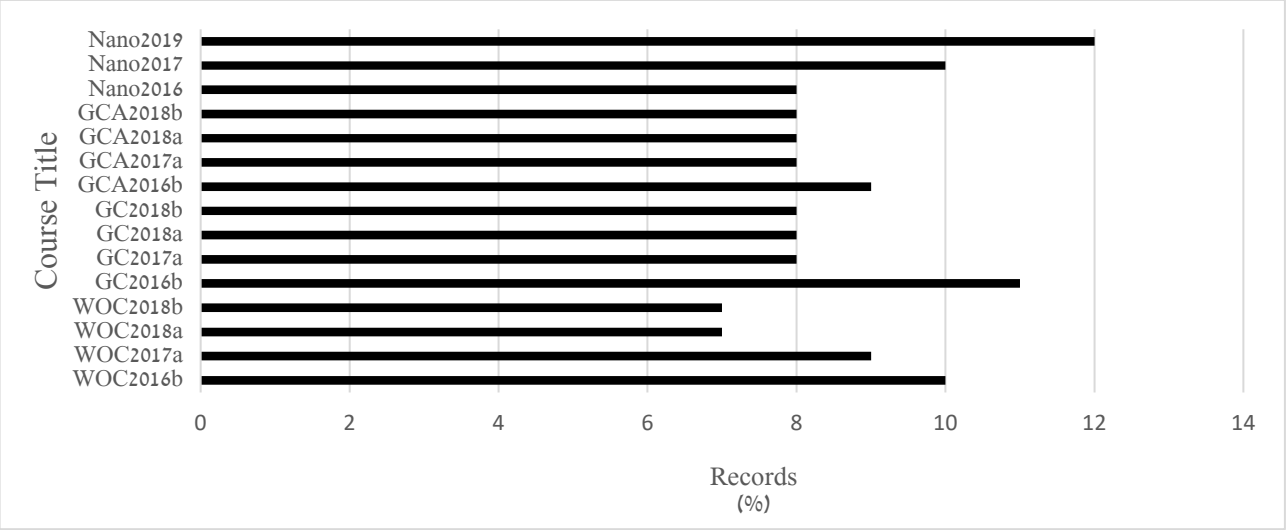


איור 1. אחוז הרשומות שאינן מייצגות סטודנטים בקורסים המשתתפים במחקר. הכותרות בציר ה-Y מייצגות את שמות הקורסים והסמסטר ממנו נלקחו הנתונים. כימיה כללית א: GCA, כימיה כללית: GC, עולם הכימיה: WOC, מבוא לחומרים וננוטכנולוגיה: Nano.

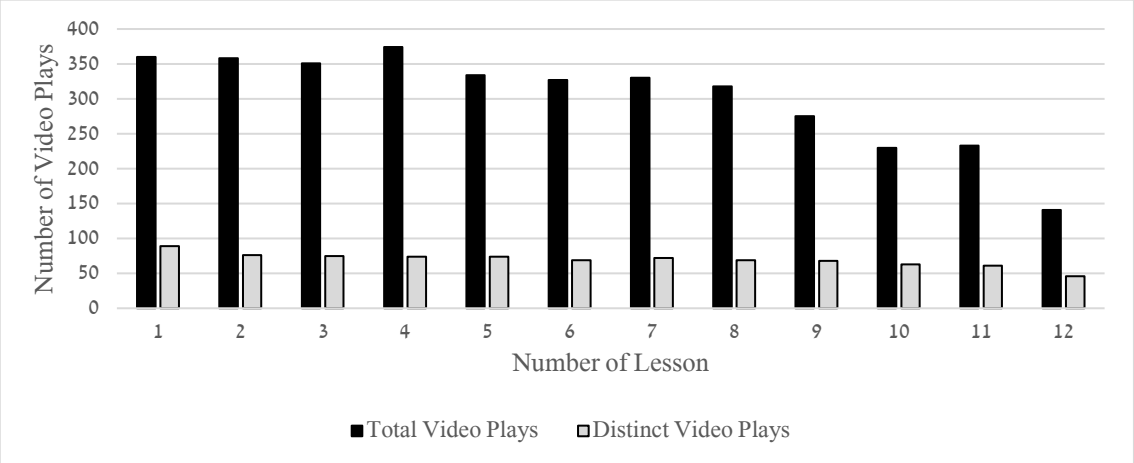
2.1.2 **חותמת זמן:** מציינת את השעה והתאריך המדויקים של התחלת כל פעילות משתמש. בהתייחסות לחותמת הזמן יש לזהות כניסות כפולות של משתמשים לפני שמשתמשים במשתנה זה בשלב ניתוח הנתונים. איור 2 מציג את אחוז הרשומות בכל קורס בהן פעולת המשתמש וחותמת הזמן הייתה זהה. הפער בין מספר הרשומות לפני ואחרי הניקוי נע בין 7% ל-12% בכל קורס בלי קשר לרמת הקורס, מספר הסטודנטים, הסמסטר בו נלמד או מספר הפעילויות המוצעות באתר הקורס. היבט נוספת נוסף בהקשר לזמן הוא חישוב זמן השהייה באתר הקורס. אפשר למדוד את הזמן הכולל בו שהה המשתמש במערכת רק אם המשתמש לחץ על כפתור "יציאה". אם המשתמש עזב את אתר הקורס בלי ללחוץ על "יציאה", זמן היציאה של המשתמש לא יירשם בשרת.

2.1.3 **כתובת IP:** מציינת את המיקום ממנו התחברו משתמשים למערכת ניהול הלמידה. עם זאת, שימוש בשרת פרוקסי או בנתב LAN (למשל, בעת גלישה ברשת פרטית – מקום עבודה או ארגון) מונע פרשנות נכונה של הנתונים.

2.1.4 **ספירת פתיחות של קבצים:** פתיחת קבצים מלמדת אותנו האם המשתמש נכנס לפעילות מקוונת כמו מצגת או הקלטת וידאו. חשוב להבדיל בין מספר הלומדים שפתחו פעילות מסוימת (פתיחות ייחודיות) ממספר הפעמים שהפעילות נפתחה (פתיחות כוללות). דוגמה לכך, מקורס "עולם הכימיה", סמסטר 2018 א מופיעה באיור 3. בשתי צורות הספירה רואים ירידה בפתיחת הקלטות במהלך הסמסטר, אולם בעוד הפתיחות הכוללות ירדו ב-62% הפתיחות הייחודיות ירדו ב-49% בלבד. נתונים דומים התקבלו גם בקורסים אחרים במחקר.



איור 2. אחוז הרשומות הכפולות בכל קורס. הכותרות בציר ה-Y מייצגות את שמות הקורסים והסמסטר ממנו נלקחו הנתונים. כימיה כללית א: GCA, כימיה כללית: GC, עולם הכימיה: WOC, מבוא לחומרים וננוטכנולוגיה: Nano



איור 3. מספר פתיחות ההקלטה הייחודית (עמודות בהירות) למספר הפתיחות הכלליות (עמודות כהות) במהלך הקורס עולם הכימיה, סמסטר 2018א.

3. **יצירת בסיס נתונים:** במחקר זה נאספו נתונים ממקורות שונים לאורך מספר שנים ולכן הקמנו מסד נתונים יחסי. זאת על פי כללי שמירה על פרטיות של וועדת האתיקה של המוסדות האקדמיים ותקנות GDPR (General Data Protection Regulation) האירופאיות (https://gdpr-info.eu/).
4. **ארגון נתונים:** בוצע במחקר זה באמצעות שאילות SQL תוך סינון ושילוב נתונים ממקורות שונים. פרמטרים אלה יהיו נקודת המוצא לבניית מודלים להתנהגות מקוונת בהמשך המחקר.

דיון ומסקנות

במערכות לניהול למידה נשמרת כמות נתונים רבה על פעולות מתוקשבות של הלומדים המאפשרים לחוקרים להשתמש בכריית נתונים על מנת לנתח את דפוסי ההתנהגות של הלומדים בסביבה מתוקשבת. במאמר זה הצגנו את התהליך המקדים לניתוח נתונים מסוג זה בצורה של שלבי עבודה. מתוכם עולים שלושה היבטים חשובים להצלחת המחקר: שיתוף פעולה של החוקרים עם מחלקות אקדמיות שונות, אוטומציה של שלבי העבודה ופרשנות. תהליכי ניקוי וסינון כמו גם בחירה נכונה של פרמטרים לניתוח הם מכריעים להבטחת איכות המחקר ומהימנותו, ועל מנת לאפשר את השוואתו לתוצאות של מחקרים אחרים. בנוסף, מומלץ לאמת את הממצאים באמצעות שיטות מחקר מגוונות, למשל, על ידי שילוב של שיטות מחקר איכותניות לחיזוק הממצאים הכמותיים (Berland, Baker & Blikstein, 2014).

מטרתנו הייתה למזער כללת נתונים לא רלוונטיים ושגויים בשלב ניתוח הנתונים תוך הדגמתם מתוך הקורסים המשתתפים במחקר זה. באמצעות המאמר נרצה להדגיש לחוקרים המתחילים לעסוק בתחום את ההתלבטויות והאתגרים הקיימים עוד לפני הפעלת אלגורתמים או מבחנים סטטיסטיים מורכבים.

תודות

המחקר מומן על ידי משרד החינוך וכן על ידי קרן המחקר של האוניברסיטה הפתוחה (תקציב מספר 507441).

מקורות

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer New York.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-Based Collaborative Filtering Analysis of Student Response Data: Machine-Learnin Item Response Theory. International Educational Data Mining Society. Retrieved October 2019 from: http://files.eric.ed.gov/fulltext/ED537194.pdf

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics Applications to constructionist research. Technology, Knowledge and Learning, 19(1-2), 205-220.

Liñán, L. C., & Pérez, Á. A. J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. International Journal of Educational Technology in Higher Education, 12(3), 98-112.

Pelánek, R., Rihák, J., & Papoušek, J. (2016, April). Impact of data collection on interpretation and evaluation of student models. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 40-47). ACM.

Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In Educational data mining (pp. 29-64). Springer International Publishing.